Appendix A
Glossary and Notation

The study of PDEs and, particularly their numerical solution has a rapacious appetite for variable names, constants, parameters, indices, and so on. This means that many symbols become overloaded—they are required to take on different meanings in different contexts. It is hoped that the list provided here will help in avoiding confusion.

$\mathcal{B}$: an operator used to represent boundary conditions
$\mathcal{B}_h$: a finite difference approximation of $\mathcal{B}$
BC: boundary condition
BVP: boundary value problem—that is a PDE together with boundary and/or initial conditions
c: the wave speed in the wave equation; the Courant number $c = ak/h$ when approximating the advection equation
∂_x, ∂_t: shorthand notation for partial derivatives with respect to x and t; so $u_x = \partial_x u$, $u_{xx} = \partial_x^2 u$, $u_{xt} = \partial_x \partial_t u$
$\triangle^+, \triangle^-, \triangle, \delta$: forward, backward and alternative central difference operators for functions of one variable; $\triangle_x^+, \triangle_x^-, \triangle_x, \delta_x$ apply to the x variable of a function of several variables
$\partial\Omega$: the boundary of a typical domain Ω
$\mathcal{F}, \mathcal{F}_h$: source term in a differential equation $\mathcal{L}u = \mathcal{F}$ or in a finite difference equation $\mathcal{L}_h U = \mathcal{F}_h$,
$\mathscr{F}, \mathscr{F}_h$: source data in a boundary value problem $\mathscr{L}u = \mathscr{F}$ or its finite difference approximation $\mathscr{L}_h U = \mathscr{F}_h$,
IC: initial condition
IVP: initial value problem
IBVP: initial-boundary value problem
κ: the conductivity coefficient in the heat equation; the wave number in a von Neumann analysis of stability
$\mathcal{L}, \mathcal{M}$: differential operators involving derivatives with respect to space variables only, such as $\mathcal{L} = -\partial_x^2$
$\mathcal{L}_h, \mathcal{M}_h$: finite difference approximations of differential operators $\mathcal{L}, \mathcal{M}$
$\mathscr{L}, \mathscr{M}$: differential operators together with associated boundary conditions

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

$\mathscr{L}_h, \mathscr{M}_h$: finite difference approximations of differential operators $\mathscr{L}, \mathscr{M}$

$\vec{n}$: the outward pointing normal direction; $\vec{n}(\vec{x})$ is the outward normal vector at a point $\vec{x}$ on the boundary

$\vec{\nabla}$: the gradient vector; $\vec{\nabla}u = [u_x, u_y]$ for a function $u(x, y)$ of two variables

∇^2: Laplacian operator; $\nabla^2 u = u_{xx} + u_{yy}$ for a function $u(x, y)$ of two variables

Ω, Ω_h: typical domain of a PDE or its approximation by a grid of points,

ODE: ordinary differential equation

PDE: partial differential equation

p: coefficient of first derivative in a differential equation; order of consistency of a finite difference approximation, as in $\mathcal{O}(h^p)$

φ, Φ: comparison functions for continuous ($\varphi \geq 0$ and $\mathscr{L}\varphi \geq 1$) and discrete problems ($\Phi \geq 0$ and $\mathscr{L}_h\Phi \geq 1$), respectively

u, U: typical solution of a boundary value problem $\mathscr{L}u = \mathscr{F}$ or its finite difference approximation $\mathscr{L}_h U = \mathscr{F}_h$,

r: the mesh ratio; $r = k/h^2$ when approximating the heat equation; the radial coordinate in polar coordinates

$\vec{r}$: position vector; $\vec{r} = (x, y)$ in two dimensions

$\mathcal{R}_h, \mathscr{R}_h$: local truncation error—$\mathcal{R}_h := \mathcal{L}_h u - \mathcal{F}_h$ and $\mathscr{R}_h := \mathscr{L}_h u - \mathscr{F}_h$, where u solves $\mathcal{L}u = \mathcal{F}$ and $\mathscr{L}u = \mathscr{F}$.

$\boldsymbol{x}^*$: Hermitian (or complex conjugate) transpose of a d-dimensional vector $\boldsymbol{x}$

ξ: the amplification factor in a von Neumann analysis of stability

$\|\cdot\|$: norm operator; maps d-dimensional vectors $\boldsymbol{x}$ (or real-valued functions) onto non-negative real numbers

$\|\cdot\|_h$: discrete norm operator; maps grid functions U onto non-negative real numbers; we use two flavours, the maximum (or ℓ_∞-) norm $\|\cdot\|_{h,\infty}$ and the ℓ_2-norm $\|\cdot\|_{h,2}$

$\langle\cdot,\cdot\rangle$: inner product; maps two d-dimensional vectors (or pairs of functions) onto real numbers

$\langle\cdot,\cdot\rangle_w$: inner product associated with a positive weight function w; maps pairs of functions onto real numbers.

Appendix B
Some Linear Algebra

B.1 Vector and Matrix Norms

Norms provide a convenient way of measuring the length of vectors and the magnifying ability of matrices.

Definition B.1 A norm on a vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|$ and is required to have the properties

(a) positivity: $\|\boldsymbol{x}\| > 0$ for $\boldsymbol{x} \neq \boldsymbol{0}$,
(b) uniqueness: $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$,
(c) scaling: $\|a\boldsymbol{x}\| = |a|\,\|\boldsymbol{x}\|$ for any complex scalar a,
(d) triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \le \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for any vector $\boldsymbol{y}$ having the same dimension as $\boldsymbol{x}$.

When $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]^\mathsf{T}$ a popular family are the so-called ℓ_p norms, and are given by

$$\|\boldsymbol{x}\|_p = \left(\sum_{j=1}^{d} |x_j|^p\right)^{1/p} \tag{B.1}$$

with $1 \le p < \infty$. Typical choices are $p = 1, 2, \infty$, where the case $p = \infty$ is interpreted as

$$\|\boldsymbol{x}\|_\infty = \max_{1 \le i \le d} |x_i|$$

and is known as the *maximum norm*. A particularly useful result is given by *Hölder's inequality* which states that, for any two complex vectors $\boldsymbol{x}$, $\boldsymbol{y}$ of the same dimension,

$$|\boldsymbol{x}^*\boldsymbol{y}| \le \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q, \quad 1/p + 1/q = 1, \tag{B.2}$$

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

where $\boldsymbol{x}^*$ denotes the Hermitian (or complex conjugate) transpose of $\boldsymbol{x}$. When $p = q = 2$ this is also known as the Cauchy–Schwarz inequality

$$|\boldsymbol{x}^*\boldsymbol{y}| \le \|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2. \tag{B.3}$$

Definition B.2 A norm on a matrix A is denoted by $\|A\|$ and is required to have the properties

(a) $\|A\| > 0$ for $A \ne 0$,
(b) $\|A\| = 0$ if and only if A is the zero matrix,
(c) $\|aA\| = |a|\,\|A\|$ for any complex scalar a,
(d) $\|A + B\| \le \|A\| + \|B\|$ for any matrix B of the same dimension as A,
(e) $\|AB\| \le \|A\|\|B\|$ for any matrix B for which the product AB is defined.

