CHAPTER 1

SOME USEFUL TOOLS

1.1 Introduction

One aim of this book is to make a significant body of mathematics accessible to people in various disciplines, including engineering, geophysics, computer science, the physical sciences, and applied mathematics. People who have had substantial mathematical training enjoy a head start in this enterprise, since they are more likely to be familiar with ideas that, too often, receive little emphasis outside departments of mathematics. The purpose of this preliminary chapter is to level the playing field by reviewing mathematical notations and concepts used throughout the book. We assume that the reader is familiar with concepts from elementary calculus, such as limits, continuity, differentiation, and integration. In three sections (2.8, 7.3, and 9.3) we refer to concepts associated with Fourier series.

Virtually every entity in mathematics is a set. If x is an element of the set S, we write $x \in S$ and say that x **belongs to** S. If every element of a set R also belongs to the set S, we say that R is a **subset** of S and write $R \subset S$. Using this concept, we

Numerical Analysis for Applied Science, Second Edition. Myron B. Allen III and Eli L. Isaacson

say that $R = S$ provided $R \subset S$ and $S \subset R$. There are several ways to specify the elements of a set. One way is simply to list them:

$$R = \Big\{2, 4, 6\Big\}, \quad S = \Big\{2, 4, 6, 8, 10, \ldots\Big\}.$$

Another is to give a rule for selecting elements from a previously defined set. For example,

$$R = \Big\{x \in S \,\Big|\, x \leqslant 6\Big\}$$

denotes the set of all elements of S that are less than or equal to 6. If the statement $x \in S$ fails for all x, then S is the **empty set**, denoted as $\varnothing$.

The notation $x = y$ should be familiar enough, but two related notions are worth mentioning. By $x \leftarrow y$, we mean "assign the value held by the variable y to the variable x." Distinguishing between $x = y$ and $x \leftarrow y$ can seem pedantic until one recalls such apparent nonsense as "$k = k + 1$" that occur in Fortran and other programming languages. Also, we use $x := y$ to indicate that x is defined to have the value y.

If R and S are sets, then $R \cup S$ is their **union**, which is the set containing all elements of R and all elements of S. The **intersection** $R \cap S$ is the set of all elements that belong to *both* R and S. If S_i is a set for each i belonging to some index set I, then

$$\bigcup_{i \in I} S_i, \quad \bigcap_{i \in I} S_i$$

denote, respectively, the set containing all elements that belong to at least one of the sets S_i and the set containing just those elements that belong to every S_i. The **set difference** $R \backslash S = \{x \in \mathbb{R} \,|\, x \notin S\}$ is the set of all elements of R that do not belong to S. If $S_1, S_2, \ldots, S_n$ are sets, then their **Cartesian product** $S_1 \times S_2 \times \cdots \times S_n$ is the set of all **ordered n-tuples** $(x_1, x_2, \ldots, x_n)$, where each $x_i \in S_i$. Two such n-tuples $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are equal precisely when $x_1 = y_1$, $x_2 = y_2$, $\ldots$, $x_n = y_n$.

Among the most commonly occurring sets in this book are $\mathbb{R}$, the set of all real numbers; $\mathbb{C}$, the set of all complex numbers $x + iy$, where $x, y \in \mathbb{R}$ and $i^2 = -1$, and

$$\mathbb{R}^n := \underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}},$$

the set of all n-tuples $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of real numbers. We often write these n-tuples as **column vectors**:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

$\mathbb{R}$ itself has several important types of subsets, including **open intervals**,

$$(a, b) := \Big\{ x \in \mathbb{R} \;\Big|\; a < x < b \Big\};$$

closed intervals,

$$[a, b] := \Big\{ x \in \mathbb{R} \;\Big|\; a \leqslant x \leqslant b \Big\};$$

and the **half-open** intervals

$$[a, b) := \Big\{ x \in \mathbb{R} \;\Big|\; a \leqslant x < b \Big\}, \quad (a, b] := \Big\{ x \in \mathbb{R} \;\Big|\; a < x \leqslant b \Big\}.$$

To extend this notation, we sometimes use the symbol ∞ in a slightly abusive fashion:

$$(a, \infty) := \Big\{ x \in \mathbb{R} \;\Big|\; a < x \Big\},$$

$$(-\infty, b] := \Big\{ x \in \mathbb{R} \;\Big|\; x \leqslant b \Big\},$$

$$(-\infty, \infty) := \mathbb{R},$$

and so forth.

In specifying functions, we write $f \colon R \to S$. This graceful notation indicates that $f(x)$ is defined for every element x belonging to R, the **domain** of f, and that each such value $f(x)$ belongs to the set S, called the **codomain** of f. The codomain of f contains as a subset the set $f(R)$ of all images $f(x)$ of points x belonging to the domain R. We call $f(R)$ the **range** of f.

The notation $f \colon x \mapsto y$ indicates that $f(x) = y$, the domain and codomain of f being understood from context. Sometimes we write $x \mapsto y$ when the function itself as well as its domain and codomain are understood from context.

Throughout this book we assume that readers are familiar with the basics of calculus and linear algebra. However, it may be useful to review a few notions from these subjects. We devote the rest of this chapter to a summary of facts about bounded sets and normed vector spaces and some frequently used results from calculus.

1.2 Bounded Sets

In numerical analysis, sets of real numbers arise in many contexts. Examples include sequences of approximate values for some quantity, ranges of values for the errors in such approximations, and so forth. It is often important to estimate where these sets lie on the real number line – for example, to guarantee that the possible values for a numerical error lie in a small region around the origin. We say that a set $S \subset \mathbb{R}$ is **bounded above** if there exists a number $B \in \mathbb{R}$ such that $x \leqslant B$ for every $x \in S$. In this case, B is an **upper bound** for S. Similarly, S is **bounded below** if, for some $b \in \mathbb{R}$, $b \leqslant x$ for every $x \in S$. In this case, b is a **lower bound** for S. A **bounded** set is one that is bounded both above and below. A set S is bounded if and only if there exists a number $M \in \mathbb{R}$ such that $|x| \leqslant M$ for every $x \in S$.

By extension, if $f\colon S \to \mathbb{R}$ is a function whose range $f(S)$ is bounded above, bounded below, or bounded, then we say that f is bounded above, bounded below, or bounded, respectively.

1.2.1 The Least Upper Bound Principle

Most upper and lower bounds give imprecise information. For example, 17 is an upper bound for the set $S = (0, 2)$, but, as Figure 1.1 illustrates, the upper bound 2 is sharper. We call B_0 a **least upper bound** or **supremum** for $S \subset \mathbb{R}$ if B_0 is an upper bound for S and $B_0 \leqslant B$ whenever B is an upper bound for S. In this case, we write $B_0 = \sup S$. Similar reasoning applies to lower bounds: -109 is a lower bound for $(0, 2)$, but so is the more informative number 0. We call b_0 a **greatest lower bound** or **infimum** for $S \subset \mathbb{R}$ if b_0 is a lower bound for S and $b_0 \geqslant b$ whenever b is also a lower bound for S. We write $b_0 = \inf S$. The notations $\inf$ and $\sup$ have obvious extensions. For example, if $S_2 := \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ denotes the unit circle in $\mathbb{R}^2$ and $f\colon S_2 \to \mathbb{R}$ is a real-valued function defined on S_2, then

$$\sup_{S_2} f := \sup_{(x,y)\in S_2} f(x, y) := \sup\Big\{ f(x, y) \in \mathbb{R} \;\Big|\; x^2 + y^2 = 1 \Big\}. \tag{1.1}$$

Shortly we discuss conditions under which this quantity exists.

Figure 1.1 The set $(0, 2) \subset \mathbb{R}$ and two of its upper bounds.

Not every set has a supremum or an infimum. For example, the set

$$\mathbb{Z} = \Big\{ \ldots, -2, -1, 0, 1, 2, \ldots \Big\}$$

of all **integers** has neither a supremum nor an infimum. The set

$$\mathbb{N} = \Big\{1, 2, 3, \ldots\Big\}$$

of **natural numbers** has infimum 1 but no supremum. One should take care to distinguish between $\sup S$ and $\inf S$ and the notions of maximum and minimum. By a **maximum** of a set $S \subset \mathbb{R}$, we mean an element $M \in S$ for which $x \leqslant M$ whenever $x \in S$, and we write $M = \max S$. Thus $\sup(0,2) = 2 = \sup[0,2] = \max[0,2]$, but $\max(0,2)$ does not exist. Similarly, an element $m \in S$ is a **minimum** of S if $m \leqslant x$ for every $x \in S$. Thus, $\inf(0,2) = 0 = \inf[0,2] = \min[0,2]$, while $\min(0,2)$ does not exist. These examples illustrate the fact that $\sup$ and $\inf$ are more general notions than $\max$ and $\min$: $\sup S = \max S$ when $\sup S \in S$, but $\sup S$ may exist even when $\max S$ does not. A corresponding statement holds for $\inf$ and $\min$.

The following principle, which one can take as a defining characteristic of $\mathbb{R}$, confirms the fundamental importance of $\sup$ and $\inf$:

LEAST-UPPER-BOUND PRINCIPLE. *If a nonempty subset of $\mathbb{R}$ is bounded above, then it has a least upper bound.*

Spivak [46, Chapter 8] gives an accessible introduction to this principle. Similarly, every nonempty subset of $\mathbb{R}$ that is bounded below has a greatest lower bound. For example,

$$\inf\Big\{\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4}, \ldots\Big\} = 0, \quad \sup(-\infty, 0) = 0, \quad \sup\Big\{2, 4, 6\Big\} = 6.$$

The set $\big\{2, 4, 6, 8, 10, \ldots\big\}$, however, is not bounded above, and it has no least upper bound. The least-upper-bound principle ensures that $\sup_{S_2} f$, defined in Eq. (1.1), exists whenever the set of real numbers

$$\Big\{f(x,y) \in \mathbb{R} \;\Big|\; (x,y) \in S_2\Big\}$$

is bounded above. However, without knowing more about f, we cannot guarantee the existence of a point $(x,y) \in S_2$ where f attains the value $\sup_{S_2} f$.

1.2.2 Bounded Sets in $\mathbb{R}^n$

Which subsets of $\mathbb{R}^n$ are bounded? Here we generally have no linear order analogous to the relation $\leqslant$ on which to base a definition of boundedness. Instead, we rely on the idea of distance, which is familiar from geometry.

DEFINITION. The **Euclidean length** of $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is

$$\|\mathbf{x}\|_2 := \sqrt{\sum_{j=1}^{n} x_j^2}.$$

The **Euclidean distance** between two points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is the Euclidean length of their difference, $\|\mathbf{y} - \mathbf{x}\|_2$.

Given a point $\mathbf{x} \in \mathbb{R}^n$ and a positive real number r, we call the set of all points in $\mathbb{R}^n$ whose Euclidean distance from $\mathbf{x}$ is less than r the **ball** of radius r about $\mathbf{x}$. We denote this set as $\mathcal{B}_r(\mathbf{x})$. Figure 1.2 depicts such a set in $\mathbb{R}^2$. A set $S \subset \mathbb{R}^n$ is **bounded** if it is a subset of $\mathcal{B}_r(\mathbf{0})$ for some $r > 0$. Observe that, if $x \in \mathbb{R} = \mathbb{R}^1$, then $\mathcal{B}_r(x) = (x - r, x + r)$. One easily checks that a subset of $\mathbb{R}$ is bounded in this sense if and only if it is bounded above and below.

