Comprehensive Introduction to Linear Algebra

PART III - OPERATORS AND TENSORS

Joel G. Broida

S. Gill Williamson

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix}$$

Comprehensive Introduction to Linear Algebra

PART III - OPERATORS AND TENSORS

Joel G. Broida
S. Gill Williamson

Preface (Part II)

This book, Part 3 - Operators and Tensors, covers Chapters 9 through 12 of the book *A Comprehensive Introduction to Linear Algebra* (Addison-Wesley, 1986), by Joel G. Broida and S. Gill Williamson. Selections from Chapters 9 and 10 are covered in most upper division courses in linear algebra. Chapters 11 and 12 introduce multilinear algebra and Hilbert space. The original Preface, Contents and Index are included. Three appendices from the original manuscript are included as well as the original Bibliography. The latter is now (2012) mostly out of date. Wikipedia articles on selected subjects are generally very informative.

Preface (Parts I, II, III)

As a text, this book is intended for upper division undergraduate and beginning graduate students in mathematics, applied mathematics, and fields of science and engineering that rely heavily on mathematical methods. However, it has been organized with particular concern for workers in these diverse fields who want to review the subject of linear algebra. In other words, we have written a book which we hope will still be referred to long after any final exam is over. As a result, we have included far more material than can possibly be covered in a single semester or quarter. This accomplishes at least two things. First, it provides the basis for a wide range of possible courses that can be tailored to the needs of the student or the desire of the instructor. And second, it becomes much easier for the student to later learn the basics of several more advanced topics such as tensors and infinite-dimensional vector spaces from a point of view coherent with elementary linear algebra. Indeed, we hope that this text will be quite useful for self-study. Because of this, our proofs are extremely detailed and should allow the instructor extra time to work out exercises and provide additional examples if desired.

A major concern in writing this book has been to develop a text that addresses the exceptional diversity of the audience that needs to know something about the subject of linear algebra. Although seldom explicitly acknowledged, one of the central difficulties in teaching a linear algebra course to advanced students is that they have been exposed to the basic background material from many different sources and points of view. An experienced mathematician will see the essential equivalence of these points of view, but these same differences seem large and very formidable to the students. An engineering student for example, can waste an inordinate amount of time because of some trivial mathematical concept missing from their background. A mathematics student might have had a concept from a different point of view and not realize the equivalence of that point of view to the one currently required. Although such problems can arise in any advanced mathematics course, they seem to be particularly acute in linear algebra.

To address this problem of student diversity, we have written a very self-contained text by including a large amount of background material necessary for a more advanced understanding of linear algebra. The most elementary of this material constitutes Chapter 0, and some basic analysis is presented in three appendices. In addition, we present a thorough introduction to those aspects of abstract algebra, including groups, rings, fields and polynomials over fields, that relate directly to linear algebra. This material includes both points that may seem "trivial" as well as more advanced background material. While trivial points can be quickly skipped by the reader who knows them already, they can cause discouraging delays for some students if omitted. It is for this reason that we have tried to err on the side of over-explaining concepts, especially when these concepts appear in slightly altered forms. The more advanced reader can gloss over these details, but they are there for those who need them. We hope that more experienced mathematicians will forgive our repetitive justification of numerous facts throughout the text.

A glance at the Contents shows that we have covered those topics normally included in any linear algebra text although, as explained above, to a greater level of detail than other books. Where we differ significantly in content from most linear algebra texts however, is in our treatment of canonical forms (Chapter 8), tensors (Chapter 11), and infinite-dimensional vector spaces (Chapter 12). In particular, our treatment of the Jordan and rational canonical forms in Chapter 8 is based entirely on invariant factors and the

Smith normal form of a matrix. We feel this approach is well worth the effort required to learn it since the result is, at least conceptually, a constructive algorithm for computing the Jordan and rational forms of a matrix. However, later sections of the chapter tie together this approach with the more standard treatment in terms of cyclic subspaces. Chapter 11 presents the basic formalism of tensors as they are most commonly used by applied mathematicians, physicists and engineers. While most students first learn this material in a course on differential geometry, it is clear that virtually all the theory can be easily presented at this level, and the extension to differentiable manifolds then becomes only a technical exercise. Since this approach is all that most scientists ever need, we leave more general treatments to advanced courses on abstract algebra. Finally, Chapter 12 serves as an introduction to the theory of infinite-dimensional vector spaces. We felt it is desirable to give the student some idea of the problems associated with infinite-dimensional spaces and how they are to be handled. And in addition, physics students and others studying quantum mechanics should have some understanding of how linear operators and their adjoints are properly defined in a Hilbert space.

One major topic we have not treated at all is that of numerical methods. The main reason for this (other than that the book would have become too unwieldy) is that we feel at this level, the student who needs to know such techniques usually takes a separate course devoted entirely to the subject of numerical analysis. However, as a natural supplement to the present text, we suggest the very readable "Numerical Analysis" by I. Jacques and C. Judd (Chapman and Hall, 1987).

The problems in this text have been accumulated over 25 years of teaching the subject of linear algebra. The more of these problems that the students work the better. Be particularly wary of the attitude that assumes that some of these problems are "obvious" and need not be written out or precisely articulated. There are many surprises in the problems that will be missed from this approach! While these exercises are of varying degrees of difficulty, we have not distinguished any as being particularly difficult. However, the level of difficulty ranges from routine calculations that everyone reading this book should be able to complete, to some that will require a fair amount of thought from most students.

Because of the wide range of backgrounds, interests and goals of both students and instructors, there is little point in our recommending a particular course outline based on this book. We prefer instead to leave it up to each teacher individually to decide exactly what material should be covered to meet the needs of the students. While at least portions of the first seven chapters should be read in order, the remaining chapters are essentially independent of each other. Those sections that are essentially applications of previous concepts, or else are not necessary for the rest of the book are denoted by an asterisk (*).

Now for one last comment on our notation. We use the symbol ■ to denote the end of a proof, and // to denote the end of an example. Sections are labeled in the format "Chapter.Section," and exercises are labeled in the format "Chapter.Section.Exercise." For example, Exercise 2.3.4 refers to Exercise 4 of Section 2.3, i.e., Section 3 of Chapter 2. Books listed in the bibliography are referred to by author and copyright date.

Contents (Part I, II, III$^{\#}$)

CHAPTER 9

Linear Forms

We are now ready to elaborate on the material of Sections 2.4, 2.5 and 5.1. Throughout this chapter, the field $\mathcal{F}$ will be assumed to be either the real or complex number system unless otherwise noted.

9.1 BILINEAR FUNCTIONALS

Recall from Section 5.1 that the vector space $V^* = L(V, \mathcal{F})$: $V \rightarrow \mathcal{F}$ is defined to be the space of linear functionals on V. In other words, if $\phi \in V^*$, then for every $u, v \in V$ and $a, b \in \mathcal{F}$ we have

$$\phi(au + bv) = a\phi(u) + b\phi(v) \in \mathcal{F} .$$

The space V* is called the **dual space** of V. If V is finite-dimensional, then viewing $\mathcal{F}$ as a one-dimensional vector space (over $\mathcal{F}$), it follows from Theorem 5.4 that dim V* = dim V. In particular, given a basis $\{e_i\}$ for V, the proof of Theorem 5.4 showed that a unique basis $\{\omega^i\}$ for V* is defined by the requirement that

$$\omega^i(e_j) = \delta^i_j$$

where we now again use superscripts to denote basis vectors in the dual space. We refer to the basis $\{\omega^i\}$ for V* as the basis **dual** to the basis $\{e_i\}$ for V.

Elements of V* are usually referred to as **1-forms**, and are commonly denoted by Greek letters such as σ, ϕ, θ and so forth. Similarly, we often refer to the ω^i as **basis 1-forms**.

Since applying Theorem 5.4 to the special case of V* directly may be somewhat confusing, let us briefly go through a slightly different approach to defining a basis for V*.

Suppose we are given a basis $\{e_1, \ldots, e_n\}$ for a finite-dimensional vector space V. Given any set of n scalars ϕ_i, we *define* the linear functionals $\phi \in V^* = L(V, \mathcal{F})$ by $\phi(e_i) = \phi_i$. According to Theorem 5.1, this mapping is unique. In particular, we *define* n linear functionals ω^i by $\omega^i(e_j) = \delta^i{}_j$. Conversely, given any linear functional $\phi \in V^*$, we *define* the n scalars ϕ_i by $\phi_i = \phi(e_i)$. Then, given any $\phi \in V^*$ and any $v = \Sigma v^j e_j \in V$, we have on the one hand

$$\phi(v) = \phi(\Sigma v^i e_i) = \Sigma v^i \phi(e_i) = \Sigma \phi_i v^i$$

while on the other hand

$$\omega^i(v) = \omega^i(\Sigma_j v^j e_j) = \Sigma_j v^j \omega^i(e_j) = \Sigma_j v^j \delta^i{}_j = v^i \;.$$

Therefore $\phi(v) = \Sigma_i \phi_i \omega^i(v)$ for any $v \in V$, and we conclude that $\phi = \Sigma_i \phi_i \omega^i$. This shows that the ω^i span V*, and we claim that they are in fact a basis for V*.

To show that the ω^i are linearly independent, suppose $\Sigma_i a_i \omega^i = 0$. We must show that every $a_i = 0$. But for any $j = 1, \ldots, n$ we have

$$0 = \Sigma_i a_i \omega^i(e_j) = \Sigma_i a_i \delta^i{}_j = a_j$$

which verifies our claim. This completes the proof that $\{\omega^i\}$ forms a basis for V*.

There is another common way of denoting the action of V* on V that is quite similar to the notation used for an inner product. In this approach, the action of the dual basis $\{\omega^i\}$ for V* on the basis $\{e_i\}$ for V is denoted by writing $\omega^i(e_j)$ as

$$\langle \omega^i, e_j \rangle = \delta^i{}_j \;.$$

However, it should be carefully noted that this is *not* an inner product. In particular, the entry on the left inside the bracket is an element of V*, while the entry on the right is an element of V. Furthermore, from the definition of V* as a linear vector space, it follows that $\langle \;,\; \rangle$ is linear in both entries. In other words, if $\phi, \theta \in V^*$, and if $u, v \in V$ and $a, b \in \mathcal{F}$, we have

$$\langle a\phi + b\theta,\, u\rangle = a\langle \phi,\, u\rangle + b\langle \theta,\, u\rangle$$
$$\langle \phi,\, au + bv\rangle = a\langle \phi,\, u\rangle + b\langle \phi,\, v\rangle\ .$$

These relations define what we shall call a **bilinear functional** $\langle\ ,\ \rangle$: $V^* \times V \to \mathcal{F}$ on V^* and V (compare this with definition IP1 of an inner product given in Section 2.4).

We summarize these results as a theorem.

Theorem 9.1 Let $\{e_1, \ldots, e_n\}$ be a basis for V, and let $\{\omega^1, \ldots, \omega^n\}$ be the corresponding dual basis for V* defined by $\omega^i(e_j) = \delta^i{}_j$. Then any $v \in V$ can be written in the forms

$$v = \sum_{i=1}^{n} v^i e_i = \sum_{i=1}^{n} \omega^i(v) e_i = \sum_{i=1}^{n} \langle \omega^i,\, v\rangle e_i$$

and any $\phi \in V^*$ can be written as

$$\phi = \sum_{i=1}^{n} \phi_i \omega^i = \sum_{i=1}^{n} \phi(e_i)\omega^i = \sum_{i=1}^{n} \langle \phi,\, e_i\rangle \omega^i\ .$$

This theorem provides us with a simple interpretation of the dual basis. In particular, since we already know that any $v \in V$ has the expansion $v = \sum v^i e_i$ in terms of a basis $\{e_i\}$, we see that $\omega^i(v) = \langle \omega^i, v\rangle = v^i$ is just the *ith* coordinate of v. In other words, ω^i is just the *ith* coordinate function on V (relative to the basis $\{e_i\}$).

Let us make another observation. If we write $v = \sum v^i e_i$ and recall that $\phi(e_i) = \phi_i$, then (as we saw above) the linearity of ϕ results in

$$\langle \phi, v\rangle = \phi(v) = \phi(\textstyle\sum v^i e_i) = \sum v^i \phi(e_i) = \sum \phi_i v^i$$

which looks very much like the standard inner product on $\mathbb{R}^n$. In fact, if V is an inner product space, we shall see that the components of an element $\phi \in V^*$ may be related in a direct way to the components of some vector in V (see Section 11.10).

It is also useful to note that given any nonzero $v \in V$, there exists $\phi \in V^*$ with the property that $\phi(v) \neq 0$. To see this, we use Theorem 2.10 to first extend v to a basis $\{v, v_2, \ldots, v_n\}$ for V. Then, according to Theorem 5.1, there exists a unique linear transformation $\phi: V \to \mathcal{F}$ such that $\phi(v) = 1$ and $\phi(v_i) = 0$ for $i = 2, \ldots, n$. This ϕ so defined clearly has the desired property. An important consequence of this comes from noting that if $v_1, v_2 \in V$ with $v_1 \neq v_2$, then $v_1 - v_2 \neq 0$, and thus there exists $\phi \in V^*$ such that

$$0 \neq \phi(v_1 - v_2) = \phi(v_1) - \phi(v_2) \ .$$

This proves our next result.

Theorem 9.2 If V is finite-dimensional and $v_1, v_2 \in V$ with $v_1 \neq v_2$, then there exists $\phi \in V^*$ with the property that $\phi(v_1) \neq \phi(v_2)$.

Example 9.1 Consider the space $V = \mathbb{R}^2$ consisting of all column vectors of the form

$$v = \begin{pmatrix} v^1 \\ v^2 \end{pmatrix} \ .$$

Relative to the standard basis we have

$$v = v^1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + v^2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = v^1 e_1 + v^2 e_2 \ .$$

If $\phi \in V^*$, then $\phi(v) = \sum \phi_i v^i$, and we may represent ϕ by the row vector $\phi = (\phi_1, \phi_2)$. In particular, if we write the dual basis as $\omega^i = (a_i, b_i)$, then we have

$$1 = \omega^1(e_1) = (a_1, b_1)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = a_1$$

$$0 = \omega^1(e_2) = (a_1, b_1)\begin{pmatrix} 0 \\ 1 \end{pmatrix} = b_1$$

$$0 = \omega^2(e_1) = (a_2, b_2)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = a_2$$

$$1 = \omega^2(e_2) = (a_2, b_2)\begin{pmatrix} 0 \\ 1 \end{pmatrix} = b_2$$

so that $\omega^1 = (1, 0)$ and $\omega^2 = (0, 1)$. Note that, for example,

$$\omega^1(v) = (1, 0)\begin{pmatrix} v^1 \\ v^2 \end{pmatrix} = v^1$$

as it should. ∕∕

Exercises

1. Find the basis dual to the given basis for each of the following:
   (a) $\mathbb{R}^2$ with basis $e_1 = (2, 1)$, $e_2 = (3, 1)$.
   (b) $\mathbb{R}^3$ with basis $e_1 = (1, -1, 3)$, $e_2 = (0, 1, -1)$, $e_3 = (0, 3, -2)$.

2. Let V be the space of all real polynomials of degree ≤ 1. Define $\omega^1, \omega^2 \in V^*$ by
$$\omega^1(f) = \int_0^1 f(x)\,dx \qquad \text{and} \qquad \omega^2(f) = \int_0^2 f(x)\,dx \ .$$
   Find a basis $\{e_1, e_2\}$ for V that is dual to $\{\omega^1, \omega^2\}$.

3. Let V be the vector space of all polynomials of degree ≤ 2. Define the linear functionals $\omega^1, \omega^2, \omega^3 \in V^*$ by
$$\omega^1(f) = \int_0^1 f(x)\,dx, \qquad \omega^2(f) = f'(1), \qquad \omega^3(f) = f(0)$$
   where $f'(x)$ is the usual derivative of $f(x)$. Find the basis $\{e_i\}$ for V which is dual to $\{\omega^i\}$.

4. (a) Let $u, v \in V$ and suppose that $\phi(u) = 0$ implies $\phi(v) = 0$ for all $\phi \in V^*$. Show that $v = ku$ for some scalar k.
   (b) Let $\phi, \sigma \in V^*$ and suppose that $\phi(v) = 0$ implies $\sigma(v) = 0$ for all $v \in V$. Show that $\sigma = k\phi$ for some scalar k.

5. Let $V = \mathcal{F}[x]$, and for $a \in \mathcal{F}$, define $\phi_a: V \to \mathcal{F}$ by $\phi_a(f) = f(a)$. Show that:
   (a) ϕ_a is linear, i.e., that $\phi_a \in V^*$.
   (b) If $a \neq b$, then $\phi_a \neq \phi_b$.

6. Let V be finite-dimensional and W a subspace of V. If $\phi \in W^*$, prove that ϕ can be extended to a linear functional $\Phi \in V^*$, i.e., $\Phi(w) = \phi(w)$ for all $w \in W$.

9.2 DOUBLE DUALS AND ANNIHILATORS

We now discuss the similarity between the dual space and inner products. To elaborate on this relationship, let V be finite-dimensional over the real field $\mathbb{R}$ with an inner product $\langle\ ,\ \rangle: V \times V \to \mathcal{F}$ defined on it. (There should be no confusion between the inner product on V and the action of a bilinear functional on $V^* \times V$ because both entries in the inner product expressions are elements of V.) In fact, throughout this section we may relax our definition of inner product somewhat as follows. Referring to our definition in Section 2.4, we keep properties (IP1) and (IP2), but instead of (IP3) we require that if $u \in V$ and $\langle u, v\rangle = 0$ for *all* $v \in V$, then $u = 0$. Such an inner product is said to be **nondegenerate**. The reader should be able to see easily that (IP3) implies nondegeneracy, and hence all inner products we have used so far in this book have been nondegenerate. (In Section 11.10 we will see an example of an inner product space with the property that $\langle u, u\rangle = 0$ for some $u \neq 0$.)

If we leave out the second vector entry in the inner product $\langle u, \;\rangle$, then what we have left is essentially a linear functional on V. In other words, given any $u \in V$, we define a linear functional $L_u \in V^*$ by

$$L_u(v) = \langle u, v\rangle$$

for all $v \in V$. From the definition of a (real) inner product, it is easy to see that this functional is indeed linear. Furthermore, it also has the property that

$$L_{au+bv} = aL_u + bL_v$$

for all $u, v \in V$ and $a, b \in \mathcal{F}$. What we have therefore done is define a linear mapping $L: V \rightarrow V^*$ by $L(u) = L_u$ for all $u \in V$. Since the inner product is nondegenerate, we see that if $u \neq 0$ then $L_u(v) = \langle u, v\rangle$ can not vanish for all $v \in V$, and hence $L_u \neq 0$. This means that $\text{Ker } L = \{0\}$, and hence the mapping must be one-to-one (Theorem 5.5). But both V and V* are of dimension n, and therefore this mapping is actually an isomorphism of V onto V*. This proves our next theorem.

Theorem 9.3 Let V be finite-dimensional over $\mathbb{R}$, and assume that V has a nondegenerate inner product defined on it. Then the mapping $u \mapsto L_u$ is an isomorphism of V onto V*.

Looking at this isomorphism as a mapping from V* onto V, we can reword this theorem as follows.

Corollary Let V be as in Theorem 9.3. Then, given any linear functional $L \in V^*$, there exists a unique $u \in V$ such that $L(v) = \langle u, v\rangle = L_u(v)$ for all $v \in V$. In other words, given any $L \in V^*$, there exists a unique $u \in V$ such that $L_u = L$.

Note that if V is a vector space over $\mathbb{C}$ with the more general Hermitian inner product defined on it, then the definition $L_u(v) = \langle u, v\rangle$ shows that $L_{au} = a^*L_u$, and the mapping $u \mapsto L_u$ is no longer an isomorphism of V onto V*. Such a mapping is not even linear, and is in fact called **antilinear** (or **conjugate linear**). We will return to this more general case later.

Let us now consider vector spaces V and V* over an arbitrary (i.e., possibly complex) field $\mathcal{F}$. Since V* is a vector space, we can equally well define the space of linear functionals on V*. By a procedure similar to that followed above, the expression $\langle\ ,\, u\rangle$ for a fixed $u \in V$ defines a linear functional on V* (note that here $\langle\ ,\ \rangle$ is a bilinear functional and not an inner product). In other words, we define the function $f_u\colon V^* \to \mathcal{F}$ by

$$f_u(\phi) = \langle\phi, u\rangle = \phi(u)$$

for all $\phi \in V^*$. It follows that for all $a, b \in \mathcal{F}$ and $\phi, \omega \in V^*$ we have

$$f_u(a\phi + b\omega) = \langle a\phi + b\omega, u\rangle = a\langle\phi, u\rangle + b\langle\omega, u\rangle = af_u(\phi) + bf_u(\omega)$$

and hence f_u is a linear functional from V* to $\mathcal{F}$. In other words, f_u is in the dual space of V*. This space is called the **double dual** (or **second dual**) of V, and is denoted by V**.

Note that Theorem 9.3 shows us that V* is isomorphic to V for any finite-dimensional V, and hence V* is also finite-dimensional. But then applying Theorem 9.3 again, we see that V** is isomorphic to V*, and therefore V is isomorphic to V**. Our next theorem verifies this fact by explicit construction of an isomorphism from V onto V**.

Theorem 9.4 Let V be finite-dimensional over $\mathcal{F}$, and for each $u \in V$ define the function $f_u\colon V^* \to \mathcal{F}$ by $f_u(\phi) = \phi(u)$ for all $\phi \in V^*$. Then the mapping f: $u \mapsto f_u$ is an isomorphism of V onto V**.

Proof We first show that the mapping f: $u \mapsto f_u$ defined above is linear. For any $u, v \in V$ and $a, b \in \mathcal{F}$ we see that

$$\begin{aligned} f_{au+bv}(\phi) &= \langle\phi, au + bv\rangle \\ &= a\langle\phi, u\rangle + b\langle\phi, v\rangle \\ &= af_u(\phi) + bf_v(\phi) \\ &= (af_u + bf_v)(\phi)\ . \end{aligned}$$

Since this holds for all $\phi \in V^*$, it follows that $f_{au+bv} = af_u + bf_v$, and hence the mapping f is indeed linear (so it defines a vector space homomorphism).

Now let $u \in V$ be an arbitrary nonzero vector. By Theorem 9.2 (with $v_1 = u$ and $v_2 = 0$) there exists a $\phi \in V^*$ such that $f_u(\phi) = \langle\phi, u\rangle \neq 0$, and hence clearly $f_u \neq 0$. Since it is obviously true that $f_0 = 0$, it follows that Ker f = {0}, and thus we have a one-to-one mapping from V into V** (Theorem 5.5).

Finally, since V is finite-dimensional, we see that dim V = dim V* = dim V**, and hence the mapping f must be onto (since it is one-to-one). ∎

The isomorphism f: $u \mapsto f_u$ defined in Theorem 9.4 is called the **natural** (or **evaluation**) **mapping** of V into V**. (We remark without proof that even if V is infinite-dimensional this mapping is linear and injective, but is not surjective.) Because of this isomorphism, we will make the identification V = V** from now on, and hence also view V as the space of linear functionals on V*. Furthermore, if $\{\omega^i\}$ is a basis for V*, then the dual basis $\{e_i\}$ for V will be taken to be the basis for V**. In other words, we may write

$$\omega^i(e_j) = e_j(\omega^i) = \delta^i_j$$

so that

$$\phi(v) = v(\phi) = \Sigma\phi_i v^i \; .$$

Now let S be an arbitrary subset of a vector space V. We call the set of elements $\phi \in V^*$ with the property that $\phi(v) = 0$ for all $v \in S$ the **annihilator** of S, and we denote it by S^0. In other words,

$$S^0 = \{\phi \in V^*\colon \phi(v) = 0 \text{ for all } v \in S\} \; .$$

It is easy to see that S^0 is a subspace of V*. Indeed, suppose that $\phi, \omega \in S^0$, let $a, b \in \mathcal{F}$ and let $v \in S$ be arbitrary. Then

$$(a\phi + b\omega)(v) = a\phi(v) + b\omega(v) = 0 + 0 = 0$$

so that $a\phi + b\omega \in S^0$. Note also that we clearly have $0 \in S^0$, and if $T \subset S$, then $S^0 \subset T^0$.

If we let $\mathcal{S}$ be the linear span of a subset $S \subset V$, then it is easy to see that $\mathcal{S}^0 = S^0$. Indeed, if $u \in \mathcal{S}$ is arbitrary, then there exist scalars $a_1, \dots, a_r$ such that $u = \Sigma a_i v^i$ for some set of vectors $\{v^1, \dots, v^r\} \in S$. But then for any $\phi \in S^0$ we have

$$\phi(u) = \phi(\Sigma a_i v^i) = \Sigma a_i \phi(v^i) = 0$$

and hence $\phi \in \mathcal{S}^0$. Conversely, if $\phi \in \mathcal{S}^0$ then ϕ annihilates every $v \in S$ and hence $\phi \in S^0$. The main conclusion to deduce from this observation is that to find the annihilator of a *subspace* W of V, it suffices to find the linear functionals that annihilate any basis for W (see Example 9.2 below).

Just as we talked about the second dual of a vector space, we may define the space S^{00} in the obvious manner by

$$S^{00} = (S^0)^0 = \{v \in V: \phi(v) = 0 \text{ for all } \phi \in S^0\} \ .$$

This is allowed because of our identification of V and V** under the isomorphism $u \mapsto f_u$. To be precise, note that if $v \in S \subset V$ is arbitrary, then for any $\phi \in S^0$ we have $f_v(\phi) = \phi(v) = 0$, and hence $f_v \in (S^0)^0 = S^{0\,0}$. But by our identification of v and f_v (i.e., the identification of V and V**) it follows that $v \in S^{00}$, and thus $S \subset S^{00}$. If S happens to be subspace of V, then we can in fact say more than this.

Theorem 9.5 Let V be finite-dimensional and W a subspace of V. Then

(a) $\dim W^0 = \dim V - \dim W$.

(b) $W^{00} = W$.

Proof (a) Assume that $\dim V = n$ and $\dim W = m \le n$. If we choose a basis $\{w_1, \ldots, w_m\}$ for W, then we may extend this to a basis

$$\{w_1, \ldots, w_m, v_1, \ldots, v_{n-m}\}$$

for V (Theorem 2.10). Corresponding to this basis for V, we define the dual basis

$$\{\phi^1, \ldots, \phi^m, \theta^1, \ldots, \theta^{n-m}\}$$

for V*. By definition of dual basis we then have $\theta^i(v_j) = \delta^i{}_j$ and $\theta^i(w_j) = 0$ for all w_j. This shows that $\theta^i \in W^0$ for each $i = 1, \ldots, n - m$. We claim that $\{\theta^i\}$ forms a basis for W^0.

Since each θ^i is an element of a basis for V*, the set $\{\theta^i\}$ must be linearly independent. Now let $\sigma \in W^0$ be arbitrary. Applying Theorem 9.1 (and remembering that $w_i \in W$) we have

$$\sigma = \sum_{i=1}^{m} \langle \sigma, w_i \rangle \phi^i + \sum_{j=1}^{n-m} \langle \sigma, v_j \rangle \theta^j = \sum_{j=1}^{n-m} \langle \sigma, v_j \rangle \theta^j \ .$$

This shows that the θ^i also span W^0, and hence they form a basis for W^0. Therefore $\dim W^0 = n - m = \dim V - \dim W$.

(b) Recall that the discussion preceding this theorem showed that $W \subset W^{00}$. To show that $W = W^{00}$, we need only show that $\dim W = \dim W^{00}$. However, since W^0 is a subspace of V* and $\dim V^* = \dim V$, we may apply part (a) to obtain

$$\begin{aligned}\dim W^{00} &= \dim V^* - \dim W^0\\ &= \dim V^* - (\dim V - \dim W)\\ &= \dim W\ .\ \blacksquare\end{aligned}$$

Example 9.2 Let W ⊂ $\mathbb{R}^4$ be the two-dimensional subspace spanned by the (column) vectors $w_1 = (1, 2, -3, 4)$ and $w_2 = (0, 1, 4, -1)$. To find a basis for W^0, we seek $\dim W^0 = 4 - 2 = 2$ independent linear functionals ϕ of the form $\phi(x, y, z, t) = ax + by + cz + dt$ such that $\phi(w_1) = \phi(w_2) = 0$. (This is just $\phi(w) = \sum\phi_i w^i$ where $w = (x, y, z, t)$ and $\phi = (a, b, c, d)$.) This means that we must solve the set of linear equations

$$\begin{aligned}\phi(1, 2, -3, 4) &= a + 2b - 3c + 4t = 0\\ \phi(0, 1, 4, -1) &= \qquad b + 4c - \ \ t = 0\end{aligned}$$

which are already in row-echelon form with c and t as free variables (see Section 3.5). We are therefore free to choose any two distinct sets of values we like for c and t in order to obtain independent solutions.

If we let c = 1 and t = 0, then we obtain a = 11 and b = −4 which yields the linear functional $\phi^1(x, y, z, t) = 11x - 4y + z$. If we let c = 0 and t = 1, then we obtain a = −6 and b = 1 so that $\phi^2(x, y, z, t) = -6x + y + t$. Therefore a basis for W^0 is given by the pair $\{\phi^1, \phi^2\}$. In component form, these basis (row) vectors are simply

$$\begin{aligned}\phi^1 &= (11, -4, 1, 0)\\ \phi^2 &= (-6, 1, 0, 1)\ .\ /\!\!/\end{aligned}$$

This example suggests a general approach to finding the annihilator of a subspace W of $\mathcal{F}^n$. To see this, first suppose that we have m ≤ n linear equations in n unknowns:

$$\sum_{j=1}^n a_{ij}x_j = 0$$

for each i = 1, . . . , m. If we define the m linear functionals ϕ^i by

$$\phi^i(x_1, \dots, x_n) = \sum_{j=1}^n a_{ij}x_j$$

then we see that the solution space of our system of equations is nothing more than the subspace of $\mathcal{F}^n$ that is annihilated by $\{\phi^i\}$. Recalling the material of Section 3.5, we know that the solution space to this system is found by row-reducing the matrix $A = (a_{ij})$. Note also that the row vectors A_i are just the coordinates of the linear functional ϕ^i relative to the basis of $\mathcal{F}^{n*}$ that is dual to the standard basis for $\mathcal{F}^n$.

Now suppose that for each i = 1, . . . , m we are given the vector n-tuple $v_i = (a_{i1}, \dots, a_{in}) \in \mathcal{F}^n$. What we would like to do is find the annihilator of the subspace $W \subset \mathcal{F}^n$ that is spanned by the vectors v_i. From the previous section (and the above example) we know that any linear functional ϕ on $\mathcal{F}^n$ must have the form $\phi(x_1, \dots, x_n) = \sum_{i=1}^n c_i x_i$, and hence the annihilator we seek satisfies the condition

$$\phi(v_i) = \phi(a_{i1}, \dots, a_{in}) = \sum_{j=1}^n a_{ij}c_j = 0$$

for each i = 1, . . . , m. In other words, the annihilator $(c_1, \dots, c_n)$ is a solution of the homogeneous system

$$\sum_{j=1}^n a_{ij}c_j = 0 \quad .$$

Example 9.3 Let $W \subset \mathbb{R}^5$ be spanned by the four vectors

$$\begin{aligned} v_1 &= (2, -2, 3, 4, -1) \qquad & v_2 &= (-1, 1, 2, 5, 2) \\ v_3 &= (0, 0, -1, -2, 3) \qquad & v_4 &= (1, -1, 2, 3, 0) \;. \end{aligned}$$

Then W^0 is found by row-reducing the matrix A whose rows are the basis vectors of W:

$$A = \begin{pmatrix} 2 & -2 & 3 & 4 & -1 \\ -1 & 1 & 2 & 5 & 2 \\ 0 & 0 & -1 & -2 & 3 \\ 1 & -1 & 2 & 3 & 0 \end{pmatrix} .$$

Using standard techniques, the reduced matrix is easily found to be

$$\begin{pmatrix} 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} .$$

This is equivalent to the equations

$$\begin{aligned} c_1 - c_2 \quad - \quad c_4 \quad &= 0 \\ c_3 + 2c_4 \quad &= 0 \\ c_5 &= 0 \end{aligned}$$

and hence the free variables are c_2 and c_4. Note that the row-reduced form of A shows that dim W = 3, and hence dim $W^0 = 5 - 3 = 2$. Choosing $c_2 = 1$ and $c_4 = 0$ yields $c_1 = 1$ and $c_3 = 0$, and hence one of the basis vectors for W^0 is given by $\phi^1 = (1, 1, 0, 0, 0)$. Similarly, choosing $c_2 = 0$ and $c_4 = 1$ results in the other basis vector $\phi^2 = (1, 0, -2, 1, 0)$. ∕∕

Exercises

1. Let U and W be subspaces of V (which may be infinite-dimensional). Prove that:
(a) $(U + W)^0 = U^0 \cap W^0$.
(b) $(U \cap W)^0 = U^0 + W^0$.
Compare with Exercise 2.5.2.

2. Let V be finite-dimensional and W a subspace of V. Prove that W* is isomorphic to V^*/W^0 and (independently of Theorem 9.5) also that

$$\dim W^0 = \dim V - \dim W \ .$$

[*Hint*: Consider the mapping $T: V^* \rightarrow W^*$ defined by $T\phi = \phi_W$ where ϕ_W is the restriction of $\phi \in V^*$ to W. Show that T is a surjective linear transformation and that Ker T = W^0. Now apply Exercise 1.5.11 and Theorems 5.4 and 7.34.]

3. Let V be an n-dimensional vector space. An (n - 1)-dimensional subspace of V is said to be a **hyperspace** (or **hyperplane**). If W is an m-dimensional subspace of V, show that W is the intersection of n - m hyperspaces in V.

4. Let U and W be subspaces of a finite-dimensional vectors space V. Prove that U = W if and only if $U^0 = W^0$.

5. Let $\{e_1, \dots, e_5\}$ be the standard basis for $\mathbb{R}^5$, and let $W \subset \mathbb{R}^5$ be spanned by the three vectors

$$\begin{aligned} w_1 &= e_1 + 2e_2 + e_3 \\ w_2 &= \quad\quad\; e_2 + 3e_3 + 3e_4 + e_5 \\ w_3 &= e_1 + 4e_2 + 6e_3 + 4e_4 + e_5 \ . \end{aligned}$$

Find a basis for W^0.

9.3 THE TRANSPOSE OF A LINEAR TRANSFORMATION

Suppose U and V are vector spaces over a field $\mathcal{F}$, and let U* and V* be the corresponding dual spaces. We will show that any T ∈ L(U, V) induces a linear transformation T* ∈ L(V*, U*) in a natural way. We begin by recalling our discussion in Section 5.4 on the relationship between two bases for a vector space. In particular, if a space V has two bases $\{e_i\}$ and $\{\bar{e}_i\}$, we seek the relationship between the corresponding dual bases $\{\omega^i\}$ and $\{\bar{\omega}^i\}$ for V*. This is given by the following theorem.

Theorem 9.6 Let $\{e_i\}$ and $\{\bar{e}_i\}$ be two bases for a finite-dimensional vector space V, and let $\{\omega^i\}$ and $\{\bar{\omega}^i\}$ be the corresponding dual bases for V*. If P is the transition matrix from the basis $\{e_i\}$ to the basis $\{\bar{e}_i\}$, then $(P^{-1})^T$ is the transition matrix from the $\{\omega^i\}$ basis to the $\{\bar{\omega}^i\}$ basis.

Proof Let dim V = n. By definition of $P = (p_{ij})$ we have

$$\bar{e}_i = \sum_{j=1}^{n} e_j p_{ji}$$

for each i = 1, . . . , n. Similarly, let us define the (transition) matrix $Q = (q_{ij})$ by the requirement that

$$\bar{\omega}_i = \sum_{j=1}^{n} \omega^j q_{ji} \quad .$$

We must show that $Q = (P^{-1})^T$. To see this, first note that the *ith* column of Q is $Q^i = (q_{1i}, \dots, q_{ni})$ and the *jth* row of P^T is $P^T{}_j = (p^T{}_{j1}, \dots, p^T{}_{jn})$. From the definition of dual bases, we then see that

$$\begin{aligned}\delta^i{}_j &= \langle \bar{\omega}_i, \bar{e}_j \rangle = \langle \Sigma_k \omega^k q_{ki}, \Sigma_r e_r p_{rj} \rangle = \Sigma_{k,\,r} q_{ki} p_{rj} \langle \omega^k, e_r \rangle \\ &= \Sigma_{k,\,r} q_{ki} p_{rj} \delta^k{}_r = \Sigma_k q_{ki} p_{kj} = \Sigma_k p^T{}_{jk} q_{ki} \\ &= (P^T Q)_{ji} \quad .\end{aligned}$$

In other words, $P^TQ = I$. Since P is a transition matrix it is nonsingular, and hence this shows that $Q = (P^T)^{-1} = (P^{-1})^T$ (Theorem 3.21, Corollary 4). ∎

Now suppose that T ∈ L(V, U). We define a mapping T*: U* → V* by the rule

$$T^*\phi = \phi \circ T$$

for all ϕ ∈ U*. (The mapping T* is frequently written T^t.) In other words, for any v ∈ V we have

$$(T^*\phi)(v) = (\phi \circ T)(v) = \phi(T(v)) \in \mathcal{F} .$$

To show that $T^*\phi$ is indeed an element of V^*, we simply note that for $v_1, v_2 \in V$ and $a, b \in \mathcal{F}$ we have (using the linearity of T and ϕ)

$$\begin{aligned}(T^*\phi)(av_1 + bv_2) &= \phi(T(av_1 + bv_2)) \\ &= \phi(aT(v_1) + bT(v_2)) \\ &= a\phi(T(v_1)) + b\phi(T(v_2)) \\ &= a(T^*\phi)(v_1) + b(T^*\phi)(v_2)\end{aligned}$$

(this also follows directly from Theorem 5.2). Furthermore, it is easy to see that the mapping T^* is linear since for any $\phi, \theta \in U^*$ and $a, b \in \mathcal{F}$ we have

$$T^*(a\phi + b\theta) = (a\phi + b\theta) \circ T = a(\phi \circ T) + b(\theta \circ T) = a(T^*\phi) + b(T^*\theta) .$$

Hence we have proven the next result.

Theorem 9.7 Suppose $T \in L(V, U)$, and define the mapping $T^*: U^* \to V^*$ by $T^*\phi = \phi \circ T$ for all $\phi \in U^*$. Then $T^* \in L(U^*, V^*)$.

The linear mapping T^* defined in this theorem is called the **transpose** of the linear transformation T. The reason for the name transpose is shown in the next theorem. Note that we make a slight change in our notation for elements of the dual space in order to keep everything as simple as possible.

Theorem 9.8 Let $T \in L(V, U)$ have matrix representation $A = (a_{ij})$ with respect to the bases $\{v_1, \dots, v_m\}$ for V and $\{u_1, \dots, u_n\}$ for U . Let the dual spaces V^* and U^* have the corresponding dual bases $\{\bar{v}^i\}$ and $\{\bar{u}^i\}$. Then the matrix representation of $T^* \in L(U^*, V^*)$ with respect to these bases for U^* and V^* is given by A^T.

Proof By definition of $A = (a_{ij})$ we have

$$Tv_i = \sum_{j=1}^{n} u_j a_{ji}$$

for each $i = 1, \dots, m$. Define the matrix representation $B = (b_{ij})$ of T^* by

$$T^*\overline{u}^i = \sum_{j=1}^{m} \overline{v}^j b_{ji}$$

for each $i = 1, \dots, n$. Applying the left side of this equation to an arbitrary basis vector v_k, we find

$$(T^*\bar{u}^i)v_k = \bar{u}^i(Tv_k) = \bar{u}^i(\Sigma_j u_j a_{jk}) = \Sigma_j \bar{u}^i(u_j)a_{jk} = \Sigma_j \delta^i{}_j a_{jk} = a_{ik}$$

while the right side yields

$$\Sigma_j b_{ji}\bar{v}^j(v_k) = \Sigma_j b_{ji}\delta^j{}_k = b_{ki} \ .$$

Therefore $b_{ki} = a_{ik} = a^T{}_{ki}$, and thus $B = A^T$. ∎

Example 9.4 If $T \in L(V, U)$, let us show that $\text{Ker } T^* = (\text{Im } T)^0$. (Remember that $T^*: U^* \rightarrow V^*$.) Let $\phi \in \text{Ker } T^*$ be arbitrary, so that $0 = T^*\phi = \phi \circ T$. If $u \in U$ is any element in Im T, then there exists $v \in V$ such that $u = Tv$. Hence

$$\phi(u) = \phi(Tv) = (T^*\phi)v = 0$$

and thus $\phi \in (\text{Im } T)^0$. This shows that $\text{Ker } T^* \subset (\text{Im } T)^0$.

Now suppose $\theta \in (\text{Im } T)^0$ so that $\theta(u) = 0$ for all $u \in \text{Im } T$. Then for any $v \in V$ we have

$$(T^*\theta)v = \theta(Tv) \in \theta(\text{Im } T) = 0$$

and hence $T^*\theta = 0$. This shows that $\theta \in \text{Ker } T^*$ and therefore $(\text{Im } T)^0 \subset \text{Ker } T^*$. Combined with the previous result, we see that $\text{Ker } T^* = (\text{Im } T)^0$. ∕∕

Example 9.5 Suppose $T \in L(V, U)$ and recall that r(T) is defined to be the number dim(Im T). We will show that $r(T) = r(T^*)$. From Theorem 9.5 we have

$$\dim(\text{Im } T)^0 = \dim U - \dim(\text{Im } T) = \dim U - r(T)$$

and from the previous example it follows that

$$\text{nul } T^* = \dim(\text{Ker } T^*) = \dim(\text{Im } T)^0 \ .$$

Therefore (using Theorem 5.6) we see that

$$\begin{aligned} r(T^*) &= \dim U^* - \text{nul} T^* = \dim U - \text{nul} T^* = \dim U - \dim(\text{Im} T)^0 \\ &= r(T) \ . \end{aligned}$$

Exercises

1. Suppose $A \in M_{m\mathrm{x}n}(\mathcal{F})$. Use Example 9.5 to give a simple proof that rr(A) = cr(A).

2. Let $V = \mathbb{R}^2$ and define $\phi \in V^*$ by $\phi(x, y) = 3x - 2y$. For each of the following linear transformations $T \in L(\mathbb{R}^3, \mathbb{R}^2)$, find $(T^*\phi)(x, y, z)$:
   (a) $T(x, y, z) = (x + y, y + z)$.
   (b) $T(x, y, z) = (x + y + z, 2x - y)$.

3. If $S \in L(U, V)$ and $T \in L(V, W)$, prove that $(T \circ S)^* = S^* \circ T^*$.

4. Let V be finite-dimensional, and suppose that $T \in L(V)$. Show that the mapping $T \mapsto T^*$ defines an isomorphism of L(V) onto L(V*).

5. Let $V = \mathbb{R}[x]$, suppose $a, b \in \mathbb{R}$ are fixed, and define $\phi \in V^*$ by
$$\phi(f) = \int_a^b f(x)\,dx \ .$$
   If D is the usual differentiation operator on V, find $D^*\phi$.

6. Let $V = M_n(\mathcal{F})$, let $B \in V$ be fixed, and define $T \in L(V)$ by
$$T(A) = AB - BA \ .$$
   If $\phi \in V^*$ is defined by $\phi(A) = \mathrm{Tr}\, A$, find $T^*\phi$.

9.4 BILINEAR FORMS

In order to facilitate our treatment of operators (as well as our later discussion of the tensor product), it is worth generalizing slightly some of what we have done so far in this chapter. Let U and V be vector spaces over $\mathcal{F}$. We say that a mapping $f: U \times V \rightarrow \mathcal{F}$ is **bilinear** if it has the following properties for all u_1, $u_2 \in U$, for all $v_1, v_2 \in V$ and all $a, b \in \mathcal{F}$:

(1) $f(au_1 + bu_2, v_1) = af(u_1, v_1) + bf(u_2, v_1)$.
(2) $f(u_1, av_1 + bv_2) = af(u_1, v_1) + bf(u_1, v_2)$.

In other words, f is bilinear if for each $v \in V$ the mapping $u \mapsto f(u, v)$ is linear, and if for each $u \in U$ the mapping $v \mapsto f(u, v)$ is linear. In the particular case that V = U, then the bilinear map $f: V \times V \rightarrow \mathcal{F}$ is called a **bilinear form** on V. (Note that a bilinear *form* is defined on V × V, while a bilinear *functional* was defined on V* × V.) Rather than write expressions like

$f(u, v)$, we will sometimes write the bilinear map as $\langle u, v\rangle$ if there is no need to refer to the mapping f explicitly. While this notation is used to denote several different operations, the context generally makes it clear exactly what is meant.

We say that the bilinear map $f: U \times V \rightarrow \mathcal{F}$ is **nondegenerate** if $f(u, v) = 0$ for all $v \in V$ implies that $u = 0$, and $f(u, v) = 0$ for all $u \in U$ implies that $v = 0$.

Example 9.6 Suppose $A = (a_{ij}) \in M_n(\mathcal{F})$. Then we may interpret A as a bilinear form on $\mathcal{F}^n$ as follows. In terms of the standard basis $\{e_i\}$ for $\mathcal{F}^n$, any $X \in \mathcal{F}^n$ may be written as $X = \sum x^i e_i$, and hence for all $X, Y \in \mathcal{F}^n$ we define the bilinear form f_A by

$$f_A(X, Y) = \textstyle\sum_{i,\, j} a_{ij} x^i y^j = X^T A Y \ .$$

Here the row vector X^T is the transpose of the column vector X, and the expression X^TAY is just the usual matrix product. It should be easy for the reader to verify that f_A is actually a bilinear form on $\mathcal{F}^n$. ∕∕

Example 9.7 Suppose $\alpha, \beta \in V^*$. Since α and β are linear, we may define a bilinear form $f: V \times V \rightarrow \mathcal{F}$ by

$$f(u, v) = \alpha(u)\beta(v)$$

for all $u, v \in V$. This form is usually denoted by $\alpha \otimes \beta$ and is called the **tensor product** of α and β. In other words, the tensor product of two elements $\alpha, \beta \in V^*$ is defined for all $u, v \in V$ by

$$(\alpha \otimes \beta)(u, v) = \alpha(u)\beta(v) \ .$$

We may also define the bilinear form $g: V \times V \rightarrow \mathcal{F}$ by

$$g(u, v) = \alpha(u)\beta(v) - \alpha(v)\beta(u) \ .$$

We leave it to the reader to show that this is indeed a bilinear form. The mapping g is usually denoted by $\alpha\wedge\beta$, and is called the **wedge product** or the **antisymmetric tensor product** of α and β. In other words

$$(\alpha\wedge\beta)(u, v) = \alpha(u)\beta(v) - \alpha(v)\beta(u) \ .$$

Note that $\alpha\wedge\beta$ is just $\alpha \otimes \beta - \beta \otimes \alpha$. ∕∕

Generalizing Example 9.6 leads to the following theorem.

Theorem 9.9 Given a bilinear map f: $\mathcal{F}^m \times \mathcal{F}^n \to \mathcal{F}$, there exists a unique matrix $A \in M_{m\mathrm{x}n}(\mathcal{F})$ such that $f = f_A$. In other words, there exists a unique matrix A such that $f(X, Y) = X^T AY$ for all $X \in \mathcal{F}^m$ and $Y \in \mathcal{F}^n$.

Proof In terms of the standard bases for $\mathcal{F}^m$ and $\mathcal{F}^n$, we have the column vectors $X = \sum_{i=1}^{m} x^i e_i \in \mathcal{F}^m$ and $Y = \sum_{j=1}^{n} y^j e_j \in \mathcal{F}^n$. Using the bilinearity of f we then have

$$f(X, Y) = f(\textstyle\sum_i x^i e_i , \sum_j y^j e_j) = \sum_{i, j} x^i y^j f(e_i, e_j) \ .$$

If we define $a_{ij} = f(e_i, e_j)$, then we see that our expression becomes

$$f(X, Y) = \textstyle\sum_{i, j} x^i a_{ij} y^j = X^T AY \ .$$

To prove the uniqueness of the matrix A, suppose there exists a matrix A′ such that $f = f_{A'}$. Then for all $X \in \mathcal{F}^m$ and $Y \in \mathcal{F}^n$ we have

$$f(X, Y) = X^T AY = X^T A'Y$$

and hence $X^T(A - A')Y = 0$. Now let $C = A - A'$ so that

$$X^T CY = \textstyle\sum_{i, j} c_{ij} x^i y^j = 0$$

for all $X \in \mathcal{F}^m$ and $Y \in \mathcal{F}^n$. In particular, choosing $X = e_i$ and $Y = e_j$, we find that $c_{ij} = 0$ for every i and j. Thus $C = 0$ so that $A = A'$. ∎

The matrix A defined in this theorem is said to **represent** the bilinear map f relative to the standard bases for $\mathcal{F}^m$ and $\mathcal{F}^n$. It thus appears that f is represented by the mn elements $a_{ij} = f(e_i, e_j)$. It is extremely important to realize that the elements a_{ij} are *defined* by the expression $f(e_i, e_j)$ and, conversely, given a matrix $A = (a_{ij})$, we *define* the expression $f(e_i, e_j)$ by requiring that $f(e_i, e_j) = a_{ij}$. In other words, to say that we are given a bilinear map f: $\mathcal{F}^m \times \mathcal{F}^n \to \mathcal{F}$ *means* that we are given values of $f(e_i, e_j)$ for each i and j. Then, given these values, we can evaluate expressions of the form $f(X, Y) = \sum_{i, j} x^i y^j f(e_i, e_j)$. Conversely, if we are given each of the $f(e_i, e_j)$, then we have defined a bilinear map on $\mathcal{F}^m \times \mathcal{F}^n$.

We denote the set of all bilinear maps on U and V by $\mathcal{B}(U \times V, \mathcal{F})$, and the set of all bilinear forms as simply $\mathcal{B}(V) = \mathcal{B}(V \times V, \mathcal{F})$. It is easy to make $\mathcal{B}(U \times V, \mathcal{F})$ into a vector space over $\mathcal{F}$. To do so, we simply define

$$(af + bg)(u, v) = af(u, v) + bg(u, v)$$

for any $f, g \in \mathcal{B}(U \times V, \mathcal{F})$ and $a, b \in \mathcal{F}$. The reader should have no trouble showing that af + bg is itself a bilinear mapping.

It is left to the reader (see Exercise 9.4.1) to show that the association $A \mapsto f_A$ defined in Theorem 9.9 is actually an isomorphism between $M_{m\text{x}n}(\mathcal{F})$ and $\mathcal{B}(\mathcal{F}^m \times \mathcal{F}^n, \mathcal{F})$. More generally, it should be clear that Theorem 9.9 applies equally well to any pair of finite-dimensional vector spaces U and V, and from now on we shall treat it as such.

Theorem 9.10 Let V be finite-dimensional over $\mathcal{F}$, and let V* have basis $\{\omega^i\}$. Define the elements $f^{ij} \in \mathcal{B}(V)$ by

$$f^{ij}(u, v) = \omega^i(u)\omega^j(v)$$

for all $u, v \in V$. Then $\{f^{ij}\}$ forms a basis for $\mathcal{B}(V)$ which thus has dimension $(\dim V)^2$.

Proof Let $\{e_i\}$ be the basis for V dual to the $\{\omega^i\}$ basis for V*, and define $a_{ij} = f(e_i, e_j)$. Given any $f \in \mathcal{B}(V)$, we claim that $f = \sum_{i, j} a_{ij}f^{ij}$. To prove this, it suffices to show that $f(e_r, e_s) = (\sum_{i, j} a_{ij}f^{ij})(e_r, e_s)$ for all r and s. We first note that

$$\begin{aligned}(\textstyle\sum_{i,j}a_{ij}f^{ij})(e_r, e_s) &= \textstyle\sum_{i,j}a_{ij}\omega^i(e_r)\omega^j(e_s) = \sum_{i,j}a_{ij}\delta^i{}_r\delta^j{}_s = a_{rs}\\ &= f(e_r, e_s)\ .\end{aligned}$$

Since f is bilinear, it follows from this that $f(u, v) = (\sum_{i, j} a_{ij}f^{\,ij})(u, v)$ for all $u, v \in V$ so that $f = \sum_{i, j} a_{ij}f^{ij}$. Hence $\{f^{ij}\}$ spans $\mathcal{B}(V)$.

Now suppose that $\sum_{i, j} a_{ij}f^{ij} = 0$ (note that this 0 is actually an element of $\mathcal{B}(V)$). Applying this to (e_r, e_s) and using the above result, we see that

$$0 = (\textstyle\sum_{i, j} a_{ij}f^{ij})(e_r\,, e_s) = a_{rs}\ .$$

Therefore $\{f^{ij}\}$ is linearly independent and hence forms a basis for $\mathcal{B}(V)$. ∎

It should be mentioned in passing that the functions f^{ij} defined in Theorem 9.10 can be written as the tensor product $\omega^i \otimes \omega^j\colon V \times V \rightarrow \mathcal{F}$ (see Example 9.7). Thus the set of bilinear forms $\omega^i \otimes \omega^j$ forms a basis for the space $V^* \otimes V^*$ which is called the **tensor product** of the two spaces V^*. This remark is not meant to be a complete treatment by any means, and we will return to these ideas in Chapter 11.

We also note that if $\{e_i\}$ is a basis for V and dim V = n, then the matrix A of any $f \in \mathcal{B}(V)$ has elements $a_{ij} = f(e_i, e_j)$, and hence $A = (a_{ij})$ has n^2 independent elements. Thus, dim $\mathcal{B}(V) = n^2$ as we saw above.

Theorem 9.11 Let P be the transition matrix from a basis $\{e_i\}$ for V to a new basis $\{e'_i\}$. If A is the matrix of $f \in \mathcal{B}(V)$ relative to $\{e_i\}$, then $A' = P^TAP$ is the matrix of f relative to the basis $\{e'_i\}$.

Proof Let $X, Y \in V$ be arbitrary. In Section 5.4 we showed that the transition matrix $P = (p_{ij})$ defined by $e'_i = P(e_i) = \sum_j e_j p_{ji}$ also transforms the components of $X = \sum_i x^i e_i = \sum_j x'^j e'_j$ as $x^i = \sum_j p_{ij} x'^j$. In matrix notation, this may be written as $[X]_e = P[X]_{e'}$ (see Theorem 5.17), and hence $[X]_e{}^T = [X]_{e'}{}^T\, P^T$. From Theorem 9.9 we then have

$$f(X, Y) = [X]_e{}^T A[Y]_e = [X]_{e'}{}^T[P]^T A[P][Y]_{e'} = [X]_{e'}{}^T A'[Y]_{e'} \ .$$

Since X and Y are arbitrary, this shows that $A' = P^TAP$ is the unique representation of f in the new basis $\{e'_i\}$. ∎

Just as the transition matrix led to the definition of a similarity transformation, we now say that a matrix B is **congruent** to a matrix A if there exists a nonsingular matrix P such that $B = P^TAP$. It was shown in Exercise 5.2.12 that if P is nonsingular, then $r(AP) = r(PA) = r(A)$. Since P is nonsingular, $r(P) = r(P^T)$, and hence $r(B) = r(P^TAP) = r(AP) = r(A)$. In other words, congruent matrices have the same rank. We are therefore justified in defining the **rank** $r(f)$ of a bilinear form f on V to be the rank of any matrix representation of f. We leave it to the reader to show that f is nondegenerate if and only if $r(f) = \dim V$ (see Exercise 9.4.3).

Exercises

1. Show that the association $A \mapsto f_A$ defined in Theorem 9.9 is an isomorphism between $M_{m \times m}(\mathcal{F})$ and $\mathcal{B}(\mathcal{F}^m \times \mathcal{F}^n, \mathcal{F})$.

2. Let $V = M_{m\text{x}n}(\mathcal{F})$ and suppose $A \in M_m(\mathcal{F})$ is fixed. Then for any $X, Y \in V$ we define the mapping $f_A\colon V \times V \to \mathcal{F}$ by $f_A(X, Y) = \mathrm{Tr}(X^T AY)$. Show that this defines a bilinear form on V.

3. Prove that a bilinear form f on V is nondegenerate if and only if $r(f) = \dim V$.

4. (a) Let $V = \mathbb{R}^3$ and define $f \in \mathcal{B}(V)$ by

$$f(X, Y) = 3x^1y^1 - 2x^1y^2 + 5x^2y^1 + 7x^2y^2 - 8x^2y^3 + 4x^3y^2 - x^3y^3 \;.$$

Write out f(X, Y) as a matrix product $X^T AY$.

(b) Suppose $A \in M_n(\mathcal{F})$ and let $f(X, Y) = X^T AY$ for $X, Y \in \mathcal{F}^n$. Show that $f \in \mathcal{B}(\mathcal{F}^n)$.

5. Let $V = \mathbb{R}^2$ and define $f \in \mathcal{B}(V)$ by

$$f(X, Y) = 2x^1y^1 - 3x^1y^2 + x^2y^2 \;.$$

(a) Find the matrix representation A of f relative to the basis $v_1 = (1, 0)$, $v_2 = (1, 1)$.

(b) Find the matrix representation B of f relative to the basis $\bar{v}_1 = (2, 1)$, $\bar{v}_2 = (1, -1)$.

(c) Find the transition matrix P from the basis $\{v_i\}$ to the basis $\{\bar{v}_i\}$ and verify that $B = P^T AP$.

6. Let $V = M_n(\mathbb{C})$, and for all $A, B \in V$ define

$$f(A, B) = n\,\mathrm{Tr}(AB) - (\mathrm{Tr}\,A)(\mathrm{Tr}\,B) \;.$$

(a) Show that this defines a bilinear form on V.

(b) Let $U \subset V$ be the subspace of traceless matrices. Show that f is degenerate, but that $f_U = f|U$ is nondegenerate.

(c) Let $W \subset V$ be the subspace of all traceless skew-Hermitian matrices A (i.e., $\mathrm{Tr}\,A = 0$ and $A^\dagger = A^{*T} = -A$). Show that $f_W = f|W$ is negative definite, i.e., that $f_W(A, A) < 0$ for all nonzero $A \in W$.

(d) Let $\widetilde{V} \subset V$ be the set of all matrices $A \in V$ with the property that $f(A, B) = 0$ for all $B \in V$. Show that $\widetilde{V}$ is a subspace of V. Give an explicit description of $\widetilde{V}$ and find its dimension.

9.5 SYMMETRIC AND ANTISYMMETRIC BILINEAR FORMS

An extremely important type of bilinear form is one for which $f(u, u) = 0$ for all $u \in V$. Such forms are said to be **alternating**. If f is alternating, then for every $u, v \in V$ we have

$$\begin{aligned} 0 &= f(u+v,\ u+v) \\ &= f(u,\ u)+f(u,\ v)+f(v,\ u)+f(v,\ v) \\ &= f(u,\ v)+f(v,\ u) \end{aligned}$$

and hence

$$f(u,\ v) = -f(v,\ u)\ \ .$$

A bilinear form that satisfies this condition is called **antisymmetric** (or **skew-symmetric**). If we let $v = u$, then this becomes $f(u, u) + f(u, u) = 0$. As long as $\mathcal{F}$ is not of characteristic 2 (see the discussion following Theorem 4.3; this is equivalent to the statement that $1 + 1 \neq 0$ in $\mathcal{F}$), we can conclude that $f(u, u) = 0$. Thus, as long as the base field $\mathcal{F}$ is not of characteristic 2, alternating and antisymmetric forms are equivalent. We will always assume that $1 + 1 \neq 0$ in $\mathcal{F}$ unless otherwise noted, and hence we always assume the equivalence of alternating and antisymmetric forms.

It is also worth pointing out the simple fact that the diagonal matrix elements of any representation of an alternating (or antisymmetric) bilinear form will necessarily be zero. This is because the diagonal elements are given by $a_{ii} = f(e_i, e_i) = 0$.

Theorem 9.12 Let $f \in \mathcal{B}(V)$ be alternating. Then there exists a basis for V in which the matrix A of f takes the block diagonal form

$$A = M \oplus \cdots \oplus M \oplus 0 \oplus \cdots \oplus 0$$

where 0 is the 1 x 1 matrix (0), and

$$M = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\ .$$

Moreover, the number of blocks consisting of the matrix M is just $(1/2)r(f)$.

Proof We first note that the theorem is clearly true if $f = 0$. Next we note that if $\dim V = 1$, then any vector $v_i \in V$ is of the form $v_i = a_i u$ for some basis vector u and scalar a_i. Therefore, for any $v_1, v_2 \in V$ we have

$$f(v_1, v_2) = f(a_1u, a_2u) = a_1a_2f(u, u) = 0$$

so that again $f = 0$. We now assume that $f \neq 0$ and that $\dim V > 1$, and proceed by induction on dim V. In other words, we assume the theorem is true for $\dim V < n$, and proceed to show that it is also true for $\dim V = n$.

Since $\dim V > 1$ and $f \neq 0$, there exist nonzero vectors $u_1, u_2 \in V$ such that $f(u_1, u_2) \neq 0$. Moreover, we can always multiply u_1 by the appropriate scalar so that

$$f(u_1, u_2) = 1 = -f(u_2, u_1) \ .$$

It is also true that u_1 and u_2 must be linearly independent because if $u_2 = ku_1$, then $f(u_1, u_2) = f(u_1, ku_1) = kf(u_1, u_1) = 0$. We can now define the two-dimensional subspace $U \subset V$ spanned by the vectors $\{u_1, u_2\}$. By definition, the matrix $(a_{ij}) \in M_2(\mathcal{F})$ of f restricted to U is given by $a_{ij} = f(u_i, u_j)$, and hence it is easy to see that (a_{ij}) is given by the matrix M defined in the statement of the theorem.

Since any $u \in U$ is of the form $u = au_1 + bu_2$, we see that

$$f(u, u_1) = af(u_1, u_1) + bf(u_2, u_1) = -b$$

and

$$f(u, u_2) = af(u_1, u_2) + bf(u_2, u_2) = a \ .$$

Now define the set

$$W = \{w \in V\colon f(w, u) = 0 \text{ for every } u \in U\} \ .$$

We claim that $V = U \oplus W$ (compare this with Theorem 2.22). To show that $U \cap W = \{0\}$, we assume that $v \in U \cap W$. Then $v \in U$ has the form $v = \alpha u_1 + \beta u_2$ for some scalars α and β. But $v \in W$ so that $0 = f(v, u_1) = -\beta$ and $0 = f(v, u_2) = \alpha$, and hence $v = 0$.

We now show that $V = U + W$. Let $v \in V$ be arbitrary, and define the vectors

$$\begin{aligned}&u = f(v, u_2)u_1 - f(v, u_1)u_2 \in U\\ &w = v - u \ .\end{aligned}$$

If we can show that $w \in W$, then we will have shown that $v = u + w \in U + W$ as desired. But this is easy to do since we have

$$\begin{aligned}f(u, u_1) &= f(v, u_2)f(u_1, u_1) - f(v, u_1)f(u_2, u_1) = f(v, u_1)\\ f(u, u_2) &= f(v, u_2)f(u_1, u_2) - f(v, u_1)f(u_2, u_2) = f(v, u_2)\end{aligned}$$

and therefore we find that

$$\begin{aligned} f(w,\, u_1) &= f(v-u,\, u_1) = f(v,\, u_1) - f(u,\, u_1) = 0 \\ f(w,\, u_2) &= f(v-u,\, u_2) = f(v,\, u_2) - f(u,\, u_2) = 0\ . \end{aligned}$$

These equations show that f(w, u) = 0 for every u ∈ U, and thus w ∈ W. This completes the proof that V = U ⊕ W, and hence it follows that dim W = dim V - dim U = n - 2 < n.

Next we note that the restriction of f to W is just an alternating bilinear form on W, and therefore, by our induction hypothesis, there exists a basis $\{u_3, \ldots, u_n\}$ for W such that the matrix of f restricted to W has the desired form. But the matrix of V is the direct sum of the matrices of U and W, where the matrix of U was shown above to be M. Therefore $\{u_1, u_2, \ldots, u_n\}$ is a basis for V in which the matrix of f has the desired form.

Finally, it should be clear that the rows of the matrix of f that are made up of the portion M ⊕ · · · ⊕ M are necessarily linearly independent (by definition of direct sum and the fact that the rows of M are independent). Since each M contains two rows, we see that r(f) = rr(f) is precisely twice the number of M matrices in the direct sum. ∎

Corollary 1 Any nonzero alternating bilinear form must have even rank.

Proof Since the number of M blocks in the matrix of f is (1/2)r(f), it follows that r(f) must be an even number. ∎

Corollary 2 If there exists a nondegenerate, alternating form on V, then dim V is even.

Proof This is Exercise 9.5.7. ∎

If f ∈ $\mathcal{B}$(V) is alternating, then the matrix elements a_{ij} representing f relative to any basis $\{e_i\}$ for V are given by

$$a_{ij} = f(e_i, e_j) = -f(e_j, e_i) = -a_{ji}\ .$$

Any matrix $A = (a_{ij}) \in M_n(\mathcal{F})$ with the property that $a_{ij} = -a_{ji}$ (i.e., $A = -A^T$) is said to be **antisymmetric**. If we are given any element a_{ij} of an antisymmetric matrix, then we automatically know a_{ji}. Because of this, we say that a_{ij} and a_{ji} are not **independent**. Since the diagonal elements of any such antisymmetric matrix must be zero, this means that the maximum number of independent elements in A is given by $(n^2 - n)/2$. Therefore, the subspace of $\mathcal{B}$(V) consisting of nondegenerate alternating bilinear forms is of dimension n(n - 1)/2.

Another extremely important class of bilinear forms on V is that for which $f(u, v) = f(v, u)$ for all $u, v \in V$. In this case we say that f is **symmetric**, and we have the matrix representation

$$a_{ij} = f(e_i, e_j) = f(e_j, e_i) = a_{ji} \ .$$

As expected, any matrix $A = (a_{ij})$ with the property that $a_{ij} = a_{ji}$ (i.e., $A = A^T$) is said to be **symmetric**. In this case, the number of independent elements of A is $[(n^2 - n)/2] + n = (n^2 + n)/2$, and hence the subspace of $\mathcal{B}(V)$ consisting of symmetric bilinear forms has dimension $n(n + 1)/2$.

It is also easy to prove generally that a matrix $A \in M_n(\mathcal{F})$ represents a symmetric bilinear form on V if and only if A is a symmetric matrix. Indeed, if f is a symmetric bilinear form, then for all $X, Y \in V$ we have

$$X^TAY = f(X, Y) = f(Y, X) = Y^TAX \ .$$

But X^TAY is just a 1 x 1 matrix, and hence $(X^TAY)^T = X^TAY$. Therefore (using Theorem 3.18) we have

$$Y^TAX = X^TAY = (X^TAY)^T = Y^TA^TX^{T\,T} = Y^TA^TX \ .$$

Since X and Y are arbitrary, this implies that $A = A^T$. Conversely, suppose that A is a symmetric matrix. Then for all $X, Y \in V$ we have

$$X^TAY = (X^TAY)^T = Y^TA^TX^{T\,T} = Y^TAX$$

so that A represents a symmetric bilinear form. The analogous result holds for antisymmetric bilinear forms as well (see Exercise 9.5.2).

Note that adding the dimensions of the symmetric and antisymmetric subspaces of $\mathcal{B}(V)$ we find

$$n(n - 1)/2 + n(n + 1)/2 = n^2 = \dim \mathcal{B}(V) \ .$$

This should not be surprising since, for an arbitrary bilinear form $f \in \mathcal{B}(V)$ and any $X, Y \in V$, we can always write

$$f(X, Y) = (1/2)[f(X, Y) + f(Y, X)] + (1/2)[f(X, Y) - f(Y, X)] \ .$$

In other words, any bilinear form can always be written as the sum of a symmetric and an antisymmetric bilinear form.

There is another particular type of form that is worth distinguishing. In particular, let V be finite-dimensional over $\mathcal{F}$, and let f = $\langle\ ,\ \rangle$ be a symmetric bilinear form on V. We define the mapping q: V → $\mathcal{F}$ by

$$q(X) = f(X, X) = \langle X, X\rangle$$

for every $X \in V$. The mapping q is called the **quadratic form associated** with the symmetric bilinear form f. It is clear that (by definition) q is represented by a symmetric matrix A, and hence it may be written in the alternative forms

$$q(X) = X^TAX = \Sigma_{i,\,j}a_{ij}x^ix^j = \Sigma_i a_{ii}(x^i)^2 + 2\Sigma_{i<j}a_{ij}x^ix^j\ .$$

This expression for q in terms of the variables x^i is called the **quadratic polynomial** corresponding to the symmetric matrix A. In the case where A happens to be a diagonal matrix, then $a_{ij} = 0$ for $i \neq j$ and we are left with the simple form $q(X) = a_{11}(x^1)^2 + \cdots + a_{nn}(x^n)^2$. In other words, the quadratic polynomial corresponding to a diagonal matrix contains no "cross product" terms.

While we will show below that every quadratic form has a diagonal representation, let us first look at a special case.

Example 9.8 Consider the *real* quadratic polynomial on $\mathcal{F}^n$ defined by

$$q(Y) = \Sigma_{i,\,j}b_{ij}y^iy^j$$

(where $b_{ij} = b_{ji}$ as usual for a quadratic form). If it happens that $b_{11} = 0$ but, for example, that $b_{12} \neq 0$, then we make the substitutions

$$\begin{aligned} y^1 &= x^1 + x^2 \\ y^2 &= x^1 - x^2 \\ y^i &= x^i \text{ for } i = 3, \dots, n\ . \end{aligned}$$

A little algebra (which you should check) then shows that q(Y) takes the form

$$q(Y) = \Sigma_{i,\,j}\, c_{ij}x^ix^j$$

where now $c_{11} \neq 0$. This means that we can focus our attention on the case $q(X) = \Sigma_{i,\,j}a_{ij}x^ix^j$ where it is assumed that $a_{11} \neq 0$.

Thus, given the real quadratic form $q(X) = \Sigma_{i,\,j}a_{ij}x^ix^j$ where $a_{11} \neq 0$, let us make the substitutions

$$\begin{aligned} x^1 &= y^1 - (1/a_{11})[a_{12}y^2 + \cdots + a_{1n}y^n] \\ x^i &= y^i \quad \text{for each } i = 2, \dots, n \ . \end{aligned}$$

Some more algebra shows that q(X) now takes the form

$$q(x^1, \dots, x^n) = a_{11}(y^1)^2 + q'(y^2, \dots, y^n)$$

where q' is a new quadratic polynomial. Continuing this process, we eventually arrive at a new set of variables in which q has a diagonal representation. This is called **completing the square**. ∕∕

Given any quadratic form q, it is possible to fully recover the values of f from those of q. To show this, let $u, v \in V$ be arbitrary. Then

$$\begin{aligned} q(u+v) &= \langle u+v,\, u+v\rangle \\ &= \langle u,\, u\rangle + \langle u,\, v\rangle + \langle v,\, u\rangle + \langle v,\, v\rangle \\ &= q(u) + 2f(u,\, v) + q(v) \end{aligned}$$

and therefore

$$f(u,\, v) = (1/2)[q(u+v) - q(u) - q(v)] \ .$$

This equation is called the **polar form** of f.

Theorem 9.13 Let f be a symmetric bilinear form on a finite-dimensional space V. Then there exists a basis $\{e_i\}$ for V in which f is represented by a diagonal matrix. Alternatively, if f is represented by a (symmetric) matrix A in one basis, then there exists a nonsingular transition matrix P to the basis $\{e_i\}$ such that P^TAP is diagonal.

Proof Since the theorem clearly holds if either $f = 0$ or $\dim V = 1$, we assume that $f \neq 0$ and $\dim V = n > 1$, and proceed by induction on dim V. If $q(u) = f(u, u) = 0$ for *all* $u \in V$, then the polar form of f shows that $f = 0$, a contradiction. Therefore, there must exist a vector $v_1 \in V$ such that $f(v_1, v_1) \neq 0$. Now let U be the (one-dimensional) subspace of V spanned by v_1, and define the subspace $W = \{u \in V\colon f(u, v_1) = 0\}$. We claim that $V = U \oplus W$.

Suppose $v \in U \cap W$. Then $v \in U$ implies that $v = kv_1$ for some scalar k, and hence $v \in W$ implies $0 = f(v, v_1) = k\ f(v_1, v_1)$. But since $f(v_1, v_1) \neq 0$ we must have $k = 0$, and thus $v = kv_1 = 0$. This shows that $U \cap W = \{0\}$.

Now let $v \in V$ be arbitrary, and define

$$w = v - [f(v, v_1)/f(v_1, v_1)]v_1 \ .$$

Then

$$f(w, v_1) = f(v, v_1) - [f(v, v_1)/f(v_1, v_1)]f(v_1, v_1) = 0$$

and hence $w \in W$. Since the definition of w shows that any $v \in V$ is the sum of $w \in W$ and an element of U, we have shown that $V = U + W$, and hence $V = U \oplus W$.

We now consider the restriction of f to W, which is just a symmetric bilinear form on W. Since $\dim W = \dim V - \dim U = n - 1$, our induction hypothesis shows there exists a basis $\{e_2, \dots, e_n\}$ for W such that $f(e_i, e_j) = 0$ for all $i \neq j$ where $i, j = 2, \dots, n$. But the definition of W shows that $f(e_i, v_1) = 0$ for each $i = 2, \dots, n$, and thus if we define $e_1 = v_1$, the basis $\{e_1, \dots, e_n\}$ for V has the property that $f(e_i, e_j) = 0$ for all $i \neq j$ where now $i , j = 1, \dots, n$. This shows that the matrix of f in the basis $\{e_i\}$ is diagonal. The alternate statement in the theorem follows from Theorem 9.11. ∎

In the next section, we shall show explicitly how this diagonalization can be carried out.

Exercises

1. (a) Show that if f is a nondegenerate, antisymmetric bilinear form on V, then $n = \dim V$ is even.
   (b) Show that there exists a basis for V in which the matrix of f takes the block matrix form

$$\begin{pmatrix} 0 & D \\ -D & 0 \end{pmatrix}$$

   where D is the $(n/2) \times (n/2)$ matrix

$$\begin{pmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \vdots & & \vdots & \vdots \\ 1 & \cdots & 0 & 0 \end{pmatrix} .$$

2. Show that a matrix $A \in M_n(\mathcal{F})$ represents an antisymmetric bilinear form on V if and only if A is antisymmetric.

3. Reduce each of the following quadratic forms to diagonal form:
   (a) $q(x, y, z) = 2x^2 - 8xy + y^2 - 16xz + 14yz + 5z^2$.
   (b) $q(x, y, z) = x^2 - xz + y^2$.
   (c) $q(x, y, z) = xy + y^2 + 4xz + z^2$.
   (d) $q(x, y, z) = xy + yz$.

4. (a) Find all antisymmetric bilinear forms on $\mathbb{R}^3$.
   (b) Find a basis for the space of all antisymmetric bilinear forms on $\mathbb{R}^n$.

5. Let V be finite-dimensional over $\mathbb{C}$. Prove:
   (a) The equation

$$(Ef)(u, v) = (1/2)[f(u, v) - f(v, u)]$$

   for every $f \in \mathcal{B}(V)$ defines a linear operator E on $\mathcal{B}(V)$.
   (b) E is a projection, i.e., $E^2 = E$.
   (c) If $T \in L(V)$, the equation

$$(T^\dagger f)(u, v) = f(Tu, Tv)$$

   defines a linear operator $T^\dagger$ on $\mathcal{B}(V)$.
   (d) $E\,T^\dagger = T^\dagger\,E$ for all $T \in \mathcal{B}(V)$.

6. Let V be finite-dimensional over $\mathbb{C}$, and suppose $f, g \in \mathcal{B}(V)$ are antisymmetric. Show there exists an invertible $T \in L(V)$ such that $f(Tu, Tv) = g(u, v)$ for all $u, v \in V$ if and only if f and g have the same rank.

7. Prove Corollary 2 of Theorem 9.12.

9.6 DIAGONALIZATION OF SYMMETRIC BILINEAR FORMS

Now that we know any symmetric bilinear form f can be diagonalized, let us look at how this can actually be carried out. After this discussion, we will give an example that should clarify everything. (The algorithm that we are about to describe may be taken as an independent proof of Theorem 9.13.) Let the (symmetric) matrix representation of f be $A = (a_{ij}) \in M_n(\mathcal{F})$, and first assume that $a_{11} \neq 0$. For each $i = 2, \ldots, n$ we multiply the *ith* row of A by a_{11}, and then add $-a_{i1}$ times the first row to this new *ith* row. In other words, this combination of two elementary row operations results in $A_i \rightarrow a_{11}A_i - a_{i1}A_1$. Following this procedure for each $i = 2, \ldots, n$ yields the first column of A in the form $A^1 = (a_{11}, 0, \dots, 0)$ (remember that this is a column vector, not a row vector). We now want to put the first row of A into the same form. However, this is easy because A is symmetric. We thus perform exactly the same operations (in the same sequence), but on columns instead of rows, resulting in $A^i \to a_{11}A^i - a_{i1}A^1$. Therefore the first row is also transformed into the form $A_1 = (a_{11}, 0, \dots, 0)$. In other words, this sequence of operations results in the transformed A having the block matrix form

$$\begin{pmatrix} a_{11} & 0 \\ 0 & B \end{pmatrix}$$

where B is a matrix of size less than that of A. We can also write this in the form $(a_{11}) \oplus B$.

Now look carefully at what we did for the case of i = 2. Let us denote the multiplication operation by the elementary matrix E_m, and the addition operation by E_a (see Section 3.8). Then what was done in performing the row operations was simply to carry out the multiplication $(E_aE_m)A$. Next, because A is symmetric, we carried out exactly the same operations but applied to the columns instead of the rows. As we saw at the end of Section 3.8, this is equivalent to the multiplication $A(E_m{}^TE_a{}^T)$. In other words, for i = 2 we effectively carried out the multiplication

$$E_aE_mAE_m{}^TE_a{}^T \ .$$

For each succeeding value of i we then carried out this same procedure, and the final net effect on A was simply a multiplication of the form

$$E_s \cdots E_1AE_1{}^T \cdots E_s{}^T$$

which resulted in the block matrix $(a_{11}) \oplus B$ shown above. Furthermore, note that if we let $S = E_1{}^T \cdots E_s{}^T = (E_s \cdots E_1)^T$, then $(a_{11}) \oplus B = S^TAS$ must be symmetric since $(S^TAS)^T = S^TA^TS = S^TAS$. This means that in fact the matrix B must also be symmetric.

We can now repeat this procedure on the matrix B and, by induction, we eventually arrive at a diagonal representation of A given by

$$D = E_r \cdots E_1AE_1{}^T \cdots E_r{}^T$$

for some set of elementary row transformations E_i. But from Theorems 9.11 and 9.13, we know that $D = P^TAP$, and therefore P^T is given by the product

$e_r \cdot\cdot\cdot e_1(I) = E_r \cdot\cdot\cdot E_1$ of elementary row operations applied to the identity matrix exactly as they were applied to A. It should be emphasized that we were able to arrive at this conclusion only because A is symmetric, thereby allowing each column operation to be the transpose of the corresponding row operation. Note however, that while the order of the row and column operations performed is important within their own group, the associativity of the matrix product allows the column operations (as a group) to be performed independently of the row operations.

We still must take into account the case where $a_{11} = 0$. If $a_{11} = 0$ but $a_{ii} \neq 0$ for some $i > 1$, then we can bring a_{ii} into the first diagonal position by interchanging the *ith* row and column with the first row and column respectively. We then follow the procedure given above. If $a_{ii} = 0$ for every $i = 1, \dots, n$, then we can pick any $a_{ij} \neq 0$ and apply the operations $A_i \rightarrow A_i + A_j$ and $A^i \rightarrow A^i + A^j$. This puts $2a_{ij} \neq 0$ into the *ith* diagonal position, and allows us to proceed as in the previous case (which then goes into the first case treated). (Note also that this last procedure requires that our field is not of characteristic 2 because we assumed that $a_{ij} + a_{ij} = 2a_{ij} \neq 0$.)

Example 9.9 Let us find the transition matrix P such that $D = P^TAP$ is diagonal, with A given by

$$\begin{pmatrix} 1 & -3 & 2 \\ -3 & 7 & -5 \\ 2 & -5 & 8 \end{pmatrix} .$$

We begin by forming the matrix (A|I):

$$\left(\begin{array}{rrr|rrr} 1 & -3 & 2 & 1 & 0 & 0 \\ -3 & 7 & -5 & 0 & 1 & 0 \\ 2 & -5 & 8 & 0 & 0 & 1 \end{array}\right) .$$

Now carry out the following sequence of elementary row operations to both A and I, and identical column operations to A only:

$$\begin{array}{r} \\ A_2 + 3A_3 \rightarrow \\ A_3 - 2A_1 \rightarrow \end{array} \left(\begin{array}{rrr|rrr} 1 & -3 & 2 & 1 & 0 & 0 \\ 0 & -2 & 1 & 3 & 1 & 0 \\ 0 & 1 & 4 & -2 & 0 & 1 \end{array}\right)$$

$$\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0\\ 0 & -2 & 1 & 3 & 1 & 0\\ 0 & 1 & 4 & -2 & 0 & 1\\ & \uparrow & \uparrow & & & \\ & A^2+3A^1 & A^3-2A^1 & & & \end{array}$$

$$\begin{array}{c|ccc|ccc} & 1 & 0 & 0 & 1 & 0 & 0\\ & 0 & -2 & 1 & 3 & 1 & 0\\ 2A^3+A^2 \rightarrow & 0 & 0 & 9 & -1 & 1 & 2 \end{array}$$

$$\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 0 & 0\\ 0 & -2 & 0 & 3 & 1 & 0\\ 0 & 0 & 18 & -1 & 1 & 2\\ & & \uparrow & & & \\ & & 2A^3+A^2 & & & \end{array}\;.$$

We have thus diagonalized A, and the final form of the matrix (A|I) is just $(D|P^T)$. //

Since Theorem 9.13 tells us that every symmetric bilinear form has a diagonal representation, it follows that the associated quadratic form q(X) has the **diagonal representation**

$$q(X) = X^TAX = a_{11}(x^1)^2 + \cdots + a_{nn}(x^n)^2$$

where A is the diagonal matrix representing the (symmetric) bilinear form.

Let us now specialize this discussion somewhat and consider only real symmetric bilinear forms. We begin by noting that in general, the diagonal representation of a symmetric bilinear form f has positive, negative, and zero entries. We can always renumber the basis vectors so that the positive entries appear first, followed by the negative entries and then the zero entries. It is in fact true, as we now show, that any other diagonal representation of f has the same number of positive and negative entries. If there are P positive entries and N negative entries, then the difference S = P - N is called the **signature** of f.

Theorem 9.14 Let $f \in \mathcal{B}(V)$ be a real symmetric bilinear form. Then every diagonal representation of f has the same number of positive and negative entries.

Proof Let $\{e_1, \dots, e_n\}$ be the basis for V in which the matrix of f is diagonal (see Theorem 9.13). By suitably numbering the e_i, we may assume that the first P entries are positive and the next N entries are negative (also note that there could be $n - P - N$ zero entries). Now let $\{e'_1, \dots, e'_n\}$ be another basis for V in which the matrix of f is also diagonal. Again, assume that the first P′ entries are positive and the next N′ entries are negative. Since the rank of f is just the rank of any matrix representation of f, and since the rank of a matrix is just the dimension of its row (or column) space, it is clear that $r(f) = P + N = P' + N'$. Because of this, we need only show that $P = P'$.

Let U be the linear span of the P vectors $\{e_1, \dots, e_P\}$, let W be the linear span of $\{e'_{P'+1}, \dots, e'_n\}$, and note that $\dim U = P$ and $\dim W = n - P'$. Then for all nonzero vectors $u \in U$ and $w \in W$, we have $f(u, u) > 0$ and $f(w, w) \le 0$ (this inequality is $\le$ and not $<$ because if $P' + N' \ne n$, then the last of the basis vectors that span W will define a diagonal element in the matrix of f that is 0). Hence it follows that $U \cap W = \{0\}$, and therefore (by Theorem 2.11)

$$\begin{aligned}\dim(U+W) &= \dim U + \dim W - \dim(U \cap W) = P + n - P' - 0 \\ &= P - P' + n \ .\end{aligned}$$

Since U and W are subspaces of V, it follows that $\dim(U + W) \le \dim V = n$, and therefore $P - P' + n \le n$. This shows that $P \le P'$. Had we let U be the span of $\{e'_1, \dots, e'_{P'}\}$ and W be the span of $\{e_{P+1}, \dots, e_n\}$, we would have found that $P' \le P$. Therefore $P = P'$ as claimed. ∎

While Theorem 9.13 showed that any quadratic form has a diagonal representation, the important special case of a real quadratic form allows an even simpler representation. This corollary is known as **Sylvester's theorem** or the **law of inertia**.

Corollary Let f be a real symmetric bilinear form. Then f has a unique diagonal representation of the form

$$\begin{pmatrix} I_r & & \\ & -I_s & \\ & & 0_t \end{pmatrix}$$

where I_r and I_s are the r x r and s x s unit matrices, and 0_t is the t x t zero matrix. In particular, the associated quadratic form q has a representation of the form

$$q(x_1, \dots, x_n) = (x^1)^2 + \cdots + (x^r)^2 - (x^{r+1})^2 - \cdots - (x^{r+s})^2 \ .$$

Proof Let f be represented by a (real) symmetric n x n matrix A. By Theorem 9.14, there exists a nonsingular matrix P_1 such that $D = P_1{}^T A P_1 = (d_{ij})$ is a diagonal representation of f with a unique number r of positive entries followed by a unique number s of negative entries. We let t = n - r - s be the unique number of zero entries in D. Now let P_2 be the diagonal matrix with diagonal entries

$$(P_2)_{ii} = \begin{cases} 1\big/\sqrt{d_{ii}} & \text{for } i = 1, \ldots, r \\ 1\big/\sqrt{-d_{ii}} & \text{for } i = r+1, \ldots, r+s \\ 1 & \text{for } i = r+s+1, \ldots, n \end{cases} .$$

Since P_2 is diagonal, it is obvious that $(P_2)^T = P_2$. We leave it to the reader to multiply out the matrices and show that

$$P_2{}^T D P_2 = P_2{}^T P_1{}^T A P_1 P_2 = (P_1 P_2)^T A (P_1 P_2)$$

is a congruence transformation that takes A into the desired form. ∎

We say that a real symmetric bilinear form $f \in \mathcal{B}(V)$ is **nonnegative** (or **positive semidefinite**) if $q(X) = X^T A X = \sum_{i,\,j} a_{ij} x^i x^j = f(X, X) \geq 0$ for all $X \in V$, and we say that f is **positive definite** if $q(X) > 0$ for all nonzero $X \in V$. In particular, from Theorem 9.14 we see that f is nonnegative semidefinite if and only if the signature $S = r(f) \leq \dim V$, and f will be positive definite if and only if $S = \dim V$.

Example 9.10 The quadratic form $(x^1)^2 - 4x^1x^2 + 5(x^2)^2$ is positive definite because it can be written in the form

$$(x^1 - 2x^2)^2 + (x^2)^2$$

which is nonnegative for all real values of x^1 and x^2, and is zero only if $x^1 = x^2 = 0$.

The quadratic form $(x^1)^2 + (x^2)^2 + 2(x^3)^2 - 2x^1x^3 - 2x^2x^3$ can be written in the form

$$(x^1 - x^3)^2 + (x^2 - x^3)^2 .$$

Since this is nonnegative for all real values of x^1, x^2 and x^3 but is zero for nonzero values (e.g., $x^1 = x^2 = x^3 \neq 0$), this quadratic form is nonnegative but not positive definite. ∕∕

Exercises

1. Determine the rank and signature of the following real quadratic forms:
(a) $x^2 + 2xy + y^2$.
(b) $x^2 + xy + 2xz + 2y^2 + 4yz + 2z^2$.

2. Find the transition matrix P such that P^TAP is diagonal where A is given by:

$$(a)\ \begin{pmatrix} 1 & 2 & -3 \\ 2 & 5 & -4 \\ -3 & -4 & 8 \end{pmatrix} \qquad (b)\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & -2 & 2 \\ 1 & 2 & -1 \end{pmatrix}$$

$$(c)\ \begin{pmatrix} 1 & 1 & -2 & -3 \\ 1 & 2 & -5 & -1 \\ -2 & -5 & 6 & 9 \\ -3 & -1 & 9 & 11 \end{pmatrix}$$

3. Let f be the symmetric bilinear form associated with the real quadratic form $q(x, y) = ax^2 + bxy + cy^2$. Show that:
(a) f is nondegenerate if and only if $b^2 - 4ac \neq 0$.
(b) f is positive definite if and only if $a > 0$ and $b^2 - 4ac < 0$.

4. If A is a real, symmetric, positive definite matrix, show there exists a nonsingular matrix P such that $A = P^T P$.

The remaining exercises are all related.

5. Let V be finite-dimensional over $\mathbb{C}$, let S be the subspace of all symmetric bilinear forms on V, and let Q be the set of all quadratic forms on V.
(a) Show that Q is a subspace of all functions from V to $\mathbb{C}$.
(b) Suppose $T \in L(V)$ and $q \in Q$. Show that the equation $(T^\dagger q)(v) = q(Tv)$ defines a quadratic form $T^\dagger q$ on V.
(c) Show that the function $T^\dagger$ is a linear operator on Q, and show that $T^\dagger$ is invertible if and only if T is invertible.

6. (a) Let q be the quadratic form on $\mathbb{R}^2$ defined by $q(x, y) = ax^2 + 2bxy + cy^2$ (where $a \neq 0$). Find an invertible $T \in L(\mathbb{R}^2)$ such that

$$(T^\dagger q)(x, y) = ax^2 + (c - b^2/a)y^2 \ .$$

[*Hint*: Complete the square to find T^{-1} (and hence T).]

(b) Let q be the quadratic form on $\mathbb{R}^2$ defined by q(x, y) = 2bxy. Find an invertible $T \in L(\mathbb{R}^2)$ such that

$$(T^\dagger q)(x, y) = 2bx^2 - 2by^2 \ .$$

(c) Let q be the quadratic form on $\mathbb{R}^3$ defined by q(x, y, z) = xy + 2xz + z^2. Find an invertible $T \in L(\mathbb{R}^3)$ such that

$$(T^\dagger q)(x, y, z) = x^2 - y^2 + z^2 \ .$$

7. Suppose $A \in M_n(\mathbb{R})$ is symmetric, and define a quadratic form q on $\mathbb{R}^n$ by

$$q(X) = \sum_{i,j=1}^{n} a_{ij}x^i x^j \ .$$

Show there exists $T \in L(\mathbb{R}^n)$ such that

$$(T^\dagger q)(X) = \sum_{i=1}^{n} c_i(x_i)^2$$

where each c_i is either 0 or ±1.