The requirement that (e) holds is unconventional. (A standard definition would only stipulate (a)–(d).) To find examples of matrix norms where (e) is satisfied, we note that analyses involving norms usually involve both vector and matrix norms, and it is standard practice in numerical analysis to use norms that are *compatible* in the sense that

$$\|A\boldsymbol{x}\| \le \|A\|\,\|\boldsymbol{x}\|. \tag{B.4}$$

One way in which this can be achieved is to first define a vector norm $\|\boldsymbol{x}\|$ and then use

$$\|A\| = \max_{\boldsymbol{x}\ne\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

to *induce* a matrix norm. When the vector norm is a p-norm, this approach leads to the matrix p-norms defined, for a $d \times d$ matrix A having a_{ij} in the ith row and jth column, by

$$\|A\|_1 = \max_{1\le j\le d} \sum_{i=1}^{d} |a_{ij}|, \quad \|A\|_\infty = \|A^\mathsf{T}\|_1, \tag{B.5}$$

while $\|A\|_2 = \sqrt{\lambda}$, where λ is the largest eigenvalue of A^*A. These matrix p-norms are examples for which property (e) automatically holds. We refer to Trefethen and Bau [26] for further discussion.

B.2 Symmetry of Matrices

A real $n \times n$ matrix A is said to be *symmetric* if it remains unchanged when its rows and columns are interchanged,[1] so that $A^\mathsf{T} = A$. If $A = (a_{ij})$, with a_{ij} denoting the entry in the ith row and jth column, then symmetry requires $a_{ij} = a_{ji}$. To exploit symmetry we note that $A^\mathsf{T} = A$ implies that, for vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$,

$$\boldsymbol{y}^*(A\boldsymbol{x}) = \boldsymbol{y}^* A\boldsymbol{x} = (A^\mathsf{T}\boldsymbol{y})^*\boldsymbol{x} = (A\boldsymbol{y})^*\boldsymbol{x}. \tag{B.6}$$

Thus $\boldsymbol{y}^*(A\boldsymbol{x}) = (A\boldsymbol{y})^*\boldsymbol{x}$ so the (complex) scalar product of $\boldsymbol{y}$ and $A\boldsymbol{x}$ is the same as the scalar product of $A\boldsymbol{y}$ and $\boldsymbol{x}$.

To take this idea further we need to introduce a special notation for scalar product—this is generally referred to as an *inner product*. Specifically, we define

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle \equiv \boldsymbol{y}^*\boldsymbol{x} \tag{B.7}$$

then, from (B.6),

$$\langle \boldsymbol{x}, A\boldsymbol{y}\rangle = \langle A\boldsymbol{x}, \boldsymbol{y}\rangle \tag{B.8}$$

for a symmetric matrix A.

B.3 Tridiagonal Matrices

Definition B.3 (*Irreducible tridiagonal matrix*) The $n \times n$ tridiagonal matrix A given by

$$A = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ 0 & \cdots & 0 & a_n & b_n \end{pmatrix} \tag{B.9}$$

is said to be *irreducible* if the off-diagonal entries a_j and c_j are all nonzero.

Suppose that $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$ is an $n \times n$ diagonal matrix and let A be an irreducible tridiagonal matrix. The product matrix

[1]We shall restrict ourselves to real matrices, the analogous property for complex matrices is that $A^* = A$, where A^* is the Hermitian (or complex conjugate) transpose.

$$DA = \begin{pmatrix} d_1b_1 & d_1c_1 & 0 & \cdots & 0 \\ d_2a_2 & d_2b_2 & d_2c_2 & & 0 \\ & \ddots & \ddots & \ddots & \\ & & d_{n-1}a_{n-1} & d_{n-1}b_{n-1} & d_{n-1}c_{n-1} \\ 0 & \cdots & 0 & d_na_n & d_nb_n \end{pmatrix}$$

will be a symmetric matrix if the entries of D are chosen so that

$$d_ja_j = d_{j-1}c_{j-1}, \quad j = 2, 3, \ldots, n.$$

Setting $d_1 = 1$, the fact that $a_j \neq 0$ means that $d_2, d_3, \ldots, d_n$ can be successively computed. The fact that $c_{j-1} \neq 0$ implies that $d_j \neq 0$, $j = 2, \ldots, n$, which ensures that the diagonal matrix D is *nonsingular*. It follows from this result that A will have real eigenvalues when $a_j/c_{j-1} > 0$, for $j = 2, 3, \ldots, n$ (see Exercise B.4). This is the linear algebra analogue of the clever change of variables for ODEs in Sect. 5.3.2 combined with (part of) Theorem 5.11.

The $n \times n$ tridiagonal matrix

$$T = \begin{bmatrix} b & c & 0 & \cdots & 0 \\ a & b & c & & \\ & \ddots & \ddots & \ddots & \\ & & a & b & c \\ 0 & & & a & b \end{bmatrix} \tag{B.10}$$

with constant diagonals, has eigenvalues

$$\lambda_j = b + 2\sqrt{ac}\,\cos\frac{j\pi}{n+1}, \quad j = 1, 2, \ldots, n. \tag{B.11}$$

with corresponding eigenvectors

$$\boldsymbol{v}_j = [\sin(\pi j x_1), \sin(\pi j x_2), \ldots, \sin(\pi j x_m), \ldots, \sin(\pi j x_n)]^{\mathsf{T}},$$

where $x_m = m/(n+1)$ (see Exercise B.5). Generalizations of this result can be found in Fletcher and Griffiths [3].

B.4 Quadratic Forms

Definition B.4 If $\boldsymbol{x} \in \mathbb{R}^n$ and A is a real $n \times n$ symmetric matrix, then the function

$$Q(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}} A \boldsymbol{x}$$

is called a *quadratic form*.