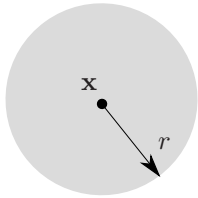

Figure 1.2 The ball $\mathcal{B}_r(\mathbf{x})$ of radius r about the point $\mathbf{x} \in \mathbb{R}^2$.

Other structural aspects of $\mathbb{R}^n$ also prove useful. Let $S \subset \mathbb{R}^n$. A point $\mathbf{x} \in S$ is an **interior point** of S if there is *some* ball $\mathcal{B}_r(\mathbf{x})$ such that $\mathcal{B}_r(\mathbf{x}) \subset S$. In Figure 1.3, the point $\mathbf{a}$ is an interior point of S, but $\mathbf{b}$ and $\mathbf{c}$ are not. A point $\mathbf{x} \in \mathbb{R}^n$ (not necessarily belonging to S) is a **limit point** of S if *every* ball $\mathcal{B}_r(\mathbf{x})$ contains at least one element of S distinct from $\mathbf{x}$. In Figure 1.4, $\mathbf{a}$ and $\mathbf{b}$ are limit points of S, but $\mathbf{c}$ is not. If every element of S is an interior point, then we call S an **open** set. If S contains all of its limit points, then we say that S is a **closed** set. The definitions are by no means mutually exclusive: $\varnothing$ and $\mathbb{R}^n$ are both open and closed.

Finally, a subset of $\mathbb{R}^n$ that is both closed and bounded is **compact**.[1] Thus the following subsets of $\mathbb{R}^2$ are compact:

$$[0,1] \times [0,1], \quad \Big\{(0,0), (0,\pi), (1,-\pi)\Big\}, \quad S_2 = \Big\{\mathbf{x} \in \mathbb{R}^2 \;\Big|\; \|\mathbf{x}\|_2 = 1\Big\},$$

while the sets

$$(0,1) \times (0,1), \quad \mathcal{B}_1(\mathbf{0}), \quad \Big\{(0,0), (1,1), (2,2), \ldots\Big\}$$

[1]This characterization of compactness is not the most general one, but it suffices for $\mathbb{R}^n$. For the more general definition, see [40, Chapter 2].

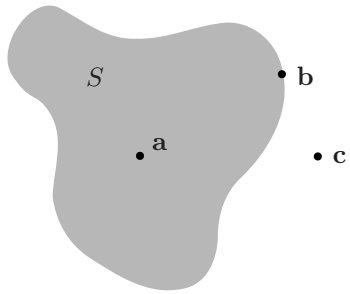

Figure 1.3 A set $S \subset \mathbb{R}^2$, showing an interior point $\mathbf{a}$ and two points $\mathbf{b}, \mathbf{c}$ that are not interior points.

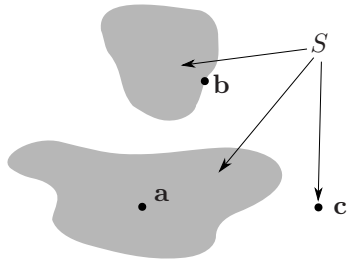

Figure 1.4 A set $S \subset \mathbb{R}^2$, along with two limit points $\mathbf{a}$ and $\mathbf{b}$ and a point $\mathbf{c}$ that is not a limit point of S.

are not. Compact sets in $\mathbb{R}^n$ have several interesting properties, one of which is especially useful in numerical analysis.

THEOREM 1.2.1 (MAXIMUM AND MINIMUM VALUES ON COMPACT SETS) *If $S \subset \mathbb{R}^n$ is nonempty and compact and $f\colon S \to \mathbb{R}$ is a continuous function, then there are points $\mathbf{a}, \mathbf{b} \in S$ for which $f(\mathbf{a})$ and $f(\mathbf{b})$ are the minimum and maximum, respectively, of the set $f(S)$.*

For a proof, see [40, Chapter 4].

This theorem partially settles an issue raised earlier: If f is a continuous, real-valued function defined on the unit circle S_2, then there is at least one point $(x, y) \in S_2$ where f takes the value $\sup_{S_2} f$ defined in Eq. (1.1). By considering the function $-f$, one can also show that f takes the value $\inf_{S_2} f$ at some point in S_2. Both of these statements hold just as well in $\mathbb{R}^n$, where $S_2 := \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1\}$. We use this generalization in the next section.

1.3 Normed Vector Spaces

1.3.1 Vector Spaces

Vector spaces are ubiquitous.

DEFINITION. A set $\mathcal{V}$ is a **vector space** over $\mathbb{R}$ if there are two operations, **addition** $(+)$ and **scalar multiplication**, that obey the following rules for any $x, y, z \in \mathcal{V}$ and $a, b \in \mathbb{R}$:

1. $x+y \in \mathcal{V}$ *and* $ax \in \mathcal{V}$*; in other words,* $\mathcal{V}$ *is* **closed algebraically** *under addition and scalar multiplication.*
2. $x + y = y + x$.
3. $x + (y + z) = (x + y) + z$.
4. *There is a unique vector* $0 \in \mathcal{V}$ *such that* $x + 0 = x$ *for all* $x \in \mathcal{V}$.
5. *For any* $x \in \mathcal{V}$*, there is a unique vector* $-x \in \mathcal{V}$ *such that* $-x + x = 0$.
6. $1x = x$.
7. $a(bx) = (ab)x$.
8. $a(x + y) = ax + ay$.
9. $(a + b)x = ax + bx$.

We refer to $\mathbb{R}$ *as the field of* **scalars**. *The elements of* $\mathcal{V}$ *are* **vectors**. *A set* $\mathcal{U}$ *is a* **subspace** *of* $\mathcal{V}$ *if every element of* $\mathcal{U}$ *belongs to* $\mathcal{V}$ *and* $\mathcal{U}$ *is a vector space under the operations that it inherits from* $\mathcal{V}$. *Analogous definitions hold for vector spaces over the field* $\mathbb{C}$ *of complex numbers.*

We denote the scalar multiple ax by juxtaposing the scalar a and the vector x. In most cases of interest in this book, the algebraic properties of addition and scalar multiplication are obvious from the definitions of the two operations, and the main issue is whether $\mathcal{V}$ is closed algebraically under these two operations.

Among the common examples of vector spaces are the finite-dimensional **Euclidean spaces** $\mathbb{R}^n$, with their familiar rules of addition and scalar multiplication:

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} := \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix};$$

$$a\mathbf{x} = a\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} := \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix}.$$

In this vector space, the zero vector is $\mathbf{0}$, the array that has 0 as each of its n entries. The real line $\mathbb{R}$ is perhaps the simplest Euclidean space.

Various sets of functions constitute another important class of vector spaces. For example, if $S \subset \mathbb{R}$ is an interval, then $C^k(S)$ signifies the vector space of all functions $f\colon S \to \mathbb{R}$ for which f and its derivatives $f', f'', \ldots, f^{(k)}$ through order k are continuous. By extension of this notation, $C^\infty(S)$ denotes the vector space of functions that have continuous derivatives of all orders on S. On all of these spaces we define addition and scalar multiplication pointwise:

$$(f+g)(x) := f(x) + g(x); \quad (af)(x) := a\,f(x).$$

Here, the vector 0 is the function that assigns the number 0 to all arguments x. A slightly more general function space is $L^2(S)$. Although the rigorous definition of this space involves some technicalities, for our purposes it suffices to think of $L^2(S)$ as the set of all functions $f\colon S \to \mathbb{R}$ for which $\int_S f^2(x)\,dx$ exists and is finite. Readers who are curious about the technicalities may consult [40, Chapter 11].

A third class of vector spaces consists of the sets $\mathbb{R}^{m\times n}$ of real $m \times n$ matrices. Our notational convention is to a use *sans-serif* capital letter, such as A, to signify the matrix whose entry in row i, column j is the number denoted by the corresponding lowercase symbol $a_{i,j}$. If C and D are two such matrices, then

$$\begin{aligned} \mathsf{C}+\mathsf{D} &= \begin{bmatrix} c_{1,1} & \cdots & c_{1,n} \\ \vdots & & \vdots \\ c_{m,1} & \cdots & c_{m,n} \end{bmatrix} + \begin{bmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & & \vdots \\ d_{m,1} & \cdots & d_{m,n} \end{bmatrix} \\ &:= \begin{bmatrix} c_{1,1}+d_{1,1} & \cdots & c_{1,n}+d_{1,n} \\ \vdots & & \vdots \\ c_{m,1}+d_{m,1} & \cdots & c_{m,n}+d_{m,n} \end{bmatrix}, \\ a\mathsf{C} &:= \begin{bmatrix} ac_{1,1} & \cdots & ac_{1,n} \\ \vdots & & \vdots \\ ac_{m,1} & \cdots & ac_{m,n} \end{bmatrix}. \end{aligned}$$

The additive identity in $\mathbb{R}^{m\times n}$ is the $m \times n$ matrix $\mathsf{0}$ all of whose entries are 0.

Finally, the set $\{\mathbf{0}\}$ is trivially a vector space.

One can use addition and scalar multiplication to construct subspaces.

DEFINITION. If $\mathcal{V}$ is a real vector space, a **linear combination** of the vectors $x_1, x_2, \ldots, x_n \in \mathcal{V}$ is a vector of the form $c_1x_1 + c_2x_2 + \cdots + c_nx_n$, where $c_1, c_2, \ldots, c_n \in \mathbb{R}$. If $S \subset \mathcal{V}$, the **span** of S, denoted $\mathrm{span}(S)$, is the set of all linear combinations of vectors belonging to S. If $\mathcal{U} = \mathrm{span}(S)$, then S **spans** $\mathcal{U}$.

Problem 1.2 asks for proof that $\mathrm{span}(S)$ is a subspace of $\mathcal{V}$ whenever $S \subset \mathcal{V}$.

DEFINITION. If $\mathcal{V}$ is a vector space, then a set $S \subset \mathcal{V}$ is **linearly independent** if no vector $x \in S$ belongs to $\mathrm{span}\,(S\backslash\{x\})$, that is, no vector in S is a linear combination of the other vectors in S. Otherwise, S is **linearly dependent**.

One can regard a linearly independent set as containing minimal information needed to determine its span.

DEFINITION. A subset S of a vector space $\mathcal{V}$ is a **basis** for $\mathcal{V}$ if S is linearly independent and $\mathrm{span}(S) = \mathcal{V}$.

A basic theorem of linear algebra asserts that, whenever two finite sets S_1 and S_2 are bases for a vector space $\mathcal{V}$, S_1 and S_2 have the same number of elements (see Ref. [48, Chapter 2]) We call this number the **dimension** of $\mathcal{V}$. For example, $\mathbb{R}^n$ has the **standard basis** $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$, where

$$\mathbf{e}_1 := \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 := \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad \mathbf{e}_n := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

If $\mathcal{V}$ has a basis containing finitely many vectors, then we say that $\mathcal{V}$ is **finite-dimensional**. If not, then $\mathcal{V}$ is **infinite-dimensional**.