9.7 HERMITIAN FORMS

Let us now briefly consider how some of the results of the previous sections carry over to the case of bilinear forms over the complex number field. Much of this material will be elaborated on in the next chapter.

We say that a mapping f: V × V → $\mathbb{C}$ is a **Hermitian form** on V if for all $u_1, u_2, v \in V$ and $a, b \in \mathbb{C}$ we have

(1) $f(au_1 + bu_2, v) = a^*f(u_1, v) + b^*f(u_2, v)$.
(2) $f(u_1, v) = f(v, u_1)^*$.

(We should point out that many authors define a Hermitian form by requiring that the scalars a and b on the right hand side of property (1) not be the complex conjugates as we have defined it. In this case, the scalars on the right hand side of property (3) below will be the complex conjugates of what we have shown.) As was the case for the Hermitian inner product (see Section 2.4), we see that

$$\begin{aligned} f(u,\, av_1+bv_2) &= f(av_1+bv_2,\, u)^* = [a^*f(v_1,\, u)+b^*f(v_2,\, u)]^* \\ &= af(v_1,\, u)^*+bf(v_2,\, u)^* = af(u,\, v_1)+bf(u,\, v_2) \end{aligned}$$

which we state as

(3) $f(u, av_1 + bv_2) = af(u, v_1) + bf(u, v_2)$.

Since $f(u, u) = f(u, u)^*$ it follows that $f(u, u) \in \mathbb{R}$ for all $u \in V$.

Along with a Hermitian form f is the **associated Hermitian quadratic form** $q\colon V \to \mathbb{R}$ defined by $q(u) = f(u, u)$ for all $u \in V$. A little algebra (Exercise 9.7.1) shows that f may be obtained from q by the **polar form** expression of f which is

$$f(u, v) = (1/4)[q(u + v) - q(u - v)] - (i/4)[q(u + iv) - q(u - iv)] \;.$$

We also say that f is **nonnegative semidefinite** if $q(u) = f(u, u) \geq 0$ for all $u \in V$, and **positive definite** if $q(u) = f(u, u) > 0$ for all nonzero $u \in V$. For example, the usual Hermitian inner product on $\mathbb{C}^n$ is a positive definite form since for every nonzero $X = (x^1, \ldots, x^n) \in \mathbb{C}^n$ we have

$$q(X) = f(X,\, X) = \langle X,\, X\rangle = \sum_{i=1}^{n}(x^i)^*x^i = \sum_{i=1}^{n}|x^i|^2 > 0 \;.$$

As we defined in Section 8.1, we say that a matrix $H = (h_{ij}) \in M_n(\mathbb{C})$ is **Hermitian** if $h_{ij} = h_{ji}^*$. In other words, H is Hermitian if $H = H^{*T}$. We denote the operation of taking the transpose along with taking the complex conjugate of a matrix A by $A^\dagger$ (read "A dagger"). In other words, $A^\dagger = A^{*T}$. For reasons that will become clear in the next chapter, we frequently call $A^\dagger$ the (**Hermitian**) **adjoint** of A. Thus H is Hermitian if $H^\dagger = H$.

Note also that for any scalar k we have $k^\dagger = k^*$. Furthermore, using Theorem 3.18(d), we see that

$$(AB)^\dagger = (AB)^{*T} = (A^*B^*)^T = B^\dagger A^\dagger \;.$$

By induction, this obviously extends to any finite product of matrices. It is also clear that

$$A^{\dagger\dagger} = A \;.$$

Example 9.11 Let H be a Hermitian matrix. We show that $f(X, Y) = X^\dagger HY$ defines a Hermitian form on $\mathbb{C}^n$. Let $X_1, X_2, Y \in \mathbb{C}^n$ be arbitrary, and let $a, b \in \mathbb{C}$. Then (using Theorem 3.18(a))

$$\begin{aligned} f(aX_1 + bX_2, Y) &= (aX_1 + bX_2)^\dagger HY \\ &= (a^*X_1{}^\dagger + b^*X_2{}^\dagger)HY \\ &= a^*X_1{}^\dagger HY + b^*X_2{}^\dagger HY \\ &= a^*f(X_1, Y) + b^*f(X_2, Y) \end{aligned}$$

which shows that f(X, Y) satisfies property (1) of a Hermitian form. Now, since $X^\dagger HY$ is a (complex) scalar we have $(X^\dagger HY)^T = X^\dagger HY$, and therefore

$$f(X, Y)^* = (X^\dagger HY)^* = (X^\dagger HY)^\dagger = Y^\dagger HX = f(Y, X)$$

where we used the fact that $H^\dagger = H$. Thus f(X, Y) satisfies property (2), and hence defines a Hermitian form on $\mathbb{C}^n$.

It is probably worth pointing out that $X^\dagger HY$ will not be a Hermitian form if the alternative definition mentioned above is used. In this case, one must use $f(X, Y) = X^T HY^*$ (see Exercise 9.7.2). ∕∕

Now let V have basis $\{e_i\}$, and let f be a Hermitian form on V. Then for any $X = \sum x^i e_i$ and $Y = \sum y^i e_i$ in V, we see that

$$f(X, Y) = f(\textstyle\sum_i x^i e_i, \sum_j y^j e_j) = \sum_{i,\,j} x^{i*} y^j f(e_i, e_j) \;.$$

Just as we did in Theorem 9.9, we define the matrix elements h_{ij} representing a Hermitian form f by $h_{ij} = f(e_i, e_j)$. Note that since $f(e_i, e_j) = f(e_j, e_i)^*$, we see that the diagonal elements of $H = (h_{ij})$ must be real. Using this definition for the matrix elements of f, we then have

$$f(X, Y) = \textstyle\sum_{i,\,j} x^{i*} h_{ij} y^j = X^\dagger HY \;.$$

Following the proof of Theorem 9.9, this shows that any Hermitian form f has a unique representation in terms of the Hermitian matrix H.

If we want to make explicit the basis referred to in this expression, we write $f(X, Y) = [X]_e{}^\dagger H[Y]_e$ where it is understood that the elements h_{ij} are defined with respect to the basis $\{e_i\}$. Finally, let us prove the complex analogues of Theorems 9.11 and 9.14.

Theorem 9.15 Let f be a Hermitian form on V, and let P be the transition matrix from a basis $\{e_i\}$ for V to a new basis $\{e'_i\}$. If H is the matrix of f with respect to the basis $\{e_i\}$ for V, then $H' = P^\dagger HP$ is the matrix of f relative to the new basis $\{e'_i\}$.

Proof We saw in the proof of Theorem 9.11 that for any $X \in V$ we have $[X]_e = P[X]_{e'}$, and hence $[X]_e{}^\dagger = [X]_{e'}{}^\dagger P^\dagger$. Therefore, for any $X, Y \in V$ we see that

$$f(X, Y) = [X]_e{}^\dagger H[Y]_e = [X]_{e'}{}^\dagger P^\dagger HP[Y]_{e'} = [X]_{e'}{}^\dagger H'[Y]_{e'}$$

where $H' = P^\dagger HP$ is the (unique) matrix of f relative to the basis $\{e'_i\}$. ∎

Theorem 9.16 Let f be a Hermitian form on V. Then there exists a basis for V in which the matrix of f is diagonal, and every other diagonal representation of f has the same number of positive and negative entries.

Proof Using the fact that f(u, u) is real for all $u \in V$ along with the appropriate polar form of f, it should be easy for the reader to follow the proofs of Theorems 9.13 and 9.14 and complete the proof of this theorem (see Exercise 9.7.3). ∎

We note that because of this result, our earlier definition for the signature of a bilinear form applies equally well to Hermitian forms.

Exercises

1. Let f be a Hermitian form on V and q the associated quadratic form. Verify the polar form

$$f(u, v) = (1/4)[q(u + v) - q(u - v)] - (i/4)[q(u + iv) - q(u - iv)] \ .$$

2. Verify the statement made at the end of Example 9.11.

3. Prove Theorem 9.16.

4. Show that the algorithm described in Section 9.6 applies to Hermitian matrices if we allow multiplication by complex numbers and, instead of multiplying by E^T on the right, we multiply by E^{*T}.

5. For each of the following Hermitian matrices H, use the results of the previous exercise to find a nonsingular matrix P such that P^THP is diagonal:

$(a)\ \begin{pmatrix} 1 & i \\ -i & 2 \end{pmatrix}$ $\qquad$ $(b)\ \begin{pmatrix} 1 & 2+3i \\ 2-3i & -1 \end{pmatrix}$

$(c)\ \begin{pmatrix} 1 & i & 2+i \\ -i & 2 & 1-i \\ 2-i & 1+i & 2 \end{pmatrix}$ $\qquad$ $(d)\ \begin{pmatrix} 1 & 1+i & 2i \\ 1-i & 4 & 2-3i \\ -2i & 2+3i & 7 \end{pmatrix}$

9.8 SIMULTANEOUS DIAGONALIZATION *

We now apply the results of Sections 8.1, 9.5 and 9.6 to the problem of simultaneously diagonalizing two real quadratic forms. After the proof we shall give an example of how this result applies to classical mechanics.

Theorem 9.17 Let X^TAX and X^TBX be two real quadratic forms on an n-dimensional Euclidean space V, and assume that X^TAX is positive definite. Then there exists a nonsingular matrix P such that the transformation X = PY reduces X^TAX to the form

$$X^TAX = Y^TY = (y^1)^2 + \cdots + (y^n)^2$$

and X^TBX to the form

$$X^TBX = Y^TDY = \lambda_1(y^1)^2 + \cdots + \lambda_n(y^n)^2$$

where $\lambda_1, \ldots, \lambda_n$ are roots of the equation

$$\det(B - \lambda A) = 0\ .$$

Moreover, the λ_i are real and positive if and only if X^TBX is positive definite.

Proof Since A is symmetric, Theorem 9.13 tells us there exists a basis for V that diagonalizes A. Furthermore, the corollary to Theorem 9.14 and the discussion following it shows that the fact A is positive definite means that the corresponding nonsingular transition matrix R may be chosen so that the transformation X = RY yields

$$X^TAX = Y^TY = (y^1)^2 + \cdots + (y^n)^2\ .$$

Note that $Y^TY = X^TAX = Y^TR^TARY$ implies that

$$R^T AR = I .$$

We also emphasize that R will not be orthogonal in general.

Now observe that $R^T BR$ is a real symmetric matrix since B is, and hence (by the corollary to Theorem 8.2) there exists an orthogonal matrix Q such that

$$Q^T R^T BRQ = (RQ)^T B(RQ) = \mathrm{diag}(\lambda_1, \dots, \lambda_n) = D$$

where the λ_i are the eigenvalues of $R^T BR$. If we define the nonsingular (and not generally orthogonal) matrix $P = RQ$, then

$$P^T BP = D$$

and

$$P^T AP = Q^T R^T ARQ = Q^T IQ = I .$$

Under the transformation $X = PY$, we are then left with

$$X^T AX = Y^T P^T APY = Y^T Y$$

as before, while

$$X^T BX = Y^T P^T BPY = Y^T DY = \lambda_1(y^1)^2 + \cdots + \lambda_n(y^n)^2$$

as desired.

Now note that by definition, the λ_i are roots of the equation

$$\det(R^T BR - \lambda I) = 0 .$$

Using $R^T AR = I$ this may be written as

$$\det[R^T(B - \lambda A)R] = 0 .$$

Since $\det R = \det R^T \neq 0$, we find that (using Theorem 4.8)

$$\det(B - \lambda A) = 0 .$$

Finally, since B is a real symmetric matrix, there exists an orthogonal matrix S that brings it into the form

$$S^T BS = \mathrm{diag}(\mu_1, \dots, \mu_n) = \tilde{D}$$

where the μ_i are the eigenvalues of B. Writing $X = SY$, we see that

$$X^T BX = Y^T S^T BSY = Y^T \tilde{D} Y = \mu_1(y^1)^2 + \cdots + \mu_n(y^n)^2$$

and thus $X^T BX$ is positive definite if and only if $Y^T\tilde{D}Y$ is positive definite, i.e., if and only if every $\mu_i > 0$. Since we saw above that

$$P^T BP = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = D$$

it follows from Theorem 9.14 that the number of positive μ_i must equal the number of positive λ_i. Therefore $X^T BX$ is positive definite if and only if every $\lambda_i > 0$. ∎

Example 9.12 Let us show how Theorem 9.17 can be of help in classical mechanics. This example requires a knowledge of both the Lagrange equations of motion and Taylor series expansions. The details of the physics are given in, e.g., the classic text by Goldstein (1980). Our purpose is simply to demonstrate the usefulness of this theorem.

Consider the small oscillations of a conservative system of N particles about a point of stable equilibrium. We assume that the position $\mathbf{r}_i$ of the *ith* particle is a function of n generalized coordinates q_i, and not explicitly on the time t. Thus we write $\mathbf{r}_i = \mathbf{r}_i(q_1, \ldots, q_n)$, and

$$\frac{d\mathbf{r}_i}{dt} = \dot{\mathbf{r}}_i = \sum_{j=1}^{n} \frac{\partial \mathbf{r}_i}{\partial q_j} \dot{q}_j$$

where we denote the derivative with respect to time by a dot.

Since the velocity v_i of the *ith* particle is given by $\dot{\mathbf{r}}_i$, the kinetic energy T of the *ith* particle is $(1/2)m_i(v_i)^2 = (1/2)m_i\dot{\mathbf{r}}_i\bullet\dot{\mathbf{r}}_i$, and hence the kinetic energy of the system of N particles is given by

$$T = \sum_{i=1}^{N} \frac{1}{2} m_i \dot{\mathbf{r}}_i \bullet \dot{\mathbf{r}}_i = \sum_{j,k=1}^{n} M_{jk} \dot{q}_j \dot{q}_k$$

where

$$M_{jk} = \sum_{i=1}^{N} \frac{1}{2} m_i \frac{\partial \mathbf{r}_i}{\partial q_j} \bullet \frac{\partial \mathbf{r}_i}{\partial q_k} = M_{kj} \quad .$$

Thus the kinetic energy is a quadratic form in the generalized velocities $\dot{q}_i$. We also assume that the equilibrium position of each q_i is at $q_i = 0$. Let the potential energy of the system be $V = V(q_1, \ldots, q_n)$. Expanding V in a Taylor series expansion about the equilibrium point, we have (using an obvious notation for evaluating functions at equilibrium)

$$V(q_1, \ldots, q_n) = V(0) + \sum_{i=1}^{n} \left(\frac{\partial V}{\partial q_i} \right)_0 q_i + \frac{1}{2} \sum_{i,j=1}^{n} \left(\frac{\partial^2 V}{\partial q_i \partial q_j} \right)_0 q_i q_j + \cdots \ .$$

At equilibrium, the force on any particle vanishes, and hence we must have $(\partial V/\partial q_i)_0 = 0$ for every i. Furthermore, we may shift the zero of potential and assume that $V(0) = 0$ because this has no effect on the force on each particle. We may therefore write the potential as the quadratic form

$$V = \sum_{i,j=1}^{n} b_{ij}q_i q_j$$

where the b_{ij} are constants, and $b_{ij} = b_{ji}$. Returning to the kinetic energy, we expand M_{ij} about the equilibrium position to obtain

$$M_{ij}(q_1, \dots, q_n) = M_{ij}(0) + \sum_{k=1}^{n}\left(\frac{\partial M_{ij}}{\partial q_k}\right)_0 q_k + \cdots \ .$$

To a first approximation, we may keep only the first (constant) term in this expansion. Then denoting $M_{ij}(0)$ by $a_{ij} = a_{ji}$ we have

$$T = \sum_{i,j=1}^{n} a_{ij}\dot{q}_i\dot{q}_j \ .$$

so that T is also a quadratic form.

The Lagrange equations of motion are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i}$$

where $L = T - V$ is called the **Lagrangian**. Since T is a function of the $\dot{q}_i$ and V is a function of the q_i , the equations of motion take the form

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) = -\frac{\partial V}{\partial q_i} \ . \qquad (*)$$

Now, the physical nature of the kinetic energy tells us that T must be a positive definite quadratic form, and hence we seek to diagonalize T as follows.

Define new coordinates $q'_1, \dots, q'_n$ by $q_i = \sum_j p_{ij}q'_j$ where $P = (p_{ij})$ is a nonsingular constant matrix. Then differentiating with respect to time yields $\dot{q}_i = \sum_j p_{ij}\dot{q}'_j$ so that the $\dot{q}_i$ are transformed in the same manner as the q_i. By Theorem 9.17, the transformation P may be chosen so that T and V take the forms

$$T = (\dot{q}'_1)^2 + \cdots + (\dot{q}'_n)^2$$

and

$$V = \lambda_1(q'_1)^2 + \cdots + \lambda_n(q'_n)^2 \ .$$

Since $V = 0$ at $q_1 = \cdots = q_n = 0$, the fact that P is nonsingular tells us that $V = 0$ at $q'_1 = \cdots = q'_n = 0$ as well. Thus we see that V is also positive definite, and hence each $\lambda_i > 0$. This means that we may write $\lambda_i = \omega_i{}^2$ where each ω_i is real and positive.

Since P is a constant matrix, the equations of motion (*) are just as valid for T and V expressed in terms of q'_i and $\dot{q}'_i$. Therefore, substituting these expressions for T and V into (*), we obtain the equations of motion

$$\frac{d^2 q_i'}{dt^2} = -\omega_i{}^2 q_i' \ \ .$$

For each i = 1, . . . , n the solution to this equation is

$$q'_i = \alpha_i \cos(\omega_i t + \beta_i)$$

where α_i and β_i are constants to be determined from the initial conditions of the problem.

The coordinates q'_i are called the **normal coordinates** for the system of particles, and the form of the solution shows that the particles move according to simple harmonic motion. ∕∕

For additional applications related to this example, we refer the reader to any advanced text on classical mechanics, such as those listed in the bibliography. (See, eg., Marion, chapter 13.6.)

CHAPTER 10

Linear Operators

Recall that a linear transformation $T \in L(V)$ of a vector space into itself is called a (linear) operator. In this chapter we shall elaborate somewhat on the theory of operators. In so doing, we will define several important types of operators, and we will also prove some important diagonalization theorems. Much of this material is directly useful in physics and engineering as well as in mathematics. While some of this chapter overlaps with Chapter 8, we assume that the reader has studied at least Section 8.1.

10.1 LINEAR FUNCTIONALS AND ADJOINTS

Recall that in Theorem 9.3 we showed that for a finite-dimensional real inner product space V, the mapping $u \mapsto L_u = \langle u, \ \rangle$ was an isomorphism of V onto V*. This mapping had the property that $L_{au}v = \langle au, v\rangle = a\langle u, v\rangle = aL_u v$, and hence $L_{au} = aL_u$ for all $u \in V$ and $a \in \mathbb{R}$. However, if V is a complex space with a Hermitian inner product, then $L_{au}v = \langle au, v\rangle = a^*\langle u, v\rangle = a^*L_u v$, and hence $L_{au} = a^*L_u$ which is not even linear (this was the definition of an antilinear (or conjugate linear) transformation given in Section 9.2). Fortunately, there is a closely related result that holds even for complex vector spaces.

Let V be finite-dimensional over $\mathbb{C}$, and assume that V has an inner product $\langle \ , \ \rangle$ defined on it (this is just a positive definite Hermitian form on V). Thus for any $X, Y \in V$ we have $\langle X, Y\rangle \in \mathbb{C}$. For example, with respect to the standard basis $\{e_i\}$ for $\mathbb{C}^n$ (which is the same as the standard basis for $\mathbb{R}^n$), we have $X = \sum x^i e_i$ and hence (see Example 2.13)

$$\begin{aligned}\langle X, Y\rangle &= \langle \Sigma_i x^i e_i, \Sigma_j y^j e_j\rangle = \Sigma_{i,j} x^i{}^* y^j \langle e_i, e_j\rangle = \Sigma_{i,j} x^i{}^* y^j \delta_{ij} \\ &= \Sigma_i x^i{}^* y^i = X^{*T} Y \ .\end{aligned}$$

Note that we are temporarily writing X^{*T} rather than $X^\dagger$. We will shortly explain the reason for this (see Theorem 10.2 below). In particular, for any $T \in L(V)$ and $X \in V$ we have the vector $TX \in V$, and hence it is meaningful to write expressions of the form $\langle TX, Y\rangle$ and $\langle X, TY\rangle$.

Since we are dealing with finite-dimensional vector spaces, the Gram-Schmidt process (Theorem 2.21) guarantees that we can always work with an orthonormal basis. Hence, let us consider a complex inner product space V with basis $\{e_i\}$ such that $\langle e_i, e_j\rangle = \delta_{ij}$. Then, just as we saw in the proof of Theorem 9.1, we now see that for any $u = \sum u^j e_j \in V$ we have

$$\langle e_i, u\rangle = \langle e_i, \Sigma_j u^j e_j\rangle = \Sigma_j u^j \langle e_i, e_j\rangle = \Sigma_j u^j \delta_{ij} = u^i$$

and thus

$$u = \Sigma_i \langle e_i, u\rangle e_i \ .$$

Now consider the vector Te_j. Applying the result of the previous paragraph we have

$$Te_j = \Sigma_i \langle e_i, Te_j\rangle e_i \ .$$

But this is precisely the definition of the matrix $A = (a_{ij})$ that represents T relative to the basis $\{e_i\}$. In other words, this extremely important result shows that the **matrix elements** a_{ij} **of the operator** $T \in L(V)$ are given by

$$a_{ij} = \langle e_i, Te_j\rangle \ .$$

It is important to note however, that this definition depended on the use of an orthonormal basis for V. To see the self-consistency of this definition, we go back to our original definition of (a_{ij}) as $Te_j = \Sigma_k e_k a_{kj}$. Taking the scalar product of both sides of this equation with e_i yields (using the orthonormality of the e_i)

$$\langle e_i, Te_j\rangle = \langle e_i, \Sigma_k e_k a_{kj}\rangle = \Sigma_k a_{kj} \langle e_i, e_k\rangle = \Sigma_k a_{kj} \delta_{ik} = a_{ij} \ .$$

We now prove the complex analogue of Theorem 9.3.

Theorem 10.1 Let V be a finite-dimensional inner product space over $\mathbb{C}$. Then, given any linear functional L on V, there exists a unique u ∈ V such that $Lv = \langle u, v \rangle$ for all v ∈ V.

Proof Let $\{e_i\}$ be an orthonormal basis for V and define $u = \sum_i (Le_i)^* e_i$. Now define the linear functional L_u on V by $L_u v = \langle u, v \rangle$ for every v ∈ V. Then, in particular, we have

$$L_u e_i = \langle u, e_i \rangle = \langle \textstyle\sum_j (Le_j)^* e_j, e_i \rangle = \sum_j Le_j \langle e_j, e_i \rangle = \sum_j Le_j \delta_{ji} = Le_i \ .$$

Since L and L_u agree on a basis for V, they must agree on any v ∈ V, and hence $L = L_u = \langle u, \ \rangle$.

As to the uniqueness of the vector u, suppose u′ ∈ V has the property that $Lv = \langle u', v \rangle$ for every v ∈ V. Then $Lv = \langle u, v \rangle = \langle u', v \rangle$ so that $\langle u - u', v \rangle = 0$. Since v was arbitrary we may choose v = u − u′. Then $\langle u - u', \ u - u' \rangle = 0$ which implies that (since the inner product is just a positive definite Hermitian form) u − u′ = 0 or u = u′. ∎

The importance of finite-dimensionality in this theorem is shown by the following example.

Example 10.1 Let $V = \mathbb{R}[x]$ be the (infinite-dimensional) space of all polynomials over $\mathbb{R}$, and define an inner product on V by

$$\langle f, g \rangle = \int_0^1 f(x) g(x) \, dx$$

for every f, g ∈ V. We will give an example of a linear functional L on V for which there does not exist a polynomial h ∈ V with the property that $Lf = \langle h, f \rangle$ for all f ∈ V.

To show this, define the nonzero linear functional L by

$$Lf = f(0) \ .$$

(L is nonzero since, e.g., L(a + x) = a.) Now suppose there exists a polynomial h ∈ V such that $Lf = f(0) = \langle h, f \rangle$ for every f ∈ V. Then, in particular, we have

$$L(xf) = 0f(0) = 0 = \langle h, xf \rangle$$

for every f ∈ V. Choosing f = xh we see that

$$0 = \langle h, x^2 h \rangle = \int_0^1 x^2 h^2 \, dx \ .$$

Since the integrand is strictly positive, this forces h to be the zero polynomial. Thus we are left with $Lf = \langle h, f\rangle = \langle 0, f\rangle = 0$ for every $f \in V$, and hence $L = 0$. But this contradicts the fact that $L \neq 0$, and hence no such polynomial h can exist.

Note the fact that V is infinite-dimensional is required when we choose $f = xh$. The reason for this is that if V consisted of all polynomials of degree ≤ some positive integer N, then $f = xh$ could have degree > N. ∕∕

Now consider an operator $T \in L(V)$, and let u be an arbitrary element of V. Then the mapping $L_u\colon V \to \mathbb{C}$ defined by $L_u v = \langle u, Tv\rangle$ for every $v \in V$ is a linear functional on V. Applying Theorem 10.1, we see that there exists a unique $u' \in V$ such that $\langle u, Tv\rangle = L_u v = \langle u', v\rangle$ for every $v \in V$. We now define the mapping $T^\dagger\colon V \to V$ by $T^\dagger u = u'$. In other words, we define the **adjoint** $T^\dagger$ of an operator $T \in L(V)$ by

$$\langle T^\dagger u, v\rangle = \langle u, Tv\rangle$$

for all $u, v \in V$. The mapping $T^\dagger$ is unique because u' is unique for a given u. Thus, if $\tilde{T}^\dagger u = u' = T^\dagger u$, then $(\tilde{T}^\dagger - T^\dagger)u = 0$ for every $u \in V$, and hence $\tilde{T}^\dagger - T^\dagger = 0$ or $\tilde{T}^\dagger = T^\dagger$.

Note further that

$$\langle Tu, v\rangle = \langle v, Tu\rangle^* = \langle T^\dagger v, u\rangle^* = \langle u, T^\dagger v\rangle \ .$$

However, it follows from the definition that $\langle u, T^\dagger v\rangle = \langle T^{\dagger\dagger}u, v\rangle$. Therefore the uniqueness of the adjoint implies that $T^{\dagger\dagger} = T$.

Let us show that the map $T^\dagger$ is linear. For all $u_1, u_2, v \in V$ and $a, b \in \mathbb{C}$ we have

$$\begin{aligned}\langle T^\dagger(au_1 + bu_2),\, v\rangle &= \langle au_1 + bu_2,\, Tv\rangle \\ &= a^*\langle u_1,\, Tv\rangle + b^*\langle u_2,\, Tv\rangle \\ &= a^*\langle T^\dagger u_1,\, v\rangle + b^*\langle T^\dagger u_2,\, v\rangle \\ &= \langle aT^\dagger u_1,\, v\rangle + \langle bT^\dagger u_2,\, v\rangle \\ &= \langle aT^\dagger u_1 + bT^\dagger u_2,\, v\rangle \ .\end{aligned}$$

Since this is true for every $v \in V$, we must have

$$T^\dagger(au_1 + bu_2) = aT^\dagger u_1 + bT^\dagger u_2 \ .$$

Thus $T^\dagger$ is linear and $T^\dagger \in L(V)$.

If $\{e_i\}$ is an orthonormal basis for V, then the matrix elements of T are given by $a_{ij} = \langle e_i, Te_j\rangle$. Similarly, the matrix elements b_{ij} of $T^\dagger$ are related to those of T because

$$b_{ij} = \langle e_i, T^\dagger e_j\rangle = \langle Te_i, e_j\rangle = \langle e_j, Te_i\rangle^* = a_{ji}^* \; .$$

In other words, if A is the matrix representation of T relative to the orthonormal basis $\{e_i\}$, then A^{*T} is the matrix representation of $T^\dagger$. This explains the symbol and terminology for the Hermitian adjoint used in the last chapter. Note that if V is a real vector space, then the matrix representation of $T^\dagger$ is simply A^T, and we may denote the corresponding operator by T^T.

We summarize this discussion in the following theorem, which is valid only in finite-dimensional vector spaces. (It is also worth pointing out that $T^\dagger$ depends on the particular inner product defined on V.)

Theorem 10.2 Let T be a linear operator on a finite-dimensional complex inner product space V. Then there exists a unique linear operator $T^\dagger$ on V defined by $\langle T^\dagger u, v\rangle = \langle u, Tv\rangle$ for all $u, v \in V$. Furthermore, if A is the matrix representation of T relative to an orthonormal basis $\{e_i\}$, then $A^\dagger = A^{*T}$ is the matrix representation of $T^\dagger$ relative to this same basis. If V is a real space, then the matrix representation of $T^\dagger$ is simply A^T.

Example 10.2 Let us give an example that shows the importance of finite-dimensionality in defining an adjoint operator. Consider the space $V = \mathbb{R}[x]$ of all polynomials over $\mathbb{R}$, and let the inner product be as in Example 10.1. Define the differentiation operator $D \in L(V)$ by $Df = df/dx$. We show that there exists no adjoint operator $D^\dagger$ that satisfies $\langle Df, g\rangle = \langle f, D^\dagger g\rangle$.

Using $\langle Df, g\rangle = \langle f, D^\dagger g\rangle$, we integrate by parts to obtain

$$\begin{aligned}\langle f, D^\dagger g\rangle &= \langle Df, g\rangle = \int_0^1 (Df)g\,dx = \int_0^1 [D(fg) - fDg]\,dx \\ &= (fg)(1) - (fg)(0) - \langle f, Dg\rangle \; .\end{aligned}$$

Rearranging, this general result may be written as

$$\langle f, (D + D^\dagger)g\rangle = (fg)(1) - (fg)(0) \; .$$

We now let $f = x^2(1 - x)^2p$ for any $p \in V$. Then $f(1) = f(0) = 0$ so that we are left with

$$\begin{aligned} 0 &= \langle f,\, (D+D^\dagger)g\rangle = \int_0^1 x^2(1-x)^2\, p(D+D^\dagger)g\, dx \\ &= \langle x^2(1-x)^2(D+D^\dagger)g,\, p\rangle\ . \end{aligned}$$

Since this is true for every $p \in V$, it follows that $x^2(1-x)^2(D + D^\dagger)g = 0$. But $x^2(1-x)^2 > 0$ except at the endpoints, and hence we must have $(D + D^\dagger)g = 0$ for all $g \in V$, and thus $D + D^\dagger = 0$. However, the above general result then yields

$$0 = \langle f, (D + D^\dagger)g\rangle = (fg)(1) - (fg)(0)$$

which is certainly not true for every $f, g \in V$. Hence $D^\dagger$ must not exist.

We leave it to the reader to find where the infinite-dimensionality of $V = \mathbb{R}[x]$ enters into this example. ∕∕

While this example shows that not every operator on an infinite-dimensional space has an adjoint, there are in fact some operators on some infinite-dimensional spaces that do indeed have an adjoint. A particular example of this is given in Exercise 10.1.3. In fact, the famous Riesz representation theorem asserts that any continuous linear functional on a Hilbert space does indeed have an adjoint. While this fact should be well known to anyone who has studied quantum mechanics, we defer further discussion until Chapter 12 (see Theorem 12.26).

As defined previously, an operator $T \in L(V)$ is **Hermitian** (or **self-adjoint**) if $T^\dagger = T$. The elementary properties of the adjoint operator $T^\dagger$ are given in the following theorem. Note that if V is a real vector space, then the properties of the matrix representing an adjoint operator simply reduce to those of the transpose. Hence, a real Hermitian operator is represented by a (real) symmetric matrix.

Theorem 10.3 Suppose $S, T \in L(V)$ and $c \in \mathbb{C}$. Then

(a) $(S + T)^\dagger = S^\dagger + T^\dagger$.
(b) $(cT)^\dagger = c^*T^\dagger$.
(c) $(ST)^\dagger = T^\dagger S^\dagger$.
(d) $T^{\dagger\dagger} = (T^\dagger)^\dagger = T$.
(e) $I^\dagger = I$ and $0^\dagger = 0$.
(f) $(T^\dagger)^{-1} = (T^{-1})^\dagger$.

Proof Let $u, v \in V$ be arbitrary. Then, from the definitions, we have

$$\begin{aligned} (a)\quad \langle (S+T)^\dagger u,\, v\rangle &= \langle u,\, (S+T)v\rangle = \langle u,\, Sv+Tv\rangle = \langle u,\, Sv\rangle + \langle u,\, Tv\rangle \\ &= \langle S^\dagger u,\, v\rangle + \langle T^\dagger u,\, v\rangle = \langle (S^\dagger + T^\dagger)u,\, v\rangle . \end{aligned}$$

$$(b)\quad \langle (cT)^\dagger u,\, v\rangle = \langle u,\, cTv\rangle = c\langle u,\, Tv\rangle = c\langle T^\dagger u,\, v\rangle = \langle c^*T^\dagger u,\, v\rangle .$$

(c) $\langle (ST)^\dagger u,\, v\rangle = \langle u,\, (ST)v\rangle = \langle u,\, S(Tv)\rangle = \langle S^\dagger u,\, Tv\rangle$

$$= \langle T^\dagger(S^\dagger u),\, v\rangle = \langle (T^\dagger S^\dagger)u,\, v\rangle .$$

(d) This was shown in the discussion preceding Theorem 10.2.

(e) $\langle Iu, v\rangle = \langle u, v\rangle = \langle u, Iv\rangle = \langle I^\dagger u, v\rangle$.
$\langle 0u, v\rangle = \langle 0, v\rangle = 0 = \langle u, 0v\rangle = \langle 0^\dagger u, v\rangle$.

(f) $I = I^\dagger = (T\,T^{-1})^\dagger = (T^{-1})^\dagger T^\dagger$ so that $(T^{-1})^\dagger = (T^\dagger)^{-1}$.

The proof is completed by noting that the adjoint and inverse operators are unique. ∎

Corollary If $T \in L(V)$ is nonsingular, then so is $T^\dagger$.

Proof This follows from Theorems 10.3(f) and 5.10. ∎

We now group together several other useful properties of operators for easy reference.

Theorem 10.4 (a) Let V be an inner product space over either $\mathbb{R}$ or $\mathbb{C}$, let $T \in L(V)$, and suppose that $\langle u, Tv\rangle = 0$ for all $u, v \in V$. Then $T = 0$.

(b) Let V be an inner product space over $\mathbb{C}$, let $T \in L(V)$, and suppose that $\langle u, Tu\rangle = 0$ for all $u \in V$. Then $T = 0$.

(c) Let V be a real inner product space, let $T \in L(V)$ be Hermitian, and suppose that $\langle u, Tu\rangle = 0$ for all $u \in V$. Then $T = 0$.

Proof (a) Let $u = Tv$. Then, by definition of the inner product, we see that $\langle Tv, Tv\rangle = 0$ implies $Tv = 0$ for all $v \in V$ which implies that $T = 0$.

(b) For any $u, v \in V$ we have (by hypothesis)

$$\begin{aligned} 0 &= \langle u+v,\, T(u+v)\rangle \\ &= \langle u,\, Tu\rangle + \langle u,\, Tv\rangle + \langle v,\, Tu\rangle + \langle v,\, Tv\rangle \\ &= 0 + \langle u,\, Tv\rangle + \langle v,\, Tu\rangle + 0 \\ &= \langle u,\, Tv\rangle + \langle v,\, Tu\rangle \end{aligned} \qquad (*)$$

Since v is arbitrary, we may replace it with iv to obtain

$$0 = i\langle u, Tv\rangle - i\langle v, Tu\rangle \ .$$

Dividing this by i and adding to (*) results in $0 = \langle u, Tv\rangle$ for any $u, v \in V$. By (a), this implies that $T = 0$.

(c) For any $u, v \in V$ we have $\langle u + v, T(u + v)\rangle = 0$ which also yields (*). Therefore, using (*), the fact that $T^\dagger = T$, and the fact that V is real, we obtain

$$\begin{aligned}0 &= \langle T^{\dagger}u,\, v\rangle + \langle v,\, Tu\rangle = \langle Tu,\, v\rangle + \langle v,\, Tu\rangle = \langle v,\, Tu\rangle + \langle v,\, Tu\rangle \\ &= 2\langle v,\, Tu\rangle.\end{aligned}$$

Since this holds for any u, v ∈ V we have T = 0 by (a). (Note that in this particular case, $T^\dagger = T^T$.) ∎

Exercises

1. Suppose S, T ∈ L(V).
(a) If S and T are Hermitian, show that ST and TS are Hermitian if and only if [S, T] = ST − TS = 0.
(b) If T is Hermitian, show that $S^\dagger TS$ is Hermitian for all S.
(c) If S is nonsingular and $S^\dagger TS$ is Hermitian, show that T is Hermitian.

2. Consider V = $M_n(\mathbb{C})$ with the inner product ⟨A, B⟩ = $\mathrm{Tr}(B^\dagger A)$. For each M ∈ V, define the operator T_M ∈ L(V) by $T_M(A)$ = MA. Show that $(T_M)^\dagger = T_{M^\dagger}$.

3. Consider the space V = $\mathbb{C}[x]$. If f = $\sum a_i x^i$ ∈ V, we define the complex conjugate of f to be the polynomial f* = $\sum a_i{}^* x^i$ ∈ V. In other words, if t ∈ $\mathbb{R}$, then f*(t) = (f(t))*. We define an inner product on V by
$$\langle f,\, g\rangle = \int_0^1 f^*(t)g(t)\, dt\ \ .$$
For each f ∈ V, define the operator T_f ∈ L(V) by $T_f(g)$ = fg. Show that $(T_f)^\dagger = T_{f^*}$.

4. Let V be the space of all real polynomials of degree ≤ 3, and define an inner product on V by
$$\langle f,\, g\rangle = \int_0^1 f(x)g(x)\, dx\ \ .$$
For any t ∈ $\mathbb{R}$, find a polynomial h_t ∈ V such that ⟨h_t, f ⟩ = f(t) for all f ∈ V.

5. If V is as in the previous exercise and D is the usual differentiation operator on V, find $D^\dagger$.

6. Let V = $\mathbb{C}^2$ with the standard inner product.
(a) Define T ∈ L(V) by Te_1 = (1, −2), Te_2 = (*i*, −1). If v = (z_1 z_2), find $T^\dagger v$.

(b) Define $T \in L(V)$ by $Te_1 = (1 + i, 2)$, $Te_2 = (i, i)$. Find the matrix representation of $T^\dagger$ relative to the usual basis for V. Is it true that $[T, T^\dagger] = 0$?

7. Let V be a finite-dimensional inner product space and suppose $T \in L(V)$. Show that $\text{Im } T^\dagger = (\text{Ker } T)^\perp$.

8. Let V be a finite-dimensional inner product space, and suppose $E \in L(V)$ is idempotent, i.e., $E^2 = E$. Prove that $E^\dagger = E$ if and only if $[E, E^\dagger] = 0$.

9. For each of the following inner product spaces V and $L \in V^*$, find a vector $u \in V$ such that $Lv = \langle u, v\rangle$ for all $v \in V$:

   (a) $V = \mathbb{R}^3$ and $L(x, y, z) = x - 2y + 4z$.

   (b) $V = \mathbb{C}^2$ and $L(z_1, z_2) = z_1 - z_2$.

   (c) V is the space of all real polynomials of degree ≤ 2 with inner product as in Exercise 4, and $Lf = f(0) + Df(1)$. (Here D is the usual differentiation operator.)

10. (a) Let $V = \mathbb{R}^2$, and define $T \in L(V)$ by $T(x, y) = (2x + y, x - 3y)$. Find $T^\dagger(3, 5)$.

    (b) Let $V = \mathbb{C}^2$, and define $T \in L(V)$ by $T(z_1, z_2) = (2z_1 + iz_2, (1 - i)z_1)$. Find $T^\dagger(3 - i, 1 + i2)$.

    (c) Let V be as in Exercise 9(c), and define $T \in L(V)$ by $Tf = 3f + Df$. Find $T^\dagger f$ where $f = 3x^2 - x + 4$.

10.2 ISOMETRIC AND UNITARY OPERATORS

Let V be a complex inner product space with the induced norm. Another important class of operators $U \in L(V)$ is that for which $\|Uv\| = \|v\|$ for all $v \in V$. Such operators are called **isometric** because they preserve the length of the vector v. Furthermore, for any $v, w \in V$ we see that

$$\|Uv - Uw\| = \|U(v - w)\| = \|v - w\|$$

so that U preserves distances as well. This is sometimes described by saying that U is an **isometry**.

If we write out the norm as an inner product and assume that the adjoint operator exists, we see that an isometric operator satisfies

$$\langle v, v\rangle = \langle Uv, Uv\rangle = \langle v, (U^\dagger U)v\rangle$$

and hence $\langle v, (U^\dagger U - 1)v\rangle = 0$ for any $v \in V$. But then from Theorem 10.4(b)) it follows that

$$U^\dagger U = 1 .$$

In fact, this is sometimes taken as the definition of an isometric operator. Note that this applies equally well to an infinite-dimensional space.

If V is finite-dimensional, then (Theorems 3.21 and 5.13) it follows that $U^\dagger = U^{-1}$, and hence

$$U^\dagger U = UU^\dagger = 1 .$$

Any operator that satisfies either $U^\dagger U = UU^\dagger = 1$ or $U^\dagger = U^{-1}$ is said to be **unitary**. It is clear that a unitary operator is necessarily isometric. If V is simply a real space, then unitary operators are called **orthogonal**.

Because of the importance of isometric and unitary operators in both mathematics and physics, it is worth arriving at both of these definitions from a slightly different viewpoint that also aids in our understanding of these operators. Let V be a complex vector space with an inner product defined on it. We say that an operator U is **unitary** if $\|Uv\| = \|v\|$ for all $v \in V$, and in addition, it has the property that it is a mapping of V *onto* itself. Since $\|Uv\| = \|v\|$, we see that $Uv = 0$ if and only if $v = 0$, and hence $\text{Ker } U = \{0\}$. Therefore U is one-to-one and U^{-1} exists (Theorem 5.5). Since U is surjective, the inverse is defined on all of V also. Note that there has been no mention of finite-dimensionality. This was avoided by requiring that the mapping be surjective.

Starting from $\|Uv\| = \|v\|$, we may write $\langle v, (U^\dagger U)v\rangle = \langle v, v\rangle$. As we did in the proof of Theorem 10.4, if we first substitute $v = v_1 + v_2$ and then $v = v_1 + iv_2$, divide the second of these equations by i and then add to the first, we find that $\langle v_1, (U^\dagger U)v_2\rangle = \langle v_1, v_2\rangle$. Since this holds for all $v_1, v_2 \in V$, it follows that $U^\dagger U = 1$. If we now multiply this equation from the left by U we have $UU^\dagger U = U$, and hence $(UU^\dagger)(Uv) = Uv$ for all $v \in V$. But as v varies over all of V, so does Uv since U is surjective. We then define $v' = Uv$ so that $(UU^\dagger)v' = v'$ for all $v' \in V$. This shows that $U^\dagger U = 1$ implies $UU^\dagger = 1$. What we have just done then, is show that a surjective norm-preserving operator U has the property that $U^\dagger U = UU^\dagger = 1$. It is important to emphasize that this approach is equally valid in infinite-dimensional spaces.

We now define an **isometric** operator Ω to be an operator defined on all of V with the property that $\|\Omega v\| = \|v\|$ for all $v \in V$. This differs from a unitary operator in that we do not require that Ω also be surjective. Again, the requirement that Ω preserve the norm tells us that Ω has an inverse (since it must be one-to-one), but this inverse is not necessarily defined on the whole of V. For example, let $\{e_i\}$ be an orthonormal basis for V, and define the "shift operator" Ω by

$$\Omega(e_i) = e_{i+1} .$$

This Ω is clearly defined on all of V, but the image of Ω is not all of V since it does not include the vector e_1. Thus, Ω^{-1} is not defined on e_1.

Exactly as we did for unitary operators, we can show that $\Omega^\dagger\Omega = 1$ for an isometric operator Ω. If V happens to be finite-dimensional, then obviously $\Omega\Omega^\dagger = 1$. Thus, on a finite-dimensional space, an isometric operator is also unitary.

Finally, let us show an interesting relationship between the inverse Ω^{-1} of an isometric operator and its adjoint $\Omega^\dagger$. From $\Omega^\dagger\Omega = 1$, we may write $\Omega^\dagger(\Omega v) = v$ for every $v \in V$. If we define $\Omega v = v'$, then for every $v' \in \operatorname{Im}\Omega$ we have $v = \Omega^{-1}v'$, and hence

$$\Omega^\dagger v' = \Omega^{-1}v' \quad \text{for } v' \in \operatorname{Im}\Omega\ .$$

On the other hand, if $w' \in (\operatorname{Im}\Omega)^\perp$, then automatically $\langle w', \Omega v\rangle = 0$ for every $v \in V$. Therefore this may be written as $\langle \Omega^\dagger w', v\rangle = 0$ for every $v \in V$, and hence (choose $v = \Omega^\dagger w'$)

$$\Omega^\dagger w' = 0 \quad \text{for } w' \in (\operatorname{Im}\Omega)^\perp\ .$$

In other words, we have

$$\Omega^\dagger = \begin{cases} \Omega^{-1} & \text{on } \operatorname{Im}\Omega \\ 0 & \text{on } (\operatorname{Im}\Omega)^\perp \end{cases}\ .$$

For instance, using our earlier example of the shift operator, we see that $\langle e_1, e_i\rangle = 0$ for $i \neq 1$, and hence $e_1 \in (\operatorname{Im}\Omega)^\perp$. Therefore $\Omega^\dagger(e_1) = 0$, so that we clearly can not have $\Omega\Omega^\dagger = 1$.

Our next theorem summarizes some of this discussion.

Theorem 10.5 Let V be a complex finite-dimensional inner product space. Then the following conditions on an operator $U \in L(V)$ are equivalent:

(a) $U^\dagger = U^{-1}$.

(b) $\langle Uv, Uw\rangle = \langle v, w\rangle$ for all $v, w \in V$.

(c) $\|Uv\| = \|v\|$.

Proof (a) ⇒ (b): $\langle Uv, Uw\rangle = \langle v, (U^\dagger U)w\rangle = \langle v, Iw\rangle = \langle v, w\rangle$.

(b) ⇒ (c): $\|Uv\| = \langle Uv, Uv\rangle^{1/2} = \langle v, v\rangle^{1/2} = \|v\|$.

(c) ⇒ (a): $\langle v, (U^\dagger U)v\rangle = \langle Uv, Uv\rangle = \langle v, v\rangle = \langle v, Iv\rangle$, and therefore $\langle v, (U^\dagger U - I)v\rangle = 0$. Hence (by Theorem 10.4(b)) we must have $U^\dagger U = I$, and thus $U^\dagger = U^{-1}$ (since V is finite-dimensional). ∎

From part (c) of this theorem we see that U preserves the length of any vector. In particular, U preserves the length of a unit vector, hence the designation "unitary." Note also that if v and w are orthogonal, then $\langle v, w\rangle = 0$ and hence $\langle Uv, Uw\rangle = \langle v, w\rangle = 0$. Thus U maintains orthogonality as well.

Condition (b) of this theorem is sometimes described by saying that a unitary transformation **preserves inner products**. In general, we say that a linear transformation (i.e., a vector space homomorphism) T of an inner product space V *onto* an inner product space W (over the same field) is an **inner product space isomorphism** of V onto W if it also preserves inner products. Therefore, one may define a unitary operator as an inner product space isomorphism.

It is also worth commenting on the case of unitary operators defined on a real vector space. Since in this case the adjoint reduces to the transpose, we have $U^\dagger = U^T = U^{-1}$. If V is a real vector space, then an operator $T = L(V)$ that satisfies $T^T = T^{-1}$ is said to be an **orthogonal** transformation. It should be clear that Theorem 10.5 also applies to real vector spaces if we replace the adjoint by the transpose. We will have more to say about orthogonal transformations below.

Theorem 10.6 Let V be finite-dimensional over $\mathbb{C}$ (*resp.* $\mathbb{R}$). A linear transformation $U \in L(V)$ is unitary (*resp.* orthogonal) if and only if it takes an orthonormal basis for V into an orthonormal basis for V.

Proof We consider the case where V is complex, leaving the real case to the reader. Let $\{e_i\}$ be an orthonormal basis for V, and assume that U is unitary. Then from Theorem 10.5(b) we have

$$\langle Ue_i, Ue_j\rangle = \langle e_i, e_j\rangle = \delta_{ij}$$

so that $\{Ue_i\}$ is also an orthonormal set. But any orthonormal set is linearly independent (Theorem 2.19), and hence $\{Ue_i\}$ forms a basis for V (since there are as many of the Ue_i as there are e_i).

Conversely, suppose that both $\{e_i\}$ and $\{Ue_i\}$ are orthonormal bases for V and let $v, w \in V$ be arbitrary. Then

$$\begin{aligned}\langle v, w\rangle &= \langle \Sigma_i v^i e_i, \Sigma_j w^j e_j\rangle = \Sigma_{i,j} v^i{*}w^j\langle e_i, e_j\rangle = \Sigma_{i,j} v^i{*}w^j\delta_{ij}\\ &= \Sigma_i v^i{*}w^i\ .\end{aligned}$$

However, we also have

$$\begin{aligned}\langle Uv,\, Uw\rangle &= \langle U(\textstyle\sum_i v^i e_i),\, U(\sum_j w^j e_j)\rangle = \sum_{i,j} v^i{}^* w^j \langle Ue_i,\, Ue_j\rangle \\ &= \textstyle\sum_{i,j} v^i{}^* w^j \delta_{ij} = \sum_i v^i{}^* w^i = \langle v,\, w\rangle\ .\end{aligned}$$

This shows that U is unitary (Theorem 10.5). ∎

Corollary Let V and W be finite-dimensional inner product spaces over $\mathbb{C}$. Then there exists an inner product space isomorphism of V onto W if and only if dim V = dim W.

Proof Clearly dim V = dim W if V and W are isomorphic. On the other hand, let $\{e_1, \dots, e_n\}$ be an orthonormal basis for V, and let $\{\bar{e}_1, \dots, \bar{e}_n\}$ be an orthonormal basis for W. (These bases exist by Theorem 2.21.) We define the (surjective) linear transformation U by the requirement $Ue_i = \bar{e}_i$. U is unique by Theorem 5.1. Since $\langle Ue_i, Ue_j\rangle = \langle \bar{e}_i, \bar{e}_j\rangle = \delta_{ij} = \langle e_i, e_j\rangle$, the proof of Theorem 10.6 shows that U preserves inner products. In particular, we see that $\|Uv\| = \|v\|$ for every $v \in V$, and hence Ker U = {0} (by property (N1) of Theorem 2.17). Thus U is also one-to-one (Theorem 5.5). ∎

From Theorem 10.2 we see that a complex matrix A represents a unitary operator relative to an orthonormal basis if and only if $A^\dagger = A^{-1}$. We therefore say that a complex matrix A is a **unitary matrix** if $A^\dagger = A^{-1}$. In the special case that A is a real matrix with the property that $A^T = A^{-1}$, then we say that A is an **orthogonal matrix**. (These classes of matrices were also discussed in Section 8.1.) The reason for this designation is shown in the next example, which is really nothing more than another way of looking at what we have done so far.

Example 10.3 Suppose $V = \mathbb{R}^n$ and $X \in V$. In terms of an orthonormal basis $\{e_i\}$ for V we may write $X = \sum_i x^i e_i$. Now suppose we are given another orthonormal basis $\{\bar{e}_i\}$ related to the first basis by $\bar{e}_i = A(e_i) = \sum_j e_j a_{ji}$ for some real matrix (a_{ij}). Relative to this new basis we have $A(X) = \bar{X} = \sum_i \bar{x}^i \bar{e}_i$ where $x^i = \sum_j a_{ij}\bar{x}^j$ (see Section 5.4). Then

$$\begin{aligned}\|X\|^2 &= \langle \textstyle\sum_i x^i e_i,\, \sum_j x^j e_j\rangle = \sum_{i,j} x^i x^j \langle e_i,\, e_j\rangle = \sum_{i,j} x^i x^j \delta_{ij} \\ &= \textstyle\sum_i (x^i)^2 = \sum_{i,j,k} a_{ij} a_{ik} \bar{x}^j \bar{x}^k = \sum_{i,j,k} a^T{}_{ji} a_{ik} \bar{x}^j \bar{x}^k \\ &= \textstyle\sum_{j,k} (A^T A)_{jk} \bar{x}^j \bar{x}^k\ .\end{aligned}$$

If A is orthogonal, then $A^T = A^{-1}$ so that $(A^T A)_{jk} = \delta_{jk}$ and we are left with

$$\|X\|^2 = \Sigma_i(x^i)^2 = \Sigma_j(\bar{x}^j)^2 = \|\bar{X}\|^2$$

so that the length of X is unchanged under an orthogonal transformation. An equivalent way to see this is to assume that A simply represents a rotation so that the length of a vector remains unchanged by definition. This then forces A to be an orthogonal transformation (see Exercise 10.2.2).

Another way to think of orthogonal transformations is the following. We saw in Section 2.4 that the angle θ between two vectors X, Y $\in \mathbb{R}^n$ is defined by

$$\cos\theta = \frac{\langle X, Y\rangle}{\|X\|\,\|Y\|}\ .$$

Under the orthogonal transformation A, we then have $\bar{X}$ = A(X) and also

$$\cos\bar{\theta} = \frac{\langle \bar{X}, \bar{Y}\rangle}{\|\bar{X}\|\,\|\bar{Y}\|}\ .$$

But $\|\bar{X}\| = \|X\|$ and $\|\bar{Y}\| = \|Y\|$, and in addition,

$$\begin{aligned}\langle X, Y\rangle &= \langle \Sigma_i x^i e_i, \Sigma_j y^j e_j\rangle = \Sigma_i x^i y^i = \Sigma_{i,j,k} a_{ij}\bar{x}^j a_{ik}\bar{y}^k \\ &= \Sigma_{j,k}\delta_{jk}\bar{x}^j\bar{y}^k = \Sigma_j \bar{x}^j\bar{y}^j = \langle \bar{X}, \bar{Y}\rangle\end{aligned}$$

so that $\theta = \bar{\theta}$ (this also follows from the real vector space version of Theorem 10.5). Therefore an orthogonal transformation also preserves the angle between two vectors, and hence is nothing more than a rotation in $\mathbb{R}^n$. ∕∕

Theorem 10.7 The following conditions on a matrix A are equivalent:

(a) A is unitary.
(b) The rows A_i of A form an orthonormal set.
(c) The columns A^i of A form an orthonormal set.

Proof We begin by by noting that, using the usual inner product on $\mathbb{C}^n$, we have

$$(AA^\dagger)_{ij} = \Sigma_k a_{ik}a^\dagger{}_{kj} = \Sigma_k a_{ik}a^*{}_{jk} = \Sigma_k a^*{}_{jk}a_{ik} = \langle A_j, A_i\rangle$$

and

$$(A^\dagger A)_{ij} = \Sigma_k a^\dagger{}_{ik}a_{kj} = \Sigma_k a^*{}_{ki}a_{kj} = \langle A^i, A^j\rangle\ .$$

Now, if A is unitary, then $AA^\dagger = I$ implies $(AA^\dagger)_{ij} = \delta_{ij}$ which then implies that $\langle A_j, A_i\rangle = \delta_{ij}$ so that (a) is equivalent to (b). Similarly, we must have

$(A^\dagger A)_{ij} = \delta_{ij} = \langle A^i, A^j\rangle$ so that (a) is also equivalent to (c). Therefore (b) must also be equivalent to (c). ∎

Note that the equivalence of (b) and (c) in this theorem means that the rows of A form an orthonormal set if and only if the columns of A form an orthonormal set. But the rows of A are just the columns of A^T, and hence A is unitary if and only if A^T is unitary.

It should be obvious that this theorem applies just as well to orthogonal matrices. Looking at this in the other direction, we see that in this case $A^T = A^{-1}$ so that $A^TA = AA^T = I$, and therefore

$$\begin{aligned}(A^TA)_{ij} &= \Sigma_k a^T{}_{ik}a_{kj} = \Sigma_k a_{ki}a_{kj} = \delta_{ij}\\ (AA^T)_{ij} &= \Sigma_k a_{ik}a^T{}_{kj} = \Sigma_k a_{ik}a_{jk} = \delta_{ij} \quad .\end{aligned}$$

Viewing the standard (orthonormal) basis $\{e_i\}$ for $\mathbb{R}^n$ as row vectors, we have $A_i = \Sigma_j a_{ij}e_j$, and hence

$$\begin{aligned}\langle A_i, A_j\rangle &= \langle \Sigma_k a_{ik}e_k, \Sigma_r a_{jr}e_r\rangle = \Sigma_{k,r}a_{ik}a_{jr}\langle e_k, e_r\rangle\\ &= \Sigma_{k,r}a_{ik}a_{jr}\delta_{kr} = \Sigma_k a_{ik}a_{jk} = \delta_{ij} \quad .\end{aligned}$$

Furthermore, it is easy to see that a similar result holds for the columns of A.

Our next theorem details several useful properties of orthogonal and unitary matrices.

Theorem 10.8 (a) If A is an orthogonal matrix, then det A = ±1.

(b) If U is a unitary matrix, then $|\det U| = 1$. Alternatively, $\det U = e^{i\phi}$ for some real number ϕ.

Proof (a) We have $AA^T = I$, and hence (from Theorems 4.8 and 4.1)

$$1 = \det I = \det(AA^T) = (\det A)(\det A^T) = (\det A)^2$$

so that det A = ±1.

(b) If $UU^\dagger = I$ then, as above, we have

$$\begin{aligned}1 &= \det I = \det(UU^\dagger) = (\det U)(\det U^\dagger) = (\det U)(\det U^T)^*\\ &= (\det U)(\det U)^* = |\det U|^2 \quad .\end{aligned}$$

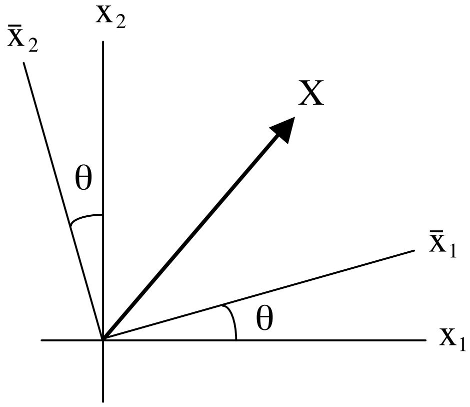
Since the absolute value is defined to be positive, this shows that $|\det U| = 1$ and hence $\det U = e^{i\phi}$ for some real ϕ. ∎

Example 10.4 Let us take a look at rotations in $\mathbb{R}^2$ as shown, for example, in the figure below. Recall from Example 10.3 that if we have two bases $\{e_i\}$ and $\{\bar{e}_i\}$, then they are related by a transition matrix $A = (a_{ij})$ defined by $\bar{e}_i = \sum_j e_j a_{ji}$. In addition, if $X = \sum x^i e_i = \sum \bar{x}^i \bar{e}_i$, then $x^i = \sum_j a_{ij} \bar{x}^j$. If both $\{e_i\}$ and $\{\bar{e}_i\}$ are orthonormal bases, then

$$\langle e_i, \bar{e}_j\rangle = \langle e_i, \textstyle\sum_k e_k a_{kj}\rangle = \textstyle\sum_k a_{kj}\langle e_i, e_k\rangle = \textstyle\sum_k a_{kj}\delta_{ik} = a_{ij} \ .$$

Using the usual dot product on $\mathbb{R}^2$ as our inner product (see Section 2.4, Lemma 2.3) and referring to the figure below, we see that the elements a_{ij} are given by (also see Section 0.6 for the trigonometric identities)i

$$\begin{aligned}
a_{11} &= e_1 \bullet \bar{e}_1 = |e_1||\bar{e}_1|\cos\theta = \cos\theta\\
a_{12} &= e_1 \bullet \bar{e}_2 = |e_1||\bar{e}_2|\cos(\pi/2+\theta) = -\sin\theta\\
a_{21} &= e_2 \bullet \bar{e}_1 = |e_2||\bar{e}_1|\cos(\pi/2-\theta) = \sin\theta\\
a_{22} &= e_2 \bullet \bar{e}_2 = |e_2||\bar{e}_2|\cos\theta = \cos\theta
\end{aligned}$$



Thus the matrix A is given by

$$(a_{ij}) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \ .$$

We leave it to the reader to compute directly that $A^T A = AA^T = I$ and $\det A = +1$. ∕∕

Example 10.5 Referring to the previous example, we can show that any (real) 2 x 2 orthogonal matrix with det A = +1 has the form

$$(a_{ij}) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

for some $\theta \in \mathbb{R}$. To see this, suppose A has the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

where a, b, c, $d \in \mathbb{R}$. Since A is orthogonal, its rows form an orthonormal set, and hence we have

$$a^2 + b^2 = 1, \quad c^2 + d^2 = 1, \quad ac + bd = 0, \quad ad - bc = 1$$

where the last equation follows from det A = 1.

If a = 0, then the first of these equations yields b = ±1, the third then yields d = 0, and the last yields $-c = 1/b = \pm 1$ which is equivalent to c = −b. In other words, if a = 0, then A has either of the following forms:

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} .$$

The first of these is of the required form if we choose $\theta = -90^\circ = -\pi/2$, and the second is of the required form if we choose $\theta = +90^\circ = +\pi/2$.

Now suppose that $a \neq 0$. From the third equation we have c = −bd/a, and substituting this into the second equation, we find $(a^2 + b^2)d^2 = a^2$. Using the first equation, this becomes $a^2 = d^2$ or a = ±d. If a = −d, then the third equation yields b = c, and hence the last equation yields $-a^2 - b^2 = 1$ which is impossible. Therefore a = d, the third equation then yields c = −b, and we are left with

$$\begin{pmatrix} a & -c \\ c & a \end{pmatrix} .$$

Since det A = $a^2 + c^2 = 1$, there exists a real number θ such that a = cos θ and c = sin θ which gives us the desired form for A. ∞

Exercises

1. Let GL(n, $\mathbb{C}$) denote the subset of $M_n(\mathbb{C})$ consisting of all nonsingular matrices, U(n) the subset of all unitary matrices, and L(n) the set of all

nonsingular lower-triangular matrices.

(a) Show that each of these three sets forms a group.

(b) Show that any nonsingular n x n complex matrix can be written as a product of a nonsingular upper-triangular matrix and a unitary matrix. [*Hint*: Use Exercises 5.4.14 and 3.7.7.]

2. Let $V = \mathbb{R}^n$ with the standard inner product, and suppose the length of any $X \in V$ remains unchanged under $A \in L(V)$. Show that A must be an orthogonal transformation.

3. Let V be the space of all continuous complex-valued functions defined on $[0, 2\pi]$, and define an inner product on V by
$$\langle f, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f^*(x)g(x)\,dx \;\; .$$
Suppose there exists $h \in V$ such that $|h(x)| = 1$ for all $x \in [0, 2\pi]$, and define $T_h \in L(V)$ by $T_h f = hf$. Prove that T is unitary.

4. Let W be a finite-dimensional subspace of an inner product space V, and recall that $V = W \oplus W^\perp$ (see Exercise 2.5.11). Define $U \in L(V)$ by
$$U(w_1 + w_2) = w_1 - w_2$$
where $w_1 \in W$ and $w_2 \in W^\perp$.

(a) Prove that U is a Hermitian operator.

(b) Let $V = \mathbb{R}^3$ have the standard inner product, and let $W \subset V$ be spanned by the vector (1, 0, 1). Find the matrix of U relative to the standard basis for V.

5. Let V be a finite-dimensional inner product space. An operator $\Omega \in L(V)$ is said to be a **partial isometry** if there exists a subspace W of V such that $\|\Omega w\| = \|w\|$ for all $w \in W$, and $\|\Omega w\| = 0$ for all $w \in W^\perp$. Let Ω be a partial isometry and suppose $\{w_1, \dots, w_k\}$ is an orthonormal basis for W.

(a) Show that $\langle \Omega u, \Omega v\rangle = \langle u, v\rangle$ for all $u, v \in W$. [*Hint*: Use Exercise 2.4.7.]

(b) Show that $\{\Omega w_1, \dots, \Omega w_k\}$ is an orthonormal basis for Im Ω.

(c) Show there exists an orthonormal basis $\{v_i\}$ for V such that the first k columns of $[\Omega]_v$ form an orthonormal set, and the remaining columns are zero.

(d) Let $\{u_1, \dots, u_r\}$ be an orthonormal basis for $(\text{Im }\Omega)^\perp$. Show that $\{\Omega w_1, \dots, \Omega w_k, u_1, \dots, u_r\}$ is an orthonormal basis for V.

(e) Suppose $T \in L(V)$ satisfies $T(\Omega w_i) = w_i$ (for $1 \le i \le k$) and $Tu_i = 0$ (for $1 \le i \le r$). Show that T is well-defined, and that $T = \Omega^\dagger$.

(f) Show that $\Omega^\dagger$ is a partial isometry.

6. Let V be a complex inner product space, and suppose $H \in L(V)$ is Hermitian. Show that:
(a) $\|v + iHv\| = \|v - iHv\|$ for all $v \in V$.
(b) $u + iHu = v + iHv$ if and only if $u = v$.
(c) $1 + iH$ and $1 - iH$ are nonsingular.
(d) If V is finite-dimensional, then $U = (1 - iH)(1 + iH)^{-1}$ is a unitary operator. (U is called the **Cayley transform** of H. This result is also true in an infinite-dimensional Hilbert space but the proof is considerably more difficult.)

10.3 NORMAL OPERATORS

We now turn our attention to characterizing the type of operator on V for which there exists an orthonormal basis of eigenvectors for V. We begin by taking a look at some rather simple properties of the eigenvalues and eigenvectors of the operators we have been discussing.

To simplify our terminology, we remark that a complex inner product space is also called a **unitary space**, while a real inner product space is sometimes called a **Euclidean space**. If H is an operator such that $H^\dagger = -H$, then H is said to be **anti-Hermitian** (or **skew-Hermitian**). Furthermore, if P is an operator such that $P = S^\dagger S$ for some operator S, then we say that P is **positive** (or **positive semidefinite** or **nonnegative**). If S also happens to be nonsingular (and hence P is also nonsingular), then we say that P is **positive definite**. Note that a positive operator is necessarily Hermitian since $(S^\dagger S)^\dagger = S^\dagger S$. The reason that P is called positive is shown in part (d) of the following theorem.

Theorem 10.9 (a) The eigenvalues of a Hermitian operator are real.

(b) The eigenvalues of an isometry (and hence also of a unitary transformation) have absolute value one.

(c) The eigenvalues of an anti-Hermitian operator are pure imaginary.

(d) A positive (positive definite) operator has eigenvalues that are real and nonnegative (positive).

Proof (a) If H is Hermitian, $v \neq 0$, and $Hv = \lambda v$, we have

$$\begin{aligned}\lambda\langle v, v\rangle &= \langle v, \lambda v\rangle = \langle v, Hv\rangle = \langle H^\dagger v, v\rangle = \langle Hv, v\rangle \\ &= \langle \lambda v, v\rangle = \lambda^*\langle v, v\rangle \;\; .\end{aligned}$$

But $\langle v, v\rangle \neq 0$, and hence $\lambda = \lambda^*$.

(b) If Ω is an isometry, $v \neq 0$, and $\Omega v = \lambda v$, then we have (using Theorem 2.17)

$$\|v\| = \|\Omega v\| = \|\lambda v\| = |\lambda|\,\|v\| \;.$$

But $\|v\| \neq 0$, and hence $|\lambda| = 1$.

(c) If $H^\dagger = -H$, $v \neq 0$, and $Hv = \lambda v$, then

$$\begin{aligned}\lambda\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Hv\rangle = \langle H^\dagger v, v\rangle = \langle -Hv, v\rangle = \langle -\lambda v, v\rangle\\ = -\lambda^*\langle v, v\rangle \;.\end{aligned}$$

But $\langle v, v\rangle \neq 0$, and hence $\lambda = -\lambda^*$. This shows that λ is pure imaginary.

(d) Let $P = S^\dagger S$ be a positive definite operator. If $v \neq 0$, then the fact that S is nonsingular means that $Sv \neq 0$, and hence $\langle Sv, Sv\rangle = \|Sv\|^2 > 0$. Then, for $Pv = (S^\dagger S)v = \lambda v$, we see that

$$\lambda\langle v, v\rangle = \langle v, \lambda v\rangle = \langle v, Pv\rangle = \langle v, (S^\dagger S)v\rangle = \langle Sv, Sv\rangle \;.$$

But $\langle v, v\rangle = \|v\|^2 > 0$ also, and therefore we must have $\lambda > 0$.

If P is positive, then S is singular and the only difference is that now for $v \neq 0$ we have $\langle Sv, Sv\rangle = \|Sv\|^2 \geq 0$ which implies that $\lambda \geq 0$. ∎

We say that an operator N is **normal** if $N^\dagger N = NN^\dagger$. Note this implies that for any $v \in V$ we have

$$\begin{aligned}\|Nv\|^2 = \langle Nv, Nv\rangle = \langle (N^\dagger N)v, v\rangle = \langle (NN^\dagger)v, v\rangle = \langle N^\dagger v, N^\dagger v\rangle\\ = \|N^\dagger v\|^2 \;.\end{aligned}$$

Now let λ be a complex number. It is easy to see that if N is normal then so is $N - \lambda 1$ since (from Theorem 10.3)

$$\begin{aligned}(N - \lambda 1)^\dagger(N - \lambda 1) = (N^\dagger - \lambda^* 1)(N - \lambda 1) = N^\dagger N - \lambda N^\dagger - \lambda^* N + \lambda^* \lambda 1\\ = (N - \lambda 1)(N^\dagger - \lambda^* 1) = (N - \lambda 1)(N - \lambda 1)^\dagger \;.\end{aligned}$$

Using $N - \lambda 1$ instead of N in the previous result we obtain

$$\|Nv - \lambda v\|^2 = \|N^\dagger v - \lambda^* v\|^2 \;.$$

Since the norm is positive definite, this equation proves the next theorem.

Theorem 10.10 Let N be a normal operator and let λ be an eigenvalue of N. Then $Nv = \lambda v$ if and only if $N^\dagger v = \lambda^* v$.

In words, if v is an eigenvector of a normal operator N with eigenvalue λ, then v is also an eigenvector of $N^\dagger$ with eigenvalue λ^*. (Note it is always true that if λ is an eigenvalue of an operator T, then λ^* will be an eigenvalue of $T^\dagger$. See Exercise 10.3.6.)

Corollary If N is normal and $Nv = 0$ for some $v \in V$, then $N^\dagger v = 0$.

Proof This follows from Theorem 10.10 by taking $\lambda = \lambda^* = 0$. Alternatively, using $N^\dagger N = NN^\dagger$ along with the fact that $Nv = 0$, we see that

$$\langle N^\dagger v, N^\dagger v\rangle = \langle v, (NN^\dagger)v\rangle = \langle v, (N^\dagger N)v\rangle = 0 \ .$$

Since the inner product is positive definite, this requires that $N^\dagger v = 0$. ∎

Theorem 10.11 (a) Eigenvectors belonging to *distinct* eigenvalues of a Hermitian operator are orthogonal.

(b) Eigenvectors belonging to *distinct* eigenvalues of an isometric operator are orthogonal. Hence the eigenvectors of a unitary operator are orthogonal.

(c) Eigenvectors belonging to *distinct* eigenvalues of a normal operator are orthogonal.

Proof As we note after the proof, Hermitian and unitary operators are special cases of normal operators, and hence parts (a) and (b) follow from part (c). However, it is instructive to give independent proofs of parts (a) and (b). Assume that T is an operator on a unitary space, and $Tv_i = \lambda_i v_i$ for $i = 1, 2$ with $\lambda_1 \neq \lambda_2$. We may then also assume without loss of generality that $\lambda_1 \neq 0$.

(a) If $T = T^\dagger$, then (using Theorem 10.9(a))

$$\begin{aligned}\lambda_2\langle v_1, v_2\rangle &= \langle v_1, \lambda_2 v_2\rangle = \langle v_1, Tv_2\rangle = \langle T^\dagger v_1, v_2\rangle = \langle Tv_1, v_2\rangle\\ &= \langle \lambda_1 v_1, v_2\rangle = \lambda_1{}^*\langle v_1, v_2\rangle = \lambda_1\langle v_1, v_2\rangle \ .\end{aligned}$$

But $\lambda_1 \neq \lambda_2$, and hence $\langle v_1, v_2\rangle = 0$.

(b) If T is isometric, then $T^\dagger T = 1$ and we have

$$\langle v_1, v_2\rangle = \langle v_1, (T^\dagger T)v_2\rangle = \langle Tv_1, Tv_2\rangle = \lambda_1{}^*\lambda_2\langle v_1, v_2\rangle \ .$$

But by Theorem 10.9(b) we have $|\lambda_1|^2 = \lambda_1{}^*\lambda_1 = 1$, and thus $\lambda_1{}^* = 1/\lambda_1$. Therefore, multiplying the above equation by λ_1, we see that $\lambda_1\langle v_1, v_2\rangle =$

$\lambda_2\langle v_1, v_2\rangle$ and hence, since $\lambda_1 \neq \lambda_2$, this shows that $\langle v_1, v_2\rangle = 0$.

(c) If T is normal, then

$$\langle v_1, Tv_2\rangle = \lambda_2\langle v_1, v_2\rangle$$

while on the other hand, using Theorem 10.10 we have

$$\langle v_1, Tv_2\rangle = \langle T^\dagger v_1, v_2\rangle = \langle \lambda_1{}^* v_1, v_2\rangle = \lambda_1\langle v_1, v_2\rangle \ .$$

Thus $\langle v_1, v_2\rangle = 0$ since $\lambda_1 \neq \lambda_2$. ∎

We note that if $H^\dagger = H$, then $H^\dagger H = HH = HH^\dagger$ so that any Hermitian operator is normal. Furthermore, if U is unitary, then $U^\dagger U = UU^\dagger$ (= 1) so that U is also normal.

A Hermitian operator T defined on a *real* inner product space is said to be **symmetric**. This is equivalent to requiring that with respect to an orthonormal basis, the matrix elements a_{ij} of T are given by

$$a_{ij} = \langle e_i, Te_j\rangle = \langle Te_i, e_j\rangle = \langle e_j, Te_i\rangle = a_{ji} \ .$$

Therefore, a symmetric operator is represented by a real symmetric matrix. It is also true that **antisymmetric** operators (i.e., $T^T = -T$) and anti-Hermitian operators ($H^\dagger = -H$) are normal. Therefore, part (a) and the unitary case in part (b) in the above theorem are really special cases of part (c).

Theorem 10.12 (a) Let T be an operator on a unitary space V, and let W be a T-invariant subspace of V. Then $W^\perp$ is invariant under $T^\dagger$.

(b) Let U be a unitary operator on a unitary space V, and let W be a U-invariant subspace of V. Then $W^\perp$ is also invariant under U.

Proof (a) For any $v \in W$ we have $Tv \in W$ since W is T-invariant. Let $w \in W^\perp$ be arbitrary. We must show that $T^\dagger w \in W^\perp$. But this is easy because

$$\langle T^\dagger w, v\rangle = \langle w, Tv\rangle = 0$$

by definition of $W^\perp$. Thus $T^\dagger w \in W^\perp$ so that $W^\perp$ is invariant under $T^\dagger$.

(b) The fact that U is unitary means $U^{-1} = U^\dagger$ exists, and hence U is nonsingular. In other words, for any $v' \in W$ there exists $v \in W$ such that $Uv = v'$. Now let $w \in W^\perp$ be arbitrary. Then

$$\langle Uw, v'\rangle = \langle Uw, Uv\rangle = \langle w, (U^\dagger U)v\rangle = \langle w, v\rangle = 0$$

by definition of $W^\perp$. Thus $Uw \in W^\perp$ so that $W^\perp$ is invariant under U. ∎

Recall from the discussion in Section 7.7 that the algebraic multiplicity of a given eigenvalue is the number of times the eigenvalue is repeated as a root of the characteristic polynomial. We also defined the geometric multiplicity as the number of linearly independent eigenvectors corresponding to this eigenvalue (i.e., the dimension of its eigenspace).

Theorem 10.13 Let H be a Hermitian operator on a finite-dimensional unitary space V. Then the algebraic multiplicity of any eigenvalue λ of H is equal to its geometric multiplicity.

Proof Let $V_\lambda = \{v \in V: Hv = \lambda v\}$ be the eigenspace corresponding to the eigenvalue λ. Furthermore, V_λ is obviously invariant under H since $Hv = \lambda v \in V_\lambda$ for every $v \in V_\lambda$. By Theorem 10.12(a), we then have that $V_\lambda{}^\perp$ is also invariant under $H^\dagger = H$. Furthermore, by Theorem 2.22 we see that $V = V_\lambda \oplus V_\lambda{}^\perp$. Applying Theorem 7.20, we may write $H = H_1 \oplus H_2$ where $H_1 = H|V_\lambda$ and $H_2 = H|V_\lambda{}^\perp$.

Let A be the matrix representation of H, and let A_i be the matrix representation of H_i $(i = 1, 2)$. By Theorem 7.20, we also have $A = A_1 \oplus A_2$. Using Theorem 4.14, it then follows that the characteristic polynomial of A is given by

$$\det(xI - A) = \det(xI - A_1)\det(xI - A_2)\ .$$

Now, H_1 is a Hermitian operator on the finite-dimensional space V_λ with only the single eigenvalue λ. Therefore λ is the only root of $\det(xI - A_1) = 0$, and hence it must occur with an algebraic multiplicity equal to the dimension of V_λ (since this is just the size of the matrix A_1). In other words, if $\dim V_\lambda = m$, then $\det(xI - A_1) = (x - \lambda^m)$. On the other hand, λ is not an eigenvalue of A_2 by definition, and hence $\det(xI - A_2) \neq 0$. This means that $\det(xI - A)$ contains $(x - \lambda)$ as a factor exactly m times. ∎

Corollary Any Hermitian operator H on a finite-dimensional unitary space V is diagonalizable.

Proof Since V is a unitary space, the characteristic polynomial of H will factor into (not necessarily distinct) linear terms. The conclusion then follows from Theorems 10.13 and 7.26. ∎

In fact, from Theorem 8.2 we know that any normal matrix is unitarily similar to a diagonal matrix. This means that given any normal operator $T \in L(V)$, there is an orthonormal basis for V that consists of eigenvectors of T.

We develop this result from an entirely different point of view in the next section.

Exercises

1. Let V be a unitary space and suppose $T \in L(V)$. Define $T_+ = (1/2)(T + T^\dagger)$ and $T_- = (1/2i)(T - T^\dagger)$.
   (a) Show that T_+ and T_- are Hermitian, and that $T = T_+ + iT_-$.
   (b) If $T|'_+$ and T'_- are Hermitian operators such that $T = T'_+ + iT'_-$, show that $T'_+ = T_+$ and $T'_- = T_-$.
   (c) Prove that T is normal if and only if $[T_+, T_-] = 0$.

2. Let N be a normal operator on a finite-dimensional inner product space V. Prove Ker N = Ker $N^\dagger$ and Im N = Im $N^\dagger$. [*Hint*: Prove that $(\text{Im } N^\dagger)^\perp$ = Ker N, and hence that Im $N^\dagger = (\text{Ker } N)^\perp$.]

3. Let V be a finite-dimensional inner product space, and suppose $T \in L(V)$ is both positive and unitary. Prove that T = 1.

4. Let $H \in M_n(\mathbb{C})$ be Hermitian. Then for any nonzero $x \in \mathbb{C}^n$ we define the **Rayleigh quotient** to be the number

$$R(x) = \frac{\langle x, Hx \rangle}{\|x\|^2}\ .$$

   Prove that max{R(x): x ≠ 0} is the largest eigenvalue of H, and that min{R(x): x ≠ 0} is the smallest eigenvalue of H.

5. Let V be a finite-dimensional unitary space, and suppose $E \in L(V)$ is such that $E^2 = E = E^\dagger$. Prove that $V = \text{Im } E \oplus (\text{Im } E)^\perp$.

6. If V is finite-dimensional and $T \in L(V)$ has eigenvalue λ, show that $T^\dagger$ has eigenvalue λ^*.

10.4 DIAGONALIZATION OF NORMAL OPERATORS

We now turn to the problem of diagonalizing operators. We will discuss several of the many ways to approach this problem. Because most commonly used operators are normal, we first treat this general case in detail, leaving unitary and Hermitian operators as obvious special cases. Next, we go back and consider the real and complex cases separately. In so doing, we will gain much insight into the structure of orthogonal and unitary transformations. While this problem was treated concisely in Chapter 8, we present an entirely different viewpoint in this section to acquaint the reader with other approaches found in the literature. If the reader has studied Chapter 8, he or she should keep in mind the rational and Jordan forms while reading this section, as many of our results (such as Theorem 10.16) follow almost trivially from our earlier work. We begin with some more elementary facts about normal transformations.

Theorem 10.14 Let V be a unitary space.

(a) If $T \in L(V)$ and $(T^\dagger T)v = 0$ for some $v \in V$, then $Tv = 0$.

(b) If H is Hermitian and $H^k v = 0$ for $k \geq 1$, then $Hv = 0$.

(c) If N is normal and $N^k v = 0$ for $k \geq 1$, then $Nv = 0$.

(d) If N is normal, and if $(N - \lambda 1)^k v = 0$ where $k \geq 1$ and $\lambda \in \mathbb{C}$, then $Nv = \lambda v$.

Proof (a) Since $(T^\dagger T)v = 0$, we have $0 = \langle v, (T^\dagger T)v\rangle = \langle Tv, Tv\rangle$ which implies that $Tv = 0$ because the inner product is positive definite.

(b) We first show that if $H^{2^m} v = 0$ for some positive integer m, then $Hv = 0$. To see this, let $T = H^{2^{m-1}}$ and note that $T^\dagger = T$ because H is Hermitian (by induction from Theorem 10.3(c)). Then $T^\dagger T = TT = H^{2^m}$, and hence

$$0 = \langle H^{2^m} v, v\rangle = \langle (T^\dagger T)v, v\rangle = \langle Tv, Tv\rangle$$

which implies that $0 = Tv = H^{2^{m-1}} v$. Repeating this process, we must eventually obtain $Hv = 0$.

Now, if $H^k v = 0$, then $H^{2^m} v = 0$ for any $2^m \geq k$, and therefore applying the above argument, we see that $Hv = 0$.

(c) Define the Hermitian operator $H = N^\dagger N$. Since N is normal, we see that

$$(N^\dagger N)^2 = N^\dagger N N^\dagger N = N^{\dagger 2} N^2$$

and by induction,

$$(N^\dagger N)^k = N^{\dagger k} N^k \; .$$

By hypothesis, we then find that

$$H^k v = (N^\dagger N)^k v = (N^{\dagger k} N^k) v = 0$$

and hence $(N^\dagger N)v = Hv = 0$ by part (b). But then $Nv = 0$ by part (a).

(d) Since N is normal, it follows that $N - \lambda 1$ is normal, and therefore by part (c) we have $(N - \lambda 1)v = 0$. ∎

Just as we did for operators, we say that a matrix N is **normal** if $N^\dagger N = NN^\dagger$. We now wish to show that any normal matrix can be diagonalized by a unitary similarity transformation. Another way to phrase this is as follows. We say that two matrices $A, B \in M_n(\mathbb{C})$ are **unitarily similar** (or **equivalent**) if there exists a unitary matrix $U \in M_n(\mathbb{C})$ such that $A = U^\dagger BU = U^{-1}BU$. Thus, we wish to show that any normal matrix is unitarily similar to a diagonal matrix. This extremely important result is quite easy to prove with what has already been shown. Let us first prove this in the case of normal operators over the complex field. (See Theorem 8.2 for another approach.)

Theorem 10.15 Let N be a normal operator on a finite-dimensional unitary space V. Then there exists an orthonormal basis for V consisting of eigenvectors of N in which the matrix of N is diagonal.

Proof Let $\lambda_1, \dots, \lambda_r$ be the distinct eigenvalues of the normal operator N. (These all exist in $\mathbb{C}$ by Theorems 6.12 and 6.13.) Then (by Theorem 7.13) the minimal polynomial m(x) for N must be of the form

$$m(x) = (x - \lambda_1)^{n_1} \cdots (x - \lambda_r)^{n_r}$$

where each $n_i \geq 1$. By the primary decomposition theorem (Theorem 7.23), we can write $V = W_1 \oplus \cdots \oplus W_r$ where $W_i = \operatorname{Ker}(N - \lambda_i 1)^{n_i}$. In other words,

$$(N - \lambda_i 1)^{n_i} v_i = 0$$

for every $v_i \in W_i$. By Theorem 10.14(d), we then have $Nv_i = \lambda_i v_i$ so that every $v_i \in W_i$ is an eigenvector of N with eigenvalue λ_i.

Now, the inner product on V induces an inner product on each subspace W_i in the usual and obvious way, and thus by the Gram-Schmidt process (Theorem 2.21), each W_i has an orthonormal basis relative to this induced inner product. Note that by the last result of the previous paragraph, this basis must consist of eigenvectors of N.

By Theorem 10.11(c), vectors in distinct W_i are orthogonal to each other. Therefore, according to Theorem 2.15, the union of the bases of the W_i forms a basis for V, which thus consists entirely of eigenvectors of N. By Theorem 7.14 then, the matrix of N is diagonal in this basis. (Alternatively, we see that the matrix elements n_{ij} of N relative to the eigenvector basis $\{e_i\}$ are given by $n_{ij} = \langle e_i, Ne_j\rangle = \langle e_i, \lambda_j e_j\rangle = \lambda_j \delta_{ij}$.) ∎

Corollary 1 Let N be a normal matrix over ℂ. Then there exists a unitary matrix U such that $U^{-1}NU = U^\dagger NU$ is diagonal. Moreover, the columns of U are just the eigenvectors of N, and the diagonal elements of $U^\dagger NU$ are the eigenvalues of N.

Proof The normal matrix N defines an operator on a finite-dimensional unitary space V with the standard orthonormal basis, and therefore by Theorem 10.15, V has an orthonormal basis of eigenvectors in which the matrix N is diagonal. By Theorem 10.6, any such change of basis in V is accomplished by a unitary transformation U, and by Theorem 5.18, the matrix of the operator relative to this new basis is related to the matrix N in the old basis by the similarity transformation $U^{-1}NU$ (= $U^\dagger NU$).

Now note that the columns of U are precisely the eigenvectors of N (see the discussion preceding Example 7.4). We also recall that Theorem 7.14 tells us that the diagonal elements of the diagonal form of N are exactly the eigenvalues of N. ∎

Corollary 2 A real symmetric matrix can be diagonalized by an orthogonal matrix.

Proof Note that a real symmetric matrix A may be considered as an operator on a finite-dimensional real inner product space V. If we think of A as a complex matrix that happens to have all real elements, then A is Hermitian and hence has all real eigenvalues. This means that all the roots of the minimal polynomial for A lie in ℝ. If $\lambda_1, \dots, \lambda_r$ are the distinct eigenvalues of A, then we may proceed exactly as in the proof of Theorem 10.15 and Corollary 1 to conclude that there exists a unitary matrix U that diagonalizes A. In this case, since $W_i = \mathrm{Ker}(A - \lambda_i I)^{n_i}$ and $A - \lambda_i I$ is real, it follows that the eigenvectors of A are real and hence U is actually an orthogonal matrix. ∎

Corollary 2 is also proved from an entirely different point of view in Exercise 10.4.9. This alternative approach has the advantage of presenting a very useful geometric picture of the diagonalization process.

Example 10.6 Let us diagonalize the real symmetric matrix

$$A = \begin{pmatrix} 2 & -2 \\ -2 & 5 \end{pmatrix}\ .$$

The characteristic polynomial of A is

$$\Delta_A(x) = \det(xI - A) = (x - 2)(x - 5) - 4 = (x - 1)(x - 6)$$

and therefore the eigenvalues of A are 1 and 6. To find the eigenvectors of A, we must solve the matrix equation $(\lambda_i I - A)v_i = 0$ for the vector v_i. For $\lambda_1 = 1$ we have $v_1 = (x_1, y_1)$, and hence we find the homogeneous system of equations

$$\begin{aligned} -x_1 + 2y_1 &= 0 \\ 2x_1 - 4y_1 &= 0\ . \end{aligned}$$

These imply that $x_1 = 2y_1$, and hence a nonzero solution is $v_1 = (2, 1)$. For $\lambda_2 = 6$ we have the equations

$$\begin{aligned} 4x_2 + 2y_2 &= 0 \\ 2x_2 + \ \ y_2 &= 0 \end{aligned}$$

which yields $v_2 = (1, -2)$.

Note that $\langle v_1, v_2\rangle = 0$ as it should according to Theorem 10.11, and that $\|v_1\| = \sqrt{5} = \|v_2\|$. We then take the normalized basis vectors to be $e_i = v_i/\sqrt{5}$ which are also eigenvectors of A. Finally, A is diagonalized by the orthogonal matrix P whose columns are just the e_i:

$$P = \begin{pmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{pmatrix}\ .$$

We leave it to the reader to show that

$$P^T AP = \begin{pmatrix} 1 & 0 \\ 0 & 6 \end{pmatrix}\ .$$

Another important point to notice is that Theorem 10.15 tells us that even though an eigenvalue λ of a normal operator N may be degenerate (i.e., have algebraic multiplicity $k > 1$), it is always possible to find k linearly independent eigenvectors belonging to λ. The easiest way to see this is to note that from Theorem 10.8 we have $|\det U| = 1 \neq 0$ for any unitary matrix U. This means that the columns of the diagonalizing matrix U (which are just the eigenvectors of N) must be linearly independent. This is in fact another proof that the algebraic and geometric multiplicities of a normal (and hence Hermitian) operator must be the same.

We now consider the case of real orthogonal transformations as independent operators, not as a special case of normal operators. First we need a general definition. Let V be an arbitrary finite-dimensional vector space over any field $\mathcal{F}$, and suppose $T \in L(V)$. A nonzero T-invariant subspace $W \subset V$ is said to be **irreducible** if the only T-invariant subspaces contained in W are $\{0\}$ and W.

Theorem 10.16 (a) Let V be a finite-dimensional vector space over an algebraically closed field $\mathcal{F}$, and suppose $T \in L(V)$. Then every irreducible T-invariant subspace W of V is of dimension 1.

(b) Let V be a finite-dimensional vector space over $\mathbb{R}$, and suppose $T \in L(V)$. Then every irreducible T-invariant subspace W of V is of dimension either 1 or 2.