The form $Q(\boldsymbol{x})$ is a homogeneous quadratic function of the independent variables $x_1, x_2, \ldots, x_n$.[2] In the simplest (two-dimensional) case, we have $\boldsymbol{x} \in \mathbb{R}^2$, and

$$Q(\boldsymbol{x}) = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = ax^2 + 2bxy + cy^2. \tag{B.12}$$

We shall be concerned with the *level curves* of the quadratic form, that is points where $Q(\boldsymbol{x}) = \text{constant}$. If we make a change of variables: $\boldsymbol{x} = V\boldsymbol{s}$, where V is a nonsingular 2×2 matrix, then the quadratic form becomes

$$Q(V\boldsymbol{s}) = \boldsymbol{s}^{\mathsf{T}}(V^{\mathsf{T}}AV)\boldsymbol{s}$$

which is a quadratic form having coefficient matrix $V^{\mathsf{T}}AV$. The idea is to choose V in such a way that $V^{\mathsf{T}}AV$ is a diagonal matrix,

$$V^{\mathsf{T}}AV = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix},$$

so that, with $\boldsymbol{s} = [s, t]^{\mathsf{T}}$ we have

$$Q(V\boldsymbol{s}) = \alpha s^2 + \beta t^2. \tag{B.13}$$

We can picture the form in (B.13) geometrically. The points $Q = \text{constant}$ in the s-t plane are (a) ellipses if $\alpha\beta > 0$ (i.e., α and β have the same sign), (b) hyperbolae if $\alpha\beta < 0$ (i.e., they have opposite signs) and (c) straight lines[3] if one of α or β is zero (i.e., $\alpha\beta = 0$).

Two basic questions need to be answered at this point:

Q1. How do we construct the matrix V?
Q2. Different matrices V will lead to different values for α and β but will their signs remain the same?

To answer the first question: a natural candidate for V is the matrix of eigenvectors of A. To confirm this choice, suppose that A has eigenvalues λ_1, λ_2 with corresponding eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

$$A\boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j, \quad j = 1, 2.$$

Since A is symmetric, the eigenvalues are real and the two eigenvectors are orthogonal: $\boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_2 = 0 = \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_1$. Setting $V = [\boldsymbol{v}_1, \boldsymbol{v}_2]$, we get

[2] Homogeneous means that $Q(c\boldsymbol{x}) = c^2 Q(\boldsymbol{x})$ for $c \in \mathbb{R}$; this is the case because there are no linear terms and no constant term.

[3] When $Q(\boldsymbol{x})$ also contains linear terms in $\boldsymbol{x}$ then $Q(V\boldsymbol{s})$ will, in general, contain linear terms in $\boldsymbol{s}$. In such cases the level curves will be parabolae.

$$V^{\mathsf{T}}AV = \begin{bmatrix} \lambda_1 \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_1 & \lambda_2 \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_2 \\ \lambda_1 \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_1 & \lambda_2 \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_1 & 0 \\ 0 & \lambda_2 \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_2 \end{bmatrix}$$

so that we have (B.13) with $\alpha = \lambda_1 \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{v}_1$ and $\beta = \lambda_2 \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{v}_2$. If the eigenvectors are normalised to have unit length then $\alpha = \lambda_1$ and $\beta = \lambda_2$. In both cases the sign of the product $\alpha\beta$ is the same as that of $\lambda_1\lambda_2$ so the level curves will be ellipses if $\lambda_1\lambda_2 > 0$ and hyperbolae if $\lambda_1\lambda_2 < 0$. Moreover, since the characteristic polynomial of the matrix A is given by

$$\det(A - \lambda I) = \lambda^2 - (a + c)\lambda + (ac - b^2),$$

then the product of the eigenvalues is given by $\lambda_1\lambda_2 = ac - b^2 = \det(A)$, so we see that it is not necessary to compute the eigenvalues in order to determine the nature of the level curves. This is formally stated in the following result.

Theorem B.5 *The level curves of the quadratic form $Q(x, y) = ax^2 + 2bxy + cy^2$ are* hyperbolae *if $b^2 - ac > 0$ (when $Q(x, y)$ has two distinct real factors) and* ellipses *if $b^2 - ac < 0$ (when $Q(x, y)$ has no real factors). In the intermediate case, $b^2 - ac = 0$ (when $Q(x, y)$ is a perfect square) the level curves are straight lines.*

If the eigenvectors are normalised then $V^{\mathsf{T}}V = I$ and $\boldsymbol{x} = V\boldsymbol{s}$ may be inverted to give

$$\boldsymbol{s} = \begin{bmatrix} s \\ t \end{bmatrix} = V^{\mathsf{T}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{v}_1^{\mathsf{T}} \\ \boldsymbol{v}_2^{\mathsf{T}} \end{bmatrix} \boldsymbol{x}$$

so that $s = \boldsymbol{v}_1^{\mathsf{T}}\boldsymbol{x}$ is the component of $\boldsymbol{x}$ in the direction of $\boldsymbol{v}_1$ and $t = \boldsymbol{v}_2^{\mathsf{T}}\boldsymbol{x}$ is the component of $\boldsymbol{x}$ in the direction of $\boldsymbol{v}_2$. In the nondenegerate case ($b^2 \neq ac$), a further rescaling of the components of $\boldsymbol{s}$ to new variables $\boldsymbol{\xi} = [\xi, \eta]^{\mathsf{T}}$ defined by $\xi = s\sqrt{|\lambda_1|}$ and $\eta = t\sqrt{|\lambda_2|}$ can be applied. This gives $\boldsymbol{s} = D\boldsymbol{\xi}$ where

$$D = \begin{bmatrix} 1/\sqrt{|\lambda_1|} & 0 \\ 0 & 1/\sqrt{|\lambda_2|} \end{bmatrix}$$

is a nonsingular diagonal matrix with $Q(VD\boldsymbol{\xi}) = \pm 1(\xi^2 + \eta^2)$ in the elliptic case and $Q(VD\boldsymbol{\xi}) = \pm 1(\xi^2 - \eta^2)$ in the hyperbolic case. Thus, with this special rescaling, the contours are either circles or rectangular hyperbolae in the ξ-η plane. The process is illustrated by the following example.

Example B.6 Determine the level curves of the quadratic form associated with the matrix

$$A = \begin{bmatrix} 13 & -4 \\ -4 & 7 \end{bmatrix}.$$

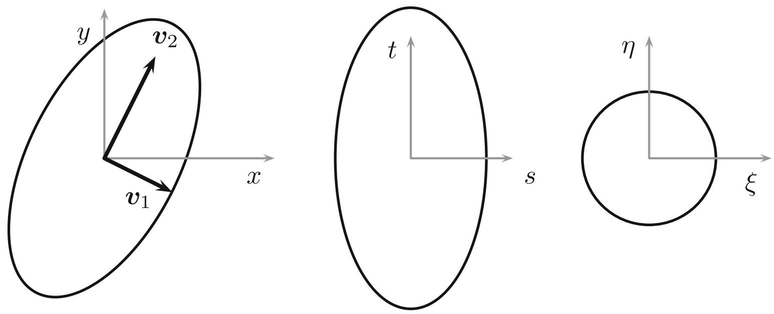

Fig. B.1 Level curves of $Q(\boldsymbol{x})$ (*left*), $Q(V\boldsymbol{s})$ (*centre*) and $Q(VD\boldsymbol{\xi})$ (*right*) for Example B.6. The mapping from s-t to x-y corresponds to a simple rotation of the coordinate system