1.3.2 Matrices as Linear Operators

Given matrices $\mathsf{A} \in \mathbb{R}^{m\times n}$ and $\mathsf{B} \in \mathbb{R}^{n\times p}$, one can compute their **matrix product**

$$\mathsf{AB} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \begin{bmatrix} b_{1,1} & \cdots & b_{1,p} \\ & & \\ \vdots & & \vdots \\ & & \\ b_{n,1} & \cdots & b_{n,p} \end{bmatrix} = \begin{bmatrix} c_{1,1} & \cdots & c_{1,p} \\ \vdots & & \vdots \\ c_{m,1} & \cdots & c_{m,p} \end{bmatrix},$$

where

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}.$$

If we identify vectors in $\mathbb{R}^n$ with matrices in $\mathbb{R}^{n\times 1}$, then the product of an $m \times n$ real matrix with a vector in $\mathbb{R}^n$ is a vector in $\mathbb{R}^m$:

$$\begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix},$$

where $b_i = a_{i,1}x_1 + \cdots + a_{i,n}x_n$. In this way, any $m \times n$ real matrix acts as a mapping $\mathsf{A}\colon \mathbb{R}^n \to \mathbb{R}^m$. It is easy to check that this mapping is a **linear operator** or **linear transformation**, that is, that it satisfies the following properties: For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any $c \in \mathbb{R}$,

1. $\mathsf{A}(\mathbf{x} + \mathbf{y}) = \mathsf{A}\mathbf{x} + \mathsf{A}\mathbf{y}$ (**additivity**).
2. $\mathsf{A}(c\mathbf{x}) = c(\mathsf{A}\mathbf{x})$ (**homogeneity**).

In this context, the **identity matrix** in $\mathbb{R}^{n\times n}$ plays a special role. This matrix has the form

$$\mathsf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

It is easy to verify that $\mathsf{IA} = \mathsf{A}$ for every matrix $\mathsf{A} \in \mathbb{R}^{n\times m}$ and that $\mathsf{AI} = \mathsf{A}$ for every matrix $\mathsf{A} \in \mathbb{R}^{m\times n}$.

Numerical analysis frequently yields problems having the following form: given a matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{b} \in \mathbb{R}^n$, find a vector $\mathbf{x} \in \mathbb{R}^n$ such that $\mathsf{A}\mathbf{x} = \mathbf{b}$.

DEFINITION. The matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ is **nonsingular** if, for any $\mathbf{b} \in \mathbb{R}^n$, there exists a unique vector $\mathbf{x} \in \mathbb{R}^n$ such that $\mathsf{A}\mathbf{x} = \mathbf{b}$. Otherwise, A is **singular**.

If A is singular, then the equation $\mathsf{A}\mathbf{x} = \mathbf{b}$ may have no solutions $\mathbf{x}$, or solutions may exist but not be unique. There are several equivalent characterizations of these notions. In the next theorem, $\det \mathsf{A}$ denotes the determinant of the matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$. Strang [48, Chapter 4] reviews the definition of this quantity.

THEOREM 1.3.1 (CONDITIONS FOR NONSINGULARITY). *If $\mathsf{A} \in \mathbb{R}^{n\times n}$, then the following statements are equivalent:*

1. A *is nonsingular.*
2. $\det \mathsf{A} \neq 0$.
3. *If* $\mathsf{A}\mathbf{x} = \mathbf{0}$*, then* $\mathbf{x} = \mathbf{0}$.
4. *The columns of* A *are linearly independent.*
5. *There is a unique matrix* $\mathsf{A}^{-1} \in \mathbb{R}^{n\times n}$ *such that* $\mathsf{A}\mathsf{A}^{-1} = \mathsf{A}^{-1}\mathsf{A} = \mathsf{I}$.

For proof of the theorem, see Ref. [48, Chapter 2]. We often rephrase condition 3 by saying that the **null space**

$$\mathcal{N}(\mathsf{A}) := \left\{ \mathbf{x} \in \mathbb{R}^n \;\middle|\; \mathsf{A}\mathbf{x} = \mathbf{0} \right\}$$

is $\{\mathbf{0}\}$. Problem 1.3 asks for proof that $\mathcal{N}(\mathsf{A})$ is a subspace of $\mathbb{R}^n$. The matrix A^{-1} in part 4 is the **inverse** of A, and its existence means that A is **invertible**.

Suppose that $\mathsf{A} \in \mathbb{R}^{m\times n}$, and denote its (i, j)th entry by $a_{i,j}$. The **transpose** of A, denoted $\mathsf{A}^\top$, is the matrix in $\mathbb{R}^{n\times m}$ whose entry in the (i, j)th position is $a_{j,i}$. A matrix A is **symmetric** when $\mathsf{A}^\top = \mathsf{A}$. This equation guarantees that A is square and that $a_{i,j} = a_{j,i}$. The transpose of a column vector $\mathbf{v} \in \mathbb{R}^m$ is a row vector,

$$\mathbf{v}^\top = (v_1, v_2, \ldots, v_m),$$

which we also say is in $\mathbb{R}^m$. Problem 1.2 asks for proof that $(\mathsf{A}\mathsf{B})^\top = \mathsf{B}^\top\mathsf{A}^\top$.

1.3.3 Norms

In analyzing errors associated with numerical approximations, we often estimate the lengths of vectors or the distances between pairs of vectors. The following concept captures the notion of length in settings even more general than $\mathbb{R}^n$.

DEFINITION. A **norm** on a vector space $\mathcal{V}$ is a function $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ that satisfies the following conditions for any $x, y \in \mathcal{V}$ and $a \in \mathbb{R}$:

1. $\|x\| \geqslant 0$, and $\|x\| = 0$ if and only if $x = 0$ (**positive definiteness**).
2. $\|ax\| = |a|\,\|x\|$ (**homogeneity**).
3. $\|x + y\| \leqslant \|x\| + \|y\|$ (**subadditivity**).

If such a function exists, then $\mathcal{V}$ is a **normed vector space**.

The third condition is the **triangle inequality**, which we use throughout this book. From the version in condition 3 there follows an alternative version.

THEOREM 1.3.2 (ALTERNATIVE TRIANGLE INEQUALITY). *If $\|\cdot\|$ is a norm on a vector space $\mathcal{V}$, then, for any $x, y \in \mathcal{V}$,*

$$|\,\|x\| - \|y\|\,| \leqslant \|x - y\|. \tag{1.2}$$

PROOF: See Problem 1.5. ∎

The prototypical norm is the absolute value function $|\cdot|\colon \mathbb{R} \to \mathbb{R}$. This familiar function has many extensions to $\mathbb{R}^n$, three of which are defined for $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ as follows:

$$\begin{aligned}
\|\mathbf{x}\|_1 &:= |x_1| + |x_2| + \cdots + |x_n|,\\
\|\mathbf{x}\|_2 &:= \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2},\\
\|\mathbf{x}\|_\infty &:= \max_{1 \leqslant i \leqslant n} |x_i|.
\end{aligned}$$

By using properties of $|\cdot|$, one easily verifies (see Problem 1.6) that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ satisfy the conditions to be norms. The function $\|\cdot\|_2$ is just the Euclidean length introduced earlier, and for this function, the first two properties of norms follow from corresponding facts for $|\cdot|$. We review below an argument establishing the triangle inequality for $\|\cdot\|_2$.

Analogous norms exist for function spaces. Consider $C^k([a, b])$, the vector space of all real-valued functions defined on the bounded, closed interval $[a, b] \subset \mathbb{R}$ whose derivatives through order k are continuous. (In this context we always assume $a \neq b$.) For $f \in C^k([a, b])$,

$$\begin{aligned}
\|f\|_1 &:= \int_a^b |f(x)|\,dx,\\
\|f\|_2 &:= \left[\int_a^b |f(x)|^2\,dx\right]^{1/2},\\
\|f\|_\infty &:= \sup_{x \in [a,b]} |f(x)|.
\end{aligned}$$

It is relatively straightforward to show that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ satisfy the properties required to be a norm. For $\|\cdot\|_2$, proving the triangle inequality requires slightly more work, which we undertake shortly.

It is also possible to construct norms for vector spaces of matrices. We explore this idea later in this chapter.

An interpretation in terms of length is natural for the norm $\|\cdot\|_2$ on $\mathbb{R}^n$, which is just the Euclidean length function. For the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $\mathbb{R}^n$ the

interpretation may be slightly less familiar. Figure 1.5 illustrates the **unit spheres**

$$\begin{aligned} S_1 &= \Big\{ \mathbf{x} \in \mathbb{R}^2 \;\Big|\; \|\mathbf{x}\|_1 = 1 \Big\}, \\ S_2 &= \Big\{ \mathbf{x} \in \mathbb{R}^2 \;\Big|\; \|\mathbf{x}\|_2 = 1 \Big\}, \\ S_\infty &= \Big\{ \mathbf{x} \in \mathbb{R}^2 \;\Big|\; \|\mathbf{x}\|_\infty = 1 \Big\}, \end{aligned} \tag{1.3}$$

in $\mathbb{R}^2$. Each unit sphere consists of all those vectors whose length, measured in the appropriate norm, is 1.

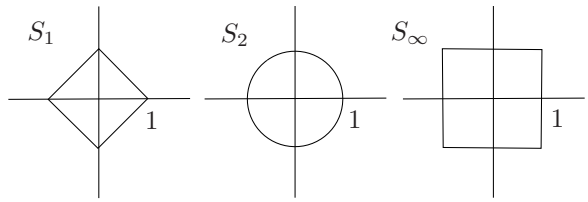

Figure 1.5 The unit spheres S_1, S_2, and S_∞ in $\mathbb{R}^2$.

In the function spaces $C^k([a,b])$, a norm typically assigns to a given function f some quantity whose interpretation as a length is more abstract. For example, $\|f\|_1$ is the average value of $|f|$ over $[a,b]$, multiplied by the length $|b-a|$ of the interval. Similarly, $\|f\|_2$ essentially gives the root-mean-square average of f over $[a,b]$, again multiplied by $|b-a|$. Finally, $\|f\|_\infty$ measures the largest excursion that f takes from the x-axis, as Figure 1.6 illustrates.

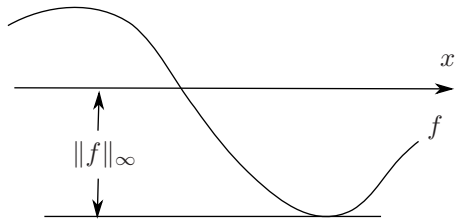

Figure 1.6 Geometric interpretation of $\|f\|_\infty$ as a measure of the largest excursion that f takes from the x-axis.

Viewing the length of a vector as its distance from 0 leads to another geometric idea: The distance between two vectors x and y in a normed vector space is the norm of their difference, $\|y - x\|$. Figure 1.7 illustrates this idea for two vectors using the Euclidean length $\|\cdot\|_2$ in $\mathbb{R}^2$, where the interpretation corresponds to familiar concepts in plane geometry. By abstracting this geometric notion to other norms and to vector spaces other than $\mathbb{R}^n$, we establish a useful means of measuring, for

example, how close an approximation – whether to an n-tuple of numbers or to a function – lies to an exact answer.

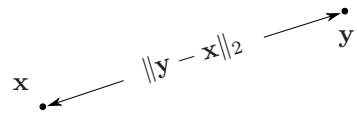

Figure 1.7 The distance $\|\mathbf{y}-\mathbf{x}\|_2$ between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$.

1.3.4 Inner Products

In many vector spaces of interest in numerical analysis there is yet another level of geometric structure.