Proof (a) Let W be an irreducible T-invariant subspace of V. Then the restriction T_W of T to W is just a linear transformation on W, where $T_W(w) = Tw \in W$ for every $w \in W$. Since $\mathcal{F}$ is algebraically closed, the characteristic polynomial of T_W has at least one root (i.e., eigenvalue λ) in $\mathcal{F}$. Therefore T has at least one (nonzero) eigenvector $v \in W$ such that $Tv = \lambda v \in W$. If we define $\mathcal{S}(v)$ to be the linear span of $\{v\}$, then $\mathcal{S}(v)$ is also a T-invariant subspace of W, and hence $\mathcal{S}(v) = W$ because W is irreducible. Therefore W is spanned by the single vector v, and hence $\dim W = 1$.

(b) Let W be an irreducible T-invariant subspace of V, and let m(x) be the minimal polynomial for T_W. Therefore, the fact that W is irreducible (so that W is not a direct sum of T-invariant subspaces) along with the primary decomposition theorem (Theorem 7.23) tells us that we must have $m(x) = f(x)^n$ where $f(x) \in \mathbb{R}[x]$ is a prime polynomial. Furthermore, if n were greater than 1, then we claim that $\operatorname{Ker} f(T)^{n-1}$ would be a T-invariant subspace of W (Theorem 7.18) that is different from $\{0\}$ and W.

To see this, first suppose that $\operatorname{Ker} f(T)^{n-1} = \{0\}$. Then the linear transformation $f(T)^{n-1}$ is one-to-one, and hence $f(T)^{n-1}(W) = W$. But then

$$0 = f(T)^n(W) = f(T)f(T)^{n-1}(W) = f(T)(W) \ .$$

However, $f(T)W \neq 0$ by definition of m(x), and hence this contradiction shows that we can not have $\operatorname{Ker} f(T)^{n-1} = \{0\}$. Next, if we had $\operatorname{Ker} f(T)^{n-1} = W$, this would imply that $f(T)^{n-1}(W) = 0$ which contradicts the definition of minimal polynomial. Therefore we must have $n = 1$ and $m(x) = f(x)$.

Since $m(x) = f(x)$ is prime, it follows from the corollary to Theorem 6.15 that we must have either $m(x) = x - a$ or $m(x) = x^2 + ax + b$ with $a^2 - 4ab < 0$. If $m(x) = x - a$, then there exists an eigenvector $v \in W$ with $Tv = av \in W$, and hence $\mathcal{S}(v) = W$ as in part (a). If $m(x) = x^2 + ax + b$, then for any nonzero $w \in W$ we have

$$0 = m(T)w = T^2w + aTw + bw$$

and hence

$$T^2w = T(Tw) = -aTw - bw \in W .$$

Thus $\mathcal{S}(w, Tw)$ is a T-invariant subspace of W with dimension either 1 or 2. However W is irreducible, and therefore we must have $W = \mathcal{S}(w, Tw)$. ∎

Theorem 10.17 Let V be a finite-dimensional Euclidean space, let $T \in L(V)$ be an orthogonal transformation, and let W be an irreducible T-invariant subspace of V. Then one of the following two conditions holds:

(a) dim W = 1, and for any nonzero $w \in W$ we have $Tw = \pm w$.

(b) dim W = 2, and there exists an orthonormal basis $\{e_1, e_2\}$ for W such that the matrix representation of T_W relative to this basis has the form

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} .$$

Proof That dim W equals 1 or 2 follows from Theorem 10.16(b). If dim W = 1, then (since W is T-invariant) there exists $\lambda \in \mathbb{R}$ such that $Tw = \lambda w$ for any (fixed) $w \in W$. But T is orthogonal so that

$$\|w\| = \|Tw\| = \|\lambda w\| = |\lambda|\|w\|$$

and hence $|\lambda| = 1$. This shows that $Tw = \lambda w = \pm w$.

If dim W = 2, then the desired form of the matrix of T_W follows essen–tially from Example 10.5. Alternatively, we know that W has an orthonormal basis $\{e_1, e_2\}$ by the Gram-Schmidt process. If we write $Te_1 = ae_1 + be_2$, then $\|Te_1\| = \|e_1\| = 1$ implies that $a^2 + b^2 = 1$. If we also write $Te_2 = ce_1 + de_2$, then similarly $c^2 + d^2 = 1$. Using $\langle Te_1, Te_2\rangle = \langle e_1, e_2\rangle = 0$ we find $ac + bd = 0$, and hence $c = -bd/a$. But then $1 = d^2(1 + b^2/a^2) = d^2/a^2$ so that $a^2 = d^2$ and $c^2 = b^2$. This means that $Te_2 = \pm(-be_1 + ae_2)$. If $Te_2 = -be_1 + ae_2$, then the matrix of T is of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

and we may choose $\theta \in \mathbb{R}$ such that $a = \cos\theta$ and $b = \sin\theta$ (since $\det T = a^2 + b^2 = 1$). However, if $Te_2 = be_1 - ae_2$, then the matrix of T is

$$\begin{pmatrix} a & b \\ b & -a \end{pmatrix}$$

which satisfies the equation $x^2 - 1 = (x - 1)(x + 1) = 0$ (and has $\det T = -1$). But if T satisfied this equation, then (by the primary decomposition theorem (Theorem 7.23)) W would be a direct sum of subspaces, in contradiction to the assumed irreducibility of W. Therefore only the first case can occur. ∎

This theorem becomes quite useful when combined with the next result.

Theorem 10.18 Let T be an orthogonal operator on a finite-dimensional Euclidean space V. Then $V = W_1 \oplus \cdots \oplus W_r$ where each W_i is an irreducible T-invariant subspace of V such that vectors belonging to distinct subspaces W_i and W_j are orthogonal.

Proof If $\dim V = 1$ there is nothing to prove, so we assume $\dim V > 1$ and that the theorem is true for all spaces of dimension less than dim V. Let W_1 be a nonzero T-invariant subspace of least dimension. Then W_1 is necessarily irreducible. By Theorem 2.22 we know that $V = W_1 \oplus W_1^\perp$ where $\dim W_1^\perp <$ dim V, and hence we need only show that $W_1^\perp$ is also T-invariant. But this follows from Theorem 10.12(b) applied to real unitary transformations (i.e., orthogonal transformations). This also means that $T(W_1^\perp) \subset W_1^\perp$, and hence T is an orthogonal transformation on $W_1^\perp$ (since it takes vectors in $W_1^\perp$ to vectors in $W_1^\perp$). By induction, $W_1^\perp$ is a direct sum of pairwise orthogonal irreducible T-invariant subspaces, and therefore so is $V = W_1 \oplus W_1^\perp$. ∎

From Theorem 10.18, we see that if we are given an orthogonal transformation T on a finite-dimensional Euclidean space V, then $V = W_1 \oplus \cdots \oplus W_r$ is the direct sum of pairwise orthogonal irreducible T-invariant subspaces W_i. But from Theorem 10.17, we see that any such subspace W_i is of dimension either 1 or 2. Moreover, Theorem 10.17 also showed that if $\dim W_i = 1$, then the matrix of $T|W_i$ is either (1) or (−1), and if $\dim W_i = 2$, then the matrix of $T|W_i$ is just the rotation matrix R_i given by

$$R_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix} .$$

Since each W_i has an orthonormal basis and the bases of distinct W_i are orthogonal, it follows that we can find an orthonormal basis for V in which the matrix of T takes the block diagonal form (see Theorem 7.20)

$$(1) \oplus \cdots \oplus (1) \oplus (-1) \oplus \cdots \oplus (-1) \oplus R_1 \oplus \cdots \oplus R_m \ .$$

These observations prove the next theorem.

Theorem 10.19 Let T be an orthogonal transformation on a finite-dimensional Euclidean space V. Then there exists an orthonormal basis for V in which the matrix representation of T takes the block diagonal form

$$M_1 \oplus \cdots \oplus M_r$$

where each M_i is one of the following: (+1), (−1), or R_i .

Exercises

1. Prove that any nilpotent normal operator is necessarily the zero operator.

2. Let A and B be normal operators on a finite-dimensional unitary space V. For notational simplicity, let v_a denote an eigenvector of A corresponding to the eigenvalue a, let v_b be an eigenvector of B corresponding to the eigenvalue b, and let v_{ab} denote a simultaneous eigenvector of A and B, i.e., $Av_{ab} = av_{ab}$ and $Bv_{ab} = bv_{ab}$.
(a) If there exists a basis for V consisting of simultaneous eigenvectors of A and B, show that the commutator $[A, B] = AB - BA = 0$.
(b) If $[A, B] = 0$, show that there exists a basis for V consisting entirely of simultaneous eigenvectors of A and B. In other words, if $[A, B] = 0$, then A and B can be simultaneously diagonalized. [*Hint*: There are several ways to approach this problem. One way follows easily from Exercise 8.1.3. Another intuitive method is as follows. First assume that at least one of the operators, say A, is nondegenerate. Show that Bv_a is an eigenvector of A, and that $Bv_a = bv_a$ for some scalar b. Next assume that both A and B are degenerate. Then $Av_{a,i} = av_{a,i}$ where the $v_{a,i}$ $(i = 1, \ldots, m_a)$ are linearly independent eigenvectors corresponding to the eigenvalue a of multiplicity m_a. What does the matrix representation of A look like in the $\{v_{a,i}\}$ basis? Again consider $Bv_{a,i}$. What does the matrix representation of B look like? Now what happens if you diagonalize B?]

3. If N_1 and N_2 are commuting normal operators, show that the product N_1N_2 is normal.

4. Let V be a finite-dimensional complex (real) inner product space, and suppose $T \in L(V)$. Prove that V has an orthonormal basis of eigenvectors of T with corresponding eigenvalues of absolute value 1 if and only if T is unitary (Hermitian and orthogonal).

5. For each of the following matrices A, find an orthogonal or unitary matrix P and a diagonal matrix D such that P†AP = D:

$$(a)\ \begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix}\qquad (b)\ \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\qquad (c)\ \begin{pmatrix}2 & 3-i3\\ 3+i3 & 5\end{pmatrix}$$

$$(d)\ \begin{pmatrix}0 & 2 & 2\\ 2 & 0 & 2\\ 2 & 2 & 0\end{pmatrix}\qquad (e)\ \begin{pmatrix}2 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 2\end{pmatrix}$$

6. Let A, B and C be normal operators on a finite-dimensional unitary space, and assume that [A, B] = 0 but [B, C] ≠ 0. If all of these operators are nondegenerate (i.e., all eigenvalues have multiplicity equal to 1), is it true that [A, C] ≠ 0? Explain. What if any of these are degenerate?

7. Let V be a finite-dimensional unitary space and suppose A ∈ L(V).
(a) Prove that Tr(AA†) = 0 if and only if A = 0.
(b) Suppose N ∈ L(V) is normal and AN = NA. Prove that AN† = N†A.

8. Let A be a positive definite real symmetric matrix on an n-dimensional Euclidean space V. Using the single variable formula (where a > 0)

$$\int_{-\infty}^{\infty}\exp(-ax^2/2)\,dx = (2\pi/a)^{1/2}$$

show that

$$\int_{-\infty}^{\infty}\exp[(-1/2)\langle\vec{x},\ A\vec{x}\rangle]\,d^n x = (2\pi)^{n/2}(\det A)^{-1/2}$$

where $d^n x = dx_1 \cdots dx_n$. [*Hint*: First consider the case where A is diagonal.]

9. (This is an independent proof of Corollary 2 of Theorem 10.15.) Let A = $(a_{ij}) \in M_3(\mathbb{R})$ be a real symmetric matrix. Thus A: $\mathbb{R}^3 \to \mathbb{R}^3$ is a Hermitian linear operator with respect to the inner product ⟨ , ⟩. Prove there exists an orthonormal basis of eigenvectors of A using the following approach. (It should be clear after you have done this that the same proof will work in $\mathbb{R}^n$ just as well.)
(a) Let S^2 be the unit sphere in $\mathbb{R}^3$, and define f: $S^2 \to \mathbb{R}$ by

$$f(x) = \langle Ax, x\rangle\ .$$

Let $M = \sup f(x)$ and $m = \inf f(x)$ where the sup and inf are taken over S^2. Show that there exist points $x_1, x_1' \in S^2$ such that $f(x_1) = M$ and $f(x_1') = m$. [*Hint*: Use Theorem A15.]

(b) Let $C = x(t)$ be any curve on S^2 such that $x(0) = x_1$, and let a dot denote differentiation with respect to the parameter t. Note that $\dot{x}(t)$ is tangent to C, and hence also to S^2. Show that $\langle Ax_1, \dot{x}(0)\rangle = 0$, and thus deduce that Ax_1 is normal to the tangent plane at x_1. [*Hint*: Consider $df(x(t))/dt|_{t=0}$ and note that C is arbitrary.]

(c) Show that $\langle \dot{x}(t), x(t)\rangle = 0$, and hence conclude that $Ax_1 = \lambda_1 x_1$. [*Hint*: Recall that S^2 is the *unit* sphere.]

(d) Argue that $Ax_1' = \lambda_1' x_1'$, and in general, that any critical point of $f(x) = \langle Ax, x\rangle$ on the unit sphere will be an eigenvector of A with critical value (i.e., eigenvalue) $\lambda_i = \langle Ax_i, x_i\rangle$. (A **critical point** of f(x) is a point x_0 where $df/dx = 0$, and the **critical value** of f is just $f(x_0)$.)

(e) Let $[x_1]$ be the 1-dimensional subspace of $\mathbb{R}^3$ spanned by x_1. Show that $[x_1]$ and $[x_1]^\perp$ are both A-invariant subspaces of $\mathbb{R}^3$, and hence that A is Hermitian on $[x_1]^\perp \subset \mathbb{R}^2$. Note that $[x_1]^\perp$ is a plane through the origin of S^2.

(f) Show that f now must achieve its maximum at a point x_2 on the unit circle $S^1 \subset [x_1]^\perp$, and that $Ax_2 = \lambda_2 x_2$ with $\lambda_2 \le \lambda_1$.

(g) Repeat this process again by considering the space $[x_2]^\perp \subset [x_1]^\perp$, and show there exists a vector $x_3 \in [x_2]^\perp$ with $Ax_3 = \lambda_3 x_3$ and $\lambda_3 \le \lambda_2 \le \lambda_1$.

10.5 THE SPECTRAL THEOREM

We now turn to another major topic of this chapter, the so-called spectral theorem. This important result is actually nothing more than another way of looking at Theorems 8.2 and 10.15. We begin with a simple version that is easy to understand and visualize if the reader will refer back to the discussion prior to Theorem 7.29.

Theorem 10.20 Suppose $A \in M_n(\mathbb{C})$ is a diagonalizable matrix with distinct eigenvalues $\lambda_1, \dots, \lambda_r$. Then A can be written in the form

$$A = \lambda_1 E_1 + \cdots + \lambda_r E_r$$

where the E_i are n x n matrices with the following properties:

(a) Each E_i is idempotent (i.e., $E_i^2 = E_i$).

(b) $E_i E_j = 0$ for $i \neq j$.

(c) $E_1 + \cdots + E_r = I$.
(d) $AE_i = E_iA$ for every E_i.

Proof Since A is diagonalizable by assumption, let $D = P^{-1}AP$ be the diagonal form of A for some nonsingular matrix P (whose columns are just the eigenvectors of A). Remember that the diagonal elements of D are just the eigenvalues λ_i of A. Let P_i be the n x n diagonal matrix with diagonal element 1 wherever a λ_i occurs in D, and 0's everywhere else. It should be clear that the collection $\{P_i\}$ obeys properties (a) – (c), and that

$$P^{-1}AP = D = \lambda_1 P_1 + \cdots + \lambda_r P_r \ .$$

If we now define $E_i = PP_iP^{-1}$, then we have

$$A = PDP^{-1} = \lambda_1 E_1 + \cdots + \lambda_r E_r$$

where the E_i also obey properties (a) – (c) by virtue of the fact that the P_i do. Using (a) and (b) in this last equation we find

$$AE_i = (\lambda_1 E_1 + \cdots + \lambda_r E_r)E_i = \lambda_i E_i$$

and similarly it follows that $E_iA = \lambda_i E_i$ so that each E_i commutes with A, i.e., $E_iA = AE_i$. ∎

By way of terminology, the collection of eigenvalues $\lambda_1, \ldots, \lambda_r$ is called the **spectrum** of A, the sum $E_1 + \cdots + E_r = I$ is called the **resolution of the identity induced** by A, and the expression $A = \lambda_1 E_1 + \cdots + \lambda_r E_r$ is called the **spectral decomposition** of A. These definitions also apply to arbitrary normal operators as in Theorem 10.22 below.

Corollary Let A be diagonalizable with spectral decomposition as in Theorem 10.20. If $f(x) \in \mathbb{C}[x]$ is any polynomial, then

$$f(A) = f(\lambda_1)E_1 + \cdots + f(\lambda_r)E_r \ .$$

Proof Using properties (a) – (c) in Theorem 10.20, it is easy to see that for any $m > 0$ we have

$$A^m = \lambda_1{}^m E_1 + \cdots + \lambda_r{}^m E_r \ .$$

The result for arbitrary polynomials now follows easily from this result. ∎

Before turning to our proof of the spectral theorem, we first prove a simple but useful characterization of orthogonal projections.

Theorem 10.21 Let V be an inner product space and suppose $E \in L(V)$. Then E is an orthogonal projection if and only if $E^2 = E = E^\dagger$.

Proof We first assume that E is an orthogonal projection. By definition this means that $E^2 = E$, and hence we must show that $E^\dagger = E$. From Theorem 7.27 we know that $V = \text{Im } E \oplus \text{Ker } E = \text{Im } E \oplus (\text{Im } E)^\perp$. Suppose $v, w \in V$ are arbitrary. Then we may write $v = v_1 + v_2$ and $w = w_1 + w_2$ where $v_1, w_1 \in \text{Im } E$ and $v_2, w_2 \in (\text{Im } E)^\perp$. Therefore

$$\langle v, Ew\rangle = \langle v_1 + v_2, w_1\rangle = \langle v_1, w_1\rangle + \langle v_2, w_1\rangle = \langle v_1, w_1\rangle$$

and

$$\langle v, E^\dagger w\rangle = \langle Ev, w\rangle = \langle v_1, w_1 + w_2\rangle = \langle v_1, w_1\rangle + \langle v_1, w_2\rangle = \langle v_1, w_1\rangle \ .$$

In other words, $\langle v, (E - E^\dagger)w\rangle = 0$ for all $v, w \in V$, and hence $E = E^\dagger$ (by Theorem 10.4(a)).

On the other hand, if $E^2 = E = E^\dagger$, then we know from Theorem 7.27 that E is a projection of V on Im E in the direction of Ker E, i.e., $V = \text{Im } E \oplus \text{Ker } E$. Therefore, we need only show that Im E and Ker E are orthogonal subspaces. To show this, let $w \in \text{Im } E$ and $w' \in \text{Ker } E$ be arbitrary. Then $Ew = w$ and $Ew' = 0$ so that

$$\langle w', w\rangle = \langle w', Ew\rangle = \langle E^\dagger w', w\rangle = \langle Ew', w\rangle = 0 \ .$$

(This was also proved independently in Exercise 10.3.5.) ∎

We are now in a position to prove the spectral theorem for normal operators. In order to distinguish projection operators from their matrix representations in this theorem, we denote the operators by π_i and the corresponding matrices by E_i.

Theorem 10.22 (Spectral Theorem for Normal Operators) Let V be a finite-dimensional unitary space, and let N be a normal operator on V with distinct eigenvalues $\lambda_1, \dots, \lambda_r$. Then

(a) $N = \lambda_1 \pi_1 + \cdots + \lambda_r \pi_r$ where each π_i is the orthogonal projection of V onto a subspace $W_i = \text{Im } \pi_i$.

(b) $\pi_i \pi_j = 0$ for $i \neq j$.

(c) $\pi_1 + \cdots + \pi_r = 1$.
(d) $V = W_1 \oplus \cdots \oplus W_r$ where the subspaces W_i are mutually orthogonal.
(e) $W_j = \operatorname{Im} \pi_j = \operatorname{Ker}(N - \lambda_j 1)$ is the eigenspace corresponding to λ_j.

Proof Choose any orthonormal basis $\{e_i\}$ for V, and let A be the matrix representation of N relative to this basis. As discussed following Theorem 7.6, the normal matrix A has the same eigenvalues as the normal operator N. By Corollary 1 of Theorem 10.15 we know that A is diagonalizable, and hence applying Theorem 10.20 we may write

$$A = \lambda_1 E_1 + \cdots + \lambda_r E_r$$

where $E_i{}^2 = E_i$, $E_i E_j = 0$ if $i \neq j$, and $E_1 + \cdots + E_r = I$. Furthermore, A is diagonalized by a unitary matrix P, and as we saw in the proof of Theorem 10.20, $E_i = PP_iP^\dagger$ where each P_i is a real diagonal matrix. Since each P_i is clearly Hermitian, this implies that $E_i{}^\dagger = E_i$, and hence each E_i is an orthogonal projection (Theorem 10.21).

Now define $\pi_i \in L(V)$ as that operator whose matrix representation relative to the basis $\{e_i\}$ is just E_i. From the isomorphism between linear transformations and their representations (Theorem 5.13), it should be clear that

$$\begin{aligned} N &= \lambda_1\pi_1 + \cdots + \lambda_r\pi_r \\ \pi_i{}^\dagger &= \pi_i \\ \pi_i{}^2 &= \pi_i \\ \pi_i\pi_j &= 0 \quad \text{for } i \neq j \\ \pi_1 + \cdots + \pi_r &= 1\ . \end{aligned}$$

Since $\pi_i{}^2 = \pi_i = \pi_i{}^\dagger$, Theorem 10.21 tells us that each π_i is an orthogonal projection of V on the subspace $W_i = \operatorname{Im} \pi_i$. Since $\pi_1 + \cdots + \pi_r = 1$, we see that for any $v \in V$ we have $v = \pi_1 v + \cdots + \pi_r v$ so that $V = W_1 + \cdots + W_r$. To show that this sum is direct suppose, for example, that

$$w_1 \in W_1 \cap (W_2 + \cdots + W_r)\ .$$

This means that $w_1 = w_2 + \cdots + w_r$ where $w_i \in W_i$ for each $i = 1, \ldots, r$. Since $w_i \in W_i = \operatorname{Im} \pi_i$, it follows that there exists $v_i \in V$ such that $\pi_i v_i = w_i$ for each i. Then

$$w_i = \pi_i v_i = \pi_i{}^2 v_i = \pi_i w_i$$

and if $i \neq j$, then $\pi_i\pi_j = 0$ implies

$$\pi_i w_j = (\pi_i \pi_j) v_j = 0 .$$

Applying π_1 to $w_1 = w_2 + \cdots + w_r$, we obtain $w_1 = \pi_1 w_1 = 0$. Hence we have shown that $W_1 \cap (W_2 + \cdots + W_r) = \{0\}$. Since this argument can clearly be applied to any of the W_i , we have proved that $V = W_1 \oplus \cdots \oplus W_r$.

Next we note that for each i, π_i is the *orthogonal* projection of V on $W_i = \operatorname{Im} \pi_i$ in the direction of $W_i{}^\perp = \operatorname{Ker} \pi_i$, so that $V = W_i \oplus W_i{}^\perp$. Therefore, since $V = W_1 \oplus \cdots \oplus W_r$, it follows that for each $j \neq i$ we must have $W_j \subset W_i{}^\perp$, and hence the subspaces W_i must be mutually orthogonal. Finally, the fact that $W_j = \operatorname{Ker}(N - \lambda_j 1)$ was proved in Theorem 7.29. ∎

The observant reader will have noticed the striking similarity between the spectral theorem and Theorem 7.29. In fact, part of Theorem 10.22 is essentially a corollary of Theorem 7.29. This is because a normal operator is diagonalizable, and hence satisfies the hypotheses of Theorem 7.29. However, note that in the present case we have used the existence of an inner product in our proof, whereas in Chapter 7, no such structure was assumed to exist. We leave it to the reader to use Theorems 10.15 and 7.28 to construct a simple proof of the spectral theorem that makes no reference to any matrix representation of the normal operator (see Exercise 10.5.1).

Theorem 10.23 Let $\sum_{j=1}^{r} \lambda_j E_j$ be the spectral decomposition of a normal operator N on a finite-dimensional unitary space. Then for each i = 1, . . . , r there exists a polynomial $f_i(x) \in \mathbb{C}[x]$ such that $f_i(\lambda_j) = \delta_{ij}$ and $f_i(N) = E_i$.

Proof For each i = 1, . . . , r we must find a polynomial $f_i(x) \in \mathbb{C}[x]$ with the property that $f_i(\lambda_j) = \delta_{ij}$. It should be obvious that the polynomials $f_i(x)$ defined by

$$f_i(x) = \prod_{j \neq i} \frac{x - \lambda_j}{\lambda_i - \lambda_j}$$

have this property. From the corollary to Theorem 10.20 we have $p(N) = \sum_j p(\lambda_j) E_j$ for any $p(x) \in \mathbb{C}[x]$, and hence

$$f_i(N) = \textstyle\sum_j f_i(\lambda_j) E_j = \sum_j \delta_{ij} E_j = E_i$$

as required. ∎

Exercises

1. Use Theorems 10.15 and 7.28 to construct a proof of Theorem 10.22 that makes no reference to any matrix representations.

2. Let N be an operator on a finite-dimensional unitary space. Prove that N is normal if and only if $N^\dagger = g(N)$ for some polynomial g. [*Hint*: If N is normal with eigenvalues $\lambda_1, \dots, \lambda_r$, use Exercise 6.4.2 to show the existence of a polynomial g such that $g(\lambda_i) = \lambda_i^*$ for each i.]

3. Let T be an operator on a finite-dimensional unitary space. Prove that T is unitary if and only if T is normal and $|\lambda| = 1$ for every eigenvalue λ of T.

4. Let H be a normal operator on a finite-dimensional unitary space. Prove that H is Hermitian if and only if every eigenvalue of H is real.

10.6 THE MATRIX EXPONENTIAL SERIES

We now use Theorem 10.20 to prove a very useful result, namely, that any unitary matrix U can be written in the form e^{iH} for some Hermitian matrix H. Before proving this however, we must first discuss some of the theory of sequences and series of matrices. In particular, we must define just what is meant by expressions of the form e^{iH}. If the reader already knows something about sequences and series of numbers, then the rest of this section should present no difficulty. However, for those readers who may need some review, we have provided all of the necessary material in Appendix B.

Let $\{S_r\}$ be a sequence of complex matrices where each $S_r \in M_n(\mathbb{C})$ has entries $s^{(r)}{}_{ij}$. We say that $\{S_r\}$ **converges** to the **limit** $S = (s_{ij}) \in M_n(\mathbb{C})$ if each of the n^2 sequences $\{s^{(r)}{}_{ij}\}$ converges to a limit s_{ij}. We then write $S_r \to S$ or $\lim_{r \to \infty} S_r = S$ (or even simply $\lim S_r = S$). In other words, a sequence $\{S_r\}$ of matrices converges if and only if every entry of S_r forms a convergent sequence.

Similarly, an infinite series of matrices

$$\sum_{r=1}^{\infty} A_r$$

where $A_r = (a^{(r)}{}_{ij})$ is said to be **convergent** to the **sum** $S = (s_{ij})$ if the sequence of partial sums

$$S_m = \sum_{r=1}^{m} A_r$$

converges to S. Another way to say this is that the series ΣA_r converges to S if and only if each of the n^2 series $\Sigma a^{(r)}{}_{ij}$ converges to s_{ij} for each i, j = 1, . . . , n. We adhere to the convention of leaving off the limits in a series if they are infinite.

Our next theorem proves several intuitively obvious properties of sequences and series of matrices.

Theorem 10.24 (a) Let $\{S_r\}$ be a convergent sequence of n x n matrices with limit S, and let P be any n x n matrix. Then $PS_r \to PS$ and $S_rP \to SP$.

(b) If $S_r \to S$ and P is nonsingular, then $P^{-1}S_rP \to P^{-1}SP$.

(c) If ΣA_r converges to A and P is nonsingular, then $\Sigma P^{-1}A_rP$ converges to $P^{-1}AP$.

Proof (a) Since $S_r \to S$, we have $\lim s^{(r)}{}_{ij} = s_{ij}$ for all i, j = 1, . . . , n. Therefore

$$\lim(PS_r)_{ij} = \lim(\Sigma_k p_{ik} s^{(r)}{}_{kj}) = \Sigma_k p_{ik} \lim s^{(r)}{}_{kj} = \Sigma_k p_{ik} s_{kj} = (PS)_{ij} \ .$$

Since this holds for all i, j = 1, . . . , n we must have $PS_r \to PS$. It should be obvious that we also have $S_r P \to SP$.

(b) As in part (a), we have

$$\begin{aligned} \lim(P^{-1}S_rP)_{ij} &= \lim(\Sigma_{k,m} p^{-1}{}_{ik} s^{(r)}{}_{km} p_{mj}) \\ &= \Sigma_{k,m} p^{-1}{}_{ik} p_{mj} \lim s^{(r)}{}_{km} \\ &= \Sigma_{k,m} p^{-1}{}_{ik} p_{mj} s_{km} \\ &= (P^{-1}SP)_{ij} \ . \end{aligned}$$

Note that we may use part (a) to formally write this as

$$\lim(P^{-1}S_rP) = P^{-1}\lim(S_rP) = P^{-1}SP \ .$$

(c) If we write the m*th* partial sum as

$$S_m = \sum_{r=1}^{m} P^{-1}A_rP = P^{-1}\left(\sum_{r=1}^{m} A_r\right)P$$

then we have

$$\begin{aligned}\lim_{m\to\infty}(S_m)_{ij} &= \sum_{k,\,l}\lim\left\{p^{-1}{}_{ik}\left(\sum_{r=1}^{m}a^{(r)}{}_{kl}\right)p_{lj}\right\}\\ &= \sum_{k,\,l}p^{-1}{}_{ik}p_{lj}\lim\sum_{r=1}^{m}a^{(r)}{}_{kl}\\ &= \sum_{k,\,l}p^{-1}{}_{ik}p_{lj}a_{kl}\\ &= P^{-1}AP\ .\ \blacksquare\end{aligned}$$

Theorem 10.25 For any $A = (a_{ij}) \in M_n(\mathbb{C})$ the following series converges:

$$\sum_{r=0}^{\infty}\frac{A_r}{r!} = I + A + \frac{A^2}{2!} + \cdots + \frac{A^r}{r!} + \cdots\ .$$

Proof Choose a positive real number $M > \max\{n, |a_{ij}|\}$ where the max is taken over all $i, j = 1, \ldots, n$. Then $|a_{ij}| < M$ and $n < M < M^2$. Now consider the term $A^2 = (b_{ij}) = (\sum_k a_{ik}a_{kj})$. We have (by Theorem 2.17, property (N3))

$$|b_{ij}| \le \sum_{k=1}^{n}|a_{ik}||a_{kj}| < \sum_{k=1}^{n}M^2 = nM^2 < M^4\ .$$

Proceeding by induction, suppose that for $A^r = (c_{ij})$, it has been shown that $|c_{ij}| < M^{2r}$. Then $A^{r+1} = (d_{ij})$ where

$$|d_{ij}| \le \sum_{k=1}^{n}|a_{ik}||c_{kj}| < nMM^{2r} = nM^{2r+1} < M^{2(r+1)}\ .$$

This proves that $A^r = (a^{(r)}{}_{ij})$ has the property that $|a^{(r)}{}_{ij}| < M^{2r}$ for every $r \ge 1$.

Now, for each of the n^2 terms $i, j = 1, \ldots, n$ we have

$$\sum_{r=0}^{\infty}\frac{|a^{(r)}{}_{ij}|}{r!} < \sum_{r=0}^{\infty}\frac{M^{2r}}{r!} = \exp(M^2)$$

so that each of these n^2 series (i.e., for each $i, j = 1, \ldots, n$) must converge (Theorem B26(a)). Hence the series $I + A + A^2/2! + \cdots$ must converge (Theorem B20). $\blacksquare$

We call the series in Theorem 10.25 the **matrix exponential series**, and denote its sum by $e^A = \exp A$. In general, the series for e^A is extremely difficult, if not impossible, to evaluate. However, there are important exceptions.

Example 10.7 Let A be the diagonal matrix

$$A = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} .$$

Then it is easy to see that

$$A^r = \begin{pmatrix} {\lambda_1}^r & 0 & \cdots & 0 \\ 0 & {\lambda_2}^r & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & {\lambda_n}^r \end{pmatrix}$$

and hence

$$\exp A = I + A + \frac{A^2}{2!} + \cdots = \begin{pmatrix} e^{\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n} \end{pmatrix} . \;/\!\!/$$

Example 10.8 Consider the 2 x 2 matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and let

$$A = \theta J = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}$$

where $\theta \in \mathbb{R}$. Then noting that $J^2 = -I$, we see that $A^2 = -\theta^2 I$, $A^3 = -\theta^3 J$, $A^4 = \theta^4 I$, $A^5 = \theta^5 J$, $A^6 = -\theta^6 I$, and so forth. From elementary calculus we know that

$$\sin\theta = \theta - \theta^3/3! + \theta^5/5! - \cdots$$

and

$$\cos\theta = 1 - \theta^2/2! + \theta^4/4! - \cdots$$

and hence

$$\begin{aligned} e^A &= I + A + A^2/2! + \cdots \\ &= I + \theta J - \theta^2 I/2! - \theta^3 J/3! + \theta^4 I/4! + \theta^5 J/5! - \theta^6 I/6! + \cdots \\ &= I(1 - \theta^2/2! + \theta^4/4! - \cdots) + J(\theta - \theta^3/3! + \theta^5/5! - \cdots) \\ &= (\cos\theta)I + (\sin\theta)J \ . \end{aligned}$$

In other words, using the explicit forms of I and J we see that

$$e^A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

so that $e^{\theta J}$ represents a rotation in $\mathbb{R}^2$ by an angle θ. ∕∕

Theorem 10.26 Let $A \in M_n(\mathbb{C})$ be diagonalizable, and let $\lambda_1, \ldots, \lambda_r$ be the distinct eigenvalues of A. Then the matrix power series

$$\sum_{s=0}^{\infty} a_s A^s$$

converges if and only if the series

$$\sum_{s=0}^{\infty} a_s \lambda_i^{\,s}$$

converges for each i = 1, . . . , r.

Proof Since A is diagonalizable, choose a nonsingular matrix P such that $D = P^{-1}AP$ is diagonal. It is then easy to see that for every s ≥ 1 we have

$$a_s D^s = a_s P^{-1} A^s P = P^{-1} a_s A^s P$$

where the n diagonal entries of D^s are just the numbers $\lambda_i^{\,s}$. By Theorem 10.24(c), we know that $\sum a_s A^s$ converges if and only if $\sum a_s D^s$ converges. But by definition of series convergence, $\sum a_s D^s$ converges if and only if $\sum a_s \lambda_i^{\,s}$ converges for every i = 1, . . . , r. ∎

Theorem 10.27 Let $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots$ be any power series with coefficients in $\mathbb{C}$, and let $A \in M_n(\mathbb{C})$ be diagonalizable with spectral decomposition $A = \lambda_1 E_1 + \cdots + \lambda_r E_r$. Then, if the series

$$f(A) = a_0 I + a_1 A + a_2 A^2 + \cdots$$

converges, its sum is

$$f(A) = f(\lambda_1)E_1 + \cdots + f(\lambda_r)E_r \ .$$

Proof As in the proof of Theorem 10.20, let the diagonal form of A be

$$D = P^{-1}AP = \lambda_1 P_1 + \cdots + \lambda_r P_r$$

so that $E_i = PP_iP^{-1}$. Now note that

$$\begin{aligned} P^{-1}f(A)P &= a_0P^{-1}P + a_1P^{-1}AP + a_2P^{-1}APP^{-1}AP + \cdots \\ &= f(P^{-1}AP) \\ &= a_0 I + a_1 D + a_2 D^2 + \cdots \\ &= f(D) \ . \end{aligned}$$

Using properties (a) – (c) of Theorem 10.20 applied to the P_i, it is easy to see that $D^k = \lambda_1{}^k P_1 + \cdots + \lambda_r{}^k P_r$ and hence

$$f(D) = f(\lambda_1)P_1 + \cdots + f(\lambda_r)P_r \ .$$

Then if $f(A) = \sum A_r$ converges, so does $\sum P^{-1}A_rP = P^{-1}f(A)P = f(D)$ (Theorem 10.24(c)), and we have

$$f(A) = f(PDP^{-1}) = Pf(D)P^{-1} = f(\lambda_1)E_1 + \cdots + f(\lambda_r)E_r \ . \ ∎$$

Example 10.9 Consider the exponential series e^A where A is diagonalizable. Then, if $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of A, we have the spectral decomposition $A = \lambda_1 E_1 + \cdots + \lambda_k E_k$. Using $f(A) = e^A$, Theorem 10.27 yields

$$e^A = e^{\lambda_1}E_1 + \cdots + e^{\lambda_k}E_k$$

in agreement with Example 10.7. ∕∕

We can now prove our earlier assertion that a unitary matrix U can be written in the form e^{iH} for some Hermitian matrix H.

Theorem 10.28 Every unitary matrix U can be written in the form e^{iH} for some Hermitian matrix H. Conversely, if H is Hermitian, then e^{iH} is unitary.

Proof By Theorem 10.9(b), the distinct eigenvalues of U may be written in the form $e^{i\lambda_1}, \ldots, e^{i\lambda_k}$ where each λ_i is real. Since U is also normal, it follows from Corollary 1 of Theorem 10.15 that there exists a unitary matrix P such that $P^\dagger UP = P^{-1}UP$ is diagonal. In fact

$$P^{-1}UP = e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k$$

where the P_i are the idempotent matrices used in the proof of Theorem 10.20. From Example 10.7 we see that the matrix $e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k$ is just e^{iD} where

$$D = \lambda_1P_1 + \cdots + \lambda_kP_k$$

is a diagonal matrix with the λ_i as diagonal entries. Therefore, using Theorem 10.24(c) we see that

$$U = Pe^{iD}P^{-1} = e^{iPDP^{-1}} = e^{iH}$$

where $H = PDP^{-1}$. Since D is a real diagonal matrix it is clearly Hermitian, and since P is unitary (so that $P^{-1} = P^\dagger$), it follows that $H^\dagger = (PDP^\dagger)^\dagger = PDP^\dagger = H$ so that H is Hermitian also.

Conversely, suppose H is Hermitian with distinct real eigenvalues $\lambda_1, \dots, \lambda_k$. Since H is also normal, there exists a unitary matrix P that diagonalizes H. Then as above, we may write this diagonal matrix as

$$P^{-1}HP = \lambda_1P_1 + \cdots + \lambda_kP_k$$

so that (from Example 10.7 again)

$$P^{-1}e^{iH}P = e^{iP^{-1}HP} = e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k \ .$$

Using the properties of the P_i, it is easy to see that the right hand side of this equation is diagonal and unitary since using

$$(e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k)^\dagger = e^{-i\lambda_1}P_1 + \cdots + e^{-i\lambda_k}P_k$$

we have

$$(e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k)^\dagger(e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k) = I$$

and

$$(e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k)(e^{i\lambda_1}P_1 + \cdots + e^{i\lambda_k}P_k)^\dagger = I \ .$$

Therefore the left hand side must also be unitary, and hence (using $P^{-1} = P^\dagger$)

$$\begin{aligned} I &= (P^{-1}e^{iH}P)^\dagger(P^{-1}e^{iH}P) \\ &= P^\dagger(e^{iH})^\dagger PP^{-1}e^{iH}P \\ &= P^\dagger(e^{iH})^\dagger e^{iH}P \end{aligned}$$

so that $PP^{-1} = I = (e^{iH})^\dagger e^{iH}$. Similarly we see that $e^{iH}(e^{iH})^\dagger = I$, and thus e^{iH} is unitary. ∎

While this theorem is also true in infinite dimensions (i.e., in a Hilbert space), its proof is considerably more difficult. The reader is referred to the books listed in the bibliography for this generalization.

Given a constant matrix A, we now wish to show that

$$\frac{de^{tA}}{dt} = Ae^{tA} \quad . \tag{1}$$

To see this, we first define the derivative of a matrix $M = M(t)$ to be that matrix whose elements are just the derivatives of the corresponding elements of M. In other words, if $M(t) = (m_{ij}(t))$, then $(dM/dt)_{ij} = dm_{ij}/dt$. Now note that (with $M(t) = tA$)

$$e^{tA} = I + tA + (tA)^2/2! + (tA)^3/3! + \cdots$$

and hence (since the a_{ij} are constant) taking the derivative with respect to t yields the desired result:

$$\begin{aligned} de^{tA}/dt &= 0 + A + tA^2 + (tA)^2A/2! + \cdots \\ &= A\{I + tA + (tA)^2/2! + \cdots\} \\ &= Ae^{tA} \quad . \end{aligned}$$

Next, given two matrices A and B (of compatible sizes), we recall that their commutator is the matrix $[A, B] = AB - BA = -[B, A]$. If $[A, B] = 0$, then $AB = BA$ and we say that A and B **commute**. Now consider the function $f(x) = e^{xA}Be^{-xA}$. Leaving it to the reader to verify that the product rule for derivatives also holds for matrices, we obtain (note that $Ae^{xA} = e^{xA}A$)

$$\begin{aligned} df/dx &= Ae^{xA}Be^{-xA} - e^{xA}Be^{-xA}A = Af - fA = [A, f] \\ d^2f/dx^2 &= [A, df/dx] = [A, [A, f]] \\ &\vdots \end{aligned}$$

Expanding f(x) in a Taylor series about x = 0, we find (using f(0) = B)

$$\begin{aligned} f(x) &= f(0) + (df/dx)_0 x + (d^2 f/dx^2)_0 x^2/2! + \cdots \\ &= B + [A,\ B]x + [A,\ [A,\ B]]x^2/2! + \cdots\ . \end{aligned}$$

Setting x = 1, we finally obtain

$$e^A B e^{-A} = B + [A,\ B] + [A,\ [A,\ B]]/2! + [A,\ [A,\ [A,\ B]]]/3! + \cdots \tag{2}$$

Note that setting B = I shows that $e^A e^{-A} = I$ as we would hope.

In the particular case that both A and B commute with their commutator [A, B], then we find from (2) that $e^A B e^{-A} = B + [A, B]$ and hence $e^A B = Be^A + [A, B]e^A$ or

$$[e^A,\ B] = [A,\ B]e^A\ . \tag{3}$$

Example 10.10 We now show that if A and B are two matrices *that both commute with their commutator* [A, B], then

$$e^A e^B = \exp\{A + B + [A,\ B]/2\}\ . \tag{4}$$

(This is sometimes referred to as **Weyl's formula**.) To prove this, we start with the function $f(x) = e^{xA}e^{xB}e^{-x(A+B)}$. Then

$$\begin{aligned} df/dx &= e^{xA}Ae^{xB}e^{-x(A+B)} + e^{xA}e^{xB}Be^{-x(A+B)} - e^{xA}e^{xB}(A+B)e^{-x(A+B)} \\ &= e^{xA}Ae^{xB}e^{-x(A+B)} - e^{xA}e^{xB}Ae^{-x(A+B)} \\ &= e^{xA}[A,\ e^{xB}]e^{-x(A+B)} \end{aligned} \tag{5}$$

As a special case, note [A, B] = 0 implies df/dx = 0 so that f is independent of x. Since f(0) = I, it follows that we may choose x = 1 to obtain $e^A e^B e^{-(A+B)} = I$ or $e^A e^B = e^{A+B}$ (as long as [A, B] = 0).

From (3) we have (replacing A by xB and B by A) $[A, e^{xB}] = x[A, B]e^{xB}$. Using this along with the fact that A commutes with the commutator [A, B] (so that $e^{xA}[A, B] = [A, B]e^{xA}$), we have

$$df/dx\ =\ xe^{xA}[A,\ B]e^{xB}e^{-x(A+B)}\ =\ x[A,\ B]f\ .$$

Since A and B are independent of x, we may formally integrate this from 0 to x to obtain

$$\ln f(x)/f(0)\ =\ [A,\ B]x^2/2\ .$$

Using f(0) = 1, this is $f(x) = \exp\{[A, B]x^2/2\}$ so that setting x = 1 we find

$$e^A e^B e^{-(A+B)} = \exp\{[A, B]/2\} \ .$$

Finally, multiplying this equation from the right by e^{A+B} and using the fact that $[[A, B]/2, A + B] = 0$ yields (4). ∕∕

Exercises

1. (a) Let N be a normal operator on a finite-dimensional unitary space. Prove that

$$\det e^N = e^{\mathrm{Tr}\, N} \ .$$

(b) Prove this holds for any $N \in M_n(\mathbb{C})$. [*Hint*: Use either Theorem 8.1 or the fact (essentially proved at the end of Section 8.6) that the diagonalizable matrices are dense in $M_n(\mathbb{C})$.]

2. If the limit of a sequence of unitary operators exists, is it also unitary? Why?

3. Let T be a unitary operator. Show that the sequence $\{T^n: n = 0, 1, 2, \ldots\}$ contains a subsequence $\{T^{n_k}: k = 0, 1, 2, \ldots\}$ that converges to a unitary operator. [*Hint*: You will need the fact that the unit disk in $\mathbb{C}^2$ is compact (see Appendix A).]

10.7 POSITIVE OPERATORS

Before proving the main result of this section (the polar decomposition theorem), let us briefly discuss functions of a linear transformation. We have already seen two examples of such a function. First, the exponential series $e^{\,A}$ (which may be defined for operators exactly as for matrices) and second, if A is a normal operator with spectral decomposition $A = \sum\lambda_i E_i$, then we saw that the linear transformation p(A) was given by $p(A) = \sum p(\lambda_i)E_i$ where p(x) is any polynomial in $\mathbb{C}[x]$ (Corollary to Theorem 10.20).

In order to generalize this notion, let N be a normal operator on a unitary space, and hence N has spectral decomposition $\sum\lambda_i E_i$. If f is an arbitrary complex-valued function (defined at least at each of the λ_i), we define a linear transformation f(N) by

$$f(N) = \sum f(\lambda_i)E_i \ .$$

What we are particularly interested in is the function $f(x) = \sqrt{x}$ defined for all real $x \geq 0$ as the positive square root of x.

Recall (see Section 10.3) that we defined a positive operator P by the requirement that $P = S^\dagger S$ for some operator S. It is then clear that $P^\dagger = P$, and hence P is normal. From Theorem 10.9(d), the eigenvalues of $P = \sum \lambda_j E_j$ are real and non-negative, and we can define $\sqrt{P}$ by

$$\sqrt{P} = \sum_j \sqrt{\lambda_j}\, E_j$$

where each $\lambda_j \geq 0$.

Using the properties of the E_j, it is easy to see that $(\sqrt{P})^2 = P$. Furthermore, since E_j is an orthogonal projection, it follows that $E_j{}^\dagger = E_j$ (Theorem 10.21), and therefore $(\sqrt{P})^\dagger = \sqrt{P}$ so that $\sqrt{P}$ is Hermitian. Note that since $P = S^\dagger S$ we have

$$\langle Pv, v\rangle = \langle (S^\dagger S)v, v\rangle = \langle Sv, Sv\rangle = \|Sv\|^2 \geq 0 .$$

Just as we did in the proof of Theorem 10.23, let us write $v = \sum E_j v = \sum v_j$ where the nonzero v_j are mutually orthogonal. Then

$$\sqrt{P}(v) = \sum \sqrt{\lambda_j}\, E_j v = \sum \sqrt{\lambda_j}\, v_j$$

and hence we also have (using $\langle v_j, v_k\rangle = 0$ if $j \neq k$)

$$\begin{aligned}\langle \sqrt{P}(v), v\rangle &= \langle \textstyle\sum_j \sqrt{\lambda_j} v_j, \sum_k v_k\rangle = \sum_{j,k} \sqrt{\lambda_j}\langle v_j, v_k\rangle = \sum_j \sqrt{\lambda_j}\langle v_j, v_j\rangle \\ &= \textstyle\sum_j \sqrt{\lambda_j}\, \|v_j\|^2 \geq 0 \ .\end{aligned}$$

In summary, we have shown that $\sqrt{P}$ satisfies

(a) $(\sqrt{P})^2 = P$
(b) $(\sqrt{P})^\dagger = \sqrt{P}$
(c) $\langle \sqrt{P}(v), v\rangle \geq 0$

and it is natural to ask about the uniqueness of any operator satisfying these three properties. For example, if we let $T = \sum \pm\sqrt{\lambda_j}\, E_j$, then we still have $T^2 = \sum \lambda_j E_j = P$ regardless of the sign chosen for each term. Let us denote the fact that $\sqrt{P}$ satisfies properties (b) and (c) above by the expression $\sqrt{P} \geq 0$. In other words, by the statement $A \geq 0$ we mean that $A^\dagger = A$ and $\langle Av, v\rangle \geq 0$ for every $v \in V$ (i.e., A is a positive Hermitian operator).

We now claim that if $P = T^2$ and $T \geq 0$, then $T = \sqrt{P}$. To prove this, we first note that $T \geq 0$ implies $T^\dagger = T$ (property (b)), and hence T must also be normal. Now let $\sum \mu_i F_i$ be the spectral decomposition of T. Then

$$\Sigma(\mu_i)^2 F_i = T^2 = P = \Sigma\lambda_j E_j \ .$$

If $v_i \neq 0$ is an eigenvector of T corresponding to μ_i, then property (c) tells us that (using the fact that each μ_i is real since T is Hermitian)

$$0 \le \langle Tv_i, v_i\rangle = \langle \mu_i v_i, v_i\rangle = \mu_i \|v_i\|^2 \ .$$

But $\|v_i\| > 0$, and hence $\mu_i \ge 0$. In other words, any operator $T \ge 0$ has nonnegative eigenvalues. Since each μ_i is distinct and nonnegative, so is each $\mu_i{}^2$, and hence each $\mu_i{}^2$ must be equal to some λ_j . Therefore the corresponding F_i and E_j must be equal (by Theorem 10.22(e)). By suitably numbering the eigenvalues, we may write $\mu_i{}^2 = \lambda_i$, and thus $\mu_i = \sqrt{\lambda_i}$. This shows that

$$T = \Sigma\mu_i F_i = \Sigma\sqrt{\lambda_i}\, E_i = \sqrt{P}$$

as claimed.

We summarize this discussion in the next result which gives us three equivalent definitions of a **positive transformation**.

Theorem 10.29 Let P be an operator on a unitary space V. Then the following conditions are equivalent:

(a) $P = T^2$ for some unique Hermitian operator $T \ge 0$.
(b) $P = S^\dagger S$ for some operator S.
(c) $P^\dagger = P$ and $\langle Pv, v\rangle \ge 0$ for every $v \in V$.

Proof (a) ⇒ (b): If $P = T^2$ and $T^\dagger = T$, then $P = TT = T^\dagger T$.

(b) ⇒ (c): If $P = S^\dagger S$, then $P^\dagger = P$ and $\langle Pv, v\rangle = \langle S^\dagger Sv, v\rangle = \langle Sv, Sv\rangle = \|Sv\|^2 \ge 0$.

(c) ⇒ (a): Note that property (c) is just our statement that $P \ge 0$. Since $P^\dagger = P$, we see that P is normal, and hence we may write $P = \Sigma\lambda_j E_j$. Defining $T = \Sigma\sqrt{\lambda_j}\, E_j$, we have $T^\dagger = T$ (since every E_j is Hermitian), and the preceding discussion shows that $T \ge 0$ is the unique operator with the property that $P = T^2$. ∎

We remark that in the particular case that P is positive definite, then $P = S^\dagger S$ where S is nonsingular. This means that P is also nonsingular.

Finally, we are in a position to prove the last result of this section, the so-called polar decomposition (or factorization) of an operator. While we state and prove this theorem in terms of matrices, it should be obvious by now that it applies just as well to operators.

Theorem 10.30 (Polar Decomposition) If $A \in M_n(\mathbb{C})$, then there exist unique positive Hermitian matrices $H_1, H_2 \in M_n(\mathbb{C})$ and (not necessarily unique) unitary matrices $U_1, U_2 \in M_n(\mathbb{C})$ such that $A = U_1H_1 = H_2U_2$. Moreover, $H_1 = (A^\dagger A)^{1/2}$ and $H_2 = (AA^\dagger)^{1/2}$. In addition, the matrices U_1 and U_2 are uniquely determined if and only if A is nonsingular.

Proof Let $\lambda_1{}^2, \dots, \lambda_n{}^2$ be the eigenvalues of the positive Hermitian matrix $A^\dagger A$, and assume the λ_i are numbered so that $\lambda_i > 0$ for $i = 1, \dots, k$ and $\lambda_i = 0$ for $i = k + 1, \dots, n$ (see Theorem 10.9(d)). (Note that if A is nonsingular, then $A^\dagger A$ is positive definite and hence $k = n$.) Applying Corollary 1 of Theorem 10.15, we let $\{v_1, \dots, v_n\}$ be the corresponding orthonormal eigenvectors of $A^\dagger A$. For each $i = 1, \dots, k$ we define the vectors $w_i = Av_i/\lambda_i$. Then

$$\begin{aligned}\langle w_i, w_j\rangle &= \langle Av_i/\lambda_i, Av_j/\lambda_j\rangle = \langle v_i, A^\dagger Av_j\rangle/\lambda_i\lambda_j\\ &= \langle v_i, v_j\rangle\lambda_j{}^2/\lambda_i\lambda_j = \delta_{ij}\lambda_j{}^2/\lambda_i\lambda_j\end{aligned}$$

so that $w_1, \dots, w_k$ are also orthonormal. We now extend these to an orthonormal basis $\{w_1, \dots, w_n\}$ for $\mathbb{C}^n$. If we define the columns of the matrices $V, W \in M_n(\mathbb{C})$ by $V^i = v_i$ and $W^i = w_i$, then V and W will be unitary by Theorem 10.7.

Defining the Hermitian matrix $D \in M_n(\mathbb{C})$ by

$$D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

it is easy to see that the equations $Av_i = \lambda_iw_i$ may be written in matrix form as $AV = WD$. Using the fact that V and W are unitary, we define $U_1 = WV^\dagger$ and $H_1 = VDV^\dagger$ to obtain

$$A = WDV^\dagger = (WV^\dagger)(VDV^\dagger) = U_1H_1 \;.$$

Since $\det(\lambda I - VDV^\dagger) = \det(\lambda I - D)$, we see that H_1 and D have the same nonnegative eigenvalues, and hence H_1 is a positive Hermitian matrix. We can now apply this result to the matrix $A^\dagger$ to write $A^\dagger = \tilde{U}_1\tilde{H}_1$ or $A = \tilde{H}_1{}^\dagger\tilde{U}_1{}^\dagger = \tilde{H}_1\tilde{U}_1{}^\dagger$. If we define $H_2 = \tilde{H}_1$ and $U_2 = \tilde{U}_1{}^\dagger$, then we obtain $A = H_2U_2$ as desired.

We now observe that using $A = U_1H_1$ we may write

$$A^\dagger A = H_1U_1{}^\dagger U_1H_1 = (H_1)^2$$

and similarly

$$AA^\dagger = H_2U_2U_2{}^\dagger H_2 = (H_2)^2$$

so that H_1 and H_2 are unique even if A is singular. Since U_1 and U_2 are unitary, they are necessarily nonsingular, and hence H_1 and H_2 are nonsingular if $A = U_1H_1 = H_2U_2$ is nonsingular. In this case, $U_1 = AH_1^{-1}$ and $U_2 = H_2^{-1}A$ will also be unique. On the other hand, suppose A is singular. Then $k \neq n$ and $w_k, \ldots, w_n$ are not unique. This means that $U_1 = WV^\dagger$ (and similarly U_2) is not unique. In other words, if U_1 and U_2 are unique, then A must be nonsingular. ∎

Exercises

1. Let V be a unitary space and let $E \in L(V)$ be an orthogonal projection.
   (a) Show directly that E is a positive transformation.
   (b) Show that $\|Ev\| \leq \|v\|$ for all $v \in V$.

2. Prove that if A and B are commuting positive transformations, then AB is also positive.

3. This exercise is related to Exercise 7.5.5. Prove that any representation of a finite group is equivalent to a unitary representation as follows:
   (a) Consider the matrix $X = \sum_{a \in G} D^\dagger(a)D(a)$. Show that X is Hermitian and positive definite, and hence that $X = S^2$ for some Hermitian S.
   (b) Show that $D(a)^\dagger XD(a) = X$.
   (c) Show that $U(a) = SD(a)S^{-1}$ is a unitary representation.

Supplementary Exercises for Chapter 10

1. Let T be a linear transformation on a space V with basis $\{e_1, \ldots, e_n\}$. If $T(e_i) = \sum_{j \geq i} a_{ji}e_j$ for all $i = 1, \ldots, n$ and $T(e_1) \neq ce_1$ for any scalar c, show that T is not normal.

2. Let A be a fixed n x n matrix, and let B be any n x n matrix such that $A = B^2$. Assume that B is similar to a diagonal matrix and has nonnegative eigenvalues $\lambda_1, \ldots, \lambda_n$. Let p(x) be a polynomial such that $p(\lambda_i{}^2) = \lambda_i$ for each $i = 1, \ldots, n$. Show that $p(A) = B$ and hence B is unique. How does this relate to our discussion of $\sqrt{P}$ for a positive operator P?

3. Describe all operators that are both unitary and positive.

4. Is it true that for any $A \in M_n(\mathbb{C})$, $AA^\dagger$ and $A^\dagger A$ are unitarily similar? Explain.

5. In each case, indicate whether or not the statement is true or false and give your reason.

   (a) For any $A \in M_n(\mathbb{C})$, $AA^\dagger$ has all real eigenvalues.

   (b) For any $A \in M_n(\mathbb{C})$, the eigenvalues of $AA^\dagger$ are of the form $|\lambda|^2$ where λ is an eigenvalue of A.

   (c) For any $A \in M_n(\mathbb{C})$, the eigenvalues of $AA^\dagger$ are nonnegative real numbers.

   (d) For any $A \in M_n(\mathbb{C})$, $AA^\dagger$ has the same eigenvalues as $A^\dagger A$ if A is nonsingular.

   (e) For any $A \in M_n(\mathbb{C})$, $\mathrm{Tr}(AA^\dagger) = |\mathrm{Tr}\, A|^2$.

   (f) For any $A \in M_n(\mathbb{C})$, $AA^\dagger$ is unitarily similar to a diagonal matrix.

   (g) For any $A \in M_n(\mathbb{C})$, $AA^\dagger$ has n linearly independent eigenvectors.

   (h) For any $A \in M_n(\mathbb{C})$, the eigenvalues of $AA^\dagger$ are the same as the eigenvalues of $A^\dagger A$.

   (i) For any $A \in M_n(\mathbb{C})$, the Jordan form of $AA^\dagger$ is the same as the Jordan form of $A^\dagger A$.

   (j) For any $A \in M_n(\mathbb{C})$, the null space of $A^\dagger A$ is the same as the null space of A.

6. Let S and T be normal operators on V. Show that there are bases $\{u_i\}$ and $\{v_i\}$ for V such that $[S]_u = [T]_v$ if and only if there are orthonormal bases $\{u'_i\}$ and $\{v'_i\}$ such that $[S]_{u'} = [T]_{v'}$.

7. Let T be normal and let $k > 0$ be an integer. Show that there is a normal S such that $S^k = T$.

8. Let N be normal and let p(x) be a polynomial over $\mathbb{C}$. Show that p(N) is also normal.

9. Let N be a normal operator on a unitary space V, let W = Ker N, and let $\tilde{N}$ be the transformation induced by N on V/W. Show that $\tilde{N}$ is normal. Show that $\tilde{N}^{-1}$ is also normal.

10. Discuss the following assertion: For any linear transformation T on a unitary space V, $TT^\dagger$ and $T^\dagger T$ have a common basis of eigenvectors.

11. Show that if A and B are real symmetric matrices and A is positive definite, then $p(x) = \det(B - xA)$ has all real roots.

CHAPTER 11

Multilinear Mappings and Tensors

In this chapter we generalize our earlier discussion of bilinear forms, which leads in a natural manner to the concepts of tensors and tensor products. While we are aware that our approach is not the most general possible (which is to say it is not the most abstract), we feel that it is more intuitive, and hence the best way to approach the subject for the first time. In fact, our treatment is essentially all that is ever needed by physicists, engineers and applied mathematicians. More general treatments are discussed in advanced courses on abstract algebra.

The basic idea is as follows. Given a vector space V with basis $\{e_i\}$, we defined the dual space V* (with basis $\{\omega^i\}$) as the space of linear functionals on V. In other words, if $\phi = \sum_i \phi_i \omega^i \in V^*$ and $v = \sum_j v^j e_j \in V$, then

$$\begin{aligned}\phi(v) = \langle \phi, v \rangle = \langle \Sigma_i \phi_i \omega^i, \Sigma_j v^j e_j \rangle = \Sigma_{i,\,j} \phi_i v^j \delta^i{}_j \\ = \Sigma_i \phi_i v^i \quad .\end{aligned}$$

Next we defined the space $\mathcal{B}(V)$ of all bilinear forms on V (i.e., bilinear mappings on V × V), and we showed (Theorem 9.10) that $\mathcal{B}(V)$ has a basis given by $\{f^{ij} = \omega^i \otimes \omega^j\}$ where

$$f^{ij}(u, v) \;=\; \omega^i \otimes \omega^j(u, v) \;=\; \omega^i(u)\omega^j(v) \;=\; u^i v^j \;\; .$$

It is this definition of the f^{ij} that we will now generalize to include linear functionals on spaces such as, for example, V* × V* × V* × V × V.

11.1 DEFINITIONS

Let V be a finite-dimensional vector space over $\mathcal{F}$, and let V^r denote the r-fold Cartesian product $V \times V \times \cdots \times V$. In other words, an element of V^r is an r-tuple $(v_1, \ldots, v_r)$ where each $v_i \in V$. If W is another vector space over $\mathcal{F}$, then a mapping $T: V^r \rightarrow W$ is said to be **multilinear** if $T(v_1, \ldots, v_r)$ is linear in each variable. That is, T is multilinear if for each i = 1, . . . , r we have

$$\begin{aligned} &T(v_1, \ldots, av_i + bv'_i, \ldots, v_r) \\ &\qquad = aT(v_1, \ldots, v_i, \ldots, v_r) + bT(v_1, \ldots, v'_i, \ldots, v_r) \end{aligned}$$

for all $v_i, v'_i \in V$ and $a, b \in \mathcal{F}$. In the particular case that $W = \mathcal{F}$, the mapping T is variously called an **r-linear form** on V, or a **multilinear form** of **degree** r on V, or an **r-tensor** on V. The set of all r-tensors on V will be denoted by $\mathcal{T}_r(V)$. (It is also possible to discuss multilinear mappings that take their values in W rather than in $\mathcal{F}$. See Section 11.5.)

As might be expected, we define addition and scalar multiplication on $\mathcal{T}_r(V)$ by

$$\begin{aligned} (S+T)(v_1, \ldots, v_r) &= S(v_1, \ldots, v_r) + T(v_1, \ldots, v_r) \\ (aT)(v_1, \ldots, v_r) &= aT(v_1, \ldots, v_r) \end{aligned}$$

for all $S, T \in \mathcal{T}_r(V)$ and $a \in \mathcal{F}$. It should be clear that S + T and aT are both r-tensors. With these operations, $\mathcal{T}_r(V)$ becomes a vector space over $\mathcal{F}$. Note that the particular case of r = 1 yields $\mathcal{T}_1(V) = V^*$, i.e., the dual space of V, and if r = 2, then we obtain a bilinear form on V.

Although this definition takes care of most of what we will need in this chapter, it is worth going through a more general (but not really more difficult) definition as follows. The basic idea is that a tensor is a scalar-valued multilinear function with variables in both V and V*. Note also that by Theorem 9.4, the space of linear functions on V* is V** which we view as simply V. For example, a tensor could be a function on the space V* × V × V. By convention, we will always write all V* variables before all V variables, so that, for example, a tensor on V × V* × V will be replaced by a tensor on V* × V × V. (However, not all authors adhere to this convention, so the reader should be very careful when reading the literature.)

Without further ado, we define a **tensor** T on V to be a multilinear map on $V^{*s} \times V^r$:

$$T\colon V^{*s} \times V^r = \underbrace{V^* \times\cdots\times V^*}_{\text{s copies}} \times \underbrace{V \times\cdots\times V}_{\text{r copies}} \to \mathcal{F}$$

where r is called the **covariant order** and s is called the **contravariant order** of T. We shall say that a tensor of covariant order r and contravariant order s is of **type** (or **rank**) $\binom{s}{r}$. If we denote the set of all tensors of type $\binom{s}{r}$ by $\mathcal{T}_r^s(V)$, then defining addition and scalar multiplication exactly as above, we see that $\mathcal{T}_r^s(V)$ forms a vector space over $\mathcal{F}$. A tensor of type $\binom{0}{0}$ is defined to be a scalar, and hence $\mathcal{T}_0^0(V) = \mathcal{F}$. A tensor of type $\binom{1}{0}$ is called a **contravariant vector**, and a tensor of type $\binom{0}{1}$ is called a **covariant vector** (or simply a **covector**). In order to distinguish between these types of vectors, we denote the basis vectors for V by a subscript (e.g., e_i), and the basis vectors for V* by a superscript (e.g., ω^j). Furthermore, we will generally leave off the V and simply write $\mathcal{T}_r$ or $\mathcal{T}_r^s$.

At this point we are virtually forced to introduce the so-called Einstein **summation convention**. This convention says that we are to sum over repeated indices in any vector or tensor expression where one index is a superscript and one is a subscript. Because of this, we write the vector components with indices in the opposite position from that of the basis vectors. This is why we have been writing $v = \sum_i v^i e_i \in V$ and $\phi = \sum_j \phi_j \omega^j \in V^*$. Thus we now simply write $v = v^i e_i$ and $\phi = \phi_j \omega^j$ where the summation is to be understood. Generally the limits of the sum will be clear. However, we will revert to the more complete notation if there is any possibility of ambiguity.

It is also worth emphasizing the trivial fact that the indices summed over are just "dummy indices." In other words, we have $v^i e_i = v^j e_j$ and so on. Throughout this chapter we will be relabelling indices in this manner without further notice, and we will assume that the reader understands what we are doing.

Suppose $T \in \mathcal{T}_r$, and let $\{e_1, \dots, e_n\}$ be a basis for V. For each i = 1, . . . , r we define a vector $v_i = e_j a^j{}_i$ where, as usual, $a^j{}_i \in \mathcal{F}$ is just the *jth* component of the vector v_i. (Note that here the subscript i is not a tensor index.) Using the multilinearity of T we see that

$$T(v_1, \dots, v_r) = T(e_{j_1} a^{j_1}{}_1, \dots, e_{j_r} a^{j_r}{}_r) = a^{j_1}{}_1 \cdots a^{j_r}{}_r\, T(e_{j_1}, \dots, e_{j_r}) \; .$$

The n^r scalars $T(e_{j_1}, \dots, e_{j_r})$ are called the **components** of T relative to the basis $\{e_i\}$, and are denoted by $T_{j_1 \cdots j_r}$. This terminology implies that there exists a basis for $\mathcal{T}_r$ such that the $T_{j_1 \cdots j_r}$ are just the components of T with respect to this basis. We now construct this basis, which will prove that $\mathcal{T}_r$ is of dimension n^r.

(We will show formally in Section 11.10 that the Kronecker symbols $\delta^i{}_j$ are in fact the components of a tensor, and that these components are the same in any coordinate system. However, for all practical purposes we continue to use the $\delta^i{}_j$ simply as a notational device, and hence we place no importance on the position of the indices, i.e., $\delta^i{}_j = \delta_j{}^i$ etc.)

For each collection $\{i_1, \dots, i_r\}$ (where $1 \le i_k \le n$), we define the tensor $\Omega^{i_1 \cdots i_r}$ (not simply the components of a tensor Ω) to be that element of $\mathcal{T}_r$ whose values on the basis $\{e_i\}$ for V are given by

$$\Omega^{i_1 \cdots i_r}(e_{j_1}, \dots, e_{j_r}) = \delta^{i_1}{}_{j_1} \cdots \delta^{i_r}{}_{j_r}$$

and whose values on an arbitrary collection $\{v_1, \dots, v_r\}$ of vectors are given by multilinearity as

$$\begin{aligned}\Omega^{i_1 \cdots i_r}(v_1, \dots, v_r) &= \Omega^{i_1 \cdots i_r}(e_{j_1} a^{j_1}{}_1, \dots, e_{j_r} a^{j_r}{}_r) \\ &= a^{j_1}{}_1 \cdots a^{j_r}{}_r \, \Omega^{i_1 \cdots i_r}(e_{j_1}, \dots, e_{j_r}) \\ &= a^{j_1}{}_1 \cdots a^{j_r}{}_r \, \delta^{i_1}{}_{j_1} \cdots \delta^{i_r}{}_{j_r} \\ &= a^{i_1}{}_1 \cdots a^{i_r}{}_r \ .\end{aligned}$$

That this does indeed define a tensor is guaranteed by this last equation which shows that each $\Omega^{i_1 \cdots i_r}$ is in fact linear in each variable (since $v_1 + v'_1 = (a^{j_1}{}_1 + a'^{j_1}{}_1)e_{j_1}$ etc.). To prove that the n^r tensors $\Omega^{i_1 \cdots i_r}$ form a basis for $\mathcal{T}_r$, we must show that they linearly independent and span $\mathcal{T}_r$.

Suppose that $\alpha_{i_1 \cdots i_r} \Omega^{i_1 \cdots i_r} = 0$ where each $\alpha_{i_1 \cdots i_r} \in \mathcal{F}$. From the definition of $\Omega^{i_1 \cdots i_r}$, we see that applying this to any r-tuple $(e_{j_1}, \dots, e_{j_r})$ of basis vectors yields $\alpha_{j_1 \cdots j_r} = 0$. Since this is true for every such r-tuple, it follows that $\alpha_{i_1 \cdots i_r} = 0$ for every r-tuple of indices $(i_1, \dots, i_r)$, and hence the $\Omega^{i_1 \cdots i_r}$ are linearly independent.

Now let $T_{i_1 \cdots i_r} = T(e_{i_1}, \dots, e_{i_r})$ and consider the tensor

$$T_{i_1 \cdots i_r} \Omega^{i_1 \cdots i_r}$$

in $\mathcal{T}_r$. Using the definition of $\Omega^{i_1 \cdots i_r}$, we see that both $T_{i_1 \cdots i_r}\Omega^{i_1 \cdots i_r}$ and T yield the same result when applied to any r-tuple $(e_{j_1}, \dots, e_{j_r})$ of basis vectors, and hence they must be equal as multilinear functions on V^r. This shows that $\{\Omega^{i_1 \cdots i_r}\}$ spans $\mathcal{T}_r$.

While we have treated only the space $\mathcal{T}_r$, it is not any more difficult to treat the general space $\mathcal{T}_r^s$. Thus, if $\{e_i\}$ is a basis for V, $\{\omega^j\}$ is a basis for V* and $T \in \mathcal{T}_r^s$, we define the **components** of T (relative to the given bases) by

$$T^{i_1 \cdots i_s}{}_{j_1 \cdots j_r} = T(\omega^{i_1}, \ldots, \omega^{i_s}, e_{j_1}, \ldots, e_{j_r}) \ .$$

Defining the n^{r+s} analogous tensors $\Omega_{i_1}{}^{j_1}{}^{\cdots}{}_{i_s}{}^{j_r}$, it is easy to mimic the above procedure and hence prove the following result.

Theorem 11.1 The set $\mathcal{T}_r^s$ of all tensors of type $\binom{s}{r}$ on V forms a vector space of dimension n^{r+s}.

Proof This is Exercise 11.1.1. ∎

Since a tensor $T \in \mathcal{T}_r^s$ is a function on $V^{*s} \times V^r$, it would be nice if we could write a basis (e.g., $\Omega_{i_1}{}^{j_1}{}^{\cdots}{}_{i_s}{}^{j_r}$) for $\mathcal{T}_r^s$ in terms of the bases $\{e_i\}$ for V and $\{\omega^j\}$ for V*. We now show that this is easy to accomplish by defining a product on $\mathcal{T}_r^s$, called the tensor product. The reader is cautioned not to be intimidated by the notational complexities, since the concepts involved are really quite simple.

Suppose that $S \in \mathcal{T}_{r_1}{}^{s_1}$ and $T \in \mathcal{T}_{r_2}{}^{s_2}$. Let $u_1, \ldots, u_{r_1}$, $v_1, \ldots, v_{r_2}$ be vectors in V, and $\alpha^1, \ldots, \alpha^{s_1}$, $\beta^1, \ldots, \beta^{s_2}$ be covectors in V*. Note that the product

$$S(\alpha^1, \ldots, \alpha^{s_1}, u_1, \ldots, u_{r_1})\, T(\beta^1, \ldots, \beta^{s_2}, v_1, \ldots, v_{r_2})$$

is linear in each of its $r_1 + s_1 + r_2 + s_2$ variables. Hence we define the **tensor product** $S \otimes T \in \mathcal{T}_{r_1+r_2}^{s_1+s_2}$ (read "S tensor T") by

$$\begin{aligned}(S \otimes T)(\alpha^1, \ldots, \alpha^{s_1}, \beta^1, \ldots, \beta^{s_2}, u_1, \ldots, u_{r_1}, v_1, \ldots, v_{r_2}) = \\ S(\alpha^1, \ldots, \alpha^{s_1}, u_1, \ldots, u_{r_1})\, T(\beta^1, \ldots, \beta^{s_2}, v_1, \ldots, v_{r_2}) \ .\end{aligned}$$

It is easily shown that the tensor product is both associative and distributive (i.e., bilinear in both factors). In other words, for any scalar $a \in \mathcal{F}$ and tensors R, S and T such that the following formulas make sense, we have

$$\begin{aligned}(R \otimes S) \otimes T &= R \otimes (S \otimes T)\\ R \otimes (S + T) &= R \otimes S + R \otimes T\\ (R + S) \otimes T &= R \otimes T + S \otimes T\\ (aS) \otimes T &= S \otimes (aT) = a(S \otimes T)\end{aligned}$$

(see Exercise 11.1.2). Because of the associativity property (which is a consequence of associativity in $\mathcal{F}$), we will drop the parentheses in expressions such as the top equation and simply write R ⊗ S ⊗ T. This clearly extends to any finite product of tensors. It is important to note, however, that the tensor product is most certainly *not* commutative.

Now let $\{e_1, \dots, e_n\}$ be a basis for V, and let $\{\omega^j\}$ be its dual basis. We claim that the set $\{\omega^{j_1} \otimes \cdots \otimes \omega^{j_r}\}$ of tensor products where $1 \le j_k \le n$ forms a basis for the space $\mathcal{T}_r$ of covariant tensors. To see this, we note that from the definitions of tensor product and dual space, we have

$$\omega^{j_1} \otimes \cdots \otimes \omega^{j_r}(e_{i_1}, \dots, e_{i_r}) = \omega^{j_1}(e_{i_1}) \cdots \omega^{j_r}(e_{i_r}) = \delta^{j_1}{}_{i_1} \cdots \delta^{j_r}{}_{i_r}$$

so that $\omega^{j_1} \otimes \cdots \otimes \omega^{j_r}$ and $\Omega^{j_1 \cdots j_r}$ take the same values on the r-tuples $(e_{i_1}, \dots, e_{i_r})$, and hence they must be equal as multilinear functions on V^r. Since we showed above that $\{\Omega^{j_1 \cdots j_r}\}$ forms a basis for $\mathcal{T}_r$, we have proved that $\{\omega^{j_1} \otimes \cdots \otimes \omega^{j_r}\}$ also forms a basis for $\mathcal{T}_r$.

The method of the previous paragraph is readily extended to the space $\mathcal{T}_r^s$. We must recall however, that we are treating V** and V as the same space. If $\{e_i\}$ is a basis for V, then the dual basis $\{\omega^j\}$ for V* was defined by $\omega^j(e_i) = \langle \omega^j, e_i \rangle = \delta^j{}_i$. Similarly, given a basis $\{\omega^j\}$ for V*, we define the basis $\{e_i\}$ for $V^{**} = V$ by $e_i(\omega^j) = \omega^j(e_i) = \delta^j{}_i$. In fact, using tensor products, it is now easy to repeat Theorem 11.1 in its most useful form. Note also that the next theorem shows that a tensor is determined by its values on the bases $\{e_i\}$ and $\{\omega^j\}$.

Theorem 11.2 Let V have basis $\{e_1, \dots, e_n\}$, and let V* have the corresponding dual basis $\{\omega^1, \dots, \omega^n\}$. Then a basis for $\mathcal{T}_r^s$ is given by the collection

$$\{e_{i_1} \otimes \cdots \otimes e_{i_s} \otimes \omega^{j_1} \otimes \cdots \otimes \omega^{j_r}\}$$

where $1 \le j_1, \dots, j_r, i_1, \dots, i_s \le n$, and hence $\dim \mathcal{T}_r^s = n^{r+s}$.

Proof In view of Theorem 11.1, all that is needed is to show that

$$e_{i_1} \otimes \cdots \otimes e_{i_s} \otimes \omega^{j_1} \otimes \cdots \otimes \omega^{j_r} = \Omega_{i_1}{}^{j_1} \dot{\cdot} \dot{\cdot} \dot{\cdot}_{i_s}{}^{j_r} .$$

The details are left to the reader (see Exercise 11.1.1). ∎

Since the components of a tensor T are defined with respect to a particular basis (and dual basis), we might ask about the relationship between the components of T relative to two different bases. Using the multilinearity of tensors, this is a simple problem to solve.

First, let $\{e_i\}$ be a basis for V and let $\{\omega^j\}$ be its dual basis. If $\{\bar{e}_i\}$ is another basis for V, then there exists a nonsingular transition matrix $A = (a^j{}_i)$ such that

$$\bar{e}_i = e_j a^j{}_i \; . \qquad (1)$$

(We emphasize that $a^j{}_i$ is only a matrix, not a tensor. Note also that our definition of the matrix of a linear transformation given in Section 5.3 shows that $a^j{}_i$ is the element of A in the *jth* row and *ith* column.) Using $\langle \omega^i, e_j \rangle = \delta^i{}_j$, we have

$$\langle \omega^i, \bar{e}_k \rangle = \langle \omega^i, e_j a^j{}_k \rangle = a^j{}_k \langle \omega^i, e_j \rangle = a^j{}_k \delta^i{}_j = a^i{}_k \; .$$

Let us denote the inverse of the matrix $A = (a^i{}_j)$ by $A^{-1} = B = (b^i{}_j)$. In other words, $a^i{}_j b^j{}_k = \delta^i{}_k$ and $b^i{}_j a^j{}_k = \delta^i{}_k$. Multiplying $\langle \omega^i, \bar{e}_k \rangle = a^i{}_k$ by $b^j{}_i$ and summing on i yields

$$\langle b^j{}_i \omega^i, \bar{e}_k \rangle = b^j{}_i a^i{}_k = \delta^j{}_k \; .$$

But the basis $\{\varpi^i\}$ dual to $\{\bar{e}_i\}$ also must satisfy $\langle \varpi^j, \bar{e}_k \rangle = \delta^j{}_k$, and hence comparing this with the previous equation shows that the dual basis vectors transform as

$$\varpi^j = b^j{}_i \omega^i \qquad (2)$$

The reader should compare this carefully with (1). We say that the dual basis vectors transform **oppositely** (i.e., use the inverse transformation matrix) to the basis vectors. It is also worth emphasizing that if the nonsingular transition matrix from the basis $\{e_i\}$ to the basis $\{\bar{e}_i\}$ is given by A, then (according to the same convention given in Section 5.4) the corresponding nonsingular transition matrix from the basis $\{\omega^i\}$ to the basis $\{\varpi^i\}$ is given by $B^T = (A^{-1})^T$. We leave it to the reader to write out equations (1) and (2) in matrix notation to show that this is true (see Exercise 11.1.3).

We now return to the question of the relationship between the components of a tensor in two different bases. For definiteness, we will consider a tensor $T \in \mathcal{T}_1{}^2$. The analogous result for an arbitrary tensor in $\mathcal{T}_r{}^s$ will be quite obvious. Let $\{e_i\}$ and $\{\omega^j\}$ be a basis and dual basis for V and V* respectively. Now consider another pair of bases $\{\bar{e}_i\}$ and $\{\varpi^j\}$ where $\bar{e}_i = e_j a^j{}_i$ and $\varpi^i = b^i{}_j \omega^j$. Then we have $T^{ij}{}_k = T(\omega^i, \omega^j, e_k)$ as well as $\bar{T}^{pq}{}_r = T(\varpi^p, \varpi^q, \bar{e}_r)$, and therefore

$$\bar{T}^{p\,q}{}_r = T(\bar{\omega}^p, \bar{\omega}^q, \bar{e}_r) = b^p{}_i b^q{}_j a^k{}_r T(\omega^i, \omega^j, e_k) = b^p{}_i b^q{}_j a^k{}_r T^{ij}{}_k \ .$$

This is the **classical law of transformation** of the components of a tensor of type $\binom{2}{1}$. It should be kept in mind that $(a^i{}_j)$ and $(b^i{}_j)$ are inverse matrices to each other. (In fact, this equation is frequently taken as the *definition* of a tensor (at least in older texts). In other words, according to this approach, any quantity with this transformation property is defined to be a tensor.)

In particular, the components v^i of a vector $v = v^i e_i$ transform as

$$\bar{v}^i = b^i{}_j v^j$$

while the components α_i of a covector $\alpha = \alpha_i \omega^i$ transform as

$$\bar{\alpha}_i = \alpha_j a^j{}_i \ .$$

We leave it to the reader to verify that these transformation laws lead to the self-consistent formulas $v = v^i e_i = \bar{v}^j \bar{e}_j$ and $\alpha = \alpha_i \omega^i = \bar{\alpha}_j \bar{\omega}^j$ as we should expect (see Exercise 11.1.4).

We point out that these transformation laws are the origin of the terms "contravariant" and "covariant." This is because the components of a vector transform oppositely ("contravariant") to the basis vectors e_i, while the components of dual vectors transform the same as ("covariant") these basis vectors.

It is also worth mentioning that many authors use a prime (or some other method such as a different type of letter) for distinguishing different bases. In other words, if we have a basis $\{e_i\}$ and we wish to transform to another basis which we denote by $\{e_{i'}\}$, then this is accomplished by a transformation matrix $(a^i{}_{j'})$ so that $e_{i'} = e_j a^j{}_{i'}$. In this case, we would write $\omega^{i'} = a^{i'}{}_j \omega^j$ where $(a^{i'}{}_j)$ is the inverse of $(a^i{}_{j'})$. In this notation, the transformation law for the tensor T used above would be written as

$$T^{p'q'}{}_{r'} = b^{p'}{}_i b^{q'}{}_j a^k{}_{r'} T^{ij}{}_k \ .$$

Note that specifying the components of a tensor with respect to one coordinate system allows the determination of its components with respect to any other coordinate system. Because of this, we shall frequently refer to a tensor by its "generic" components. In other words, we will refer to e.g., $T^{ij}{}_k$, as a "tensor" and not the more accurate description as the "components of the tensor T."

Example 11.1 For those readers who may have seen a classical treatment of tensors and have had a course in advanced calculus, we will now show how our more modern approach agrees with the classical.

If $\{x^i\}$ is a local coordinate system on a differentiable manifold X, then a (tangent) vector field v(x) on X is defined as the derivative function $v = v^i(\partial/\partial x^i)$, so that $v(f) = v^i(\partial f/\partial x^i)$ for every smooth function $f: X \rightarrow \mathbb{R}$ (and where each v^i is a function of position $x \in X$, i.e., $v^i = v^i(x)$). Since every vector at $x \in X$ can in this manner be written as a linear combination of the $\partial/\partial x^i$, we see that $\{\partial/\partial x^i\}$ forms a basis for the tangent space at x.

We now define the **differential** df of a function by df(v) = v(f) and thus df(v) is just the directional derivative of f in the direction of v. Note that

$$dx^i(v) = v(x^i) = v^j(\partial x^i/\partial x^j) = v^j\delta^i{}_j = v^i$$

and hence $df(v) = v^i(\partial f/\partial x^i) = (\partial f/\partial x^i)dx^i(v)$. Since v was arbitrary, we obtain the familiar elementary formula $df = (\partial f/\partial x^i)dx^i$. Furthermore, we see that

$$dx^i(\partial/\partial x^j) = \partial x^i/\partial x^j = \delta^i{}_j$$

so that $\{dx^i\}$ forms the basis dual to $\{\partial/\partial x^i\}$. In summary then, relative to the local coordinate system $\{x^i\}$, we define a basis $\{e_i = \partial/\partial x^i\}$ for a (tangent) space V along with the dual basis $\{\omega^j = dx^j\}$ for the (cotangent) space V*.

If we now go to a new coordinate system $\{\bar{x}^i\}$ in the same coordinate patch, then from calculus we obtain

$$\partial/\partial\bar{x}^i = (\partial x^j/\partial\bar{x}^i)\partial/\partial x^j$$

so that the expression $\bar{e}_i = e_j a^j{}_i$ implies $a^j{}_i = \partial x^j/\partial\bar{x}^i$. Similarly, we also have

$$d\bar{x}^i = (\partial\bar{x}^i/\partial x^j)dx^j$$

so that $\bar{\omega}^i = b^i{}_j\omega^j$ implies $b^i{}_j = \partial\bar{x}^i/\partial x^j$. Note that the chain rule from calculus shows us that

$$a^i{}_k b^k{}_j = (\partial x^i/\partial\bar{x}^k)(\partial\bar{x}^k/\partial x^j) = \partial x^i/\partial x^j = \delta^i{}_j$$

and thus $(b^i{}_j)$ is indeed the inverse matrix to $(a^i{}_j)$.

Using these results in the above expression for $\bar{T}^{pq}{}_r$, we see that

$$\bar{T}^{pq}{}_r = \frac{\partial \bar{x}^p}{\partial x^i}\frac{\partial \bar{x}^q}{\partial x^j}\frac{\partial x^k}{\partial \bar{x}^r}T^{ij}{}_k$$

which is just the classical definition of the transformation law for a tensor of type $\binom{2}{1}$.

We also remark that in older texts, a contravariant vector is *defined* to have the same transformation properties as the expression $d\bar{x}^i = (\partial\bar{x}^i/\partial x^j)dx^j$, while a covariant vector is *defined* to have the same transformation properties as the expression $\partial/\partial\bar{x}^i = (\partial x^j/\partial\bar{x}^i)\partial/\partial x^j$. ∕∕

Finally, let us define a simple classical tensor operation that is frequently quite useful. To begin with, we have seen that the result of operating on a vector $v = v^i e_i \in V$ with a dual vector $\alpha = \alpha_j\omega^j \in V^*$ is just $\langle\alpha, v\rangle = \alpha_j v^i\langle\omega^j, e_i\rangle = \alpha_j v^i\delta^j{}_i = \alpha_i v^i$. This is sometimes called the **contraction** of α with v. We leave it to the reader to show that the contraction is independent of the particular coordinate system used (see Exercise 11.1.5).

If we start with tensors of higher order, then we can perform the same sort of operation. For example, if we have $S \in \mathcal{T}_2{}^1$ with components $S^i{}_{jk}$ and $T \in \mathcal{T}^2$ with components T^{pq}, then we can form the $\binom{2}{1}$ tensor with components $S^i{}_{jk}T^{jq}$, or a different $\binom{2}{1}$ tensor with components $S^i{}_{jk}T^{pj}$ and so forth. This operation is also called **contraction**. Note that if we start with a $\binom{1}{1}$ tensor T, then we can contract the components of T to obtain the scalar $T^i{}_i$. This is called the **trace** of T.

Exercises

1. (a) Prove Theorem 11.1.
   (b) Prove Theorem 11.2.

2. Prove the four associative and distributive properties of the tensor product given in the text following Theorem 11.1.

3. If the nonsingular transition matrix from a basis $\{e_i\}$ to a basis $\{\bar{e}_i\}$ is given by $A = (a^i{}_j)$, show that the transition matrix from the corresponding dual bases $\{\omega^i\}$ and $\{\bar{\omega}^i\}$ is given by $(A^{-1})^T$.

4. Using the transformation matrices $(a^i{}_j)$ and $(b^i{}_j)$ for the bases $\{e_i\}$ and $\{\bar{e}_i\}$ and the corresponding dual bases $\{\omega^i\}$ and $\{\bar{\omega}^i\}$, verify that $v = v^i e_i = \bar{v}^j\,\bar{e}_j$ and $\alpha = \alpha_i\omega^i = \bar{\alpha}_j\bar{\omega}^j$.

5. If $v \in V$ and $\alpha \in V^*$, show that $\langle \alpha, v \rangle$ is independent of the particular basis chosen for V. Generalize this to arbitrary tensors.

6. Let A_i be a covariant vector field (i.e., $A_i = A_i(x)$) with the transformation rule

$$\bar{A}_i = \frac{\partial x^j}{\partial \bar{x}^i} A_j \enspace .$$

Show that the quantity $\partial_j A_i = \partial A_i / \partial x^j$ does not define a tensor, but that $F_{ij} = \partial_i A_j - \partial_j A_i$ is in fact a second-rank tensor.

11.2 SPECIAL TYPES OF TENSORS

In order to obtain some of the most useful results concerning tensors, we turn our attention to the space $\mathcal{T}_r$ of covariant tensors on V. Generalizing our earlier definition for bilinear forms, we say that a tensor $S \in \mathcal{T}_r$ is **symmetric** if for each pair (i, j) with $1 \le i, j \le r$ and all $v_i \in V$ we have

$$S(v_1, \dots, v_i, \dots, v_j, \dots, v_r) = S(v_1, \dots, v_j, \dots, v_i, \dots, v_r) \enspace .$$

Similarly, $A \in \mathcal{T}_r$ is said to be **antisymmetric** (or **skew-symmetric** or **alternating**) if

$$A(v_1, \dots, v_i, \dots, v_j, \dots, v_r) = -A(v_1, \dots, v_j, \dots, v_i, \dots, v_r) \enspace .$$

Note this definition implies that $A(v_1, \dots, v_r) = 0$ if any two of the v_i are identical. In fact, this was the original definition of an alternating bilinear form. Furthermore, we also see that $A(v_1, \dots, v_r) = 0$ if any v_i is a linear combination of the rest of the v_j. In particular, this means that we must always have $r \le \dim V$ if we are to have a nonzero antisymmetric tensor of type $\binom{0}{r}$ on V.

It is easy to see that if $S_1, S_2 \in \mathcal{T}_r$ are symmetric, then so is $aS_1 + bS_2$ where $a, b \in \mathcal{F}$. Similarly, $aA_1 + bA_2$ is antisymmetric. Therefore the symmetric tensors form a subspace of $\mathcal{T}_r$ which we denote by $\Sigma^r(V)$, and the antisymmetric tensors form another subspace of $\mathcal{T}_r$ which is denoted by $\bigwedge^r(V)$ (some authors denote this space by $\bigwedge^r(V^*)$). Elements of $\bigwedge^r(V)$ are generally called **exterior r-forms**, or simply **r-forms**. According to this terminology, the basis vectors $\{\omega^i\}$ for V^* are referred to as **basis 1-forms**. Note that the only element common to both of these subspaces is the zero tensor.