Computing the eigenvalues and eigenvectors of A gives

$$\lambda_1 = 15, \quad \boldsymbol{v}_1 = \tfrac{1}{\sqrt{5}}\begin{bmatrix} 2 \\ -1 \end{bmatrix}, \qquad \lambda_2 = 5, \quad \boldsymbol{v}_2 = \tfrac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Setting $V = [\boldsymbol{v}_1, \boldsymbol{v}_2]$ and changing variables gives $s = (2x - y)/\sqrt{5}$, $t = (x + 2y)/\sqrt{5}$, $\xi = (2x - y)\sqrt{3}$ and $\eta = x + 2y$ with associated quadratic forms

$$Q(V\boldsymbol{s}) = 15s^2 + 5t^2, \quad Q(VD\boldsymbol{\xi}) = \xi^2 + \eta^2.$$

We note that both of these correspond to writing

$$Q(x, y) = 3\,(2x - y)^2 + (x + 2y)^2.$$

The level curves $Q(\boldsymbol{x}) = 25$ in the x-y plane, $Q(V\boldsymbol{s}) = 25$ in the s-t plane and $Q(VD\boldsymbol{\xi}) = 1$ in the ξ-η plane are shown in Fig. B.1. The ellipse in the s-t plane is a rotation of the ellipse in the x-y plane anticlockwise through the angle $\tan^{-1}(1/2)$. The eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ can be seen to be directed along the minor and major axes of the ellipse associated with $Q(\boldsymbol{x})$. ♢

Returning to the second question above, we will see that suitable transformation matrices V can be defined without knowledge of the eigenvalues or eigenvectors of A. The key element in the construction of such matrices is Sylvester's law of inertia, stated below. Two definitions will be needed beforehand.

Definition B.7 (*Congruence*) Assuming that V is a nonsingular matrix, the matrix (triple-) product $V^\top AV$ is called a *congruence transformation* of A.

The two matrices $V^\top AV$ and A are said to be congruent. The set of eigenvalues of two congruent matrices will generally be different (unless V is an orthogonal matrix). A congruence transformation does, however, retain just enough information about the eigenvalues to be useful in the current context. This information is called the *inertia* of a matrix.

Definition B.8 (*Inertia*) The *inertia* of a symmetric matrix is a triple of integers (p, z, n) giving the number of positive, zero and negative eigenvalues, respectively.

Theorem B.9 (Sylvester's law of inertia) *The inertia of a symmetric matrix is invariant under a congruence transformation.*

Sylvester's law of inertia guarantees that the qualitative nature of the level curves are invariant when the matrix is subject to congruence transformations. Thus in two dimensions the level curves of a matrix having inertia $(2, 0, 0)$ or $(0, 0, 2)$ will be ellipses while an inertia of $(1, 0, 1)$ will lead to hyperbolae.

One natural possibility for a congruence transformation is to use Gaussian elimination. When $a \neq 0$ we subtract b/a times the 1st row of A from the second row in order to create a zero in the $(2, 1)$ position. The elimination process can be represented by defining V^T so that

$$V^\mathsf{T} = \begin{bmatrix} 1 & 0 \\ -b/a & 1 \end{bmatrix}, \quad V^\mathsf{T}A = \begin{bmatrix} a & b \\ 0 & (ac-b^2)/a \end{bmatrix}, \quad V^\mathsf{T}AV = \begin{bmatrix} a & 0 \\ 0 & (ac-b^2)/a \end{bmatrix}$$

so we generate (B.13) with $\alpha = a$ and $\beta = (ac - b^2)/a$. The congruence transformation explicitly highlights the role of the discriminant $(ac - b^2)$ of the underlying quadratic form. Next, computing the inverse transformation matrix gives

$$V^{-1} = \begin{bmatrix} 1 & b/a \\ 0 & 1 \end{bmatrix},$$

and setting $\boldsymbol{s} = V^{-1}\boldsymbol{x}$, we find that $s = x + by/a$, $t = y$. Thus, from (B.13),

$$Q(x, y) = a\left(x + \frac{by}{a}\right)^2 + \left(c - \frac{b^2}{a}\right)y^2$$

which can be seen to be equivalent to "completing the square".

Example B.10 (Example B.6 revisited) Completing the square in the quadratic form associated with the matrix

$$A = \begin{bmatrix} 13 & -4 \\ -4 & 7 \end{bmatrix}$$

we find that

$$Q(x, y) = 13x^2 - 8xy + 7y^2 = 13\left(x - \frac{4}{13}y\right)^2 + \frac{75}{13}y^2.$$

This suggests an alternative mapping from (x, y) to (s, t) via $s = x - (4/13)\,y$ $(= x + by/a)$, $t = y$. The change of variables leads to

$$Q(V\boldsymbol{s}) = 13s^2 + \frac{75}{13}t^2,$$

and defines the ellipse in the s-t plane that is associated with the Gaussian elimination congruence transformation $V^{\mathsf{T}}AV$. Making the further scaling $\xi = s/\sqrt{13}$, $\eta = t\sqrt{13/75}$ gives $Q(VD\boldsymbol{\xi}) = \xi^2 + \eta^2$, whose level curves in the ξ-η plane are again circles. ♢

B.5 Inverse Monotonicity for Matrices

A real matrix A or a real vector $\boldsymbol{x}$ which has entries that are all nonnegative numbers is called a *nonnegative* matrix or vector. They can be identified by writing $A \geq 0$ and $\boldsymbol{x} \geq 0$. The matrix interpretation of an inverse monotone discrete operator is called a monotone matrix.

Definition B.11 (*Monotone matrix*) A nonsingular real $n \times n$ matrix A is said to be *monotone* if the inverse matrix A^{-1} is nonnegative: equivalently, $A\boldsymbol{x} \geq 0$ implies that $\boldsymbol{x} \geq 0$.

The standard way of showing that a matrix is monotone is to show that it is an M-matrix.

Definition B.12 (*M-matrix*) A real $n \times n$ nonsingular matrix A is an M-matrix if

(a) $a_{ij} \leq 0$ for $i \neq j$ (this means that A is a Z-matrix), and
(b) the real part of every eigenvalue of A is positive.

An M-matrix is guaranteed to be monotone. To expand on condition (b): first, a symmetric matrix A has real eigenvalues. They are all positive numbers if and only if the quadratic form $Q(\boldsymbol{x}) = \boldsymbol{x}^{\mathsf{T}}A\boldsymbol{x}$ is positive for every nonzero vector $\boldsymbol{x}$ (such a matrix is said to be *positive definite*). Second, an irreducible nonsymmetric matrix will satisfy (b) if it is also diagonally dominant; that is if $a_{ii} \geq \sum_{j=1, j\neq i}^{n} |a_{ij}|$ with strict inequality for at least one value of i. (This result immediately follows from the Gershgorin circle theorem.[4])

[4]This states that: let $r_i = \sum_{i\neq j} |a_{i,j}|$ denote the sum of the moduli of the off-diagonal entries of an $n \times n$ matrix A, then every eigenvalue of A lies in at least one of the disks of radius r_i centered on a_{ii}.