DEFINITION. If $\mathcal{V}$ is a vector space, a function $\langle \cdot, \cdot \rangle \colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is an **inner product** on $\mathcal{V}$ if, for all $x, y, z \in \mathcal{V}$,

1. $\langle x, x \rangle \geqslant 0$, *and* $\langle x, x \rangle = 0$ *only if* $x = 0$ (**positive definiteness**).
2. $\langle x, y \rangle = \langle y, x \rangle$ (**symmetry**).
3. $\langle x, ay + bz \rangle = a\langle x, y \rangle + b\langle x, z \rangle$ *for any* $a, b \in \mathbb{R}$ (**linearity**).

If such a function exists, then $\mathcal{V}$ is an **inner-product space**.

The ordinary dot product on $\mathbb{R}^n$ is an inner product: If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then

$$\langle \mathbf{x}, \mathbf{y} \rangle := \mathbf{x} \cdot \mathbf{y} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} x_j y_j.$$

The notation for matrix transposes allows us to write the dot product $\mathbf{u} \cdot \mathbf{v}$ as $\mathbf{u}^\top \mathbf{v}$, using the rules for array multiplication.

Each of the function spaces $C^k([a, b])$ also possesses an inner product, defined for two functions f, g as follows:

$$\langle f, g \rangle := \int_a^b f(x) g(x)\, dx.$$

The extra geometry associated with inner-product spaces stems from the following concept.

DEFINITION. Two vectors $x, y \in \mathcal{V}$ are **orthogonal** if $\langle x, y\rangle = 0$.

When $\mathcal{V} = \mathbb{R}^n$ and $\langle \cdot, \cdot \rangle$ is the ordinary dot product, this definition of orthogonality coincides with the usual notion of perpendicularity. In the function spaces $C^k([a, b])$ and in most other examples of inner-product spaces, the picture is more abstract, but the geometric analogy remains just as profitable.

Any inner-product space is a normed vector space, the natural norm being defined (and denoted) by analogy with the Euclidean length:

$$\|x\|_2 := \sqrt{\langle x, x\rangle}.$$

This definition includes the norms $\|\cdot\|_2$ defined on the vector spaces $\mathbb{R}^n$ and $C^k([a, b])$. To show that $\| \cdot \|_2$ indeed defines a norm, we must establish the triangle inequality. The argument hinges on the following fact.

THEOREM 1.3.3 (CAUCHY–SCHWARZ INEQUALITY). *If $\mathcal{V}$ is an inner-product space with inner product $\langle \cdot, \cdot \rangle$, then, for any $x, y \in \mathcal{V}$,*

$$\big|\langle x, y\rangle\big| \leqslant \|x\|_2 \|y\|_2. \tag{1.4}$$

PROOF: If $y = 0$, then both sides of the inequality (1.4) vanish, and the theorem is true trivially. Assume that $y \neq 0$. In this case, for any $r \in \mathbb{R}$, positive definiteness of the inner product implies that

$$0 \leqslant \langle x + ry, x + ry\rangle = \|x\|_2^2 + 2r\langle x, y\rangle + r^2\|y\|_2^2.$$

The expression on the right is quadratic in r, and the fact that it is nonnegative implies that the discriminant $4\langle x, y\rangle^2 - 4\|x\|_2^2\|y\|_2^2 \leqslant 0$. The inequality (1.4) follows. ∎

COROLLARY 1.3.4 (TRIANGLE INEQUALITY FOR $\|\cdot\|_2$). *If $\mathcal{V}$ is an inner-product space, then*

$$\|x + y\|_2 \leqslant \|x\|_2 + \|y\|_2$$

for every $x, y \in \mathcal{V}$.

PROOF: Observe that

$$(\|x\|_2 + \|y\|_2)^2 = \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2.$$

The Cauchy–Schwarz inequality guarantees that the middle term on the right side of this identity is at least as large as $\langle x, y\rangle$, so

$$(\|x\|_2 + \|y\|_2)^2 \geqslant \|x\|_2^2 + 2\langle x, y\rangle + \|y\|_2^2 = \|x + y\|_2^2.$$

Taking square roots completes the argument. ∎

The connection between norms and orthogonality in inner-product spaces allows us to specify a particularly useful type of basis.

DEFINITION. A basis S for an inner-product space $\mathcal{V}$ is an **orthonormal basis** if the following conditions hold:

1. Whenever $x, y \in S$, and $x \neq y$, $\langle x, y \rangle = 0$.
2. For every $x \in S$, $\|x\|_2 = 1$.

When $\mathcal{V}$ is a finite-dimensional inner-product space, one can always construct an orthonormal basis from an arbitrary basis for $\mathcal{V}$ using an algorithm known as the Gram–Schmidt procedure. See Algorithm 6.3.2 for details.

1.3.5 Norm Equivalence

While one can define infinitely many norms on $\mathbb{R}^n$, they impose essentially the same structures, in a sense defined below. We devote the rest of this section to a discussion of this remarkable fact, which does not hold for normed vector spaces in general. We begin with the following general property of norms.

THEOREM 1.3.5 (UNIFORM CONTINUITY OF NORMS) *Let $\mathcal{V}$ be a normed vector space over $\mathbb{R}$.*

1. *Any norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ is uniformly continuous.*
2. *In the special case $\mathcal{V} = \mathbb{R}^n$, any norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is uniformly continuous with respect to the Euclidean norm $\|\cdot\|_2$.*

PROOF: To prove 1 we must show that, for any $\epsilon > 0$, there exists a number $\delta > 0$ such that, whenever the vectors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ satisfy $\|\mathbf{x} - \mathbf{y}\| < \delta$, $|\, \|\mathbf{x}\| - \|\mathbf{y}\| \,| < \epsilon$. By the version (1.2) of the triangle inequality, we choose $\delta = \epsilon$.

To establish 2, let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ be the standard basis for $\mathbb{R}^n$, and suppose that $\epsilon > 0$. For $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$ and $\mathbf{y} = y_1\mathbf{e}_1 + y_2\mathbf{e}_2 + \cdots + y_n\mathbf{e}_n$, we

have

$$\begin{aligned}\big|\,\|\mathbf{x}\|-\|\mathbf{y}\|\,\big| \;\leqslant\; \|\mathbf{x}-\mathbf{y}\| &= \left\|\sum_{j=1}^{n}(x_j-y_j)\mathbf{e}_j\right\| \\ &\leqslant \sum_{j=1}^{n}|x_j-y_j|\,\|\mathbf{e}_j\| \\ &\leqslant \left(\sum_{i=1}^{n}|x_i-y_i|^2\right)^{1/2}\underbrace{\left(\sum_{i=1}^{n}\|\mathbf{e}_i\|^2\right)^{1/2}}_{M} = M\,\|\mathbf{x}-\mathbf{y}\|_2,\end{aligned}$$

the number M being independent of $\mathbf{x}$ and $\mathbf{y}$. The third inequality in this chain follows from the Cauchy–Schwarz inequality. Choosing $\delta = \epsilon/M$ guarantees that $\big|\,\|\mathbf{x}\|-\|\mathbf{y}\|\,\big| < \epsilon$ whenever $\|\mathbf{x}-\mathbf{y}\|_2 < \delta$. ∎

The crucial question for norm equivalence is whether inequalities derived using one norm $\|\cdot\|_{\mathrm{I}}$ can be converted to analogous inequalities expressed in a different norm $\|\cdot\|_{\mathrm{II}}$.

DEFINITION. Let $\|\cdot\|_{\mathrm{I}}$ and $\|\cdot\|_{\mathrm{II}}$ be norms on a vector space $\mathcal{V}$. Then $\|\cdot\|_{\mathrm{I}}$ and $\|\cdot\|_{\mathrm{II}}$ are **equivalent** if there exist constants $m, M > 0$ such that

$$m\|\mathbf{x}\|_{\mathrm{I}} \leqslant \|\mathbf{x}\|_{\mathrm{II}} \leqslant M\|\mathbf{x}\|_{\mathrm{I}} \tag{1.5}$$

for all $\mathbf{x} \in \mathcal{V}$. If this relationship holds, then we write $\|\cdot\|_{\mathrm{I}} \approx \|\cdot\|_{\mathrm{II}}$.

THEOREM 1.3.6 (NORM EQUIVALENCE AS AN EQUIVALENCE RELATION). *The relation $\approx$ of norm equivalence is an* **equivalence relation***, that is,*

1. *The relation is* **reflexive***:* $\|\cdot\| \approx \|\cdot\|$.
2. *The relation is* **symmetric***:* $\|\cdot\|_{\mathrm{I}} \approx \|\cdot\|_{\mathrm{II}}$ *implies* $\|\cdot\|_{\mathrm{II}} \approx \|\cdot\|_{\mathrm{I}}$.
3. *The relation is* **transitive***: If* $\|\cdot\|_{\mathrm{I}} \approx \|\cdot\|_{\mathrm{II}}$ *and* $\|\cdot\|_{\mathrm{II}} \approx \|\cdot\|_{\mathrm{III}}$*, then* $\|\cdot\|_{\mathrm{I}} \approx \|\cdot\|_{\mathrm{III}}$.

PROOF: This is Problem 1.7. ∎

Symmetry implies that one can reverse the roles of the two norms in the inequalities (1.5), possibly using different values for the constants m and M.

THEOREM 1.3.7 (NORM EQUIVALENCE IN $\mathbb{R}^n$). *All norms on $\mathbb{R}^n$ are equivalent.*

PROOF: It suffices to show that any norm on $\mathbb{R}^n$ is equivalent to $\|\cdot\|_2$ by finding appropriate constants m and M, as stipulated in (1.5). Let $\|\cdot\|$ be such a norm. The

unit sphere S_2 defined in Eq. (1.3) is compact, that is, it is closed and bounded in $\mathbb{R}^n$. Moreover, the function $\|\cdot\|$ is continuous with respect to $\|\cdot\|_2$ on S_2 by part 2 of Theorem 1.3.5. From these two facts and Theorem 1.2.1 it follows that $\|\cdot\|$ attains maximum and minimum values at some points $\mathbf{x}_{\max}$ and $\mathbf{x}_{\min}$, respectively, on S_2. This means that, for any $\mathbf{x} \in S_2$,

$$\|\mathbf{x}_{\min}\| \leqslant \|\mathbf{x}\| \leqslant \|\mathbf{x}_{\max}\|.$$

We claim that we can choose $m = \|\mathbf{x}_{\min}\|$ and $M = \|\mathbf{x}_{\max}\|$.

First, since $\mathbf{x}_{\min} \in S_2$, $\|\mathbf{x}_{\min}\| > 0$. Next, select an arbitrary vector $\mathbf{x} \in \mathbb{R}^n$. If $\mathbf{x} = \mathbf{0}$, then the claim is trivially true. Otherwise, $\mathbf{x}/\|\mathbf{x}\|_2 \in S_2$, which implies that

$$\|\mathbf{x}_{\min}\| \leqslant \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\| \leqslant \|\mathbf{x}_{\max}\|.$$

Multiplying these inequalities through by $\|\mathbf{x}\|_2$ establishes the claim and hence the theorem. ∎

1.4 Eigenvalues and Matrix Norms

This section develops tools from linear algebra required to analyze direct and iterative methods for linear systems, discussed in Chapters 3 and 5. Two concepts – eigenvalues and matrix norms – enable us to measure relationships between the sizes of vectors and the sizes of their images under matrix multiplication.