A particularly important example of an antisymmetric tensor is the determinant function $\det \in \mathcal{T}_n(\mathbb{R}^n)$ (see Theorem 4.9 and the discussion preceding it). Note also that the definition of a symmetric tensor translates into the obvious requirement that (e.g., in the particular case of $\mathcal{T}_2$) $S_{ij} = S_{ji}$, while an antisymmetric tensor obeys $A_{ij} = -A_{ji}$. These definitions can also be extended to include contravariant tensors, although we shall have little need to do so.

It will be extremely convenient for us to now incorporate the treatment of permutation groups given in Section 1.2. In terms of any permutation $\sigma \in S_r$, we may rephrase the above definitions as follows. We say that $S \in \mathcal{T}_r$ is **symmetric** if for every collection $v_1, \dots, v_r \in V$ and each $\sigma \in S_r$ we have

$$S(v_1, \dots, v_r) = S(v_{\sigma 1}, \dots, v_{\sigma r}) .$$

Similarly, we say that $A \in \mathcal{T}_r$ is **antisymmetric** (or **alternating**) if either

$$A(v_1, \dots, v_r) = (\operatorname{sgn} \sigma)A(v_{\sigma 1}, \dots, v_{\sigma r})$$

or

$$A(v_{\sigma 1}, \dots, v_{\sigma r}) = (\operatorname{sgn} \sigma)A(v_1, \dots, v_r)$$

where the last equation follows from the first since $(\operatorname{sgn} \sigma)^2 = 1$. Note that even if $S, T \in \Sigma^r(V)$ are both symmetric, it need not be true that $S \otimes T$ be symmetric (i.e., $S \otimes T \notin \Sigma^{r+r}(V)$). For example, if $S_{ij} = S_{ji}$ and $T_{pq} = T_{qp}$, it does not necessarily follow that $S_{ij}T_{pq} = S_{ip}T_{jq}$. It is also clear that if $A, B \in \bigwedge^r(V)$, then we do not necessarily have $A \otimes B \in \bigwedge^{r+r}(V)$.

Example 11.2 Suppose $\alpha \in \bigwedge^n(V)$, let $\{e_1, \dots, e_n\}$ be a basis for V, and for each $i = 1, \dots, n$ let $v_i = e_j a^j{}_i$ where $a^j{}_i \in \mathcal{F}$. Then, using the multilinearity of α, we may write

$$\alpha(v_1, \dots, v_n) = a^{j_1}{}_1 \cdots a^{j_n}{}_n \alpha(e_{j_1}, \dots, e_{j_n})$$

where the sums are over all $1 \le j_k \le n$. But $\alpha \in \bigwedge^n(V)$ is antisymmetric, and hence $(e_{j_1}, \dots, e_{j_n})$ must be a permutation of $(e_1, \dots, e_n)$ in order that the e_{j_k} all be distinct (or else $\alpha(e_{j_1}, \dots, e_{j_n}) = 0$). This means that we are left with

$$\alpha(v_1, \dots, v_n) = \textstyle\sum_J a^{j_1}{}_1 \cdots a^{j_n}{}_n \, \alpha(e_{j_1}, \dots, e_{j_n})$$

where $\sum_J$ denotes the fact that we are summing over only those values of j_k such that $(j_1, \dots, j_n)$ is a permutation of $(1, \dots, n)$. In other words, we have

$$\alpha(v_1, \dots, v_n) = \textstyle\sum_{\sigma\in S_n} a^{\sigma 1}{}_1 \cdots a^{\sigma n}{}_n\, \alpha(e_{\sigma 1}, \dots, e_{\sigma n}) \ .$$

But now, by the antisymmetry of α, we see that $\alpha(e_{\sigma 1}, \dots, e_{\sigma n}) = (\operatorname{sgn} \sigma)\alpha(e_1, \dots, e_n)$ and hence we are left with

$$\alpha(v_1, \dots, v_n) = \textstyle\sum_{\sigma\in S_n} (\operatorname{sgn} \sigma)\, a^{\sigma 1}{}_1 \cdots a^{\sigma n}{}_n\, \alpha(e_1, \dots, e_n) \ . \qquad (*)$$

Using the definition of determinant and the fact that $\alpha(e_1, \dots, e_n)$ is just some scalar, we finally obtain

$$\alpha(v_1, \dots, v_n) = \det(a^j{}_i)\, \alpha(e_1, \dots, e_n) \ .$$

Referring back to Theorem 4.9, let us consider the special case where $\alpha(e_1, \dots, e_n) = 1$. Note that if $\{\omega^j\}$ is a basis for V^*, then

$$\omega^{\sigma j}(v_i) = \omega^{\sigma j}(e_k a^k{}_i) = a^k{}_i\, \omega^{\sigma j}(e_k) = a^k{}_i\, \delta^{\sigma j}{}_k = a^{\sigma j}{}_i \ .$$

Using the definition of tensor product, we can therefore write (*) as

$$\det(a^j{}_i) = \alpha(v_1, \dots, v_n) = \textstyle\sum_{\sigma\in S_n} (\operatorname{sgn} \sigma)\omega^{\sigma 1} \otimes \cdots \otimes \omega^{\sigma n}(v_1, \dots, v_n)$$

which implies that the determinant function is given by

$$\alpha = \textstyle\sum_{\sigma\in S_n} (\operatorname{sgn} \sigma)\omega^{\sigma 1} \otimes \cdots \otimes \omega^{\sigma n} \ .$$

In other words, if A is a matrix with columns given by $v_1, \dots, v_n$ then $\det A = \alpha(v_1, \dots, v_n)$.

While we went through many detailed manipulations in arriving at these equations, we will assume from now on that the reader understands what was done in this example, and henceforth leave out some of the intermediate steps in such calculations. ∕∕

At the risk of boring some readers, let us very briefly review the meaning of the binomial coefficient $\binom{n}{r} = n!/[r!(n - r)!]$. The idea is that we want to know the number of ways of picking r distinct objects out of a collection of n distinct objects. In other words, how many combinations of n things taken r at a time are there? Well, to pick r objects, we have n choices for the first, then n - 1 choices for the second, and so on down to n - (r - 1) = n - r + 1 choices for the r*th* . This gives us

$$n(n - 1) \cdots (n - r + 1) = n!/(n - r)!$$

as the number of ways of picking r objects out of n if we take into account the order in which the r objects are chosen. In other words, this is the number of injections INJ(r, n) (see Section 4.6). For example, to pick three numbers in order out of the set {1, 2, 3, 4}, we might choose (1, 3, 4), or we could choose (3, 1, 4). It is this kind of situation that we must take into account. But for each distinct collection of r objects, there are r! ways of arranging these, and hence we have over-counted each collection by a factor of r!. Dividing by r! then yields the desired result.

If $\{e_1, \dots, e_n\}$ is a basis for V and $T \in \bigwedge^r(V)$, then T is determined by its values $T(e_{i_1}, \dots, e_{i_r})$ for $i_1 < \cdots < i_r$. Indeed, following the same procedure as in Example 11.2, we see that if $v_i = e_j a^j{}_i$ for $i = 1, \dots, r$ then

$$T(v_1, \dots, v_r) = a^{i_1}{}_1 \cdots a^{i_r}{}_r T(e_{i_1}, \dots, e_{i_r})$$

where each sum is over $1 \leq i_k \leq n$. Furthermore, each collection $\{e_{i_1}, \dots, e_{i_r}\}$ must consist of distinct basis vectors in order that $T(e_{i_1}, \dots, e_{i_r}) \neq 0$. But the antisymmetry of T tells us that for any $\sigma \in S_r$, we must have

$$T(e_{\sigma i_1}, \dots, e_{\sigma i_r}) = (\operatorname{sgn} \sigma) T(e_{i_1}, \dots, e_{i_r})$$

where we may choose $i_1 < \cdots < i_r$. Thus, since the number of ways of choosing r distinct basis vectors $\{e_{i_1}, \dots, e_{i_r}\}$ out of the basis $\{e_1, \dots, e_n\}$ is $\binom{n}{r}$, it follows that

$$\dim \bigwedge^r(V) = \binom{n}{r} = n!/[r!(n-r)!] \;.$$

We will prove this result again when we construct a specific basis for $\bigwedge^r(V)$ (see Theorem 11.8 below).

In order to define linear transformations on $\mathcal{T}_r$ that preserve symmetry (or antisymmetry), we define the **symmetrizing mapping** $\mathcal{S}\colon \mathcal{T}_r \to \mathcal{T}_r$ and **alternation mapping** $\mathcal{A}\colon \mathcal{T}_r \to \mathcal{T}_r$ by

$$(\mathcal{S}T)(v_1, \dots, v_r) = (1/r!)\textstyle\sum_{\sigma \in S_r} T(v_{\sigma 1}, \dots, v_{\sigma r})$$

and

$$(\mathcal{A}T)(v_1, \dots, v_r) = (1/r!)\textstyle\sum_{\sigma \in S_r} (\operatorname{sgn} \sigma) T(v_{\sigma 1}, \dots, v_{\sigma r})$$

where $T \in \mathcal{T}_r(V)$ and $v_1, \dots, v_r \in V$. That these are in fact linear transformations on $\mathcal{T}_r$ follows from the observation that the mapping T_σ defined by

$$T_\sigma(v_1, \dots, v_r) = T(v_{\sigma 1}, \dots, v_{\sigma r})$$

is linear, and any linear combination of such mappings is again a linear transformation.

Given any $\sigma \in S_r$, it will be convenient for us to define the mapping $\hat{\sigma}: V^r \to V^r$ by

$$\hat{\sigma}(v_1, \ldots, v_r) = (v_{\sigma 1}, \ldots, v_{\sigma r}) \ .$$

This mapping permutes the *order* of the vectors in its argument, not the labels (i.e. the indices), and hence its argument must always be $(v_1, v_2, \ldots, v_r)$ or $(w_1, w_2, \ldots, w_r)$ and so forth. Then for any $T \in \mathcal{T}_r(V)$ we define $\sigma T \in \mathcal{T}_r(V)$ by

$$\sigma T = T \circ \hat{\sigma}$$

which is the mapping T_σ defined above. It should be clear that $\sigma(T_1 + T_2) = \sigma T_1 + \sigma T_2$. Note also that if we write

$$\hat{\sigma}(v_1, \ldots, v_r) = (v_{\sigma 1}, \ldots, v_{\sigma r}) = (w_1, \ldots, w_r)$$

then $w_i = v_{\sigma i}$ and therefore for any other $\tau \in S_r$ we have

$$\begin{aligned}\hat{\tau} \circ \hat{\sigma}(v_1, \ldots, v_r) &= \hat{\tau}(w_1, \ldots, w_r)\\ &= (w_{\tau 1}, \ldots, w_{\tau r})\\ &= (v_{\sigma\tau 1}, \ldots, v_{\sigma\tau r})\\ &= \widehat{\sigma\circ\tau}\,(v_1, \ldots, v_r) \ .\end{aligned}$$

This shows that

$$\hat{\tau} \circ \hat{\sigma} = \widehat{\sigma\circ\tau}$$

and hence

$$\sigma(\tau T) = \sigma(T \circ \hat{\tau}) = T \circ (\hat{\tau} \circ \hat{\sigma}) = T \circ (\widehat{\sigma\circ\tau}) = (\sigma \circ \tau)T \ .$$

Note also that in this notation, the alternation mapping is defined as

$$\mathcal{A}T = (1/r!)\textstyle\sum_{\sigma\in S_r} (\operatorname{sgn} \sigma)(\sigma T) \ .$$

Theorem 11.3 The linear mappings $\mathcal{A}$ and $\mathcal{S}$ have the following properties:

(a) $T \in \bigwedge^r(V)$ if and only if $\mathcal{A}T = T$, and $T \in \Sigma^r(V)$ if and only if $\mathcal{S}T = T$.

(b) $\mathcal{A}(\mathcal{T}_r(V)) = \bigwedge^r(V)$ and $\mathcal{S}(\mathcal{T}_r(V)) = \Sigma^r(V)$.

(c) $\mathcal{A}^2 = \mathcal{A}$ and $\mathcal{S}^2 = \mathcal{S}$, i.e., $\mathcal{A}$ and $\mathcal{S}$ are projections.

Proof Since the mapping $\mathcal{A}$ is more useful, we will prove the theorem only for this case, and leave the analogous results for $\mathcal{S}$ to the reader (see Exercise 11.2.1). Furthermore, all three statements of the theorem are interrelated, so we prove them together.

First suppose that $T \in \bigwedge^r(V)$. From the definition of antisymmetric tensor we have $T(v_{\sigma 1}, \dots, v_{\sigma r}) = (\operatorname{sgn} \sigma)T(v_1, \dots, v_r)$, and thus using the fact that the order of S_r is r!, we see that

$$\begin{aligned}\mathcal{A}T(v_1, \dots, v_r) &= (1/r!)\Sigma_{\sigma \in S_r}(\operatorname{sgn}\sigma)T(v_{\sigma 1}, \dots, v_{\sigma r})\\ &= (1/r!)\Sigma_{\sigma \in S_r} T(v_1, \dots, v_r)\\ &= T(v_1, \dots, v_r)\ .\end{aligned}$$

This shows that $T \in \bigwedge^r(V)$ implies $\mathcal{A}T = T$.

Next, let T be any element of $\mathcal{T}_r(V)$. We may fix any particular element $\theta \in S_r$ and then apply $\mathcal{A}T$ to the vectors $v_{\theta 1}, \dots, v_{\theta r}$ to obtain

$$\begin{aligned}\mathcal{A}T(v_{\theta 1}, \dots, v_{\theta r}) &= \mathcal{A}T_\theta(v_1, \dots, v_r)\\ &= (1/r!)\Sigma_{\sigma \in S_r}(\operatorname{sgn}\sigma)T_\theta(v_{\sigma 1}, \dots, v_{\sigma r})\\ &= (1/r!)\Sigma_{\sigma \in S_r}(\operatorname{sgn}\sigma)T(v_{\sigma\theta 1}, \dots, v_{\sigma\theta r})\ .\end{aligned}$$

Now note that $\operatorname{sgn} \sigma = (\operatorname{sgn} \sigma)(\operatorname{sgn} \theta)(\operatorname{sgn} \theta) = (\operatorname{sgn} \sigma\theta)(\operatorname{sgn} \theta)$, and that $S_r = \{\phi = \sigma\theta\colon \sigma \in S_r\}$ (this is essentially what was done in Theorem 4.3). We now see that the right hand side of the above equation is just

$$\begin{aligned}&(1/r!)(\operatorname{sgn}\theta)\Sigma_{\sigma \in S_r}(\operatorname{sgn}\sigma\theta)T(v_{\sigma\theta 1}, \dots, v_{\sigma\theta r})\\ &\qquad= (1/r!)(\operatorname{sgn}\theta)\Sigma_{\phi \in S_r}(\operatorname{sgn}\phi)T(v_{\phi 1}, \dots, v_{\phi r})\\ &\qquad= (\operatorname{sgn}\theta)\mathcal{A}T(v_1, \dots, v_r)\end{aligned}$$

which shows (by definition) that $\mathcal{A}T$ is antisymmetric. In other words, this shows that $\mathcal{A}T \in \bigwedge^r(V)$ for any $T \in \mathcal{T}_r(V)$, or $\mathcal{A}(\mathcal{T}_r(V)) \subset \bigwedge^r(V)$.

Since the result of the earlier paragraph showed that $T = \mathcal{A}T \in \mathcal{A}(\mathcal{T}_r(V))$ for every $T \in \bigwedge^r(V)$, we see that $\bigwedge^r(V) \subset \mathcal{A}(\mathcal{T}_r(V))$, and therefore $\mathcal{A}(\mathcal{T}_r(V)) = \bigwedge^r(V)$. This also shows that if $\mathcal{A}T = T$, then T is necessarily an element of $\bigwedge^r(V)$. It then follows that for any $T \in \mathcal{T}_r(V)$ we have $\mathcal{A}T \in \bigwedge^r(V)$, and hence $\mathcal{A}^2T = \mathcal{A}(\mathcal{A}T) = \mathcal{A}T$ so that $\mathcal{A}^2 = \mathcal{A}$. ∎

Suppose $A_{i_1 \cdots i_r}$ and $T^{i_1 \cdots i_r}$ (where $r \le n = \dim V$ and $1 \le i_k \le n$) are both antisymmetric tensors, and consider their contraction $A_{i_1 \cdots i_r} T^{i_1 \cdots i_r}$. For any particular set of indices $i_1, \dots, i_r$ there will be r! different ordered sets $(i_1, \dots, i_r)$. But by antisymmetry, the values of $A_{i_1 \cdots i_r}$ corresponding to each ordered set will differ only by a sign, and similarly for $T^{i_1 \cdots i_r}$. This means that the product of $A_{i_1 \cdots i_r}$ times $T^{i_1 \cdots i_r}$ summed over the r! ordered sets $(i_1, \ldots, i_r)$ is the same as r! times a single product which we choose to be the indices $i_1, \ldots, i_r$ taken in increasing order. In other words, we have

$$A_{i_1 \cdots i_r} T^{i_1 \cdots i_r} = r!\, A_{|i_1 \cdots i_r|} T^{i_1 \cdots i_r}$$

where $|i_1 \cdots i_r|$ denotes the fact that we are summing over increasing sets of indices only. For example, if we have antisymmetric tensors A_{ijk} and T^{ijk} in $\mathbb{R}^3$, then

$$A_{ijk}T^{ijk} = 3!A_{|ijk|}T^{ijk} = 6A_{123}T^{123}$$

(where, in this case of course, A_{ijk} and T^{ijk} can only differ by a scalar).

There is a simple but extremely useful special type of antisymmetric tensor that we now wish to define. Before doing so however, it is first convenient to introduce another useful notational device. Note that if $T \in \mathcal{T}_r$ and we replace $v_1, \ldots, v_r$ in the definitions of $\mathcal{S}$ and $\mathcal{A}$ by basis vectors e_i, then we obtain an expression in terms of components as

$$\mathcal{S}T_{1 \cdots r} = (1/r!)\Sigma_{\sigma \in S_r} T_{\sigma 1 \cdots \sigma r}$$

and

$$\mathcal{A}T_{1 \cdots r} = (1/r!)\Sigma_{\sigma \in S_r} (\operatorname{sgn} \sigma) T_{\sigma 1 \cdots \sigma r} \;.$$

We will write $T_{(1 \cdots r)} = \mathcal{S}T_{1 \cdots r}$ and $T_{[1 \cdots r]} = \mathcal{A}T_{1 \cdots r}$. For example, we have

$$T_{(ij)} = (1/2!)(T_{ij} + T_{ji})$$

and

$$T_{[ij]} = (1/2!)(T_{ij} - T_{ji}) \;.$$

A similar definition applies to any mixed tensor such as

$$\begin{aligned} T^{k(pq)}{}_{[ij]} &= (1/2!)\{T^{k(pq)}{}_{ij} - T^{k(pq)}{}_{ji}\} \\ &= (1/4)\{T^{kpq}{}_{ij} + T^{kqp}{}_{ij} - T^{kpq}{}_{ji} - T^{kqp}{}_{ji}\} \;. \end{aligned}$$

Note that if $T \in \Sigma^r(V)$, then $T_{(i_1 \cdots i_r)} = T_{i_1 \cdots i_r}$, while if $T \in \bigwedge^r(V)$, then $T_{[i_1 \cdots i_r]} = T_{i_1 \cdots i_r}$.

Now consider the vector space $\mathbb{R}^3$ with the standard orthonormal basis $\{e_1, e_2, e_3\}$. We define the antisymmetric tensor $\varepsilon \in \bigwedge^3(\mathbb{R}^3)$ by the requirement that

$$\varepsilon_{123} = \varepsilon(e_1, e_2, e_3) = +1 \;.$$

Since dim $\bigwedge^3(\mathbb{R}^3) = 1$, this defines all the components of ε by antisymmetry: $\varepsilon_{213} = -\varepsilon_{231} = \varepsilon_{321} = -1$ etc. If $\{\bar{e}_i = e_j a^j{}_i\}$ is any other orthonormal basis for $\mathbb{R}^3$ related to the first basis by an (orthogonal) transition matrix $A = (a^j{}_i)$ with determinant equal to +1, then it is easy to see that

$$\varepsilon(\bar{e}_1, \bar{e}_2, \bar{e}_3) = \det A = +1$$

also. This is because $\varepsilon(e_i, e_j, e_k) = \operatorname{sgn} \sigma$ where σ is the permutation that takes (1, 2, 3) to (i, j, k). Since $\varepsilon \in \bigwedge^3(\mathbb{R}^3)$, we see that $\varepsilon_{[ijk]} = \varepsilon_{ijk}$. The tensor ε is frequently called the **Levi-Civita tensor**. However, we stress that in a non-orthonormal coordinate system, it will not generally be true that $\varepsilon_{123} = +1$.

While we have defined the ε_{ijk} as the components of a tensor, it is just as common to see the **Levi-Civita** (or **permutation**) **symbol** ε_{ijk} *defined* simply as an antisymmetric *symbol* with $\varepsilon_{123} = +1$. In fact, from now on we shall use it in this way as a convenient notation for the sign of a permutation. For notational consistency, we also define the permutation symbol ε^{ijk} to have the same values as ε_{ijk}. A simple calculation shows that $\varepsilon_{ijk}\,\varepsilon^{ijk} = 3! = 6$.

It should be clear that this definition can easily be extended to an arbitrary number of dimensions. In other words, we define

$$\varepsilon_{i_1 \cdots i_s} = \begin{cases} +1 \text{ if } (i_1, \dots, i_s) \text{ is an even permutation of } (1,2, \dots, n) \\ -1 \text{ if } (i_1, \dots, i_s) \text{ is an odd permutation of } (1,2, \dots, n) \\ \ 0 \text{ otherwise} \end{cases} .$$

This is just another way of writing sgn σ where $\sigma \in S_n$. Therefore, using this symbol, we have the convenient notation for the determinant of a matrix $A = (a^i{}_j) \in M_n(\mathcal{F})$ as

$$\det A = \varepsilon_{i_1 \cdots i_n} a^{i_1}{}_1 \cdots a^{i_n}{}_n \ .$$

We now wish to prove a very useful result. To keep our notation from getting too cluttered, it will be convenient for us write $\delta^{ijk}_{pqr} = \delta^i_p \delta^j_q \delta^k_r$. Now note that $\varepsilon_{pqr} = 6\delta^{[123]}_{pqr}$. To see that this true, simply observe that both sides are antisymmetric in (p, q, r), and are equal to 1 if (p, q, r) = (1, 2, 3). (This also gives us a formula for ε_{pqr} as a 3 x 3 determinant with entries that are all Kronecker delta's. See Exercise 11.2.4) Using $\varepsilon^{123} = 1$ we may write this as $\varepsilon^{123}\, \varepsilon_{pqr} = 6\delta^{[123]}_{pqr}$. But now the antisymmetry in (1, 2, 3) yields the general result

$$\varepsilon^{ijk}\varepsilon_{pqr} = 6\delta^{[ijk]}_{pqr} \tag{1}$$

which is what we wanted to show. It is also possible to prove this in another manner that serves to illustrate several useful techniques in tensor algebra.

Example 11.3 Suppose we have an arbitrary tensor $A \in \bigwedge^3(\mathbb{R}^3)$, and hence $A_{[ijk]} = A_{ijk}$. As noted above, the fact that $\dim \bigwedge^3(\mathbb{R}^3) = 1$ means that we must have $A_{ijk} = \lambda\varepsilon_{ijk}$ for some scalar $\lambda \in \mathbb{R}$. Then

$$A_{ijk}\varepsilon^{ijk}\varepsilon_{pqr} = \lambda\varepsilon_{ijk}\varepsilon^{ijk}\varepsilon_{pqr} = 6\lambda\varepsilon_{pqr} = 6\lambda\varepsilon_{ijk}\delta^i_p\delta^j_q\delta^k_r \;.$$

Because $\varepsilon_{ijk} = \varepsilon_{[ijk]}$, we can antisymmetrize over the indices i, j, k on the right hand side of the above equation to obtain $6\lambda\varepsilon_{ijk}\delta^{[ijk]}_{pqr}$ (write out the 6 terms if you do not believe it, and see Exercise 11.2.7). This gives us

$$A_{ijk}\varepsilon^{ijk}\varepsilon_{pqr} = 6\lambda\varepsilon_{ijk}\delta^{[ijk]}_{pqr} = 6A_{ijk}\delta^{[ijk]}_{pqr} \;.$$

Since the antisymmetric tensor A_{ijk} is contracted with another antisymmetric tensor on both sides of this equation, the discussion following the proof of Theorem 11.3 shows that we may write

$$A_{|ijk|}\varepsilon^{ijk}\varepsilon_{pqr} = 6A_{|ijk|}\delta^{[ijk]}_{pqr}$$

or

$$A_{123}\varepsilon^{123}\varepsilon_{pqr} = 6A_{123}\delta^{[123]}_{pqr} \;.$$

Cancelling A_{123} then yields $\varepsilon^{123}\varepsilon_{pqr} = 6\delta^{[123]}_{pqr}$ as we had above, and hence (1) again follows from this.

In the particular case that p = k, we leave it to the reader to show that

$$\varepsilon^{ijk}\varepsilon_{kqr} = 6\delta^{[ijk]}_{kqr} = \delta^{ij}_{qr} - \delta^{ji}_{qr} = \delta^i_q\delta^j_r - \delta^j_q\delta^i_r$$

which is very useful in manipulating vectors in $\mathbb{R}^3$. As an example, consider the vectors $\vec{A}$, $\vec{B}$, $\vec{C} \in \mathbb{R}^3$ equipped with a *Cartesian* coordinate system. Abusing our notation for simplicity (alternatively, we will see formally in Example 11.12 that $A^i = A_i$ for such an $\vec{A}$), we have

$$(\vec{B} \times \vec{C})^i = \varepsilon_{ijk}B^jC^k$$

and hence

$$\vec{A} \bullet (\vec{B} \times \vec{C}) = A^i\varepsilon_{ijk}B^jC^k = +\varepsilon_{jki}B^jC^kA^i = \vec{B} \bullet (\vec{C} \times \vec{A}) \;.$$

Other examples are given in the exercises. ∕∕

Exercises

1. Prove Theorem 11.3 for the symmetrizing mapping $\mathcal{S}$.

2. Using the Levi-Civita symbol, prove the following vector identities in $\mathbb{R}^3$ equipped with a Cartesian coordinate system (where the vectors are actually vector fields where necessary, f is differentiable, and $\nabla^i = \partial_i = \partial/\partial x^i$):
(a) $\vec{A} \times (\vec{B} \times \vec{C}) = (\vec{A} \bullet \vec{C})\vec{B} - (\vec{A} \bullet \vec{B})\vec{C}$
(b) $(\vec{A} \times \vec{B}) \bullet (\vec{C} \times \vec{D}) = (\vec{A} \bullet \vec{C})(\vec{B} \bullet \vec{D}) - (\vec{A} \bullet \vec{D})(\vec{B} \bullet \vec{C})$
(c) $\nabla \times \nabla f = 0$
(d) $\nabla \bullet (\nabla \times \vec{A}) = 0$
(e) $\nabla \times (\nabla \times \vec{A}) = \nabla(\nabla \bullet \vec{A}) - \nabla^2 \vec{A}$
(f) $\nabla \bullet (\vec{A} \times \vec{B}) = \vec{B} \bullet (\nabla \times \vec{A}) - \vec{A} \bullet (\nabla \times \vec{B})$
(g) $\nabla \times (\vec{A} \times \vec{B}) = \vec{A}(\nabla \bullet \vec{B}) - \vec{B}(\nabla \bullet \vec{A}) + (\vec{B} \bullet \nabla)\vec{A} - (\vec{A} \bullet \nabla)\vec{B}$

3. Using the divergence theorem ($\int_V \nabla \bullet \vec{A}\, d^3 x = \int_S \vec{A} \bullet \vec{n}\, da$), prove that

$$\int_V \nabla \times \vec{A}\, d^3x = \int_S \vec{n} \times \vec{A}\, da \ .$$

[*Hint*: Let $\vec{C}$ be a constant vector and show that

$$\vec{C} \bullet \int_V \nabla \times \vec{A}\, d^3x = \int_S (\vec{n} \times \vec{A}) \bullet \vec{C}\, da = \vec{C} \bullet \int_S \vec{n} \times \vec{A}\, da \ .]$$

4. (a) Find the expression for ε_{pqr} as a 3 x 3 determinant with all Kronecker delta's as entries.
(b) Write $\varepsilon^{ijk}\varepsilon_{pqr}$ as a 3 x 3 determinant with all Kronecker delta's as entries.

5. Suppose $V = \mathbb{R}^3$ and let A_{ij} be antisymmetric and S^{ij} be symmetric. Show that $A_{ij}S^{ij} = 0$ in two ways:
(a) Write out all terms in the sum and show that they cancel in pairs.
(b) Justify each of the following equalities:

$$A_{ij}S^{ij} \ = \ A_{ij}S^{ji} \ = \ -A_{ji}S^{ji} \ = \ -A_{ij}S^{ij} \ = \ 0 \ .$$

6. Show that a second-rank tensor T_{ij} can be written in the form $T_{ij} = T_{(ij)} + T_{[ij]}$, but that a third-rank tensor can not. (The complete solution for ten-

sors of rank higher than two requires a detailed study of representations of the symmetric group and Young diagrams.)

7. Let $A_{i_1 \cdots i_r}$ be antisymmetric, and suppose $T^{i_1 \cdots i_r \cdots}{}_{\cdots}$ is an arbitrary tensor. Show that

$$A_{i_1 \cdots i_r} T^{i_1 \cdots i_r \cdots}{}_{\cdots} = A_{i_1 \cdots i_r} T^{[i_1 \cdots i_r] \cdots}{}_{\cdots} \ .$$

8. (a) Let $A = (a^i{}_j)$ be a 3 x 3 matrix. Show

$$\varepsilon_{ijk}\, a^i{}_p\, a^j{}_q\, a^k{}_r = (\det A)\varepsilon_{pqr} \ .$$

(b) Let A be a linear transformation, and let $y_{(i)} = Ax_{(i)}$. Show

$$\det[y_{(1)}, \ldots, y_{(n)}] = (\det A) \det[x_{(1)}, \ldots, x_{(n)}] \ .$$

9. Show that under orthogonal transformations $A = (a^i{}_j)$ in $\mathbb{R}^3$, the vector cross product $\vec{x} = \vec{y} \times \vec{z}$ transforms as $\bar{x}^i = \varepsilon_{ijk}\, \bar{y}^j \bar{z}^k = (\det A) a^i{}_j x^j$. Discuss the difference between the cases det A = +1 and det A = −1.

11.3 THE EXTERIOR PRODUCT

We have seen that the tensor product of two elements of $\bigwedge^r(V)$ is not generally another element of $\bigwedge^{r+r}(V)$. However, using the mapping $\mathcal{A}$ we can define another product on $\bigwedge^r(V)$ that turns out to be of great use. We adopt the convention of denoting elements of $\bigwedge^r(V)$ by Greek letters such as α, β etc., which should not be confused with elements of the permutation group S_r.

If $\alpha \in \bigwedge^r(V)$ and $\beta \in \bigwedge^s(V)$, we define their **exterior product** (or **wedge product**) $\alpha\wedge\beta$ to be the mapping from $\bigwedge^r(V) \times \bigwedge^s(V) \to \bigwedge^{r+s}(V)$ given by

$$\alpha \wedge \beta = \frac{(r+s)!}{r!s!} \mathcal{A}(\alpha \otimes \beta) \ .$$

In other words, the wedge product is just an antisymmetrized tensor product. The reader may notice that the numerical factor is just the binomial coefficient $\binom{r+s}{r} = \binom{r+s}{s}$. It is also worth remarking that many authors leave off this coefficient entirely. While there is no fundamental reason for following either convention, our definition has the advantage of simplifying expressions involving volume elements as we shall see later.

A very useful formula for computing exterior products for small values of r and s is given in the next theorem. By way of terminology, a permutation $\sigma \in S_{r+s}$ such that $\sigma 1 < \cdots < \sigma r$ and $\sigma(r + 1) < \cdots < \sigma(r + s)$ is called an (r, s)-**shuffle**. The proof of the following theorem should help to clarify this definition.

Theorem 11.4 Suppose $\alpha \in \bigwedge^r(V)$ and $\beta \in \bigwedge^s(V)$. Then for any collection of r + s vectors $v_i \in V$ (with $r + s \le \dim V$), we have

$$\alpha\wedge\beta(v_1, \dots, v_{r+s}) = \Sigma^* (\operatorname{sgn} \sigma)\alpha(v_{\sigma 1}, \dots, v_{\sigma r})\beta(v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)})$$

where Σ^* denotes the sum over all permutations $\sigma \in S_{r+s}$ such that $\sigma 1 < \cdots < \sigma r$ and $\sigma(r + 1) < \cdots < \sigma(r + s)$ (i.e., over all (r, s)-shuffles).

Proof The proof is simply a careful examination of the terms in the definition of $\alpha\wedge\beta$. By definition, we have

$$\begin{aligned}\alpha \wedge \beta(v_1, \dots, v_{r+s}) & \\ &= [(r+s)!/r!s!]\mathcal{A}(\alpha \otimes \beta)(v_1, \dots, v_{r+s}) \qquad (*)\\ &= [1/r!s!]\Sigma_\sigma(\operatorname{sgn}\sigma)\alpha(v_{\sigma 1}, \dots, v_{\sigma r})\beta(v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)})\end{aligned}$$

where the sum is over all $\sigma \in S_{r+s}$. Now note that there are only $\binom{r+s}{r}$ *distinct* collections $\{\sigma 1, \dots, \sigma r\}$, and hence there are also only $\binom{r+s}{s} = \binom{r+s}{r}$ distinct collections $\{\sigma(r + 1), \dots, \sigma(r + s)\}$. Let us call the set $\{v_{\sigma 1}, \dots, v_{\sigma r}\}$ the "α-variables," and the set $\{v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)}\}$ the "β-variables." For any of the $\binom{r+s}{r}$ distinct collections of α- and β-variables, there will be r! ways of ordering the α-variables within themselves, and s! ways of ordering the β-variables within themselves. Therefore, there will be r!s! possible arrangements of the α- and β-variables within themselves for each of the $\binom{r+s}{r}$ distinct collections. Let $\sigma \in S_{r+s}$ be a permutation that yields one of these distinct collections, and assume it is the one with the property that $\sigma 1 < \cdots < \sigma r$ and $\sigma(r + 1) < \cdots < \sigma(r + s)$. The proof will be finished if we can show that all the rest of the r!s! members of this collection are the same.

Let T denote the term in (*) corresponding to our chosen σ. Then T is given by

$$T = (\operatorname{sgn} \sigma)\alpha(v_{\sigma 1}, \dots, v_{\sigma r})\beta(v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)}) .$$

This means that every other term t in the distinct collection containing T will be of the form

$$t = (\operatorname{sgn} \theta)\alpha(v_{\theta 1}, \dots, v_{\theta r})\beta(v_{\theta(r+1)}, \dots, v_{\theta(r+s)})$$

where the permutation $\theta \in S_{r+s}$ is such that the set $\{\theta 1, \dots, \theta r\}$ is the same as the set $\{\sigma 1, \dots, \sigma r\}$ (although possibly in a different order), and similarly, the set $\{\theta(r+1), \dots, \theta(r+s)\}$ is the same as the set $\{\sigma(r+1), \dots, \sigma(r+s)\}$. Thus the α- and β-variables are permuted within themselves. But we may then write $\theta = \phi\sigma$ where $\phi \in S_{r+s}$ is again such that the two sets $\{\sigma 1, \dots, \sigma r\}$ and $\{\sigma(r+1), \dots, \sigma(r+s)\}$ are permuted within themselves. Because none of the transpositions that define the permutation ϕ interchange α- and β-variables, we may use the antisymmetry of α and β separately to obtain

$$\begin{aligned} t &= (\operatorname{sgn}\phi\sigma)\alpha(v_{\phi\sigma 1}, \dots, v_{\phi\sigma r})\beta(v_{\phi\sigma(r+1)}, \dots, v_{\phi\sigma(r+s)}) \\ &= (\operatorname{sgn}\phi\sigma)(\operatorname{sgn}\phi)\alpha(v_{\sigma 1}, \dots, v_{\sigma r})\beta(v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)}) \\ &= (\operatorname{sgn}\sigma)\alpha(v_{\sigma 1}, \dots, v_{\sigma r})\beta(v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)}) \\ &= T \ . \end{aligned}$$

(It was in bringing out only a single factor of sgn ϕ that we used the fact that there is no mixing of α- and β-variables.) In other words, the original sum over all $(r+s)!$ possible permutations $\sigma \in S_{r+s}$ has been reduced to a sum over $\binom{r+s}{r} = (r+s)!/r!s!$ distinct terms, each one of which is repeated $r!s!$ times. We are thus left with

$$\alpha\wedge\beta(v_1, \dots, v_{r+s}) = \sum{}^{*} (\operatorname{sgn}\sigma)\alpha(v_{\sigma 1}, \dots, v_{\sigma r})\beta(v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)})$$

where the sum is over the $(r+s)!/r!s!$ distinct collections $\{v_{\sigma 1}, \dots, v_{\sigma r}\}$ and $\{v_{\sigma(r+1)}, \dots, v_{\sigma(r+s)}\}$ subject to the requirements $\sigma 1 < \cdots < \sigma r$ and $\sigma(r+1) < \cdots < \sigma(r+s)$. ∎

Let us introduce some convenient notation for handling multiple indices. Instead of writing the ordered set $(i_1, \dots, i_r)$, we simply write I where the exact range will be clear from the context. Furthermore, we write $\underrightarrow{I}$ to denote the increasing sequence $(i_1 < \cdots < i_r)$. Similarly, we shall also write v_I instead of $(v_{i_1}, \dots, v_{i_r})$. To take full advantage of this notation, we first define the **generalized permutation symbol** ε by

$$\varepsilon^{j_1\cdots j_r}_{i_1\cdots i_r} = \begin{cases} +1 \text{ if } (j_1, \dots, j_r) \text{ is an even permutation of } (i_1, \dots, i_r) \\ -1 \text{ if } (j_1, \dots, j_r) \text{ is an odd permutation of } (i_1, \dots i_r) \\ 0 \text{ otherwise} \end{cases} \quad .$$

For example, $\varepsilon^{352}_{235} = +1$, $\varepsilon^{431}_{341} = -1$, $\varepsilon^{142}_{231} = 0$ etc. In particular, if $A = (a^j{}_i)$ is an $n \times n$ matrix, then

$$\det A = \varepsilon^{i_1\cdots i_n}_{1\cdots n} a^1{}_{i_1} \cdots a^n{}_{i_n} = \varepsilon^{1\cdots n}_{i_1\cdots i_n} a^{i_1}{}_1 \cdots a^{i_n}{}_n$$

because

$$\varepsilon^{j_1\cdots j_n}_{1\cdots n} = \varepsilon_{j_1\cdots j_n} = \varepsilon^{j_1\cdots j_n} \quad .$$

Using this notation and Theorem 11.4, we may write the wedge product of α and β as

$$\alpha \wedge \beta(v_{i_1}, \dots, v_{i_{r+s}}) = \sum_{\underset{\rightarrow}{J},\, \underset{\rightarrow}{K}} \varepsilon^{j_1\cdots j_r k_1\cdots k_s}_{i_1\cdots i_r\cdots i_{r+s}} \alpha(v_{j_1}, \dots, v_{j_r})\beta(v_{k_1}, \dots, v_{k_s})$$

or most simply in the following form, which we state as a corollary to Theorem 11.4 for easy reference.

Corollary Suppose $\alpha \in \bigwedge^r(V)$ and $\beta \in \bigwedge^s(V)$. Then for any collection of $r + s$ vectors $v_i \in V$ (with $r + s \leq \dim V$) we have

$$\alpha \wedge \beta(v_I) = \sum_{\underset{\rightarrow}{J},\underset{\rightarrow}{K}} \varepsilon^{JK}_I \alpha(v_J)\beta(v_K) \quad .$$

Example 11.4 Suppose $\dim V = 5$ and $\{e_1, \dots, e_5\}$ is a basis for V. If $\alpha \in \bigwedge^2(V)$ and $\beta \in \bigwedge^1(V)$, then

$$\begin{aligned}
&\alpha \wedge \beta(e_5, e_2, e_3)\\
&\quad = \textstyle\sum_{j_1<j_2,k} \varepsilon^{j_1 j_2 k}_{523} \alpha(e_{j_1}, e_{j_2})\beta(e_k)\\
&\quad = \varepsilon^{235}_{523}\alpha(e_2, e_3)\beta(e_5) + \varepsilon^{253}_{523}\alpha(e_2, e_5)\beta(e_3) + \varepsilon^{352}_{523}\alpha(e_3, e_5)\beta(e_2)\\
&\quad = \alpha(e_2, e_3)\beta(e_5) - \alpha(e_2, e_5)\beta(e_3) + \alpha(e_3, e_5)\beta(e_2) \quad . \ /\!\!/
\end{aligned}$$

Our next theorem is a useful result in many computations. It is simply a contraction of indices in the permutation symbols.

Theorem 11.5 Let $I = (i_1, \dots, i_q)$, $J = (j_1, \dots, j_{r+s})$, $K = (k_1, \dots, k_r)$ and $L = (l_1, \dots, l_s)$. Then

$$\sum_{\underset{\rightarrow}{J}} \varepsilon^{IJ}_{1\cdots q+r+s}\, \varepsilon^{KL}_J = \varepsilon^{IKL}_{1\cdots q+r+s}$$

where I, K and L are fixed quantities, and J is summed over all increasing subsets $j_1 < \cdots < j_{r+s}$ of $\{1, \dots, q + r + s\}$.

Proof The only nonvanishing terms on the left hand side can occur when J is a permutation of KL (or else $\varepsilon_J{}^{KL} = 0$), and of these possible permutations, we only have one in the sum, and that is for the increasing set $\underset{\rightarrow}{J}$. If J is an even permutation of KL, then $\varepsilon^{KL}_J = +1$, and $\varepsilon^{IJ}_{1\cdots q+r+s} = \varepsilon^{IKL}_{1\cdots q+r+s}$ since an even number of permutations is required to go from J to KL. If J is an odd permutation of KL, then $\varepsilon_J^{KL} = -1$, and $\varepsilon^{IJ}_{1\,.\,.\,.\,q+r+s} = -\varepsilon^{IKL}_{1\,.\,.\,.\,q+r+s}$ since an odd number of permutations is required to go from J to KL. The conclusion then follows immediately. ∎

Note that we could have let $J = (j_1, \dots, j_r)$ and left out L entirely in Theorem 11.5. The reason we included L is shown in the next example.

Example 11.5 Let us use Theorem 11.5 to give a simple proof of the associativity of the wedge product. In other words, we want to show that

$$\alpha\wedge(\beta\wedge\gamma) = (\alpha\wedge\beta)\wedge\gamma$$

for any $\alpha \in \bigwedge^q(V)$, $\beta \in \bigwedge^r(V)$ and $\gamma \in \bigwedge^s(V)$. To see this, let $I = (i_1, \dots, i_q)$, $J = (j_1, \dots, j_{r+s})$, $K = (k_1, \dots, k_r)$ and $L = (l_1, \dots, l_s)$. Then we have

$$\begin{aligned}\alpha\wedge(\beta\wedge\gamma)(v_1, \dots, v_{q+r+s}) &= \textstyle\sum_{\underrightarrow{I},\underrightarrow{J}}\varepsilon^{IJ}_{1\cdots q+r+s}\alpha(v_I)(\beta\wedge\gamma)(v_J)\\ &= \textstyle\sum_{\underrightarrow{I},\underrightarrow{J}}\varepsilon^{IJ}_{1\cdots q+r+s}\alpha(v_I)\sum_{\underrightarrow{K},\underrightarrow{L}}\varepsilon_J^{KL}\beta(v_K)\gamma(v_L)\\ &= \textstyle\sum_{\underrightarrow{I},\underrightarrow{K},\underrightarrow{L}}\varepsilon^{IKL}_{1\cdots q+r+s}\alpha(v_I)\beta(v_K)\gamma(v_L)\ .\end{aligned}$$

It is easy to see that had we started with $(\alpha\wedge\beta)\wedge\gamma$, we would have arrived at the same sum.

As was the case with the tensor product, we simply write $\alpha\wedge\beta\wedge\gamma$ from now on. Note also that a similar calculation can be done for the wedge product of any number of terms. //

We now wish to prove the basic algebraic properties of the wedge product. This will be facilitated by a preliminary result on the alternation mapping.

Theorem 11.6 If $S \in \mathcal{T}_r$ and $T \in \mathcal{T}_s$, then

$$\mathcal{A}((\mathcal{A}S)\otimes T) = \mathcal{A}(S\otimes T) = \mathcal{A}(S\otimes(\mathcal{A}T))\ .$$

Proof Using the bilinearity of the tensor product and the definition of $\mathcal{A}S$ we may write

$$(\mathcal{A}S)\otimes T = (1/r!)\textstyle\sum_{\sigma\in S_r}(\operatorname{sgn}\sigma)[(\sigma S)\otimes T]\ .$$

For each $\sigma \in S_r$, let $G \subset S_{r+s}$ be the set of permutations ϕ defined by

$$(\phi 1, \dots, \phi(r+s)) = (\sigma 1, \dots, \sigma r, r+1, \dots, r+s)\ .$$

In other words, G consists of all permutations $\phi \in S_{r+s}$ that have the same effect on 1, . . . , r as $\sigma \in S_r$, but leave the remaining terms r + 1, . . . , r + s unchanged. This means that sgn ϕ = sgn σ, and $\phi(S \otimes T) = (\sigma S) \otimes T$ (see Exercise 11.3.1). We then have

$$(\mathcal{A}S) \otimes T = (1/r!)\Sigma_{\phi\in G}\,(\operatorname{sgn}\phi)\phi(S \otimes T)$$

and therefore

$$\begin{aligned}\mathcal{A}((\mathcal{A}S)\otimes T) &= [1/(r+s)!]\Sigma_{\tau\in S_{r+s}}(\operatorname{sgn}\tau)\tau((1/r!)\Sigma_{\phi\in G}(\operatorname{sgn}\phi)\phi(S\otimes T))\\ &= [1/(r+s)!](1/r!)\Sigma_{\phi\in G}\Sigma_{\tau\in S_{r+s}}(\operatorname{sgn}\tau)(\operatorname{sgn}\phi)\tau\phi(S\otimes T)\ .\end{aligned}$$

But for each $\phi \in G$, we note that $S_{r+s} = \{\theta = \tau\phi\colon \tau \in S_{r+s}\}$, and hence

$$\begin{aligned}&[1/(r+s)!]\Sigma_{\tau\in S_{r+s}}(\operatorname{sgn}\tau)(\operatorname{sgn}\phi)\tau\phi(S\otimes T)\\ &\qquad= [1/(r+s)!]\Sigma_{\tau\in S_{r+s}}(\operatorname{sgn}\tau\phi)\tau\phi(S\otimes T)\\ &\qquad= [1/(r+s)!]\Sigma_{\theta\in S_{r+s}}(\operatorname{sgn}\theta)\theta(S\otimes T)\\ &\qquad= \mathcal{A}(S\otimes T)\ .\end{aligned}$$

Since this is independent of the particular $\phi \in G$ and there are r! elements in G, we then have

$$\begin{aligned}\mathcal{A}((\mathcal{A}S)\otimes T) &= (1/r!)\Sigma_{\phi\in G}\mathcal{A}(S\otimes T)\\ &= \mathcal{A}(S\otimes T)(1/r!)\Sigma_{\phi\in G}1\\ &= \mathcal{A}(S\otimes T)\ .\end{aligned}$$

The proof that $\mathcal{A}(S \otimes T) = \mathcal{A}(S \otimes (\mathcal{A}T))$ is similar, and we leave it to the reader (see Exercise 11.3.2). ∎

Note that in defining the wedge product $\alpha\wedge\beta$, there is really nothing that requires us to have $\alpha \in \bigwedge^r(V)$ and $\beta \in \bigwedge^s(V)$. We could just as well be more general and let $\alpha \in \mathcal{T}_r(V)$ and $\beta \in \mathcal{T}_s(V)$. However, if this is the case, then the formula given in Theorem 11.4 most certainly is *not* valid. However, we do have the following corollary to Theorem 11.6.

Corollary For any $S \in \mathcal{T}_r$ and $T \in \mathcal{T}_s$ we have $\mathcal{A}S\wedge T = S\wedge T = S\wedge \mathcal{A}T$.

Proof This follows directly from Theorem 11.6 and the wedge product definition $S\wedge T = [(r + s)!/r!s!]\, \mathcal{A}(S \otimes T)$. ∎

We are now in a position to prove some of the most important properties of the wedge product. Note that this next theorem is stated in terms of the more general definition of the wedge product.

Theorem 11.7 Suppose $\alpha, \alpha_1, \alpha_2 \in \mathcal{T}_q(V)$, $\beta, \beta_1, \beta_2 \in \mathcal{T}_r(V)$, $\gamma \in \mathcal{T}_s(V)$ and $a \in \mathcal{F}$. Then

(a) The wedge product is bilinear. That is,

$$\begin{aligned}(\alpha_1 + \alpha_2)\wedge\beta &= \alpha_1\wedge\beta + \alpha_2\wedge\beta\\ \alpha\wedge(\beta_1 + \beta_2) &= \alpha\wedge\beta_1 + \alpha\wedge\beta_2\\ (a\alpha)\wedge\beta &= \alpha\wedge(a\beta) = a(\alpha\wedge\beta)\end{aligned}$$

(b) $\alpha\wedge\beta = (-1)^{qr}\beta\wedge\alpha$.

(c) The wedge product is associative. That is,

$$\alpha\wedge(\beta\wedge\gamma) = (\alpha\wedge\beta)\wedge\gamma = [(q + r + s)!/q!r!s!]\mathcal{A}(\alpha \otimes \beta \otimes \gamma) \ .$$

Proof (a) This follows from the definition of wedge product, the fact that $\otimes$ is bilinear and $\mathcal{A}$ is linear. This result may also be shown directly in the case that $\alpha, \alpha_1, \alpha_2 \in \bigwedge^q(V)$ and $\beta, \beta_1, \beta_2 \in \bigwedge^r(V)$ by using the corollary to Theorem 11.4 (see Exercise 11.3.3).

(b) This can also be shown directly from the corollary to Theorem 11.4 (see Exercise 11.3.4). Alternatively, we may proceed as follows. First note that for $\sigma \in S_r$, we see that (since for any other $\tau \in S_r$ we have $\tau(\sigma\alpha) = (\tau \circ \sigma)\alpha$, and hence $\tau(\sigma\alpha)(v_1, \ldots, v_r) = \alpha(v_{\tau\sigma 1}, \ldots, v_{\tau\sigma r})$)

$$\begin{aligned}\mathcal{A}(\sigma\alpha)(v_1, \ldots, v_r) &= (1/r!)\Sigma_{\tau\in S_r}(\operatorname{sgn}\tau)\tau(\sigma\alpha)(v_1, \ldots, v_r)\\ &= (1/r!)\Sigma_{\tau\in S_r}(\operatorname{sgn}\tau)\alpha(v_{\tau\sigma 1}, \ldots, v_{\tau\sigma r})\\ &= (1/r!)\Sigma_{\tau\in S_r}(\operatorname{sgn}\tau\sigma)(\operatorname{sgn}\sigma)\alpha(v_{\tau\sigma 1}, \ldots, v_{\tau\sigma r})\\ &= (\operatorname{sgn}\sigma)(1/r!)\Sigma_{\theta\in S_r}(\operatorname{sgn}\theta)\alpha(v_{\theta 1}, \ldots, v_{\theta r})\\ &= (\operatorname{sgn}\sigma)\mathcal{A}\sigma(v_1, \ldots, v_r) \ .\end{aligned}$$

Hence $\mathcal{A}(\sigma\alpha) = (\operatorname{sgn}\sigma)\,\mathcal{A}\alpha$.

Now define $\sigma_0 \in S_{q+r}$ by

$$\sigma_0(1, \ldots, q + r) = (q + 1, \ldots, q + r, 1, \ldots, q) \ .$$

Since σ_0 is just a product of qr transpositions, it follows that $\operatorname{sgn}\sigma_0 = (-1)^{qr}$. We then see that

$$\begin{aligned}\alpha \otimes \beta(v_1, \ldots, v_{q+r}) &= (\beta \otimes \alpha)(v_{\sigma_0 1}, \ldots, v_{\sigma_0(q+r)})\\ &= \sigma_0(\beta \otimes \alpha)(v_1, \ldots, v_{q+r}) \ .\end{aligned}$$

Therefore (ignoring the factorial multiplier which will cancel out from both sides of this equation),

$$\begin{aligned}\alpha\wedge\beta &= \mathcal{A}(\alpha\otimes\beta) = \mathcal{A}(\sigma_0(\beta\otimes\alpha)) = (\operatorname{sgn}\sigma_0)\mathcal{A}(\beta\otimes\alpha)\\ &= (-1)^{qr}\beta\wedge\alpha\ \ .\end{aligned}$$

(c) Using Theorem 11.6, we simply calculate

$$\begin{aligned}(\alpha\wedge\beta)\wedge\gamma &= [(q+r+s)!/(q+r)!s!]\mathcal{A}((\alpha\wedge\beta)\otimes\gamma)\\ &= [(q+r+s)!/(q+r)!s!][(q+r)!/q!r!]\mathcal{A}(\mathcal{A}(\alpha\otimes\beta)\otimes\gamma)\\ &= [(q+r+s)!/(q!r!s!]\mathcal{A}(\alpha\otimes\beta\otimes\gamma)\ \ .\end{aligned}$$

Similarly, we find that $\alpha\wedge(\beta\wedge\gamma)$ yields the same result. We are therefore justified (as we also saw in Example 11.5) in writing simply $\alpha\wedge\beta\wedge\gamma$. Furthermore, it is clear that this result can be extended to any finite number of products. ∎

Example 11.6 Suppose $\alpha \in \mathcal{T}_r$ and $\beta \in \mathcal{T}_s$. Since $\alpha\wedge\beta = (-1)^{rs}\beta\wedge\alpha$, we see that if either r or s is even, then $\alpha\wedge\beta = \beta\wedge\alpha$, but if both r and s are odd, then $\alpha\wedge\beta = -\beta\wedge\alpha$. Therefore if r is odd we have $\alpha\wedge\alpha = 0$, but if r is even, then $\alpha\wedge\alpha$ is not necessarily zero. In particular, any 1-form α always has the property that $\alpha\wedge\alpha = 0$. ∕∕

Example 11.7 If $\alpha_1, \dots, \alpha_5$ are 1-forms on $\mathbb{R}^5$, let us define

$$\beta = \alpha_1\wedge\alpha_3 + \alpha_3\wedge\alpha_5$$

and

$$\gamma = 2\alpha_2\wedge\alpha_4\wedge\alpha_5 - \alpha_1\wedge\alpha_2\wedge\alpha_4\ \ .$$

Using the properties of the wedge product given in Theorem 11.7 we then have

$$\begin{aligned}\beta\wedge\gamma &= (\alpha_1\wedge\alpha_3 + \alpha_3\wedge\alpha_5)\wedge(2\alpha_2\wedge\alpha_4\wedge\alpha_5 - \alpha_1\wedge\alpha_2\wedge\alpha_4)\\ &= 2\alpha_1\wedge\alpha_3\wedge\alpha_2\wedge\alpha_4\wedge\alpha_5 - \alpha_1\wedge\alpha_3\wedge\alpha_1\wedge\alpha_2\wedge\alpha_4\\ &\qquad + 2\alpha_3\wedge\alpha_5\wedge\alpha_2\wedge\alpha_4\wedge\alpha_5 - \alpha_3\wedge\alpha_5\wedge\alpha_1\wedge\alpha_2\wedge\alpha_4\\ &= -2\alpha_1\wedge\alpha_2\wedge\alpha_3\wedge\alpha_4\wedge\alpha_5 - 0 + 0 - \alpha_1\wedge\alpha_2\wedge\alpha_3\wedge\alpha_4\wedge\alpha_5\\ &= -3\alpha_1\wedge\alpha_2\wedge\alpha_3\wedge\alpha_4\wedge\alpha_5\ .\ /\!/\end{aligned}$$

Example 11.8 Suppose $\alpha_1, \dots, \alpha_r \in \bigwedge^1(V)$ and $v_1, \dots, v_r \in V$. Using Theorem 11.5, it is easy to generalize the corollary to Theorem 11.4 to obtain (see Exercise 11.3.5)

$$\begin{aligned}\alpha_1\wedge\cdots\wedge\alpha_r(v_1,\dots,v_r) &= \Sigma_{i_1\cdots i_r}\varepsilon^{i_1\cdots i_r}_{1\cdots r}\,\alpha_1(v_{i_1})\cdots\alpha_r(v_{i_r})\\ &= \det(\alpha_i(v_j))\ \ .\end{aligned}$$

(Note that the sum is not over any increasing indices because each α_i is only a 1-form.) As a special case, suppose $\{e_i\}$ is a basis for V and $\{\omega^j\}$ is the corresponding dual basis. Then $\omega^j(e_i) = \delta^j{}_i$ and hence

$$\begin{aligned}\omega^{i_1}\wedge\cdots\wedge\omega^{i_r}(e_{j_1},\dots,e_{j_r}) &= \Sigma_{k_1\cdots k_r}\varepsilon^{k_1\cdots k_r}_{j_1\cdots j_r}\,\omega^{i_1}(e_{k_1})\cdots\omega^{i_r}(e_{k_r})\\ &= \varepsilon^{i_1\cdots i_r}_{j_1\cdots j_r}\ \ .\end{aligned}$$

In particular, if dim V = n, choosing the indices $(i_1, \dots, i_n) = (1, \dots, n) = (j_1, \dots, j_n)$, we see that

$$\omega^1\wedge\cdots\wedge\omega^n(e_1,\dots,e_n) = 1\ \ .\ /\!\!/$$

Exercises

1. Show that $\phi(S\otimes T) = (\sigma S)\otimes T$ in the proof of Theorem 11.6.

2. Finish the proof of Theorem 11.6 by showing that

$$\mathcal{A}(S\otimes T) = \mathcal{A}(S\otimes(\mathcal{A}T)).$$

3. Using $\alpha_i\in\bigwedge^q(V)$ and $\beta_i\in\bigwedge^r(V)$, prove Theorem 11.7(a) directly from the corollary to Theorem 11.4.

4. Use the corollary to Theorem 11.4 to prove Theorem 11.7(b).

5. Suppose $\alpha_1, \dots, \alpha_r\in\bigwedge^1(V)$ and $v_1, \dots, v_r\in V$. Show that

$$\alpha_1\wedge\cdots\wedge\alpha_r(v_1,\dots,v_r) = \det(\alpha_i(v_j))\ \ .$$

6. Suppose $\{e_1, \dots, e_n\}$ is a basis for V and $\{\omega^1, \dots, \omega^n\}$ is the corresponding dual basis. If $\alpha\in\bigwedge^r(V)$ (where $r\le n$), show that

$$\alpha = \Sigma_{\underline{I}}\,\alpha(e_I)\omega^I = \Sigma_{i_1<\cdots<i_r}\,\alpha(e_{i_1},\dots,e_{i_r})\omega^{i_1}\wedge\cdots\wedge\omega^{i_r}$$

by applying both sides to $(e_{j_1}, \dots, e_{j_r})$. (See also Theorem 11.8.)

7. (**Interior Product**) Suppose $\alpha \in \bigwedge^r(V)$ and $v, v_2, \ldots, v_r \in V$. We define the $(r - 1)$-form $i_v\alpha$ by

$$\begin{array}{ll} i_v\alpha = 0 & \text{if } r = 0. \\ i_v\alpha = \alpha(v) & \text{if } r = 1. \\ i_v\alpha(v_2, \ldots, v_r) = \alpha(v, v_2, \ldots, v_r) & \text{if } r > 1. \end{array}$$

(a) Prove that $i_{u+v} = i_u + i_v$.

(b) If $\alpha \in \bigwedge^r(V)$ and $\beta \in \bigwedge^s(V)$, prove that $i_v\colon \bigwedge^{r+s}(V) \to \bigwedge^{r+s-1}(V)$ is an **anti-derivation**, i.e.,

$$i_v(\alpha\wedge\beta) = (i_v\,\alpha)\wedge\beta + (-1)^r\alpha\wedge(i_v\,\beta)\ .$$

(c) If $v = v^i e_i$ and $\alpha = \sum_{\underline{I}} a_{i_1 \cdots i_r}\omega^{i_1}\wedge \cdots \wedge\omega^{i_r}$ where $\{\omega^i\}$ is the basis dual to $\{e_i\}$, show that

$$i_v\alpha = \sum\nolimits_{i_2< \cdots <i_r} b_{i_2 \cdots i_r}\omega^{i_2}\wedge\cdots\wedge\omega^{i_r}$$

where

$$b_{i_2 \cdots i_r} = \sum\nolimits_j v^j a_{ji_2 \cdots i_r}\ .$$

(d) If $\alpha = f^1\wedge \cdots \wedge f^r$, show that

$$\begin{aligned} i_v\alpha &= \sum_{k=1}^{r}(-1)^{k-1}f^k(v)f^1\wedge\cdots\wedge f^{k-1}\wedge f^{k+1}\wedge\cdots\wedge f^r \\ &= \sum_{k=1}^{r}(-1)^{k-1}f^k(v)f^1\wedge\cdots\wedge \widehat{f^k}\wedge\cdots\wedge f^r \end{aligned}$$

where the ˆ means that the term f^k is to be deleted from the expression.

8. Let $V = \mathbb{R}^n$ have the standard basis $\{e_i\}$, and let the corresponding dual basis for V^* be $\{\omega^i\}$.

(a) If $u, v \in V$, show that

$$\omega^i\wedge\omega^j(u, v) = \begin{vmatrix} u^i & v^i \\ u^j & v^j \end{vmatrix}$$

and that this is ± the area of the parallelogram spanned by the projection of u and v onto the $x^i x^j$-plane. What do you think is the significance of the different signs?

(b) Generalize this to $\omega^{i_1}\wedge \cdots \wedge\omega^{i_r}$ where $r \le n$.

9. Suppose $V = \mathcal{F}^n$, and let $v_1, \dots, v_n \in V$ have components relative to the standard basis $\{e_i\}$ defined by $v_i = e_j v^j{}_i$. For any $1 \le r < n$, let $s = n - r$ and define the r-form α by

$$\alpha(v_1, \dots, v_r) = \begin{vmatrix} v^1{}_1 & \cdots & v^1{}_r \\ \vdots & & \vdots \\ v^r{}_1 & \cdots & v^r{}_r \end{vmatrix}$$

and the s-form β by

$$\beta(v_1, \dots, v_s) = \begin{vmatrix} v^1{}_{r+1} & \cdots & v^1{}_n \\ \vdots & & \vdots \\ v^s{}_{r+1} & \cdots & v^s{}_n \end{vmatrix} .$$

(a) Use Theorem 4.9 to show that $\alpha\wedge\beta$ is the determinant function D on $\mathcal{F}^n$.

(b) Show that the sign of an (r, s)-shuffle is given by

$$\varepsilon^{i_1 \cdots i_r j_1 \cdots j_s}_{1 \cdots r\ r+1 \cdots r+s} = (-1)^{i_1 + \cdots + i_r + r(r+1)/2}$$

where $i_1, \dots, i_r$ and $j_1, \dots, j_s$ are listed in increasing order.

(c) If $A = (a^i{}_j) \in M_n(\mathcal{F})$, prove the **Laplace expansion** formula

$$\det A = \Sigma_{\underset{\rightarrow}{I}} (-1)^{i_1 + \cdots + i_r + r(r+1)/2} \begin{vmatrix} a^{i_1}{}_1 & \cdots & a^{i_1}{}_r \\ \vdots & & \vdots \\ a^{i_r}{}_1 & \cdots & a^{i_r}{}_r \end{vmatrix} \begin{vmatrix} a^{j_1}{}_{r+1} & \cdots & a^{j_1}{}_n \\ \vdots & & \vdots \\ a^{j_r}{}_{r+1} & \cdots & a^{j_r}{}_n \end{vmatrix}$$

where $I = \{i_1, \dots, i_r\}$ and $J = \{j_1, \dots, j_s\}$ are “complementary” sets of indices, i.e., $I \cap J = \varnothing$ and $I \cup J = \{1, 2, \dots, n\}$.

10. Let $\mathcal{B} = r!\mathcal{A}$ where $\mathcal{A}\colon \mathcal{T}_r \to \mathcal{T}_r$ is the alternation mapping. Define $\alpha\wedge\beta$ in terms of $\mathcal{B}$. What is $\mathcal{B}(f^1 \otimes \cdots \otimes f^r)$ where $f^i \in V^*$?

11. Let $I = (i_1, \dots, i_q)$, $J = (j_1, \dots, j_p)$, and $K = (k_1, \dots, k_q)$. Prove the following generalization of Example 11.3:

$$\sum_{\underset{\rightarrow}{J}} \varepsilon^{JI}_{1 \cdots p+q} \varepsilon^{1 \cdots p+q}_{JK} = \varepsilon^I_K = n!\delta^{[i_1}_{k_1} \cdots \delta^{i_q]}_{k_q}$$

11.4 TENSOR ALGEBRAS

We define the direct sum of all tensor spaces $\mathcal{T}_r(V)$ to be the (infinite-dimensional) space $\mathcal{T}_0(V) \oplus \mathcal{T}_1(V) \oplus \cdots \oplus \mathcal{T}_r(V) \oplus \cdots$, and $\mathcal{T}(V)$ to be all elements of this space with finitely many nonzero components. This means that every element $T \in \mathcal{T}(V)$ has a unique expression of the form (ignoring zero summands)

$$T = T^{(1)}{}_{i_1} + \cdots + T^{(r)}{}_{i_r}$$

where each $T^{(k)}{}_{i_k} \in \mathcal{T}_{i_k}(V)$ and $i_1 < \cdots < i_r$. The tensors $T^{(k)}{}_{i_k}$ are called the **graded components** of T. In the special case that $T \in \mathcal{T}_r(V)$ for some r, then T is said to be of **order** r. We define addition in $\mathcal{T}(V)$ componentwise, and we also define multiplication in $\mathcal{T}(V)$ by defining $\otimes$ to be distributive on all of $\mathcal{T}(V)$. We have therefore made $\mathcal{T}(V)$ into an associative algebra over $\mathcal{F}$ which is called the **tensor algebra**.

We have seen that $\bigwedge^r(V)$ is a subspace of $\mathcal{T}_r(V)$ since $\bigwedge^r(V)$ is just the image of $\mathcal{T}_r(V)$ under $\mathcal{A}$. Recall also that $\bigwedge^0(V) = \mathcal{T}_0(V)$ is defined to be the scalar field $\mathcal{F}$. As might therefore be expected, we define $\bigwedge(V)$ to be the direct sum

$$\bigwedge(V) = \bigwedge^0(V) \oplus \bigwedge^1(V) \oplus \cdots \oplus \bigwedge^r(V) \oplus \cdots \subset \mathcal{T}(V) \ .$$

Note that $\bigwedge^r(V) = 0$ if $r > \dim V$.

It is important to realize that if $\alpha \in \bigwedge^r(V) \subset \mathcal{T}_r(V)$ and $\beta \in \bigwedge^s(V) \subset \mathcal{T}_s(V)$, then even though $\alpha \otimes \beta \in \mathcal{T}_{r+s}(V)$, it is not generally true that $\alpha \otimes \beta \in \bigwedge^{r+s}(V)$. Therefore $\bigwedge(V)$ is not a subalgebra of $\mathcal{T}(V)$. However, the wedge product is a mapping from $\bigwedge^r(V) \times \bigwedge^s(V) \rightarrow \bigwedge^{r+s}(V)$, and hence if we extend this product in the obvious manner to a bilinear mapping $\bigwedge(V) \times \bigwedge(V) \rightarrow \bigwedge(V)$, then $\bigwedge(V)$ becomes an algebra over $\mathcal{F} = \bigwedge^0(V)$. In other words, if $\alpha = \alpha_1 + \cdots + \alpha_r$ with each $\alpha_i \in \bigwedge^{r_i}(V)$, and $\beta = \beta_1 + \cdots + \beta_s$ with each $\beta_i \in \bigwedge^{s_i}(V)$, then we define

$$\alpha \wedge \beta = \sum_{i=1}^{r} \sum_{j=1}^{s} \alpha_i \wedge \beta_j \ .$$

This algebra is called the **Grassmann** (or **exterior**) **algebra**.

The astute reader may be wondering exactly how we add together the elements $\alpha_1 \in \bigwedge^{r_1}(V)$ and $\alpha_2 \in \bigwedge^{r_2}(V)$ (with $r_1 \neq r_2$) when none of the operations $(\alpha_1 + \alpha_2)(v_1, \ldots, v_{r_1})$, $(\alpha_1 + \alpha_2)(v_1, \ldots, v_{r_2})$ nor $(\alpha_1 + \alpha_2)(v_1, \ldots, v_{r_1+r_2})$ makes any sense. The answer is that for purposes of the Grassmann algebra, we consider both α_1 and α_2 to be elements of $\bigwedge(V)$. For example, if

α_1 is a 1-form and α_2 is a 2-form, then we write $\alpha_1 = 0 + \alpha_1 + 0 + 0 + \cdots$ and $\alpha_2 = 0 + 0 + \alpha_2 + 0 + \cdots$, and hence $a_1\alpha_1 + a_2\alpha_2$ (where $a_i \in \mathcal{F}$) makes sense in $\bigwedge(V)$. In this way, every element of $\bigwedge(V)$ has a degree (recall that an r-form is said to be of degree r), and we say that $\bigwedge(V)$ is a **graded associative algebra**.

Unlike the infinite-dimensional algebra $\mathcal{T}(V)$, the algebra $\bigwedge(V)$ is finite-dimensional. This should be clear from the discussion following Example 11.2 where we showed that $\dim \bigwedge^r(V) = \binom{n}{r}$ (where $n = \dim V$), and hence that $\dim \bigwedge^r(V) = 0$ if $r > n$. Let us now prove this result again by constructing a specific basis for $\bigwedge^r(V)$.

Theorem 11.8 Suppose $\dim V = n$. Then for $r > n$ we have $\bigwedge^r(V) = \{0\}$, and if $0 \le r \le n$, then $\dim \bigwedge^r(V) = \binom{n}{r}$. Therefore $\dim \bigwedge(V) = 2^n$. Moreover, if $\{\omega^1, \dots, \omega^n\}$ is a basis for $V^* = \bigwedge^1(V)$, then a basis for $\bigwedge^r(V)$ is given by the set

$$\{\omega^{i_1} \wedge \cdots \wedge \omega^{i_r} : 1 \le i_1 < \cdots < i_r \le n\} \ .$$

Proof Suppose $\alpha \in \bigwedge^r(V)$ where $r > \dim V = n$. By multilinearity, α is determined by its values on a basis $\{e_1, \dots, e_n\}$ for V (see Example 11.2). But then we must have $\alpha(e_{i_1}, \dots, e_{i_r}) = 0$ because at least two of the e_{i_k} are necessarily the same and α is antisymmetric. This means that $\alpha(v_1, \dots, v_r) = 0$ for all $v_i \in V$, and hence $\alpha = 0$. Thus $\bigwedge^r(V) = \{0\}$ if $r > n$.

Now suppose that $\{\omega^1, \dots, \omega^n\}$ is the basis for V^* dual to $\{e_i\}$. From Theorem 11.2, we know that $\{\omega^{i_1} \otimes \cdots \otimes \omega^{i_r} : 1 \le i_1, \dots, i_r \le n\}$ forms a basis for $\mathcal{T}_r(V)$, and since the alternation mapping $\mathcal{A}$ maps $\mathcal{T}_r(V)$ onto $\bigwedge^r(V)$ (Theorem 11.3(b)), it follows that the image of the basis $\{\omega^{i_1} \otimes \cdots \otimes \omega^{i_r}\}$ for $\mathcal{T}_r(V)$ must span $\bigwedge^r(V)$. If $\alpha \in \bigwedge^r(V)$, then $\alpha \in \mathcal{T}_r(V)$, and hence

$$\alpha = \alpha_{i_1 \cdots i_r} \omega^{i_1} \otimes \cdots \otimes \omega^{i_r}$$

where the sum is over all $1 \le i_1, \dots, i_r \le n$ and $\alpha_{i_1 \cdots i_r} = \alpha(e_{i_1}, \dots, e_{i_r})$. Using Theorems 11.3(a) and 11.7(c) we have

$$\begin{aligned} \alpha = \mathcal{A}\alpha &= \alpha_{i_1 \cdots i_r} \mathcal{A}(\omega^{i_1} \otimes \cdots \otimes \omega^{i_r}) \\ &= \alpha_{i_1 \cdots i_r} (1/r!) \omega^{i_1} \wedge \cdots \wedge \omega^{i_r} \end{aligned}$$

where the sum is still over all $1 \le i_1, \dots, i_r \le n$. However, by the antisymmetry of the wedge product, the collection $\{i_1, \dots, i_r\}$ must all be different, and hence the sum can only be over the $\binom{n}{r}$ distinct such combinations. For each such combination there will be r! permutations $\sigma \in S_r$ of the basis vectors. If we write each of these permutations in increasing order $i_1 < \cdots < i_r$, then the wedge product changes by a factor sgn σ, as does $\alpha_{i_1 \cdots i_r} = \alpha(e_{i_1}, \ldots, e_{i_r})$. Therefore the signs cancel and we are left with

$$\alpha = \alpha_{|i_1 \cdots i_r|}\,\omega^{i_1} \wedge \cdots \wedge \omega^{i_r}$$

where, as mentioned previously, we use the notation $\alpha_{|\, i_1 \cdots i_r|}$ to mean that the sum is over increasing sets $i_1 < \cdots < i_r$. Thus we have shown that the $\binom{n}{r}$ elements $\omega^{i_1} \wedge \cdots \wedge \omega^{i_r}$ with $1 \le i_1 < \cdots < i_r \le n$ span $\bigwedge^r(V)$. We must still show that they are linearly independent.

Suppose $\alpha_{|i_1 \cdots i_r|}\,\omega^{i_1} \wedge \cdots \wedge \omega^{i_r} = 0$. Then for any set $\{e_{j_1}, \ldots, e_{j_r}\}$ with $1 \le j_1 < \cdots < j_r \le n$ we have (using Example 11.8)

$$\begin{aligned} 0 &= \alpha_{|i_1 \cdots i_r|}\omega^{i_1} \wedge \cdots \wedge \omega^{i_r}(e_{j_1}, \ldots, e_{j_r}) \\ &= \alpha_{|i_1 \cdots i_r|}\varepsilon^{i_1 \cdots i_r}_{j_1 \cdots j_r} \\ &= \alpha_{j_1 \cdots j_r} \end{aligned}$$

since the only nonvanishing term occurs when $\{i_1, \ldots, i_r\}$ is a permutation of $\{j_1, \ldots, j_r\}$ and both are increasing sets. This proves linear independence.

Finally, using the binomial theorem, we now see that

$$\dim \bigwedge(V) = \sum_{r=0}^{n} \dim \bigwedge^{\mathrm{r}}(V) = \sum_{r=0}^{n}\binom{n}{r} = (1+1)^n = 2^n \quad . \ ∎$$

Example 11.9 Another useful result is the following. Suppose dim V = n, and let $\{\omega^1, \ldots, \omega^n\}$ be a basis for V*. If $\alpha^1, \ldots, \alpha^n$ are any other 1-forms in $\bigwedge^1(V) = V^*$, then we may expand each α^i in terms of the ω^j as $\alpha^i = a^i{}_j\omega^j$. We then have

$$\begin{aligned} \alpha^1 \wedge \cdots \wedge \alpha^n &= a^1{}_{i_n} \cdots a^n{}_{i_n}\omega^{i_1} \wedge \cdots \wedge \omega^{i_n} \\ &= a^1{}_{i_n} \cdots a^n{}_{i_n}\varepsilon^{i_1 \cdots i_n}_{1 \cdots n}\omega^1 \wedge \cdots \wedge \omega^n \\ &= \det(a^i{}_j)\omega^1 \wedge \cdots \wedge \omega^n \quad . \end{aligned}$$

Recalling Example 11.1, if $\{\omega^i = dx^i\}$ is a local basis for a cotangent space V* and $\{\alpha^i = dy^i\}$ is any other local basis, then $dy^i = (\partial y^i/\partial x^j)dx^j$ and

$$\det(a^i{}_j) = \frac{\partial(y^1\cdots y^n)}{\partial(x^1\cdots x^n)}$$

is just the usual Jacobian of the transformation. We then have

$$dy^1 \wedge\cdots\wedge dy^n = \frac{\partial(y^1\cdots y^n)}{\partial(x^1\cdots x^n)} dx^1 \wedge\cdots\wedge dx^n \quad .$$

The reader may recognize $dx^1\wedge \cdots \wedge dx^n$ as the volume element on $\mathbb{R}^n$, and hence differential forms are a natural way to describe the change of variables in multiple integrals. ∕∕

Theorem 11.9 If $\alpha^1, \ldots, \alpha^r \in \bigwedge^1(V)$, then $\{\alpha^1, \ldots, \alpha^r\}$ is a linearly dependent set if and only if $\alpha^1\wedge \cdots \wedge\alpha^r = 0$.

Proof If $\{\alpha^1, \ldots, \alpha^r\}$ is linearly dependent, then there exists at least one vector, say α^1, such that $\alpha^1 = \Sigma_{j\neq1}a_j\alpha^j$. But then

$$\begin{aligned}\alpha^1 \wedge\cdots\wedge \alpha^r &= (\Sigma_{j\neq1}a_j\alpha^j)\wedge\alpha^2 \wedge\cdots\wedge\alpha^r\\ &= \Sigma_{j\neq1}a_j(\alpha^j\wedge\alpha^2\wedge\cdots\wedge\alpha^r)\\ &= 0\end{aligned}$$

since every term in the sum contains a repeated 1-form and hence vanishes.

Conversely, suppose that $\alpha^1, \ldots, \alpha^r$ are linearly *independent*. We can then extend them to a basis $\{\alpha^1, \ldots, \alpha^n\}$ for V* (Theorem 2.10). If $\{e_i\}$ is the corresponding dual basis for V, then $\alpha^1\wedge \cdots \wedge\alpha^n(e_1, \ldots, e_n) = 1$ which implies that $\alpha^1\wedge \cdots \wedge\alpha^r \neq 0$. Therefore $\{\alpha^1, \ldots, \alpha^r\}$ must be linearly *dependent* if $\alpha^1\wedge \cdots \wedge\alpha^r = 0$. ∎

11.5 THE TENSOR PRODUCT OF VECTOR SPACES

We now discuss the notion of the tensor product of vector spaces. Our reason for presenting this discussion is that it provides the basis for defining the Kronecker (or direct) product of two matrices, a concept which is very useful in the theory of group representations.

It should be remarked that there are many ways of defining the tensor product of vector spaces. While we will follow the simplest approach, there is another (somewhat complicated) method involving quotient spaces that is also frequently used. This other method has the advantage that it includes infinite-dimensional spaces. The reader can find a treatment of this alternative method in, e.g., in the book by Curtis (1984).

By way of nomenclature, we say that a mapping f: U × V → W of vector spaces U and V to a vector space W is **bilinear** if f is linear in each variable. This is exactly the same as we defined in Section 9.4 except that now f takes its values in W rather than the field $\mathcal{F}$. In addition, we will need the concept of a vector space generated by a set. In other words, suppose $S = \{s_1, \dots , s_n\}$ is some finite set of objects, and $\mathcal{F}$ is a field. While we may have an intuitive sense of what it should mean to write formal linear combinations of the form $a_1s_1 + \cdots + a_ns_n$, we should realize that the + sign as used here has no meaning for an arbitrary set S. We now go through the formalities involved in defining such terms, and hence make the set S into a vector space T over $\mathcal{F}$.

The basic idea is that we want to recast each element of S into the form of a function from S to $\mathcal{F}$. This is because we already know how to add functions as well as multiply them by a scalar. With these ideas in mind, for each $s_i \in S$ we define a function $s_i: S \to \mathcal{F}$ by

$$s_i(s_j) = 1\delta_{ij}$$

where 1 is the multiplicative identity of $\mathcal{F}$. Since addition in $\mathcal{F}$ is well-defined as is the addition of functions and multiplication of functions by elements of $\mathcal{F}$, we see that for any $a, b \in \mathcal{F}$ and $s_i, s_j \in S$ we have

$$\begin{aligned}(a+b)s_i(s_j) &= (a+b)\delta_{ij} = a\delta_{ij} + b\delta_{ij} = as_i(s_j) + bs_i(s_j)\\ &= (as_i + bs_i)(s_j)\end{aligned}$$

and therefore $(a + b)s_i = as_i + bs_i$. Similarly, it is easy to see that $a(bs_i) = (ab)s_i$.

We now define T to be the set of all functions from S to $\mathcal{F}$. These functions can be written in the form $a_1s_1 + \cdots + a_ns_n$ with $a_i \in \mathcal{F}$. It should be clear that with our definition of the terms a_is_i, T forms a vector space over $\mathcal{F}$. In fact, it is easy to see that the functions $1s_1 , \dots , 1s_n$ are linearly independent. Indeed, if 0 denotes the zero function, suppose $a_1s_1 + \cdots + a_ns_n = 0$ for some set of scalars a_i. Applying this function to s_i (where $1 \le i \le n$) we obtain $a_i = 0$. As a matter of course, we simply write s_i rather than $1s_i$.

The linear combinations just defined are called **formal** linear combinations of the elements of S, and T is the vector space **generated** by the set S. T is therefore the vector space of all such formal linear combinations, and is sometimes called the **free** vector space of S over $\mathcal{F}$.

Theorem 11.10 Let U, V and W be finite-dimensional vector spaces over $\mathcal{F}$. Then there exists a finite-dimensional vector space over $\mathcal{F}$ denoted by T and a bilinear mapping t: U × V → T denoted by $t(u, v) = u \otimes v$ satisfying the following properties:

(a) For every bilinear mapping f: U × V → W, there exists a unique linear transformation $\tilde{f} : T \to W$ such that $f = \tilde{f} \circ t$. In other words, for all $u \in U$ and $v \in V$ we have

$$f(u, v) = \tilde{f}(t(u, v)) = \tilde{f}(u \otimes v) .$$

(b) If $\{u_1, \dots, u_m\}$ is a basis for U and $\{v_1, \dots, v_n\}$ is a basis for V, then $\{u_i \otimes v_j\}$ is a basis for T and therefore

$$\dim T = mn = (\dim U)(\dim V) .$$

Proof Let $\{u_1, \dots, u_m\}$ be a basis for U and let $\{v_1, \dots, v_n\}$ be a basis for V. For each pair of integers (i, j) with $1 \le i \le m$ and $1 \le j \le n$ we let t_{ij} be a letter (i.e., an element of some set). We now define T to be the vector space over $\mathcal{F}$ consisting of all formal linear combinations of the elements t_{ij}. In other words, every element of T is of the form $a^{ij} t_{ij}$ where $a^{ij} \in \mathcal{F}$.

Define the bilinear map t: U × V → T by

$$u_i \otimes v_j \equiv t(u_i, v_j) = t_{ij}$$

and hence to all of U × V by "bilinear extension." In particular, if $u = x^i u_i \in U$ and $v = y^j v_j \in V$, let us define $u \otimes v$ to be that element of T given by

$$u \otimes v = t(u, v) = x^i y^j t_{ij} .$$

It should be obvious that this does indeed define a bilinear map.

Now suppose that f: U × V → W is any bilinear map, and remember that every element of T is a linear combination of the t_{ij}. According to Theorem 5.1, we may define a unique linear transformation $\tilde{f} : T \to W$ by

$$\tilde{f}(t_{ij}) = f(u_i, v_j) .$$

Using the bilinearity of f and the linearity of $\tilde{f}$ we then have

$$\begin{aligned} f(u, v) &= f(x^i u_i, y^j v_j) = x^i y^j f(u_i, v_j) = x^i y^j \tilde{f}(t_{ij}) = \tilde{f}(x^i y^j t_{ij}) \\ &= \tilde{f}(u \otimes v) = \tilde{f}(t(u, v)) . \end{aligned}$$

This proves the existence and uniqueness of the mapping $\tilde{f}$ such that $f = \tilde{f} \circ t$ as specified in (a).

We have defined T to be the vector space generated by the mn elements $t_{ij} = u_i \otimes v_j$ where $\{u_1, . . . , u_m\}$ and $\{v_1, . . . , v_n\}$ were particular bases for U and V respectively. We now want to show that in fact $\{u'_i \otimes v'_j\}$ forms a basis for T where $\{u'_i\}$ and $\{v'_j\}$ are arbitrary bases for U and V. For any $u = x'^i u'_i \in U$ and $v = y'^j v'_j \in V$, we have (using the bilinearity of $\otimes$)

$$u \otimes v = x'^i y'^j (u'_i \otimes v'_j)$$

which shows that the mn elements $u'_i \otimes v'_j$ span T. If these mn elements are linearly dependent, then dim T < mn which contradicts the fact that the mn elements t_{ij} form a basis for T. Hence $\{u'_i \otimes v'_j\}$ is a basis for T. ∎

The space T defined in this theorem is denoted by $U \otimes V$ and called the **tensor product** of U and V. Note that T can be any mn dimensional vector space. For example, if m = n we could take $T = \mathcal{T}_2(V)$ with basis $t_{ij} = \omega^i \otimes \omega^j$, $1 \le i, j \le n$. The map $t(u_i, v_j) = u_i \otimes v_j$ then defines $u_i \otimes v_j = \omega^i \otimes \omega^j$.

Example 11.10 To show how this formalism relates to our previous treatment of tensors, consider the following example of the mapping $\tilde{f}$ defined in Theorem 11.10. Let $\{e_i\}$ be a basis for a real inner product space U, and let us define the real numbers $g_{ij} = \langle e_i, e_j \rangle$. If $\bar{e}_i = e_j p^j{}_i$ is another basis for U, then

$$\bar{g}_{ij} = \langle \bar{e}_i, \bar{e}_j \rangle = p^r{}_i p^s{}_j \langle e_r, e_s \rangle = p^r{}_i p^s{}_j \, g_{rs}$$

so that the g_{ij} transform like the components of a covariant tensor of order 2. This means that we may define the tensor $g \in \mathcal{T}_2(U)$ by $g(u, v) = \langle u, v \rangle$. This tensor is called the **metric tensor** on U (see Section 11.10).

Now suppose that we are given a positive definite symmetric bilinear form (i.e., an inner product) $g = \langle\ ,\ \rangle: U \times U \to \mathcal{F}$. Then the mapping $\tilde{g}$ is just the metric because

$$\tilde{g}(e_i \otimes e_j) = g(e_i, e_j) = \langle e_i, e_j \rangle = g_{ij} \; .$$

Therefore, if $u = u^i e_i$ and $v = v^j e_j$ are vectors in U, we see that

$$\tilde{g}(u \otimes v) = g(u, v) = \langle u, v \rangle = u^i v^j \langle e_i, e_j \rangle = g_{ij} u^i v^j \; .$$

If $\{\omega^i\}$ is the basis for U* dual to $\{e_i\}$, then according to our earlier formalism, we would write this as $\tilde{g} = g_{ij} \omega^i \otimes \omega^j$. ∕∕

Some of the main applications in mathematics and physics (e.g., in the theory of group representations) of the tensor product of two vector spaces are contained in the next two results. While the ideas are simple enough, the notation becomes somewhat awkward because of the double indices.

Theorem 11.11 Let U and V have the respective bases $\{u_1, \dots, u_m\}$ and $\{v_1, \dots, v_n\}$, and suppose the linear operators $S \in L(U)$ and $T \in L(V)$ have matrix representations $A = (a^i{}_j)$ and $B = (b^i{}_j)$ respectively. Then there exists a linear transformation $S \otimes T: U \otimes V \to U \otimes V$ such that for all $u \in U$ and $v \in V$ we have $(S \otimes T)(u \otimes v) = S(u) \otimes T(v)$.

Furthermore, the matrix C of $S \otimes T$ relative to the ordered basis

$$\{u_1 \otimes v_1, \dots, u_1 \otimes v_n, u_2 \otimes v_1, \dots, u_2 \otimes v_n, \dots, u_m \otimes v_1, \dots, u_m \otimes v_n\}$$

for $U \otimes V$ is the mn x mn matrix given in block matrix form as

$$C = \begin{pmatrix} a^1{}_1 B & a^1{}_2 B & \cdots & a^1{}_m B \\ \vdots & \vdots & & \vdots \\ a^m{}_1 B & a^m{}_2 B & \cdots & a^m{}_m B \end{pmatrix} .$$

The matrix C is called the **Kronecker** (or **direct** or **tensor**) **product** of the matrices A and B, and will also be written as $C = A \otimes B$.

Proof Since S and T are linear and $\otimes$ is bilinear, it is easy to see that the mapping $f: U \times V \to U \otimes V$ defined by $f(u, v) = S(u) \otimes T(v)$ is bilinear. Therefore, according to Theorem 11.10, there exists a unique linear transformation $\tilde{f} \in L(U \otimes V)$ such that $\tilde{f}(u \otimes v) = S(u) \otimes T(v)$. We denote the mapping $\tilde{f}$ by $S \otimes T$. Thus, $(S \otimes T)(u \otimes v) = S(u) \otimes T(v)$.

To find the matrix C of $S \otimes T$ is straightforward enough. We have $S(u_i) = u_j\, a^j{}_i$ and $T(v_i) = v_j b^j{}_i$, and hence

$$(S \otimes T)(u_i \otimes v_j) = S(u_i) \otimes T(v_j) = a^r{}_i b^s{}_j (u_r \otimes v_s) .$$

Now recall that the *ith* column of the matrix representation of an operator is just the image of the *ith* basis vector under the transformation (see Theorem 5.11). In the present case, we will have to use double pairs of subscripts to label the matrix elements. Relative to the ordered basis

$$\{u_1 \otimes v_1, \dots, u_1 \otimes v_n, u_2 \otimes v_1, \dots, u_2 \otimes v_n, \dots, u_m \otimes v_1, \dots, u_m \otimes v_n\}$$

for $U \otimes V$, we then see that, for example, the (1, 1)*th* column of C is the vector $(S \otimes T)(u_1 \otimes v_1) = a^r{}_1 b^s{}_1(u_r \otimes v_s)$ given by

$$(a^1{}_1 b^1{}_1, \dots, a^1{}_1 b^n{}_1, a^2{}_1 b^1{}_1, \dots, a^2{}_1 b^n{}_1, \dots, a^m{}_1 b^1{}_1, \dots, a^m{}_1 b^n{}_1)$$

and in general, the (i, j)*th* column is given by

$$(a^1{}_i b^1{}_j, \dots, a^1{}_i b^n{}_j, a^2{}_i b^1{}_j, \dots, a^2{}_i b^n{}_j, \dots, a^m{}_i b^1{}_j, \dots, a^m{}_i b^n{}_j) \ .$$

This shows that the matrix C has the desired form. ∎

Theorem 11.12 Let U and V be finite-dimensional vector spaces over $\mathcal{F}$.