Exercises

B.1 Prove that an $n \times n$ matrix A is symmetric if and only if $\langle \boldsymbol{x}, A\boldsymbol{y} \rangle = \langle A\boldsymbol{x}, \boldsymbol{y} \rangle$ for all vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. (Hint: Choose $\boldsymbol{x} = \boldsymbol{e}_i$ and $\boldsymbol{y} = \boldsymbol{e}_j$, where $\boldsymbol{e}_k$ is the kth unit vector in $\mathbb{R}^n$, that is, the vectors whose only nonzero entry is one in the kth position.)

B.2 Suppose that A is an irreducible $n \times n$ tridiagonal matrix. Construct the nonsingular diagonal matrix $D = \operatorname{diag}(1, d_2, \ldots, d_n)$ that makes the product matrix AD symmetric.

B.3 Suppose that A is a tridiagonal matrix and let D be a nonsingular diagonal matrix that makes the matrix DA symmetric. Prove that AD^{-1} is a symmetric matrix.

B.4 Suppose that the matrix A in (B.9) is irreducible and that its elements are real with $a_j/c_{j-1} > 0$ for $j = 2, 3, \ldots, n$. Show that its eigenvalues are real.

[Hint: $A\boldsymbol{v} = \lambda\boldsymbol{v}$ for an eigenvector $\boldsymbol{v} \in \mathbb{C}^n$ and corresponding eigenvalue λ. Now consider the inner product $\langle \boldsymbol{v}, AD\boldsymbol{v} \rangle$.]

B.5 Suppose that $U_m = \sin(\pi jm/(n+1))$ so that $U_0 = 0$ and $U_{n+1} = 0$ when j is an integer. Show, using the trigonometric identity for $\sin(A \pm B)$, that

$$aU_{m-1} + bU_m + cU_{m+1} = \lambda_j U_m,$$

where λ_j is the jth eigenvalue of the tridiagonal matrix (B.10).

B.6 Sketch the level curves of the quadratic form $Q(\boldsymbol{x})$ when

$$A = \begin{bmatrix} -10 & 10 \\ 10 & 5 \end{bmatrix}.$$

B.7 Consider a quadratic form $Q(\boldsymbol{x})$ of the form (B.12) with $a \geq c > 0$ and $b^2 < ac$. Suppose that $\hat{a} = \frac{1}{2}(a+c)$ and that r, θ are defined by $a = \hat{a}(1 + r\cos\theta)$, $c = \hat{a}(1 - r\cos\theta)$ and $b = \hat{a}r\sin\theta$ ($0 \leq r < 1$, $-\pi/2 \leq \theta \leq \pi/2$). Show that

$$r^2 = 1 - 4\frac{ac - b^2}{(a+c)^2}$$

and that the matrix of the quadratic form has eigenvalues $\lambda_\pm = \hat{a}(1 \pm r)$ with corresponding eigenvectors

$$\boldsymbol{v}_+ = (\cos\tfrac{1}{2}\theta, \sin\tfrac{1}{2}\theta), \quad \boldsymbol{v}_- = (\sin\tfrac{1}{2}\theta, -\cos\tfrac{1}{2}\theta),$$

where $\tan\theta = 2b/(a-c)$.

If the major axis of the ellipse $Q(\boldsymbol{x}) = \text{constant}$ makes an angle ϕ with the y-axis, show that $-\pi/4 \leq \phi \leq \pi/4$.

B.8 This builds on the previous exercise. Attention is drawn in Sect. 10.2.5 to quadratic forms where the coefficients satisfy $b^2 < ac \le a|b|$. Show that the minimum value of r, subject to these inequalities, is $1/\sqrt{2}$. Hence show that the ratio of the lengths of the two axes of the ellipse $Q =$ constant (that is, $\sqrt{\lambda_+/\lambda_-}$) is greater than, or equal to, $\sqrt{2} + 1$.

B.9 Find the stationary points of the function $\phi(x, y) = 2x + 2ye^x + y^2$.

The leading terms in the Taylor expansion of a function of two variables about a point $\boldsymbol{x} = \boldsymbol{a}$ may be expressed in the matrix-vector form

$$\phi(\boldsymbol{x}) = \phi(\boldsymbol{a}) + (\boldsymbol{x} - \boldsymbol{a})^\mathsf{T}\boldsymbol{g} + (\boldsymbol{x} - \boldsymbol{a})^\mathsf{T} H(\boldsymbol{x} - \boldsymbol{a}) + \cdots$$

where $\boldsymbol{x} = (x, y)^\mathsf{T}$, $\boldsymbol{g} = (\phi_x(\boldsymbol{a}), \phi_y(\boldsymbol{a}))^\mathsf{T}$ is the gradient of ϕ at $\boldsymbol{a}$ and

$$H = \begin{bmatrix} \phi_{xx}(\boldsymbol{a}) & \phi_{xy}(\boldsymbol{a}) \\ \phi_{yx}(\boldsymbol{a}) & \phi_{yy}(\boldsymbol{a}) \end{bmatrix}$$

is the matrix of second derivatives (or Hessian matrix) evaluated at $\boldsymbol{a}$. Use the expansion to determine whether the stationary points of $\phi(x, y)$ are maxima, minima or saddle points.

Appendix C
Integral Theorems

This appendix reviews the most important results in vector calculus: these are generalisations of the *fundamental theorem of integral calculus*, that is

$$\int_a^b f'(x)\,\mathrm{d}x = f(b) - f(a).$$

Theorem C.1 *Suppose that $\Omega \subset \mathbb{R}^3$ is a closed bounded region with a piecewise smooth boundary $\partial\Omega$ and let $\vec{n}$ denote the unit outward normal vector to $\partial\Omega$. If ϕ and $\vec{F}$ denote, respectively, scalar and vector fields defined on a region that contains Ω, then*

$$\text{(i)} \quad \iiint\limits_{\Omega} \operatorname{grad}\phi\,\mathrm{d}V = \iint\limits_{\partial\Omega} \phi\,\vec{n}\,\mathrm{d}S$$

$$\text{(ii)} \quad \iiint\limits_{\Omega} \operatorname{div}\vec{F}\,\mathrm{d}V = \iint\limits_{\partial\Omega} \vec{F}\cdot\vec{n}\,\mathrm{d}S$$

$$\text{(iii)} \quad \iiint\limits_{\Omega} \operatorname{curl}\vec{F}\,\mathrm{d}V = \iint\limits_{\partial\Omega} \vec{n}\times\vec{F}\,\mathrm{d}S.$$

These three identities are associated with the famous names of George Green and Carl Friedrich Gauss. The result (ii) is perhaps of greatest importance in applications and is often referred to as the *divergence theorem*. What might not be obvious is that if any one of the identities in Theorem C.1 is true then the other parts follow. This is established below. It will be assumed that, in Cartesian coordinates, $\vec{F} = [F_x, F_y, F_z]$ with a similar notation for $\vec{n}$.