1.4.1 Eigenvalues and Eigenvectors

DEFINITION. A number $\lambda \in \mathbb{C}$ is an **eigenvalue** of $\mathsf{A} \in \mathbb{R}^{n\times n}$ if there is a nonzero vector $\boldsymbol{v} \in \mathbb{C}^n$ for which $\mathsf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$. Any such vector $\boldsymbol{v}$ is an **eigenvector** of A associated with λ. The collection σ_A of all eigenvalues of A is the **spectrum** of A, and the number

$$\varrho(\mathsf{A}) = \max_{\lambda\in\sigma_A} |\lambda|$$

is the **spectral radius** of A.

Eigenvalues are the (possibly complex-valued) factors by which A stretches the associated eigenvectors. The requirement $\mathsf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$ with $\boldsymbol{v} \neq \mathbf{0}$ implies that the matrix $\lambda\mathsf{I} - \mathsf{A}$ is singular, and hence any eigenvalue λ of A is a zero of the **characteristic polynomial** $\det(\lambda\mathsf{I} - \mathsf{A})$, which has degree n in λ. Chapter 6 discusses numerical methods for computing eigenvalues and eigenvectors of matrices.

DEFINITION. A matrix $\mathsf{A} \in \mathbb{R}^{n \times n}$ is **nonnegative** if $\mathbf{x} \cdot (\mathsf{A}\mathbf{x}) = \mathbf{x}^{\top}\mathsf{A}\mathbf{x} \geqslant 0$ for every $\mathbf{x} \in \mathbb{R}^n$. A is **positive definite** if $\mathbf{x}^{\top}\mathsf{A}\mathbf{x} > 0$ for every nonzero $\mathbf{x} \in \mathbb{R}^n$.

The following theorem summarizes important properties of eigenvalues and eigenvectors. For proofs, see Ref. [48, Chapter 5].

THEOREM 1.4.1 (PROPERTIES OF EIGENVALUES AND EIGENVECTORS). *Let* $\mathsf{A} \in \mathbb{R}^{n \times n}$. *Then*

1. A *is singular if and only if* 0 *is an eigenvalue of* A.
2. *If* A *is upper or lower triangular, then its eigenvalues are its diagonal entries.*
3. *If* A *is symmetric, then all of its eigenvalues are real numbers.*
4. *If* A *is symmetric and nonnegative, then all eigenvalues of* A *are nonnegative.*
5. *If* A *is symmetric and positive definite, then all of its eigenvalues are positive.*
6. *If* A *is symmetric, then there exists an orthonormal basis for* $\mathbb{R}^n$*, each of whose elements is an eigenvector of* A.

The sixth assertion means that there exists a set $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ of eigenvectors of A such that:

1. Each eigenvector $\boldsymbol{v}_i$ has unit Euclidean length $\|\boldsymbol{v}_i\|_2 = \langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle^{1/2} = 1$.
2. Distinct eigenvectors $\boldsymbol{v}_i, \boldsymbol{v}_j$ in the set are orthogonal, that is, $\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = 0$ when $i \neq j$.
3. Any vector $\mathbf{y} \in \mathbb{R}^n$ has an expansion

$$\mathbf{y} = \sum_{j=1}^{n} c_j \boldsymbol{v}_j, \tag{1.6}$$

   for some real coefficients $c_1, c_2, \ldots, c_n$.

Problem 1.13 shows that the coefficients in such an expansion are $c_j = \boldsymbol{v}_j^{\top}\mathbf{y}$. Moreover,

$$\|\mathbf{y}\|_2^2 = \langle \mathbf{y}, \mathbf{y} \rangle = \sum_{j=1}^{n} c_j^2. \tag{1.7}$$

The last identity generalizes the Pythagorean theorem.

The following real-valued function plays an important role in the theory of eigenvalues and eigenvectors.

DEFINITION. The **Rayleigh quotient** for a matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$ is defined for nonzero vectors $\mathbf{x} \in \mathbb{R}^n$ as follows:

$$R_A(\mathbf{x}) := \frac{\langle \mathbf{x}, \mathsf{A}\mathbf{x}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle} = \frac{\mathbf{x}\cdot \mathsf{A}\mathbf{x}}{\|\mathbf{x}\|_2^2}. \tag{1.8}$$

Among the many important consequences of Theorem 1.4.1 is the following characterization of the Rayleigh quotient.

THEOREM 1.4.2 (EXTREMA OF THE RAYLEIGH QUOTIENT). *If $\mathsf{A} \in \mathbb{R}^{n\times n}$ is symmetric with eigenvalues $\lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_n$, then the Rayleigh quotient (1.8) has minimum value λ_1 and maximum value λ_n.*

PROOF: Any vector $\mathbf{x} \in \mathbb{R}^n$ has an expansion in the orthonormal basis of eigenvectors $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ guaranteed by Theorem 1.4.1:

$$\mathbf{x} = \sum_{j=1}^{n} x_j \boldsymbol{v}_j.$$

The orthonormality of the eigenvectors implies that

$$\mathbf{x}\cdot \mathsf{A}\mathbf{x} = \sum_{j=1}^{n} x_j \boldsymbol{v}_j \cdot \sum_{k=1}^{n} x_k \mathsf{A}\boldsymbol{v}_k = \sum_{j=1}^{n} x_j^2 \lambda_j.$$

Therefore,

$$R_A(\mathbf{x}) = \frac{\sum_{j=1}^{n} x_j^2 \lambda_j}{\sum_{j=1}^{n} x_j^2}.$$

The expression on the right represents a weighted average of the eigenvalues λ_j. This average takes its minimum value λ_1 when $x_2 = x_3 = \cdots = x_n = 0$ and its maximum value λ_n when $x_1 = x_2 = \cdots = x_{n-1} = 0$. ∎

1.4.2 Matrix Norms

In some cases, the spectrum of a matrix A yields scanty information about the relationship between the size of $\mathbf{x}$ and the size of $\mathsf{A}\mathbf{x}$. For example, consider the matrix

$$\mathsf{A}_1 = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}.$$

This matrix has characteristic polynomial $\det(\lambda \mathsf{I} - \mathsf{A}_1) = \lambda^2$, which has a double root $\lambda = 0$. Therefore the spectrum of A_1 is $\{0\}$, and $\varrho(\mathsf{A}_1) = 0$. However, for any

vector $\mathbf{x} = (0, x_2)^\top \in \mathbb{R}^2$, the image vector $\mathsf{A}_1\mathbf{x} = (2x_2, 0)^\top$ has Euclidean length twice that of $\mathbf{x}$. In this case, eigenvalues reveal very little about how multiplication by the matrix changes the size of an arbitrary vector.

An extension of the concept of norms allows us to gauge the size of $\mathsf{A}\mathbf{x} \in \mathbb{R}^n$ in terms of the size of $\mathbf{x} \in \mathbb{R}^n$, for any matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$. The following definition captures the idea.

DEFINITION. If $\mathsf{A} \in \mathbb{R}^{n\times n}$ and $\|\cdot\|\colon \mathbb{R}^n \to \mathbb{R}$ is a norm, then the **subordinate matrix norm** $\|\cdot\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ is defined as follows:

$$\|\mathsf{A}\| := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathsf{A}\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{1.9}$$

As an immediate consequence of this definition,

$$\|\mathsf{A}\mathbf{x}\| \leqslant \|\mathsf{A}\|\,\|\mathbf{x}\|, \tag{1.10}$$

for any vector $\mathbf{x} \in \mathbb{R}^n$. Problem 1.10 asks for proof that that for any matrices $\mathsf{A}, \mathsf{B} \in \mathbb{R}^{n\times n}$ for which AB makes sense,

$$\|\mathsf{AB}\| \leqslant \|\mathsf{A}\|\,\|\mathsf{B}\|. \tag{1.11}$$

Also, the following three formulas for $\|\mathsf{A}\|$ are equivalent to Eq. (1.9):

$$\begin{aligned}
\|\mathsf{A}\| &= \sup_{\|\mathbf{x}\|=1} \|\mathsf{A}\mathbf{x}\|,\\
\|\mathsf{A}\| &= \inf\Big\{M \geqslant 0 \;\Big|\; \|\mathsf{A}\mathbf{x}\| \leqslant M\|\mathbf{x}\| \text{ for all } \mathbf{x} \in \mathbb{R}^n\Big\},\\
\|\mathsf{A}\| &= \inf\Big\{M \geqslant 0 \;\Big|\; \|\mathsf{A}\mathbf{x}\| \leqslant M \text{ for all } \mathbf{x} \in \mathbb{R}^n \text{ with } \|\mathbf{x}\| = 1\Big\}.
\end{aligned} \tag{1.12}$$

Therefore, if $\|\mathsf{A}\mathbf{x}\| \leqslant M\|\mathbf{x}\|$ for all $\mathbf{x}$, then $\|\mathsf{A}\| \leqslant M$. On the other hand, if $\|\mathsf{A}\mathbf{x}\| \geqslant M\|\mathbf{x}\|$ for some $\mathbf{x} \neq \mathbf{0}$, then $\|\mathsf{A}\| \geqslant M$.

Problem 1.11 asks for verification that subordinate matrix norms satisfy the three conditions required to be a norm on the vector space $\mathbb{R}^{n\times n}$. (However, not every norm on $\mathbb{R}^{n\times n}$ is subordinate to a vector norm. Problem 1.16 examines this fact.) In particular, any subordinate matrix norm obeys the triangle inequality. Here lies a crucial defect in the spectral radius as a measure of size: If $n > 1$, it is possible to find matrices $\mathsf{A}, \mathsf{B} \in \mathbb{R}^{n\times n}$ for which $\varrho(\mathsf{A}+\mathsf{B}) > \varrho(\mathsf{A}) + \varrho(\mathsf{B})$, and consequently the triangle inequality fails. Problem 1.15 asks for details.

While matrix norms typically give better characterizations than the spectrum of the stretching power of a matrix, one can derive a simple lower bound for $\|\mathsf{A}\|$ if one

knows an eigenvalue λ of A. Since $\mathsf{A}\mathbf{x} = \lambda\mathbf{x}$, $\|\mathsf{A}\mathbf{x}\| = |\lambda|\,\|\mathrm{x}\|$. From the inequality (1.10) it follows that $\|\mathsf{A}\| \geqslant |\lambda|$ and hence that

$$\|\mathsf{A}\| \geqslant \varrho(\mathsf{A}). \tag{1.13}$$

Each of the vector norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ gives rise to a useful subordinate matrix norm. Shortly we prove the following characterizations:

$$\|\mathsf{A}\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{i,j}|, \text{ the maximum row sum of A.}$$

$$\|\mathsf{A}\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|, \text{ the maximum column sum of A.}$$

$$\|\mathsf{A}\|_2 = \sqrt{\varrho(\mathsf{A}^\top\mathsf{A})}.$$

When A is symmetric, one can calculate $\|\mathsf{A}\|_2$ more simply. Symmetry implies that $\mathsf{A}^\top\mathsf{A} = \mathsf{A}^2$. But the eigenvalues of A^2 are the squares of the eigenvalues of A (see Problem 1.14). Therefore, when A is symmetric,

$$\|\mathsf{A}\|_2 = \sqrt{\varrho(\mathsf{A}^2)} = \varrho(\mathsf{A}). \tag{1.14}$$

Simple examples illustrate these norms. Consider the matrices

$$\mathsf{A}_1 = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}, \quad \mathsf{A}_1^\top\mathsf{A}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}.$$

The eigenvalues of $\mathsf{A}_1^\top\mathsf{A}_1$ are 0 and 4, so $\|\mathsf{A}_1\|_2 = \sqrt{4} = 2$. Checking column and row sums, we find that $\|\mathsf{A}_1\|_1 = \|\mathsf{A}\|_2 = \|\mathsf{A}_1\|_\infty = 2$. However, both eigenvalues of A_1 are 0, so none of these norms equals $\varrho(\mathsf{A})$.