(a) If $S_1, S_2 \in L(U)$ and $T_1, T_2 \in L(V)$, then

$$(S_1 \otimes T_1)(S_2 \otimes T_2) = S_1 S_2 \otimes T_1 T_2 \ .$$

Moreover, if A_i and B_i are the matrix representations of S_i and T_i respectively (relative to some basis for $U \otimes V$), then $(A_1 \otimes B_1)(A_2 \otimes B_2) = A_1 A_2 \otimes B_1 B_2$.

(b) If $S \in L(U)$ and $T \in L(V)$, then $\mathrm{Tr}(S \otimes T) = (\mathrm{Tr}\, S)(\mathrm{Tr}\, T)$.

(c) If $S \in L(U)$ and $T \in L(V)$, and if S^{-1} and T^{-1} exist, then

$$(S \otimes T)^{-1} = S^{-1} \otimes T^{-1} \ .$$

Conversely, if $(S \otimes T)^{-1}$ exists, then S^{-1} and T^{-1} also exist, and $(S \otimes T)^{-1} = S^{-1} \otimes T^{-1}$.

Proof (a) For any $u \in U$ and $v \in V$ we have

$$\begin{aligned}(S_1 \otimes T_1)(S_2 \otimes T_2)(u \otimes v) &= (S_1 \otimes T_1)(S_2(u) \otimes T_2(v)) \\ &= S_1 S_2(u) \otimes T_1 T_2(v) \\ &= (S_1 S_2 \otimes T_1 T_2)(u \otimes v) \ .\end{aligned}$$

As to the matrix representations, simply note that $A_i \otimes B_i$ is the representation of $S_i \otimes T_i$, and $A_1 A_2 \otimes B_1 B_2$ is the representation of $S_1 S_2 \otimes T_1 T_2$ (since the representation of a product of linear operators is the product of their matrices).

(b) Recall that the trace of a linear operator is defined to be the trace of any matrix representation of the operator (see Theorem 5.19). Therefore, if $A = (a^i{}_j)$ is the matrix of S and $B = (b^i{}_j)$ is the matrix of T, we see from Theorem 11.11 that the diagonal blocks of $A \otimes B$ are the matrices $a^1{}_1 B, \dots, a^m{}_m B$ and hence the diagonal elements of $A \otimes B$ are $a^1{}_1 b^1{}_1, \dots, a^1{}_1 b^n{}_n, \dots, a^m{}_m b^1{}_1, \dots, a^m{}_m b^n{}_n$. Therefore the sum of these diagonal elements is just

$$\begin{aligned}\mathrm{Tr}(A\otimes B) &= a^1{}_1(\Sigma_i b^i{}_i) + \cdots + a^m{}_m(\Sigma_i b^i{}_i)\\ &= (\Sigma_j a^j{}_j)(\Sigma_i b^i{}_i)\\ &= (\mathrm{Tr}\,A)(\mathrm{Tr}\,B)\enspace .\end{aligned}$$

(c) We first note that if 1 denotes the identity transformation, then

$$(1 \otimes 1)(\mathrm{u} \otimes \mathrm{v}) \;=\; \mathrm{u} \otimes \mathrm{v}$$

and hence $1 \otimes 1 = 1$. Next note that $\mathrm{u} \otimes \mathrm{v} = (\mathrm{u} + 0) \otimes \mathrm{v} = \mathrm{u} \otimes \mathrm{v} + 0 \otimes \mathrm{v}$, and hence $0 \otimes \mathrm{v} = 0$. Similarly, it is clear that $\mathrm{u} \otimes 0 = 0$. This then shows that

$$(\mathrm{S} \otimes 0)(\mathrm{u} \otimes \mathrm{v}) \;=\; \mathrm{S}(\mathrm{u}) \otimes 0 \;=\; 0$$

so that $\mathrm{S} \otimes 0 = 0$, and similarly $0 \otimes \mathrm{T} = 0$.
Now, if S and T are invertible, then by part (a) we see that

$$(\mathrm{S}^{-1} \otimes \mathrm{T}^{-1})(\mathrm{S} \otimes \mathrm{T}) \;=\; \mathrm{SS}^{-1} \otimes \mathrm{T}^{-1}\mathrm{T} \;=\; 1 \otimes 1 \;=\; 1$$

and similarly for $(\mathrm{S} \otimes \mathrm{T})(\mathrm{S}^{-1} \otimes \mathrm{T}^{-1})$. Therefore $(\mathrm{S} \otimes \mathrm{T})^{-1} = \mathrm{S}^{-1} \otimes \mathrm{T}^{-1}$.

Conversely, suppose that $\mathrm{S} \otimes \mathrm{T}$ is invertible. To prove that S and T are also invertible we use Theorem 5.9. In other words, a surjective linear operator is invertible if and only if its kernel is zero. Since $\mathrm{S} \otimes \mathrm{T}$ is invertible we must have $\mathrm{T} \neq 0$, and hence there exists $\mathrm{v} \in \mathrm{V}$ such that $\mathrm{T}(\mathrm{v}) \neq 0$. Suppose $\mathrm{u} \in \mathrm{U}$ is such that $\mathrm{S}(\mathrm{u}) = 0$. Then

$$0 \;=\; \mathrm{S}(\mathrm{u}) \otimes \mathrm{T}(\mathrm{v}) \;=\; (\mathrm{S} \otimes \mathrm{T})(\mathrm{u} \otimes \mathrm{v})$$

which implies that $\mathrm{u} \otimes \mathrm{v} = 0$ (since $\mathrm{S} \otimes \mathrm{T}$ is invertible). But $\mathrm{v} \neq 0$, and hence we must have $\mathrm{u} = 0$. This shows that S is invertible. Similarly, had we started with $\mathrm{S} \neq 0$, we would have found that T is invertible. ∎

Exercises

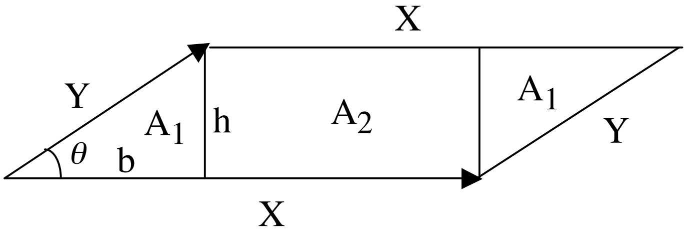
1. Give a direct proof of the matrix part of Theorem 11.12(a) using the definition of the Kronecker product of two matrices.

2. Suppose $\mathrm{A} \in L(\mathrm{U})$ and $\mathrm{B} \in L(\mathrm{V})$ where $\dim \mathrm{U} = \mathrm{n}$ and $\dim \mathrm{V} = \mathrm{m}$. Show that
$$\det(\mathrm{A} \otimes \mathrm{B}) \;=\; (\det \mathrm{A})^{\mathrm{m}}\,(\det \mathrm{B})^{\mathrm{n}}\enspace .$$

11.6 VOLUMES IN $\mathbb{R}^3$

Instead of starting out with an abstract presentation of volumes, we shall first go through an intuitive elementary discussion beginning with $\mathbb{R}^2$, then going to $\mathbb{R}^3$, and finally generalizing to $\mathbb{R}^n$ in the next section.

First consider a parallelogram in $\mathbb{R}^2$ (with the usual norm) defined by the vectors X and Y as shown.



Note that h = ‖Y‖ sin θ and b = ‖Y‖ cos θ, and also that the area of each triangle is given by $A_1 = (1/2)bh$. Then the area of the rectangle is given by $A_2 = (\|X\| - b)h$, and the area of the entire parallelogram is given by

$$A = 2A_1 + A_2 = bh + (\|X\| - b)h = \|X\|h = \|X\|\|Y\|\sin\theta \;. \tag{1}$$

The reader should recognize this as the magnitude of the elementary "vector cross product" X × Y of the ordered pair of vectors (X, Y) that is *defined* to have a direction normal to the plane spanned by X and Y, and given by the "right hand rule" (i.e., *out* of the plane in this case).

If we define the usual orthogonal coordinate system with the x-axis parallel to the vector X, then

$$X = (x^1, x^2) = (\|X\|, 0)$$

and

$$Y = (y^1, y^2) = (\|Y\|\cos\theta, \|Y\|\sin\theta)$$

and hence we see that the determinant with columns formed from the vectors X and Y is just

$$\begin{vmatrix} x^1 & y^1 \\ x^2 & y^2 \end{vmatrix} = \begin{vmatrix} \|X\| & \|Y\|\cos\theta \\ 0 & \|Y\|\sin\theta \end{vmatrix} = \|X\|\|Y\|\sin\theta = A \;. \tag{2}$$

Notice that if we interchanged the vectors X and Y in the diagram, then the determinant would change sign and the vector X × Y (which by definition has a direction dependent on the *ordered* pair (X, Y)) would point *into* the page.

Thus the area of a parallelogram (which is always positive by definition) defined by two vectors in $\mathbb{R}^2$ is in general given by the absolute value of the determinant (2).

In terms of the usual inner product (or "dot product") $\langle\ ,\ \rangle$ on $\mathbb{R}^2$, we have $\langle X, X\rangle = \|X\|^2$ and $\langle X, Y\rangle = \langle Y, X\rangle = \|X\|\|Y\| \cos\theta$, and hence

$$\begin{aligned} A^2 &= \|X\|^2\|Y\|^2 \sin^2\theta \\ &= \|X\|^2\|Y\|^2(1-\cos^2\theta) \\ &= \|X\|^2\|Y\|^2 - \langle X, Y\rangle^2 \ . \end{aligned}$$

Therefore we see that the area is also given by the positive square root of the determinant

$$A^2 = \begin{vmatrix} \langle X, X\rangle & \langle X, Y\rangle \\ \langle Y, X\rangle & \langle Y, Y\rangle \end{vmatrix} \ . \tag{3}$$

It is also worth noting that the inner product may be written in the form $\langle X, Y\rangle = x^1y^1 + x^2y^2$, and thus in terms of matrices we may write

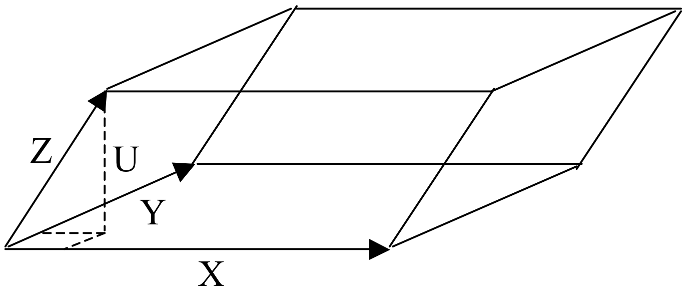
$$\begin{pmatrix} \langle X, X\rangle & \langle X, Y\rangle \\ \langle Y, X\rangle & \langle Y, Y\rangle \end{pmatrix} = \begin{pmatrix} x^1 & x^2 \\ y^1 & y^2 \end{pmatrix} \begin{pmatrix} x^1 & y^1 \\ x^2 & y^2 \end{pmatrix} \ .$$

Hence taking the determinant of this equation (using Theorems 4.8 and 4.1), we find (at least in $\mathbb{R}^2$) that the determinant (3) also implies that the area is given by the absolute value of the determinant in equation (2).

It is now easy to extend this discussion to a parallelogram in $\mathbb{R}^3$. Indeed, if $X = (x^1, x^2, x^3)$ and $Y = (y^1, y^2, y^3)$ are vectors in $\mathbb{R}^3$, then equation (1) is unchanged because any two vectors in $\mathbb{R}^3$ define the plane $\mathbb{R}^2$ spanned by the two vectors. Equation (3) also remains unchanged since its derivation did not depend on the specific coordinates of X and Y in $\mathbb{R}^2$. However, the left hand part of equation (2) does not apply (although we will see below that the three-dimensional version determines a volume in $\mathbb{R}^3$).

As a final remark on parallelograms, note that if X and Y are linearly dependent, then $aX + bY = 0$ so that $Y = -(a/b)X$, and hence X and Y are colinear. Therefore θ equals 0 or π so that all equations for the area in terms of sin θ are equal to zero. Since X and Y are dependent, this also means that the determinant in equation (2) equals zero, and everything is consistent.

We now take a look at volumes in $\mathbb{R}^3$. Consider three linearly independent vectors $X = (x^1, x^2, x^3)$, $Y = (y^1, y^2, y^3)$ and $Z = (z^1, z^2, z^3)$, and consider the parallelepiped with edges defined by these three vectors (in the given order (X, Y, Z)).



We claim that the volume of this parallelepiped is given by both the positive square root of the determinant

$$\begin{vmatrix} \langle X, X\rangle & \langle X, Y\rangle & \langle X, Z\rangle \\ \langle Y, X\rangle & \langle Y, Y\rangle & \langle Y, Z\rangle \\ \langle Z, X\rangle & \langle Z, Y\rangle & \langle Z, Z\rangle \end{vmatrix} \tag{4}$$

and the absolute value of the determinant

$$\begin{vmatrix} x^1 & y^1 & z^1 \\ x^2 & y^2 & z^2 \\ x^3 & y^3 & z^3 \end{vmatrix} . \tag{5}$$

To see this, first note that the volume of the parallelepiped is given by the product of the area of the base times the height, where the area A of the base is given by equation (3) and the height $\|U\|$ is just the projection of Z onto the orthogonal complement in $\mathbb{R}^3$ of the space spanned by X and Y. In other words, if W is the subspace of $V = \mathbb{R}^3$ spanned by X and Y, then (by Theorem 2.22) $V = W^\perp \oplus W$, and hence by Theorem 2.12 we may write

$$Z = U + aX + bY$$

where $U \in W^\perp$ and $a, b \in \mathbb{R}$ are uniquely determined (the uniqueness of a and b actually follows from Theorem 2.3 together with Theorem 2.12).

By definition we have $\langle X, U\rangle = \langle Y, U\rangle = 0$, and therefore

$$\begin{aligned}\langle X, Z\rangle &= a\|X\|^2 + b\langle X, Y\rangle\\ \langle Y, Z\rangle &= a\langle Y, X\rangle + b\|Y\|^2\\ \langle U, Z\rangle &= \|U\|^2\ .\end{aligned}\tag{6}$$

We now wish to solve the first two of these equations for a and b by Cramer's rule (Theorem 4.13). Note that the determinant of the matrix of coefficients is just equation (3), and hence is just the square of the area A of the base of the parallelepiped. Applying Cramer's rule we have

$$aA^2 = \begin{vmatrix}\langle X, Z\rangle & \langle X, Y\rangle\\ \langle Y, Z\rangle & \langle Y, Y\rangle\end{vmatrix} = -\begin{vmatrix}\langle X, Y\rangle & \langle X, Z\rangle\\ \langle Y, Y\rangle & \langle Y, Z\rangle\end{vmatrix}$$

$$bA^2 = \begin{vmatrix}\langle X, X\rangle & \langle X, Z\rangle\\ \langle Y, X\rangle & \langle Y, Z\rangle\end{vmatrix}\ .$$

Denoting the volume by Vol(X, Y, Z), we now have (using the last of equations (6) together with U = Z - aX - bY)

$$\mathrm{Vol}^2(\mathrm{X, Y, Z}) \ =\ \mathrm{A}^2\|\mathrm{U}\|^2 \ =\ \mathrm{A}^2\langle \mathrm{U, Z}\rangle \ =\ \mathrm{A}^2(\langle \mathrm{Z, Z}\rangle - \mathrm{a}\langle \mathrm{X, Z}\rangle - \mathrm{b}\langle \mathrm{Y, Z}\rangle)$$

so that substituting the expressions for A^2, aA^2 and bA^2, we find

$$\begin{aligned}\mathrm{Vol}^2(X, Y, Z) = \langle Z, Z\rangle\begin{vmatrix}\langle X, X\rangle & \langle X, Y\rangle\\ \langle Y, X\rangle & \langle Y, Y\rangle\end{vmatrix} + \langle X, Z\rangle\begin{vmatrix}\langle X, Y\rangle & \langle X, Z\rangle\\ \langle Y, Y\rangle & \langle Y, Z\rangle\end{vmatrix}\\ - \langle Y, Z\rangle\begin{vmatrix}\langle X, X\rangle & \langle X, Z\rangle\\ \langle Y, X\rangle & \langle Y, Z\rangle\end{vmatrix}\ .\end{aligned}$$

Using ⟨X, Y⟩ = ⟨Y, X⟩ etc., we see that this is just the expansion of a determinant by minors of the third row, and hence (using det A^T = det A)

$$\begin{aligned}\mathrm{Vol}^2(X, Y, Z) &= \begin{vmatrix}\langle X, X\rangle & \langle Y, X\rangle & \langle Z, X\rangle\\ \langle X, Y\rangle & \langle Y, Y\rangle & \langle Z, Y\rangle\\ \langle X, Z\rangle & \langle Y, Z\rangle & \langle Z, Z\rangle\end{vmatrix}\\ &= \begin{vmatrix}x^1 & x^2 & x^3\\ y^1 & y^2 & y^3\\ z^1 & z^2 & z^3\end{vmatrix}\begin{vmatrix}x^1 & y^1 & z^1\\ x^2 & y^2 & z^2\\ x^3 & y^3 & z^3\end{vmatrix} = \begin{vmatrix}x^1 & y^1 & z^1\\ x^2 & y^2 & z^2\\ x^3 & y^3 & z^3\end{vmatrix}^2\ .\end{aligned}$$

We remark that if the collection {X, Y, Z} is linearly dependent, then the volume of the parallelepiped degenerates to zero (since at least one of the parallelograms that form the sides will have zero area). This agrees with the fact that the determinant (5) will vanish if two rows are linearly dependent. We also note that the area of the base is given by

$$|X \times Y| = \|X\|\|Y\| \sin\measuredangle(X, Y)$$

where the direction of the vector X × Y is up (in this case). Therefore the projection of Z in the direction of X × Y is just Z dotted into a unit vector in the direction of X × Y, and hence the volume of the parallelepiped is given by the number Z • (X × Y). This is the so-called **scalar triple product** that should be familiar from elementary courses. We leave it to the reader to show that the scalar triple product is given by the determinant (5) (see Exercise 11.6.1).

Finally, note that if any two of the vectors X, Y, Z in equation (5) are interchanged, then the determinant changes sign even though the volume is unaffected (since it must be positive). This observation will form the basis for the concept of "orientation" to be defined later.

Exercises

1. Show that Z • (X × Y) is given by the determinant in equation (5).

2. Find the area of the parallelogram whose vertices are:
   (a) (0, 0), (1, 3), (−2, 1), and (−1, 4).
   (b) (2, 4), (4, 5), (5, 2), and (7, 3).
   (c) (−1, 3), (1, 5), (3, 2), and (5, 4).
   (d) (0, 0, 0), (1, −2, 2), (3, 4, 2), and (4, 2, 4).
   (e) (2, 2, 1), (3, 0, 6), (4, 1, 5), and (1, 1, 2).

3. Find the volume of the parallelepipeds whose adjacent edges are the vectors:
   (a) (1, 1, 2), (3, −1, 0), and 5, 2, −1).
   (b) (1, 1, 0), (1, 0, 1), and (0, 1, 1).

4. Prove both algebraically and geometrically that the parallelogram with edges X and Y has the same area as the parallelogram with edges X and Y + aX for any scalar a.

5. Prove both algebraically and geometrically that the volume of the parallelepiped in $\mathbb{R}^3$ with edges X, Y and Z is equal to the volume of the parallelepiped with edges X, Y and Z + aX + bY for any scalars a and b.

6. Show that the parallelepiped in $\mathbb{R}^3$ defined by the three vectors (2, 2, 1), (1, −2, 2) and (−2, 1, 2) is a cube. Find the volume of this cube.

11.7 VOLUMES IN $\mathbb{R}^n$

Now that we have a feeling for volumes in $\mathbb{R}^3$ expressed as determinants, let us prove the analogous results in $\mathbb{R}^n$. To begin with, we note that parallelograms defined by the vectors X and Y in either $\mathbb{R}^2$ or $\mathbb{R}^3$ contain all points (i.e., vectors) of the form aX + bY for any a, b ∈ [0, 1]. Similarly, given three linearly independent vectors X, Y, Z ∈ $\mathbb{R}^3$, we may define the parallelepiped with these vectors as edges to be that subset of $\mathbb{R}^3$ containing all vectors of the form aX + bY + cZ where $0 \le a, b, c \le 1$. The corners of the parallelepiped are the points $\delta_1X + \delta_2Y + \delta_3Z$ where each δ_i is either 0 or 1.

Generalizing these observations, given any r linearly independent vectors $X_1, \dots, X_r \in \mathbb{R}^n$, we define an **r-dimensional parallelepiped** as the set of all vectors of the form $a_1X_1 + \cdots + a_rX_r$ where $0 \le a_i \le 1$ for each i = 1, . . . , r. In $\mathbb{R}^3$, by a **1-volume** we mean a length, a **2-volume** means an area, and a **3-volume** is just the usual volume. To define the volume of an r-dimensional parallelepiped we proceed by induction on r. In particular, if X is a nonzero vector (i.e., a 1-dimensional parallelepiped) in $\mathbb{R}^n$, we define its 1-volume to be its length $\langle X, X\rangle^{1/2}$. Proceeding, suppose the (r − 1)-dimensional volume of an (r − 1)-dimensional parallelepiped has been defined. If we let P_r denote the r-dimensional parallelepiped defined by the r linearly independent vectors $X_1, \dots, X_r$, then we say that the **base** of P_r is the (r − 1)-dimensional parallelepiped defined by the r − 1 vectors $X_1, \dots, X_{r-1}$, and the **height** of P_r is the length of the projection of X_r onto the orthogonal complement in $\mathbb{R}^n$ of the space spanned by $X_1, \dots, X_{r-1}$. According to our induction hypothesis, the volume of an (r − 1)-dimensional parallelepiped has already been defined. Therefore we define the **r-volume** of P_r to be the product of its height times the (r − 1)-dimensional volume of its base.

The reader may wonder whether or not the r-volume of an r-dimensional parallelepiped in any way depends on which of the r vectors is singled out for projection. We proceed as if it does not and then, after the next theorem, we shall show that this is indeed the case.

Theorem 11.13 Let P_r be the r-dimensional parallelepiped defined by the r linearly independent vectors $X_1, \dots, X_r \in \mathbb{R}^n$. Then the r-volume of P_r is the positive square root of the determinant

$$\begin{vmatrix} \langle X_1, X_1\rangle & \langle X_1, X_2\rangle & \cdots & \langle X_1, X_r\rangle \\ \langle X_2, X_1\rangle & \langle X_2, X_2\rangle & \cdots & \langle X_2, X_r\rangle \\ \vdots & \vdots & & \vdots \\ \langle X_r, X_1\rangle & \langle X_r, X_2\rangle & \cdots & \langle X_r, X_r\rangle \end{vmatrix} . \tag{7}$$

Proof For the case of r = 1, we see that the theorem is true by the definition of length (or 1-volume) of a vector. Proceeding by induction, we assume the theorem is true for an (r − 1)-dimensional parallelepiped, and we show that it is also true for an r-dimensional parallelepiped. Hence, let us write

$$A^2 = \mathrm{Vol}^2(P_{r-1}) = \begin{vmatrix} \langle X_1, X_1\rangle & \langle X_1, X_2\rangle & \cdots & \langle X_1, X_{r-1}\rangle \\ \langle X_2, X_1\rangle & \langle X_2, X_2\rangle & \cdots & \langle X_2, X_{r-1}\rangle \\ \vdots & \vdots & & \vdots \\ \langle X_{r-1}, X_1\rangle & \langle X_{r-1}, X_2\rangle & \cdots & \langle X_{r-1}, X_{r-1}\rangle \end{vmatrix}$$

for the volume of the (r − 1)-dimensional base of P_r. Just as we did in our discussion of volumes in $\mathbb{R}^3$, we write X_r in terms of its projection U onto the orthogonal complement of the space spanned by the r − 1 vectors $X_1, \dots, X_r$. This means that we can write

$$X_r = U + a_1 X_1 + \cdots + a_{r-1}X_{r-1}$$

where $\langle U, X_i\rangle = 0$ for i = 1, . . . , r − 1, and $\langle U, X_r\rangle = \langle U, U\rangle$. We thus have the system of equations

$$\begin{aligned} a_1\langle X_1, X_1\rangle &+ a_2\langle X_1, X_2\rangle + \cdots + a_{r-1}\langle X_1, X_{r-1}\rangle = \langle X_1, X_r\rangle \\ a_1\langle X_2, X_1\rangle &+ a_2\langle X_2, X_2\rangle + \cdots + a_{r-1}\langle X_2, X_{r-1}\rangle = \langle X_2, X_r\rangle \\ &\vdots \\ a_1\langle X_{\mathrm{r}-1}, X_1\rangle &+ a_2\langle X_{\mathrm{r}-1}, X_2\rangle + \cdots + a_{r-1}\langle X_{\mathrm{r}-1}, X_{r-1}\rangle = \langle X_{\mathrm{r}-1}, X_r\rangle \end{aligned}$$

We write $M_1, \dots, M_{r-1}$ for the minors of the first r − 1 elements of the last row in (7). Solving the above system for the a_i using Cramer's rule, we obtain

$$\begin{aligned} A^2a_1 &= (-1)^{r-2}M_1 \\ A^2a_2 &= (-1)^{r-3}M_2 \\ &\vdots \\ A^2a_{r-1} &= M_{r-1} \end{aligned}$$

where the factors of $(-1)^{r-k-1}$ in A^2a_k result from moving the last column of (7) over to become the k*th* column of the k*th* minor matrix.

Using this result, we now have

$$\begin{aligned} A^2U &= A^2(-a_1X_1 - a_2X_2 - \cdots - a_{r-1}X_{r-1} + X_r) \\ &= (-1)^{r-1}M_1X_1 + (-1)^{r-2}M_2X_2 + \cdots + (-1)M_{r-1}X_{r-1} + A^2X_r \end{aligned}$$

and hence, using $\|U\|^2 = \langle U, U\rangle = \langle U, X_r\rangle$, we find that (since $(-1)^{-k} = (-1)^k$)

$$\begin{aligned} A^2\|U\|^2 &= A^2\langle U, X_r\rangle \\ &= (-1)^{r-1}M_1\langle X_r, X_1\rangle + (-1)(-1)^{r-1}M_2\langle X_r, X_2\rangle \\ &\qquad + \cdots + A^2\langle X_r, X_r\rangle \\ &= (-1)^{r-1}[M_1\langle X_r, X_1\rangle - M_2\langle X_r, X_2\rangle \\ &\qquad + \cdots + (-1)^{r-1}A^2\langle X_r, X_r\rangle] \ . \end{aligned}$$

Now note that the right hand side of this equation is precisely the expansion of (7) by minors of the last row, and the left hand side is by definition the square of the r-volume of the r-dimensional parallelepiped P_r. This also shows that the determinant (7) is positive. ∎

This result may also be expressed in terms of the matrix $(\langle X_i, X_j\rangle)$ as

$$\mathrm{Vol}(P_r) = [\det(\langle X_i, X_j\rangle)]^{1/2} \ .$$

The most useful form of this theorem is given in the following corollary.

Corollary The n-volume of the n-dimensional parallelepiped in ℝ^n defined by the vectors $X_1, \dots, X_n$ where each X_i has coordinates $(x^1{}_i, \dots, x^n{}_i)$ is the absolute value of the determinant of the matrix X given by

$$X=\begin{pmatrix} x^1{}_1 & x^1{}_2 & \cdots & x^1{}_n \\ x^2{}_1 & x^2{}_2 & \cdots & x^2{}_n \\ \vdots & \vdots & & \vdots \\ x^n{}_1 & x^n{}_2 & \cdots & x^n{}_n \end{pmatrix}.$$

Proof Note that $(\det X)^2 = (\det X)(\det X^T) = \det XX^T$ is just the determinant (7) in Theorem 11.13, which is the square of the volume. In other words, $\mathrm{Vol}(P_n) = |\det X|$. ∎

Prior to this theorem, we asked whether or not the r-volume depended on which of the r vectors is singled out for projection. We can now easily show that it does not. Suppose that we have an r-dimensional parallelepiped defined by r linearly independent vectors, and let us label these vectors $X_1, \ldots, X_r$. According to Theorem 11.13, we project X_r onto the space orthogonal to the space spanned by $X_1, \ldots, X_{r-1}$, and this leads to the determinant (7). If we wish to project any other vector instead, then we may simply relabel these r vectors to put a different one into position r. In other words, we have made some permutation of the indices in (7). However, remember that any permutation is a product of transpositions (Theorem 1.2), and hence we need only consider the effect of a single interchange of two indices.

Notice, for example, that the indices 1 and r only occur in rows 1 and r as well as in columns 1 and r. And in general, indices i and j only occur in the i*th* and j*th* rows and columns. But we also see that the matrix corresponding to (7) is symmetric about the main diagonal in these indices, and hence an interchange of the indices i and j has the effect of interchanging *both* rows i and j *as well as* columns i and j in exactly the same manner. Thus, because we have interchanged the same rows and columns there will be no sign change, and therefore the determinant (7) remains unchanged. In particular, it always remains positive. It now follows that the volume we have defined is indeed independent of which of the r vectors is singled out to be the height of the parallelepiped.

Now note that according to the above corollary, we know that $\mathrm{Vol}(P_n) = \mathrm{Vol}(X_1, \ldots, X_n) = |\det X|$ which is always positive. While our discussion just showed that $\mathrm{Vol}(X_1, \ldots, X_n)$ is independent of any permutation of indices, the actual value of det X can change sign upon any such permutation. Because of this, we say that the vectors $(X_1, \ldots, X_n)$ are **positively oriented** if $\det X > 0$, and **negatively oriented** if $\det X < 0$. Thus the orientation of a set of vectors depends on the order in which they are written. To take into account the sign of det X, we define the **oriented volume** $\mathrm{Vol}_o(X_1, \ldots, X_n)$ to be $+\mathrm{Vol}(X_1, \ldots, X_n)$ if $\det X \geq 0$, and $-\mathrm{Vol}(X_1, \ldots, X_n)$ if $\det X < 0$. We will return to a careful discussion of orientation in a later section. We also remark that det X is always nonzero as long as the vectors $(X_1, \dots, X_n)$ are linearly independent. Thus the above corollary may be expressed in the form

$$\mathrm{Vol}_o(X_1, \dots, X_n) = \det(X_1, \dots, X_n)$$

where $\det(X_1, \dots, X_n)$ means the determinant as a function of the column vectors X_i.

Exercises

1. Find the 3-volume of the three-dimensional parallelepipeds in $\mathbb{R}^4$ defined by the vectors:
   (a) (2, 1, 0, −1), (3, −1, 5, 2), and (0, 4, −1, 2).
   (b) (1, 1, 0, 0), (0, 2, 2, 0), and (0, 0, 3, 3).

2. Find the 2-volume of the parallelogram in $\mathbb{R}^4$ two of whose edges are the vectors (1, 3, −1, 6) and (−1, 2, 4, 3).

3. Prove that if the vectors $X_1, X_2, \dots, X_r$ are mutually orthogonal, the r-volume of the parallelepiped defined by them is equal to the product of their lengths.

4. Prove that r vectors $X_1, X_2, \dots, X_r$ in $\mathbb{R}^n$ are linearly dependent if and only if the determinant (7) is equal to zero.

11.8 LINEAR TRANSFORMATIONS AND VOLUMES

One of the most useful applications of Theorem 11.13 and its corollary relates to linear mappings. In fact, this is the approach usually followed in deriving the change of variables formula for multiple integrals. Let $\{e_i\}$ be an ortho–normal basis for $\mathbb{R}^n$, and let C_n denote the **unit cube** in $\mathbb{R}^n$. In other words,

$$C_n = \{t_1e_1 + \cdots + t_ne_n \in \mathbb{R}^n\colon 0 \le t_i \le 1\} .$$

This is similar to the definition of P_r given previously.

Now let $A\colon \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation. Then the matrix of A relative to the basis $\{e_i\}$ is defined by $A(e_i) = e_ja^j{}_i$. Let us write the image of e_i as X_i, so that $X_i = A(e_i) = e_ja^j{}_i$. This means that the column vector X_i has components $(a^1{}_i , \ldots , a^n{}_i)$. Under the transformation A, the image of C_n becomes

$$A(C_n) = A(\textstyle\sum t_i e_i) = \sum t_i A(e_i) = \sum t_i X_i$$

(where $0 \le t_i \le 1$) which is just the parallelepiped P_n spanned by the vectors $(X_1, \ldots , X_n)$. Therefore the volume of $P_n = A(C_n)$ is given by

$$|\det(X_1 , \ldots , X_n)| = |\det (a^j{}_i)| \; .$$

Recalling that the determinant of a linear transformation is defined to be the determinant of its matrix representation, we have proved the next result.

Theorem 11.14 Let C_n be the unit cube in $\mathbb{R}^n$ spanned by the orthonormal basis vectors $\{e_i\}$. If $A: \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation and $P_n = A(C_n)$, then $\mathrm{Vol}(P_n) = \mathrm{Vol}\, A(C_n) = |\det A|$.

It is quite simple to generalize this result somewhat to include the image of an n-dimensional parallelepiped under a linear transformation A. First, we note that any parallelepiped P_n is just the image of C_n under some linear transformation B. Indeed, if $P_n = \{t_1X_1 + \cdots + t_n X_n\colon 0 \le t_i \le 1\}$ for some set of vectors X_i, then we may define the transformation B by $B(e_i) = X_i$, and hence $P_n = B(C_n)$. Thus

$$A(P_n) = A(B(C_n)) = (A \circ B)(C_n)$$

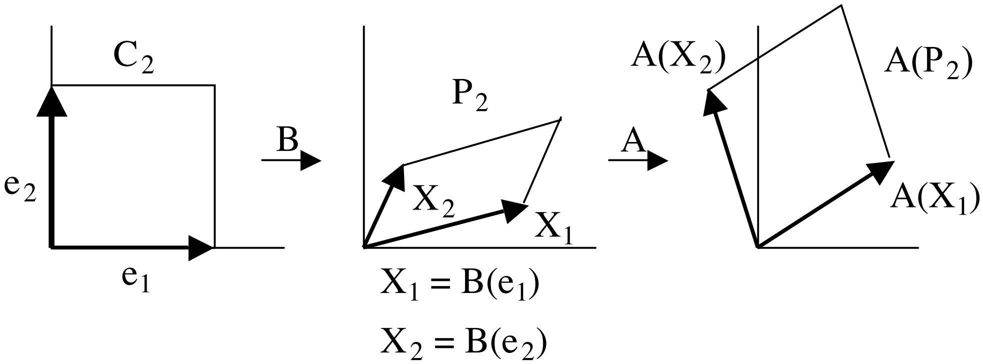
and therefore (using Theorem 11.14 along with the fact that the matrix of the composition of two transformations is the matrix product)

$$\begin{aligned}\mathrm{Vol}\, A(P_n) &= \mathrm{Vol}[(A \circ B)(C_n)] = |\det(A \circ B)| = |\det A||\det B| \\ &= |\det A|\mathrm{Vol}(P_n) \; .\end{aligned}$$

In other words, $|\det A|$ is a measure of how much the volume of the parallelepiped changes under the linear transformation A. See the figure below for a picture of this in $\mathbb{R}^2$.

We summarize this discussion as a corollary to Theorem 11.14.

Corollary Suppose P_n is an n-dimensional parallelepiped in $\mathbb{R}^n$, and let $A: \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation. Then $\mathrm{Vol}\, A(P_n) = |\det A|\mathrm{Vol}(P_n)$.



Now that we have an intuitive grasp of these concepts, let us look at this material from the point of view of exterior algebra. This more sophisticated approach is of great use in the theory of integration.

Let U and V be real vector spaces. Recall from Theorem 9.7 that given a linear transformation $T \in L(U, V)$, we defined the transpose mapping $T^* \in L(V^*, U^*)$ by

$$T^*\omega = \omega \circ T$$

for all $\omega \in V^*$. By this we mean if $u \in U$, then $T^*\omega(u) = \omega(Tu)$. As we then saw in Theorem 9.8, if U and V are finite-dimensional and A is the matrix representation of T, then A^T is the matrix representation of T^*, and hence certain properties of T^* follow naturally. For example, if $T_1 \in L(V, W)$ and $T_2 \in L(U, V)$, then $(T_1 \circ T_2)^* = T_2^* \circ T_1^*$ (Theorem 3.18), and if T is nonsingular, then $(T^{-1})^* = (T^*)^{-1}$ (Corollary 4 of Theorem 3.21).

Now suppose that $\{e_i\}$ is a basis for U and $\{f_j\}$ is a basis for V. To keep the notation simple and understandable, let us write the corresponding dual bases as $\{e^i\}$ and $\{f^j\}$. We define the matrix $A = (a^j{}_i)$ of T by $Te_i = f_j a^j{}_i$. Then (just as in the proof of Theorem 9.8)

$$\begin{aligned}(T^*f^i)e_j &= f^i(Te_j) = f^i(f_k a^k{}_j) = a^k{}_j f^i(f_k) = a^k{}_j \delta^i{}_k = a^i{}_j = a^i{}_k \delta^k{}_j \\ &= a^i{}_k e^k(e_j)\end{aligned}$$

which shows that

$$T^*f^i = a^i{}_k e^k\ . \tag{8}$$

We will use this result frequently below.

We now generalize our definition of the transpose. If $\phi \in L(U, V)$ and $T \in \mathcal{T}_r(V)$, we define the **pull-back** $\phi^* \in L(\mathcal{T}_r(V), \mathcal{T}_r(U))$ by

$$(\phi^*T)(u_1, \dots, u_r) = T(\phi(u_1), \dots, \phi(u_r))$$

where $u_1, \dots, u_r \in U$. Note that in the particular case of $r = 1$, the mapping ϕ^* is just the transpose of ϕ. It should also be clear from the definition that ϕ^* is indeed a linear transformation, and hence

$$\phi^*(aT_1 + bT_2) = a\phi^*T_1 + b\phi^*T_2 \ .$$

We also emphasize that ϕ need not be an isomorphism for us to define ϕ^*.

The main properties of the pull-back are given in the next theorem.

Theorem 11.15 If $\phi \in L(U, V)$ and $\psi \in L(V, W)$, then

(a) $(\psi \circ \phi)^* = \phi^* \circ \psi^*$.

(b) If $I \in L(U)$ is the identity map, then I^* is the identity in $L(\mathcal{T}_r(U))$.

(c) If ϕ is an isomorphism, then so is ϕ^*, and $(\phi^*)^{-1} = (\phi^{-1})^*$.

(d) If $T_1 \in \mathcal{T}_{r_1}(V)$ and $T_2 \in \mathcal{T}_{r_2}(V)$, then

$$\phi^*(T_1 \otimes T_2) = (\phi^*T_1) \otimes (\phi^*T_2) \ .$$

(e) Let U have basis $\{e_1, \dots, e_m\}$, V have basis $\{f_1, \dots, f_n\}$ and suppose that $\phi(e_i) = f_j a^j{}_i$. If $T \in \mathcal{T}_r(V)$ has components $T_{i_1 \cdots i_r} = T(f_{i_1}, \dots, f_{i_r})$, then the components of ϕ^*T relative to the basis $\{e_i\}$ are given by

$$(\phi^*T)_{j_1 \cdots j_r} = T_{i_1 \cdots i_r} a^{i_1}{}_{j_1} \cdots a^{i_r}{}_{j_r} \ .$$

Proof (a) Note that $\psi \circ \phi: U \to W$, and hence $(\psi \circ \phi)^*: \mathcal{T}_r(W) \to \mathcal{T}_r(U)$. Thus for any $T \in \mathcal{T}_r(W)$ and $u_1, \dots, u_r \in U$ we have

$$\begin{aligned}((\psi \circ \phi)^*T)(u_1, \dots, u_r) &= T(\psi(\phi(u_1)), \dots, \psi(\phi(u_r))) \\ &= (\psi^*T)(\phi(u_1), \dots, \phi(u_r)) \\ &= ((\phi^* \circ \psi^*)T)(u_1, \dots, u_r) \ .\end{aligned}$$

(b) Obvious from the definition of I^*.

(c) If ϕ is an isomorphism, then ϕ^{-1} exists and we have (using (a) and (b))

$$\phi^* \circ (\phi^{-1})^* = (\phi^{-1} \circ \phi)^* = I^* \ .$$

Similarly $(\phi^{-1})^* \circ \phi^* = I^*$. Hence $(\phi^*)^{-1}$ exists and is equal to $(\phi^{-1})^*$.

(d) This follows directly from the definitions (see Exercise 11.8.1).

(e) Using the definitions, we have

$$\begin{aligned}(\phi^*T)_{j_1\cdots j_r} &= (\phi^*T)(e_{j_1},\dots,e_{j_r})\\ &= T(\phi(e_{j_1}),\dots,\phi(e_{j_r}))\\ &= T(f_{i_1}a^{i_1}{}_{j_1},\dots,f_{i_r}a^{i_r}{}_{j_r})\\ &= T(f_{i_1},\dots,f_{i_r})a^{i_1}{}_{j_1}\cdots a^{i_r}{}_{j_r}\\ &= T_{i_1\cdots i_r}a^{i_1}{}_{j_1}\cdots a^{i_r}{}_{j_r}\quad .\end{aligned}$$

Alternatively, if $\{e^i\}$and $\{f^j\}$ are the bases dual to $\{e_i\}$ and $\{f_j\}$ respectively, then $T = T_{i_1\cdots i_r}e^{i_1}\otimes\cdots\otimes e^{i_r}$ and consequently (using the linearity of ϕ^*, part (d) and equation (8)),

$$\begin{aligned}\phi^*T &= T_{i_1\cdots i_r}\phi^*e^{i_1}\otimes\cdots\otimes\phi^*e^{i_r}\\ &= T_{i_1\cdots i_r}a^{i_1}{}_{j_1}\cdots a^{i_r}{}_{j_r}f^{j_1}\otimes\cdots\otimes f^{j_r}\end{aligned}$$

which therefore yields the same result. ∎

For our present purposes, we will only need to consider the pull-back as defined on the space $\bigwedge^r(V)$ rather than on $\mathcal{T}_r(V)$. Therefore, if $\phi \in L(U, V)$ then $\phi^* \in L(\mathcal{T}_r(V), \mathcal{T}_r(U))$, and hence we see that for $\omega \in \bigwedge^r(V)$ we have $(\phi^*\omega)(u_1, \dots, u_r) = \omega(\phi(u_1), \dots, \phi(u_r))$. This shows that $\phi^*(\bigwedge^r(V)) \subset \bigwedge^r(U)$. Parts (d) and (e) of Theorem 11.15 applied to the space $\bigwedge^r(V)$ yield the following special cases. (Recall that $|i_1 \cdots i_r|$ means the sum is over increasing indices $i_1 < \cdots < i_r$.)

Theorem 11.16 Suppose $\phi \in L(U, V)$, $\alpha \in \bigwedge^r(V)$ and $\beta \in \bigwedge^s(V)$. Then

(a) $\phi^*(\alpha\wedge\beta) = (\phi^*\alpha)\wedge(\phi^*\beta)$.

(b) Let U and V have bases $\{e_i\}$ and $\{f_i\}$ respectively, and let U^* and V^* have bases $\{e^i\}$ and $\{f^i\}$. If we write $\phi(e_i) = f_j a^j{}_i$ and $\phi^*(f^i) = a^i{}_j e^j$, and if $\alpha = a_{|i_1\cdots i_r|}f^{i_1}\wedge\cdots\wedge f^{i_r} \in \bigwedge^r(V)$, then

$$\phi^*\alpha = \hat{a}_{|k_1\cdots k_r|}e^{k_1}\wedge\cdots\wedge e^{k_r}$$

where

$$\hat{a}_{|k_1\cdots k_r|} = a_{|i_1\cdots i_r|}\varepsilon^{j_1\cdots j_r}_{k_1\cdots k_r}a^{i_1}{}_{j_1}\cdots a^{i_r}{}_{j_r}\quad .$$

Thus we may write

$$\hat{a}_{k_1\cdots k_r} = a_{|i_1\cdots i_r|}\det(a^I{}_K)$$

where

$$\det(a^I{}_K) = \begin{vmatrix} a^{i_1}{}_{k_1} & \cdots & a^{i_1}{}_{k_r} \\ \vdots & & \vdots \\ a^{i_r}{}_{k_1} & \cdots & a^{i_r}{}_{k_r} \end{vmatrix} \;.$$

Proof (a) For simplicity, let us write $(\phi^*\alpha)(u_J) = \alpha(\phi(u_J))$ instead of $(\phi^*\alpha)(u_1, \ldots, u_r) = \alpha(\phi(u_1), \ldots, \phi(u_r))$ (see the discussion following Theorem 11.4). Then, in an obvious notation, we have

$$\begin{aligned} [\phi^*(\alpha \wedge \beta)](u_I) &= (\alpha \wedge \beta)(\phi(u_I)) \\ &= \Sigma_{\underrightarrow{J},\underrightarrow{K}} \varepsilon_I^{JK} (\alpha(\phi(u_J))\beta(\phi(u_K)) \\ &= \Sigma_{\underrightarrow{J},\underrightarrow{K}} \varepsilon_I^{JK} (\phi^*\alpha)(u_J)(\phi^*\beta)(u_K) \\ &= [(\phi^*\alpha) \wedge (\phi^*\beta)](u_I) \;. \end{aligned}$$

By induction, this also obviously applies to the wedge product of a finite number of forms.

(b) From $\alpha = a_{|i_1 \cdots i_r|} f^{i_1} \wedge \cdots \wedge f^{i_r}$ and $\phi^*(f^i) = a^i{}_j e^j$, we have (using part (a) and the linearity of ϕ^*)

$$\begin{aligned} \phi^*\alpha &= a_{|i_1 \cdots i_r|} \phi^*(f^{i_1}) \wedge \cdots \wedge \phi^*(f^{i_r}) \\ &= a_{|i_1 \cdots i_r|} a^{i_1}{}_{j_1} \cdots a^{i_r}{}_{j_r} e^{j_1} \wedge \cdots \wedge e^{j_r} \;. \end{aligned}$$

But

$$e^{j_1} \wedge \cdots \wedge e^{j_r} = \Sigma_{\underrightarrow{K}} \varepsilon^{j_1 \cdots j_r}_{k_1 \cdots k_r} e^{k_1} \wedge \cdots \wedge e^{k_r}$$

and hence we have

$$\begin{aligned} \phi^*\alpha &= a_{|i_1 \cdots i_r|} \Sigma_{\underrightarrow{K}} \varepsilon^{j_1 \cdots j_r}_{k_1 \cdots k_r} a^{i_1}{}_{j_1} \cdots a^{i_r}{}_{j_r} e^{k_1} \wedge \cdots \wedge e^{k_r} \\ &= \hat{a}_{|k_1 \cdots k_r|} e^{k_1} \wedge \cdots \wedge e^{k_r} \end{aligned}$$

where

$$\hat{a}_{|k_1 \cdots k_r|} = a_{|i_1 \cdots i_r|} \varepsilon^{j_1 \cdots j_r}_{k_1 \cdots k_r} a^{i_1}{}_{j_1} \cdots a^{i_r}{}_{j_r} \;.$$

Finally, from the definition of determinant we see that

$$\varepsilon^{j_1 \cdots j_r}_{k_1 \cdots k_r} a^{i_1}{}_{j_1} \cdots a^{i_r}{}_{j_r} = \begin{vmatrix} a^{i_1}{}_{k_1} & \cdots & a^{i_1}{}_{k_r} \\ \vdots & & \vdots \\ a^{i_r}{}_{k_1} & \cdots & a^{i_r}{}_{k_r} \end{vmatrix} \;. \;\blacksquare$$

Example 11.11 (This is a continuation of Example 11.1.) An important example of $\phi^*\alpha$ is related to the change of variables formula in multiple integrals. While we are not in any position to present this material in detail, the idea is this. Suppose we consider the spaces U = $\mathbb{R}^3$(u, v, w) and V = $\mathbb{R}^3$(x, y, z) where the letters in parentheses tell us the coordinate system used for that particular copy of $\mathbb{R}^3$. Note that if we write (x, y, z) = (x^1, x^2, x^3) and (u, v, w) = (u^1, u^2, u^3), then from elementary calculus we know that $dx^i = (\partial x^i/\partial u^j)du^j$ and $\partial/\partial u^i = (\partial x^j/\partial u^i)(\partial/\partial x^j)$.

Now recall from Example 11.1 that at each point of $\mathbb{R}^3$(u, v, w), the tangent space has the basis $\{e_i\} = \{\partial/\partial u^i\}$ and the cotangent space has the corresponding dual basis $\{e^i\} = \{du^i\}$, with a similar result for $\mathbb{R}^3$(x, y, z). Let us define ϕ: $\mathbb{R}^3$(u, v, w) → $\mathbb{R}^3$(x, y, z) by

$$\phi(\partial/\partial u^i) = (\partial x^j/\partial u^i)(\partial/\partial x^j) = a^j{}_i(\partial/\partial x^j) \; .$$

It is then apparent that (see equation (8))

$$\phi^*(dx^i) = a^i{}_j du^j = (\partial x^i/\partial u^j)du^j$$

as we should have expected. We now apply this to the 3-form

$$\alpha = a_{123}\, dx^1 \wedge dx^2 \wedge dx^3 = dx \wedge dy \wedge dz \in \textstyle\bigwedge^3(V) \; .$$

Since we are dealing with a 3-form in a 3-dimensional space, we must have

$$\phi^*\alpha = \hat{a}\, du \wedge dv \wedge dw$$

where $\hat{a} = \hat{a}_{123}$ consists of the single term given by the determinant

$$\begin{vmatrix} a^1{}_1 & a^1{}_2 & a^1{}_3 \\ a^2{}_1 & a^2{}_2 & a^2{}_3 \\ a^3{}_1 & a^3{}_2 & a^3{}_3 \end{vmatrix} = \begin{vmatrix} \partial x^1/\partial u^1 & \partial x^1/\partial u^2 & \partial x^1/\partial u^3 \\ \partial x^2/\partial u^1 & \partial x^2/\partial u^2 & \partial x^2/\partial u^3 \\ \partial x^3/\partial u^1 & \partial x^3/\partial u^2 & \partial x^3/\partial u^3 \end{vmatrix}$$

which the reader may recognize as the so-called Jacobian of the transformation. This determinant is usually written as $\partial(x, y, z)/\partial(u, v, w)$, and hence we see that

$$\phi^*(dx \wedge dy \wedge dz) = \frac{\partial(x, y, z)}{\partial(u, v, w)} du \wedge dv \wedge dw \; .$$

This is precisely how volume elements transform (at least locally), and hence we have formulated the change of variables formula in quite general terms. //

This formalism allows us to define the determinant of a linear transformation in an interesting abstract manner. To see this, suppose $\phi \in L(V)$ where $\dim V = n$. Since $\dim \bigwedge^n(V) = 1$, we may choose any nonzero $\omega_0 \in \bigwedge^n(V)$ as a basis. Then $\phi^*: \bigwedge^n(V) \to \bigwedge^n(V)$ is linear, and hence for any $\omega = c_0\omega_0 \in \bigwedge^n(V)$ we have

$$\phi^*\omega = \phi^*(c_0\omega_0) = c_0\phi^*\omega_0 = c_0c\omega_0 = c(c_0\omega_0) = c\omega$$

for some scalar c (since $\phi^*\omega_0 \in \bigwedge^n(V)$ is necessarily of the form $c\omega_0$). Noting that this result did not depend on the scalar c_0 and hence is independent of $\omega = c_0\omega_0$, we see that the scalar c must be unique. We therefore define the **determinant** of ϕ to be the unique scalar, denoted by $\det \phi$, such that

$$\phi^*\omega = (\det \phi)\omega \ .$$

It is important to realize that this definition of the determinant does not depend on any choice of basis for V. However, let $\{e_i\}$ be a basis for V, and define the matrix $(a^i{}_j)$ of ϕ by $\phi(e_i) = e_ja^j{}_i$. Then for any nonzero $\omega \in \bigwedge^n(V)$ we have

$$(\phi^*\omega)(e_1, \dots, e_n) = (\det \phi)\omega(e_1, \dots, e_n) \ .$$

On the other hand, Example 11.2 shows us that

$$\begin{aligned}(\phi^*\omega)(e_1, \dots, e_n) &= \omega(\phi(e_1), \dots, \phi(e_n))\\ &= a^{i_1}{}_1 \cdots a^{i_n}{}_n\omega(e_{i_1}, \dots, e_{i_n})\\ &= (\det(a^i{}_j))\omega(e_1, \dots, e_n) \ .\end{aligned}$$

Since $\omega \neq 0$, we have therefore proved the next result.

Theorem 11.17 If V has basis $\{e_1, \dots, e_n\}$ and $\phi \in L(V)$ has the matrix representation $(a^i{}_j)$ defined by $\phi(e_i) = e_ja^j{}_i$, then $\det \phi = \det(a^i{}_j)$.

In other words, our abstract definition of the determinant is exactly the same as our earlier classical definition. In fact, it is now easy to derive some of the properties of the determinant that were not exactly simple to prove in the more traditional manner.

Theorem 11.18 If V is finite-dimensional and $\phi, \psi \in L(V, V)$, then

(a) $\det(\phi \circ \psi) = (\det \phi)(\det \psi)$.

(b) If ϕ is the identity transformation, then $\det \phi = 1$.

(c) ϕ is an isomorphism if and only if $\det \phi \neq 0$, and if this is the case, then $\det \phi^{-1} = (\det \phi)^{-1}$.

Proof (a) By definition we have $(\phi \circ \psi)^*\omega = \det(\phi \circ \psi)\omega$. On the other hand, by Theorem 11.15(a) we know that $(\phi \circ \psi)^* = \psi^* \circ \phi^*$, and hence

$$\begin{aligned}(\phi \circ \psi)^*\omega &= \psi^*(\phi^*\omega) = \psi^*[(\det\phi)\omega] = (\det\phi)\psi^*\omega \\ &= (\det\phi)(\det\psi)\omega\ \ .\end{aligned}$$

(b) If $\phi = 1$ then $\phi^* = 1$ also (by Theorem 11.15(b)), and thus $\omega = \phi^*\omega = (\det \phi)\omega$ implies $\det \phi = 1$.

(c) First assume that ϕ is an isomorphism so that ϕ^{-1} exists. Then by parts (a) and (b) we see that

$$1 = \det(\phi\phi^{-1}) = (\det \phi)(\det \phi^{-1})$$

which implies $\det \phi \neq 0$ and $\det \phi^{-1} = (\det \phi)^{-1}$. Conversely, suppose that ϕ is not an isomorphism. Then $\operatorname{Ker} \phi \neq 0$ and there exists a nonzero $e_1 \in V$ such that $\phi(e_1) = 0$. By Theorem 2.10, we can extend this to a basis $\{e_1, \dots, e_n\}$ for V. But then for any nonzero $\omega \in \bigwedge^n(V)$ we have

$$\begin{aligned}(\det\phi)\omega(e_1, \dots, e_n) &= (\phi^*\omega)(e_1, \dots, e_n) \\ &= \omega(\phi(e_1), \dots, \phi(e_n)) \\ &= \omega(0, \phi(e_2), \dots, \phi(e_n)) \\ &= 0\end{aligned}$$

and hence we must have $\det \phi = 0$. ∎

Exercises

1. Prove Theorem 11.15(d).

2. Show that the matrix ϕ^*T defined in Theorem 11.15(e) is just the r-fold Kronecker product $A \otimes \cdots \otimes A$ where $A = (a^i{}_j)$.

The next three exercises are related.

3. Let $\phi \in L(U, V)$ be an isomorphism, and suppose $T \in \mathcal{T}_r{}^s(U)$. Define the **push-forward** $\phi_* \in L(\mathcal{T}_r^s(U), \mathcal{T}_r^s(V))$ by

$$\phi_*T(\alpha^1, \dots, \alpha^s, u_1, \dots, u_r) = T(\phi^*\alpha^1, \dots, \phi^*\alpha^s, \phi^{-1}u_1, \dots, \phi^{-1}u_r)$$

where $\alpha^1, \dots, \alpha^s \in U^*$ and $u_1, \dots, u_r \in U$. If $\psi \in L(V, W)$ is also an isomorphism, prove the following:
(a) $(\psi \circ \phi)_* = \psi_* \circ \phi_*$.
(b) If $I \in L(U)$ is the identity map, then so is $I_* \in L(\mathcal{T}_r^s(U))$.
(c) ϕ_* is an isomorphism, and $(\phi_*)^{-1} = (\phi^{-1})_*$.
(d) If $T_1 \in \mathcal{T}_{r_1}{}^{s_1}(U)$ and $T_2 \in \mathcal{T}_{r_2}{}^{s_2}(U)$, then

$$\phi_*(T_1 \otimes T_2) = (\phi_*T_1) \otimes (\phi_*T_2) \ .$$

4. Let $\phi \in L(U, V)$ be an isomorphism, and let U and V have bases $\{e_i\}$ and $\{f_i\}$ respectively. Define the matrices $(a^i{}_j)$ and $(b^i{}_j)$ by $\phi(e_i) = f_j a^j{}_i$ and $\phi^{-1}(f_i) = e_j b^j{}_i$. Suppose $T \in \mathcal{T}_r{}^s(U)$ has components $T^{i_1 \cdots i_s}{}_{j_1 \cdots j_r}$ relative to $\{e_i\}$, and $S \in \mathcal{T}_r{}^s(V)$ has components $S^{i_1 \cdots i_s}{}_{j_1 \cdots j_r}$ relative to $\{f_i\}$. Show that the components of ϕ_*T and ϕ_*S are given by

$$\begin{aligned}(\phi_*T)^{i_1 \cdots i_s}{}_{j_1 \cdots j_r} &= a^{i_1}{}_{p_1} \cdots a^{i_s}{}_{p_s} T^{p_1 \cdots p_s}{}_{q_1 \cdots q_r} b^{q_1}{}_{j_1} \cdots b^{q_r}{}_{j_r} \\ (\phi_*S)^{i_1 \cdots i_s}{}_{j_1 \cdots j_r} &= b^{i_1}{}_{p_1} \cdots b^{i_s}{}_{p_s} S^{p_1 \cdots p_s}{}_{q_1 \cdots q_r} a^{q_1}{}_{j_1} \cdots a^{q_r}{}_{j_r} \ .\end{aligned}$$

5. Let $\{\omega^i\}$ be the basis dual to $\{e_i\}$ for $\mathbb{R}^2$. Let

$$T = 2e_1 \otimes \omega^1 - e_2 \otimes \omega^1 + 3e_1 \otimes \omega^2$$

and suppose $\phi \in L(\mathbb{R}^2)$ and $\psi \in L(\mathbb{R}^3, \mathbb{R}^2)$ have the matrix representations

$$\phi = \begin{pmatrix} 2 & 1 \\ -1 & 1 \end{pmatrix} \qquad \text{and} \qquad \psi = \begin{pmatrix} 0 & 1 & -1 \\ 1 & 0 & 2 \end{pmatrix} \ .$$

Compute Tr T, ϕ^*T, ψ^*T, $\mathrm{Tr}(\psi^*T)$, and ϕ_*T.

11.9 ORIENTATIONS AND VOLUMES

Suppose dim V = n and consider the space $\bigwedge^n(V)$. Since this space is 1-dimensional, we consider the n-form

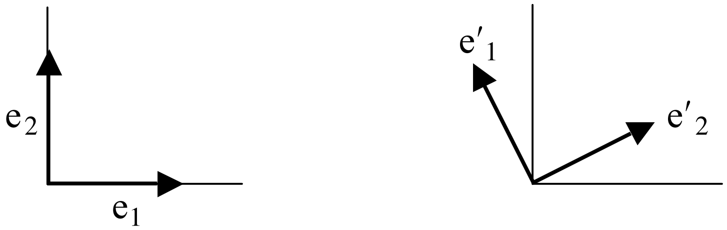
$$\omega = e^1 \wedge \cdots \wedge e^n \in \bigwedge^n(V)$$

where the basis $\{e^i\}$ for V* is dual to the basis $\{e_i\}$ for V. If $\{v_i = e_j v^j{}_i\}$ is any set of n linearly independent vectors in V then, according to Examples 11.2 and 11.8, we have

$$\omega(v_1, \dots, v_n) = \det(v^j{}_i)\omega(e_1, \dots, e_n) = \det(v^j{}_i) \;.$$

However, from the corollary to Theorem 11.13, this is just the oriented n-volume of the n-dimensional parallelepiped in $\mathbb{R}^n$ spanned by the vectors $\{v_i\}$. Therefore, we see that an n-form in some sense represents volumes in an n-dimensional space. We now proceed to make this definition precise, beginning with a careful definition of the notion of orientation on a vector space.

In order to try and make the basic idea clear, let us first consider the space $\mathbb{R}^2$ with all possible orthogonal coordinate systems. For example, we may consider the usual "right-handed" coordinate system $\{e_1, e_2\}$ shown below, or we may consider the alternative "left-handed" system $\{e'_1, e'_2\}$ also shown.



In the first case, we see that rotating e_1 into e_2 through the smallest angle between them involves a counterclockwise rotation, while in the second case, rotating e'_1 into e'_2 entails a clockwise rotation. This effect is shown in the elementary vector cross product, where the direction of $e_1 \times e_2$ is defined by the "right-hand rule" to point out of the page, while $e'_1 \times e'_2$ points into the page.

We now ask whether or not it is possible to continuously rotate e'_1 into e_1 and e'_2 into e_2 while maintaining a basis at all times. In other words, we ask if these two bases are in some sense equivalent. Without being rigorous, it should be clear that this can not be done because there will always be one point where the vectors e'_1 and e'_2 will be co-linear, and hence linearly dependent. This observation suggests that we consider the determinant of the matrix representing this change of basis.

In order to formulate this idea precisely, let us take a look at the matrix relating our two bases $\{e_i\}$ and $\{e'_i\}$ for $\mathbb{R}^2$. We thus write $e'_i = e_j a^j{}_i$ and investigate the determinant $\det(a^i{}_j)$. From the above figure, we see that

$$\begin{aligned} e'_1 &= e_1 a^1{}_1 + e_2 a^2{}_1 \qquad \text{where } a^1{}_1 < 0 \text{ and } a^2{}_1 > 0 \\ e'_2 &= e_1 a^1{}_2 + e_2 a^2{}_2 \qquad \text{where } a^1{}_2 < 0 \text{ and } a^2{}_2 > 0 \end{aligned}$$

and hence $\det(a^i{}_j) = a^1{}_1\, a^2{}_2 - a^1{}_2\, a^2{}_1 < 0$.

Now suppose that we view this transformation as a continuous modification of the identity transformation. This means we consider the basis vectors e'_i to be continuous functions $e'_i(t)$ of the matrix $a^j{}_i(t)$ for $0 \le t \le 1$ where $a^j{}_i(0) = \delta^j{}_i$ and $a^j{}_i(1) = a^j{}_i$, so that $e'_i(0) = e_i$ and $e'_i(1) = e'_i$. In other words, we write $e'_i(t) = e_j a^j{}_i(t)$ for $0 \le t \le 1$. Now note that $\det(a^i{}_j(0)) = \det(\delta^i{}_j) = 1 > 0$, while $\det(a^i{}_j(1)) = \det(a^i{}_j) < 0$. Therefore, since the determinant is a continuous function of its entries, there must be some value $t_0 \in (0, 1)$ where $\det(a^i{}_j(t_0)) = 0$. It then follows that the vectors $e'_i(t_0)$ will be linearly dependent.

What we have just shown is that if we start with any pair of linearly independent vectors, and then transform this pair into another pair of linearly independent vectors by moving along any continuous path of linear transformations that always maintains the linear independence of the pair, then every linear transformation along this path must have positive determinant. Another way of saying this is that if we have two bases that are related by a transformation with negative determinant, then it is impossible to continuously transform one into the other while maintaining their independence. This argument clearly applies to $\mathbb{R}^n$ and is not restricted to $\mathbb{R}^2$.

Conversely, suppose we had assumed that $e'_i = e_j a^j{}_i$, but this time with $\det(a^i{}_j) > 0$. We want to show that $\{e_i\}$ may be continuously transformed into $\{e'_i\}$ while maintaining linear independence all the way. We first assume that both $\{e_i\}$ and $\{e'_i\}$ are orthonormal bases. After treating this special case, we will show how to take care of arbitrary bases.

(Unfortunately, the argument we are about to give relies on the topological concept of path connectedness. Since a complete discussion of this topic would lead us much too far astray, we shall be content to present only the fundamental concepts in Appendix C. Besides, this discussion is only motivation, and the reader should not get too bogged down in the details of this argument. Those readers who know some topology should have no trouble filling in the necessary details if desired.)

Since $\{e_i\}$ and $\{e'_i\}$ are orthonormal, it follows from Theorem 10.6 (applied to $\mathbb{R}$ rather than $\mathbb{C}$) that the transformation matrix $A = (a^i{}_j)$ defined by $e'_i = e_j a^j{}_i$ must be orthogonal, and hence $\det A = +1$ (by Theorem 10.8(a) and the fact that we are assuming $\{e_i\}$ and $\{e'_i\}$ are related by a transformation with positive determinant). By Theorem 10.19, there exists a nonsingular matrix S such that $S^{-1}AS = M_\theta$ where M_θ is the block diagonal canonical form consisting of +1's, −1's, and 2 x 2 rotation matrices $R(\theta_i)$ given by

$$R(\theta_i) = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix} .$$

It is important to realize that if there are more than two +1's or more than two −1's, then each pair may be combined into one of the $R(\theta_i)$ by choosing either $\theta_i = \pi$ (for each pair of −1's) or $\theta_i = 0$ (for each pair of +1's). In this manner, we view M_θ as consisting entirely of 2 x 2 rotation matrices, and at most a single +1 and/or −1. Since $\det R(\theta_i) = +1$ for any θ_i, we see that (using Theorem 4.14) $\det M_\theta = +1$ if there is no −1, and $\det M_\theta = -1$ if there is a single −1. From $A = SM_\theta S^{-1}$, we see that $\det A = \det M_\theta$, and since we are requiring that $\det A > 0$, we must have the case where there is no −1 in M_θ.

Since $\cos\theta_i$ and $\sin\theta_i$ are continuous functions of $\theta_i \in [0, 2\pi)$ (where the interval $[0, 2\pi)$ is a path connected set), we note that by parametrizing each θ_i by $\theta_i(t) = (1 - t)\theta_i$, the matrix M_θ may be continuously connected to the identity matrix I (i.e., at t = 1). In other words, we consider the matrix $M_{\theta(t)}$ where $M_{\theta(0)} = M_\theta$ and $M_{\theta(1)} = I$. Hence every such M_θ (i.e., any matrix of the same form as our particular M_θ, but with a different set of θ_i's) may be continuously connected to the identity matrix. (For those readers who know some topology, note all we have said is that the torus $[0, 2\pi) \times \cdots \times [0, 2\pi)$ is path connected, and hence so is its continuous image which is the set of all such M_θ.)

We may write the (infinite) collection of all such M_θ as $M = \{M_\theta\}$. Clearly M is a path connected set. Since $A = SM_\theta S^{-1}$ and $I = SIS^{-1}$, we see that both A and I are contained in the collection $SMS^{-1} = \{SM_\theta S^{-1}\}$. But SMS^{-1} is also path connected since it is just the continuous image of a path connected set (matrix multiplication is obviously continuous). Thus we have shown that both A and I lie in the path connected set SMS^{-1}, and hence A may be continuously connected to I. Note also that every transformation along this path has positive determinant since $\det SM_\theta S^{-1} = \det M_\theta = 1 > 0$ for every $M_\theta \in M$.

If we now take any path in SMS^{-1} that starts at I and goes to A, then applying this path to the basis $\{e_i\}$ we obtain a continuous transformation from $\{e_i\}$ to $\{e'_i\}$ with everywhere positive determinant. This completes the proof for the special case of orthonormal bases.

Now suppose that $\{v_i\}$ and $\{v'_i\}$ are arbitrary bases related by a transformation with positive determinant. Starting with the basis $\{v_i\}$, we first apply the Gram-Schmidt process (Theorem 2.21) to $\{v_i\}$ to obtain an orthonormal basis $\{e_i\} = \{v_j b^j{}_i\}$. This orthonormalization process may be visualized as a sequence $v_i(t) = v_j b^j{}_i(t)$ (for $0 \le t \le 1$) of continuous scalings and rotations that always maintain linear independence such that $v_i(0) = v_i$ (i.e., $b^j{}_i(0) = \delta^j{}_i$) and $v_i(1) = e_i$ (i.e., $b^j{}_i(1) = b^j{}_i$). Hence we have a continuous transformation $b^j{}_i(t)$ taking $\{v_i\}$ into $\{e_i\}$ with $\det(b^j{}_i(t)) > 0$ (the transformation starts with $\det(b^j{}_i(0)) = \det I > 0$, and since the vectors are always independent, it must maintain $\det((b^j{}_i(t)) \neq 0)$. Similarly, we may transform $\{v'_i\}$ into an orthonormal basis $\{e'_i\}$ by a continuous transformation with positive determinant. (Alternatively, it was shown in Exercise 5.4.14 that the Gram-Schmidt process is represented by an upper-triangular matrix with all positive diagonal elements, and hence its determinant is positive.) Now $\{e_i\}$ and $\{e'_i\}$ are related by an orthogonal transformation that must also have determinant equal to $+1$ because $\{v_i\}$ and $\{v'_i\}$ are related by a transformation with positive determinant, and both of the Gram-Schmidt transformations have positive determinant. This reduces the general case to the special case treated above.

With this discussion as motivation, we make the following definition. Let $\{v_1, \dots, v_n\}$ and $\{v'_1, \dots, v'_n\}$ be two ordered bases for a real vector space V, and assume that $v'_i = v_j a^j{}_i$. These two bases are said to be **similarly oriented** if $\det(a^i{}_j) > 0$, and we write this as $\{v_i\} \approx \{v'_i\}$. In other words, $\{v_i\} \approx \{v'_i\}$ if $v'_i = \phi(v_i)$ with $\det \phi > 0$. We leave it to the reader to show that this defines an equivalence relation on the set of all ordered bases for V (see Exercise 11.9.1). We denote the equivalence class of the basis $\{v_i\}$ by $[v_i]$.

It is worth pointing out that had we instead required $\det(a^i{}_j) < 0$, then this would not have defined an equivalence relation. This is because if $(b^i{}_j)$ is another such transformation with $\det(b^i{}_j) < 0$, then

$$\det(a^i{}_j b^j{}_k) = \det(a^i{}_j)\det(b^j{}_k) > 0 \ .$$

Intuitively this is quite reasonable since a combination of two reflections (each of which has negative determinant) is not another reflection.

We now define an **orientation** of V to be an equivalence class of ordered bases. The space V together with an orientation $[v_i]$ is called an **oriented vector space** (V, $[v_i]$). Since the determinant of a linear transformation that relates any two bases must be either positive or negative, we see that V has exactly two orientations. In particular, if $\{v_i\}$ is any given basis, then every other basis belonging to the equivalence class $[v_i]$ of $\{v_i\}$ will be related to $\{v_i\}$ by a transformation with positive determinant, while those bases related to $\{v_i\}$ by a transformation with negative determinant will be related to each other by a transformation with positive determinant (see Exercise 11.9.1).

Now recall we have seen that n-forms seem to be related to n-volumes in an n-dimensional space V. To precisely define this relationship, we formulate orientations in terms of n-forms. To begin with, the nonzero elements of the 1-dimensional space $\bigwedge^n(V)$ are called **volume forms** (or sometimes **volume elements**) on V. If ω_1 and ω_2 are volume forms, then ω_1 is said to be **equivalent** to ω_2 if $\omega_1 = c\omega_2$ for some real $c > 0$, and in this case we also write $\omega_1 \approx \omega_2$. Since every element of $\bigwedge^n(V)$ is related to every other element by a relationship of the form $\omega_1 = a\omega_2$ for some real a (i.e., $-\infty < a < \infty$), it is clear that this equivalence relation divides the set of all nonzero volume forms into two distinct groups (i.e., equivalence classes). We can relate any ordered basis $\{v_i\}$ for V to a specific volume form by defining

$$\omega = v^1\wedge \cdots \wedge v^n$$

where $\{v^i\}$ is the basis dual to $\{v_i\}$. That this association is meaningful is shown in the next result.

Theorem 11.19 Let $\{v_i\}$ and $\{\bar{v}_i\}$ be bases for V, and let $\{v^i\}$ and $\{\bar{v}^i\}$ be the corresponding dual bases. Define the volume forms

$$\omega = v^1\wedge \cdots \wedge v^n$$

and

$$\bar{\omega} = \bar{v}^1\wedge \cdots \wedge \bar{v}^n \ .$$

Then $\{v_i\} \approx \{\bar{v}_i\}$ if and only if $\omega \approx \bar{\omega}$.

Proof First suppose that $\{v_i\} \approx \{\bar{v}_i\}$. Then $\bar{v}_i = \phi(v_i)$ where $\det \phi > 0$, and hence (using

$$\omega(v_1, \dots, v_n) = v^1\wedge \cdots \wedge v^n(v_1, \dots, v_n) = 1$$

as shown in Example 11.8)

$$\begin{aligned}\omega(\bar{v}_1, \dots, \bar{v}_n) &= \omega(\phi(v_1), \dots, \phi(v_n))\\ &= (\phi^*\omega)(v_1, \dots, v_n)\\ &= (\det\phi)\omega(v_1, \dots, v_n)\\ &= \det\phi \ .\end{aligned}$$

If we assume that $\omega = c\bar{\omega}$ for some $-\infty < c < \infty$, then using $\bar{\omega}(\bar{v}_1, \dots, \bar{v}_n) = 1$ we see that our result implies $c = \det \phi > 0$ and thus $\omega \approx \bar{\omega}$.

Conversely, if $\omega = c\,\bar{\omega}$ where $c > 0$, then assuming that $\bar{v}_i = \phi(v_i)$, the above calculation shows that $\det \phi = c > 0$, and hence $\{v_i\} \approx \{\bar{v}_i\}$. ∎

What this theorem shows us is that an equivalence class of bases uniquely determines an equivalence class of volume forms and conversely. Therefore it is consistent with our earlier definitions to say that an equivalence class $[\omega]$ of volume forms on V defines an **orientation** on V, and the space V together with an orientation $[\omega]$ is called an **oriented vector space** $(V, [\omega])$. A basis $\{v_i\}$ for $(V, [\omega])$ is now said to be **positively oriented** if $\omega(v_1, \dots, v_n) > 0$. Not surprisingly, the equivalence class $[-\omega]$ is called the **reverse orientation**, and the basis $\{v_i\}$ is said to be **negatively oriented** if $\omega(v_1, \dots, v_n) < 0$. Note that if the ordered basis $\{v_1, v_2, \dots, v_n\}$ is negatively oriented, then the basis $\{v_2, v_1, \dots, v_n\}$ will be positively oriented because $\omega(v_2, v_1, \dots, v_n) = -\,\omega(v_1, v_2, \dots, v_n) > 0$. By way of additional terminology, the **standard orientation** on $\mathbb{R}^n$ is that orientation defined by either the standard ordered basis $\{e_1, \dots, e_n\}$, or the corresponding volume form $e^1 \wedge \cdots \wedge e^n$.

In order to proceed any further, we must introduce the notion of a metric on V. This is the subject of the next section.

Exercises

1. (a) Show that the collection of all similarly oriented bases for V defines an equivalence relation on the set of all ordered bases for V.
   (b) Let $\{v_i\}$ be a basis for V. Show that all other bases related to $\{v_i\}$ by a transformation with negative determinant will be related to each other by a transformation with positive determinant.

2. Let (U, ω) and (V, μ) be oriented vector spaces with chosen volume elements. We say that $\phi \in L(U, V)$ is **volume preserving** if $\phi^*\mu = \omega$. If $\dim U = \dim V$ is finite, show that ϕ is an isomorphism.

3. Let $(U, [\omega])$ and $(V, [\mu])$ be oriented vector spaces. We say that $\phi \in L(U, V)$ is **orientation preserving** if $\phi^*\mu \in [\omega]$. If $\dim U = \dim V$ is finite, show that ϕ is an isomorphism. If $U = V = \mathbb{R}^3$, give an example of a linear transformation that is orientation preserving but not volume preserving.

11.10 THE METRIC TENSOR AND VOLUME FORMS

We now generalize slightly our definition of inner products on V. In particular, recall from Section 2.4 (and the beginning of Section 9.2) that property (IP3) of an inner product requires that $\langle u, u\rangle \geq 0$ for all $u \in V$ and $\langle u, u\rangle = 0$ if and only if $u = 0$. If we drop this condition entirely, then we obtain an **indefinite inner product** on V. (In fact, some authors define an inner product as obeying only (IP1) and (IP2), and then refer to what we have called an inner product as a "positive definite inner product.") If we replace (IP3) by the weaker requirement

(IP3′) $\langle u, v\rangle = 0$ for all $v \in V$ if and only if $u = 0$

then our inner product is said to be **nondegenerate**. (Note that every example of an inner product given in this book up to now has been nondegenerate.) Thus a real nondegenerate indefinite inner product is just a real nondegenerate symmetric bilinear map. We will soon see an example of an inner product with the property that $\langle u, u\rangle = 0$ for some $u \neq 0$ (see Example 11.13 below).

Throughout the remainder of this chapter, we will assume that our inner products are indefinite and nondegenerate unless otherwise noted. We furthermore assume that we are dealing exclusively with real vector spaces.

Let $\{e_i\}$ be a basis for an inner product space V. Since in general we will not have $\langle e_i, e_j\rangle = \delta_{ij}$, we define the scalars g_{ij} by

$$g_{ij} = \langle e_i, e_j \rangle \ .$$

In terms of the g_{ij}, we have for any $X, Y \in V$

$$\langle X, Y\rangle = \langle x^i e_i, y^j e_j\rangle = x^i y^j \langle e_i, e_j\rangle = g_{ij}x^i y^j \ .$$

If $\{\bar{e}_i\}$ is another basis for V, then we will have $\bar{e}_i = e_j a^j{}_i$ for some nonsingular transition matrix $A = (a^j{}_i)$. Hence, writing $\bar{g}_{ij} = \langle \bar{e}_i, \bar{e}_j\rangle$ we see that

$$\bar{g}_{ij} = \langle \bar{e}_i, \bar{e}_j\rangle = \langle e_r a^r{}_i, e_s a^s{}_j\rangle = a^r{}_i a^s{}_j \langle e_r, e_s\rangle = a^r{}_i a^s{}_j g_{rs}$$

which shows that the g_{ij} transform like the components of a second-rank covariant tensor. Indeed, defining the tensor $g \in \mathcal{T}_2(V)$ by

$$g(X, Y) = \langle X, Y\rangle$$

results in

$$g(e_i, e_j) = \langle e_i, e_j\rangle = g_{ij}$$

as it should. We are therefore justified in defining the (**covariant**) **metric tensor**

$$g = g_{ij}\omega^i \otimes \omega^j \in \mathcal{T}_2(V)$$

(where $\{\omega^i\}$ is the basis dual to $\{e_i\}$) by $g(X, Y) = \langle X, Y\rangle$. In fact, since the inner product is nondegenerate and symmetric (i.e., $\langle X, Y\rangle = \langle Y, X\rangle$), we see that g is a nondegenerate symmetric tensor (i.e., $g_{ij} = g_{ji}$).

Next, we notice that given any vector $A \in V$, we may define a linear functional $\langle A, \ \rangle$ on V by the assignment $B \mapsto \langle A, B\rangle$. In other words, for any $A \in V$, we associate the 1-form α defined by $\alpha(B) = \langle A, B\rangle$ for every $B \in V$. Note that the kernel of the mapping $A \mapsto \langle A, \rangle$ (which is easily seen to be a vector space homomorphism) consists of only the zero vector (since $\langle A, B\rangle = 0$ for every $B \in V$ implies that $A = 0$), and hence this association is an iso-morphism. Given any basis $\{e_i\}$ for V, the components a_i of $\alpha \in V^*$ are given in terms of those of $A = a^ie_i \in V$ by

$$a_i = \alpha(e_i) = \langle A, e_i\rangle = \langle a^j e_j, e_i\rangle = a^j\langle e_j, e_i\rangle = a^jg_{ji}$$

Thus, to any contravariant vector $A = a^i e_i \in V$, we can associate a unique covariant vector $\alpha \in V^*$ by

$$\alpha = a_i\omega^i = (a^jg_{ji})\omega^i$$

where $\{\omega^i\}$ is the basis for V^* dual to the basis $\{e_i\}$ for V. In other words, we write

$$a_i = a^jg_{ji}$$

and we say that a_i arises by **lowering the index** j of a^j.

Example 11.12 If we consider the space $\mathbb{R}^n$ with a Cartesian coordinate system $\{e_i\}$, then we have $g_{ij} = \langle e_i, e_j \rangle = \delta_{ij}$, and hence $a_i = \delta_{ij}a^j = a^i$. Therefore, *in a Cartesian coordinate system*, there is no distinction between the components of covariant and contravariant vectors. This explains why 1-forms never arise in elementary treatments of vector analysis. ∕∕∕

Since the metric tensor is nondegenerate, the matrix (g_{ij}) must be nonsingular (or else the mapping $a^j \mapsto a_i$ would not be an isomorphism). We can therefore define the inverse matrix (g^{ij}) by

$$g^{ij}g_{jk} = g_{kj}g^{ji} = \delta^i{}_k \ .$$

Using (g^{ij}), we see that the inverse of the mapping $a^j \mapsto a_i$ is given by

$$g^{ij}a_j = a^i \ .$$

This is called, naturally enough, **raising an index**. We will show below that the g^{ij} do indeed form the components of a tensor.

It is worth remarking that the "tensor" $g^i{}_j = g^{ik}g_{kj} = \delta^i{}_j$ $(= \delta_j{}^i)$ is unique in that it has the same components in any coordinate system. Indeed, if $\{e_i\}$ and $\{\bar{e}_i\}$ are two bases for a space V with corresponding dual bases $\{\omega^i\}$ and $\{\bar{\omega}^i\}$, then $\bar{e}_i = e_j a^j{}_i$ and $\bar{\omega}^j = b^j{}_i\omega^i = (a^{-1})^j{}_i\omega^i$ (see the discussion following Theorem 11.2). Therefore, if we define the tensor δ to have the same values in the first coordinate system as the Kronecker delta, then $\delta^i{}_j = \delta(\omega^i, e_j)$. If we now define the symbol $\bar{\delta}^i{}_j$ by $\bar{\delta}^i{}_j = \delta(\bar{\omega}^i, \bar{e}_j)$, then we see that

$$\begin{aligned}\bar{\delta}^i{}_j &= \delta(\bar{\omega}^i, \bar{e}_j) = \delta((a^{-1})^i{}_k\omega^k, e_r a^r{}_j) = (a^{-1})^i{}_k a^r{}_j \delta(\omega^k, e_r) \\ &= (a^{-1})^i{}_k a^r{}_j \delta^k{}_r = (a^{-1})^i{}_k a^k{}_j = \delta^i{}_j \ .\end{aligned}$$

This shows that the $\delta^i{}_j$ are in fact the components of a tensor, and that these components are the same in any coordinate system.

We would now like to show that the scalars g^{ij} are indeed the components of a tensor. There are several ways that this can be done. First, let us write $g_{ij}g^{jk} = \delta^k{}_i$ where we know that both g_{ij} and $\delta^k{}_i$ are tensors. Multiplying both sides of this equation by $(a^{-1})^r{}_k a^i{}_s$ and using $(a^{-1})^r{}_k a^i{}_s \delta^k{}_i = \delta^r{}_s$ we find

$$g_{ij}g^{jk}(a^{-1})^r{}_k a^i{}_s = \delta^r{}_s \ .$$

Now substitute $g_{ij} = g_{it}\delta^t{}_j = g_{it}a^t{}_q(a^{-1})^q{}_j$ to obtain

$$[a^i{}_s a^t{}_q g_{it}][(a^{-1})^q{}_j(a^{-1})^r{}_k g^{jk}] = \delta^r{}_s \ .$$

Since g_{it} is a tensor, we know that $a^i{}_s a^t{}_q g_{it} = \bar{g}_{sq}$. If we write

$$\bar{g}^{qr} = (a^{-1})^q{}_j(a^{-1})^r{}_k g^{jk}$$

then we will have defined the g^{jk} to transform as the components of a tensor, and furthermore, they have the requisite property that $\bar{g}_{sq}\,\bar{g}^{qr} = \delta^r{}_s$. Therefore we have defined the (**contravariant**) **metric tensor** $G \in \mathcal{T}_0^{\,2}(V)$ by

$$G = g^{ij} e_i \otimes e_j$$

where $g^{ij}g_{jk} = \delta^i{}_k$.

There is another interesting way for us to define the tensor G. We have already seen that a vector $A = a^i e_i \in V$ defines a unique linear form $\alpha = a_j\omega^j \in V^*$ by the association $\alpha = g_{ij}a^i\omega^j$. If we denote the inverse of the matrix (g_{ij}) by (g^{ij}) so that $g^{ij}g_{jk} = \delta^i{}_k$, then to any linear form $\alpha = a_i\omega^i \in V^*$ there corresponds a unique vector $A = a^i e_i \in V$ defined by $A = g^{ij}a_ie_j$. We can now use this isomorphism to define an inner product on V*. In other words, if $\langle\ ,\ \rangle$ is an inner product on V, we define an inner product $\langle\ ,\ \rangle$ on V* by

$$\langle\alpha, \beta\rangle = \langle A, B\rangle$$

where $A, B \in V$ are the vectors corresponding to the 1-forms $\alpha, \beta \in V^*$.

Let us write an arbitrary basis vector e_i in terms of its components relative to the basis $\{e_i\}$ as $e_i = \delta^j{}_ie_j$. Therefore, in the above isomorphism, we may define the linear form $\hat{e}_i \in V^*$ corresponding to the basis vector e_i by

$$\hat{e}_i = g_{jk}\delta^j{}_i\omega^k = g_{ik}\omega^k$$

and hence using the inverse matrix, we find that

$$\omega^k = g^{ki}\hat{e}_i\ .$$

Applying our definition of the inner product in V* we have $\langle\hat{e}_i, \hat{e}_j\rangle = \langle e_i, e_j\rangle = g_{ij}$, and therefore we obtain

$$\langle\omega^i, \omega^j\rangle = \langle g^{ir}\hat{e}_r, g^{js}\hat{e}_s\rangle = g^{ir}g^{js}\langle\hat{e}_r, \hat{e}_s\rangle = g^{ir}g^{js}g_{rs} = g^{ir}\delta^j{}_r = g^{ij}$$

which is the analogue in V* of the definition $g_{ij} = \langle e_i, e_j\rangle$ in V.

Lastly, since $\bar{\omega}^j = (a^{-1})^j{}_i\omega^i$, we see that

$$\begin{aligned}\bar{g}^{ij} &= \langle\bar{\omega}^i, \bar{\omega}^j\rangle = \langle(a^{-1})^i{}_r\omega^r, (a^{-1})^j{}_s\omega^s\rangle = (a^{-1})^i{}_r(a^{-1})^j{}_s\langle\omega^r, \omega^s\rangle \\ &= (a^{-1})^i{}_r(a^{-1})^j{}_s g^{rs}\end{aligned}$$

so the scalars g^{ij} may be considered to be the components of a symmetric tensor $G \in \mathcal{T}_2{}^0(V)$ defined as above by $G = g^{ij}e_i \otimes e_j$.

Now let $g = \langle\ ,\ \rangle$ be an arbitrary (i.e., possibly degenerate) real symmetric bilinear form on the inner product space V. It follows from the corollary to Theorem 9.14 that there exists a basis $\{e_i\}$ for V in which the matrix (g_{ij}) of g takes the unique diagonal form

$$g_{ij} = \begin{pmatrix} I_r & & \\ & -I_s & \\ & & 0_t \end{pmatrix}$$

where r + s + t = dim V = n. Thus

$$g(e_i,\, e_i) = \begin{cases} 1 & \text{for } 1 \le i \le r \\ -1 & \text{for } r+1 \le i \le r+s \\ 0 & \text{for } r+s+1 \le i \le n \end{cases} \;.$$

If r + s < n, the inner product is degenerate and we say that the space V is **singular** (with respect to the given inner product). If r + s = n, then the inner product is nondegenerate, and the basis $\{e_i\}$ is orthonormal. In the orthonormal case, if either r = 0 or r = n, the space is said to be **ordinary Euclidean**, and if 0 < r < n, then the space is called **pseudo-Euclidean**. Recall that the number r - s = r - (n - r) = 2r - n is called the **signature** of g (which is therefore just the trace of (g_{ij})). Moreover, the number of -1's is called the **index** of g, and is denoted by Ind(g). If g = ⟨ , ⟩ is to be a metric on V, then by definition, we must have r + s = n so that the inner product is nondegenerate. In this case, the basis $\{e_i\}$ is called **g-orthonormal**.

Example 11.13 If the metric g represents a positive definite inner product on V, then we must have Ind(g) = 0, and such a metric is said to be **Riemannian**. Alternatively, another well-known metric is the Lorentz metric used in the theory of special relativity. By definition, a **Lorentz** metric η has Ind(η) = 1. Therefore, if η is a Lorentz metric, an η-orthonormal basis $\{e_1, \dots, e_n\}$ ordered in such a way that $\eta(e_i, e_i) = +1$ for i = 1, . . . , n - 1 and $\eta(e_n, e_n) = -1$ is called a **Lorentz frame**.

Thus, in terms of a g-orthonormal basis, a Riemannian metric has the form

$$(g_{ij}) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

while in a Lorentz frame, a Lorentz metric takes the form

$$(\eta_{ij}) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -1 \end{pmatrix} \;.$$

It is worth remarking that a Lorentz metric is also frequently defined as having $\mathrm{Ind}(\eta) = n - 1$. In this case we have $\eta(e_1, e_1) = 1$ and $\eta(e_i, e_i) = -1$ for each $i = 2, \dots, n$. We also point out that a vector $v \in V$ is called **timelike** if $\eta(v, v) < 0$, **lightlike** (or **null**) if $\eta(v, v) = 0$, and **spacelike** if $\eta(v, v) > 0$. Note that a Lorentz inner product is clearly indefinite since, for example, the nonzero vector v with components $v = (0, 0, 1, 1)$ has the property that $\langle v, v\rangle = \eta(v, v) = 0$. ∕∕

We now show that introducing a metric on V leads to a unique volume form on V.