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

- (i)⇒(ii) If we take $\phi = F_x$ and write down the first (or the x-) component of the vector-valued identity (i), we get

$$\iiint\limits_{\Omega} \frac{\partial F_x}{\partial x}\,\mathrm{d}V = \iint\limits_{\partial\Omega} F_x n_x\,\mathrm{d}S$$

where n_x is the first component of $\vec{n}$. Moreover, taking $\phi = F_y$ and $\phi = F_z$ and writing down the second and third components of $\vec{n}$ gives

$$\iiint\limits_{\Omega} \frac{\partial F_y}{\partial y}\,\mathrm{d}V = \iint\limits_{\partial\Omega} F_y n_y\,\mathrm{d}S, \quad \iiint\limits_{\Omega} \frac{\partial F_z}{\partial z}\,\mathrm{d}V = \iint\limits_{\partial\Omega} F_z n_z\,\mathrm{d}S.$$

Adding these results gives identity (ii) written in Cartesian coordinates.

- (ii)⇒(i) If we take $\vec{F} = [\phi, 0, 0]$ then (ii) gives

$$\iiint\limits_{\mathrm{V}} \frac{\partial \phi}{\partial x}\,\mathrm{d}V = \iint\limits_{\mathrm{S}} \phi\, n_x\,\mathrm{d}S,$$

which is the first component of result (i). Taking $\vec{F} = [0, \phi, 0]$ and $\vec{F} = [0, 0, \phi]$ give the second and third components. Hence, (i) follows from (ii).

- (i)⇒(iii) The first component of (iii) is

$$\iiint\limits_{\mathrm{V}} \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right)\mathrm{d}V = \iint\limits_{\mathrm{S}} \left(F_z n_y - F_y n_z\right)\mathrm{d}S.$$

To establish this result, the second component of (i) with $\phi = F_z$ must be subtracted from the third component of (i) with $\phi = F_y$, that is

$$\iiint\limits_{\mathrm{V}} \frac{\partial F_z}{\partial y}\,\mathrm{d}V = \iint\limits_{\mathrm{S}} F_z n_y\,\mathrm{d}S \quad \text{and} \quad \iiint\limits_{\mathrm{V}} \frac{\partial F_y}{\partial z}\,\mathrm{d}V = \iint\limits_{\mathrm{S}} F_y n_z\,\mathrm{d}S.$$

The other two components of (iii) can be established in exactly the same way.

- (iii)⇒(i) Suppose that $\vec{v}$ is a constant vector. Taking $\vec{F} = \phi\vec{v}$ in identity (iii) gives

$$\iiint\limits_{\mathrm{V}} \mathrm{curl}(\phi\vec{v})\,\mathrm{d}V = \iint\limits_{\mathrm{S}} \vec{n} \times (\vec{v}\phi)\,\mathrm{d}S.$$

Next, using the vector identity

$$\mathrm{curl}(\phi\vec{v}) = \phi\,\mathrm{curl}\,\vec{v} - \vec{v} \times \mathrm{grad}\,\phi$$

the left hand side simplifies (since $\vec{v}$ is a constant vector, its curl is zero) to give

$$-\iiint\limits_{V} \vec{v} \times \operatorname{grad}\phi \, \mathrm{d}V = \iint\limits_{S} \vec{n} \times (\vec{v}\phi) \, \mathrm{d}S = -\iint\limits_{S} \phi \, \vec{v} \times \vec{n} \, \mathrm{d}S$$

where we have used $\vec{a} \times \vec{b} = -\vec{b} \times \vec{a}$ on the right hand side. Since $\vec{v}$ is a constant vector, it can be taken outside the integrals to give

$$-\vec{v} \times \iiint\limits_{V} \operatorname{grad}\phi \, \mathrm{d}V = -\vec{v} \times \iint\limits_{S} \phi \, \vec{n} \, \mathrm{d}S.$$

Finally, since the above result is valid *for all* constant vectors $\vec{v}$, identity (i) must always hold. □

Appendix D
Bessel Functions

The ODE

$$x^2u'' + xu' + (x^2 - \nu^2)u = 0 \tag{D.1}$$

defined on the semi-infinite real line $0 < x < \infty$ and involving a real parameter ν is known as Bessel's equation. This is a linear second-order ODE and therefore has two linearly independent solutions. These are usually denoted by $J_\nu(x)$ and $Y_\nu(x)$ and are referred to as Bessel functions of the first and second kind, respectively, of order ν.

The equation cannot be solved in terms of standard elementary functions and so the method of Frobenius is used to construct a series solution. That is, coefficients $\{a_n\}$ are sought such that a solution may be expressed in the form

$$u(x) = \sum_{n=0}^{\infty} a_n x^n \tag{D.2}$$

It is sufficient for our purposes to consider the case $\nu = 0$. Thus, on substituting (D.2) into (D.1) and collecting terms with like powers of x, we find

$$x^2u'' + xu' + x^2u = a_1x + \sum_{n=2}^{\infty}\Big(n^2a_n + a_{n-2}\Big)x^n$$

and the right hand side is zero when $a_1 = 0$ and

$$a_n = -\frac{1}{n^2}a_{n-2}, \quad n = 2, 3, \ldots. \tag{D.3}$$

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

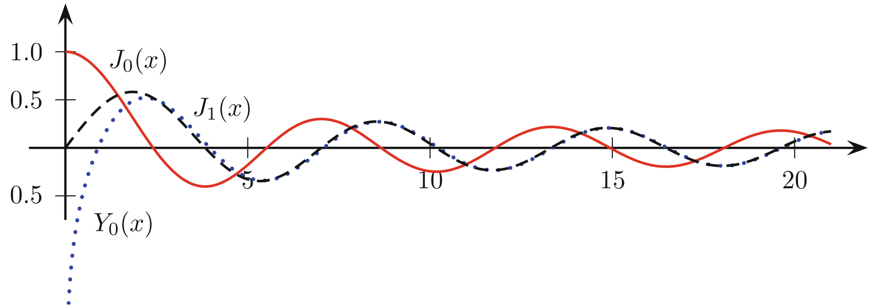

Fig. D.1 The Bessel functions $J_0(x)$ (*solid*), $J_1(x)$ (*dashed*) and $Y_0(x)$ (*dotted*) for $0 \le x \le 21$

When $n = 3, 5, \dots$ we find that all odd-numbered terms in (D.2) vanish while, writing $n = 2m$, the even-numbered coefficients are

$$a_{2m} = (-1)^m \frac{1}{2^{2m}(m!)^2} a_0, \quad m = 0, 1, 2, \dots \tag{D.4}$$

in which a_0 is an arbitrary constant The remaining coefficient a_0 is used to normalise the solution and, by choosing $a_0 = 1$, we find that $u(x) = J_0(x)$, where

$$J_0(x) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{(n!)^2} \left(\frac{x}{2}\right)^{2n} \tag{D.5}$$

is the Bessel function of the first kind of order zero. Note that $J_0(0) = 1$ and $J_0'(0) = 0$.