Next consider

$$\mathsf{A}_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \mathsf{A}_2^\top\mathsf{A}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this case, $\mathsf{A}_2^\top\mathsf{A}_2$ is the identity matrix I, both of whose eigenvalues are 1. Therefore

$$\|\mathsf{A}_2\|_2 = \sqrt{\varrho(\mathsf{A}_2^\top\mathsf{A}_2)} = 1 = \|\mathsf{A}_2\|_1 = \|\mathsf{A}_2\|_\infty.$$

The eigenvalues of A_2 are the purely imaginary numbers $\pm i$, where $i^2 = -1$, so in this case all three matrix norms equal $\varrho(\mathsf{A})$. In the geometric view, multiplying $\mathbf{x}$

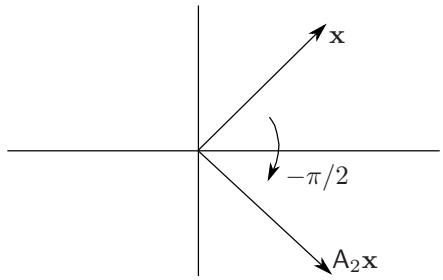

Figure 1.8 Geometric action of the matrix A_2.

on the left by A_2 rotates $\mathbf{x}$ about the origin by $-\pi/2$ radians without changing its Euclidean length, as shown in Figure 1.8.

Finally, consider the 3×3 matrix

$$A_3 = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

In this case A_3 is symmetric, so $\|A_3\|_2$ is just the largest value of $|\lambda|$, where λ ranges over the eigenvalues of A_3. Solving the cubic equation $\det(\lambda I - A_3) = 0$, we find that the eigenvalues of A_3 are 0, 1, and 3. Therefore, $\|A_3\|_2 = 3$. However, in this case $\|A_3\|_1 = \|A_3\|_\infty = 4$.

We now prove the characterizations of the matrix norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ stated earlier.

THEOREM 1.4.3 (CHARACTERIZATIONS OF MATRIX NORMS). *Let* $A \in \mathbb{R}^{n\times n}$. *Then*

1. $\|A\|_\infty = \max\limits_{1\leqslant i\leqslant n} \sum\limits_{j=1}^{n} |a_{i,j}|.$
2. $\|A\|_1 = \max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{n} |a_{i,j}|.$
3. $\|A\|_2 = \sqrt{\varrho(A^\top A)}.$

PROOF: To prove 1, let $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, and call

$$N := \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{i,j}|.$$

The definition of the vector norm $\|\cdot\|_\infty$ and the triangle inequality imply that

$$\|\mathsf{A}\mathbf{x}\|_\infty = \max_{1\leqslant i\leqslant n}\left|\sum_{j=1}^n a_{i,j}x_j\right| \leqslant \max_{1\leqslant i\leqslant n}\sum_{j=1}^n |a_{i,j}|\ \max_{1\leqslant j\leqslant n}|x_j| = N\|\mathbf{x}\|_\infty.$$

Therefore $\|\mathsf{A}\|_\infty \leqslant N$. It now suffices to show that $\|\mathsf{A}\| \geqslant N$, which we do by showing that $\|\mathsf{A}\mathbf{x}\|_\infty$ actually attains the value N for some unit vector $\mathbf{x}$. If $\mathsf{A} = \mathsf{0}$, the result is clear, so assume that $\mathsf{A} \neq \mathsf{0}$. Choose i so that

$$\sum_{j=1}^n |a_{i,j}| = N,$$

and define $\mathbf{x}$ by

$$x_j := \begin{cases} a_{i,j}/|a_{i,j}|, & \text{if} \quad a_{i,j} \neq 0, \\ 0, & \text{if} \quad a_{i,j} = 0. \end{cases}$$

It is now straightforward to check that $\|\mathbf{x}\|_\infty = 1$ and $\|\mathsf{A}\mathbf{x}\|_\infty = N$.

The characterization 2 is Problem 1.17, the argument being similar in spirit to the one just given.

To prove 3, choose $\mathbf{y} \in \mathbb{R}^n$ such that $\|\mathbf{y}\|_2 = 1$ and $\|\mathsf{A}\mathbf{y}\|_2 = \|\mathsf{A}\|_2$. (This is possible because $\|\cdot\|_2$ is a continuous function on the compact set S_2 defined in Eq. (1.3); hence $\|\cdot\|_2$ attains a maximum value at some point $\mathbf{x} \in S_2$.) Since $(\mathsf{A}\mathbf{y})^\top = \mathbf{y}^\top\mathsf{A}^\top$,

$$\|\mathsf{A}\|_2^2 = \|\mathsf{A}\mathbf{y}\|_2^2 = (\mathsf{A}\mathbf{y})^\top(\mathsf{A}\mathbf{y}) = \mathbf{y}^\top\mathsf{A}^\top\mathsf{A}\mathbf{y}. \tag{1.15}$$

But $\mathsf{A}^\top\mathsf{A} \in \mathbb{R}^{n\times n}$ is symmetric, so Theorem 1.4.1 guarantees that there exists an orthonormal basis $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ for $\mathbb{R}^n$ consisting entirely of eigenvectors of $\mathsf{A}^\top\mathsf{A}$. Moreover, all of the corresponding eigenvalues are nonnegative, since $\mathsf{A}^\top\mathsf{A}\mathbf{x} = \lambda\mathbf{x}$ implies that

$$0 \leqslant \|\mathsf{A}\mathbf{x}\|_2^2 = (\mathsf{A}\mathbf{x})^\top(\mathsf{A}\mathbf{x}) = \mathbf{x}^\top\mathsf{A}^\top\mathsf{A}\mathbf{x} = \mathbf{x}^\top\lambda\mathbf{x} = \lambda\|\mathbf{x}\|_2^2.$$

Denote by λ_i the eigenvalue of $\mathsf{A}^\top\mathsf{A}$ associated with $\boldsymbol{v}_i$. If we substitute an expansion of the form (1.6) for the unit-length vector $\mathbf{y}$ into Eq. (1.15), we obtain

$$\begin{aligned} \|\mathsf{A}\|_2^2 &= \sum_{i=1}^n c_i\boldsymbol{v}_i^\top\mathsf{A}^\top\mathsf{A}\sum_{j=1}^n c_j\boldsymbol{v}_j \\ &= \sum_{i=1}^n c_i\boldsymbol{v}_i^\top\sum_{j=1}^n c_j\lambda_j\boldsymbol{v}_j \\ &= \sum_{j=1}^n \lambda_j c_j^2 \ \leqslant\ \varrho(\mathsf{A}^\top\mathsf{A})\sum_{j=1}^n c_j^2 \ =\ \varrho(\mathsf{A}^\top\mathsf{A}), \end{aligned}$$

the last step following from the fact that $\|\mathbf{y}\|_2^2 = 1$. Hence $\|\mathsf{A}\|_2^2 \leqslant \varrho(\mathsf{A}^\top\mathsf{A})$.

To finish the proof, we show that $\|\mathsf{A}\|_2^2 \geqslant \varrho(\mathsf{A}^\top\mathsf{A})$. Suppose that $\boldsymbol{v}_k$ is an eigenvector of $\mathsf{A}^\top\mathsf{A}$, chosen from the orthonormal basis, and that its associated eigenvalue $\lambda_k = \varrho(\mathsf{A}^\top\mathsf{A})$. The inequality (1.6) and the fact that $\boldsymbol{v}_k$ has unit length imply that

$$\|\mathsf{A}\|_2^2 = \|\mathsf{A}\|_2^2 \, \|\boldsymbol{v}_k\|_2^2 \geqslant \|\mathsf{A}\boldsymbol{v}_k\|_2^2 = \boldsymbol{v}_k^\top \mathsf{A}^\top \mathsf{A} \boldsymbol{v}_k = \lambda_k \, \boldsymbol{v}_k^\top \boldsymbol{v}_k = \varrho(\mathsf{A}^\top\mathsf{A}),$$

as claimed. ∎

The focus so far on the vector spaces $\mathbb{R}^{n\times n}$ may obscure the fact that one can define norms for the more general vector spaces $\mathbb{R}^{m\times n}$. Let $\mathsf{A} \in \mathbb{R}^{m\times n}$, and suppose that $\|\cdot\|_{\mathrm{I}}$ is a norm on $\mathbb{R}^m$ and $\|\cdot\|_{\mathrm{II}}$ is a norm on $\mathbb{R}^n$. Since the mapping $\mathbf{x} \mapsto \mathsf{A}\mathbf{x}$ sends vectors $\mathbf{x} \in \mathbb{R}^n$ to images $\mathsf{A}\mathbf{x} \in \mathbb{R}^m$, the definition of subordinate matrix norms extends as follows:

$$\|\mathsf{A}\|_{\mathrm{I,II}} := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathsf{A}\mathbf{x}\|_{\mathrm{I}}}{\|\mathbf{x}\|_{\mathrm{II}}}.$$

Much of the theory developed in this section translates in a straightforward manner to this more general setting. In particular,

$$\|\mathsf{A}\mathbf{x}\|_{\mathrm{II}} \leqslant \|\mathsf{A}\|_{\mathrm{I,II}} \, \|\mathbf{x}\|_{\mathrm{I}}.$$

Also, if $\mathsf{B} \in \mathbb{R}^{p\times m}$ and $\|\cdot\|_{\mathrm{III}}$ is a norm on $\mathbb{R}^p$, then $\mathsf{BA} \in \mathbb{R}^{p\times n}$, and

$$\|\mathsf{BA}\|_{\mathrm{III,II}} \leqslant \|\mathsf{B}\|_{\mathrm{III,I}} \, \|\mathsf{A}\|_{\mathrm{I,II}}. \tag{1.16}$$

See Problem 1.18.

1.5 Results from Calculus

1.5.1 Seven Theorems

We conclude this chapter with a review of basic results from calculus, leading to several versions of the Taylor theorem. We begin with seven theorems.

THEOREM 1.5.1 (INTERMEDIATE VALUE THEOREM). *Let $f \in C^0([a,b])$, and suppose that $f(a) < c < f(b)$. Then there exists a point $\zeta \in (a,b)$ such that $f(\zeta) = c$.*

See Ref. [46, Chapter 7].

THEOREM 1.5.2 (SEQUENTIAL CRITERION FOR CONTINUITY). *If $f \in C^0([a,b])$ and $\{x_m\}$ is a sequence such that every $x_m \in [a,b]$ and $x_m \to x^*$, then $f(x_m) \to f(x^*)$.*

See Ref. [46, Chapter 22].