Theorem 11.20 Let g be a metric on an n-dimensional oriented vector space $(V, [\omega])$. Then, corresponding to the metric g, there exists a unique volume form $\mu = \mu(g) \in [\omega]$ such that $\mu(e_1, \dots, e_n) = 1$ for every positively oriented g-orthonormal basis $\{e_i\}$ for V. Moreover, if $\{v_i\}$ is any (not necessarily g-orthonormal) positively oriented basis for V with dual basis $\{v^i\}$, then

$$\mu = |\det(g(v_i, v_j))|^{1/2}\, v^1 \wedge \cdots \wedge v^n \;.$$

In particular, if $\{v_i\} = \{e_i\}$ is a g-orthonormal basis, then $\mu = e^1 \wedge \cdots \wedge e^n$.

Proof Since $\omega \neq 0$, there exists a positively oriented g-orthonormal basis $\{e_i\}$ such that $\omega(e_1, \dots, e_n) > 0$ (we can multiply e_1 by -1 if necessary in order that $\{e_i\}$ be positively oriented). We now define $\mu \in [\omega]$ by

$$\mu(e_1, \dots, e_n) = 1 \;.$$

That this defines a unique μ follows by multilinearity. We claim that if $\{f_i\}$ is any other positively oriented g-orthonormal basis, then $\mu(f_1, \dots, f_n) = 1$ also. To show this, we first prove a simple general result.

Suppose $\{v_i\}$ is any other basis for V related to the g-orthonormal basis $\{e_i\}$ by $v_i = \phi(e_i) = e_j a^j{}_i$ where, by Theorem 11.17, we have $\det \phi = \det(a^i{}_j)$. We then have $g(v_i, v_j) = a^r{}_i a^s{}_j g(e_r, e_s)$ which in matrix notation is $[g]_v = A^T[g]_e A$, and hence

$$\det(g(v_i, v_j)) = (\det\phi)^2 \det(g(e_r, e_s)) \;. \tag{9}$$

However, since $\{e_i\}$ is g-orthonormal we have $g(e_r, e_s) = \pm\delta_{rs}$, and therefore $|\det(g(e_r, e_s))| = 1$. In other words

$$|\det(g(v_i, v_j))|^{1/2} = |\det\phi| \;. \tag{10}$$

Returning to our problem, we have $\det(g(f_i, f_j)) = \pm 1$ also since $\{f_i\} = \{\phi(e_i)\}$ is g-orthonormal. Thus (10) implies that $|\det \phi| = 1$. But $\{f_i\}$ is positively oriented so that $\mu(f_1, \dots, f_n) > 0$ by definition. Therefore

$$\begin{aligned} 0 < \mu(f_1, \dots, f_n) &= \mu(\phi(e_1), \dots \mu(\phi(e_n)) = (\phi^*\mu)(e_1, \dots, e_n) \\ &= (\det\phi)\mu(e_1, \dots, e_n) = \det\phi \end{aligned}$$

so that we must in fact have $\det \phi = +1$. In other words, $\mu(f_1, \dots, f_n) = 1$ as claimed.

Now suppose that $\{v_i\}$ is an arbitrary positively oriented basis for V such that $v_i = \phi(e_i)$. Then, analogously to what we have just shown, we see that $\mu(v_1, \dots, v_n) = \det \phi > 0$. Hence (10) shows that (using Example 11.8)

$$\begin{aligned} \mu(v_1, \dots, v_n) &= \det\phi \\ &= |\det(g(v_i, v_j))|^{1/2} \\ &= |\det(g(v_i, v_j))|^{1/2} v^1 \wedge \cdots \wedge v^n(v_1, \dots, v_n) \end{aligned}$$

which implies

$$\mu = |\det(g(v_i, v_j))|^{1/2}\, v^1 \wedge \cdots \wedge v^n \;. \;\blacksquare$$

The unique volume form μ defined in Theorem 11.20 is called the **g-volume**, or sometimes the **metric volume form**. A common (although rather careless) notation is to write $|\det(g(v_i, v_j))|^{1/2} = \sqrt{|g|}$. In this notation, the g-volume is written as

$$\sqrt{|g|}\, v^1 \wedge \cdots \wedge v^n$$

where $\{v_1, \dots, v_n\}$ must be positively oriented. If the basis $\{v_1, v_2, \dots, v_n\}$ is negatively oriented, then clearly $\{v_2, v_1, \dots, v_n\}$ will be positively oriented. Furthermore, even though the matrix of g relative to each of these oriented bases will be different, the determinant actually remains unchanged (see the discussion following the corollary to Theorem 11.13). Therefore, for this negatively oriented basis, the g-volume is

$$\sqrt{|g|}\, v^2 \wedge v^1 \wedge \cdots \wedge v^n = -\sqrt{|g|}\, v^1 \wedge v^2 \wedge \cdots \wedge v^n \;.$$

We thus have the following corollary to Theorem 11.20.

Corollary Let $\{v_i\}$ be any basis for the n-dimensional oriented vector space $(V, [\omega])$ with metric g. Then the g-volume form on V is given by

$$\pm\sqrt{|g|}\, v^1 \wedge \cdots \wedge v^n$$

where the "+" sign is for $\{v_i\}$ positively oriented, and the "−" sign is for $\{v_i\}$ negatively oriented.

Example 11.14 From Example 11.13, we see that for a Riemannian metric g and g-orthonormal basis $\{e_i\}$ we have $\det(g(e_i, e_j)) = +1$. Hence, from equation (9), we see that $\det(g(v_i, v_j)) > 0$ for any basis $\{v_i = \phi(e_i)\}$. Thus the g-volume form on a Riemannian space is given by $\pm\sqrt{g}\, v^1 \wedge \cdots \wedge v^n$.

For a Lorentz metric we have det(ƒ(e_i, e_j)) = −1 in a Lorentz frame, and therefore $\det(g(v_i, v_j)) < 0$ in an arbitrary frame. Thus the g-volume in a Lorentz space is given by $\pm\sqrt{-g}\, v^1 \wedge \cdots \wedge v^n$.

Let us point out that had we defined $\operatorname{Ind}(\eta) = n - 1$ instead of $\operatorname{Ind}(\eta) = 1$, then $\det(\eta(e_i, e_j)) < 0$ only in an even dimensional space. In this case, we would have to write the g-volume as in the above corollary. ∕∕

Example 11.15 (This example is a continuation of Example 11.11.) We remark that these volume elements are of great practical use in the theory of integration on manifolds. To see an example of how this is done, let us use Examples 11.1 and 11.11 to write the volume element as (remember that this applies only locally, and hence the metric depends on the coordinates)

$$d\tau = \sqrt{|g|}\, dx^1 \wedge \cdots \wedge dx^n \ .$$

If we go to a new coordinate system $\{\bar{x}^i\}$, then

$$\bar{g}_{ij} = \frac{\partial x^r}{\partial \bar{x}^i} \frac{\partial x^s}{\partial \bar{x}^j} g_{rs}$$

so that $|\bar{g}| = (J^{-1})^2 |g|$ where $J^{-1} = \det(\partial x^r/\partial \bar{x}^i)$ is the determinant of the inverse Jacobian matrix of the transformation. But using $d\bar{x}^i = (\partial \bar{x}^i/\partial x^j)dx^j$ and the properties of the wedge product, it is easy to see that

$$\begin{aligned} d\bar{x}^1 \wedge \cdots \wedge d\bar{x}^n &= \frac{\partial \bar{x}^1}{\partial x^{i_1}} \cdots \frac{\partial \bar{x}^n}{\partial x^{i_n}} dx^{i_1} \wedge \cdots \wedge dx^{i_n} \\ &= \det\left(\frac{\partial \bar{x}^i}{\partial x^j} \right) dx^1 \wedge \cdots \wedge dx^n \end{aligned}$$

and hence

$$d\bar{x}^1 \wedge \cdots \wedge d\bar{x}^n = J\, dx^1 \wedge \cdots \wedge dx^n$$

where J is the determinant of the Jacobian matrix. (Note that the proper transformation formula for the volume element in multiple integrals arises naturally in the algebra of exterior forms.) We now have

$$\begin{aligned} d\bar{\tau} &= \sqrt{|\bar{g}|}\, d\bar{x}^1 \wedge\cdots\wedge d\bar{x}^n = J^{-1}\sqrt{|g|}\, J\, dx^1 \wedge\cdots\wedge dx^n \\ &= \sqrt{|g|}\, dx^1 \wedge\cdots\wedge dx^n = d\tau \end{aligned}$$

and hence $d\tau$ is a scalar called the **invariant volume element**. In the case of $\mathbb{R}^4$ as a Lorentz space, this result is used in the theory of relativity. ∕∕

Exercises

1. Suppose V has a metric g_{ij} defined on it. Show that for any $A, B \in V$ we have $\langle A, B\rangle = a_ib^i = a^ib_i$.

2. According to the special theory of relativity, the speed of light is the same for all unaccelerated observers regardless of the motion of the source of light relative to the observer. Consider two observers moving at a constant velocity β with respect to each other, and assume that the origins of their respective coordinate systems coincide at $t = 0$. If a spherical pulse of light is emitted from the origin at $t = 0$, then (in units where the speed of light is equal to 1) this pulse satisfies the equation $x^2 + y^2 + z^2 - t^2 = 0$ for the first observer, and $\bar{x}^2 + \bar{y}^2 + \bar{z}^2 - \bar{t}^{\,2} = 0$ for the second observer. We shall use the common notation $(t, x, y, z) = (x^0, x^1, x^2, x^3)$ for our coordinates, and hence the Lorentz metric takes the form

$$\eta_{\mu\nu} = \begin{pmatrix} -1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

where $0 \le \mu, \nu \le 3$.

(a) Let the Lorentz transformation matrix be Λ so that $\bar{x}^\mu = \Lambda^\mu{}_\nu\, x^\nu$. Show that the Lorentz transformation must satisfy $\Lambda^T\eta\Lambda = \eta$.

(b) If the $\{\bar{x}^\mu\}$ system moves along the x^1-axis with velocity β, then it turns out that the Lorentz transformation is given by

$$\begin{aligned} \bar{x}^0 &= \gamma(x^0 - \beta x^1) \\ \bar{x}^1 &= \gamma(x^1 - \beta x^0) \end{aligned}$$

$$\bar{x}^2 = x^2$$
$$\bar{x}^3 = x^3$$

where $\gamma^2 = 1/(1 - \beta^2)$. Using $\Lambda^\mu{}_\nu = \partial\bar{x}^\mu/\partial x^\nu$, write out the matrix $(\Lambda^\mu{}_\nu)$, and verify explicitly that $\Lambda^T\eta\Lambda = \eta$.

(c) The electromagnetic field tensor is given by

$$F_{\mu\nu} = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix} .$$

Using this, find the components of the electric field $\vec{E}$ and magnetic field $\vec{B}$ in the $\{\bar{x}^\mu\}$ coordinate system. In other words, find $\bar{F}_{\mu\nu}$. (The actual definition of $F_{\mu\nu}$ is given by $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ where $\partial_\mu = \partial/\partial x^\mu$ and $A_\mu = (\phi, A_1, A_2, A_3)$ is related to $\vec{E}$ and $\vec{B}$ through the classical equations $\vec{E} = -\nabla\phi - \partial\vec{A}/\partial t$ and $\vec{B} = \nabla \times \vec{A}$. See also Exercise 11.1.6.)

3. Let V be an n-dimensional vector space with a Lorentz metric η, and let W be an $(n - 1)$-dimensional subspace of V. Note that

$$W^\perp = \{v \in V: \eta(v, w) = 0 \text{ for all } w \in W\}$$

is the 1-dimensional subspace of all normal vectors for W. We say that W is **timelike** if every normal vector is spacelike, **null** if every normal vector is null, and **spacelike** if every normal vector is timelike. Prove that η restricted to W is

(a) Positive definite if W is spacelike.

(b) A Lorentz metric if W is timelike.

(c) Degenerate if W is null.

4. (a) Let D be a 3 x 3 determinant considered as a function of three contravariant vectors $A^i{}_{(1)}$, $A^i{}_{(2)}$, and $A^i{}_{(3)}$. Show that under a change of coordinates, D does not transform as a scalar, but that $D\sqrt{|g|}$ does transform as a proper scalar. [*Hint*: Use Exercise 11.2.8.]

(b) Show that $\varepsilon_{ijk}\sqrt{|g|}$ transforms like a tensor. (This is the Levi-Civita tensor in general coordinates. Note that in a g-orthonormal coordinate system this reduces to the Levi-Civita symbol.)

(c) What is the contravariant version of the tensor in part (b)?

CHAPTER 12

Hilbert Spaces

The material to be presented in this chapter is essential for all advanced work in physics and analysis. We have attempted to present several relatively difficult theorems in sufficient detail that they are readily understandable by readers with less background than normally might be required for such results. However, we assume that the reader is quite familiar with the contents of Appendices A and B, and we will frequently refer to results from these appendices. Essentially, this chapter serves as an introduction to the theory of infinite-dimensional vector spaces. Throughout this chapter we let E, F and G denote normed vector spaces over the real or complex number fields only.

12.1 MATHEMATICAL PRELIMINARIES

This rather long first section presents the elementary properties of limits and continuous functions. While most of this material properly falls under the heading of analysis, we do not assume that the reader has already had such a course. However, if these topics are familiar, then the reader should briefly scan the theorems of this section now, and return only for details if and when it becomes necessary.

For ease of reference, we briefly repeat some of our earlier definitions and results. By a **norm** on a vector space E, we mean a mapping $\| \,\|: E \to \mathbb{R}$ satisfying:

(N1) $\|u\| \geq 0$ for every $u \in E$ and $\|u\| = 0$ if and only if $u = 0$ (positive definiteness).

(N2) $\|cu\| = |c|\,\|u\|$ for every $u \in E$ and $c \in \mathcal{F}$.

(N3) $\|u + v\| \leq \|u\| + \|v\|$ (triangle inequality).

If there is more than one norm on E under consideration, then we may denote them by subscripts such as $\|\ \|_2$ etc. Similarly, if we are discussing more than one space, then the norm associated with a space E will sometimes be denoted by $\|\ \|_E$. We call the pair $(E, \|\ \|)$ a **normed vector space**.

If E is a complex vector space, we define the **Hermitian inner product** as the mapping $\langle\ ,\ \rangle$: $E \times E \to \mathbb{C}$ such that for all $u, v, w \in E$ and $c \in \mathbb{C}$ we have:

(IP1) $\langle u, v + w\rangle = \langle u, v\rangle + \langle u, w\rangle$.

(IP2) $\langle cu, v\rangle = c^*\langle u, v\rangle$.

(IP3) $\langle u, v\rangle = \langle v, u\rangle^*$.

(IP4) $\langle u, u\rangle \geq 0$ and $\langle u, u\rangle = 0$ if and only if $u = 0$.

where * denotes complex conjugation. A Hermitian inner product is sometimes called a **sesquilinear form**. Note that (IP2) and (IP3) imply

$$\langle u, cv\rangle = \langle cv, u\rangle^* = c\langle v, u\rangle^* = c\langle u, v\rangle$$

and that (IP3) implies $\langle v, v\rangle$ is real.

As usual, if $\langle u, v\rangle = 0$ we say that u and v are orthogonal, and we sometimes write this as $u \perp v$. If we let S be a subset of E, then the collection

$$\{u \in E\colon \langle u, v\rangle = 0 \text{ for every } v \in S\}$$

is a subspace of E called the **orthogonal complement** of S, and is denoted by $S^\perp$.

It should be remarked that many authors define $\langle cu, v\rangle = c\langle u, v\rangle$ rather than our (IP2), and the reader must be careful to note which definition is being followed. Furthermore, there is no reason why we could not have defined a mapping $E \times E \to \mathbb{R}$, and in this case we have simply an **inner product** on E (where obviously there is now no complex conjugation).

The most common example of a Hermitian inner product is the standard inner product on $\mathbb{C}^n = \mathbb{C} \times \cdots \times \mathbb{C}$ defined for all $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{C}^n$ by

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i{}^* y_i \ .$$

We leave it to the reader to verify conditions (IP1) – (IP4). Before defining a norm on $\mathbb{C}^n$, we again prove (in a slightly different manner from that in Chapter 2) the **Cauchy-Schwartz inequality**.

Example 12.1 Let E be a complex (or real) inner product space, and let u, v ∈ E be nonzero vectors. Then for any a, b ∈ $\mathbb{C}$ we have

$$0 \le \langle au + bv, au + bv\rangle = |a|^2\langle u, u\rangle + a^*b\langle u, v\rangle + b^*a\langle v, u\rangle + |b|^2\langle v, v\rangle \ .$$

Now note that the middle two terms are complex conjugates of each other, and hence their sum is $2\mathrm{Re}(a^*b\langle u, v\rangle)$. Therefore, letting $a = \langle v, v\rangle$ and $b = -\langle v, u\rangle$, we have

$$0 \le \langle v, v\rangle^2\langle u, u\rangle - 2\langle v, v\rangle|\langle u, v\rangle|^2 + |\langle u, v\rangle|^2\langle v, v\rangle$$

which is equivalent to

$$\langle v, v\rangle|\langle u, v\rangle|^2 \le \langle v, v\rangle^2\langle u, u\rangle \ .$$

Since $v \ne 0$ we have $\langle v, v\rangle \ne 0$, and hence dividing by $\langle v, v\rangle$ and taking the square root yields the desired result

$$|\langle u, v\rangle| \le \langle u, u\rangle^{1/2} \langle v, v\rangle^{1/2} \ . \ /\!\!/$$

If a vector space E has an inner product defined on it, then we may define a norm on E by

$$\|v\| = \langle v, v\rangle^{1/2}$$

for all v ∈ E. Properties (N1) and (N2) for this norm are obvious, and (N3) now follows from the Cauchy-Schwartz inequality and the fact that $\mathrm{Re}\langle u, v\rangle \le |\langle u, v\rangle|$:

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v\rangle \\ &= \|u\|^2 + 2\,\mathrm{Re}\langle u, v\rangle + \|v\|^2 \\ &\le \|u\|^2 + 2\,|\langle u, v\rangle| + \|v\|^2 \\ &\le \|u\|^2 + 2\,\|u\|\|v\| + \|v\|^2 \\ &= (\|u\| + \|v\|)^2 \ . \end{aligned}$$

We leave it to the reader (Exercise 12.1.1) to prove the so-called **parallelogram law** in an inner product space (E, ⟨ , ⟩):

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2 \ .$$

The geometric meaning of this formula in $\mathbb{R}^2$ is that the sum of the squares of the diagonals of a parallelogram is equal to the sum of the squares of the sides. If $\langle u, v\rangle = 0$, then the reader can also easily prove the **Pythagorean theorem**:

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2 \quad \text{if } u \perp v \ .$$

In terms of the standard inner product on $\mathbb{C}^n$, we now define a norm on $\mathbb{C}^n$ by

$$\|x\|^2 = \langle x, x\rangle = \sum_{i=1}^{n} |x_i|^2 \ .$$

The above results now show that this does indeed satisfy the requirements of a norm.

Continuing, if $(E, \| \ \|)$ is a normed space, then we may make E into a metric space (E, d) by defining

$$d(u, v) = \|u - v\| \ .$$

Again, the only part of the definition of a metric space (see Appendix A) that is not obvious is (M4), and this now follows from (N3) because

$$\begin{aligned} d(u, v) &= \|u - v\| = \|u - w + w - v\| \le \|u - w\| + \|w - v\| \\ &= d(u, w) + d(w, v) \ . \end{aligned}$$

The important point to get from all this is that normed vector spaces form a special class of metric spaces. This means that all the results from Appendix A and many of the results from Appendix B will carry over to the case of normed spaces. In Appendix B we presented the theory of sequences and series of numbers. As we explained there however, many of the results are valid as well for normed vector spaces if we simply replace the absolute value by the norm.

For example, suppose $A \subset E$ and let $v \in E$. Recall that v is said to be an **accumulation point** of A if every open ball centered at v contains a point of A distinct from v. In other words, given $\varepsilon > 0$ there exists $u \in A$, $u \neq v$, such that $\|u - v\| < \varepsilon$. As expected, if $\{v_n\}$ is a sequence of vectors in E, then we say that $\{v_n\}$ **converges** to the **limit** $v \in E$ if given $\varepsilon > 0$, there exists an integer $N > 0$ such that $n \ge N$ implies $\|v_n - v\| < \varepsilon$. As usual, we write $\lim v_n = \lim_{n\to\infty} v_n = v$. If there exists a neighborhood of v (i.e., an open ball containing v) such that v is the only point of A in this neighborhood, then we say that v is an **isolated point** of A.

Example 12.2 Suppose $\lim v_n = v$. Then for every $\varepsilon > 0$, there exists N such that $n \geq N$ implies $\|v - v_n\| < \varepsilon$. From Example 2.11 we then see that

$$|\,\|v\| - \|v_n\|\,| \leq \|v - v_n\| < \varepsilon$$

and hence directly from the definition of $\lim \|v_n\|$ we have

$$\|\lim v_n\| = \|v\| = \lim \|v_n\| \;.$$

This result will be quite useful in several later proofs. ∕∕

Note that if v is an accumulation point of A, then for every $n > 0$ there exists $v_n \in A$, $v_n \neq v$, such that $\|v_n - v\| < 1/n$. In particular, for any $\varepsilon > 0$, choose N so that $1/N < \varepsilon$. Then for all $n \geq N$ we have $\|v_n - v\| < 1/n < \varepsilon$ so that $\{v_n\}$ converges to v. Conversely, it is clear that if $v_n \to v$ with $v_n \in A$, $v_n \neq v$, then v must necessarily be an accumulation point of A. This proves the following result.

Theorem 12.1 If $A \subset E$, then v is an accumulation point of A if and only if it is the limit of some sequence in $A - \{v\}$.

A function $f\colon (X, d_X) \to (Y, d_Y)$ is said to be **continuous** at $x_0 \in X$ if for each $\varepsilon > 0$ there exists $\delta > 0$ such that $d_X(x, x_0) < \delta$ implies $d_Y(f(x), f(x_0)) < \varepsilon$. Note though, that for any given ε, the δ required will in general be different for each point x_0 chosen. If f is continuous at each point of X, then we say that f is “continuous on X.”

A function f as defined above is said to be **uniformly continuous** on X if for each $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x, y \in X$ with $d_X(x, y) < \delta$ we have $d_Y(f(x), f(y)) < \varepsilon$. The important difference between continuity and uniform continuity is that for a uniformly continuous function, once ε is chosen, there is a *single* δ (which will generally still depend on ε) such that this definition applies to *all* points $x, y \in X$ subject only to the requirement that $d_X(x, y) < \delta$. It should be clear that a uniformly continuous function is necessarily continuous, but the converse is not generally true. We do though have the next very important result. However, since we shall not have any occasion to refer to it in this text, we present it only for its own sake and as an (important and useful) illustration of the concepts involved.

Theorem 12.2 Let $A \subset (X, d_X)$ be compact, and let $f\colon A \to (Y, d_Y)$ be continuous. Then f is uniformly continuous. In other words, a continuous function on a compact set is uniformly continuous.

Proof Fix $\varepsilon > 0$. Since f is continuous on A, for each point $x \in A$ there exists $\delta_x > 0$ such that for all $y \in A$, $d_X(x, y) < \delta_x$ implies $d_Y(f(x), f(y)) < \varepsilon/2$. The collection $\{B(x, \delta_x/2): x \in A\}$ of open balls clearly covers A, and since A is compact, a finite number will cover A. Let $\{x_1, \dots, x_n\}$ be the finite collection of points such that $\{B(x_i, \delta_{x_i}/2)\}$, $i = 1, \dots, n$ covers A, and define $\delta = (1/2)\min(\{\delta_{x_i}\})$. Since each $\delta_{x_i} > 0$, δ must also be > 0. (Note that if A were not compact, then $\delta = \inf(\{\delta_x\})$ taken over all $x \in A$ could be equal to 0.)

Now let $x, y \in A$ be any two points such that $d_X(x, y) < \delta$. Since the collection $\{B(x_i, \delta_{x_i}/2)\}$ covers A, x must lie in some $B(x_i, \delta_{x_i}/2)$, and hence $d_X(x, x_i) < \delta_{x_i}/2$ for this particular x_i. Then we also have

$$d_X(y, x_i) \le d_X(x, y) + d_X(x, x_i) < \delta + \delta_{x_i}/2 \le \delta_{x_i} .$$

But f is continuous at x_i, and δ_{x_i} was defined so that the set of points z for which $d_X(z, x_i) < \delta_{x_i}$ satisfies $d_Y(f(z), f(x_i)) < \varepsilon/2$. Since we just showed that x and y satisfy $d_X(x, x_i) < \delta_{x_i}/2 < \delta_{x_i}$ and $d_X(y, x_i) < \delta_{x_i}$, we must have

$$d_Y(f(x), f(y)) \le d_Y(f(x), f(x_i)) + d_Y(f(y), f(x_i)) < \varepsilon/2 + \varepsilon/2 = \varepsilon .$$

In other words, for our given ε, we found a δ such that for all $x, y \in A$ with $d_X(x, y) < \delta$, we have $d_Y(f(x), f(y)) < \varepsilon$. ∎

Example 12.3 Consider the function $f: E \to \mathbb{R}$ defined by $f(u) = \|u\|$. In other words, f is just the norm function on E. Referring to the above discussion, we say that a function g is uniformly continuous if given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|u - v\| < \delta$ implies that $|g(u) - g(v)| < \varepsilon$ (note that the norm on E is $\| \ \|$ while the norm on $\mathbb{R}$ is $| \ |$). But for our norm function f and for any $\varepsilon > 0$, we see that for all $u, v \in E$, if we choose $\delta = \varepsilon$ then $\|u - v\| < \delta = \varepsilon$ implies

$$|f(u) - f(v)| = |\, \|u\| - \|v\| \,| \le \|u - v\| < \varepsilon$$

(where we used Example 2.11). Thus the norm is in fact uniformly continuous on E.

We leave it to the reader (see Exercise 12.1.2) to show (using the Cauchy-Schwartz inequality) that the inner product on E is also continuous in both variables. ∕∕

There is an equivalent way to define continuous functions in terms of limits that is also of great use. Let X and Y be metric spaces, and suppose $f: A \subset X \to Y$. Then if x_0 is an accumulation point of A, we say that a point $L \in$

Y is the **limit** of f at x_0 if, given $\varepsilon > 0$, there exists $\delta > 0$ (which may depend on f, x_0 and ε) such that for all $x \in A$ we have $0 < d_X(x, x_0) < \delta$ implies $d_Y(f(x), L) < \varepsilon$. This is written as $\lim_{x \to x_0} f(x) = L$ or simply "$f(x) \to L$ as $x \to x_0$."

Note that while x_0 is an accumulation point of A, x_0 is not necessarily an element of A, and hence $f(x_0)$ might not be defined. In addition, even if $x_0 \in A$, it is *not* necessarily true that $\lim_{x \to x_0} f(x) = f(x_0)$. However, we do have the following result.

Theorem 12.3 If $f\colon A \subset (X, d_X) \to (Y, d_Y)$ and $x_0 \in A$ is an accumulation point of A, then f is continuous at x_0 if and only if

$$\lim_{x \to x_0} f(x) = f(x_0) \;.$$

Proof This obvious by comparison of the definition of continuity of f at x_0, and the definition of the limit of f at x_0. ∎

Before we can prove the basic properties of continuous functions, we must prove some elementary properties of limits. First we need a definition. A **product** on $E \times F \to G$ is a mapping denoted by $(u, v) \mapsto uv$ that is bilinear and satisfies $\|uv\|_G \le \|u\|_E \|v\|_F$. For example, using the Cauchy-Schwartz inequality, we see that the usual inner product on $\mathbb{R}^n$ is just a product on $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

Example 12.4 We say that a function $f\colon S \to F$ is **bounded** if there exists $M > 0$ such that $\|f(x)\| \le M$ for all $x \in S$. Now consider the space $E = \mathcal{B}(S, \mathbb{R})$ of real-valued bounded functions on any nonempty set S. Let us define a norm $\|\ \|_\infty$ on $\mathcal{B}(S, \mathbb{R})$ by

$$\|f\|_\infty = \sup_{x \in S} |f(x)|$$

for any $f \in \mathcal{B}(S, \mathbb{R})$. This important norm is called the **sup norm**. For any f, $g \in \mathcal{B}(S, \mathbb{R})$ suppose $\|f\|_\infty = C_1$ and $\|g\|_\infty = C_2$. Then it follows that $|f(x)| \le C_1$ and $|g(x)| \le C_2$ for all $x \in S$. But then for all $x \in S$ we have

$$|f(x)g(x)| = |f(x)|\,|g(x)| \le C_1C_2 = \|f\|_\infty \|g\|_\infty$$

so that the usual product of (real-valued) functions is also bounded. Therefore we see that

$$\|fg\|_\infty \le \|f\|_\infty \|g\|_\infty$$

and since the usual product is obviously bilinear, we have a (general) product on $E \times E \to E$. ∕∕

With the notion of a product carefully defined, we can repeat parts (a) – (c) of Theorem B2 in a more general form as follows. The proof is virtually identical to that of Theorem B2 except that here we replace the absolute value by the norm.

Theorem 12.4 (a) Let $u_n \to u$ and $v_n \to v$ be convergent sequences of vectors in E. Then $\lim_{n\to\infty}(u_n + v_n) = u + v$.

(b) Let $v_n \to v$ be a convergent sequence of vectors in E, and let c be a scalar. Then $\lim_{n\to\infty}(cv_n) = cv$.

(c) Let $u_n \to u$ and $v_n \to v$ be convergent sequences of vectors in E and F respectively, and let $E \times F \to G$ be a product. Then $\lim_{n\to\infty}(u_n v_n) = uv$.

Theorem 12.5 (a) Suppose that $A \subset E$ and v is an accumulation point of A. Let f and g be mappings of A into F, and assume that $\lim_{u\to v} f(u) = w_1$ and $\lim_{u\to v} g(u) = w_2$. Then

$$\lim_{u\to v}(f + g)(u) = w_1 + w_2 \ .$$

(b) Let A be a subset of some normed space, and let v be an accumulation point of A. Let $f\colon A \to E$ and $g\colon A \to F$ be mappings, and assume further that $\lim_{u\to v} f(u) = w_1$ and $\lim_{u\to v} g(u) = w_2$. If $E \times F \to G$ is a product, then

$$\lim_{u\to v} f(u)g(u) = w_1 w_2 \ .$$

Proof (a) Given $\varepsilon > 0$, there exists $\delta_1 > 0$ such that if $u \in A$ with $\|u - v\| < \delta_1$ then $|f(u) - w_1| < \varepsilon/2$. Similarly, there exists $\delta_2 > 0$ such that $\|u - v\| < \delta_2$ implies $|g(u) - w_2| < \varepsilon/2$. Choosing $\delta = \min\{\delta_1, \delta_2\}$ we see that if $u \in A$ and $\|u - v\| < \delta$ we have

$$\begin{aligned}\|(f+g)(u) - (w_1 + w_2)\| &= \|f(u) - w_1 + g(u) - w_2\| \\ &\le \|f(u) - w_1\| + \|g(u) - w_2\| \\ &< \varepsilon/2 + \varepsilon/2 \\ &= \varepsilon \ .\end{aligned}$$

(b) Given $\varepsilon > 0$, there exists $\delta_1 > 0$ such that $\|u - v\| < \delta_1$ implies

$$\|f(u) - w_1\| < \varepsilon/[2(1 + \|w_2\|)] \ .$$

Similarly, there exists $\delta_2 > 0$ such that $\|u - v\| < \delta_2$ implies

$$\|g(u) - w_2\| < \varepsilon/[2(1 + \|w_1\|)] \ .$$

From the definition of limit, given $\varepsilon = 1$ there exists $\delta_3 > 0$ such that $\|u - v\| < \delta_3$ implies

$$\|f(u) - w_1\| < 1 \; .$$

But from Example 2.11 we see that

$$\|f(u)\| - \|w_1\| \le \|f(u) - w_1\| < 1$$

which implies

$$\|f(u)\| < 1 + \|w_1\| \; .$$

If we let $\delta = \min\{\delta_1, \delta_2, \delta_3\}$, then for all $u \in A$ with $\|u - v\| < \delta$ we have

$$\begin{aligned}\|f(u)g(u) - w_1w_2\| &= \|f(u)g(u) - f(u)w_2 + f(u)w_2 - w_1w_2\|\\ &\le \|f(u)[g(u) - w_2]\| + \|[f(u) - w_1]w_2\|\\ &\le \|f(u)\|\,\|g(u) - w_2\| + \|f(u) - w_1\|\,\|w_2\|\\ &< (1 + \|w_1\|)\varepsilon/[2(1 + \|w_1\|)] + \varepsilon/[2(1 + \|w_2\|)]\,\|w_2\|\\ &< \varepsilon/2 + \varepsilon/2\\ &= \varepsilon \; . \; ∎\end{aligned}$$

The reader should realize that the norms used in the last proof are not defined on the same normed space. However, it would have been too cluttered for us to distinguish between them, and this practice is usually followed by most authors.

It will also be of use to formulate the limit of a composition of mappings.

Theorem 12.6 Suppose $A \subset E$ and $B \subset F$, and let $f\colon A \to B$ and $g\colon B \to G$ be mappings. Assume that u is an accumulation point of A and that $\lim_{x\to u} f(x) = v$. Assume also that v is an accumulation point of B and that $\lim_{y\to v} g(y) = w$. Then

$$\lim_{x\to u} (g \circ f)(x) = \lim_{x\to u} g(f(x)) = w \; .$$

Proof Given $\varepsilon > 0$, there exists $\delta_1 > 0$ such that for all $y \in B$ with $\|y - v\| < \delta_1$, we have $\|g(y) - w\| < \varepsilon$. Then given this δ_1, there exists $\delta_2 > 0$ such that for all $x \in A$ with $\|x - u\| < \delta_2$, we have $\|f(x) - v\| < \delta_1$. But now letting $y = f(x)$, we see that for such an $x \in A$ we must have $\|g(f(x)) - w\| < \varepsilon$. ∎

We are now in a position to prove the basic properties of continuous functions on a normed space.

Theorem 12.7 (a) If $A \subset E$ and $f, g: A \to F$ are continuous at $v \in A$, then the sum $f + g$ is continuous at v.

(b) Let $f: A \to E$ and $g: A \to F$ be continuous at $v \in A$ and suppose that $E \times F \to G$ is a product. Then the product map fg is continuous at v.

(c) Suppose $A \subset E$, $B \subset F$, and let $f: A \to B$ and $g: B \to G$ be mappings. Assume that f is continuous at $v \in A$ with $f(v) = w$, and assume that g is continuous at w. Then the composite map $g \circ f$ is continuous at v.

(d) A mapping $f: E \to F$ is continuous at v if and only if for every sequence $\{v_n\}$ in E we have $v_n \to v$ implies $f(v_n) \to f(v)$. In other words, we have $\lim f(v_n) = f(\lim v_n)$.

Proof (a) If v is an isolated point there is nothing to prove, so assume that v is an accumulation point. Then by Theorems 12.5 and 12.3 we have

$$\begin{aligned}\lim_{u\to v}(f+g)(u) &= \lim_{u\to v} f(u) + \lim_{u\to v} g(u)\\ &= f(v)+g(v)\\ &= (f+g)(v)\ .\end{aligned}$$

(b) This also follows from Theorems 12.5 and 12.3.

(c) Left to the reader (see Exercise 12.1.3).

(d) We first assume that f is continuous at v, and that the sequence $\{v_n\}$ converges to v. We must show that $f(v_n) \to f(v)$. Now, since f is continuous, we know that given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|u - v\| < \delta$ implies $\|f(u) - f(v)\| < \varepsilon$. Furthermore, the convergence of $\{v_n\}$ means that given $\delta > 0$, there exists N such that $\|v_n - v\| < \delta$ for every $n \geq N$. Therefore, for every $n \geq N$ we have $\|v_n - v\| < \delta$ implies $\|f(v_n) - f(v)\| < \varepsilon$.

We now prove that if f is not continuous at v, then there exists a convergent sequence $v_n \to v$ for which $f(v_n) \not\to f(v)$. It will be notationally simpler for us to formulate this proof in terms of open balls defined by the induced metric (see Appendix A). If f is not continuous at v, then there exists $B(f(v), \varepsilon)$ with no corresponding $B(v, \delta)$ such that $f(B(v, \delta)) \subset B(f(v), \varepsilon)$. Consider the sequence of open balls $\{B(v, 1/n)\}$ for $n = 1, 2, \dots$. Since f is not continuous, we can find $v_n \in B(v, 1/n)$ such that $f(v_n) \notin B(f(v), \varepsilon)$. It is clear that the sequence $\{v_n\}$ converges to v (given ε, choose $n \geq N = 1/\varepsilon$), but that $f(v_n)$ does not converge to f(v) since by construction $B(f(v), \varepsilon)$ contains none of the $f(v_n)$. ∎

Since the notion of open sets is extremely important in much of what follows, it is natural to wonder whether different norms defined on a space lead to different open sets (through their induced metrics). We shall say that two norms $\| \|_1$ and $\| \|_2$ defined on E are **equivalent** if there exists a number C such that for all $u \in E$ we have

$$C^{-1}\|u\|_1 \le \|u\|_2 \le C\|u\|_1 \ .$$

We leave it to the reader to show that this defines an equivalence relation on the set of all norms on E (see Exercise 12.1.4).

Example 12.5 It is easy to see that this definition does exactly what we want it to do. For example, suppose $U \subset E$ is open relative to a norm $\| \|_1$. This means that for any $u \in U$, there exists $\varepsilon_1 > 0$ such that $\|u - v\|_1 < \varepsilon_1$ implies $v \in U$. We would like to show that given an equivalent norm $\| \|_2$, then there exists $\varepsilon_2 > 0$ such that $\|u - v\|_2 < \varepsilon_2$ implies $v \in U$. We know there exists $C > 0$ such that $C^{-1}\| \|_1 \le \| \|_2 \le C \| \|_1$, and hence choosing $\varepsilon_2 = \varepsilon_1/C$, it follows that for all $v \in E$ with $\|u - v\|_2 < \varepsilon_2$ we have

$$\|u - v\|_1 \le C \|u - v\|_2 < C\varepsilon_2 = \varepsilon_1$$

so that $v \in U$. Therefore we have shown that if a set is open with respect to one norm, then it is open with respect to any equivalent norm. ∕∕

Example 12.6 It is also easy to give an example of two non-equivalent norms defined on a space. To see this, consider the space E of all real-valued continuous functions defined on [0, 1]. We define a norm on E by means of the scalar product. Thus for any $f, g \in E$ we define the scalar product by

$$\langle f, g\rangle = \int_0^1 f(x)g(x)\,dx$$

and the associated norm by $\|f\|_2 = \langle f, f\rangle^{1/2}$. This norm is usually called the **L^2-norm**. Alternatively, we note that any continuous real function defined on [0, 1] must be bounded (Theorems A8 and A14). Hence we may also define the **sup norm** $\|f\|_\infty$ by

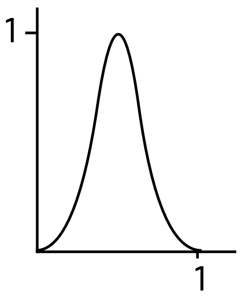
$$\|f\|_\infty = \sup |f(x)|$$

where the sup is taken over all $x \in [0, 1]$.

Now suppose $f \in E$ and write $\|f\|_\infty = C$. Then we have

$$\int_0^1 [f(x)]^2\,dx \le \int_0^1 C^2\,dx = C^2$$

and hence $\|f\|_2 \le \|f\|_\infty$. However, this is only half of the inequalities required by the definition. Consider the peaked function defined on [0, 1] by



If we let this function become arbitrarily narrow while maintaining the height, it is clear that the sup norm will always be equal to 1, but that the L^2-norm can be made arbitrarily small. ∕∕

The source of the problem that arose in this example is a result of the fact that the space E of continuous functions defined on [0, 1] is infinite-dimensional. In fact, we will soon prove that this can not occur in finite-dimensional spaces. In other words, we will see that all norms are equivalent in a finite-dimensional space.

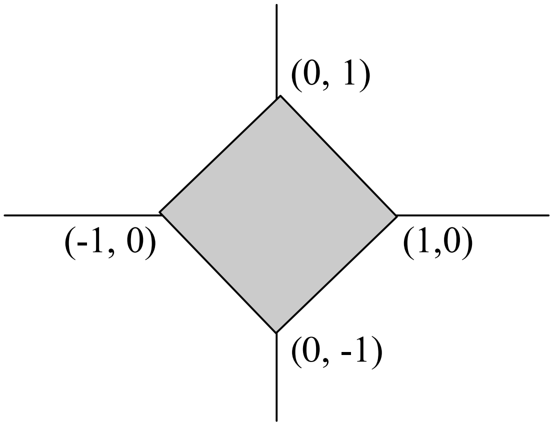
The reader may wonder whether or not the limits we have defined depend in any way on the particular norm being used. It is easy to show that if the limit of a sequence exists with respect to one norm, then it exists with respect to any other equivalent norm, and in fact the limits are equal (see Exercise 12.1.5). It should now be clear that a function that is continuous at a point v with respect to a norm $\| \|_1$ is also continuous at v with respect to any equivalent norm $\| \|_2$.

Now recall from Appendix B that a metric space in which every Cauchy sequence converges to a point in the space is said to be **complete**. It was also shown there that the space $\mathbb{R}^n$ is complete with respect to the standard norm (Theorem B8), and hence so is the space $\mathbb{C}^n$ (since $\mathbb{C}^n$ may be thought of as $\mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$). Recall also that a **Banach space** is a normed vector space $(E, \| \|)$ that is complete as a metric space (where as usual, the metric is that induced by the norm). If an inner product space $(E, \langle\ ,\ \rangle)$ is complete as a metric space (again with the metric defined by the norm induced by the inner product), then E is called a **Hilbert space**.

It is natural to wonder whether a space that is complete relative to one norm is necessarily complete relative to any other equivalent norm. This is answered by the next theorem. In the proof that follows, it will be convenient to use the **nonstandard norm** $\| \|_N$ defined on $\mathbb{R}^n$ (or $\mathbb{C}^n$) by

$$\|(u^1, \dots, u^n)\|_N = \sum_{i=1}^{n} |u^i|$$

where $(u^1, \ldots, u^n)$ is a vector n-tuple in $\mathbb{R}^n$ (or $\mathbb{C}^n$). In $\mathbb{R}^2$, the unit ball $\{(u_1, u_2): \|(u^1, u^2)\|_N \leq 1\}$ looks like



Theorem 12.8 Let E be a finite-dimensional vector space over either $\mathbb{R}$ or $\mathbb{C}$. Then

(a) There exists a norm on E.
(b) All norms on E are equivalent.
(c) All norms on E are complete.

Proof Let $\{e_1, \ldots, e_n\}$ be a basis for E so that any $u \in E$ may be written as $u = \Sigma u^i e_i$.

(a) We define the norm $\| \ \|_1$ on E by

$$\|u\|_1 = \sum_{i=1}^{n} |u^i| \ .$$

Properties (N1) and (N2) are trivial to verify, and if $v = \Sigma v^i e_i$ is any other vector in E, then $u + v = \Sigma(u^i + v^i)e_i$, and hence

$$\begin{aligned} \|u+v\|_1 &= \Sigma\,|\,u^i + v^i\,| \leq \Sigma(|\,u^i\,|+|\,v^i\,|) \ = \Sigma|\,u^i\,|+\Sigma|\,v^i\,| \\ &= \|u\|_1 + \|v\|_1 \end{aligned}$$

so that (N3) is also satisfied.

This norm is quite convenient for a number of purposes. Note that it yields the same result for any $v = \Sigma v^i e_i \in E$ as does the nonstandard norm $\| \ \|_N$ for the corresponding $(v^1, \ldots, v^n) \in \mathbb{R}^n$ (or $\mathbb{C}^n$).

(b) Let $\| \ \|_2$ be any other norm on E, and let $u, v \in E$ be arbitrary. Using Example 2.11 and properties (N3) and (N2), we see that

$$\begin{aligned}\big|\|u\|_2 - \|v\|_2\big| &\le \|u - v\|_2 \\ &= \|\Sigma(u^i - v^i)e_i\|_2 \\ &\le \Sigma\,\|(u^i - v^i)e_i\|_2 \\ &= \Sigma|u^i - v^i|\,\|e_i\|_2 \\ &\le \max_{1\le i\le n}\{\|e_i\|_2\}\Sigma|u^i - v^i| \\ &= \max_{1\le i\le n}\{\|e_i\|_2\}\|(u^1, \dots, u^n) - (v^1, \dots, v^n)\|_N\end{aligned} \qquad (*)$$

Define the mapping f: $\mathbb{C}^n \to \mathbb{R}$ by $x = (x^1, \dots, x^n) \mapsto \|\Sigma x^i e_i\|_2 \in [0, \infty)$. To say that f is uniformly continuous on $\mathbb{C}^n$ with respect to the norm $\|\ \|_N$ means that given $\varepsilon > 0$, we can find $\delta > 0$ such that for all $x, y \in \mathbb{C}^n$ with

$$\|x - y\|_N = \|(x^1, \dots, x^n) - (y^1, \dots, y^n)\|_N < \delta$$

we have

$$|f(x) - f(y)| = |\,\|\Sigma x^i e_i\|_2 - \|\Sigma y^i e_i\|_2\,| < \varepsilon\ .$$

If we define $B = \max_{1\le i\le n}\{\|e_i\|_2\}$, then choosing $\delta = \varepsilon/B$, we see (*) shows that f is (uniformly) continuous with respect to the norm $\|\ \|_N$ on $\mathbb{C}^n$.

We now note that the unit sphere $S = \{x \in \mathbb{C}^n: \|x\|_N = 1\}$ is closed and bounded, and hence S is compact (Theorem A14). The restriction of the function f to S is then continuous and strictly positive (by (N1)), and hence according to Theorem A15, f attains both its minimum m and maximum M on S. In other words, for every $x = (x^1, \dots, x^n) \in \mathbb{C}^n$ with $\|x\|_N = 1$ we have $0 < m \le \|\Sigma x^i e_i\|_2 \le M$. Since $\|x\|_N = \|\,(x^1, \dots, x^n)\,\|_N = 1$, we may write

$$m\|(x^1, \dots, x^n)\|_N \le \|\Sigma x^i e_i\|_2 \le M\|(x^1, \dots, x^n)\|_N$$

or, using part (a) with $u = \Sigma x^i e_i \in E$, we find that

$$m\|u\|_1 \le \|u\|_2 \le M\|u\|_1\ .$$

Choosing $C = \max\{1/m, M\}$, we see that $\|\ \|_1$ and $\|\ \|_2$ are equivalent. The fact that $\|\ \|_2$ was arbitrary combined with the fact that equivalent norms form an equivalence class completes the proof that all norms on E are equivalent.

(c) It suffices to show that E is complete with respect to any particular norm on E. This is because part (b) together with the fact that a sequence that converges with respect to one norm must converge with respect to any equivalent norm then shows that E will be complete with respect to any norm.

Recall from the corollary to Theorem 2.8 that E is isomorphic to either $\mathbb{R}^n$ or $\mathbb{C}^n$. The result then follows from Theorem B8 or its obvious extension to $\mathbb{C}^n$ (see Exercise 12.1.6). ∎

We shall see that closed subspaces play an important role in the theory of Hilbert spaces. Because of this, we must make some simple observations. Suppose that Y is a closed subset of a complete space (X, d), and let $\{x_n\}$ be a Cauchy sequence in Y. Then $\{x_n\}$ is also obviously a Cauchy sequence in X, and hence $x_n \to x \in X$. But this means that $x \in \mathrm{Cl}\, Y = Y$ (Theorem B13(b) or B14(a)) so that $\{x_n\}$ converges in Y.

On the other hand, suppose that Y is a complete subset of an arbitrary metric space (X, d) and let $\{x_n\}$ be any sequence in Y that converges to an element $x \in X$. We claim that in fact $x \in Y$ which will prove that Y is closed (Theorem B14(a)). Since $x_n \to x \in X$, it follows that $\{x_n\}$ is a Cauchy sequence in X (since any convergent sequence is necessarily Cauchy). In other words, given $\varepsilon > 0$ there exists $N > 0$ such that $m, n \geq N$ implies $\|x_m - x_n\| < \varepsilon$. But then $\{x_n\}$ is just a Cauchy sequence in Y (which is complete), and hence $x_n \to x \in Y$.

This discussion proves the next result.

Theorem 12.9 Any closed subset of a complete metric space is also a complete metric space. On the other hand, if a subset of an arbitrary metric space is complete, then it is closed.

Corollary A finite-dimensional subspace of any real or complex normed vector space is closed.

Proof This follows from Theorems 12.8 and 12.9. ∎

Now suppose that $A \subset E$ and that we have a mapping $f: A \to F$ where $F = F_1 \times \cdots \times F_n$ is the Cartesian product of normed spaces. Then for any $v \in A$ we have $f(v) = (f_1(v), \ldots, f_n(v))$ where each $f_i: A \to F_i$ is called the *ith* **coordinate function** of f. In other words, we write $f = (f_1, \ldots, f_n)$.

If $w = (w_1, \ldots, w_n) \in F$, then one possible norm on F, the **sup norm**, is defined by

$$\|w\| = \sup_{1\leq i\leq n}\{\|w_i\|\}$$

where $\|w_i\|$ denotes the norm in F_i. However, this is not the only possible norm. Recall that if $x = (x_1, \ldots, x_n) \in \mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$, then the standard norm in $\mathbb{R}^n$ is given by $\|x\|^2 = \sum |x_i|^2$. The analogous "Pythagorean" norm $\|\ \|_p$ on F would then be defined by $\|w\|_p{}^2 = \sum \|w_i\|^2$. Alternatively, we could also define the "nonstandard" norm $\|w\|_N = \sum \|w_i\|$. We leave it to the reader to show that these three norms are equivalent on F (see Exercise 12.1.7).

The next result should come as no surprise. Note that we will use the sup norm on F as defined above.

Theorem 12.10 Suppose $A \subset E$ and let $f: A \to F = F_1 \times \cdots \times F_n$ be a mapping. If v is an accumulation point of A, then $\lim_{u\to v} f(u)$ exists if and only if $\lim_{u\to v} f_i(u)$ exists for each $i = 1, \ldots, n$. If this is the case and if $\lim_{u\to v} f(u) = w = (w_1, \ldots, w_n)$, then $\lim_{u\to v} f_i(u) = w_i$ for each $i = 1, \ldots, n$.

Proof First assume that $\lim_{u\to v} f(u) = w = (w_1, \ldots, w_n)$. This means that given $\varepsilon > 0$, there exists δ such that $\|u - v\| < \delta$ implies $\|f(u) - w\| < \varepsilon$. If we write $f(u) = (f_i(u), \ldots, f_n(u))$, then for all $u \in A$ with $\|u - v\| < \delta$, the definition of sup norm tells us that

$$\|f_i(u) - w_i\| \le \|f(u) - w\| < \varepsilon \ .$$

This proves that $\lim_{u\to v} f_i(u) = w_i$.

Conversely, suppose $\lim_{u\to v} f_i(u) = w_i$ for each $i = 1, \ldots, n$. Then given $\varepsilon > 0$, there exists δ_i such that $\|u - v\| < \delta_i$ implies $\|f_i(u) - w_i\| < \varepsilon$. Defining $\delta = \min\{\delta_i\}$, we see that for all $u \in A$ with $\|u - v\| < \delta$, we have $\|f_i(u) - w_i\| < \varepsilon$ for each $i = 1, \ldots\ , n$ and therefore

$$\|f(u) - w\| = \sup\{\|f_i(u) - w_i\|\} < \varepsilon \ .$$

This shows that $\lim_{u\to v} f(u) = w$. ∎

Corollary The mapping f defined in Theorem 12.10 is continuous if and only if each f_i is continuous.

Proof Obvious from the definitions. ∎

Exercises

1. If $u, v \in (E, \langle\ ,\ \rangle)$ prove:
   (a) The parallelogram law: $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$.
   (b) The Pythagorean theorem: $\|u + v\|^2 = \|u\|^2 + \|v\|^2$ if $u \perp v$.

2. Show that an inner product on E is continuous in both variables by showing that $\lim_{y\to y_0} \langle x, y\rangle = \langle x, y_0\rangle$ and $\lim_{x\to x_0} \langle x, y\rangle = \langle x_0, y\rangle$.

3. Prove part (c) of Theorem 12.7.

4. Show that equivalent norms define an equivalence relation on the set of all norms on E.

5. (a) Suppose $\{v_n\}$ is a sequence in a normed space E. Show that if $v_n \to v$ with respect to one norm on E, then $v_n \to v$ with respect to any equivalent norm on E.
   (b) Show that if a function is continuous at a point with respect to one norm, then it is continuous at that point with respect to any equivalent norm.

6. Fill in the details in the proof of Theorem 12.8(c).

7. Let $F = F_1 \times \cdots \times F_n$ be a Cartesian product of normed spaces, and suppose $w = (w_1, \ldots, w_n) \in F$. If $\|w_i\|$ denotes the norm on F_i, show that the norms $\|w\| = \sup_{1\leq i\leq n}\{\|w_i\|\}$, $\|w\|_p{}^2 = \sum_{i=1}^{n}\|w_i\|^2$ and $\|w\|_N = \sum_{i=1}^{n}\|w_i\|$ are equivalent on F.

8. Show that the set $\mathcal{B}(S, E)$ of all bounded functions from a nonempty set S to a normed vector space E forms a vector space (over the same field as E).

12.2 OPERATOR NORMS

Suppose E and F are normed spaces, and let $A: E \to F$ be a linear map. If there exists a number $M > 0$ such that $\|Av\|_F \leq M\|v\|_E$ for all $v \in E$, then A is said to be **bounded**, and the number M is called a **bound** for A. In other words, to say that A is bounded means that it takes bounded values on bounded sets. Note that we labeled our norms in a way that denotes which space they are defined on. From now on though, we shall not complicate our notation by this designation unless it is necessary. However, the reader should be careful to note that the symbol $\| \ \|$ may mean two different things within a single equation.

Recall also that a linear map $A: E \to F$ is said to be **continuous** at $v_0 \in E$ if given $\varepsilon > 0$, there exists $\delta > 0$ such that for all $v \in E$ with $\|v_0 - v\| < \delta$, we have $\|Av_0 - Av\| = \|A(v_0 - v)\| < \varepsilon$.

Theorem 12.11 Let E be a finite-dimensional normed space, and let A: E → F be a linear map of E into a normed space F (not necessarily finite-dimensional). Then A is bounded.

Proof Let $\{e_1, \dots, e_n\}$ be a basis for E so that any $v \in E$ may be written in the form $v = \sum v^i e_i$. Using the defining properties of the norm and the linearity of A, we then have

$$\|Av\| = \|A\textstyle\sum v^i e_i\| = \|\sum v^i Ae_i\| \le \sum \|v^i Ae_i\| = \sum |v^i|\,\|Ae_i\| \ .$$

Since all norms on E are equivalent (Theorem 12.8), we use the norm $\|\ \|_1$ defined by $\|v\|_1 = \sum|v^i|$. Thus any other norm $\|\ \|_2$ on E will be related to $\|\ \|_1$ by $C^{-1}\|\ \|_2 \le \|\ \|_1 \le C\|\ \|_2$ for some number C. Since $\|Ae_i\| < \infty$ for each i, we define the real number $M = \max\{\|Ae_i\|\}$. Then

$$\|Av\| \le M\textstyle\sum|v^i| = M\|v\|_1 \le MC\|v\|_2 \ . \ ∎$$

Our next result is quite fundamental, and will be referred to again several times.

Theorem 12.12 Let A: E → F be a linear map of normed spaces. If A is bounded, then A is uniformly continuous, and if A is continuous at 0 then A is bounded.

Proof If A is bounded, then there exists $M > 0$ such that $\|Av\| \le M\|v\|$ for every $v \in E$. Then for any $\varepsilon > 0$, we choose $\delta = \varepsilon/M$ so that for all $u, v \in E$ with $\|u - v\| < \delta$, we have

$$\|Au - Av\| = \|A(u - v)\| \le M\|u - v\| < \varepsilon \ .$$

This proves that A is uniformly continuous.

Conversely, suppose A is continuous at $0 \in E$. Then given $\varepsilon = 1$, there exists $\delta > 0$ such that $\|v\| < \delta$ implies $\|Av\| < \varepsilon = 1$. In particular, we see that for any nonzero $v \in E$ we have $\|\delta v/(2\|v\|)\| = \delta/2 < \delta$ which implies $\|A(\delta v/(2\|v\|))\| < 1$. Taking out the constants yields $\|Av\| < (2/\delta)\|v\|$. This shows that A is bounded with bound $M = 2/\delta$. ∎

As shown in Exercise 12.2.1, there is nothing special about the continuity of A at the origin.

Corollary 1 A linear map A: E → F is bounded if and only if A is continuous.

Proof Obvious. ∎

Corollary 2 Let E be finite-dimensional, and let A: E → F be a linear map. Then A is uniformly continuous.

Proof This follows directly from Theorems 12.11 and 12.12. ∎

Let E and F be normed spaces, and let A: E → F be a continuous (and hence bounded) linear map (if E is finite-dimensional, then the continuity requirement is redundant). We define the **operator norm** of A by

$$\begin{aligned}\|A\| &= \sup\{\|Av\|/\|v\|: v \in E,\ v \neq 0\}\\ &= \sup\{\|Av\|: \|v\| = 1\}\ .\end{aligned}$$

If $\|v\| \leq 1$, then we may write $v = c\hat{v}$ where $\|\hat{v}\| = 1$ and $|c| \leq 1$. Then $\|Av\| = |c|\ \|A\hat{v}\| \leq \|A\hat{v}\|$ and therefore, since we are using the sup, an equivalent definition of $\|A\|$ is

$$\|A\| \;=\; \sup\{\|Av\|: \|v\| \leq 1\}\ .$$

From the first definition, we see that for any $v \in E$ we have $\|Av\|/\|v\| \leq \|A\|$, and hence we have the important result

$$\|Av\| \;\leq\; \|A\|\ \|v\|\ .$$

This shows that another equivalent definition of $\|A\|$ is

$$\|A\| \;=\; \inf\{M > 0: \|Av\| \leq M\|v\| \text{ for all } v \in E\}\ .$$

Another useful result follows by noting that if A: E → F and B: F → G, then for any $v \in E$ we have

$$\|(B \circ A)v\| \;=\; \|B(Av)\ \| \;\leq\; \|B\|\ \|Av\| \;\leq\; \|B\|\ \|A\|\ \|v\|$$

and hence from the definition of the operator norm we have

$$\|B \circ A\| \;\leq\; \|B\|\ \|A\|\ .$$

We denote the space of all continuous linear maps from E to F by $\mathcal{L}$(E, F). That $\mathcal{L}$(E, F) is in fact a vector space will be shown below. Since for any A ∈ $\mathcal{L}$(E, F) we have $\|A\| = \sup\{\|Av\|: \|v\| \le 1\}$, we see that by restricting A to the unit ball in E, the space $\mathcal{L}$(E, F) is just a subspace of the space $\mathcal{B}$(S, F) of all bounded maps from S into F that was defined in Example 12.4 (where S is just the unit ball in E).

Theorem 12.13 The space $\mathcal{L}$(E, F) with the operator norm is a normed vector space. Moreover, if F is a Banach space, then so is $\mathcal{L}$(E, F).

Proof Suppose that A ∈ $\mathcal{L}$(E, F). We first verify requirements (N1) – (N3) for a norm. From the definitions, it is obvious that $\|A\| \ge 0$ and $\|0\| = 0$. In addition, if $\|A\| = 0$ then for any v ∈ E we have $\|Av\| \le \|A\|\,\|v\| = 0$ which implies that A = 0. This verifies (N1). If $c \in \mathcal{F}$, then

$$\begin{aligned}\|cA\| &= \sup\{\|(cA)v\|: \|v\| \le 1\}\\ &= |c|\sup\{\|Av\|: \|v\| = 1\}\\ &= |c|\,\|A\|\end{aligned}$$

which verifies (N2). Now let A, B ∈ $\mathcal{L}$(E, F). Then using Theorem 0.5 we see that (leaving out the restriction on $\|v\|$)

$$\begin{aligned}\|A+B\| &= \sup\{\|(A+B)v\|\}\\ &= \sup\{\|Av+Bv\|\}\\ &\le \sup\{\|Av\|+\|\mathrm{Bv}\|\}\\ &= \sup\{\|Av\|\}+\sup\{\|Bv\|\}\\ &= \|A\|+\|B\|\end{aligned}$$

which proves (N3). That $\mathcal{L}$(E, F) is in fact a vector space follows from Theorem 12.7(a) and (b).

Now suppose that F is a Banach space and let $\{A_n\}$ be a Cauchy sequence in $\mathcal{L}$(E, F). This means that for every ε > 0 there exists N such that m, n ≥ N implies $\|A_m - A_n\| < \varepsilon$. In particular, for any v ∈ E and ε > 0, there exists N such that for all m, n ≥ N we have

$$\|A_m v - A_n v\| = \|(A_m - A_n)v\| \le \|A_m - A_n\|\,\|v\| < (\varepsilon/\|v\|)\|v\| = \varepsilon$$

so that $\{A_n v\}$ is a Cauchy sequence in F. Since F is a Banach space this sequence converges, and hence we define Av ∈ F by

$$Av = \lim_{n\to\infty} A_n v\ .$$

This defines a map $A: E \to F$. Since each A_n is linear, it should be clear (using Theorem 12.4) that A is linear. We must still show that A is continuous (so that $A \in \mathcal{L}(E, F)$) and that $A_n \to A$.

Given $\varepsilon > 0$, there exists N such that $\|A_m - A_n\| < \varepsilon$ for all $m, n \geq N$. If $v \in E$ is such that $\|v\| \leq 1$, then

$$\|A_m v - A_n v\| \leq \|A_m - A_n\| \, \|v\| < \varepsilon \ .$$

But $\{A_m v\}$ converges to Av, and hence letting $m \to \infty$ yields

$$\|(A - A_n)v\| = \|Av - A_n v\| \leq \varepsilon$$

for every $v \in E$ with $\|v\| \leq 1$. This shows that $A - A_n$ is continuous at 0, and hence $A - A_n$ is in $\mathcal{L}(E, F)$ (by Theorem 12.12 and its Corollary 1). Thus $A \in \mathcal{L}(E, F)$ (since each A_n is). Finally, since $A - A_n \in \mathcal{L}(E, F)$, we may apply the definition of operator norm to obtain

$$\|A - A_n\| = \sup\{\|(A - A_n)v\| : \|v\| \leq 1\} \leq \varepsilon$$

for every $n \geq N$, and hence $A_n \to A$. ∎

Exercises

1. Let $A: E \to F$ be linear and continuous a some point $v_0 \in E$. Prove directly that A is continuous at every $v \in E$.

2. (**Linear Extension Theorem**) Let E be a normed vector space, F a subspace of E, and G a Banach space. Suppose $A \in \mathcal{L}(F, G)$ and assume that $\|A\| = M$. Prove:
(a) The closure $\bar{F}$ of F in E is a subspace of E.
(b) There exists a unique extension $\bar{A} \in \mathcal{L}(\bar{F}, G)$ of A. [*Hint*: If $v \in \bar{F}$, then there exists $\{v_n\} \in F$ such that $v_n \to v$ (why?). Show that $\{Av_n\}$ is Cauchy in G, and converges to a unique limit that is independent of $\{v_n\}$. Define $\bar{A}v = \lim Av_n$, and show that $\bar{A}$ is linear. Also show that $\bar{A}v = Av$ for any $v \in F$. Next, show that $\bar{A}$ is bounded so that $\bar{A} \in \mathcal{L}(\bar{F}, G)$ (why?), and finally show that the $\bar{A}$ so defined is unique.]
(c) $\|\bar{A}\| = \|A\|$.

3. Let E be a normed vector space. A **completion** $(\bar{E}, A)$ of E is a Banach space $\bar{E}$ together with a continuous injective linear mapping $A\colon E \to \bar{E}$ that preserves the norm and is such that A(E) is dense in $\bar{E}$. Show that $(\bar{E}, A)$ is unique in the sense that if (F, B) is another completion of E, then there exists a unique invertible element $C \in \mathcal{L}(\bar{E}, F)$ such that $B = C \circ A$. [*Hint*: Apply the previous exercise to the mappings $B \circ A^{-1}$ and $A \circ B^{-1}$.]

12.3 HILBERT SPACES

Discussing infinite-dimensional vector spaces requires a certain amount of care that was not needed in our treatment of finite-dimensional spaces. For example, how are we to express an arbitrary vector as a linear combination of basis vectors? For that matter, how do we define a basis in an infinite-dimensional space? As another example, recall that in our treatment of operator adjoints, we restricted our discussion to finite-dimensional spaces (see Theorems 10.1 and 10.2). While we cannot define the adjoint in an arbitrary infinite-dimensional space (e.g., a Banach space), we shall see that it is nevertheless possible to make such a definition in a Hilbert space.

Unfortunately, a thorough treatment of Hilbert spaces requires a knowledge of rather advanced integration theory (i.e., the Lebesgue theory). However, it is quite easy to present a fairly complete discussion of many of the basic and important properties of Hilbert spaces without using the general theory of integration.

As in the previous sections of this chapter, we consider only vector spaces over the real and complex fields. In fact, unless otherwise noted we shall always assume that our scalars are complex numbers. Recall from Section 12.1 that a **Hilbert space** is an inner product space which is complete as a metric space. We shall generally denote a Hilbert space by the letter H.

To begin with, we recall that a linear space E is n-dimensional if it contains a set of n linearly independent vectors, but every set of n + 1 vectors is linearly dependent. If E contains n linearly independent vectors for every positive integer n, then we say that E is **infinite-dimensional**. Let us rephrase some of our earlier discussion of (infinite) series in a terminology that fits in with the concept of an infinite-dimensional space.

We have already seen that a sequence $\{v_n\}$ of vectors in a space $(E, \|\ \|)$ converges to $v \in E$ if for each $\varepsilon > 0$, there exists an N such that $n \geq N$ implies $\|v_n - v\| < \varepsilon$. We sometimes write this as $v_n \to v$ or $\|v_n - v\| \to 0$. Similarly, we say that an infinite linear combination $\sum_{k=1}^{\infty} a_k w_k$ of vectors in E **converges** if the sequence of partial sums $v_n = \sum_{k=1}^{n} a_k w_k$ converges. In other words, to write $v = \sum_{k=1}^{\infty} a_k w_k$ means that $v_n \to v$. If no explicit limits on the sum are given, then we assume that the sum is over an infinite number of vectors.

Just as we did with finite linear combinations, we define the addition of infinite linear combinations componentwise. Thus, if $x = \sum a_n v_n$ and $y = \sum b_n v_n$, then we define the sum $x + y$ by $x + y = \sum(a_n + b_n)v_n$. If $x_n = \sum_{k=1}^{n} a_k v_k$ converges to x, and $y_n = \sum_{k=1}^{n} b_k v_k$ converges to y, then it is quite easy to see that $x_n + y_n = \sum_{k=1}^{n}(a_k + b_k)v_k$ converges to $x + y$ (see Exercise 12.3.1). Furthermore, we define scalar multiplication of an infinite linear combination $x = \sum a_n v_n$ by $cx = \sum(ca_n)v_n$. It is also easy to see that if the n*th* partial sum x_n converges to x, then cx_n converges to cx.

In our next two examples we define the general Banach spaces l_p^n and l_p, and we then show that both l_2^n and l_2 may be made into Hilbert spaces. (In more advanced work, the space l_p may be generalized to include measure spaces.) Remember that our scalars may be either real or complex numbers.

Example 12.7 If p is any real number such that $1 \le p < \infty$, we let l_p^n denote the space of all scalar n-tuples $x = (x_1, \ldots, x_n)$ with the norm $\| \, \|_p$ defined by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \; .$$

We first show that this does indeed define a norm on l_p^n. Properties (N1) and (N2) of the norm are obvious, so it remains to show that property (N3) is also obeyed. To show this, we will prove two general results that are of importance in their own right. In the derivation to follow, if p occurs by itself, it is defined as above. If the numbers p and q occur together, then q is defined the same way as p, but we also assume that $1/p + 1/q = 1$. (If p and q satisfy the relation $1/p + 1/q = 1$, then p and q are said to be **conjugate exponents**. Note that in this case both p and q are strictly greater than 1.)

Let α and β be real numbers ≥ 0. We first show that

$$\alpha^{1/p}\beta^{1/q} \le \alpha/p + \beta/q \; . \qquad (1)$$

This result is clear if either α or β is zero, so assume that both α and β are greater than zero. For any real $k \in (0, 1)$ define the function f(t) for $t \ge 1$ by

$$f(t) = k(t - 1) - t^k + 1 \; .$$

From elementary calculus, we see that $f'(t) = k(1 - t^{k-1})$, and hence $f'(t) \ge 0$ for every $t \ge 1$ and $k \in (0, 1)$. Since $f(1) = 0$, this implies that $f(t) \ge 0$, and thus the definition of f(t) shows that

$$t^k \le k(t-1)+1 = kt+(1-k) \ .$$

If $\alpha \ge \beta$, we let $t = \alpha/\beta$ and $k = 1/p$ to obtain

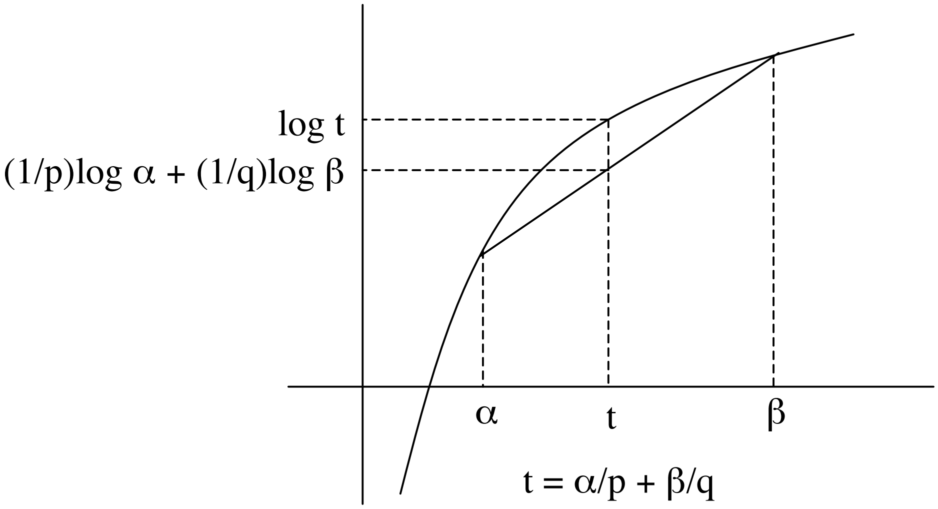
$$(\alpha/\beta)^{1/p} \le \alpha/\beta p + (1-1/p) = \alpha/\beta p + 1/q \ .$$

Multiplying through by β and using $\beta^{1-1/p} = \beta^{1/q}$ yields the desired result. Similarly, if $\alpha < \beta$ we let $t = \beta/\alpha$ and $k = 1/q$.

To help see the meaning of (1), note that taking the logarithm of both sides of (1) yields

$$\frac{1}{p}\log\alpha + \frac{1}{q}\log\beta \le \log\left(\frac{\alpha}{p}+\frac{\beta}{q}\right) \ .$$

The reader should recognize this as the statement that the logarithm is a “convex function” (see the figure below).



We now use (1) to prove **Hölder’s inequality**:

$$\sum_{i=1}^{n}|x_i y_i| \le \|x\|_p \|y\|_q \ .$$

Again, we assume that x and y are both nonzero. Define $\alpha_i = (|x_i|\,/\,\|x\|_p)^p$ and $\beta_i = (|y_i|\,/\,\|y\|_q)^q$. From (1) we see that

$$|x_i y_i|\,/\,(\|x\|_p\,\|y\|_q) \le \alpha_i/p + \beta_i/q \ .$$

Using the definition of $\|\ \|_p$, it follows that $\Sigma^n_{i=1}\alpha_i = 1$ and similarly for β_i. Hence summing the previous inequality over $i = 1, \ldots, n$ and using the fact that 1/p + 1/q = 1 yields Hölder’s inequality. We remark that the particular case of p = q = 2 yields

$$\sum_{i=1}^{n}|x_i y_i| \le \left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2}\left(\sum_{i=1}^{n}|y_i|^2\right)^{1/2}$$

which is called **Cauchy’s inequality**.

Finally, we use Hölder’s inequality to prove **Minkowski’s inequality**:

$$\|x + y\|_p \le \|x\|_p + \|y\|_p \quad .$$

If p = 1 this is obvious since $|x_i + y_i| \le |x_i| + |y_i|$, so we may assume that p > 1. In this case we have

$$\begin{aligned}(\|x+y\|_p)^p &= \sum_{i=1}^{n}|x_i + y_i|^p\\ &= \sum_{i=1}^{n}|x_i + y_i|\,|x_i + y_i|^{p-1} \qquad (2)\\ &\le \sum_{i=1}^{n}(|x_i| + |y_i|)|x_i + y_i|^{p-1} \quad .\end{aligned}$$

Using Hölder’s inequality with y_i replaced by $|x_i + y_i|^{p/q}$ results in the inequality

$$\sum_{i=1}^{n}|x_i|\,|x_i + y_i|^{p/q} \le \|x\|_p(\|x+y\|_p)^{p/q}$$

with a similar result if we interchange x_i and y_i. Since 1/p + 1/q = 1 implies p/q = p − 1, we now see that (2) yields

$$(\|x + y\|_p)^p \le (\|x\|_p + \|y\|_p)(\|x + y\|_p)^{p-1} \quad .$$

Dividing this by $(\|x + y\|_p)^{p-1}$ yields Minkowski’s inequality.

We now see that Minkowski’s inequality is just the requirement (N3), and thus we have shown that our norm on l_p^n is indeed a norm. Finally, it follows from Theorem 12.8 that l_p^n is complete and is thus a Banach space.

We now consider the particular case of l_2^n, and define an inner product in the expected manner by

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i{}^* y_i \quad .$$

Defining the norm by

$$\|x\| = \langle x,\ x\rangle^{1/2} = \left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2}$$

it is easy to see that l_2^n satisfies all of the requirements for a Hilbert space. ∕∕

In our next example, we generalize this result to the case of infinite-dimensional spaces.

Example 12.8 As in the previous example, let p be any real number such that $1 \le p < \infty$. We let l_p denote the space of all sequences $x = \{x_1, x_2, \dots\}$ of scalars such that $\sum_{k=1}^{\infty}|x_k|^p < \infty$, and we define a norm on l_p by

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

We must show that this definition also satisfies the properties of a norm, which means that we need only verify the not entirely obvious condition (N3).

From the previous example, we may write Minkowski's inequality for the space l_p^n in the form

$$\left(\sum_{k=1}^{n}|x_k+y_k|^p\right)^{1/p} \le \left(\sum_{k=1}^{n}|x_k|^p\right)^{1/p} + \left(\sum_{k=1}^{n}|y_k|^p\right)^{1/p} .$$

Now, if $x, y \in l_p$, then both $(\sum_{k=1}^{\infty}|x_k|^p)^{1/p}$ and $(\sum_{k=1}^{\infty}|y_k|^p)^{1/p}$ exist since they are convergent by definition of l_p. Hence taking the limit of Minkowski's inequality as $n \to \infty$ shows that this equation also applies to infinite series as well. (This requires the observation that the p*th* root is a continuous function so that, by Theorem 12.7(d), the limit may be taken inside the root.) In other words, the equation $\|x + y\|_p \le \|x\|_p + \|y\|_p$ also applies to the space l_p. This shows that our definition of a norm is satisfactory. It should also be clear that Hölder's inequality similarly applies to the space l_p.

It is more difficult to show that l_p is complete as a metric space. The origin of the problem is easily seen by referring to Theorems B3 and B8. In these theorems, we showed that a Cauchy sequence $\{x_k\}$ in $\mathbb{R}^n$ led to n distinct Cauchy sequences $\{x_k{}^j\}$ in $\mathbb{R}$, each of which then converged to a number x^j by the completeness of $\mathbb{R}$. This means that for each $j = 1, \dots, n$ there exists an N_j such that $|x_k{}^j - x^j| < \varepsilon/\sqrt{n}$ for all $k \ge N_j$. Letting $N = \max\{N_j\}$, we see that

$$\|x_k - x\|^2 = \sum_{j=1}^{n} |x_k{}^j - x^j|^2 < n(\varepsilon^2/n) = \varepsilon^2$$

for all k ≥ N, and hence $x_k \rightarrow x$. However, the case of l_p we cannot take the max of an infinite number of integers. To circumvent this problem we may proceed as follows.

To keep the notation as simple as possible and also consistent with most other authors, we let $x = \{x_1, x_2, \dots\}$ be an element of l_p with components x_i, and we let $\{x^{(n)}\}$ be a sequence in l_p. Thus, the k*th* component of the vector $x^{(n)} \in l_p$ is given by $x_k{}^{(n)}$. Note that this is the opposite of our notation in the finite-dimensional case.

Let $\{x^{(n)}\}$ be a Cauchy sequence in l_p. This means that for any ε > 0, there exists M > 0 such that m, n ≥ M implies $\|x^{(m)} - x^{(n)}\|_p < \varepsilon$. Then, exactly as in the finite-dimensional case, for any k = 1, 2, . . . we have

$$|x_k{}^{(m)} - x_k{}^{(n)}|^p \le \sum_{j=1}^{\infty} |x_j{}^{(m)} - x_j{}^{(n)}|^p = (\|x^{(m)} - x^{(n)}\|_p)^p < \varepsilon^p$$

and hence $|x_k{}^{(m)} - x_k{}^{(n)}| < \varepsilon$. Therefore, for each k the sequence $\{x_k{}^{(n)}\}$ of the k*th* component forms a Cauchy sequence. Since ℝ (or ℂ) is complete, these sequences converge to a number which we denote by x_k. In other words, for every k we have

$$\lim_{n\rightarrow\infty} x_k{}^{(n)} = x_k \ .$$

To show that l_p is complete, we will show that the sequence $x = \{x_k\}$ is an element of l_p, and that in fact $\|x^{(n)} \rightarrow x\|_p \rightarrow 0$.

Using Minkowski's inequality, we have for every N and any n,

$$\begin{aligned}\left(\sum_{k=1}^{N} |x_k|^p\right)^{1/p} &= \left(\sum_{k=1}^{N} |x_k - x_k{}^{(n)} + x_k{}^{(n)}|^p\right)^{1/p} \\ &\le \left(\sum_{k=1}^{N} |x_k - x_k{}^{(n)}|^p\right)^{1/p} + \left(\sum_{k=1}^{N} |x_k{}^{(n)}|^p\right)^{1/p} \qquad (3) \\ &\le \left(\sum_{k=1}^{N} |x_k - x_k{}^{(n)}|^p\right)^{1/p} + \|x^{(n)}\|_p \ .\end{aligned}$$

Now write the n*th* term of the sequence $\{x^{(n)}\}$ as $x^{(n)} - x^{(m)} + x^{(m)}$ to obtain

$$\|x^{(n)}\|_p = \|x^{(n)} - x^{(m)} + x^{(m)}\|_p \le \|x^{(n)} - x^{(m)}\|_p + \|x^{(m)}\|_p \ .$$

Since $\{x^{(n)}\}$ is a Cauchy sequence , we know that given any $\varepsilon > 0$, there exists M_0 such that $m, n \ge M_0$ implies $\|x^{(n)} - x^{(m)}\|_p < \varepsilon$. Thus for any fixed $m \ge M_0$, the set $\{\|x^{(n)}\|_p\colon n \ge M_0\}$ of real numbers is bounded by $\varepsilon + \|x^{(m)}\|_p$. Moreover, we may take the max of the (finite) set of all $\|x^{(n)}\|_p$ with $n < M_0$. In other words, we have shown that the norms of every term in any Cauchy sequence are bounded, and hence we may write (3) as

$$\left(\sum_{k=1}^{N}|x_k|^p\right)^{1/p} \le \left(\sum_{k=1}^{N}|x_k - x_k{}^{(n)}|^p\right)^{1/p} + B \tag{4}$$

where $\|x^{(n)}\|_p \le B$ for all n.

Since $x_k{}^{(n)} \to x_k$ for each of the finite number of terms $k = 1, \ldots, N$, we can choose n sufficiently large (but depending on N) that the first term on the right hand side of (4) is ≤ 1, and hence for every N we have

$$\sum_{k=1}^{N}|x_k|^p \le (1+B)^p \ .$$

This shows that the series $\sum_{k=1}^{\infty} |x_k|^p$ converges, and thus by definition of l_p, the corresponding sequence $x = \{x_k\}$ is an element of l_p. We must still show that $x^{(n)} \to x$.

Since $\{x^{(n)}\}$ is a Cauchy sequence, it follows that given $\varepsilon > 0$, there exists M such that $m, n \ge M$ implies $\|x^{(m)} - x^{(n)}\|_p < \varepsilon$. Then for any N and all m, $n \ge M$ we have (using the Minkowski inequality again)

$$\begin{aligned}
&\left(\sum_{k=1}^{N}|x_k - x_k{}^{(n)}|^p\right)^{1/p} \\
&\qquad \le \left(\sum_{k=1}^{N}|x_k - x_k{}^{(m)}|^p\right)^{1/p} + \left(\sum_{k=1}^{N}|x_k{}^{(m)} - x_k{}^{(n)}|^p\right)^{1/p} \\
&\qquad \le \left(\sum_{k=1}^{N}|x_k - x_k{}^{(m)}|^p\right)^{1/p} + \|x^{(m)} - x^{(n)}\|_p
\end{aligned}$$

$$\le \left(\sum_{k=1}^{N} |x_k - x_k{}^{(m)}|^p\right)^{1/p} + \varepsilon \ . \tag{5}$$

But $x_k{}^{(m)} \to x_k$ for each k = 1, . . . , N and hence (by the same argument used above) we can choose m sufficiently large that the first term in the last line of (5) is < ε. This means that for every N and all n ≥ m (where m is independent of N) we have

$$\left(\sum_{k=1}^{N} |x_k - x_k{}^{(n)}|^p\right)^{1/p} < 2\varepsilon$$

and hence taking the limit as N → ∞ yields

$$\|x - x^{(n)}\|_p = \left(\sum_{k=1}^{\infty} |x_k - x_k{}^{(n)}|^p\right)^{1/p} \le 2\varepsilon \ .$$

Since this inequality holds for all n ≥ M, we have shown that $\|x - x^{(n)}\|_p \to 0$ or, alternatively, that $x^{(n)} \to x$. We have therefore shown that the space l_p is complete, i.e., it is a Banach space.

It is now easy to show that l_2 is a Hilbert space. To see this, we define the inner product on l_2 in the usual way by

$$\langle x, y\rangle = \sum_{k=1}^{\infty} x_k{}^* y_k \ .$$

Using the infinite-dimensional version of Hölder's inequality with p = q = 2 (i.e., Cauchy's inequality), we see that this series converges absolutely, and hence the series converges to a complex number (see Theorem B20). This shows that the inner product so defined is meaningful. The rest of the verification that l_2 is a Hilbert space is straightforward and left to the reader (see Exercise 12.3.2). ⫽

Recall that a subset A of a metric space X is said to be **dense** if Cl A = X. Intuitively, this simply means that any neighborhood of any x ∈ X contains points of A (see Theorems B13 and B15). A space is said to be **separable** if it contains a countable dense subset. An important class of Hilbert spaces are those that are separable.

Example 12.9 Let us show that the space l_2 is actually separable. In other words, we shall show that l_2 contains a countable dense subset. To see this, we say that a point $x = \{x_1, x_2, \dots\} \in l_2$ is a **rational point** if $x_n \neq 0$ for only a finite number of the indices n, and if each of these nonzero components is a rational (complex) number. It should be clear that the set of all such rational points is countably infinite. We must now show that any neighborhood of any point in l_2 contains at least one rational point.

To do this, we show that given any $x \in l_2$ and any $\varepsilon > 0$, there exists a rational point $r = \{r_1, r_2, \dots, 0, 0, \dots\} \in l_2$ such that $\|r - x\|_2 < \varepsilon$. Since $x \in l_2$ the series $\sum_{k=1}^{\infty} |x_k|^2$ converges, and hence there exists N such that

$$\sum_{k=N+1}^{\infty} |x_k|^2 < \varepsilon^2/2$$

(see Theorem B17). Next, for each $k = 1, \dots, N$ we find a rational number r_k with the property that

$$|r_k - x_k| < \frac{\varepsilon}{\sqrt{2N}} \;.$$

(That this can be done follows from Theorem 0.4 applied to both the real and imaginary parts of x_k.) Then the distance between x and the rational point $r = \{r_1, r_2, \dots, r_N, 0, 0, \dots\}$ is given by

$$\begin{aligned} \|r - x\|_2 &= \left(\sum_{k=1}^{N} |r_k - x_k|^2 + \sum_{k=N+1}^{\infty} |x_k|^2 \right)^{1/2} \\ &< [N(\varepsilon^2/2N) + \varepsilon^2/2]^{1/2} = \varepsilon \;. \;/\!\!/ \end{aligned}$$

As the last remark of this section, the reader should note that the proof of the Cauchy-Schwartz inequality in Example 12.1 made no reference whatsoever to any components, and thus it clearly holds in any Hilbert space, as does the parallelogram law. Furthermore, as mentioned in Example 12.3, the Cauchy-Schwartz inequality also shows that the inner product is continuous in each variable. Indeed, applying Theorem 12.7(d) we see that if $x_n \to x$ and $y_n \to y$, then

$$\begin{aligned} |\langle x_n, y_n\rangle - \langle x, y\rangle| &= |\langle x_n - x, y_n - y\rangle + \langle x_n - x, y\rangle + \langle x, y_n - y\rangle| \\ &\le \|x_n - x\|\,\|y_n - y\| + \|x_n - x\|\,\|y\| + \|x\|\,\|y_n - y\| \to 0 \end{aligned}$$

This is sometimes expressed by saying that the inner product is **jointly continuous**. Alternatively, we can note that

$$|\langle x_1, y\rangle - \langle x_2, y\rangle| = |\langle x_1 - x_2, y\rangle| \le \|x_1 - x_2\|\,\|y\|$$

which shows that the map x → ⟨x, y⟩ is actually uniformly continuous, with the same result holding for y → ⟨x, y⟩.

Exercises

1. If $x_n = \sum_{k=1}^{n} a_k v_k \to x$ and $y_n = \sum_{k=1}^{n} b_k v_k \to y$, show that $x_n + y_n \to x + y$.

2. Complete the proof (begun in Example 12.8) that l_2 is a Hilbert space. [*Hint*: Note that if $x = \{x_1, x_2, \dots\}$ and $y = \{y_1, y_2, \dots\}$ are vectors in l_2, then you must show that $x + y \in l_2$ also.]

3. Prove that every compact metric space (X, d) is separable. [*Hint*: For each integer $n \ge 1$ consider the collection U_n of open spheres
$$U_n = \{B(x, 1/n) \colon x \in X\} \ .]$$

4. Let H be a Hilbert space and suppose $A \in \mathcal{L}(H)$ is a positive symmetric operator. Prove the **generalized Schwartz inequality**:
$$|\langle Ax, y\rangle|^2 \le \langle Ax, x\rangle\langle Ay, y\rangle$$
where $x, y \in H$. [*Hint*: Let c be a real number and consider the vector $z = x + c\langle Ax, y\rangle y$.]

5. Let l_∞ denote the linear space consisting of all bounded sequences $x = \{x_1, x_2, \dots, x_n, \dots\}$ of scalars with norm $\|x\| = \sup |x_n|$. Show that l_∞ is a Banach space.

12.4 CLOSED SUBSPACES

Since the norm on a vector space induces a metric topology on the space (i.e., defines the open sets in terms of the induced metric), it makes sense to define a **closed subspace** as a subspace which is a closed set relative to the metric topology. In view of Theorem B14, we say that a set A of vectors is **closed** if every convergent sequence of vectors in A converges to a vector in A. If E is a vector space, many authors define a **linear manifold** to be a subset $S \subset E$ of vectors such that S is also a linear space. In this case, a **subspace** is defined to be a closed linear manifold. From the corollary to Theorem 12.9, we then see that any finite-dimensional linear manifold over either $\mathbb{C}$ or $\mathbb{R}$ is a subspace. We mention this terminology only in passing, and will generally continue to use the word "subspace" in our previous context (i.e., as a linear manifold). As a simple example, let $V = \mathbb{R}$ be a vector space over the field $\mathbb{Q}$. Then the subspace $W \subset V$ defined by $W = \mathbb{Q}$ is not closed (why?).

Recall from Theorem 2.22 that if W is a subspace of a finite-dimensional inner product space V, then $V = W \oplus W^\perp$. We now wish to prove that if M is a closed subspace of a Hilbert space H, then $H = M \oplus M^\perp$. Unfortunately, this requires that we prove several preliminary results along the way. We begin with a brief discussion of convex sets.

We say that a subset S of a vector space V is **convex** if for every pair $x, y \in S$ and any real number $t \in [0, 1]$, the vector

$$z = (1 - t)x + ty$$

is also an element of S. Intuitively, this is just says that the straight line segment from x to y in V is in fact contained in S. It should be obvious that the intersection of any collection of convex sets is convex, and that every subspace of V is necessarily convex.

It follows by induction that if S is convex and $x_1, \dots, x_n \in S$, then the vector $t_1x_1 + \cdots + t_nx_n$ where $0 \le t_i \le 1$ and $t_1 + \cdots + t_n = 1$ is also in S. Conversely, the set of all such linear combinations forms a convex set. It is trivial to verify that if S is convex, then so is any translate

$$S + z = \{x + z\colon z \in V \text{ is fixed and } x \in S\} .$$

Moreover, if $\lambda\colon V \rightarrow W$ is a linear map and $S \subset V$ is convex, then $\lambda(S)$ is a convex subset of W, and if $T \subset W$ is convex, then $\lambda^{-1}(T)$ is convex in V. We leave the proofs of these elementary facts to the reader (see Exercise 12.4.1).

The main result dealing with convex sets that we shall need is given in the next theorem.

Theorem 12.14 Every nonempty closed convex subset S of a Hilbert space H contains a unique vector of smallest norm. In other words, there exists a unique $x_0 \in S$ with the property that $\|x_0\| \le \|x\|$ for every $x \in S$.