Bessel functions of the second kind have the property that $Y_\nu(x) \to -\infty$ as $x \to 0^+$ and, since we make no direct use of them in this book, they will not be discussed further. Graphs of $J_0(x)$, $J_1(x)$ and $Y_0(x)$ are shown in Fig. D.1. The graphs of $J_1(x)$ and $Y_0(x)$ are indistinguishable for $x > 3$. Their oscillatory behaviour is evident and the zeros play an important role when using separation of variables, as illustrated in Example 8.5. A selection of the zeros of $J_0(x)$ are given in Table D.1. It can be shown that, for large values of x,

Table D.1 Selected zeros ξ_k of $J_0(x)$ compared with $(k - \frac{1}{4})\pi$ from the approximation (D.6) with $\nu = 0$

k	1	2	3	5	10	20
ξ_k	2.4048	5.5201	8.6537	14.9309	30.6346	62.0485
$(k - \frac{1}{4})\pi$	2.3562	5.4978	8.6394	14.9226	30.6305	62.0465
	0.0486	0.0223	0.0143	0.0084	0.0041	0.0020

The bottom row gives the difference between the preceding two rows

$$J_\nu(x) \approx \sqrt{\frac{2}{\pi x}} \cos\Big(x - (\nu + \tfrac{1}{2})\frac{\pi}{2}\Big), \tag{D.6}$$

which suggests that the zeros of $J_0(x)$ approach $(k - \frac{1}{4})\pi$ as $k \to \infty$. The table confirms that this is a reasonable approximation even for small values of k.

Bessel functions have played an important role in applied mathematics for more than a century. One reason for this prominent position is that those equations that can be solved in Cartesian coordinates by the use of sines and cosines require Bessel functions when the equations are expressed in polar coordinates. Bessel functions have been studied extensively over the years. A comprehensive review can be found in the celebrated book by G.N. Watson [27] (originally published in 1922), whereas Kreyszig [12] provides an accessible introduction to their properties.

Exercises

D.1 Use the result of Exercise 5.4 to show that solutions of (D.1) are oscillatory.

D.2 By differentiating (D.1) when $\nu = 0$, show that $u = cJ_0'(x)$ satisfies Bessel's equation with $\nu = 1$ and the initial condition $u(0) = 0$ for any constant c.

D.3 Follow the process leading to (D.5) to show that (D.1) with $\nu = 1$ has a solution $u(x) = J_1(x)$, where

$$J_1(x) = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!(n+1)!} \left(\frac{x}{2}\right)^{2n+1}. \tag{$\star$}$$

Verify, by term by term differentiation, that $J_1(x) = -J_0'(x)$, confirming the result of the previous exercise with $c = -1$.

D.4 Show that equation (D.1) with $\nu = 0$ can be rewritten in the form $(xu')' + xu = 0$. Hence show that

$$\int_0^x s J_0(s)\,\mathrm{d}s = x J_1(x).$$

D.5 By multiplying (D.1) with $\nu = 0$ by $2u'$, establish the result

$$\frac{\mathrm{d}}{\mathrm{d}x}\big(x^2 u(x)'^2 + x^2 u(x)^2\big) = 2xu^2(x).$$

Then, using the fact that $u = J_0(x)$ is a solution of (D.1), show that

$$\int_0^\xi x J_0^2(x)\,\mathrm{d}x = \tfrac{1}{2}\xi^2 J_1^2(\xi), \tag{$\ddagger$}$$

where ξ is any zero of J_0.

D.6 By making an appropriate change of variable in (‡) establish the relation

$$\int_0^a r J_0^2(r\xi/a)\,\mathrm{d}r = \tfrac{1}{2}a^2 J_1^2(\xi).$$

D.7 Using a search engine, or otherwise, explore the *ratio test* for convergence of an infinite series. Use the ratio test to show that the series for $J_0(x)$ in equation (D.5) and $J_1(x)$ in Exercise D.3 converge for all values of x.

Appendix E
Fourier Series

Consider the case of a real-valued function $u(x)$ that is defined on the real line $-\infty < x < \infty$ and is *periodic*, of period L. Thus $u(x + L) = u(x)$ for all x and knowledge of u on any interval of length L is sufficient to define it on the entire real line. Let us suppose that $u(x)$ is to be determined on the interval $(0, L)$. Such a function $u(x)$ has a well-defined Fourier series expansion into complex exponentials (sines and cosines)

$$u(x) = \sum_{j=-\infty}^{\infty} c_j \mathrm{e}^{2\pi \mathrm{i} jx/L}, \tag{E.1}$$

with complex-valued coefficients c_j that are constructed to ensure that the series converges at almost every point x. Since we know (from Exercise 5.19) that the functions $\{\exp(2\pi \mathrm{i} jx/L)\}$ $(j = 0, \pm 1, \pm 2, \ldots)$ are mutually orthogonal with respect to the complex $L_2(0, L)$ inner product (5.17), we can determine a general Fourier coefficient c_k by multiplying (E.1) by $\exp(-2\pi kx/L)$ and integrating with respect to x over the interval. This gives

$$c_k = \frac{1}{L}\int_0^L \mathrm{e}^{-2\pi \mathrm{i} kx/L} u(x)\,\mathrm{d}x. \tag{E.2}$$

Expanding (E.2) in terms of sines and cosines, and then taking the complex conjugate gives

$$c_k = \frac{1}{L}\int_0^L \{\cos(2\pi \mathrm{i} kx/L) - \mathrm{i}\sin(2\pi \mathrm{i} kx/L)\}\; u(x)\,\mathrm{d}x \tag{E.3}$$

$$c_k^* = c_{-k} = \frac{1}{L}\int_0^L \{\cos(2\pi \mathrm{i} kx/L) + \mathrm{i}\sin(2\pi \mathrm{i} kx/L)\}\; u(x)\,\mathrm{d}x. \tag{E.4}$$