THEOREM 1.5.3 (MONOTONICITY OF INTEGRATION). *If $f, g \in C^0([a,b])$ and $f(x) \leqslant g(x)$ for every $x \in [a,b]$, then*

$$\int_a^b f(x)\,dx \leqslant \int_a^b g(x)\,dx.$$

For a proof, see Ref. [46, Chapter 12]. This theorem has a remarkably useful corollary, which serves as a continuous analog of the triangle inequality:

THEOREM 1.5.4 (ABSOLUTE VALUE OF AN INTEGRAL). *If $f \in C^0([a,b])$, then*

$$\left|\int_a^b f(x)\,dx\right| \leqslant \int_a^b |f(x)|\,dx.$$

PROOF: Choose $s = \pm 1$ so that

$$s\int_a^b f(x)\,dx \geqslant 0.$$

This choice guarantees that $sf(x) \leqslant |f(x)|$ for all $x \in [a,b]$. It follows from Theorem 1.5.3 that

$$\left|\int_a^b f(x)\,dx\right| \leqslant s\int_a^b f(x)\,dx = \int_a^b sf(x)\,dx \leqslant \int_a^b |f(x)|\,dx. \qquad \blacksquare$$

THEOREM 1.5.5 (FUNDAMENTAL THEOREM OF CALCULUS). *If $f \in C^1([a,b])$ and $x \in [a,b]$, then*

$$f(x) = f(a) + \int_a^x f'(t)\,dt.$$

Again, Ref. [46, Chapter 14] gives a proof.

THEOREM 1.5.6 (INTEGRATION BY PARTS). *If $u, v \in C^1([a,b])$, then*

$$\int_a^b u(x)v'(x)\,dx = u(x)v(x)\Big|_a^b - \int_a^b u'(x)v(x)\,dx, \tag{1.17}$$

where

$$u(x)v(x)\Big|_a^b := u(b)v(b) - u(a)v(a).$$

PROOF: Eq. (1.17) follows directly from Theorem 1.5.5 and the product rule for differentiation. ∎

THEOREM 1.5.7 (MEAN VALUE THEOREM FOR INTEGRALS). *Let $f \in C^0([a,b])$, and suppose that g is integrable on $[a,b]$ and does not change signs there. Then there exists a number $\zeta \in [a,b]$ such that*

$$\int_a^b f(x)g(x)\,dx = f(\zeta)\int_a^b g(x)\,dx.$$

See Ref. [46, p. 277].

1.5.2 The Taylor Theorem

We now have the tools needed to prove one of the cornerstones of numerical analysis.

THEOREM 1.5.8 (TAYLOR THEOREM). *Let $f \in C^{n+1}([a,b])$ for some $n \geqslant 0$, and let $c, x \in [a,b]$ be distinct points. There is a point ζ, lying strictly between c and x, such that*

$$f(x) = \underbrace{\sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!}(x-c)^k}_{T_n(x,c)} + \underbrace{\frac{f^{(n+1)}(\zeta)}{(n+1)!}(x-c)^{n+1}}_{R_{n+1}(x,c)}. \tag{1.18}$$

Several comments are in order before the proof. The idea of the theorem is to approximate f near a point c, where we have information about the values of f and its first few derivatives. The **Taylor polynomial** $T_n(x,c)$ in Eq. (1.18) is a polynomial of degree at most n in the difference $x - c$, which we often regard as a small parameter. We view $T_n(x,c)$ as a polynomial approximation to $f(x)$ valid for x close to c, where we expect the **remainder** $R_{n+1}(x,c)$ to be small.

The success of this idea depends upon whether R_{n+1} is indeed small. One difficulty is the fact that ζ, while guaranteed to exist, remains unknown except for the stipulation that it lies between c and x. To circumvent this problem, observe that

$$|R_{n+1}(x,c)| \leqslant \underbrace{\sup_{y\in[a,b]} \left|f^{(n+1)}(y)\right|}_{M_{n+1}} |x-c|^{n+1},$$

the constant M_{n+1} being independent of ζ and hence of the choice of x. This estimate shows, heuristically, that R_{n+1} shrinks at least as fast as $(x-c)^{n+1}$. The latter magnitude of the latter quantity grows smaller either as $x \to c$ or as the allowable order $n+1$ of differentiation increases, provided M_{n+1} is bounded as $n \to \infty$. To express succinctly the rate at which R_{n+1} shrinks with the small parameter $(x-c)^{n+1}$, we write $R_{n+1} = \mathcal{O}((x-c)^{n+1})$.

The notation $\mathcal{O}(\cdot)$ appears in so many contexts that it warrants a formal definition.

DEFINITION. Let $\alpha(\epsilon)$ and $\beta(\epsilon)$ depend on some parameter ϵ. The notation $\alpha(\epsilon) = \mathcal{O}(\beta(\epsilon))$ as $\epsilon \to 0$ means there exist positive constants M and $\epsilon_{\max}$ such that $|\alpha(\epsilon)| \leqslant M|\beta(\epsilon)|$ whenever $0 < |\epsilon| < \epsilon_{\max}$. Similarly, $\alpha(\epsilon) = \mathcal{O}(\beta(\epsilon))$ as $\epsilon \to \infty$ if there exist positive constants M and $\epsilon_{\min}$ such that $|\alpha(\epsilon)| \leqslant M|\beta(\epsilon)|$ whenever $\epsilon > \epsilon_{\min}$.

Whether $\epsilon \to 0$ or $\epsilon \to \infty$ is often clear from context, and in these cases we typically omit explicit mention of the limits. This notation uses the symbol $=$ in an unusual way. For example, the definition implies the following:

1. If $\alpha(\epsilon) = \mathcal{O}(\gamma(\epsilon))$ and $\beta(\epsilon) = \mathcal{O}(\gamma(\epsilon))$, then
$$\alpha(\epsilon) \pm \beta(\epsilon) = \mathcal{O}(\gamma(\epsilon)). \tag{1.19}$$
2. If $0 < p < q$ and $\alpha(\epsilon) = \mathcal{O}(\epsilon^q)$ as $\epsilon \to 0$, then
$$\alpha(\epsilon) = \mathcal{O}(\epsilon^p) \quad as \quad \epsilon \to 0. \tag{1.20}$$
3. If $0 < p < q$ and $\alpha(\epsilon) = \mathcal{O}(\epsilon^p)$ as $\epsilon \to \infty$, then
$$\alpha(\epsilon) = \mathcal{O}(\epsilon^q) \quad as \quad \epsilon \to \infty. \tag{1.21}$$

Problem 1.19 asks for proofs.

PROOF OF THEOREM 1.5.8: Assume that $x \neq c$, the case $x = c$ being trivial. According to Theorem 1.5.5,

$$f(x) = f(c) + \int_c^x f'(t)\,dt.$$

If $n = 0$, letting $T_n(x,c) = f(c)$ completes the proof. If $n \geqslant 1$, integrate by parts, using $u(t) = f'(t)$ and $v(t) = -(x-t)$ in Theorem 1.17 to get

$$f(x) = f(c) + f'(c)(x-c) + \int_c^x (x-t)f''(t)\,dt.$$

Continue to integrate by parts in this way, using $u(t) = f^{(k)}(t)$ and $v(t) = -(x-t)^k/k!$ at the kth stage, until the allowable derivatives of f are exhausted. We then

have

$$f(x) = T_n(x, c) + \int_c^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt. \tag{1.22}$$

This equation, useful in its own right, is the **Taylor theorem with integral remainder**.

It remains to show that the integral on the right of this identity equals R_{n+1}, as defined in Eq. (1.18). We argue for the case when $c < x$, the case $c > x$ being similar. Call

$$m := \min_{t\in[c,x]} f^{(n+1)}(t), \quad M := \max_{t\in[c,x]} f^{(n+1)}(t),$$

which exist since $f^{(n+1)}$ is continuous on the interval $[c, x]$. By Theorem 1.5.3,

$$m \int_c^x \frac{(x-t)^n}{n!}\, dt \leqslant \int_c^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt \leqslant M \int_c^x \frac{(x-t)^n}{n!}\, dt.$$

Computing the integrals on the left and right and rearranging gives

$$m \leqslant \frac{(n+1)!}{(x-c)^{n+1}} \int_c^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt \leqslant M.$$

But $f^{(n+1)}$ is continuous, so the intermediate value theorem guarantees that there is a point $\zeta \in (c, x)$ such that

$$f^{(n+1)}(\zeta) = \frac{(n+1)!}{(x-c)^{n+1}} \int_c^x \frac{(x-t)^n}{n!} f^{(n+1)}(t)\, dt.$$

Solving this identity for the integral shows that it is identical to R_{n+1}. ∎

The Taylor theorem admits two special cases important enough to have their own names.

THEOREM 1.5.9 (MEAN VALUE THEOREM). *If $f \in C^1([a, b])$, then there is a point $\zeta \in (a, b)$ such that*

$$f'(\zeta) = \frac{f(b) - f(a)}{b - a}. \tag{1.23}$$

Equation (1.23) is just the Taylor theorem for the case $n = 0$. It guarantees the existence of a point $\zeta \in (a, b)$ where the derivative of f equals the average slope of f over $[a, b]$, as shown in Figure 1.9.

COROLLARY 1.5.10 (ROLLE'S THEOREM). *If $f \in C^1([a, b])$ has zeros at a and b, then there is a point $\zeta \in (a, b)$ where $f'(\zeta) = 0$.*

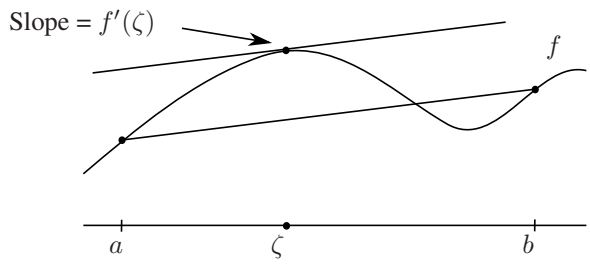

Figure 1.9 Graphic example of the mean value theorem. At the point ζ, the value of f' equals the average slope of f over the interval $[a, b]$.

PROOF: This is the mean value theorem for the case $f(a) = f(b) = 0$. ∎

The Taylor theorem extends to functions of several real variables. Instead of introducing the most general statement of the theorem, we examine two useful cases.

DEFINITION. Let $\Omega \subset \mathbb{R}^n$ be an open set, with $f\colon \Omega \to \mathbb{R}$. We say that $f \in C^1(\Omega)$ if f is continuous at each point $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \Omega$ and each of the partial derivatives $\partial f/\partial x_1, \partial f/\partial x_2, \ldots, \partial f/\partial x_n$ exists and is continuous at each $\mathbf{x} \in \Omega$. The vector-valued function $\nabla f\colon \Omega \to \mathbb{R}^n$ defined by

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \frac{\partial f}{\partial x_2}(\mathbf{x}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}) \right)$$

is the **gradient** of f.

The first extension of the Taylor theorem is the following:

THEOREM 1.5.11 (TAYLOR THEOREM TO ORDER 1). *Let $f \in C^1(\Omega)$, and suppose that $\mathbf{c}, \mathbf{x} \in \Omega$ and that the line segment connecting $\mathbf{c}$ and $\mathbf{x}$ lies entirely in Ω. Then there is a point $\boldsymbol{\zeta}$ lying on that line segment such that*

$$f(\mathbf{x}) = f(\mathbf{c}) + \nabla f(\boldsymbol{\zeta}) \cdot (\mathbf{x} - \mathbf{c}).$$

Figure 1.10 illustrates the theorem. Think of the line segment connecting $\mathbf{c}$ and $\mathbf{x}$ as an analog of the interval (c, x) in the one-dimensional Theorem 1.5.8.