Proof Let $\delta = \inf\{\|x\|\colon x \in S\}$. By definition of inf, this implies the existence of a sequence $\{x_n\}$ of vectors in S such that $\|x_n\| \rightarrow \delta$. Since S is convex, $(x_n + x_m)/2$ is also in S (take $t = 1/2$ in the definition of convex set), and hence $\|(x_n + x_m)/2\| \ge \delta$ or $\|x_n + x_m\| \ge 2\delta$. Applying the parallelogram law we see that

$$\begin{aligned}\|x_n - x_m\|^2 &= 2\|x_n\|^2 + 2\|x_m\|^2 - \|x_n + x_m\|^2 \\ &\le 2\|x_n\|^2 + 2\|x_m\|^2 - 4\delta^2 \ .\end{aligned}$$

Taking the limit of the right hand side of this equation as m, n → ∞ shows that $\|x_n - x_m\| \to 0$, and hence $\{x_n\}$ is a Cauchy sequence in S. By Theorem 12.9, S is complete, and thus there exists a vector $x \in S$ such that $x_n \to x$. Since the norm is a continuous function, we see that (see Examples 12.2 and 12.3 or Theorem 12.7(d))

$$\|x\| = \|\lim x_n\| = \lim \|x_n\| = \delta \ .$$

Thus $x \in S$ is a vector with smallest norm $\delta = \inf\{\|x\|: x \in S\}$.

To show that this x is unique, suppose that $y \in S$ is such that $\|y\| = \delta$. Applying the parallelogram law again to the vectors x/2 and y/2 yields

$$\|x - y\|^2/4 = \|x\|^2/2 + \|y\|^2/2 - \|(x + y)/2\|^2 \ .$$

But $(x + y)/2 \in S$ implies $\|(x + y)/2\| \ge \delta$, and thus we have

$$\|x - y\|^2 \le 2\|x\|^2 + 2\|y\|^2 - 4\delta^2 \ .$$

If $\|x\| = \|y\| = \delta$, then this equation implies that x = y. ∎

The notion of orthogonality is extremely important in the theory of Hilbert spaces. We recall from Section 12.1 that two vectors x and y in a Hilbert space H are said to be orthogonal if $\langle x, y\rangle = 0$, and we write this as $x \perp y$. If S is a (nonempty) subset of H and $x \in H$ is orthogonal to every $y \in S$, then we express this by writing $x \perp S$. Thus the orthogonal complement $S^\perp$ of S is defined by $S^\perp = \{x \in H: x \perp S\}$.

As an example, we consider the orthogonal complement $x^\perp$ of any $x \in H$. If $x \perp y$ and $x \perp z$, then $x \perp (y + z)$ and $x \perp (\alpha y)$ for any scalar α. Therefore $x^\perp$ is actually a subspace of H. If we define a continuous linear map $f_x: H \to \mathbb{C}$ by $f_x(y) = \langle x, y\rangle$, then $x^\perp = \{y \in H: f_x(y) = 0\}$. In fact, if $\{y_n\}$ is a sequence in $x^\perp$ that converges to an element $y \in H$, then the continuity of the inner product yields

$$\langle x, y\rangle = \langle x, \lim y_n\rangle = \lim \langle x, y_n\rangle = 0$$

and hence $y \in x^\perp$ also. This proves that $x^\perp$ is in fact a closed subspace of H. Carrying this idea a little farther, if S is a subset of H, then we can clearly write $S^\perp = \cap_{x \in S}\, x^\perp$. Since this shows that $S^\perp$ is the intersection of closed subspaces, it follows that $S^\perp$ must also be a closed subspace (see Exercise 12.4.2). Alternatively, if $y \in S$ and $\{x_n\} \subset S^\perp$ with $x_n \to x$, then we again have $\langle x, y\rangle = \lim \langle x_n, y\rangle = 0$ so that $x \in S^\perp$ and $S^\perp$ is therefore closed. We leave it to the reader to prove the following (see Exercise 12.4.3):

$$0^\perp = H \text{ and } H^\perp = 0.$$
$$S \cap S^\perp \subset \{0\}.$$
$$S \subset S^{\perp\perp} = (S^\perp)^\perp.$$
$$S_1 \subset S_2 \text{ implies } S_2{}^\perp \subset S_1{}^\perp.$$

Furthermore, using the next theorem, it is not hard to show that a subset M of a Hilbert space H is closed if and only if $M^{\perp\perp} = M$ (see Exercise 12.4.6).

Theorem 12.15 Let M be a proper closed subspace of a Hilbert space H (i.e., $M \neq H$). Then there exists a nonzero vector $x_0 \in H$ such that $x_0 \perp M$.

Proof Suppose $x \in H$ and $x \notin M$. Since any subspace is automatically convex, it follows that the set $x - M = \{x - y: y \in M\}$ is closed and convex. By Theorem 12.14, this set contains a unique vector $x_0 = x - y_0 \in x - M$ of smallest norm. By definition, this means that $\|x - y_0\| = \inf\{\|x - y\|: y \in M\}$. If we had $\|x - y_0\| = 0$, then x would be an accumulation point of M, contradicting the assumption that M is closed and $x \notin M$. Thus we must have $x_0 \neq 0$, and we claim that $x_0 \perp M$.

Since x_0 is of smallest norm, we see that for any $y \in M$ and any $\alpha \in \mathbb{C}$ we have

$$\|x_0\|^2 \le \|x_0 + \alpha y\|^2 \ .$$

Expanding this out in terms of the inner product on H we find that

$$0 \le 2\,\mathrm{Re}\{\alpha\langle x_0, y\rangle\} + |\alpha|^2\,\|y\|^2 \ .$$

In particular, if we let $\alpha = c\langle y, x_0\rangle$ where $c \in \mathbb{R}$ is nonzero, then this equation becomes

$$0 \le c|\langle x_0, y\rangle|^2\,(2 + c\|y\|^2) \ .$$

If $y \in M$ is such that $\langle x_0, y\rangle \neq 0$, then the fact that this equation holds for all nonzero $c \in \mathbb{R}$ leads to a contradiction if we choose c such that $-2/\|y\|^2 < c < 0$. It therefore follows that we must have $\langle x_0, y\rangle = 0$ for every $y \in M$, and hence $x_0 \perp M$. ∎

We are now in a position to prove our earlier assertion. After the proof we shall give some background as to why this result is important.

Theorem 12.16 Let M be a closed subspace of a Hilbert space H. Then $H = M \oplus M^\perp$.

Proof We first show that $M + M^\perp$ is a closed subspace of H. To see this, we note that $M \cap M^\perp = \{0\}$ (since M is a subspace and hence contains the zero vector), and every $z \in M + M^\perp$ may be written in the unique form $z = x + y$ with $x \in M$ and $y \in M^\perp$. (See the proof of Theorem 2.12. Note also that we have not yet shown that every $z \in H$ is of this form.) Now let $\{z_n\} = \{x_n + y_n\}$ be any sequence in $M + M^\perp$ that converges to an element $z \in H$. We must show that $z \in M + M^\perp$. Using the Pythagorean theorem we see that

$$\begin{aligned}\|z_m - z_n\|^2 &= \|(x_m + y_m) - (x_n + y_n)\|^2\\ &= \|(x_m - x_n) + (y_m - y_n)\|^2\\ &= \|x_m - x_n\|^2 + \|y_m - y_n\|^2\end{aligned}$$

and therefore $\{z_n\}$ is a Cauchy sequence in $M + M^\perp$ if and only if $\{x_n\}$ is a Cauchy sequence in M and $\{y_n\}$ is a Cauchy sequence in $M^\perp$. Since both M and $M^\perp$ are closed they are complete (Theorem 12.9). Therefore, since $\{z_n\}$ is a convergent sequence in H it is a Cauchy sequence in H, and in fact it is a Cauchy sequence in $M + M^\perp$ since every $z_n = x_n + y_n \in M + M^\perp$. But then $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences in M and $M^\perp$ which must converge to points $x \in M$ and $y \in M^\perp$. Hence

$$z = \lim z_n = \lim(x_n + y_n) = \lim x_n + \lim y_n = x + y \in M + M^\perp .$$

This shows that $M + M^\perp$ is a closed subspace of H.

We now claim that $H = M + M^\perp$. Since we already know that $M \cap M^\perp = \{0\}$, this will complete the proof that $H = M \oplus M^\perp$. If $H \neq M + M^\perp$, then according to Theorem 12.15 there exists a nonzero $z_0 \in H$ with the property that $z_0 \perp (M + M^\perp)$. But this implies that $z_0 \in M^\perp$ and $z_0 \in M^{\perp\perp}$, and hence $\|z_0\|^2 = \langle z_0, z_0\rangle = 0$ (or observe that $z_0 \in M^\perp \cap M^{\perp\perp} = \{0\}$) which contradicts the assumption that $z_0 \neq 0$. ∎

To gain a little insight as to why this result is important, we recall our discussion of projections in Section 7.8. In particular, Theorem 7.27 shows that a linear transformation E on a finite-dimensional vector space V is idempotent (i.e., $E^2 = E$) if and only if $V = U \oplus W$ where E is the projection of V on $U = \operatorname{Im} E$ in the direction of $W = \operatorname{Ker} E$. In order to generalize this result to Banach spaces, we define an **operator** on a Banach space B to be an element of $\mathcal{L}(B, B)$. In other words, an operator on B is a continuous linear transformation of B into itself. A **projection** on B is an idempotent operator on B.

Thus, in order that an operator P on B be idempotent, it must obey both the algebraic requirement that $P^2 = P$ as well as the topological condition of continuity. The generalization of Theorem 7.27 to Banach spaces is given by the next two theorems, the proofs of which are left to the reader since we will not be needing them again.

Theorem 12.17 Let P be a projection on a Banach space B and let M = Im P and N = Ker P. Then M and N are closed subspaces of B, and $B = M \oplus N$.

Proof See Exercise 12.4.5. ∎

Theorem 12.18 Let B be a Banach space and let M, N be closed subspaces of B such that $B = M \oplus N$. Then for any $z = x + y \in M \oplus N$, the mapping P defined by $P(z) = x$ is a projection on B with Im P = M and Ker P = N.

Proof The only difficult part of this theorem is the proof that P is continuous. While this may be proved using only what has been covered in this book (including the appendices), it is quite involved since it requires proving both Baire's theorem and the open mapping theorem. Since these are essentially purely topological results whose proofs are of no benefit to us at this point, we choose to refer the interested reader to, e.g., the very readable treatment by Simmons (1963). ∎

As mentioned in Section 7.8, if we are given a space V and subspace U, there may be many subspaces W with the property that $V = U \oplus W$. Thus, if we are given a closed subspace M of a Banach space B, then there could be many *algebraic* projections defined on B with image M, and in fact none of them may be projections as defined above (i.e., they may not be continuous). In other words, there may not exist any closed subspace N such that $B = M \oplus N$. However, Theorem 12.16 together with Theorem 12.18 shows that if we have a Hilbert space H together with a closed subspace M, then there always exists a projection P defined on $H = M \oplus M^\perp$ with Im P = M and Ker P = $M^\perp$.

Exercises

1. Let V and W be vector spaces, and let $S \subset V$ be convex.
   (a) Show that the intersection of any collection of convex sets is convex.
   (b) Show that every subspace of V is convex.

(c) Show that any translate $S + z = \{x + z: z \in V$ is fixed and $x \in S\}$ is convex.
(d) If $\lambda: V \to W$ is linear, show that $\lambda(S)$ is a convex subset of W, and if $T \subset W$ is convex, then $\lambda^{-1}(T)$ is convex in V.

2. Let H be a Hilbert space and S a nonempty subset of H. Show that $S^\perp = \cap_{x \in S} x^\perp$ is a closed subspace of H.

3. Let H be a Hilbert space and S a nonempty subset of H. Prove the following:
(a) $0^\perp = H$ and $H^\perp = 0$.
(b) $S \cap S^\perp \subset \{0\}$.
(c) $S \subset S^{\perp\perp}$.
(d) $S_1 \subset S_2$ implies $S_2{}^\perp \subset S_1{}^\perp$.

4. Show that a subset M of a Hilbert space H is closed if and only if $M^{\perp\perp} = M$.

5. Prove Theorem 12.17.

12.5 HILBERT BASES

Let us now turn our attention to the infinite-dimensional analogue of the expansion of a vector in terms of a basis. (We recommend that the reader first review Sections 0.3 and 0.4 before continuing on with this material.) Suppose that x is a vector in a Hilbert space H such that $\|x\| \neq 0$, and let $y \in H$ be arbitrary. We claim there exists a unique scalar c such that $y - cx$ is orthogonal to x. Indeed, if $(y - cx) \perp x$, then

$$0 = \langle x, y - cx\rangle = \langle x, y\rangle - c\langle x, x\rangle$$

implies that

$$c = \langle x, y\rangle/\langle x, x\rangle$$

while if $c = \langle x, y\rangle/\langle x, x\rangle$, then reversing the argument shows that $(y - cx) \perp x$. The scalar c is usually called the **Fourier coefficient** of y with respect to (or relative to) x.

To extend this idea to finite sets of vectors, let $\{x_i\} = \{x_1, \ldots, x_n\}$ be a collection of vectors in H. Furthermore assume that the x_i are mutually orthogonal, i.e., $\langle x_i, x_j\rangle = 0$ if $i \neq j$. If $c_i = \langle x_i, y\rangle/\langle x_i, x_i\rangle$ is the Fourier coefficient of $y \in H$ with respect to x_i, then

$$\begin{aligned}\langle x_i,\, y-\sum_{j=1}^{n} c_j x_j\rangle &= \langle x_i,\, y\rangle-\sum_{j=1}^{n} c_j\langle x_i,\, x_j\rangle\\ &= \langle x_i,\, y\rangle-c_i\langle x_i,\, x_i\rangle\\ &=0\end{aligned}$$

which shows that $y - \sum_{j=1}^{n} c_j x_j$ is orthogonal to each of the x_i. Geometrically, this result says that if we subtract off the components of a vector y in the direction of n orthogonal vectors x_i, then the resulting vector is orthogonal to each of the vectors x_i.

We can easily simplify many of our calculations be requiring that our finite set $\{x_i\}$ be orthonormal instead of just orthogonal. In other words, we assume that $\langle x_i, x_j\rangle = \delta_{ij}$, which is equivalent to requiring that $i \neq j$ implies that $x_i \perp x_j$ and $\|x_i\| = 1$. Note that given any $x_i \in H$ with $\|x_i\| \neq 0$, we can normalize x_i by forming the vector $e_i = x_i/\|x_i\|$. It is then easy to see that the above calculations remain unchanged except that now we simply have $c_i = \langle x_i, y\rangle$. We will usually denote such an orthonormal set by $\{e_i\}$, and hence we write

$$\langle e_i, e_j\rangle = \delta_{ij}\ .$$

Suppose $\{e_i\}$ is a finite orthonormal set in a Hilbert space H and x is any element of H. We claim that the expression

$$\|x-\sum_{k=1}^{n} a_k e_k\|$$

achieves its minimum value in the case where each of the scalars a_k is equal to the Fourier coefficient $c_k = \langle e_k, x\rangle$. To see this, we note that the above discussion showed that $x - \sum_{k=1}^{n} c_k e_k$ is orthogonal to each e_i for i = 1, . . . , n and hence we may apply the Pythagorean theorem to obtain

$$\begin{aligned}\|x-\sum_{k=1}^{n} a_k e_k\|^2 &= \|x-\sum_{k=1}^{n} c_k e_k+\sum_{k=1}^{n}(c_k-a_k)e_k\|^2\\ &= \|x-\sum_{k=1}^{n} c_k e_k\|^2+\|\sum_{k=1}^{n}(c_k-a_k)e_k\|^2\ .\end{aligned}$$

It is clear that the right hand side of this equation takes its minimum value at $a_k = c_k$ for k = 1, . . . , n and hence we see that in general

$$\|x-\sum_{k=1}^{n} c_k e_k\| \leq \|x-\sum_{k=1}^{n} a_k e_k\|$$

for any set of scalars a_k. Moreover, we see that (using $c_k = \langle e_k, x\rangle$)

$$\begin{aligned}
0 &\le \|x - \sum_{k=1}^{n} c_k e_k\|^2 \\
&= \langle x - \sum_{k=1}^{n} c_k e_k ,\ x - \sum_{r=1}^{n} c_r e_r \rangle \\
&= \|x\|^2 - \sum_{k=1}^{n} c_k{}^*\langle e_k ,\ x\rangle - \sum_{r=1}^{n} c_r \langle x,\ e_r\rangle + \sum_{k=1}^{n} |c_k|^2 \\
&= \|x\|^2 - \sum_{k=1}^{n} |c_k|^2
\end{aligned}$$

which implies

$$\sum_{k=1}^{n} |c_k|^2 = \sum_{k=1}^{n} |\langle e_k ,\ x\rangle|^2 \le \|x\|^2 \ \ .$$

This relationship is frequently called **Bessel's inequality**, although this designation also applies to the infinite-dimensional version to be proved below.

We now seek to generalize these last two results to the case of arbitrary (i.e., possibly uncountable) orthonormal sets. We begin with a simple theorem.

Theorem 12.19 Let $\{e_i\}$, $i \in I$ (where I may be uncountable) be an arbitrary orthonormal set in a Hilbert space H. Then if x is any vector in H, the set $S = \{e_i : \langle e_i, x\rangle \ne 0\}$ is countable (but possibly empty).

Proof For each $n \in \mathbb{Z}^+$ define the set

$$S_n = \{e_i : |\langle e_i, x\rangle|^2 > \|x\|^2/n\} \ \ .$$

We claim that each S_n can contain at most $n - 1$ vectors. To see this, suppose S_n contains N vectors, i.e., $S_n = \{e_1, \ldots, e_N\}$. Then from the definition of S_n we have

$$\sum_{i=1}^{N} |\langle e_i ,\ x\rangle|^2 > (\|x\|^2/n)N$$

while Bessel's inequality shows that

$$\sum_{i=1}^{N} |\langle e_i ,\ x\rangle|^2 \le \|x\|^2 \ \ .$$

Thus we must have $N < n$ which is the same as requiring that $N \le n - 1$. The theorem now follows if we note that each S_n consists of a finite number of vectors, and that $S = \cup_{n=1}^{\infty} S_n$ since $S_n \to S$ as $n \to \infty$. ∎

Theorem 12.19 now allows us to prove the general (i.e., infinite-dimensional) version of **Bessel's inequality**. Keep in mind that an arbitrary orthonormal set may consist of an uncountable number of elements, and in this case we do not write any limits in the sum $\sum|\langle e_i, x\rangle|^2$.

Theorem 12.20 If $\{e_i\}$ is any orthonormal set in a Hilbert space H and x is any vector in H, then $\sum|\langle e_i, x\rangle|^2 \le \|x\|^2$.

Proof First note that if $e_\alpha \in \{e_i\}$ is such that $\langle e_\alpha, x\rangle = 0$, then this e_α will not contribute to $\sum|\langle e_i, x\rangle|^2$. As in Theorem 12.19, we again consider the set

$$S = \{e_i\colon \langle e_i, x\rangle \neq 0\} \ .$$

If $S = \varnothing$, then we have $\sum|\langle e_i, x\rangle|^2 = 0$ and the conclusion is obviously true. If $S \neq \varnothing$, then according to Theorem 12.19 it must contain a countable number of vectors. If S is in fact finite, then we write $S = \{e_1, \dots, e_n\}$ and the theorem follows from the finite-dimensional Bessel inequality proved above. Thus we need only consider the case where S is countably infinite.

We may consider the vectors in S to be arranged in any arbitrary (but now fixed) order $\{e_1, e_2, \dots, e_n, \dots\}$. From the corollary to Theorem B20 we know that if $\sum_{i=1}^\infty|\langle e_i, x\rangle|^2$ converges, then this sum is independent of any rearrangement of the terms in the series. This then gives an unambiguous meaning to the expression $\sum|\langle e_i, x\rangle|^2 = \sum_{i=1}^\infty|\langle e_i, x\rangle|^2$. Therefore we see that the sum is a nonnegative (extended) real number that depends only on the set S and not on the order in which the vectors in S are written. If we let

$$s_n = \sum_{i=1}^n |\langle e_i, x\rangle|^2$$

be the n*th* partial sum of the series, then the finite-dimensional version of Bessel's inequality shows that $s_n \le \|x\|^2$ for every n, and hence we must have

$$\sum_{i=1}^\infty |\langle e_i, x\rangle|^2 \le \|x\|^2 \ . \ \blacksquare$$

Theorem 12.21 Let $\{e_i\}$ be an orthonormal set in a Hilbert space H, and let x be any vector in H. Then $(x - \sum\langle e_i, x\rangle e_i) \perp e_j$ for each j.

Proof Just as we did in the proof of Theorem 12.20, we must first make precise the meaning of the expression $\sum\langle e_i, x\rangle e_i$. Therefore we again define the set $S = \{e_i\colon \langle e_i, x\rangle \neq 0\}$. If $S = \varnothing$, then we have $\sum\langle e_i, x\rangle e_i = 0$ so that our theorem is obviously true since the definition of S then means that $x \perp e_j$ for every j. If S is finite but nonempty, then the theorem reduces to the finite case proved in the discussion prior to Theorem 12.19. Thus, by Theorem 12.19, we are again left with the case where S is countably infinite. We first prove the result for a particular ordering $S = \{e_1, e_2, \dots\}$, and afterwards we will show that our result is independent of the ordering chosen.

Let $s_n = \sum_{i=1}^{n}\langle e_i, x\rangle e_i$. Since the Pythagorean theorem may be applied to any finite collection of orthogonal vectors, the fact that $\{e_i\}$ is an orthonormal set allows us to write (for $m > n$)

$$\|s_m - s_n\|^2 = \|\sum_{i=n+1}^{m}\langle e_i, x\rangle e_i\|^2 = \sum_{i=n+1}^{m}|\langle e_i, x\rangle|^2 \ .$$

Now, Bessel's inequality shows that $\sum_{i=1}^{\infty}|\langle e_i, x\rangle|^2$ must converge, and hence for any $\varepsilon > 0$ there exists N such that $m > n \geq N$ implies $\sum_{i=n+1}^{m} |\langle e_i, x\rangle|^2 < \varepsilon^2$ (this is just Theorem B17). This shows that $\{s_n\}$ is a Cauchy sequence in H, and thus the fact that H is complete implies that $s_n \to s = \sum_{i=1}^{\infty}\langle e_i, x\rangle e_i \in H$. If we define $\sum\langle e_i, x\rangle e_i = \sum_{i=1}^{\infty}\langle e_i, x\rangle e_i = s$, then the continuity of the inner product yields

$$\begin{aligned}\langle e_j, x - s\rangle &= \langle e_j, x\rangle - \langle e_j, s\rangle = \langle e_j, x\rangle - \langle e_j, \lim s_n\rangle \\ &= \langle e_j, x\rangle - \lim\langle e_j, s_n\rangle = \langle e_j, x\rangle - \lim\sum_{i=1}^{n}\langle e_i, x\rangle\langle e_j, e_i\rangle \\ &= \langle e_j, x\rangle - \langle e_j, x\rangle = 0 \ .\end{aligned}$$

Thus we have shown that $(x - s) \perp e_j$ for every j.

We now show that this result is independent of the particular order chosen for the $\{e_i\}$ in the definition of s. Our proof of this fact is similar to the proof of the corollary to Theorem B20. Let $\{e'_i\}$ be any other arrangement of the set $\{e_i\}$, and let $s'_n = \sum_{i=1}^{n}\langle e'_i, x\rangle e'_i$. Repeating the above argument shows that s'_n converges to a limit $s' = \sum_{i=1}^{\infty}\langle e'_i, x\rangle e'_i$. We must show that $s' = s$. Since $\{s_n\}$, $\{s'_n\}$ and $\sum_{i=1}^{\infty}|\langle e_i, x\rangle|^2$ all converge, we see that for any $\varepsilon > 0$, there exists $N > 0$ such that $n \geq N$ implies

$$\|s_n - s\| < \varepsilon$$

$$\|s'_n - s'\| < \varepsilon$$

and

$$\sum_{i=N+1}^{\infty}|\langle e_i, x\rangle|^2 < \varepsilon^2$$

(this last inequality follows by letting $m \to \infty$ in Theorem B17). We now note that since there are only a finite number of terms in s_N and $\{e'_i\}$ is just a rearrangement of $\{e_i\}$, there must exist an integer $M > N$ such that every term in s_N also occurs in s'_M. Then $s'_M - s_N$ contains a finite number of terms, each of which is of the form $\langle e_i, x\rangle e_i$ for $i = N + 1, N + 2, \dots$. We therefore have

$$\|s'_M - S_N\|^2 \le \sum_{i=N+1}^{\infty} |\langle e_i, x\rangle|^2 < \varepsilon^2$$

and hence $\|s'_M - s_N\| < \varepsilon$. Putting all of this together, we have

$$\|s' - s\| \le \|s' - s'_M\| + \|s'_M - s_N\| + \|s_N - s\| < 3\varepsilon$$

and hence $s' = s$. ∎

At last we are in a position to describe the infinite-dimensional analogue of the expansion of a vector in terms of a basis. Let H be a nonzero Hilbert space, and let $\{x_i\}$ be an arbitrary collection of vectors in H such that $\|x_i\| \neq 0$ for each i. (We required H to be nonzero so that such a collection will exist.) For *each* finite subcollection $\{x_{i_1}, \dots, x_{i_n}\}$ of $\{x_i\}$, we can form the vector space spanned by this subcollection of vectors. In other words, we can consider the space consisting of all linear combinations of the form $c_1x_{i_1} + \cdots + c_nx_{i_n}$ where each c_i is a complex number. In order to simplify our notation, we will generally omit the double indices and write simply $\{x_1, \dots, x_n\}$.

Now consider the union of *all* vector spaces generated by such finite subcollections of $\{x_i\}$. This union is clearly a vector space itself, and is called the subspace **generated** by the collection $\{x_i\}$. Let us denote this space by E. We say that the collection $\{x_i\}$ is **total** in H if E is dense in H (i.e., Cl E = H). In other words, $\{x_i\}$ is total in H if every vector in H is the limit of a sequence of vectors in E (see Theorem B14(b)). A total orthonormal set $\{e_i\}$ is called a **Hilbert basis** (or an **orthonormal basis**). Be careful to note however, that this is not the same as a basis in a finite-dimensional space. This is because not every vector in H can be written as a linear combination of a *finite* number of elements in a Hilbert basis.

An equivalent way of formulating this property that is frequently used is the following. Consider the family of all orthonormal subsets of a nonzero Hilbert space H. We can order this family by ordinary set inclusion, and the result is clearly a partially (but not totally) ordered set. In other words, if S_1 and S_2 are orthonormal sets, we say that $S_1 \le S_2$ if $S_1 \subset S_2$. We say that an orthonormal set $\{e_i\}$ is **complete** if it is maximal in this partially ordered set. This means that there is no nonzero vector $x \in H$ such that if we adjoin $e = x/\|x\|$ to $\{e_i\}$, the resulting set $\{e_i, e\}$ is also orthonormal and contains $\{e_i\}$ as a proper subset. We now show the equivalence of this approach to the previous paragraph.

Let $\{e_i\}$ be a complete orthonormal set in a Hilbert space H, and let E be the subspace generated by $\{e_i\}$. If Cl E ≠ H, then by Theorem 12.15 there exists a nonzero vector $x \in H$ such that $x \perp \text{Cl } E$. In particular, this means that $x \perp E$ and hence the set $\{e_i, e = x/\|x\|\}$ would be a larger orthonormal set than $\{e_i\}$, contradicting the maximality of $\{e_i\}$.

Conversely, suppose that $\{e_i\}$ is a Hilbert basis for H (i.e., a total orthonormal set). If $\{e_i\}$ is not complete, then there exists a nonzero vector $x \in H$ such that $\{e_i, e = x/\|x\|\}$ is an orthonormal set that contains $\{e_i\}$ as a proper subset. Then $e \perp \{e_i\}$, and hence the subspace E generated by $\{e_i\}$ must be a subset of $e^\perp$. Since $e^\perp$ is closed, it follows that $\text{Cl } E \subset e^\perp$. But then $e \perp \text{Cl } E$ which contradicts the assumption that Cl E = H.

Theorem 12.22 Every nonzero Hilbert space H contains a complete orthonormal set. Alternatively, every such H has a Hilbert basis.

Proof Note that every chain of orthonormal sets in H has an upper bound given by the union of the sets in the chain. By Zorn's lemma, the set of all orthonormal sets thus has a maximal element. This shows that H contains a complete orthonormal set. That H has an orthonormal basis then follows from the above discussion on the equivalence of a complete orthonormal set and a Hilbert basis. ∎

Some of the most important basic properties of Hilbert spaces are contained in our next theorem.

Theorem 12.23 Let $\{e_i\}$ be an orthonormal set in a Hilbert space H. Then the following conditions are equivalent:

(1) $\{e_i\}$ is complete.

(2) $x \perp \{e_i\}$ implies $x = 0$.

(3) For any $x \in H$ we have $x = \sum \langle e_i, x\rangle e_i$.

(4) For any $x \in H$ we have $\|x\|^2 = \sum |\langle e_i, x\rangle|^2$.

Proof (1) ⇒ (2): If (2) were not true, then there would exist a nonzero vector $e = x/\|x\| \in H$ such that $e \perp \{e_i\}$, and hence $\{e_i, e\}$ would be an orthonormal set larger than $\{e_i\}$, contradicting the completeness of $\{e_i\}$.

(2) ⇒ (3): By Theorem 12.21, the vector $y = x - \sum\langle e_i, x\rangle e_i$ is orthogonal to $\{e_i\}$, and hence (2) implies that $y = 0$.

(3) ⇒ (4): Using the joint continuity of the inner product (so that the sum as a limit of partial sums can be taken outside the inner product), we simply calculate

$$\begin{aligned}\|x\|^2 = \langle x, x\rangle &= \langle \Sigma\langle e_i, x\rangle e_i, \Sigma\langle e_j, x\rangle e_j\rangle \\ &= \Sigma\langle e_i, x\rangle^* \Sigma\langle e_j, x\rangle\langle e_i, e_j\rangle \\ &= \Sigma\langle e_i, x\rangle^*\langle e_i, x\rangle \\ &= \Sigma|\langle e_i, x\rangle|^2 \ .\end{aligned}$$

(4) ⇒ (1): If $\{e_i\}$ is not complete, then there exists $e \in H$ such that $\{e_i, e\}$ is a larger orthonormal set in H. Since this means that $e \perp \{e_i\}$, statement (4) yields $\|e\|^2 = \Sigma|\langle e_i, e\rangle|^2 = 0$ which contradicts the assumption that $\|e\| = 1$. ∎

Note that the equivalence of (1) and (3) in this theorem is really just our earlier statement that an orthonormal set is complete if and only if it is a Hilbert basis. We also remark that statement (4) is sometimes called **Parseval's equation**, although this designation also applies to the more general result

$$\langle x, y\rangle = \Sigma\langle x, e_i\rangle\langle e_i, y\rangle$$

(see Exercise 12.5.1).

It should be emphasized that we have so far considered the general case where an arbitrary Hilbert space H has a possibly uncountable orthonormal set. However, if H happens to be separable (i.e., H contains a countable dense subset), then we can show that every orthonormal set in H is in fact countable.

Theorem 12.24 Every orthonormal set $\{e_i\}$ in a separable Hilbert space H contains at most a countable number of elements.

Proof We first note that by the Pythagorean theorem we have

$$\|e_i - e_j\|^2 = \|e_i\|^2 + \|e_j\|^2 = 2$$

and hence $\|e_i - e_j\| = \sqrt{2}$ for every $i \neq j$. If we consider the set $\{B(e_i, 1/2)\}$ of open balls of radius 1/2, then the fact that $2(1/2) = 1 < \sqrt{2}$ implies that these balls are pairwise disjoint. Now let $\{x_n\}$ be a countable dense subset of H. This means that any neighborhood of any element of H must contain at least one of the x_n. In particular, each of the open balls $B(e_i, 1/2)$ must contain at least one of the x_n, and hence there can be only a countable number of such balls (since distinct balls are disjoint). Therefore the set $\{e_i\}$ must in fact be countable. ∎

It is worth remarking that if we are given any countable set of linearly independent vectors $\{x_i\}$ in a Hilbert space H, then the Gram-Schmidt procedure (see the corollary to Theorem 2.21) may be applied to yield a countable orthonormal set $\{e_i\}$ such that for any n, the space spanned by $\{e_1 , . . . , e_n\}$ is the same as the space spanned by $\{x_1, \ldots, x_n\}$. It then follows that $\{e_i\}$ is complete if and only if $\{x_i\}$ is complete.

Finally, suppose that we have a countable (but not necessarily complete) orthonormal set $\{e_i\}$ in a Hilbert space H. From Bessel's inequality, it follows that a necessary condition for a set of scalars $c_1, c_2, \ldots$ to be the Fourier coefficients of some $x \in H$ is that $\sum_{k=1}^{\infty}|c_k|^2 \le \|x\|^2$. In other words, the series $\sum_{k=1}^{\infty}|c_k|^2$ must converge. That this is also a sufficient condition is the content of our next result, which is a special case of the famous Riesz-Fischer theorem.

Theorem 12.25 (**Riesz-Fischer**) Let $\{e_i\}$ be an orthonormal set in a Hilbert space H, and let $\{c_i\}$ be a collection of scalars such that the series $\sum_{k=1}^{\infty}|c_k|^2$ converges. Then there exists a vector $x \in H$ with $\{c_i\}$ as its Fourier coefficients. In other words, $\sum_{k=1}^{\infty}|c_k|^2 = \|x\|^2$ where $c_k = \langle e_k, x\rangle$.

Proof For each n, define the vector

$$x_n = \sum_{k=1}^{n} c_k e_k$$

and note that $c_k = \langle e_k, x_n\rangle$ for $k \le n$. Since $\sum_{k=1}^{\infty}|c_k|^2$ converges, it follows from Theorem B17 that for each $\varepsilon > 0$, there exists N such that $n > m \ge N$ implies

$$\sum_{k=m+1}^{n} |c_k|^2 < \varepsilon \ .$$

Using the Pythagorean theorem, we then see that $n > m \ge N$ implies

$$\|x_n - x_m\|^2 = \|\sum_{k=m+1}^{n} c_k e_k\|^2 = \sum_{k=m+1}^{n} \|c_k e_k\|^2 = \sum_{k=m+1}^{n} |c_k|^2 < \varepsilon$$

and hence $\{x_n\}$ is a Cauchy sequence in H. Since H is complete, there exists a vector $x \in H$ such that $\|x_n - x\| \to 0$. In addition, we note that we may write

$$\langle e_k, x\rangle = \langle e_k, x_n\rangle + \langle e_k, x - x_n\rangle \qquad (6)$$

where the first term on the right hand side is just c_k. From the Cauchy-Schwartz inequality we see that

$$|\langle e_k, x - x_n\rangle| \le \|e_k\|\,\|x - x_n\| = \|x - x_n\|$$

and thus letting $n \to \infty$ shows that $\langle e_k, x - x_n\rangle \to 0$. Since the left hand side of (6) is independent of n, we then see that

$$\langle e_k, x\rangle = c_k \quad .$$

Using this result, we now let n → ∞ to obtain (since $\|x - x_n\| \to 0$)

$$\begin{aligned}\|x - x_n\|^2 &= \langle x - \sum_{k=1}^{n} c_k e_k,\ x - \sum_{k=1}^{n} c_k e_k\rangle \\ &= \|x\|^2 - \sum_{k=1}^{n} |c_k|^2 \ \to\ 0 \quad .\end{aligned}$$

In other words, we have

$$\lim_{n\to\infty} \sum_{k=1}^{n} |c_k|^2 = \sum_{k=1}^{\infty} |c_k|^2 = \|x\|^2 \quad . \ ∎$$

Exercises

1. If $\{e_i\}$ is a complete orthonormal set in a Hilbert space H and x, y ∈ H, prove that $\langle x, y\rangle = \sum_i \langle x, e_i\rangle\langle e_i, y\rangle$.

2. Let e_n denote the sequence with a 1 in the n*th* position and 0's elsewhere. Show that $\{e_1, e_2, \ldots, e_n, \ldots\}$ is a complete orthonormal set in l_2.

3. Prove that a Hilbert space H is separable if and only if every orthonormal set in H is countable.

4. (a) Show that an orthonormal set in a Hilbert space is linearly independent.
(b) Show that a Hilbert space is finite-dimensional if and only if every complete orthonormal set is a basis.

5. Let S be a nonempty set, and let $l_2(S)$ denote the set of all complex-valued functions f defined on S with the property that:
(i) $\{s \in S: f(s) \neq 0\}$ is countable (but possibly empty).
(ii) $\sum |f(s)|^2 < \infty$.
It should be clear that $l_2(S)$ forms a complex vector space with respect to pointwise addition and scalar multiplication.

(a) Show that $l_2(S)$ becomes a Hilbert space if we define the norm and inner product by $\|f\| = (\sum|f(s)|^2)^{1/2}$ and $\langle f, g\rangle = \sum f(s)^*g(s)$.
(b) Show that the subset of $l_2(S)$ consisting of functions that have the value 1 at a single point and 0 elsewhere forms a complete orthonormal set.
(c) Now let $S = \{e_i\}$ be a complete orthonormal set in a Hilbert space H. Each $x \in H$ defines a function f on S by $f(e_i) = \langle e_i, x\rangle$. Show that f is in $l_2(S)$.
(d) Show that the mapping $x \mapsto f$ is an isometric (i.e., norm preserving) isomorphism of H onto $l_2(S)$.

12.6 BOUNDED OPERATORS ON A HILBERT SPACE

One of the most important concepts in quantum theory is that of self-adjoint operators on a Hilbert space. We now begin a discussion on the existence of operator adjoints. While the existence of the adjoint in a finite-dimensional space was easy enough to prove, the infinite-dimensional analogue requires slightly more care. Therefore, our first goal is to prove one version of a famous result known as the Riesz representation theorem, which is the Hilbert space analogue of Theorem 10.1.

As usual, we let E* denote the **dual space** (which is also frequently called the **conjugate space**) to the Banach space E. In other words, E* is just the space $\mathcal{L}(E, \mathbb{C})$ of continuous linear maps of E into $\mathbb{C}$. Elements of E* are called **functionals**, and it is important to remember that this designation implies that the map is continuous (and hence bounded). If $f \in E^*$, we may define the norm of f as usual by

$$\|f\| = \sup\{|f(x)|: \|x\| = 1\} \;.$$

Since $\mathbb{C}$ is clearly a Banach space, it follows from Theorem 12.13 that E* is also a Banach space (even if E is not).

If y is any (fixed) vector in a Hilbert space H, then we define the function $f_y\colon H \to \mathbb{C}$ by $f_y\colon x \mapsto \langle y, x\rangle$. Since the inner product is continuous, it follows that f_y is continuous. Furthermore, we note that for any $x_1, x_2 \in H$ and $\alpha \in \mathbb{C}$ we have

$$f_y(x_1 + x_2) = \langle y, x_1 + x_2\rangle = \langle y, x_1\rangle + \langle y, x_2\rangle = f_y(x_1) + f_y(x_2)$$

and

$$f_y(\alpha x_1) = \langle y, \alpha x_1\rangle = \alpha\langle y, x_1\rangle = \alpha f_y(x_1)$$

and hence f_y is linear. This shows that $f_y \in H^* = \mathcal{L}(H, \mathbb{C})$, and therefore we may define

$$\|f_y\| = \sup\{|f_y(x)|: \|x\| = 1\} .$$

Using Example 12.1, we see that $|\langle y, x\rangle| \le \|y\| \, \|x\|$, and thus (by definition of sup)

$$\|f_y\| = \sup\{|\langle y, x\rangle|: \|x\| = 1\} \le \|y\| .$$

On the other hand, we see that $y = 0$ implies $\|f_y\| = 0 = \|y\|$, while if $y \ne 0$ then (again by the definition of sup)

$$\|f_y\| = \sup\{|f_y(x)|: \|x\| = 1\} \ge |f_y(y/\|y\|)| = |\langle y, y/\|y\|\rangle| = \|y\| .$$

We thus see that in fact $\|f_y\| = \|y\|$, and hence the map $y \mapsto f_y$ preserves the norm.

However, the mapping $y \mapsto f_y$ is not linear. While it is true that

$$f_{y_1 + y_2}(x) = \langle y_1 + y_2, x\rangle = (f_{y_1} + f_{y_2})(x)$$

and hence $f_{y_1 + y_2} = f_{y_1} + f_{y_2}$, we also have

$$f_{\alpha y}(x) = \langle \alpha y, x\rangle = \alpha^*\langle y, x\rangle = \alpha^* f_y(x)$$

so that $f_{\alpha y} = \alpha^* f_y$. This shows that the map $y \mapsto f_y$ is really a norm preserving *antilinear* mapping of H into H*. We also note that

$$\|f_{y_1} - f_{y_2}\| = \|f_{y_1 - y_2}\| = \|y_1 - y_2\|$$

which shows that the map $y \mapsto f_y$ is an isometry.

What we have shown so far is that given any $y \in H$, there exists a linear functional $f_y \in H^*$ where the association $y \mapsto f_y = \langle y, \ \rangle$ is a norm preserving antilinear map. It is of great importance that this mapping is actually an isomorphism of H *onto* H*. In other words, any element of H* may be written in the form $f_y = \langle y, \ \ \rangle$ for a unique $y \in H$. We now prove this result, which is a somewhat restricted form of the Riesz representation theorem.

Theorem 12.26 (**Riesz Representation Theorem**) Let H be a Hilbert space. Then given any $f \in H^*$, there exists a unique $y \in H$ such that

$$f(x) = \langle y, x\rangle \tag{7}$$

for every $x \in H$. Moreover we have $\|y\| = \|f\|$.

Proof Assuming the existence of such a $y \in H$, it is easy to show that it must be unique. To see this, we simply note that if $f(x) = \langle y_1, x\rangle$ and $f(x) = \langle y_2, x\rangle$ for every $x \in H$, then $0 = \langle y_1, x\rangle - \langle y_2, x\rangle$ implies $0 = \langle y_1 - y_2, x\rangle$. But this holds for all $x \in H$, and hence we must have $y_1 - y_2 = 0$ which implies $y_1 = y_2$. (If $\langle y, x\rangle = 0$ for all $x \in H$, then in particular, $0 = \langle y, y\rangle = \|y\|^2$ implies $y = 0$.)

We now prove that such a y does indeed exist. First note that if $f = 0$, then we may choose $y = 0$ to satisfy the theorem. Therefore we now assume that $f \neq 0$. Let $M = \text{Ker } f$. We know that Ker f is a subspace of H, and since $f \neq 0$ we must have $M \neq H$. Furthermore, if $\{x_n\}$ is a sequence in M that converges to some $x \in H$, then the continuity of f shows that

$$f(x) = f(\lim x_n) = \lim f(x_n) = 0$$

and hence $x \in M$. This shows that M is a proper closed subspace of H (by Theorem B14(a)). By Theorem 12.15, there exists a nonzero $y_0 \in H$ such that $y_0 \perp M$. We claim that $y = \alpha y_0$ will satisfy our requirements for a suitably chosen scalar α.

First note that for any scalar α and any $x \in M$, we have $f(x) = 0$ on the one hand, while $\langle \alpha y_0, x\rangle = \alpha^*\langle y_0, x\rangle = 0$ on the other (since $y_0 \perp x$). This shows that (7) is true for every $x \in M$ no matter how we choose α. However, if we now require that (7) hold for the vector $x = y_0$ (where $y_0 \notin M$ by definition), then we must also have

$$f(y_0) = \langle \alpha y_0, y_0\rangle = \alpha^*\|y_0\|^2$$

which leads us to choose $\alpha = f(y_0)^*/\ \|y_0\|^2$. With this choice of α, we have then shown that (7) holds for all $x \in M$ and for the vector $x = y_0$. We now show that in fact (7) holds for every $x \in H$.

We observe that any $x \in H$ may be written as

$$x = x - [f(x)/f(y_0)]y_0 + [f(x)/f(y_0)]y_0$$

where $x - [f(x)/f(y_0)]y_0 \in M$. In other words, any $x \in H$ may be written in the form $x = m + \beta y_0$ where $m \in M$ and $\beta = f(x)/f(y_0)$. Since f is linear, we now see that our previously shown special cases result in (setting $y = \alpha y_0$)

$$\begin{aligned} f(x) &= f(m+\beta y_0) = f(m) + \beta f(y_0) = \langle y, m\rangle + \beta\langle y, y_0\rangle \\ &= \langle y, m + \beta y_0\rangle = \langle y, x\rangle \end{aligned}$$

and hence $f(x) = \langle y, x\rangle$ for every $x \in H$.

Finally, the fact that $\|y\| = \|f\|$ was shown in the discussion prior to the theorem. ∎

If H is a Hilbert space, we define an inner product on H* by

$$\langle f_y , f_x \rangle = \langle x, y \rangle \ .$$

Note the order of the vectors x and y in this definition. This is to ensure that the inner product on H* has the correct linearity properties. In other words, using the fact that the mapping $y \mapsto f_y$ is antilinear, we have

$$\langle \alpha f_y, f_x \rangle = \langle f_{\alpha^* y}, f_x \rangle = \langle x, \alpha^* y \rangle = \alpha^* \langle x, y \rangle = \alpha^* \langle f_y, f_x \rangle$$

and

$$\langle f_y, \alpha f_x \rangle = \langle f_y, f_{\alpha^* x} \rangle = \langle \alpha^* x, y \rangle = \alpha \langle x, y \rangle = \alpha \langle f_y, f_x \rangle \ .$$

Using $f_{y_1} + f_{y_2} = f_{y_1 + y_2}$ it is trivial to verify that

$$\langle f_{y_1} + f_{y_2}, f_x \rangle = \langle f_{y_1}, f_x \rangle + \langle f_{y_2}, f_x \rangle \ .$$

This inner product induces a norm on H* in the usual way.

We claim that H* is also a Hilbert space. To see this, let $\{f_{x_i}\}$ be a Cauchy sequence in H*. Given $\varepsilon > 0$, there exists N such that $m > n \geq N$ implies that $\|f_{x_m} - f_{x_n}\| < \varepsilon$. Then, since every f_{x_n} corresponds to a unique x_n (the kernel of the mapping $x \rightarrow f_x$ is $\{0\}$), it should be obvious that $\{x_n\}$ will be a Cauchy sequence in H. However, we can also show this directly as follows:

$$\begin{aligned}
\|f_{x_m} - f_{x_n}\|^2 &= \langle f_{x_m} - f_{x_n}, f_{x_m} - f_{x_n} \rangle \\
&= \langle f_{x_m}, f_{x_m} \rangle - \langle f_{x_m}, f_{x_n} \rangle - \langle f_{x_n}, f_{x_m} \rangle + \langle f_{x_n}, f_{x_n} \rangle \\
&= \langle x_m, x_m \rangle - \langle x_m, x_n \rangle - \langle x_n, x_m \rangle + \langle x_n, x_n \rangle \\
&= \langle x_m - x_n, x_m \rangle - \langle x_m - x_n, x_n \rangle \\
&= \langle x_m - x_n, x_m - x_n \rangle \\
&= \|x_m - x_n\|^2 \ .
\end{aligned}$$

This shows that $\{x_n\}$ is a Cauchy sequence in H, and hence $x_n \rightarrow x \in$ H. But then $f_{x_n} \rightarrow f_x \in$ H* which shows that H* is complete, and hence H* is also a Hilbert space.

Example 12.10 Recalling the Banach space l_p^n defined in Example 12.7, we shall show that if $1 < p < \infty$ and $1/p + 1/q = 1$, then $(l_p^n)^* = l_q^n$. By this equality sign, we mean there exists a norm preserving isomorphism of l_q^n onto $(l_p^n)^*$. If $\{e_i\}$ is the standard basis for $\mathbb{R}^n$, then any $x = (x_1, \ldots, x_n) \in l_p^n$ may be written in the form

$$x = \sum_{i=1}^{n} x_i e_i \ .$$

Now let f be a linear mapping of l_p^n into any normed space (although we shall be interested only in the normed space $\mathbb{C}$). By Corollary 2 of Theorem 12.12, we know that f is (uniformly) continuous. Alternatively, we can show this directly as follows. The linearity of f shows that

$$f(x) = \sum_{i=1}^{n} x_i f(e_i)$$

and hence

$$\|f(x)\| = \|\sum_{i=1}^{n} x_i f(e_i)\| \le \sum_{i=1}^{n} |x_i|\,\|f(e_i)\| \le \max\{\|f(e_i)\|\}\sum_{i=1}^{n}|x_i| \ .$$

But

$$|x_i|^p \le \sum_{i=1}^{n} |x_i|^p = (\|x\|_p)^p$$

and therefore $|\, x_i| \le \|x\|_p$. If we write $K = \max\{\|f(e_i)\|\}$, then this leaves us with

$$\|f(x)\ \| \ \le \ nK\ \|x\|_p$$

which shows that f is bounded and hence continuous (Theorem 12.12).

We now restrict ourselves to the particular case that $f\colon l_p^n \to \mathbb{C}$, and we then see that the set of all such f's is just the dual space $(l_p^n)^*$. Since f is bounded, we can define the norm of f in the usual way by

$$\|f\| \ = \ \inf\{K > 0\colon |f(x)| \le K\|x\|_p \text{ for all } x \in l_p^n\} \ .$$

Now note that for each $i = 1, \ldots, n$ the result of f applied to e_i is just some scalar $y_i = f(e_i) \in \mathbb{C}$. Since $f(x) = \Sigma_{i=1}^n x_i f(e_i) = \Sigma_{i=1}^n x_i y_i$, we see that specifying each of the y_i's will determine f, and conversely, f determines each of the y_i's. We therefore have an isomorphism $y = (y_1, \ldots, y_n) \to f$ of the space of all n-tuples $y = (y_1, \ldots, y_n) \in \mathbb{C}$ of scalars onto the space $(l_p^n)^*$ of all linear functionals f on l_p^n defined by $f(x) = \Sigma_{i=1}^n x_i y_i$. Because of this isomorphism, we want to know what norm to define on the set of all such y's so that the mapping $y \to f$ is an isometry.

For any $x \in l_p^n$, Hölder's inequality (see Example 12.7) yields

$$\begin{aligned}
|f(x)| = \left|\sum_{i=1}^{n} x_i y_i\right| &\le \sum_{i=1}^{n} |x_i y_i| \\
&\le \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q} \\
&= \|x\|_p \|y\|_q \quad .
\end{aligned}$$

By definition, this implies that $\|f\| \le \|y\|_q$ (since $\|f\|$ is just the greatest lower bound of the set of all bounds for f). We will show that in fact $\|f\| = \|y\|_q$.

Consider the vector $x = (x_1, \ldots, x_n)$ defined by $x_i = 0$ if $y_i = 0$, and $x_i = |y_i|^q/y_i$ if $y_i \neq 0$. Then using the fact that $1/p = 1 - 1/q$ we find

$$\begin{aligned}
\|x\|_p \|y\|_q &= \left(\sum_{i=1}^{n} \frac{|y_i|^{pq}}{|y_i|^p}\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q} \\
&= \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q} \\
&= \left(\sum_{i=1}^{n} |y_i|^q\right)^{1-1/q} \left(\sum_{i=1}^{n} |y_i|^q\right)^{1/q} \\
&= \sum_{i=1}^{n} |y_i|^q \quad .
\end{aligned}$$

On the other hand, we also see that for this x we have

$$|f(x)| = \left|\sum_{i=1}^{n} x_i y_i\right| = \left|\sum_{i=1}^{n} |y_i|^q\right| = \sum_{i=1}^{n} |y_i|^q \quad .$$

Thus, for this particular x, we find that $|f(x)| = \|x\|_p \, \|y\|_q$, and hence in general we must have $\|f\| = \|y\|_q$ (since it should now be obvious that nothing smaller than $K = \|y\|_q$ can satisfy $|f(x)| \le K \, \|x\|_p$ for *all* x).

In summary, defining a norm on the space of all n-tuples $y = (y_1, \ldots, y_n)$ by $\|y\|_q$, we have constructed a norm preserving isomorphism of l_q^n onto $(l_p^n)^*$ as desired.

In the particular case of p = q = 2, we know that l_2^n is a Hilbert space, and hence $(l_2^n)^* = l_2^n$. Note also that in general,

$$(l_p^n)^{**} = (l_q^n)^* = l_p^n \ .$$

Any normed linear space E for which E** = E is said to be **reflexive**. Thus we have shown that l_p^n is a reflexive Banach space, and hence l_2^n is a reflexive Hilbert space. In fact, it is not difficult to use the Riesz representation theorem to show that any Hilbert space is reflexive (see Exercise 12.6.1). ∕∕

Example 12.11 Recall the space l_∞ defined in Exercise 12.3.5, and let c_0 denote the subspace consisting of all convergent sequences with limit 0. In other words, $x = \{x_1, x_2, \dots, x_n, \dots\}$ has the property that $x_n \to 0$ as $n \to \infty$. We shall show that $c_0^{**} = l_1^* = l_\infty$, and hence c_0 is not reflexive.

Let us first show that any bounded linear functional f on c_0 is expressible in the form

$$f(x) = \sum_{i=1}^{\infty} f_i x_i$$

where

$$\sum_{i=1}^{\infty} |f_i| < \infty \ .$$

To see this, let $e_i = \{0, 0, \dots, 1, 0, \dots\}$ be the sequence with a 1 in the *ith* position and 0's elsewhere. Now let f(x) be any bounded linear functional on c_0, and define $f_i = f(e_i)$. Note that if

$$x = \{x_1, x_2, \dots, x_n, 0, \dots\} \qquad (*)$$

then

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$

and

$$f(x) = \sum_{i=1}^{n} f_i x_i \ .$$

Observe that if $\sum_{i=1}^{\infty} |f_i| = \infty$, then for every real B it would be possible to find an integer N such that

$$\sum_{i=1}^{N} |f_i| > B \ .$$

But then consider an element x defined by

$$x_i = \begin{cases} 1 & \text{if } i \le N \text{ and } f_i > 0 \\ -1 & \text{if } i \le N \text{ and } f_i < 0 \\ 0 & \text{if } i \le N \text{ and } f_i = 0 \\ 0 & \text{if } i > N \end{cases} \; .$$

Clearly $\|x\| = \sup |x_i| = 1$, and

$$|f(x)| = \left|\sum_{i=1}^{N} f_i x_i\right| = \sum_{i=1}^{N} |f_i| > B = B\|x\|$$

which contradicts the assumed boundedness of f. Therefore we must have

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

It is not hard to see that the set of all elements of the form (*) is dense in c_0. Indeed, suppose we are given any $z = \{z_1, z_2, \ldots, z_n, \ldots\} \in c_0$. Then given $\varepsilon > 0$, we must find an x of the above form with the property that

$$\|z - x\| = \sup_i |z_i - x_i| < \varepsilon \; .$$

Since any sequence $x \in c_0$ has the property that $x_n \to 0$ as $n \to \infty$, it follows that given $\varepsilon > 0$, there exists M such that $|x_n| < \varepsilon$ for all $n \ge M$. If we choose $x = \{z_1, z_2, \ldots, z_M, 0, \ldots\}$, then clearly we will have $\|z - x\| < \varepsilon$.

By Corollary 1 of Theorem 12.12 any bounded linear functional is continuous. Together with Theorem 12.7(d), this shows that any bounded linear functional on c_0 is uniquely defined by its values on the dense set of elements of the form (*), and hence for every $x \in c_0$ we must have

$$f(x) = \sum_{i=1}^{\infty} f_i x_i \; .$$

We now claim that the norm of any such linear functional is given by

$$\|f\| = \sum_{i=1}^{\infty} |f_i| \; .$$

First note that

$$|f(x)| \le \sum_{i=1}^{\infty} |f_i||x_i| \le \|x\| \sum_{i=1}^{\infty} |f_i| = \|x\| a$$

where

$$a = \sum_{i=1}^{\infty} |f_i| \;.$$

Then

$$\frac{|f(x)|}{\|x\|} \le \sum_{i=1}^{\infty} |f_i| = a$$

and hence

$$\|f\| = \sup\left\{\frac{|f(x)|}{\|x\|} : x \ne 0\right\} \le a \;.$$

On the other hand, it follows from Theorem B17 (see Appendix B) that given ε > 0, there exists N such that

$$a - \varepsilon < \sum_{i=1}^{N} |f_i| \;.$$

If we define x again by

$$x_i = \begin{cases} 1 & \text{if } i \le N \text{ and } f_i > 0 \\ -1 & \text{if } i \le N \text{ and } f_i < 0 \\ 0 & \text{if } i \le N \text{ and } f_i = 0 \\ 0 & \text{if } i > N \end{cases} \;.$$

then $\|x\| = 1$ and

$$f(x) = \sum_{i=1}^{\infty} f_i x_i = \sum_{i=1}^{N} |f_i|$$

so that $|f(x)| > a - \varepsilon$. Therefore $|f(x)|/\|x\| \ge a$ since $\varepsilon > 0$ was arbitrary. But then $\|f\| \ge a$, and hence we must have $\|f\| = a$ as claimed.

In summary, we have shown that $c_0{}^* = l_1$. In Exercise 12.6.2 the reader is asked to show that $l_1{}^* = l_\infty$, and hence this shows that $c_0{}^{**} = l_\infty$. //

We now proceed to define the adjoint of an operator exactly as we did in Section 10.1. Therefore, let H be a Hilbert space and let $T \in \mathcal{L}(H, H)$ be a linear operator on H. If $y \in H$ is an arbitrary but fixed vector, then the mapping $x \mapsto \langle y, Tx\rangle$ is just a linear functional on H. We can thus apply the Riesz representation theorem to conclude that there exists a unique vector $z \in H$ such that $\langle y, Tx\rangle = \langle z, x\rangle$. We now define the **adjoint** $T^\dagger$ of T by $T^\dagger y = z$. Since y was arbitrary, we see that the definition of $T^\dagger$ may be stated as

$$\langle T^\dagger y, x\rangle = \langle y, Tx\rangle$$

for all $x, y \in H$.

To prove $T^\dagger$ is unique, suppose that $T' \in \mathcal{L}(H, H)$ is defined by $\langle T'y, x\rangle = \langle y, Tx\rangle$. Then $\langle T'y, x\rangle = \langle T^\dagger y, x\rangle$ and hence $\langle T'y - T^\dagger y, x\rangle = 0$. But this must hold for all $x \in H$, and therefore $T'y - T^\dagger y = 0$, i.e., $T'y = T^\dagger y$. Since this is true for all y, it then follows that $T' = T^\dagger$.

Let us show that $T^\dagger$ as we have defined it is really an element of $\mathcal{L}(H, H)$. In other words, we must show that $T^\dagger$ is both linear and continuous. But this is easy since for any $x, y, z \in H$ and $\alpha \in \mathbb{C}$ we have

$$\begin{aligned}\langle T^\dagger(x+y), z\rangle &= \langle x+y, Tz\rangle = \langle x, Tz\rangle + \langle y, Tz\rangle = \langle T^\dagger x, z\rangle + \langle T^\dagger y, z\rangle\\ &= \langle T^\dagger x + T^\dagger y, z\rangle\end{aligned}$$

and

$$\langle T^\dagger(\alpha x), y\rangle = \langle \alpha x, Ty\rangle = \alpha^*\langle x, Ty\rangle = \alpha^*\langle T^\dagger x, y\rangle = \langle(\alpha T^\dagger)x, y\rangle\ .$$

Therefore

$$T^\dagger(x + y) = T^\dagger x + T^\dagger y$$

and

$$T^\dagger(\alpha x) = (\alpha T^\dagger)x\ .$$

To prove the continuity of $T^\dagger$, we first show that it is bounded. Using the Cauchy-Schwartz inequality we have

$$\|T^\dagger x\|^2 = \langle T^\dagger x, T^\dagger x\rangle = \langle x, TT^\dagger x\rangle \le \|x\|\,\|TT^\dagger x\| \le \|x\|\,\|T\|\,\|T^\dagger x\|$$

which shows that $\|T^\dagger x\| \le \|T\|\,\|x\|$ for all $x \in H$. Since $\|T\| < \infty$, this shows that $T^\dagger$ is continuous (Theorem 12.12). We can therefore define the norm of $T^\dagger$ in the usual manner to obtain

$$\|T^\dagger\| = \sup\{\|T^\dagger x\|: \|x\| = 1\} \le \|T\|\ .$$

In fact, we will show in the next theorem that $\|T^\dagger\| = \|T\|$.

Theorem 12.27 Suppose S and T are operators on a Hilbert space H. Then there exists a unique linear operator $T^\dagger$ on H defined by $\langle T^\dagger y, x\rangle = \langle y, Tx\rangle$ for all $x, y \in H$. Moreover, this operator has the following properties:

(a) $(S + T)^\dagger = S^\dagger + T^\dagger$.
(b) $(\alpha T)^\dagger = \alpha^* T^\dagger$.
(c) $(ST)^\dagger = T^\dagger S^\dagger$.
(d) $T^{\dagger\dagger} = (T^\dagger)^\dagger = T$.
(e) $\|T^\dagger\| = \|T\|$.

(f) $\|T^\dagger T\| = \|T\|^2$.
(g) $0^\dagger = 0$ and $1^\dagger = 1$.
(h) If T is invertible, then $(T^\dagger)^{-1} = (T^{-1})^\dagger$.

Proof The existence and uniqueness of $T^\dagger$ was shown in the above discussion. Properties (a) – (d) and (g) – (h) follow exactly as in Theorem 10.3. As to property (e), we just showed that $\|T^\dagger\| \le \|T\|$, and hence together with property (d), this also shows that $\|T\| = \|(T^\dagger)^\dagger\| \le \|T^\dagger\|$. To prove (f), we first note that the basic properties of the norm along with property (e) show that

$$\|T^\dagger T\| \le \|T^\dagger\|\,\|T\| = \|T\|^2 \ .$$

To show that $\|T\|^2 \le \|T^\dagger T\|$, we observe that the Cauchy-Schwartz inequality yields

$$\|Tx\|^2 = \langle Tx, Tx\rangle = \langle T^\dagger Tx, x\rangle \le \|T^\dagger Tx\|\,\|x\| \le \|T^\dagger T\|\,\|x\|^2$$

which (by definition of $\|T\|$) implies $\|T\| \le \|T^\dagger T\|^{1/2}$. ∎

While we have defined the adjoint in the most direct manner possible, we should point out that there is a more general approach that is similar to our discussion of the transpose mapping defined in Theorem 9.7. This alternative method shows that a linear operator T defined on a Banach space E leads to a "conjugate" operator T* defined on the dual space E*. Furthermore, the mapping $T \to T^*$ is a norm preserving isomorphism of $\mathcal{L}(E, E)$ into $\mathcal{L}(E^*, E^*)$. However, in the case of a Hilbert space, Theorem 12.26 gives us an isomorphism between H and H*, and hence we can consider T* to be an operator on H itself, and we therefore define $T^\dagger = T^*$. For the details of this approach, the reader is referred to, e.g., the very readable treatment by Simmons (1963).

Exercises

1. Let H be a Hilbert space. We define a mapping $H \to H^{**}$ by $x \mapsto F_x$ where $F_x \in H^{**}$ is defined by $F_x(f) = f(x)$. We can also consider the composite mapping $H \to H^* \to H^{**}$ defined by $x \mapsto f_x \mapsto F_{f_x}$ where $f_x(y) = \langle y, x\rangle$ and $F_{f_x}(f) = \langle f, f_x\rangle$.
   (a) Show that H** is a Hilbert space with the inner product $\langle F_f, F_g\rangle = \langle g, f\rangle$.
   (b) By considering the two mappings defined above, show that H is reflexive.
   (c) Show $\langle F_x, F_y\rangle = \langle x, y\rangle$.

2. (a) Show $(l_1{}^n)^* = l_\infty^n$ and $(l_\infty^n)^* = l_1$.
   (b) Show $l_1{}^* = l_\infty$ and $l_\infty{}^* = l_1$. [*Hint*: Refer to Example 12.11.]

3. Let V be infinite-dimensional with orthonormal basis $\{e_i\}$. Define $T \in L(V)$ by $Te_i = e_{i-1}$. Show $T^\dagger e_i = e_{i+1}$.

12.7 HERMITIAN, NORMAL AND UNITARY OPERATORS

Let us denote the space $\mathcal{L}(H, H)$ of continuous linear maps of H into itself by $\mathcal{L}(H)$. In other words, $\mathcal{L}(H)$ consists of all operators on H. As any physics student knows, the important operators $A \in \mathcal{L}(H)$ are those for which $A^\dagger = A$. These operators are called **self-adjoint** or **Hermitian**. In fact, we now show that the set of all Hermitian operators on H is a closed subspace of $\mathcal{L}(H)$.

Theorem 12.28 The set of all Hermitian operators on a Hilbert space H forms a real closed subspace of $\mathcal{L}(H)$. Moreover, this subspace is a real Banach space containing the identity transformation on H.

Proof We showed in Theorem 12.27 that 0 and 1 are Hermitian. If $A, B \in \mathcal{L}(H)$ are Hermitian operators and $\alpha, \beta \in \mathbb{R}$, then

$$(\alpha A + \beta B)^\dagger = (\alpha A)^\dagger + (\beta B)^\dagger = \alpha A^\dagger + \beta B^\dagger = \alpha A + \beta B$$

so that $\alpha A + \beta B$ is also Hermitian, and hence the set of Hermitian operators forms a real subspace of $\mathcal{L}(H)$. If $\{A_n\}$ is a sequence of Hermitian operators with the property that $A_n \to A \in \mathcal{L}(H)$, then (using $A_n{}^\dagger = A_n$ and Theorem 12.27(e))

$$\begin{aligned} \|A - A^\dagger\| &\le \|A - A_n\| + \|(A_n - A)^\dagger\| \\ &= \|A - A_n\| + \|A_n - A\| \\ &= 2\|A_n - A\| \;. \end{aligned}$$

Since this shows that $\|A - A^\dagger\| \to 0$ as $n \to \infty$, we see that $A = A^\dagger$ and hence A is also Hermitian. Therefore the subspace of all Hermitian operators on H is closed (Theorem B14(a)).

Finally, since $\mathcal{L}(H)$ is a Banach space (Theorem 12.13), the fact that the closed subspace of Hermitian operators forms a real Banach space follows from Theorem 12.9. ∎

It should be clear by now that most of the basic properties of Hermitian operators on an infinite-dimensional Hilbert space are exactly the same as in the finite-dimensional case discussed in Chapter 10. In particular, the proofs of Theorems 10.4, 10.9(a) and 10.11(a) all carry over verbatim to the infinite-dimensional case.

Recall from Section 10.3 that an operator N on H is said to be **normal** if N and $N^\dagger$ commute, i.e., if $N^\dagger N = NN^\dagger$. It should be obvious that any Hermitian operator is necessarily normal, and that αN is normal for any scalar α. However, even if N_1 and N_2 are normal, it is not generally true that either $N_1 + N_2$ or N_1N_2 are normal, and hence the subset of $\mathcal{L}(H)$ consisting of normal operators is not a subspace. We do however have the following two results.

Theorem 12.29 The set of all normal operators on H is a closed subset of $\mathcal{L}(H)$ that is also closed under scalar multiplication.

Proof All that remains to be shown is that if $\{N_k\}$ is a sequence of normal operators that converges to an operator $N \in \mathcal{L}(H)$, then N is normal. Since $N_k \to N$, it follows from Theorem 12.27(a) and (e) that $N_k{}^\dagger \to N^\dagger$. We thus have (using the fact that each N_k is normal)

$$\begin{aligned}\|NN^\dagger - N^\dagger N\| &\le \|NN^\dagger - N_kN_k{}^\dagger\| + \|N_kN_k{}^\dagger - N_k{}^\dagger N_k\| + \|N_k{}^\dagger N_k - N^\dagger N\| \\ &= \|NN^\dagger - N_kN_k{}^\dagger\| + \|N_k{}^\dagger N_k - N^\dagger N\| \to 0\end{aligned}$$

which shows that $NN^\dagger = N^\dagger N$, and hence N is normal. ∎

Theorem 12.30 Let N_1 and N_2 be normal operators with the property that one of them commutes with the adjoint of the other. Then $N_1 + N_2$ and N_1N_2 are normal.

Proof Suppose that $N_1N_2{}^\dagger = N_2{}^\dagger N_1$. Taking the adjoint of both sides of this equation then shows that $N_2N_1{}^\dagger = N_1{}^\dagger N_2$. In other words, the hypothesis of the theorem is equivalent to the statement that both operators commute with the adjoint of the other. The rest of the proof is left to the reader (see Exercise 12.7.1). ∎

Probably the most important other type of operator that is often defined on a Hilbert space is the unitary operator. We recall that unitary and isometric operators were defined in Section 10.2, and we suggest that the reader again go through that discussion. Here we will repeat the essential content of that earlier treatment in a concise manner.

We say that an operator $\Omega \in \mathcal{L}$ (H) is **isometric** if $\|\Omega x\| = \|x\|$ for every $x \in$ H. Note that we do not require that Ω map H *onto* H, and hence Ω^{-1} need not exist (at least if we assume that Ω^{-1} must be defined on all of H). The definition of an isometric operator shows that $\|\Omega x\| = 0$ if and only if $x = 0$, and hence an isometric operator Ω is a one-to-one mapping of H *into* H (since $\Omega x = \Omega y$ implies $\|\Omega(x - y)\| = 0$ which then implies $x = y$).

Theorem 12.31 If $\Omega \in \mathcal{L}(H)$, then the following conditions are equivalent:

(a) $\Omega^\dagger\Omega = 1$.

(b) $\langle \Omega x, \Omega y\rangle = \langle x, y\rangle$ for all $x, y \in H$.

(c) $\|\Omega x\| = \|x\|$.

Proof Let $x, y \in H$ be arbitrary.

(a) ⇒ (b): $\langle \Omega x, \Omega y\rangle = \langle x, \Omega^\dagger\Omega y\rangle = \langle x, 1y\rangle = \langle x, y\rangle$.

(b) ⇒ (c): $\|\Omega x\|^2 = \langle \Omega x, \Omega x\rangle = \langle x, x\rangle = \|x\|^2$.

(c) ⇒ (a): $\|\Omega x\|^2 = \|x\|^2$ implies $\langle \Omega x, \Omega x\rangle = \langle x, x\rangle$ or $\langle x, \Omega^\dagger\Omega x\rangle = \langle x, x\rangle$, and hence $\langle x, (\Omega^\dagger\Omega - 1)x\rangle = 0$. It now follows from Theorem 10.4(b) that $\Omega^\dagger\Omega - 1 = 0$, and hence $\Omega^\dagger\Omega = 1$. ∎

Isometric operators are sometimes defined by the relationship $\Omega^\dagger\Omega = 1$, and we saw in Section 10.2 that in a finite-dimensional space this implies that $\Omega\Omega^\dagger = 1$ also. However, in an infinite-dimensional space, the property $\Omega\Omega^\dagger = 1$ must be imposed as an additional condition on Ω in one way or another. A **unitary** operator $U \in \mathcal{L}(H)$ is an operator that satisfies $U^\dagger U = UU^\dagger = 1$. Since inverses are unique (if they exist), this implies that an equivalent definition of unitary operators is that they map H onto itself and satisfy $U^\dagger = U^{-1}$. Alternatively, our next theorem shows that we can define a unitary operator as an isometric operator that maps H *onto* H.

Theorem 12.32 An operator $U \in \mathcal{L}(H)$ is unitary if and only if it is a one-to-one isometry of H onto itself. In other words, $U \in \mathcal{L}(H)$ is unitary if and only if it is a bijective isometry.

Proof If U is unitary then it maps H onto itself, and since $U^\dagger U = 1$, we see from Theorem 12.31 that $\|Ux\| = \|x\|$. Therefore U is an isometric isomorphism of H onto H.

Conversely, if U is an isomorphism of H onto H then U^{-1} exists, and the fact that U is isometric shows that $U^\dagger U = 1$ (Theorem 12.31). Multiplying from the right by U^{-1} shows that $U^\dagger = U^{-1}$, and hence $U^\dagger U = UU^\dagger = 1$ so that U is unitary. ∎

One reassuring fact about unitary operators in a Hilbert space is that they also obey the analogue of Theorem 10.6. In other words, an operator U on a Hilbert space H is unitary if and only if $\{Ue_i\}$ is a complete orthonormal set whenever $\{e_i\}$ is (see Exercise 12.7.2 for a proof).

There is no all-encompassing treatment of eigenvalues (i.e., like Theorems 10.21 or 10.26) for Hermitian or unitary operators in an infinite-dimensional space even close to that for finite-dimensional spaces. Unfortunately, most of the general results that are known are considerably more difficult to treat in the infinite-dimensional case. In fact, a proper treatment involves a detailed discussion of many subjects which the ambitious reader will have to study on his or her own.

Exercises

1. Finish the proof of Theorem 12.30.

2. Prove that $U \in \mathcal{L}(H)$ is unitary if and only if $\{Ue_i\}$ is a complete orthonormal set if $\{e_i\}$ is.

APPENDIX A

Metric Spaces

For those readers not already familiar with the elementary properties of metric spaces and the notion of compactness, this appendix presents a sufficiently detailed treatment for a reasonable understanding of this subject matter. However, for those who have already had some exposure to elementary point set topology (or even a solid introduction to real analysis), then the material in this appendix should serve as a useful review of some basic concepts. Besides, any mathematics or physics student should become thoroughly familiar with all of this material.

Let S be any set. Then a function $d: S \times S \rightarrow \mathbb{R}$ is said to be a **metric** on S if it has the following properties for all $x, y, z \in S$:

(M1) $d(x, y) \geq 0$;
(M2) $d(x, y) = 0$ if and only if $x = y$;
(M3) $d(x, y) = d(y, x)$;
(M4) $d(x, y) + d(y, z) \geq d(x, z)$.

The real number $d(x, y)$ is called the **distance** between x and y, and the set S together with a metric d is called a **metric space** (S, d).

As a simple example, let $S = \mathbb{R}$ and let $d(x, y) = |x - y|$ for all $x, y \in \mathbb{R}$. From the properties of the absolute value, conditions (M1) - (M3) should be obvious, and (M4) follows by simply noting that

$$|x - z| = |x - y + y - z| \leq |x - y| + |y - z| \ .$$

For our purposes, we point out that given any normed vector space $(V, \| \ \|)$ we may treat V as a metric space by defining

$$d(x, y) = \|x - y\|$$

for every $x, y \in V$. Using Theorem 2.17, the reader should have no trouble showing that this does indeed define a metric space (V, d). In fact, it is easy to see that $\mathbb{R}^n$ forms a metric space relative to the standard inner product and its associated norm.

Given a metric space (X, d) and any real number $r > 0$, the **open ball** of **radius** r and **center** x_0 is the set $B_d(x_0, r) \subset X$ defined by

$$B_d(x_0, r) = \{x \in X: d(x, x_0) < r\} \ .$$

Since the metric d is usually understood, we will generally leave off the subscript d and simply write $B(x_0, r)$. Such a set is frequently referred to as an **r-ball**. We say that a subset U of X is **open** if, given any point $x \in U$, there exists $r > 0$ and an open ball $B(x, r)$ such that $B(x, r) \subset U$.

Probably the most common example of an open set is the open unit disk D_1 in $\mathbb{R}^2$ defined by

$$D_1 = \{(x, y) \in \mathbb{R}^2: x^2 + y^2 < 1\} \ .$$

We see that given any point $x_0 \in D_1$, we can find an open ball $B(x_0, r) \subset D_1$ by choosing $r = 1 - d(x_0, 0)$. The set

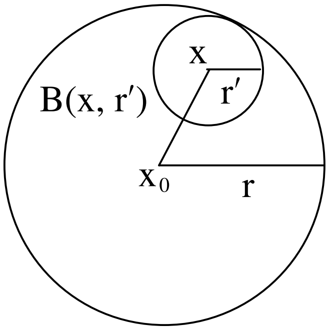
$$D_2 = \{(x, y) \in \mathbb{R}^2: x^2 + y^2 \leq 1\}$$

is not open because there is no open ball centered on any of the boundary points $x^2 + y^2 = 1$ that is contained entirely within D_2.

The fundamental characterizations of open sets are contained in the following three theorems.

Theorem A1 Let (X, d) be a metric space. Then any open ball is an open set.

Proof Let $B(x_0, r)$ be an open ball in X and let x be any point in $B(x_0, r)$. We must find a $B(x, r')$ contained in $B(x_0, r)$.



$B(x_0, r)$

Since $d(x, x_0) < r$, we define $r' = r - d(x, x_0)$. Then for any $y \in B(x, r')$ we have $d(y, x) < r'$, and hence

$$d(y, x_0) \le d(y, x) + d(x, x_0) < r' + d(x, x_0) = r$$

which shows that $y \in B(x_0, r)$. Therefore $B(x, r') \subset B(x_0, r)$. ∎

Theorem A2 Let (X, d) be a metric space. Then

(a) Both X and $\varnothing$ are open sets.
(b) The intersection of a finite number of open sets is open.
(c) The union of an arbitrary number of open sets is open.

Proof (a) X is clearly open since for any $x \in X$ and $r > 0$ we have $B(x, r) \subset X$. The statement that $\varnothing$ is open is also automatically satisfied since for any $x \in \varnothing$ (there are none) and $r > 0$, we again have $B(x, r) \subset \varnothing$.

(b) Let $\{U_i\}$, $i \in I$, be a finite collection of open sets in X. Suppose $\{U_i\}$ is empty. Then $\cap U_i = X$ because a point is in the intersection of a collection of sets if it belongs to each set in the collection, so if there are no sets in the collection, then every point of X satisfies this requirement. Hence $\cap U_i = X$ is open by (a). Now assume that $\{U_i\}$ is not empty, and let $U = \cap U_i$. If $U = \varnothing$ then it is open by (a), so assume that $U \ne \varnothing$. Suppose $x \in U$ so that $x \in U_i$ for every $i \in I$. Therefore there exists $B(x, r_i) \subset U_i$ for each i, and since there are only a *finite* number of the r_i we may let $r = \min\{r_i\}$. It follows that

$$B(x, r) \subset B(x, r_i) \subset U_i$$

for every i, and hence $B(x, r) \subset \cap U_i = U$. In other words, we have found an open ball centered at each point of U and contained in U, thus proving that U is open.

(c) Let $\{U_i\}$ be an arbitrary collection of open sets. If $\{U_i\}$ is empty, then $U = \cup U_i = \varnothing$ is open by (a). Now suppose that $\{U_i\}$ is not empty and $x \in \cup U_i$. Then $x \in U_i$ for some i, and hence there exists $B(x, r_i) \subset U_i \subset \cup U_i$ so that $\cup U_i$ is open. ∎

Notice that part (b) of this theorem requires that the collection be finite. To see the necessity of this, consider the infinite collection of intervals in $\mathbb{R}$ given by $(-1/n, 1/n)$ for $1 \leq n < \infty$. The intersection of these sets is the point $\{0\}$ which is not open in $\mathbb{R}$.

In an arbitrary metric space the structure of the open sets can be very complicated. However, the most general description of an open set is contained in the following.

Theorem A3 A subset U of a metric space (X, d) is open if and only if it is the union of open balls.

Proof Assume U is the union of open balls. By Theorem A1 each open ball is an open set, and hence U is open by Theorem A2(c). Conversely, let U be an open subset of X. For each $x \in U$ there exists at least one $B(x, r) \subset U$, so that $\cup_{x \in U} B(x, r) \subset U$. On the other hand each $x \in U$ is contained in at least $B(x, r)$ so that $U \subset \cup_{x \in U} B(x, r)$. Therefore $U = \cup B(x, r)$. ∎

As a passing remark, note that a set is never open in and of itself. Rather, a set is open only with respect to a specific metric space containing it. For example, the set of numbers $[0, 1)$ is not open when considered as a subset of the real line because any open interval about the point 0 contains points not in $[0, 1)$. However, if $[0, 1)$ is considered to be the entire space X, then it is open by Theorem A2(a).

If U is an open subset of a metric space (X, d), then its complement $U^c = X - U$ is said to be **closed**. In other words, a set is closed if and only if its complement is open. For example, a moments thought should convince you that the subset of $\mathbb{R}^2$ defined by $\{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 \leq 1\}$ is a closed set. The **closed ball** of radius r centered at x_0 is the set $B[x_0, r]$ defined in the obvious way by

$$B[x_0, r] = \{x \in X \colon d(x_0, x) \leq r\} \;.$$

We leave it to the reader (see Exercise A.3) to prove the closed set analogue of Theorem A2. The important difference to realize is that the intersection of an arbitrary number of closed sets is closed, while only the union of a finite number of closed sets is closed.

If (X, d) is a metric space and $Y \subset X$, then Y may be considered a metric space in its own right with the same metric d used on X. In other words, if we let $d|Y$ denote the metric d restricted to points in Y, then the space $(Y, d|Y)$ is said to be a **subspace** of the metric space (X, d).

Theorem A4 Let (X, d) be a metric space and $(Y, d|Y)$ a metric subspace of X. Then a subset $W \subset Y$ is open in Y (i.e., open with respect to the metric $d|Y$) if and only if $W = Y \cap U$ where U is open in X.

Proof Let $W \subset Y$ be open in Y and suppose $x \in W$. Then there exists $r > 0$ such that the set

$$B_{d|Y}(x, r) = \{y \in Y: (d|Y)(x, y) < r\}$$

is a subset of W. But this is clearly the same as the open set

$$B_d(x, r) = \{y \in X: d(x, y) < r\}$$

restricted to only those points y that are in Y. Another way of saying this is that

$$B_{d|Y}(x, r) = B_d(x, r) \cap Y \ .$$

Since $W = \cup_{x \in W} B_{d|Y}(x, r)$, it follows that (see Exercise 0.1.1(b)) $W = U \cap Y$ where $U = \cup_{x \in W} B_d(x, r)$ is open in X (by Theorem A2(c)).

On the other hand, let $W = Y \cap U$ where U is open in X, and suppose $x \in W$. Then $x \in U$ so there exists $r > 0$ with

$$B_d(x, r) = \{y \in X: d(x, y) < r\} \subset U \ .$$

But $Y \cap B_d(x, r)$ is just $B_{d|Y}(x, r) = \{y \in Y: (d|Y)(x, y) < r\} \subset W$ which shows that W is open in Y. ∎

Note that all of our discussion on metric spaces also applies to normed vector spaces where $d(x, y) = \|x - y\|$. Because of this, we can equally well discuss open sets in any normed space V.

Let $f: (X, d_X) \rightarrow (Y, d_Y)$ be a mapping. We say that f is **continuous at** $x_0 \in X$ if, given any real number $\varepsilon > 0$, there exists a real number $\delta > 0$ such that $d_X(f(x), f(x_0)) < \varepsilon$ for every $x \in X$ with $d_Y(x, x_0) < \delta$. Equivalently, f is continuous at x_0 if for each $B(f(x_0), \varepsilon)$ there exists $B(x_0, \delta)$ such that $f(B(x_0, \delta)) \subset B(f(x_0), \varepsilon)$. (Note that these open balls are defined with respect to two different metrics since they are in different spaces. We do not want to clutter the notation by adding subscripts such as d_X and d_Y to B.) In words, “if you tell me how close you wish to get to the number $f(x_0)$, then I will tell you how close x must be to x_0 in order that $f(x)$ be that close.” If f is defined on a subset $S \subset X$, then f is said to be **continuous on** S if f is continuous at every point of S.

For example, consider the mapping $f: (0, \infty) \subset \mathbb{R} \rightarrow (0, \infty) \subset \mathbb{R}$ defined by $f(x) = 1/x$. For any $x_0 \in (0, \infty)$ we have (using the absolute value as our metric)

$$|f(x) - f(x_0)| = |1/x - 1/x_0| = |x_0 - x|/|x\, x_0| \ .$$

If x is such that $|x - x_0| < \delta$, then we see that

$$|f(x) - f(x_0)| < \delta/|x\, x_0| = \delta/(x\, x_0) \ .$$

In particular, choosing $\delta \le x_0/2$, it follows that $x \ge x_0/2$ (since $|x - x_0| < \delta = x_0/2$), and hence $\delta/(x\, x_0) < 2\delta/x_0^2$. Therefore, given any $\varepsilon > 0$, if we pick $\delta = \min\{x_0/2, \varepsilon x_0^2/2\}$ then we will have $|f(x) - f(x_0)| < \varepsilon$.

Fortunately one can usually tell by inspection (i.e., by drawing a picture if necessary) whether or not a particular function is continuous without resorting to clever calculations. The general definition is a powerful technique for proving theorems about classes of continuous functions satisfying given properties. Moreover, there is an intrinsic way to characterize continuous mappings that is of the utmost importance.

Theorem A5 Let $f: (X, d_X) \rightarrow (Y, d_Y)$. Then f is continuous if and only if $f^{-1}(U)$ is open in X for all open sets U in Y.

Proof Suppose f is continuous and U is an open subset of Y. If $x \in f^{-1}(U)$, then $f(x) \in U$ so there exists $B(f(x), \varepsilon) \subset U$ (since U is open). But the continuity of f then implies that there exists $B(x, \delta)$ such that

$$f(B(x, \delta)) \subset B(f(x), \varepsilon) \subset U \ .$$

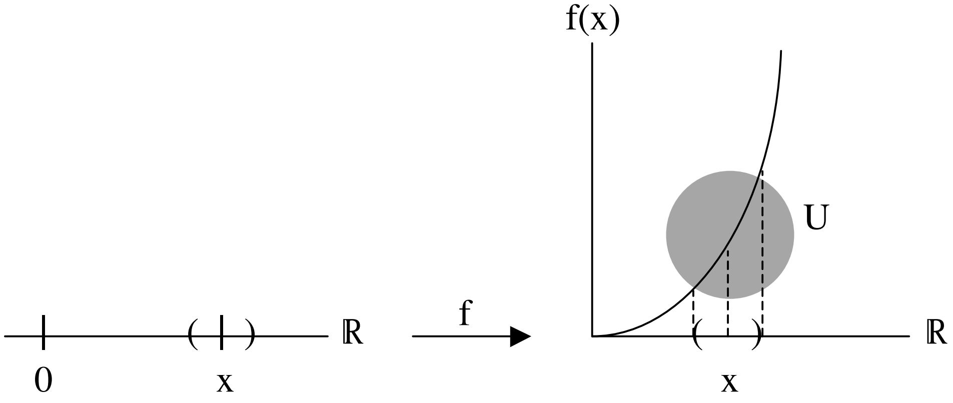
Therefore $B(x, \delta)$ is an r-ball centered on x and contained in $f^{-1}(U)$, and hence $f^{-1}(U)$ is open.

Conversely, assume that $f^{-1}(U)$ is open whenever U is, and let $x \in X$ be arbitrary. Then the open ball $B(f(x), \varepsilon)$ is an open set, so its inverse image is an open set containing x. Therefore there exists an open ball $B(x, \delta)$ contained in this inverse image, and it clearly has the property that $f(B(x, \delta)) \subset B(f(x), \varepsilon)$, hence proving that f is continuous. ∎

Corollary If $f: (X, d_X) \rightarrow (Y, d_Y)$, then f is continuous if and only if $f^{-1}(F)$ is closed in X whenever F is closed in Y.

Proof It was shown in Exercise 0.2.1 that if $A \subset Y$, then $f^{-1}(A^c) = f^{-1}(A)^c$. Therefore if $F \subset Y$ is closed, then $F = U^c$ for some open set $U \subset Y$ and so by Theorem A5, $f^{-1}(F) = f^{-1}(U^c) = f^{-1}(U)^c$ must be closed if and only if f is continuous. ∎

Note that if $f: X \to Y$ is continuous and $U \subset Y$ is open, then $f^{-1}(U)$ is open, but if $A \subset X$ is open, then it is not necessarily true that f(A) is open. As a simple example, consider the function $f: \mathbb{R} \to \mathbb{R}^2$ defined by $f(x) = (x, x^2)$. It should be clear that the open ball $U \subset \mathbb{R}^2$ shown below is an open set whose inverse image is an open interval on $\mathbb{R} \cup \{\varnothing\}$ (since some points of U are not the image under f of any point in $\mathbb{R}$), but that the image under f of an open interval is part of the parabola $y = x^2$ which is not open as a subset of $\mathbb{R}^2$.



Now suppose that (X, d) is a metric space, and let $\{U_i\}$ be a collection of open subsets of X such that $\cup U_i = X$. Such a collection of subsets is called an **open cover** of X. A subcollection $\{V_j\}$ of the collection $\{U_i\}$ is said to be an **open subcover** of X if $\cup V_j = X$. A space (X, d) is said to be **compact** if *every* open cover has a finite subcover. Similarly, given a subset $A \subset X$, a collection $\{U_i\}$ of open subsets of X with the property that $A \subset \cup U_i$ is said to be an **open cover** of A. Equivalently, the collection $\{U_i\}$ of open subsets of X is an open cover of A in X if the collection $\{U_i \cap A\}$ is an open cover of the subset A in the metric d|A (i.e., in the subspace A). We then say that A is **compact** if every open cover of A has a finite subcover, or equivalently, A is compact if the subspace A is compact. While this is not a particularly easy concept to thoroughly understand and appreciate without detailed study, its importance to us is based on the following two examples.

Example A1 Consider the subset $A = (0, 1)$ of the real line $\mathbb{R}$. We define the collection $\{U_1, U_2, \ldots\}$ of open sets by

$$U_n = (1/2^{n+1}, 1 - 1/2^{n+1}) \ .$$

Thus $U_1 = (1/4, 3/4)$, $U_2 = (1/8, 7/8)$ etc. The collection $\{U_n\}$ clearly covers A since for any $x \in (0, 1)$ we can always find some U_n such that $x \in U_n$. However, A is not compact since given any finite number of the U_n there exists $\varepsilon > 0$ (so that $\varepsilon \in (0, 1)$) which is not in any of the U_n. ∕∕

Example A2 Let us show that the subspace [0, 1] of the real line is compact. This is sometimes called the **Heine-Borel theorem**, although we shall prove a more general version below.

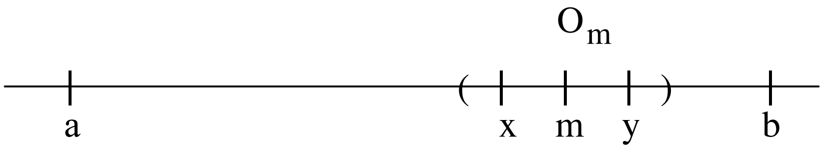
First note that the points 0 and 1 which are included in the subspace [0, 1] are not in the set (0, 1) discussed in the previous example. However, if we have positive real numbers a and b with $a \le b < 1$, then the collection $\{U_n\}$ defined above together with the sets [0, a) and (b, 1] does indeed form an open cover for [0, 1] (the sets [0, a) and (b, 1] are open by Theorem A4). It should be clear that given the sets [0, a) and (b, 1] we can now choose a finite cover of [0, 1] by including these sets along with a finite number of the U_n. To prove that [0, 1] is compact however, we must show that *any* open cover has a finite subcover.

Somewhat more generally, let $\{O_n\}$ be any open cover of the interval [a, b] in ℝ. Define

$$A = \{x \in [a, b]: [a, x] \text{ is covered by a finite number of the } O_n\} \ .$$

We see that $A \neq \varnothing$ since clearly $a \in A$, and furthermore A is bounded above by b. Therefore (by the Archimedean axiom) A must have a least upper bound $m = \sup A \le b$. If A is to be compact, then we must have $b \in A$. We will show that this is true by first proving that $m \in A$, and then that $m = b$.