PROOF: Define a function $\phi\colon [0,1] \to \mathbb{R}$ by setting $\phi(t) := f(\mathbf{c} + t(\mathbf{x} - \mathbf{c}))$. By the chain rule,

$$\begin{aligned} \phi'(t) &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{c} + t(\mathbf{x} - \mathbf{c})) \frac{d}{dt} \left[c_i + t(x_i - c_i) \right] \\ &= \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{c} + t(\mathbf{x} - \mathbf{c})) \, (x_i - c_i), \end{aligned}$$

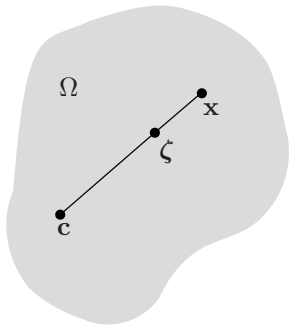

Figure 1.10 An open set $\Omega \subset \mathbb{R}^2$, with the points $\mathbf{c}$, $\mathbf{x}$, and $\boldsymbol{\zeta}$ referred to in Theorem 1.5.11.

the continuity of the individual terms in the sum guaranteeing that $\phi \in C^1([0,1])$. The mean value theorem yields a point $\zeta \in (0,1)$ such that $\phi(1) = \phi(0) + \phi'(\zeta)$. Therefore, by the definition of ϕ,

$$f(\mathbf{x}) = f(\mathbf{c}) + \nabla f\left(\mathbf{c} + \zeta(\mathbf{x} - \mathbf{c})\right) \cdot (\mathbf{x} - \mathbf{c}).$$

The vector $\boldsymbol{\zeta} := \mathbf{c} + \zeta(\mathbf{x} - \mathbf{c})$, which lies on the line segment between $\mathbf{c}$ and $\mathbf{x}$, is the desired point. ∎

To carry the Taylor expansion for $f\colon \Omega \to \mathbb{R}$ one term further, it is necessary to introduce more notation. We say that $f \in C^2(\Omega)$ if $f \in C^1(\Omega)$ and each of the second partial derivatives $\partial^2 f/\partial x_i \partial x_j$, $i, j = 1, 2, \ldots, n$, exists and is continuous at every $\mathbf{x} \in \Omega$. The matrix $\mathsf{H}_f(\mathbf{x}) \in \mathbb{R}^{n \times n}$ whose (i,j)th entry is

$$\frac{\partial^2 f}{\partial x_i \, \partial x_j}(\mathbf{x}) \tag{1.24}$$

is the **Hessian matrix** of f at $\mathbf{x}$. The continuity of the second partial derivatives guarantees that $\partial^2 f/\partial x_i \, \partial x_j = \partial^2 f/\partial x_j \, \partial x_i$, so the Hessian matrix is symmetric.

THEOREM 1.5.12 (TAYLOR THEOREM TO ORDER 2). *Let $f \in C^2(\Omega)$, and suppose that $\mathbf{c}, \mathbf{x} \in \Omega$ and that the line segment connecting $\mathbf{c}$ and $\mathbf{x}$ lies entirely in Ω. Then there exists a point $\boldsymbol{\zeta}$ lying on that line segment such that*

$$f(\mathbf{x}) = f(\mathbf{c}) + \nabla f(\mathbf{c}) \cdot (\mathbf{x} - \mathbf{c}) + \tfrac{1}{2}(\mathbf{x} - \mathbf{c}) \cdot \mathsf{H}_f(\boldsymbol{\zeta})(\mathbf{x} - \mathbf{c}).$$

The proof, which uses the one-dimensional Taylor expansion through order 2, is the subject of Problem 1.20.

While it is possible to extend the Taylor expansion for functions $f\colon \Omega \to \mathbb{R}$ to any order, depending upon the smoothness of f, for functions of several variables

we do not use expansions past the second order. One can also prove analogs of the Taylor theorem for vector-valued functions, a task that we postpone until Chapter 5.

1.6 Problems

1.1 For each of the following subsets of $\mathbb{R}$, determine the least upper bound and greatest lower bound, if they exist.

(A) $(0,1) \cup (2,2.1) \cup (3,3.01) \cup (4,4.001) \cup \cdots$.

(B) $\left\{1, -1, \frac{1}{2}, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{3}, \ldots\right\}$.

(C) $\left\{\exp(-x^2) \in \mathbb{R} \;\middle|\; x \in \mathbb{R}\right\}$.

1.2 Let $\mathcal{V}$ be a vector space. Show that $\operatorname{span}(S)$ is a vector space for any subset $S \subset \mathcal{V}$.

1.3 Prove that the null space $\mathcal{N}(\mathsf{A})$ is a subspace of $\mathbb{R}^n$ for any matrix $\mathsf{A} \in \mathbb{R}^{n\times n}$. This theorem extends to matrices $\mathsf{A} \in \mathbb{R}^{m\times n}$ for which $m \neq n$.

1.4 Given matrices A, B for which AB makes sense, show that $(\mathsf{AB})^\top = \mathsf{B}^\top \mathsf{A}^\top$.

1.5 Prove the alternative triangle inequality, Theorem 1.3.2.

1.6 Prove that $\|\cdot\|_1$, $\|\cdot\|_1$, and $\|\cdot\|_\infty$ are norms on $\mathbb{R}^n$.

1.7 Show that norm equivalence ($\|\cdot\|_{\mathrm{I}} \approx \|\cdot\|_{\mathrm{II}}$) is an equivalence relation.

1.8 Show that, for any $\mathbf{x} \in \mathbb{R}^n$,

(A) $\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_2 \leqslant \sqrt{n}\,\|\mathbf{x}\|_\infty$.

(B) $\sqrt{1/n}\,\|\mathbf{x}\|_1 \leqslant \|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\|_1$.

(C) $\|\mathbf{x}\|_\infty \leqslant \|\mathbf{x}\|_1 \leqslant n\|\mathbf{x}\|_\infty$.

Also show that these inequalities are sharp, in the sense that each inequality becomes an equality for some appropriate nonzero vector $\mathbf{x}$.

1.9 With respect to a given norm $\|\cdot\|$, a sequence $\{\mathbf{x}_k\}$ of vectors in $\mathbb{R}^n$ **converges** to $\mathbf{x} \in \mathbb{R}^n$ (written $\mathbf{x}_k \to \mathbf{x}$) under the following condition: For any $\epsilon > 0$, there is a number $N > 0$ such that $\|\mathbf{x}_k - \mathbf{x}\| < \epsilon$ whenever $k > N$. Let $\|\cdot\|_{\mathrm{I}}$ and $\|\cdot\|_{\mathrm{II}}$ be two norms on $\mathbb{R}^n$. Show that $\mathbf{x}_k \to \mathbf{x}$ with respect to $\|\cdot\|_{\mathrm{I}}$ if and only if $\mathbf{x}_k \to \mathbf{x}$ with respect to $\|\cdot\|_{\mathrm{II}}$.

1.10 Prove the inequality (1.11).

1.11 Matrix norms $\|\cdot\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ inherit nice properties of the vector norms that define them:

(A) Prove that any subordinate matrix norm satisfies the requirements to be a norm.

(B) Prove that any matrix norm $\|\mathsf{A}\|$ is a continuous function of the n^2 entries of A.

(C) Prove that all matrix norms on $\mathbb{R}^{n\times n}$ are equivalent.

(Propositions (B) and (C) do not require the norm to be subordinate to a vector norm.)

1.12 Prove the characterization

$$\|\mathsf{A}\| = \sup_{\|\mathbf{x}\|=1} \|\mathsf{A}\mathbf{x}\|$$

stated in Eqs. (1.12).

1.13 Prove that the coefficients in the expansion (1.6) are $c_j = \boldsymbol{\upsilon}_j \cdot \mathbf{y}$.

1.14 Suppose that $\mathsf{A} \in \mathbb{R}^{n\times n}$ and that p is a polynomial. Show the following:

(A) If λ is an eigenvalue of A, then $p(\lambda)$ is an eigenvalue of $p(\mathsf{A})$.

(B) If $\mathbb{R}^n$ has a basis consisting of eigenvectors of A and μ is an eigenvalue of $p(\mathsf{A})$, then there is an eigenvalue λ of A for which $\mu = p(\lambda)$. (Actually, the assumption that eigenvectors of A form a basis is not necessary.)

1.15 This problem examines properties of the spectral radius.

(A) Show that $\varrho(\mathsf{A}) \leqslant \|\mathsf{A}\|$ for any subordinate matrix norm $\|\cdot\|$. *Hint:* Consider eigenvectors having unit length.

(B) Show that the spectral radius $\varrho\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ is not a matrix norm by finding matrices A and B such that $\varrho(\mathsf{A}+\mathsf{B}) > \varrho(\mathsf{A}) + \varrho(\mathsf{B})$.

(C) Show that $1/\|\mathsf{A}^{-1}\| = \inf_{\|\mathbf{x}\|=1}\|\mathsf{A}\mathbf{x}\|$.

(D) Let A be symmetric and positive definite with smallest eigenvalue $\lambda_{\min}$. Show that $\|\mathsf{A}^{-1}\|_2 = 1/\lambda_{\min}$.

1.16 Not all matrix norms are subordinate:

(A) Show that the **Frobenius norm**

$$\|\mathsf{A}\|_F := \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{i,j}|^2\right)^{1/2}$$

satisfies the three conditions required of norms, as does the function

$$\|\mathsf{A}\|_{\max} := \max\ |a_{i,j}|\,.$$

(B) Neither of the norms in (A) is subordinate to a vector norm when $n > 1$. Therefore, we have no guarantee that the inequality (1.11) holds. Show that it fails for the norm $\|\cdot\|_{\max}$.

(C) Show that $\|\cdot\|_F$ is not subordinate to any vector norm for $n > 1$. *Hint:*: Consider $\|\mathsf{I}\|_F$.

1.17 Prove the matrix norm characterization

$$\|\mathsf{A}\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{i,j}|.$$

1.18 Prove the inequality (1.16).

1.19 Prove the statements (1.19) through (1.21).

1.20 Prove Theorem 1.5.12.

1.21 Prove that if $\|\cdot\|$ is a norm on $\mathbb{R}^n$, then the unit sphere $S := \left\{\mathbf{x}\in\mathbb{R}^n \;\middle|\; \|\mathbf{x}\| = 1\right\}$ is compact.

1.22 Let $\mathsf{A}\in\mathbb{R}^{n\times n}$, and let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Prove that the linear map defined by $\mathbf{x}\mapsto\mathsf{A}\mathbf{x}$ is uniformly continuous with respect to $\|\cdot\|$.

1.23 Let $\mathsf{H}\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ be a continuous, matrix-valued function. (The Hessian matrix of a function $f\colon \mathbb{R}^n \to \mathbb{R}$, defined in Eq. (1.24), is such a function provided f is twice continuously differentiable.) Show that, if $\mathsf{H}(\mathbf{y})$ is positive definite at a point $\mathbf{y}$, then there is a radius $\epsilon > 0$ such that $\mathsf{H}(\mathbf{x})$ is positive definite at every $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{y})$.