Since $\{O_n\}$ covers [a, b] and $m \in [a, b]$, it follows that $m \in O_m$ for some $O_m \in \{O_n\}$. Now, O_m is an open subset of [a, b], and hence there are points in O_m that are less than m, and points in O_m that are greater than m.



Since $m = \sup A$, there is an $x < m$ with $x \in O_m$ such that the interval [a, x] is covered by a finite number of the O_n, while [x, m] is covered by the single set O_m. Therefore [a, m] is covered by a finite number of open sets so that $m \in A$.

Now suppose that $m < b$. Then there is a point y with $m < y < b$ such that $[m, y] \subset O_m$. But we just showed that $m \in A$, so the interval $[a, m]$ is covered by finitely many O_n while $[m, y]$ is covered by O_m. Therefore $y \in A$ which contradicts the definition of m, and hence we must have $m = b$. ∕∕

An important property of metric spaces is the following. Given two distinct points $x, y \in X$, there exist disjoint open sets U and V in X such that $x \in U$ and $y \in V$. That this does indeed hold for metric spaces is easy to prove by considering open balls of radius $d(x, y)/2$ centered on each of the points x and y. This property is called the **Hausdorff property**. We sometimes refer to a metric space as a "Hausdorff space" if we wish to emphasize this property.

The following theorems describe some of the most fundamental properties of compact spaces.

Theorem A6 Any closed subset of a compact space is compact.

Proof Let $F \subset X$ be a closed subset of a compact space X. If $\{U_i\}$ is any open cover of F, then $(\cup U_i) \cup F^c$ is an open cover of X. Since X is compact, we may select a finite subcover by choosing F^c along with a finite number of the U_i. But then F is covered by this finite subcollection of the U_i, and hence F is compact. ∎

Theorem A7 Any compact subset of a metric space is closed.

Proof Let F be a compact subset of a metric space X. We will show that F^c is open. Fix any $x \in F^c$ and suppose $y \in F$. Since X is Hausdorff, there exist open sets U_y and V_y such that $x \in U_y$, $y \in V_y$ and $U_y \cap V_y = \varnothing$. As the point y varies over F, we see that $\{V_y\colon y \in F\}$ is an open cover for F. Since F is compact, a finite number, say $V_{y_1}, \dots, V_{y_n}$, will cover F. Corresponding to each V_{y_i} there is a U_{y_i}, and we let $U = \cap_i U_{y_i}$ and $V = \cup_i V_{y_i}$. By construction $x \in U$, $F \subset V$ and $U \cap V = \varnothing$. But then U is an open set containing x such that $U \cap F = \varnothing$, and hence F^c is open. ∎

Theorem A8 Let (X, d_X) be a compact space and let f be a continuous function from X onto a space (Y, d_Y). Then Y is compact.

Proof Let $\{U_i\}$ be any open cover of Y. Since f is continuous, each $f^{-1}(U_i)$ is open in X, and hence $\{f^{-1}(U_i)\}$ is an open cover for X. But X is compact, so that a finite number of the $f^{-1}(U_i)$, say $\{f^{-1}(U_{i_1}), \dots, f^{-1}(U_{i_n})\}$ cover X. Therefore $\{U_{i_1}, \dots, U_{i_n}\}$ form a finite subcover for Y, and hence Y is compact. ∎

Theorem A9 Let $\{K_i\}$ be a collection of compact subsets of a metric space X, such that the intersection of every finite subcollection of $\{K_i\}$ is nonempty. Then $\cap K_i$ is nonempty.

Proof Fix any $K_1 \in \{K_i\}$ and assume that $K_1 \cap (\cap_{i \neq 1} K_i) = \varnothing$. We will show that this leads to a contradiction. First note that by our assumption we have $(\cap_{i \neq 1} K_i) \subset {K_1}^c$, and hence from Example 0.1 and Theorem 0.1 we see that

$$K_1 \subset \cap_{i \neq 1} {K_i}^c = \cup_{i \neq 1} {K_i}^c \ .$$

Thus $\{{K_i}^c\}$, $i \neq 1$, is an open cover of K_1. But K_1 is compact so that a finite number of these sets, say ${K_{i_1}}^c, \ldots, {K_{i_n}}^c$ cover K_1. Then

$$K_1 \subset (\cup_{\alpha=1}^n {K_{i_\alpha}}^c) = (\cap_{\alpha=1}^n K_{i_\alpha})^c$$

which implies

$$K_1 \cap (\cap_{\alpha=1}^n K_{i_\alpha}) = \varnothing \ .$$

However, this contradicts the hypothesis of the theorem. ∎

Corollary If $\{K_n\}$ is a *sequence* of nonempty compact sets such that $K_n \supset K_{n+1}$, then $\cap K_n \neq \varnothing$.

Proof This is an obvious special case of Theorem A9. ∎

As a particular application of this corollary we see that if $\{I_n\}$ is a nonempty sequence of intervals $[a_n, b_n] \subset \mathbb{R}$ such that $I_n \supset I_{n+1,}$ then $\cap I_n \neq \varnothing$. While this result is based on the fact that each I_n is compact (Example A2), we may also prove this directly as follows. If $I_n = [a_n, b_n]$, we let $S = \{a_n\}$. Then $S \neq \varnothing$ and is bounded above by b_1. By the Archimedean axiom, we let $x = \sup S$. For any $m, n \in \mathbb{Z}^+$ we have $a_n \leq a_{m+n} \leq b_{m+n} \leq b_m$ so that $x \leq b_m$ for all m. Since $a_m \leq x$ for all m, we must have $x \in [a_m, b_m] = I_m$ for each $m = 1, 2, \ldots$ so that $\cap I_n \neq \varnothing$. We now show that this result holds in $\mathbb{R}^n$ as well.

Suppose $a, b \in \mathbb{R}^n$ where $a^i < b^i$ for each $i = 1, \ldots, n$. By an **n-cell** we mean the set of all points $x \in \mathbb{R}^n$ such that $a^i \leq x^i \leq b^i$ for every i. In other words, an n-cell is just an n-dimensional rectangle.

Theorem A10 Let $\{I_k\}$ be a sequence of n-cells such that $I_k \supset I_{k+1}$. Then $\cap I_k \neq \varnothing$.

Proof For each $k = 1, 2, \dots$ the n-cell I_k consists of all points $x \in \mathbb{R}^n$ with the property that $a_k{}^i \le x^i \le b_k{}^i$ for every $1 \le i \le n$, so we let $I_k{}^i = [a_k{}^i, b_k{}^i]$. Now, for each $i = 1, \dots, n$ the sequence $\{I_k{}^i\}$ satisfies the hypotheses of the corollary to Theorem A9. Hence for each $i = 1, \dots, n$ there exists $z^i \in [a_k{}^i, b_k{}^i]$ for every $k = 1, 2, \dots$. If we define $z = (z^1, \dots, z^n) \in \mathbb{R}^n$, we see that $z \in I_k$ for every $k = 1, 2, \dots$. ∎

Theorem A11 Every n-cell is compact.

Proof Let I be an n-cell as defined above, and set $\delta = [\sum_{i=1}^n (b_i - a_i)^2]^{1/2}$. Then if $x, y \in I$ we have $\|x - y\| \le \delta$ (see Example 2.9). Let $\{U_i\}$ be any open cover of I and assume that it contains no finite subcover. We will show that this leads to a contradiction.

Let $c^j = (a^j + b^j)/2$ for each $j = 1, \dots, n$. Then we have 2^n n-cells Q_i defined by the intervals $[a^j, c^j]$ and $[c^j, b^j]$ such that $\cup Q_i = I$. Since I has no finite subcover, at least one of the Q_i, which we call I_1, can not be covered by any finite number of the U_i. Next we subdivide I_1 into another 2^n n-cells and continue in the same manner. We thus obtain a sequence $\{I_\alpha\}$ of n-cells with the following properties:

(a) $I \supset I_1 \supset I_2 \supset \cdots$;
(b) I_α is not covered by any finite subcollection of the U_k;
(c) $x, y \in I_\alpha$ implies $\|x - y\| \le 2^{-\alpha}\delta$.

By (a) and Theorem A10, there exists $z \in \cap I_\alpha$, and since $\{U_i\}$ covers I, we must have $z \in U_k$ for some k. Now, U_k is an open set in the metric space $\mathbb{R}^n$, so there exists $\varepsilon > 0$ such that $\|z - y\| < \varepsilon$ implies that $y \in U_k$. If we choose α sufficiently large that $2^{-\alpha}\delta < \varepsilon$ (that this can be done follows from Theorem 0.3), then (c) implies that $I_\alpha \subset U_k$ which contradicts (b). ∎

We are now in a position to prove the generalized Heine-Borel theorem. Before doing so however, we first prove a simple result which is sometimes taken as the definition of a compact set. By way of terminology, any open set U containing a point x is said to be a **neighborhood** of x, and the set $U - \{x\}$ is called a **deleted neighborhood** of x. We say that a point $x \in (X, d)$ is an **accumulation point** of $A \subset X$ if every *deleted* neighborhood of x intersects A.

Theorem A12 Any infinite subset A of a compact set K has a point of accumulation in K.

Proof Suppose every point $x \in K$ is not an accumulation point of A. Then there exists a neighborhood U_x of x such that U_x contains at most a single point of A, namely x itself if $x \in A$. Then clearly no finite subcollection of $\{U_x\}$ covers $A \subset K$ so that K can not possibly be compact. ∎

Theorem A13 A subset A of a metric space (X, d) is closed if and only if A contains all of its accumulation points.

Proof First suppose that A is closed. Let $x \in X$ be an accumulation point of A and assume that $x \notin A$. Then $x \in A^c$ which is an open set containing x that does not intersect A, and hence contradicts the fact that x is an accumulation point of A. Therefore x must be an element of A.

Conversely, suppose A contains all of its accumulation points. We show that A^c is open. If $x \in A^c$ and hence is not an accumulation point of A, then there exists an open set U containing x such that $A \cap U = \varnothing$. But then $x \in U \subset A^c$ which implies that A^c is open. ∎

We say that a subset $A \subset \mathbb{R}^n$ is **bounded** if it can be enclosed in some n-cell. The equivalence of (a) and (b) in the next theorem is called the (**generalized**) **Heine-Borel theorem**, while the equivalence of (a) and (c) is a general version of the **Bolzano-Weierstrass theorem**.

Theorem A14 Let A be a subset of $\mathbb{R}^n$. Then the following three properties are equivalent:

(a) A is closed and bounded .

(b) A is compact.

(c) Every infinite subset of A has a point of accumulation in A.

Proof (a) ⇒ (b): If (a) holds, then A can be enclosed by some n-cell which is compact by Theorem A11. But then A is compact by Theorem A6.

(b) ⇒ (c): This follows from Theorem A12.

(c) ⇒ (a): We assume that every infinite subset of A has an accumulation point in A. Let us first show that A must be bounded. If A is not bounded, then for each positive integer k = 1, 2, . . . we can find an $x_k \in A$ such that $\|x_k\| > k$. Then the set $\{x_k\}$ is clearly infinite but contains no point of accumulation in $\mathbb{R}^n$, so it certainly contains none in A. Hence A must be bounded.

We now show that A must be closed. Again assume the contrary. Then there exists $x_0 \in \mathbb{R}^n$ which is an accumulation point of A but which does not belong to A (Theorem A13). This means that for each k = 1, 2, . . . there exists $x_k \in A$ such that $\|x_k - x_0\| < 1/k$. The set $S = \{x_k\}$ is then an infinite subset of

A with x_0 as an accumulation point. Since $x_0 \notin A$, we will be finished if we can show that S has no accumulation point in A (because the assumption that A is not closed then leads to a contradiction with the property described in (c)).

First note that if a , b $\in \mathbb{R}^n$, then Example 2.11 shows us that

$$\|a + b\| = \|a - (-b)\| \geq \|a\| - \|b\| .$$

Using this result, if y is any point of $\mathbb{R}^n$ other than x_0 we have

$$\begin{aligned} \|x_k - y\| &= \|x_k - x_0 + x_0 - y\| \\ &\geq \|x_0 - y\| - \|x_k - x_0\| \\ &> \|x_0 - y\| - 1/k \ . \end{aligned}$$

No matter how large (or small) $\|x_0 - y\|$ is, we can always find a $k_0 \in \mathbb{Z}^+$ such that $1/k \leq (1/2)\|x_0 - y\|$ for every $k \geq k_0$ (this is just Theorem 0.3). Hence

$$\|x_k - y\| > (1/2)\|x_0 - y\|$$

for every $k \geq k_0$. This shows that y can not possibly be an accumulation point of $\{x_k\} = S$ (because the open ball of radius $(1/2)\|x_0 - y\|$ centered at y can contain at most a finite number of elements of S). ∎

We remark that the implication "(a) implies (b)" in this theorem is not true in an arbitrary metric space (see Exercise A.5).

Let f be a mapping from a set A into $\mathbb{R}^n$. Then f is said to be **bounded** if there exists a real number M such that $\|f(x)\| \leq M$ for all $x \in A$. If f is a continuous mapping from a compact space X into $\mathbb{R}^n$, then f(X) is compact (Theorem A8) and hence closed and bounded (Theorem A14). Thus we see that any continuous function from a compact set into $\mathbb{R}^n$ is bounded. On the other hand, note that the function f: $\mathbb{R} \to \mathbb{R}$ defined by $f(x) = 1/x$ is not bounded on the interval (0, 1). We also see that the function g: $\mathbb{R} \to \mathbb{R}$ defined by $g(x) = x$ for $x \in [0, 1)$ never attains a maximum value, although it gets arbitrarily close to 1. Note that both f and g are defined on non-compact sets.

We now show that a continuous function defined on a compact space takes on its maximum and minimum values at some point of the space.

Theorem A15 Let f be a continuous real-valued function defined on a compact space X, and let $M = \sup_{x \in X} f(x)$ and $m = \inf_{x \in X} f(x)$. Then there exist points $p, q \in X$ such that $f(p) = M$ and $f(q) = m$.

Proof The above discussion showed that f(X) is a closed and bounded subset of $\mathbb{R}$. Hence by the Archimedean axiom, f(X) must have a sup and an inf. Let $M = \sup f(x)$. This means that given $\varepsilon > 0$, there exists $x \in X$ such that

$$M - \varepsilon < f(x) < M$$

(or else M would not be the least upper bound of f(X)). This just says that any open ball centered on M intersects f(X), and hence M is an accumulation point of f(X). But f(X) is closed so that Theorem A13 tells us that $M \in f(X)$. In other words, there exists $p \in X$ such that $M = f(p)$. The proof for the minimum is identical. ∎

As an application of these ideas, we now prove the Fundamental Theorem of Algebra.

Theorem A16 (**Fundamental Theorem of Algebra**) The complex number field $\mathbb{C}$ is algebraically closed.

Proof Consider the non-constant polynomial

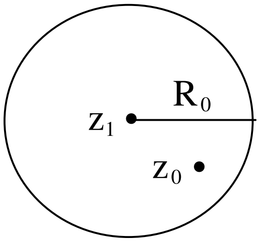
$$f(z) = a_0 + a_1 z + \cdots + a_n z^n \in \mathbb{C}[z]$$

where $a_n \neq 0$. Recall that we view $\mathbb{C}$ as the set $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, and let R be any (finite) real number. Then the absolute value function $|f|: \mathbb{C} \rightarrow \mathbb{R}$ that takes any $z \in \mathbb{C}$ to the real number $|f(z)|$ is continuous on the closed ball $B[0, R]$ of radius R centered at the origin. But $B[0, R]$ is compact (Theorem A14) so that $|f(z)|$ takes its minimum value at some point on the ball (Theorem A15). On the other hand, if we write f(z) in the form

$$f(z) = a_n z^n(a_0/a_n z^n + a_1/a_n z^{n-1} + \cdots + a_{n-1}/a_n z + 1)$$

we see that $|f(z)|$ becomes arbitrarily large as $|z|$ becomes large. To be precise, given any real $C > 0$ there exists $R > 0$ such that $|z| > R$ implies $|f(z)| > C$.

We now combine these two facts as follows. Let z_1 be arbitrary, and define $C = |f(z_1)|$. Then there exists $R_0 > 0$ such that $|f(z)| > |f(z_1)|$ for all $z \in \mathbb{C}$ such that $|z - z_1| > R_0$ (i.e., for all z outside $B[z_1, R_0]$). Since $B[z_1, R_0]$ is compact, there exists a point $z_0 \in B[z_1, R_0]$ such that $|f(z_0)| \leq |f(z)|$ for all $z \in B[z_1, R_0]$.



$B[z_1, R_0]$

In particular, $|f(z_0)| \le |f(z_1)|$ and hence we see that $|f(z_0)| < |f(z)|$ for all $z \in \mathbb{C}$. In other words, z_0 is an absolute minimum of $|f|$. We claim that $f(z_0) = 0$.

To show that $f(z_0) = 0$, we assume that $f(z_0) \ne 0$ and arrive at a contradiction. By a suitable choice of constants c_i, we may write f in the form

$$f(z) = c_0 + c_1(z - z_0) + \cdots + c_n(z - z_0)^n \ .$$

If $f(z_0) \ne 0$ then $c_0 = f(z_0) \ne 0$. By assumption, $\deg f \ge 1$, so we let m be the smallest integer greater than 0 such that $c_m \ne 0$. Defining the new variable $w = z - z_0$, we may define the polynomial function g by

$$f(z) = g(w) = c_0 + c_m w^m + w^{m+1}h(w)$$

for some polynomial h.

Now let w_1 be a complex number such that $w_1{}^m = -c_0/c_m$ and consider all values of $w = \lambda w_1$ for real λ with $0 \le \lambda \le 1$. Then

$$c_m w^m = c_m \lambda^m w_1{}^m = -c_0 \lambda^m$$

and hence

$$\begin{aligned} f(z) = g(\lambda w_1) &= c_0 - \lambda^m c_0 + \lambda^{m+1} w_1{}^{m+1} h(\lambda w_1) \\ &= c_0[1 - \lambda^m + \lambda^{m+1} w_1{}^{m+1} c_0{}^{-1} h(\lambda w_1)] \ . \end{aligned}$$

But $\lambda \in [0, 1]$ which is compact, and hence $|w_1{}^{m+1} c_0{}^{-1} h(\lambda w_1)|$ is a continuous function defined on a compact set. Then the image of this function is a compact subset of $\mathbb{R}$ (Theorem A8) and so is closed and bounded (Theorem A14). This means there exists a number $B > 0$ such that

$$|w_1{}^{m+1} c_0{}^{-1} h(\lambda w_1)| \le B$$

for all $\lambda \in [0, 1]$, and therefore (since $0 \le \lambda \le 1$ implies that $0 \le \lambda^m \le 1$)

$$\begin{aligned}|g(\lambda w_1)| &= |c_0|\,|1-\lambda^m+\lambda^{m+1}w_1{}^{m+1}c_0{}^{-1}h(\lambda w_1)| \\ &\le |c_0|\,\{|1-\lambda^m| + \lambda^{m+1}|w_1{}^{m+1}c_0{}^{-1}h(\lambda w_1)|\} \\ &\le |c_0|\,(1-\lambda^m+\lambda^{m+1}B)\ .\end{aligned}$$

Now recall that $|c_0| = |f(z_0)| \le |f(z)|$ for all $z \in \mathbb{C}$. If we can show that

$$0 < 1-\lambda^m+\lambda^{m+1}B < 1$$

for sufficiently small λ with $0 < \lambda < 1$, then we will have shown that $|f(z)| = |g(\lambda w_1)| < |c_0|$, a contradiction. But it is obvious that λ can be chosen so that $0 < 1 - \lambda^m + \lambda^{m+1}B$. And to require that $1 - \lambda^m + \lambda^{m+1}B < 1$ is the same as requiring that $\lambda B < 1$ which can certainly be satisfied for small enough λ. ∎

Exercises

1. Show the absolute value function is continuous on $\mathbb{R}$.

2. Show that the norm on a vector space V defines a metric on V.

3. Let (X, d) be a metric space. Prove:
   (a) Both X and $\varnothing$ are closed sets.
   (b) The intersection of an arbitrary number of closed sets is closed.
   (c) The union of a finite number of closed sets is closed.

4. Let A be the subset of [0, 1] consisting of all $x \in [0, 1]$ whose decimal expansion contains only the digits 4 and 7. Explain whether or not A is countable, dense in [0, 1], or compact.

5. Show that $\{x\colon \|x\|_2 \le 1\}$ is closed and bounded but not compact in the space l_2 (see Example 12.8).

6. A metric space is said to be **separable** if it contains a countable dense subset. Prove that $\mathbb{R}^n$ is separable. [*Hint*: Consider the set of all points in $\mathbb{R}^n$ with rational coordinates.]

APPENDIX B

Sequences and Series

In this appendix we briefly go through all of the theory necessary for an understanding of Section 10.6 and Chapter 12. Furthermore, as we mentioned at the beginning of Appendix A, rather than being simply an overview, we want the reader to understand this material even if it has not been studied before. We do assume however, that the reader has studied Appendix A, and part of this appendix is a direct extension of that material.

A **sequence** $\{x_n\} = \{x_1, x_2, \dots\}$ in a set S is any function from the set $\mathbb{Z}^+$ of positive integers into S. If (X, d) is a metric space, then a sequence $\{x_n\}$ in X is said to **converge** to the **limit** x if for each $B(x, \varepsilon)$ there exists $N \in \mathbb{Z}^+$ such that $x_n \in B(x, \varepsilon)$ for every $n \geq N$. In other words, given $\varepsilon > 0$, there exists a positive integer N such that $n \geq N$ implies $d(x_n, x) < \varepsilon$. This is usually written as $\lim x_n = x$ or $x_n \to x$. If a sequence $\{x_n\}$ does not converge, then it is said to **diverge**. Furthermore, if for every real number M there exists an integer N such that $n \geq N$ implies $x_n \geq M$, then we write $x_n \to +\infty$. Similarly, if for every real number M there exists an integer N such that $n \geq N$ implies $x_n \leq M$, then we write $x_n \to -\infty$.

(We remark that the small, *always positive* number ε will be used extensively in many proofs, and it is important to realize that for proofs of convergence there is no real difference between the number ε and certain simple functions of ε such as 2ε. For example, suppose we can show that given ε, there exists N such that $n \geq N$ implies $d(x_n, x) < 2\varepsilon$. We claim this also proves that $x_n \to x$. Indeed, let $\varepsilon' = \varepsilon/2$. Then, by assumption, there exists

N' such that $n \geq N'$ implies $d(x_n, x) < 2\varepsilon' = \varepsilon$ which shows that x is the limit of the sequence. It should be clear that the statement "given $\varepsilon > 0$" is equivalent to saying "given $C\varepsilon > 0$" for any finite $C > 0$.)

The set of all points x_n for $n = 1, 2, \ldots$ is called the **range** of the sequence $\{x_n\}$. This may be either a finite or an infinite set of points. A set $A \subset (X, d)$ is said to be **bounded** if there exists a real number M and a point $x_0 \in X$ such that $d(x, x_0) \leq M$ for all $x \in A$. (The point x_0 is almost always taken to be the origin of any given coordinate system in X.) The sequence $\{x_n\}$ is said to be **bounded** if its range is a bounded set. It is easy to show that any convergent sequence is bounded. Indeed, if $x_n \to x$ then, given 1, there exists N such that $n \geq N$ implies $d(x_n, x) < 1$. This shows that $\{x_n: n \geq N\}$ is bounded. To show that $\{x_n: n = 1, \ldots, N - 1\}$ is bounded, let

$$r = \max\{1, d(x_1, x), \ldots, d(x_{N-1}, x)\} .$$

Since $x_n \in X$, it must be true that each $d(x_n, x)$ is finite, and hence $d(x_n, x) \leq r$ for each $n = 1, \ldots, N - 1$.

We now prove several elementary properties of sequences, starting with the uniqueness of the limit.

Theorem B1 If $\{x_n\}$ is a sequence in a metric space (X, d) such that $x_n \to x$ and $x_n \to y$, then $x = y$.

Proof Given $\varepsilon > 0$, there exists N such that $n \geq N$ implies $d(x_n, x) < \varepsilon$ and $d(x_n, y) < \varepsilon$. But then $d(x, y) \leq d(x_n, x) + d(x_n, y) < 2\varepsilon$. Since this holds for all $\varepsilon > 0$, we must have $x = y$ (see Appendix A, definition (M2)). ∎

Theorem B2 Let $s_n \to s$ and $t_n \to t$ be convergent sequences of complex numbers. Then

(a) $\lim (s_n + t_n) = s + t$.

(b) $\lim cs_n = cs$ and $\lim (c + s_n) = c + s$ for any $c \in \mathbb{C}$.

(c) $\lim s_n t_n = st$.

(d) $\lim 1/s_n = 1/s$ if $s \neq 0$ and $s_n \neq 0$ for all n.

Proof (a) Given $\varepsilon > 0$, there exists N_1 and N_2 such that $n \geq N_1$ implies that $|s_n - s| < \varepsilon/2$ and $n \geq N_2$ implies that $|t_n - t| < \varepsilon/2$. Let $N = \max\{N_1, N_2\}$. Then $n \geq N$ implies

$$|(s_n - s) + (t_n - t)| \leq |s_n - s| + |t_n - t| < \varepsilon .$$

(b) This is Exercise B.1.

(c) Note the algebraic identity

$$s_n t_n - st = (s_n - s)(t_n - t) + s(t_n - t) + t(s_n - s) \ .$$

By parts (a) and (b) we see that $\lim s(t_n - t) = 0 = \lim t(s_n - s)$. Now, given $\varepsilon > 0$, there exists N_1 and N_2 such that $n \geq N_1$ implies

$$|s_n - s| < \sqrt{\varepsilon}$$

and $n \geq N_2$ implies

$$|t_n - t| < \sqrt{\varepsilon} \ .$$

If $N = \max\{N_1, N_2\}$, then $n \geq N$ implies that

$$|(s_n - s)(t_n - t)| < \varepsilon$$

and hence $\lim (s_n t_n - st) = \lim (s_n - s)(t_n - t) = 0$.

(d) At the end of Section 0.4 we showed that $|a| - |b| \leq |a + b|$ for any $a, b \in \mathbb{R}$. Reviewing the proof shows that this applies to complex numbers as well. Letting $b \rightarrow -b$ then shows that $|a| - |b| \leq |a - b|$.

Given $|s|/2$, there exists m such that $n \geq m$ implies $|s_n - s| < |s|/2$ (this follows from the fact that $s_n \rightarrow s$). But $|s| - |s_n| \leq |s_n - s| < |s|/2$ implies $|s_n| > |s|/2$.

Alternatively, given $\varepsilon > 0$, there exists N (which we can always choose greater than m) such that $n \geq N$ implies $|s_n - s| < |s|^2\varepsilon/2$. Combining these results, we see that for all $n \geq N$ we have

$$\left|\frac{1}{s_n} - \frac{1}{s}\right| = \left|\frac{s_n - s}{ss_n}\right| = \frac{|s_n - s|}{|s||s_n|} < \frac{2|s_n - s|}{|s|^2} < \varepsilon \ . \ \blacksquare$$

Intuitively, we expect that a sequence $\{x_k\}$ of points in $\mathbb{R}^n$ converges to a point in $\mathbb{R}^n$ if and only if each of the n coordinates converges on its own. That this is indeed true is shown in our next theorem.

Theorem B3 Suppose $x_k = (x_k{}^1, \ldots, x_k{}^n) \in \mathbb{R}^n$. Then $\{x_k\}$ converges to $x = (x^1, \ldots, x^n) \in \mathbb{R}^n$ if and only if $x_k{}^i \rightarrow x^i$ for each $i = 1, \ldots, n$.

Proof First note that for any $j = 1, \ldots, n$ we have

$$|x - y|^2 = \sum_{i=1}^{n} (x^i - y^i)^2 \geq (x^j - y^j)^2$$

which implies $|x^j - y^j| \le |x - y|$. Now assume that $x_k \to x$. Then given $\varepsilon > 0$, there exists N such that $k \ge N$ implies $|x_k{}^j - y^j| \le |x_k - x| < \varepsilon$. This shows that $x_k \to x$ implies $x_k{}^j \to x^j$.

Conversely, assume $x_k{}^j \to x^j$ for every $j = 1, \dots, n$. Then given $\varepsilon > 0$, there exists N such that $k \ge N$ implies $|x_k{}^j - x^j| < \varepsilon/\sqrt{n}$. Hence $k \ge N$ implies

$$|x_k - x| = \left(\sum_{j=1}^{n} (x_k{}^j - x^j)^2 \right)^{1/2} < (n\varepsilon^2/n)^{1/2} = \varepsilon$$

so that $x_k \to x$. ∎

A sequence $\{x_k\}$ in a metric space (X, d) is said to be a **Cauchy sequence** if given $\varepsilon > 0$, there exists N such that $n, m \ge N$ implies $d(x_n, x_m) < \varepsilon$. It is easy to see that every convergent sequence is in fact a Cauchy sequence. Indeed, simply note that if $x_k \to x$, then given $\varepsilon > 0$, there exists N such that $n \ge N$ implies $d(x_n, x) < \varepsilon/2$. Hence if $n, m \ge N$ we have

$$d(x_n, x_m) \le d(x_n, x) + d(x_m, x) < \varepsilon/2 + \varepsilon/2 = \varepsilon .$$

However, it is not true that a Cauchy sequence need converge. For example, suppose $X = (0, 1] \subset \mathbb{R}$ and let $\{x_k\} = \{1/k\}$ for $k = 1, 2, \dots$. This is a Cauchy sequence that wants to converge to the point 0 (choose $N = 1/\varepsilon$ so that $|1/n - 1/m| \le |1/n| + |1/m| < 2\varepsilon$ for all $m, n \ge N$). But $0 \notin (0, 1]$ so that the limit of the sequence is not in the space. This example shows that convergence is not an intrinsic property of sequences, but rather depends on the space in which the sequence lies. A metric space in which every Cauchy sequence converges is said to be **complete** (see Appendix A).

We have shown that any convergent sequence is necessarily a Cauchy sequence, but that the converse is not true in general. However, in the case of $\mathbb{R}^n$, it is in fact true that every Cauchy sequence does indeed converge, i.e., $\mathbb{R}^n$ is a complete metric space. This is easy to prove using the fact that any n-cell in $\mathbb{R}^n$ is compact (Theorem A11), and we outline the proof in Exercise B.10. However, it is worth proving that $\mathbb{R}^n$ is complete without using this result. We begin by proving several other facts dealing with the real number line $\mathbb{R}$. By way of terminology, a sequence $\{x_n\}$ of real numbers with the property that $x_n \le x_{n+1}$ is said to be **increasing**. Similarly, if $x_n \ge x_{n+1}$ then the sequence is said to be **decreasing**. We will sometimes use the term **monotonic** to refer to a sequence that is either increasing or decreasing.

Theorem B4 Let $\{x_k\}$ be an increasing sequence (i.e., $x_k \le x_{k+1}$) of real numbers that is bounded above. Then the least upper bound b of the set $\{x_k\}$ is the limit of the sequence.

Proof It should be remarked that the existence of the least upper bound is guaranteed by the Archimedean axiom. Given $\varepsilon > 0$, the number $b - \varepsilon/2$ is not an upper bound for $\{x_k\}$ since b is by definition the least upper bound. Therefore there exists N such that $b - \varepsilon/2 \le x_N \le b$ (for otherwise $b - \varepsilon/2$ would be the least upper bound). Since $\{x_k\}$ is increasing, we have

$$b - \varepsilon/2 \le x_N \le x_n \le b$$

for every $n \ge N$. Rearranging, this is just $b - x_n \le \varepsilon/2 < \varepsilon$ which is the same as $|x_n - b| < \varepsilon$. ∎

Since the Archimedean axiom also refers to the greatest lower bound of a set of real numbers, it is clear that Theorem B4 may be applied equally well to the greatest lower bound of a decreasing sequence.

Let (X, d) be a metric space, and suppose A is a subset of X. Recall from Appendix A that a point $x \in X$ is said to be an **accumulation point** of A if every deleted neighborhood of x contains a point of A. The analogous term for sequences is the following. A number x is said to be a **cluster point** (or **limit point**) of a sequence $\{x_n\}$ if given $\varepsilon > 0$, there exist infinitely many integers n such that $|x_n - x| < \varepsilon$. Equivalently, x is a cluster point if given $\varepsilon > 0$ and given N, there exists *some* $n \ge N$ such that $|x_n - x| < \varepsilon$. Note that this does not say that there are infinitely many *distinct* x_n such that $|x_n - x| < \varepsilon$. In fact, all the x_n could be identical. It is important to distinguish between the indices n and the actual elements x_n of the sequence. It is also important to realize that a limit point of a sequence is not the same as the limit of a sequence (why?). Note also that a sequence in X may be considered to be a subset of X, and in this context we may also refer to the accumulation points of a sequence.

Our next result is known as the Bolzano-Weierstrass Theorem.

Theorem B5 (**Bolzano-Weierstrass**) Let $\{x_k\}$ be a sequence of real numbers, and let $a, b \in \mathbb{R}$ be such that $a \le x_k \le b$ for all positive integers k. Then there exists a cluster point c of the sequence with $a \le c \le b$.

Proof For each n, the sequence $\{x_n, x_{n+1}, \dots\}$ is bounded below (by a), and hence has a greatest lower bound (whose existence is guaranteed by the Archimedean axiom) which we denote by c_n. Then $\{c_n\}$ forms an increasing sequence $c_n \le c_{n+1} \le \cdots$ which is bounded above by b. Theorem B4 now shows that the sequence $\{c_n\}$ has a least upper bound c (with $a \le c \le b$) which is in fact the limit of the sequence $\{c_n\}$. We must show that c is a cluster point of the sequence $\{x_k\}$.

To say that c is the limit of the sequence $\{c_n\}$ means that given $\varepsilon > 0$ and given any N, there exists some $m \ge N$ such that

$$|c_m - c| < \varepsilon/2 \ .$$

By definition, c_m is the greatest lower bound of the set $\{x_m, x_{m+1}, \ldots\}$ which means there exists $k \ge m$ such that $c_m \le x_k < c_m + \varepsilon/2$ or

$$|x_k - c_m| < \varepsilon/2 \ .$$

Therefore $k \ge m \ge N$ and

$$|x_k - c| = |x_k - c_m + c_m - c| \le |x_k - c_m| + |c_m - c| < \varepsilon$$

which shows that c is a cluster point of the sequence $\{x_k\}$. ∎

Note that Theorem B5 also follows from Theorem A12.

Theorem B6 If $\{x_n\}$ is a Cauchy sequence of numbers, then it is bounded.

Proof By definition of a Cauchy sequence, given $\varepsilon = 1$ there exists N such that $n \ge N$ implies $|x_n - x_N| < 1$. Hence $|x_n| - |x_N| \le |x_n - x_N| < 1$ implies $|x_n| \le |x_N| + 1$ for every $n \ge N$. Define $B = \max\{|x_1|, \ldots, |x_N|, |x_N| + 1\}$. Then B is clearly a bound for $\{x_n\}$. ∎

Theorem B7 Any Cauchy sequence $\{x_n\}$ of numbers converges.

Proof From Theorem B6 the sequence $\{x_n\}$ has a bound B, and hence we have $-B \le x_n \le B$ for all n. Hence by Theorem B5, the sequence $\{x_n\}$ has a cluster point c. We claim that c is the limit of the sequence. Since the sequence is Cauchy, given $\varepsilon > 0$ there exists N such that $m, n \ge N$ implies $|x_m - x_n| < \varepsilon/2$. Using this ε, we see that because c is a cluster point, there exists $m \ge N$ such that $|x_m - c| < \varepsilon/2$. Combining these last two results shows that for all $n \ge N$

$$|x_n - c| \le |x_n - x_m| + |x_m - c| < \varepsilon \ . \ ∎$$

We are now in a position to prove our principal assertion.

Theorem B8 $\mathbb{R}^n$ is a complete metric space. In other words, every Cauchy sequence in $\mathbb{R}^n$ converges to a point in $\mathbb{R}^n$.

Proof Let $\{x_k\}$ be a Cauchy sequence in $\mathbb{R}^n$. Then $|x_m{}^j - x_n{}^j| \le |x_m - x_n|$ (see the proof of Theorem B3) so that $\{x_k{}^j\}$ is also a Cauchy sequence in $\mathbb{R}$ for each $j = 1, \dots, n$. Hence by Theorem B7 each of the sequences $\{x_k{}^j\}$ also converges in $\mathbb{R}$. Therefore (by Theorem B3) the sequence $\{x_k\}$ must converge in $\mathbb{R}^n$. ∎

We have seen that any convergent sequence is a Cauchy sequence and hence bounded. However, the converse is not generally true. (For example, the sequence $\{1, 2, 1, 2, \dots\}$ is clearly bounded but does not converge to either 1 or 2.) There is, however, a special case in which the converse is true that will be of use to us.

Theorem B9 A monotonic sequence $\{x_n\}$ converges if and only if it is bounded.

Proof We consider increasing sequences. The proof for decreasing sequences is similar. It was shown above that any convergent sequence is bounded, so we need only consider the converse. But this was proved in Theorem B4. ∎

Finally, accumulation points are useful in determining whether or not a set is closed. The principal result relating these concepts is the following.

Theorem B10 A subset A of a metric space (X, d) is closed if and only if A contains all of its accumulation points.

Proof This is also Theorem A13. ∎

Before continuing, we must make a digression to discuss some more basic properties of metric spaces. If the reader already knows that a point x in a subset A of a metric space X is in the closure of A if and only if every neighborhood of x intersects A, then he/she may skip to Theorem B16 below.

Let (X, d) be a metric space, and suppose $A \subset X$. We define

(a) The **closure** of A, denoted by Cl A, to be the intersection of all closed supersets of A;

(b) The **interior** of A, denoted by Int A (or A^o), to be the union of all open subsets of A;

(c) The **boundary** of A, denoted by Bd A, to be the set of all $x \in X$ such that every open set containing x contains both points of A and points of $A^c = X - A$;

(d) The **exterior** of A, denoted by Ext A, to be $(\text{Cl } A)^c = X - \text{Cl } A$;

(e) The **derived set** of A, denoted by A′, to be the set of all accumulation points of A.

Example B1 Let $X = \mathbb{R}^2$ with the Pythagorean metric. Let A be the open unit ball defined by $A = \{(x, y): x^2 + y^2 < 1\}$. Then the following sets should be intuitively clear to the reader:

$$\begin{aligned}
&\text{Cl } A = \{(x, y): x^2 + y^2 \le 1\};\\
&\text{Int } A = A;\\
&\text{Bd } A = \{(x, y): x^2 + y^2 = 1\};\\
&\text{Ext } A = \{(x, y): x^2 + y^2 > 1\};\\
&A' = \text{Cl } A. \;/\!\!/
\end{aligned}$$

Theorem B11 Let (X, d) be a metric space and suppose $A, B \subset X$. Then

(a) $A \subset \text{Cl } A$.

(b) $\text{Cl}(\text{Cl } A) = \text{Cl } A$.

(c) $\text{Cl}(A \cup B) = (\text{Cl } A) \cup (\text{Cl } B)$.

(d) $\text{Cl } \varnothing = \varnothing$.

(e) A is closed if and only if $A = \text{Cl } A$.

(f) $\text{Cl}(A \cap B) \subset (\text{Cl } A) \cap (\text{Cl } B)$.

Proof Parts (a), (b), (d) and (e) are essentially obvious from the definition of Cl A, the fact that the intersection of an arbitrary number of closed sets is closed (see Exercise A.3), the fact that the empty set is a subset of every set, and the fact that if A is closed, then A is one of the closed sets which contains A.

(c) First note that if $S \subset T$, then any closed superset of T is also a closed superset of S, and therefore $\text{Cl } S \subset \text{Cl } T$. Next, observe that $A \subset A \cup B$ and $B \subset A \cup B$, so that taking the closure of both sides of each of these relations yields

$$\text{Cl } A \subset \text{Cl}(A \cup B)$$

and

$$\text{Cl } B \subset \text{Cl}(A \cup B)\ .$$

Together these show that $(\text{Cl } A) \cup (\text{Cl } B) \subset \text{Cl}(A \cup B)$. Since Cl A and Cl B are both closed, $(\text{Cl } A) \cup (\text{Cl } B)$ must also be closed and contain $A \cup B$. Hence we also have

$$\mathrm{Cl}(A \cup B) \subset (\mathrm{Cl}\,A) \cup (\mathrm{Cl}\,B)\ .$$

This shows that $\mathrm{Cl}(A \cup B) = (\mathrm{Cl}\,A) \cup (\mathrm{Cl}\,B)$.

(f) By (a) we have $A \cap B \subset A \subset \mathrm{Cl}\,A$ and $A \cap B \subset B \subset \mathrm{Cl}\,B$, and hence $A \cap B$ is a subset of the *closed* set $(\mathrm{Cl}\,A) \cap (\mathrm{Cl}\,B)$. But by definition of closure this means that

$$\mathrm{Cl}(A \cap B) \subset (\mathrm{Cl}\,A) \cap (\mathrm{Cl}\,B)\ . \ ∎$$

Theorem B12 Let (X, d) be a metric space and suppose $A \subset X$. Then

(a) $\mathrm{Cl}\,A = A \cup A'$.

(b) $\mathrm{Cl}\,A = \mathrm{Int}\,A \cup \mathrm{Bd}\,A$.

(c) $\mathrm{Bd}\,A = \mathrm{Bd}\,A^c$.

(d) $\mathrm{Int}\,A = \mathrm{Cl}\,A - \mathrm{Bd}\,A$.

Proof (a) Assume $x \in \mathrm{Cl}\,A$ but that $x \notin A$. We first show that $x \in A'$. Let U be any open set containing x. If $U \cap A = \varnothing$, then U^c is a closed superset of A which implies $\mathrm{Cl}\,A \subset U^c$. But this contradicts the assumption that $x \in U$ since it was assumed that $x \in \mathrm{Cl}\,A$. Therefore, since $x \notin A$, we must have

$$(U - \{x\}) \cap A \neq \varnothing$$

so that $x \in A'$. This shows that $\mathrm{Cl}\,A \subset A \cup A'$.

Now assume that $x \in A \cup A'$. If $x \in A$, then obviously $x \in \mathrm{Cl}\,A$ (since $A \subset \mathrm{Cl}\,A$), so suppose $x \in A'$. We will show that x is contained in any closed superset of A. Let F be any closed superset of A not containing x. Then F^c is an open set containing x and such that $(F^c - \{x\}) \cap A = \varnothing$ which says that $x \notin A'$, a contradiction. Thus x is contained in any closed superset of A so that $x \in \mathrm{Cl}\,A$. Since this shows that $A \cup A' \subset \mathrm{Cl}\,A$, it follows that $A \cup A' = \mathrm{Cl}\,A$.

(b) We first suppose that $x \in \mathrm{Cl}\,A$ but $x \notin \mathrm{Bd}\,A$. Since $x \notin \mathrm{Bd}\,A$, there exists an open set U containing x such that either $U \subset A$ or $U \subset A^c$. If it were true that $U \subset A^c$, then U^c would be a closed superset of A (see Example 0.1) so that $\mathrm{Cl}\,A \subset U^c$ which contradicts the assumption that $x \in \mathrm{Cl}\,A$. We must therefore have $U \subset A$, and hence $x \in \mathrm{Int}\,A$. Since the assumption that $x \in \mathrm{Cl}\,A$ but $x \notin \mathrm{Bd}\,A$ led to the requirement that $x \in \mathrm{Int}\,A$, we must have

$$\mathrm{Cl}\,A \subset \mathrm{Int}\,A \cup \mathrm{Bd}\,A\ .$$

Now assume $x \in \text{Int } A \cup \text{Bd } A$, but that $x \notin \text{Cl } A$. Note $\text{Int } A \subset A \subset \text{Cl } A$ so that $x \notin \text{Int } A$, and hence it must be true that $x \in \text{Bd } A$. However, since $x \notin \text{Cl } A$, $(\text{Cl } A)^c$ is an open set containing x with the property that $(\text{Cl } A)^c \cap A = \varnothing$. But this says that $x \notin \text{Bd } A$ which contradicts our original assumption. In other words, we must have $\text{Int } A \cup \text{Bd } A \subset \text{Cl } A$, so that

$$\text{Cl } A = \text{Int } A \cup \text{Bd } A \ .$$

(c) If $x \in \text{Bd } A$ and U is any open set containing x, then $U \cap A \neq \varnothing$ and $U \cap A^c \neq \varnothing$. But $A = (A^c)^c$ so that we also have $U \cap (A^c)^c \neq \varnothing$. Together with $U \cap A^c \neq \varnothing$, this shows that $x \in \text{Bd } A^c$. Reversing the argument shows that if $x \in \text{Bd } A^c$, then $x \in \text{Bd } A$. Hence $\text{Bd } A = \text{Bd } A^c$.

(d) This will follow from part (b) if we can show that $\text{Int } A \cap \text{Bd } A = \varnothing$. Now suppose that $x \in \text{Int } A \cap \text{Bd } A$. Then since $x \in \text{Bd } A$, it must be true that every open set containing x intersects A^c. But this contradicts the assumption that $x \in \text{Int } A$ (since by definition there must exist an open set U containing x such that $U \subset A$), and hence we must have $\text{Int } A \cap \text{Bd } A = \varnothing$. ∎

It should be remarked that some authors *define* Bd A as Cl A − Int A so that our definition of Bd A follows as a theorem. This fact, along with some additional insight, is contained in the following theorem.

Theorem B13 Let A be a subset of a metric space (X, d), and suppose $x \in X$. Then

(a) $x \in \text{Int } A$ if and only if *some* neighborhood of x is a subset of A.
(b) $x \in \text{Cl } A$ if and only if *every* neighborhood of x intersects A.
(c) $\text{Bd } A = \text{Cl } A - \text{Int } A$.

Proof (a) If $x \in \text{Int } A$, then by definition of Int A there exists a neighborhood U of x such that $U \subset A$. On the other hand, if there exists an open set U such that $x \in U \subset A$, then $x \in \text{Int } A$.

(b) By Theorem B12(a), $\text{Cl } A = A \cup A'$. If $x \in \text{Cl } A$ and $x \in A$, then clearly every neighborhood of x intersects A. If $x \in \text{Cl } A$ and $x \in A'$, then every neighborhood of x also intersects A. Conversely, suppose that every neighborhood of x intersects A. Then either $x \in A$ or $x \in A'$, and hence $x \in A \cup A' = \text{Cl } A$.

(c) By definition, $x \in \text{Bd } A$ if and only if every open set containing x contains both points of A and points of A^c. In other words, $x \in \text{Bd } A$ if and only if every open set containing x contains points of A but is not a subset of A. By parts (a) and (b), this is just $\text{Bd } A = \text{Cl } A - \text{Int } A$. ∎

Example B2 An elementary fact that will be referred to again is the following. Let A be a nonempty set of real numbers that is bounded above. Then by the Archimedean axiom, A has a least upper bound b = sup A. Given $\varepsilon > 0$, there must exist $x \in A$ such that $b - \varepsilon < x < b$, for otherwise $b - \varepsilon$ would be an upper bound of A. But this means that any neighborhood of b intersects A, and hence $b \in \mathrm{Cl}\, A$. Thus $b \in A$ if A is closed. ∕∕

Our next example yields an important basic result.

Example B3 Let $X = \mathbb{R}$ with the standard (absolute value) metric, and let $\mathbb{Q} \subset \mathbb{R}$ be the subset of all rational numbers. In Theorem 0.4 it was shown that given any two distinct real numbers there always exists a rational number between them. This may also be expressed by stating that any neighborhood of any real number always contains a rational number. In other words, we have $\mathrm{Cl}\, \mathbb{Q} = \mathbb{R}$. ∕∕

From Theorems B10 and B12(a), we might guess that there is a relationship between sequences and closed sets. This is indeed the case, and our next theorem provides a very useful description of the closure of a set.

Theorem B14 (a) A set $A \subset (X, d)$ is closed if and only if for every sequence $\{x_n\}$ in A that converges, the limit is an element of A.

(b) If $A \subset (X, d)$, then $x \in \mathrm{Cl}\, A$ if and only if there is a sequence $\{x_n\}$ in A such that $x_n \to x$.

Proof (a) Suppose that A is closed, and let $x_n \to x$. Since any neighborhood of x must contain all x_n for n sufficiently large, it follows from Theorem B10 that $x \in A$.

Conversely, assume that any sequence in A converges to an element of A, and let x be any accumulation point of A. We will construct a sequence in A that converges to x. To construct such a sequence, choose $x_n \in B(x, 1/n) \cap A$. This is possible since x is an accumulation point of A. Then given $\varepsilon > 0$, choose $N \geq 1/\varepsilon$ so that $x_n \in B(x, \varepsilon)$ for every $n \geq N$. Hence $x_n \to x$ so that $x \in A$. Theorem B10 then shows that A is closed.

(b) This is Exercise B.4. ∎

If (X, d) is a metric space, then $A \subset X$ is said to be **somewhere dense** if $\mathrm{Int}(\mathrm{Cl}\, A) \neq \varnothing$. The set A is said to be **nowhere dense** if it is not somewhere dense. If $\mathrm{Cl}\, A = X$, then A is said to be **dense** in X.

Example B4 Let X = $\mathbb{R}$ with the usual metric. The set A = [a, b) has closure Cl A = [a, b], and therefore Int(Cl A) = (a, b) ≠ ∅. Hence A is somewhere dense. Example B3 showed that the set $\mathbb{Q}$ is dense in $\mathbb{R}$. Now let A = $\mathbb{Z}$, the set of all integers. $\mathbb{Z}^c = \mathbb{R} - \mathbb{Z}$ is the union of open sets of the form (n, n + 1) where n is an integer, and hence $\mathbb{Z}$ is closed. It should also be clear that $\mathbb{Z}' = \varnothing$ since there clearly exist deleted neighborhoods of any integer that do not contain any other integers. By Theorem B13(a), we also see that Int(Cl $\mathbb{Z}$) = Int $\mathbb{Z}$ = ∅ so that $\mathbb{Z}$ is nowhere dense. ∕∕

Theorem B15 A subset A of a metric space (X, d) is dense if and only if every open subset U of X contains some point of A.

Proof Suppose A is dense so that Cl A = X. If U ⊂ X is open, then the fact that any x ∈ U ⊂ Cl A implies that every neighborhood of x intersects A (Theorem B13). In particular, U is a neighborhood of x so that U ⋂ A ≠ ∅. On the other hand, suppose that every open set U intersects A. If x ∈ X, then every neighborhood U of x must intersect A so that x ∈ Cl A. Since x was arbitrary, it must be true that Cl A = X. ∎

After this topological digression, let us return to sequences of numbers. Given a sequence $\{x_n\}$, we may consider a sequence $\{n_k\}$ of positive integers that forms a subset of $\mathbb{Z}^+$ such that $n_1 < n_2 < \cdots$. The corresponding subset $\{x_{n_k}\}$ of $\{x_n\}$ is called a **subsequence** of $\{x_n\}$. If $\{x_{n_k}\}$ converges, then its limit is called a **subsequential limit** of $\{x_n\}$. From the definitions, it should be clear that any cluster point of $\{x_n\}$ is the limit of a convergent subsequence. It should also be clear that a sequence $\{x_n\}$ converges to x if and only if every subsequence also converges to x (see Exercise B.5).

Theorem B16 The set of all subsequential limits of a sequence $\{x_n\}$ in a metric space (X, d) forms a closed subset of X.

Proof Let S be the set of all subsequential limits of $\{x_n\}$. If y is an accumulation point of S, we must show that y ∈ S (Theorem B10), and hence that some subsequence of $\{x_n\}$ converges to y. Choose n_1 such that $x_{n_1} \neq y$ (why can this be done?), and let $\delta = d(x_{n_1}, y)$. Now suppose that $n_1, n_2, \dots, n_{k-1}$ have been chosen. Since y is an accumulation point of S, there exists z ∈ S, z ≠ y, such that $d(z, y) < 2^{-k}\delta$. But z ∈ S implies that z is a subsequential limit, and hence there exists $n_k > n_{k-1}$ such that $d(z, x_{n_k}) < 2^{-k}\delta$. Therefore, for each k = 1, 2, . . . we have

$$d(x_{n_k}, y) \le d(x_{n_k}, z) + d(z, y) < 2^{1-k}\delta$$

so that the subsequence $\{x_{n_k}\}$ converges to y. ∎

Given a sequence $\{a_n\}$ of complex numbers (which can be regarded as points in $\mathbb{R}^2$), we may define the **infinite series** (generally called simply a **series**)

$$\sum_{i=1}^{\infty} a_n = a_1 + a_2 + \cdots$$

as the sequence $\{s_n\}$ where

$$s_n = \sum_{k=1}^{n} a_k \ .$$

If the *sequence* $\{s_n\}$ converges to a number s, we say that the series **converges** and we write this as $\Sigma_{n=1}^{\infty} a_n = s$. Note that the number s is the limit of a sequence of partial sums. For notational convenience, the lower limit in the sum may be taken to be 0 instead of 1, and we will frequently write just Σa_n when there is no danger of ambiguity and the proper limits are understood.

The reader should note that $\Sigma_{n=1}^{\infty} a_n$ stands for two very different things. On the one hand it is used to stand for the sequence $\{s_n\}$ of partial sums, and on the other hand it stands for $\lim_{n\to\infty} s_n$. This is a common abuse of notation, and the context usually makes it clear which meaning is being used.

We have seen that any convergent sequence is Cauchy, and in Theorem B8 we showed that any Cauchy sequence in $\mathbb{R}^n$ converges. Thus, a sequence in $\mathbb{R}^n$ converges if and only if it is Cauchy. This is called the **Cauchy criterion**. Since the convergence of a series is defined in terms of the sequence of partial sums, we see that the Cauchy criterion may be restated as follows.

Theorem B17 A series of numbers Σa_n converges if and only if given $\varepsilon > 0$, there exists an integer N such that $m \geq n \geq N$ implies

$$\left| \sum_{k=n}^{m} a_k \right| < \varepsilon \ .$$

Proof If the series Σa_n converges, then the sequence $\{s_k\}$ of partial sums $s_k = \Sigma_{n=1}^{k} a_n$ converges, and hence $\{s_k\}$ is Cauchy. Conversely, if the sequence of partial sums s_k is Cauchy, then $\{s_k\}$ converges (Theorem B8). In either case, this means that given $\varepsilon > 0$, there exists N such that $p \geq q \geq N$ implies

$$\left|s_p - s_q\right| = \left|\sum_{k=1}^{p} a_k - \sum_{k=1}^{q} a_k\right| = \left|\sum_{k=q+1}^{p} a_k\right| < \varepsilon \ .$$

The result now follows by choosing m = p and n = q + 1. ∎

Another useful way of stating Theorem B17 is to say that a series Σa_n converges if and only if given ε > 0, there exists N such that k ≥ N implies that $|a_k + \cdots + a_{k+p}| < \varepsilon$ for all positive integers p = 0, 1, 2, ...

Corollary If Σa_n converges, then given ε > 0 there exists N such that $|a_n| < \varepsilon$ for all n ≥ N. In other words, if Σa_n converges, then $\lim_{n\to\infty} a_n = 0$.

Proof This follows from Theorem B17 by letting m = n. ∎

While this corollary says that a *necessary* condition for Σa_n to converge is that $a_n \to 0$, this is not a *sufficient* condition (see Example B5 below).

If we have a series of nonnegative real numbers, then each partial sum is clearly nonnegative, and the sequence of partial sums forms a non-decreasing sequence. Thus, directly from Theorem B9 we have the following.

Theorem B18 A series of nonnegative real numbers converges if and only if its partial sums form a bounded sequence.

One consequence of this theorem is the next result.

Theorem B19 Suppose Σa_n is a series such that $a_1 \geq a_2 \geq \cdots \geq 0$. Then Σa_n converges if and only if

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + \cdots$$

converges.

Proof Let $s_n = a_1 + a_2 + \cdots + a_n$ and let $t_k = a_1 + 2a_2 + \cdots + 2^k a_{2^k}$. Since all terms in the series Σa_n are nonnegative we may write for $n < 2^k$

$$s_n \leq a_1 + (a_2 + a_3) + (a_4 + a_5 + a_6 + a_7) + \cdots + (a_{2^k} + \cdots + a_{2^{k+1}-1})$$

since this is just adding the nonnegative term $a_{2^k+1} + \cdots + a_{2^{k+1}-1}$ to $s_{2^k} \geq s_n$. But $\{a_n\}$ is a decreasing sequence, so noting that the last term in parentheses consists of 2^k terms, we have

$$s_n \le a_1 + 2a_2 + 4a_4 + \cdots + 2^k a_{2^k} = t_k \ .$$

Similarly, if $n > 2^k$ we have

$$\begin{aligned} s_n &\ge a_1 + a_2 + (a_3 + a_4) + \cdots + (a_{2^{k-1}+1} + \cdots + a_{2^k}) \\ &\ge (1/2)a_1 + a_2 + 2a_4 + \cdots + 2^{k-1} a_{2^k} \\ &= (1/2)t_k \ . \end{aligned}$$

We have now shown that $n < 2^k$ implies $s_n \le t_k$, and $n > 2^k$ implies $2s_n \ge t_k$. Thus the sequences $\{s_n\}$ and $\{t_k\}$ are either both bounded or both unbounded. Together with Theorem B18, this completes the proof. ∎

The interesting aspect of this theorem is that the convergence of $\sum a_n$ is determined by the convergence of a rather "sparse" subsequence.

Example B5 Let us show that the series $\sum n^{-p}$ converges if $p > 1$ and diverges if $p \le 1$. Indeed, suppose $p > 0$. Then by Theorem B19 we consider the series

$$\sum_{k=0}^{\infty} 2^k \cdot 2^{-kp} = \sum_{k=0}^{\infty} 2^{k(1-p)} \ .$$

By the corollary to Theorem B17, we must have $1 - p < 0$ so that $p > 1$. In this case, $\sum 2^{k(1-p)} = \sum (2^{1-p})^k$ is a geometric series which converges as in Example B1, and hence Theorem B19 shows that $\sum n^{-p}$ converges for $p > 1$. If $p \le 0$, then $\sum n^{-p}$ diverges by the corollary to Theorem B17. ∕∕

If we are given a series $\sum a_n$, we could rearrange the terms in this series to obtain a new series $\sum a'_n$. Formally, we define this rearrangement by letting $\{k_n\}$ be a sequence in which every positive integer appears exactly once. In other words, $\{k_n\}$ is a one-to-one mapping from $\mathbb{Z}^+$ onto $\mathbb{Z}^+$. If we now define $a'_n = a_{k_n}$ for $n = 1, 2, \ldots$, then the corresponding series $\sum a'_n$ is called a **rearrangement** of the series $\sum a_n$.

For each of the series $\sum a_n$ and $\sum a'_n$, we form the respective sequences of partial sums $\{s_k\}$ and $\{s'_k\}$. Since these sequences are clearly different in general, it is not generally the case that they both converge to the same limit. While we will not treat this problem in any detail, there is one special case that we will need. This will be given as a corollary to the following theorem.

A series $\sum a_n$ is said to **converge absolutely** if the series $\sum |a_n|$ converges.

Theorem B20 If $\sum a_n$ converges absolutely, then $\sum a_n$ converges.

Proof Note $|\sum_{k=n}^{m} a_k| \le \sum_{k=n}^{m} |a_k|$ and apply Theorem B17. ∎

Corollary If $\sum a_n$ converges absolutely, then every rearrangement of $\sum a_n$ converges to the same sum.

Proof Let $\sum a'_n$ be a rearrangement with partial sums s'_k. Since $\sum a_n$ converges absolutely, we may apply Theorem B17 to conclude that for every $\varepsilon > 0$ there exists N such that $m \ge n \ge N$ implies

$$\sum_{i=n}^{m} |a_i| < \varepsilon \ . \qquad (*)$$

Using the notation of the discussion preceding the theorem, we let $p \in \mathbb{Z}^+$ be such that the integers $1, \ldots, N$ are contained in the collection $k_1, \ldots, k_p$ (note that we must have $p \ge N$). If for any $n > p$ we now form the difference $s_n - s'_n$, then the numbers $a_1, \ldots, a_N$ will cancel out (as may some other numbers if $p > N$) and hence, since (*) applies to all $m \ge n \ge N$, we are left with $|s_n - s'_n| < \varepsilon$. This shows that $\{s_n\}$ and $\{s'_n\}$ both converge to the same sum (since if $s_n \to s$, then $|s'_n - s| \le |s'_n - s_n| + |s_n - s| < 2\varepsilon$ which implies that $s'_n \to s$ also). ∎

We remark that Theorems B17 and B20 apply equally well to any complete normed space if we replace the absolute value by the appropriate norm.

Before presenting any examples of series, we first compute the limits of some commonly occurring sequences of real numbers.

Theorem B21 (a) If $p > 0$, then $\lim_{n\to\infty} 1/n^p = 0$.

(b) If $p > 0$, then $\lim_{n\to\infty} p^{1/n} = 1$.

(c) $\lim_{n\to\infty} n^{1/n} = 1$.

(d) If $p > 0$ and r is real, then $\lim_{n\to\infty} n^r/(1 + p)^n = 0$.

(e) If $|x| < 1$, then $\lim_{n\to\infty} x^n = 0$.

Proof (a) Given $\varepsilon > 0$, we seek an N such that $n \ge N$ implies $1/n^p < \varepsilon$. Then choose $N \ge (1/\varepsilon)^{1/p}$.

(b) If $p = 1$, there is nothing to prove. For $p > 1$, define $x_n = p^{1/n} - 1 > 0$ so that by the binomial theorem (Example 0.7) we have

$$p = (1 + x_n)^n \ge 1 + nx_n \ .$$

Thus $0 < x_n \le (p - 1)/n$ so that $\lim x_n = 0$, and hence $\lim p^{1/n} = 1$. If $0 < p < 1$, we define $y_n = (1/p)^{1/n} - 1 > 0$. Then

$$p = (1 + y_n)^n \geq 1 + ny_n$$

so that $y_n \to 0$, and hence we again have $\lim p^{1/n} = 1$.

(c) Let $x_n = n^{1/n} - 1 \geq 0$, so that using only the quadratic term in the binomial expansion yields

$$n = (1 + x_n)^n \geq \binom{n}{2} x_n{}^2 = \frac{n(n-1)}{2} x_n{}^2 \quad .$$

Thus (for $n \geq 2$) we have $0 \leq x_n \leq [2/(n - 1)]^{1/2}$ so that $x_n \to 0$. Therefore $\lim n^{1/n} = 1$.

(d) Let k be any integer > 0 such that k > r. Choosing the k*th* term in the binomial expansion we have (since p > 0)

$$(1+p)^n \geq \binom{n}{p} p^k = \frac{n(n-1)\cdots(n-(k-1))}{k!} p^k \quad .$$

If we let $n > 2k$, then $k < n/2$ so that $n > n/2$, $n - 1 > n/2, \dots, n - (k - 1) > n/2$ and hence $(1 + p)^n > (n/2)^k p^k / k!$. Thus (for $n > 2k$)

$$0 < \frac{n^r}{(1+p)^n} < \frac{2^k k!}{p^k} n^{r-k} \quad .$$

Since $r - k < 0$, it follows that $n^{r-k} \to 0$ by (a).

(e) Choose $r = 0$ in (d). ∎

Corollary If $N > 0$ is any finite integer, then $\lim_{n\to\infty} (n^N)^{1/n} = 1$.

Proof $(n^N)^{1/n} = n^{N/n} = (n^{1/n})^N$ so that (by Theorem B2(c))

$$\lim(n^{1/n})^N = (\lim n^{1/n})^N = 1^N = 1 \ . \ ∎$$

Example B6 The geometric series $\sum_{n=0}^{\infty} x^n$ converges for $|x| < 1$ and diverges for $|x| \geq 1$. Indeed, from elementary algebra we see that (for $x \neq 1$)

$$1 + x + x^2 + \cdots + x^n = \frac{1 - x^{n+1}}{1 - x} \quad .$$

If $|x| < 1$, we clearly have $\lim x^{n+1} = 0$, and hence

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \ .$$

If $|x| > 1$, then $|x|^{n+1} \to \infty$ and the series diverges. In the case that $|x| = 1$, we see that $x^n \nrightarrow 0$ so the series diverges. //

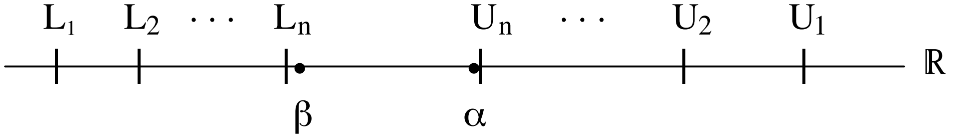
Let $\{x_n\}$ be a sequence of real numbers. In general this sequence may have many cluster points, in which case it will not converge to a definite limit. Now define the sequences

$$U_n = \sup_{k \geq n} x_k$$

and

$$L_n = \inf_{k \geq n} x_k \ .$$

Note that U_n is a decreasing sequence and L_n is an increasing sequence. If α is the largest cluster point of $\{x_n\}$, then clearly the U_n will approach α as n increases. Furthermore, no matter how large n gets, U_n will always remain $\geq \alpha$. Similarly, if β is the smallest cluster point of $\{x_n\}$, then all the L_n must be $\leq \beta$. This situation is represented schematically in the figure below.



Let $U = \inf_n U_n$. By Theorem B4 and the remarks following it we see that U_n converges to U. The limit U is called the **upper limit** (or **limit superior**) of the sequence $\{x_n\}$ and will be denoted by $\bar{x}$. In other words,

$$\bar{x} = \inf_n \sup_{k \geq n} x_k = \lim_{n \to \infty} \sup_{k \geq n} x_k \ .$$

The upper limit is frequently also written as $\lim\sup x_n$.

Similarly, L_n converges to $L = \sup_n L_n$. The limit L is called the **lower limit** (or **limit inferior**) of the sequence $\{x_n\}$, and is denoted by $\underline{x}$. Thus

$$\underline{x} = \sup_n \inf_{k \geq n} x_k = \lim_{n \to \infty} \inf_{k \geq n} x_k$$

which is also written as $\lim\inf x_n$. Note that either or both $\bar{x}$ and $\underline{x}$ could be $\pm\infty$.

Theorem B22 If $x_n \leq y_n$ for all n greater than or equal to some fixed N, then $\lim\sup x_n \leq \lim\sup y_n$ and $\lim\inf x_n \leq \lim\inf y_n$.

Proof This is Exercise B.6. ∎

We have already remarked that in general a sequence may have many (or no) cluster points, and hence will not converge. However, suppose $\{x_n\}$ converges to x, and let $\lim U_n = U$. We claim that $x = U$.

To see this, we simply use the definitions involved. Given $\varepsilon > 0$, we may choose N such that for all $n \geq N$ we have both $|x - x_n| < \varepsilon$ and $|U - U_n| < \varepsilon$. Since $U_N = \sup_{k\geq N} x_k$, we see that given this ε, there exists $k \geq N$ such that $U_N - \varepsilon < x_k$ or $U_N - x_k < \varepsilon$. But then we have

$$|U - x| \leq |U - U_N| + |U_N - x_k| + |x_k - x| < 3\varepsilon$$

which proves that $U = x$. In an exactly analogous way, it is easy to prove that $L = \lim L_n = x$ (see Exercise B.7). We have therefore shown that $x_n \to x$ implies $\lim \sup x_n = \lim \inf x_n = x$. That the converse of this statement is also true is given in the next theorem. It should be clear however, that all but a finite number of terms in the sequence $\{x_n\}$ will be caught between U and L, and hence if $U = L$ it must be true that $x_n \to x = U = L$.

Theorem B23 A real-valued sequence $\{x_n\}$ converges to the number x if and only if $\lim \sup x_n = \lim \inf x_n = x$.

Proof Let $U_n = \sup_{k\geq n} x_k$ and $L_n = \inf_{k\geq n} x_k$, and first suppose that $\lim U_n = \lim L_n = x$. Given $\varepsilon > 0$, there exists N such that $|U_n - x| < \varepsilon$ for all $n \geq N$, and there exists M such that $|L_n - x| < \varepsilon$ for all $n \geq M$. These may be written as (see Example 0.6) $x - \varepsilon < U_n < x + \varepsilon$ for all $n \geq N$, and $x - \varepsilon < L_n < x + \varepsilon$ for all $n \geq M$. But from the definitions of U_n and L_n we know that $x_n \leq U_n$ and $L_n \leq x_n$. Hence $x_n < x + \varepsilon$ for all $n \geq N$ and $x - \varepsilon < x_n$ for all $n \geq M$. Therefore $|x_n - x| < \varepsilon$ for all $n \geq \max\{N, M\}$ so that $x_n \to x$.

The converse was shown in the discussion preceding the theorem. ∎

Define S to be the set of all cluster points of $\{x_n\}$. Since any cluster point is the limit of some subsequence, it follows that S is just the set of all subsequential limits of $\{x_n\}$. From the figure above, we suspect that $\sup S = \bar{x}$ and $\inf S = \underline{x}$. It is not hard to prove that this is indeed the case.

Theorem B24 Let $\{x_n\}$ be a sequence of real numbers and let S, $\bar{x}$ and $\underline{x}$ be defined as above. Then $\sup S = \bar{x}$ and $\inf S = \underline{x}$.

Proof This is Exercise B.8. ∎

Example B7 Let $x_n = (-1)^n/(1 + 1/n)$. Then it should be clear that we have $\lim\sup x_n = 1$ and $\lim\inf x_n = -1$. ∕∕

Our next theorem will be very useful in proving several tests for the convergence of series.

Theorem B25 Let $\{x_n\}$ be a sequence of real numbers, let S be the set of all subsequential limits of $\{x_n\}$, and let $\bar{x} = \lim\sup x_n$ and $\underline{x} = \lim\inf x_n$. Then

(a) $\bar{x} \in S$.

(b) If $r > \bar{x}$, then there exists N such that $n \geq N$ implies $x_n < r$.

(c) $\bar{x}$ is unique.

Of course, the analogous results hold for $\underline{x}$ as well.

Proof We will show only the results for $\bar{x}$, and leave the case of $\underline{x}$ to the reader.

(a) Since S (the set of all subsequential limits) lies in the extended number system, we must consider three possibilities. If $-\infty < \bar{x} < +\infty$, then S is bounded above so that at least one subsequential limit exists. Then the set S is closed (by Theorem B16), and hence $\bar{x} = \sup S \in S$ (see Example B2).

If $\bar{x} = +\infty$, then S is not bounded above so that $\{x_n\}$ is not bounded above. Thus there exists a subsequence $\{x_{n_k}\}$ such that $x_{n_k} \to +\infty$. But then $+\infty \in S$ so that $\bar{x} \in S$.

If $\bar{x} = -\infty$, then there is no finite subsequential limit (since $\bar{x}$ is the least upper bound of the set of such limits), and hence S consists solely of the element $-\infty$. This means that given any real M, $x_n > M$ for at most a finite number of indices n so that $x_n \to -\infty$, and hence $\bar{x} = -\infty \in S$.

(b) If there existed an $r > \bar{x}$ such that $x_n \geq r$ for an infinite number of n, then there would be a subsequential limit x' of $\{x_n\}$ such that $x' \geq r > \bar{x}$. This contradicts the definition of $\bar{x}$.

(c) Let $\bar{x}$ and $\bar{y}$ be distinct numbers that satisfy (a) and (b), and suppose $\bar{x} < \bar{y}$. Let r be any number such that $\bar{x} < r < \bar{y}$ (that such an r exists was shown in Theorem 0.4). Since $\bar{x}$ satisfies (b), there exists N such that $x_n < r$ for all $n \geq N$. But then $\bar{y}$ can not possibly satisfy (a). ∎

We now have the background to prove three basic tests for the convergence of series.

Theorem B26 (a) (**Comparison test**) If $\sum b_n$ converges, and if $|a_n| \leq b_n$ for $n \geq N_0$ (N_0 fixed), then $\sum a_n$ converges. If $\sum c_n$ diverges, and if $a_n \geq c_n \geq 0$ for $n \geq N_0$, then $\sum a_n$ diverges.

(b) (**Root test**) Given the series $\sum a_n$, let $\bar{a} = \lim\sup|a_n|^{1/n}$. If $\bar{a} < 1$, then $\sum a_n$ converges, and if $\bar{a} > 1$, then $\sum a_n$ diverges.

(c) (**Ratio test**) The series $\sum a_n$ converges if lim sup $|a_{n+1}/a_n| < 1$, and diverges if $|a_{n+1}/a_n| \geq 1$ for $n \geq N$ (N fixed).

Proof (a) Given $\varepsilon > 0$, there exists $N \geq N_0$ such that $m \geq n \geq N$ implies that $|\sum_{k=n}^{m} b_k| < \varepsilon$ (Theorem B17). Hence $\sum a_n$ converges since

$$\left|\sum_{k=n}^{m} a_k\right| \leq \sum_{k=n}^{m} |a_k| \leq \sum_{k=n}^{m} b_k = \left|\sum_{k=n}^{m} b_k\right| < \varepsilon \ \ .$$

By what has just been shown, we see that if $0 \leq c_n \leq a_n$ and $\sum a_n$ converges, then $\sum c_n$ must also converge. But the contrapositive of this statement is then that if $\sum c_n$ diverges, then so must $\sum a_n$.

(b) First note that $\bar{a} \geq 0$ since $|a_n|^{1/n} \geq 0$. Now suppose that $\bar{a} < 1$. By Theorem B25(b), for any r such that $\bar{a} < r < 1$, there exists N such that $n \geq N$ implies $|a_n|^{1/n} < r$, and thus $|a_n| < r^n$. But $\sum r^n$ converges (Example B5) so that $\sum a_n$ must also converge by the comparison test.

If $\bar{a} > 1$, then by Theorems B22(a) and B14(b), there must exist a sequence $\{n_k\}$ such that $|a_{n_k}|^{1/n_k} \to \bar{a}$. But this means that $|a_n| > 1$ for infinitely many n so that $a_n \nrightarrow 0$ and $\sum a_n$ does not converge (corollary to Theorem B17).

(c) If lim sup$|a_{n+1}/a_n| < 1$ then, by Theorem B25(b), we can find a number $r < 1$ and an integer N such that $n \geq N$ implies $|a_{n+1}/a_n| < r$. We then see that

$$\begin{aligned} &|a_{N+1}| < r|a_N| \\ &|a_{N+2}| < r|a_{N+1}| < r^2|a_N| \\ &\qquad\vdots \\ &|a_{N+p}| < r^p|a_N| \ \ . \end{aligned}$$

Therefore, letting $n = N + p$ we have

$$|a_n| \ < \ r^{n-N}|a_N| \ = \ r^{-N}|a_N|r^n$$

for $n \geq N$, and hence $\sum a_n$ converges by the comparison test and Example B6.

If $|a_{n+1}| \geq |a_n|$ for $n \geq N$ (N fixed), then clearly $a_n \nrightarrow 0$ so that $\sum a_n$ can not converge (corollary to Theorem B17). ∎

Note that if $\bar{a} = 1$ when applying the root test we get no information since, for example, $\sum 1/n$ and $\sum 1/n^2$ both have $\bar{a} = 1$ (corollary to Theorem B21), but the first diverges whereas the second converges (see Example B5).

Example B8 Consider the function e^x (= exp x) defined by

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

where x may be real or complex. To test for convergence, we observe that $|a_{n+1}/a_n| = |x/(n+1)|$ so that

$$\begin{aligned}\lim\sup\left|\frac{a_{n+1}}{a_\mathrm{n}}\right| &= \lim_{n\to\infty} \sup_{k\geq n}\left|\frac{x}{k+1}\right| \\ &= \lim_{n\to\infty}\left|\frac{x}{n+1}\right| \\ &= 0\end{aligned}$$

and hence the series converges by the ratio test.

It is of great use in both mathematics and physics to prove that

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n \quad .$$

While this can be proved by taking the logarithm of both sides and using l'Hospital's rule, we shall follow a direct approach. Let $x_n = (1 + x/n)^n$. Expanding x_n by the binomial theorem we have (for n > 2)

$$\begin{aligned}x_n &= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\left(\frac{x}{n}\right)^k \\ &= 1 + x + \frac{n(n-1)}{2!}\frac{x^2}{n^2} + \frac{n(n-1)(n-2)}{3!}\frac{x^3}{n^3} + \cdots + \frac{x^n}{n^n} \quad .\end{aligned}$$

If we write

$$\begin{aligned}\frac{1}{n^n} = \frac{n!}{n!}\frac{1}{n^n} &= \frac{1}{n!}\frac{n(n-1)(n-2)\cdots(n-(n-1))}{n^n} \\ &= \frac{1}{n!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{n-1}{n}\right)\end{aligned}$$

then

$$\begin{aligned} x_n = 1 + x &+ \frac{1}{2!}\left(1-\frac{1}{n}\right)x^2 + \frac{1}{3!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)x^3 \\ &+\cdots+\frac{1}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)x^k \\ &+\cdots+\frac{1}{n!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{n-1}{n}\right)x^n\ . \end{aligned}$$

We now treat each x_n as an infinite series by defining all terms with $k > n$ to be zero, and we consider the difference

$$e^x - x_n = \sum_{k=2}^{\infty}\frac{1}{k!}\left[1-\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right]x^k\ . \qquad (*)$$

Applying Theorem B17 to the convergent series $e^x = \sum|x|^n/n!$, we see that for fixed x and $\varepsilon > 0$, we can choose an integer m sufficiently large that

$$\sum_{k=m+1}^{\infty}\frac{|x|^k}{k!} < \varepsilon/2\ .$$

Writing (*) in the form $\sum_{k=2}^{\infty} = \sum_{k=2}^{m} + \sum_{k=m+1}^{\infty}$ and noting that the coefficient of x^k in the (second) sum is ≥ 0 but $\leq 1/k!$, we obtain (for $n > m$)

$$|e^x - x_n| \leq \sum_{k=2}^{m}\frac{1}{k!}\left[1-\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right]|x^k| + \varepsilon/2\ .$$

Since the sum in this expression consists of a finite number of terms, each of which approaches 0 as $n \to \infty$, we may choose an $N > 0$ such that the sum is less than $\varepsilon/2$ for $n > N$. Therefore, for $n > N$ we have $|e^x - x_n| < \varepsilon$ which proves that $x_n \to e^x$. ˙ //

Exercises

1. Prove Theorem B2(b).

2. Let $\{x_n\}$ and $\{y_n\}$ be sequences of real numbers such that $x_n \leq y_n$ for all $n \geq N$ where N is fixed. If $x_n \to x$ and $y_n \to y$, prove that $x \leq y$.
3. If A is a subset of a metric space X and $x \in X$, prove that $x \in \operatorname{Ext} A$ if and only if x has *some* neighborhood disjoint from A.

4. Prove Theorem B14(b).

5. Prove that any cluster point is the limit of a subsequence.

6. Prove Theorem B22.

7. If $\{x_n\}$ is a sequence of real numbers converging to x, and $L_n = \inf_{k\geq n} x_k$ converges to L, show that x = L.

8. Prove Theorem B24.

9. Let $\bar{\mathbb{R}}$ denote the extended real number system, and let $f\colon (a, b) \subset \mathbb{R} \to \bar{\mathbb{R}}$. Define

$$\lim_{x\to y} \sup f(x) = \inf_{\delta > 0} \sup_{0< |x-y| < \delta} f(x)$$

and suppose that $\lim_{x\to y} f(x) = L$ (i.e., $\lim_{x\to y} f(x)$ exists). Show

$$\lim_{x\to y} \sup f(x) = L \ .$$

[*Hint*: Let $S_\delta = \sup_{|x-y| < \delta} f(x)$ and define $S = \inf_\delta S_\delta$. Then note that

$$|S - L| \leq |S - S_\delta| + |S_\delta - f(x)| + |f(x) - L| \ .]$$

10. (a) Let $\{x_n\}$ be a Cauchy sequence in $\mathbb{R}^n$, and assume that $\{x_n\}$ has a cluster point c. Prove that $\{x_n\}$ converges to c.

(b) Using this result, prove that any Cauchy sequence in $\mathbb{R}^n$ converges to a point of $\mathbb{R}^n$.

11. Show that Bd A = Cl A ∩ Cl A^c.

12. If U ⊂ (X, d) is open and A ⊂ X is dense, show that Cl U = Cl(U ∩ A).

13. If $\{x_n\}$ is a Cauchy sequence with a convergent subsequence $\{x_{n_k}\}$, show that $\{x_n\}$ converges.

APPENDIX C

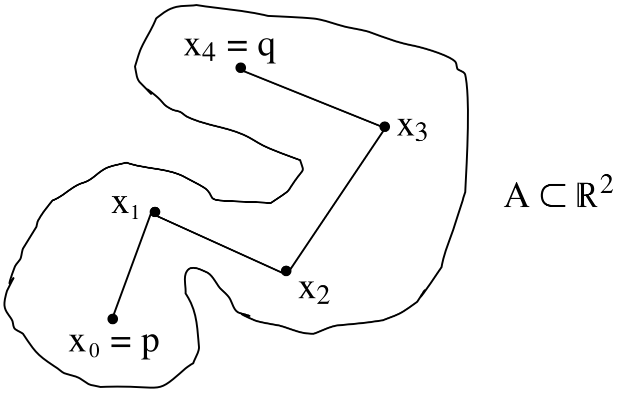
Path Connectedness

In order to avoid having to define a general topological space, we shall phrase this appendix in terms of metric spaces. However, the reader should be aware that this material is far more general than we are presenting it. We assume that the reader has studied Appendix A.

In elementary analysis and geometry, one thinks of a curve as a collection of points whose coordinates are continuous functions of a real variable t. For example, a curve in the plane $\mathbb{R}^2$ may be specified by giving its coordinates $(x = f(t), y = g(t))$ where f and g are continuous functions of the parameter t. If we require that the curve join two points p and q, then the parameter can always be adjusted so that $t = 0$ at p and $t = 1$ at q. Thus we see that the curve is described by a continuous mapping from the unit interval $I = [0, 1]$ into the plane.

Let X be a metric space, and let $I = [0, 1]$ be a subspace of $\mathbb{R}$ with the usual metric. We define a **path** in X, joining two points p and q of X, to be a continuous mapping $f: I \to X$ such that $f(0) = p$ and $f(1) = q$. This path will be said to lie in a subset $A \subset X$ if $f(I) \subset A$. It is important to realize that *the path is the mapping* f, and not the set of image points f(I). The space X is said to be **path connected** if for every $p, q \in X$ there exists a path in X joining p and q. If $A \subset X$, then A is path connected if every pair of points of A can be joined by a path in A. (We should note that what we have called path connected is sometimes called **arcwise connected**.)

Let us consider for a moment the space $\mathbb{R}^n$. If x_i, $x_j \in \mathbb{R}^n$, then we let $\overline{x_i x_j}$ denote the closed line segment joining x_i and x_j. A subset $A \subset \mathbb{R}^n$ is said to be **polygonally connected** if given any two points p, $q \in A$ there are points $x_0 = p, x_1, x_2, \dots, x_m = q$ in A such that $\cup_{i=1}^{m} \overline{x_{i-1} x_i} \subset A$.



Just because a subset of $\mathbb{R}^n$ is path connected does not mean that it is polygonally connected. For example, the unit circle in $\mathbb{R}^2$ is path connected since it is actually a path itself, but it is not polygonally connected.

Example C1 The space $\mathbb{R}^n$ is path connected. Indeed, if $p \in \mathbb{R}^n$ has coordinates $(x^1, \dots, x^n)$ and $q \in \mathbb{R}^n$ has coordinates $(y^1, \dots, y^n)$, then we define the mapping $f: I \rightarrow \mathbb{R}^n$ by $f(t) = (f^1(t), \dots, f^n(t))$ where $f^i(t) = (1 - t)x^i + ty^i$. This mapping is clearly continuous and satisfies $f(0) = p$ and $f(1) = q$. Thus f is a path joining the arbitrary points p and q of $\mathbb{R}^n$, and hence $\mathbb{R}^n$ is path connected. //

The following is a simple consequence of Theorem A5 that we shall need for our main result (i.e., Theorem C2).

Theorem C1 Let $f: (X_1, d_1) \rightarrow (X_2, d_2)$ and $g: (X_2, d_2) \rightarrow (X_3, d_3)$ both be continuous functions. Then $g \circ f: (X_1, d_1) \rightarrow (X_3, d_3)$ is a continuous function.

Proof If $U \subset X_3$ is open, then the continuity of g shows that $g^{-1}(U) \subset X_2$ is open. Therefore $(g \circ f)^{-1}(U) = (f^{-1} \circ g^{-1})(U) = f^{-1}(g^{-1}(U))$ is open by the continuity of f. ∎

Theorem C2 Let f be a continuous mapping from a metric space X *onto* a metric space Y. Then Y is path connected if X is.

Proof Let x', y' be any two points of Y. Then (since f is surjective) there exist $x, y \in X$ such that $f(x) = x'$ and $f(y) = y'$. Since X is path connected, there exists a path g joining x and y such that $g(0) = x$ and $g(1) = y$. But then $f \circ g$ is a continuous function (Theorem C1) from I into Y such that $(f \circ g)(0) = x'$ and $(f \circ g)(1) = y'$. In other words, $f \circ g$ is a path joining x' and y', and hence Y is path connected. ∎

It is an obvious corollary of Theorem C2 that if f is a continuous mapping from the path connected space X *into* Y, then f(X) is path connected in Y since f maps X *onto* the subspace f(X).

Bibliography

This bibliography lists those books referred to in the text as well as many of the books that we found useful in our writing.

Abraham, R., Marsden, J. E. and Ratiu, T., *Manifolds, Tensor Analysis, and Applications*, Addison-Wesley, Reading, MA, 1983.

Adler, R., Bazin, M. and Schiffer, M., *Introduction to General Relativity*, McGraw-Hill, New York, 1965.

Arnold, V. I., *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.

Biggs, N. L., *Discrete Mathematics*, Clarendon Press, Oxford, 1985.

Bishop, R. L. and Goldberg, S. I., *Tensor Analysis on Manifolds*, Macmillan, New York, 1968.

Boothby, W. M., *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

Byron, F. W. Jr., and Fuller, R. W., *Mathematics of Classical and Quantum Physics*, Addison-Wesley, Reading, MA, 1969.

Curtis, C. W., *Linear Algebra*, Springer-Verlag, New York, 1984.

Curtis, W. D. and Miller, F. R., *Differential Manifolds and Theoretical Physics*, Academic Press, Orlando, FL, 1985.

Dennery, P. and Krzywicki, A., *Mathematics for Physicists*, Harper & Row, New York, 1967.

Durbin, J. R., *Modern Algebra*, John Wiley & Sons, New York, 1985.

Flanders, H., *Differential Forms*, Academic Press, New York, 1963.

Frankel, T., *The Geometry of Physics*, 2nd edition, Cambridge University Press, New York, NY, 2004.

Frankel, T., *Linear Algebra*, unpublished lecture notes, University of California, San Diego, 1980.

Friedberg, S. H. and Insel, A. J., *Introduction to Linear Algebra with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

Friedberg, S. H., Insel, A. J. and Spence, L. E., *Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

Gemignani, M. C., *Elementary Topology*, 2nd edition, Addison-Wesley, Reading, MA, 1972.

Geroch, R., *Mathematical Physics*, University of Chicago Press, Chicago, IL, 1985.

Goldstein, H., *Classical Mechanics*, 2nd edition, Addison-Wesley, Reading, MA, 1980.

Halmos, P. R., *Finite-Dimensional Vector Spaces*, Springer-Verlag, New York, 1974.

Herstein, I. N., *Topics in Algebra*, Xerox College Publishing, Lexington, MA, 1964.

Hoffman, K. and Kunze, R., *Linear Algebra*, 2nd edition, Prentice-Hall, Englewood Cliffs, NJ, 1971.

Jackson, J. D., *Classical Electrodynamics*, 2nd edition, John Wiley & Sons, New York, 1975.

Johnson, R. E., *Linear Algebra*, Prindle, Weber & Schmidt, Boston, MA, 1967.

Knopp, K., *Infinite Sequences and Series*, Dover Publications, New York, 1956.

Kolmogorov, A. N. and Fomin, S. V., *Functional Analysis*, Graylock Press, Rochester, NY, 1957.

Lang, S., *Analysis I*, Addison-Wesley, Reading, MA, 1968.

Lang, S., *Linear Algebra*, 2nd edition, Addison-Wesley, Reading, MA, 1971.

Lang, S., *Real Analysis*, 2nd edition, Addison-Wesley, Reading, MA, 1983.

Lipschutz, S., *Linear Algebra*, Schaum's Outline Series, McGraw-Hill, New York, 1968.

Marcus, M. and Minc, H., *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA, 1964.

Marcus, M., *Finite-Dimensional Multilinear Algebra, Parts I and II*, Marcel Dekker, New York, 1973.

Marcus, M., *Introduction to Linear Algebra*, Dover Publications, New York, 1988.

Marion, J. B., *Classical Dynamics of Particles and Systems*, 2nd edition, Academic Press, Orlando, FL, 1970.

Marsden, J. E., *Elementary Classical Analysis*, W. H. Freeman, San Francisco, CA, 1974.

Misner, C. W., Thorne, K. S. and Wheeler, J. A., *Gravitation*, W. H. Freeman, San Francisco, CA, 1973.

Munkres, J. R., *Topology*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

Murdoch, D. C., *Linear Algebra*, John Wiley & Sons, New York, 1970.

Prugovecki, E., *Quantum Mechanics in Hilbert Space*, 2nd edition, Academic Press, New York, 1981.

Reed, M. and Simon, B., *Functional Analysis*, Revised and Enlarged edition, Academic Press, Orlando, FL, 1980.

Roach, G. F., *Green's Functions*, 2nd edition, Cambridge University Press, Cambridge, 1982.

Royden, H. L., *Real Analysis*, 2nd edition, Macmillan, New York, 1968.

Rudin, W., *Functional Analysis*, 2nd edition, McGraw-Hill, New York, 1974.

Rudin, W., *Principles of Mathematical Analysis*, 3rd edition, McGraw-Hill, New York, 1976.

Ryder, L. H., *Quantum Field Theory*, Cambridge University Press, Cambridge, 1985.

Schutz, B., *Geometrical Methods of Mathematical Physics*, Cambridge University Press, Cambridge, 1980.

Shankar, R., *Principles of Quantum Mechanics*, Plenum Press, New York, 1980.

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

Talman, J. D., *Special Functions*, W. A. Benjamin, New York, 1968.

Taylor, J. R., *Scattering Theory*, John Wiley & Sons, New York, 1972.

Tung, W., *Group Theory in Physics*, World Scientific, Philadelphia, PA, 1985.

Index