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

Combining these results gives new coefficients a_k, b_k, so that

$$a_k := c_k + c_{-k} = \frac{2}{L}\int_0^L \cos(2\pi kx/L)\, u(x)\, \mathrm{d}x \tag{E.5}$$

$$b_k := \mathrm{i}\,(c_k - c_{-k}) = \frac{2}{L}\int_0^L \sin(2\pi kx/L)\, u(x)\, \mathrm{d}x. \tag{E.6}$$

Next, rearranging the expressions in (E.5) and (E.6) gives $c_k = \frac{1}{2}(a_k - \mathrm{i}b_k)$ and $c_{-k} = \frac{1}{2}(a_k + \mathrm{i}b_k)$, and substituting these expressions into (E.1) and rearranging (see Exercise E.1) generates the *standard form* (no complex numbers!)

$$u(x) = \frac{1}{2}a_0 + \sum_{k=1}^{\infty}\{a_k\cos(2\pi kx/L) + b_k\sin(2\pi kx/L)\}\,. \tag{E.7}$$

Another direct consequence of the mutual orthogonality of the complex exponentials is *Parseval's relation* (see Exercise E.2)

$$\sum_{k=-\infty}^{\infty}|c_k|^2 = \frac{1}{L}\int_0^L |u(x)|^2\, \mathrm{d}x. \tag{E.8}$$

This implies that the Fourier coefficients are square summable if, and only if, the function is square integrable. This also shows that the coefficients a_k and b_k will need to decay with increasing k if the Fourier series (E.1) is to converge: coefficients of the form $(1, 1, 1, 1, \ldots)$ will not be allowed, whereas $(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$ will be just fine (because $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots < \infty$).

Example E.1 Construct a piecewise continuous function (see Definition 8.3) that has Fourier coefficients that decay harmonically, that is $|c_k| = \frac{a}{|k|}$ for $k \neq 0$, where a is a constant (so that the series (E.1) converges).

Integrating the right hand side of (E.2) by parts we get

$$c_k = -\frac{1}{2\pi\mathrm{i}k}\, u(x)\,\mathrm{e}^{-2\pi\mathrm{i}kx/L}\Big|_0^L + \frac{1}{2\pi\mathrm{i}k}\int_0^L \mathrm{e}^{-2\pi\mathrm{i}kx/L}u'(x)\,\mathrm{d}x. \tag{E.9}$$

The factor $1/k$ in the first term of (E.9) suggests the specific choice $u(x) = x$ (the constant derivative means that the second integral is zero). Note that the periodic extension of this function is *discontinuous* at the boundary points: $u(0^+) = 0 \neq u(L^-) = L$. When these two limits are substituted into the first term in (E.9) we get $c_k = -\frac{L}{2\pi\mathrm{i}k}$ so that $|c_k| = \frac{L}{2\pi|k|}$, as required.

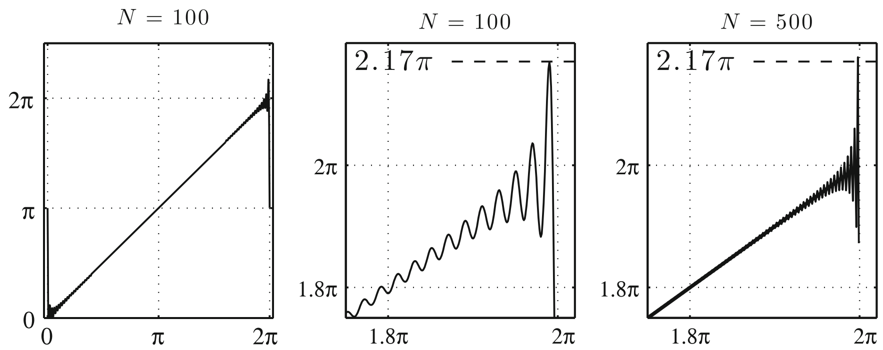

Fig. E.1 The Fourier series of the function $u(x) = x$ for $x \in (0, 2\pi)$ truncated to $N = 100$ real terms (*left*). An expanded view in the *square* around the point $(2\pi, 2\pi)$ is shown with $N = 100$ (*centre*) and $N = 500$ (*right*)

This slow decay of the coefficients has significant implications. The Nth partial sum with $L = 2\pi$ is (since $c_0 = \pi$)

$$S_N(x) = \sum_{k=-N}^{N} c_k e^{ikx} = \pi - \sum_{k=1}^{N} \frac{2}{k} \sin(kx)$$

and is shown on the left of Fig. E.1 with $N = 100$. It is seen to give a coherent approximation of $u(x) = x$ except at points close to the discontinuities at $x = 0, 2\pi$. The expanded view in the central figure reveals a highly oscillatory behaviour. When N is increased to 500, the oscillations in the rightmost figure appear to be confined to a smaller interval but their amplitude is undiminished—a dashed horizontal line is drawn at $y = 2.17\pi$ for reference. This is an illustration of Gibb's phenomenon and is a consequence of using continuous functions in order to approximate a discontinuous one. The "overshoot" in the partial sum is at least 8.5 % (closer, in fact, to 9 %) of the total jump (2π) at $x = 2\pi$. This, and other features, are succinctly explained by Körner [11, pp. 62–66]. ♢

The following theorem relates the smoothness of the function u to the rate of decay of the coefficients c_k as the index k increases.

Theorem E.2 *Let $s \in \{1, 2, \ldots\}$ be a parameter. If the function u and its first $(s-1)$ derivatives are continuous and its sth derivative is* piecewise *continuous on $(-\infty, \infty)$, then the Fourier coefficients satisfy*

$$|c_k| \le a/|k|^{s+1} \tag{E.10}$$

for $k \neq 0$, where a is a constant.

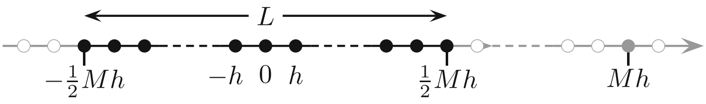

Fig. E.2 Grid points over a full period $-\frac{1}{2}Mh \le x \le \frac{1}{2}Mh$ when M is even

Proof Suppose that $s = 1$. The function u is continuous (and periodic) so $\lim_{x\to 0^+} u(x) = \lim_{x\to L^-} u(x)$. Thus, unlike the previous example, the first term in (E.9) is zero, so that

$$|c_k| = \frac{1}{k}\cdot\Bigl|\underbrace{\frac{1}{2\pi \mathrm{i}}\int_0^L \mathrm{e}^{-2\pi \mathrm{i}kx/L} u'(x)\,\mathrm{d}x}_{c_k'}\Bigr|.$$

Note that c_k' is a scalar multiple of the Fourier coefficient of the (piecewise continuous) derivative function, so the coefficients $|c_k'|$ will decay harmonically (at worst). The general result may be obtained by repeating the above argument $s-1$ times (each integration by parts gives another factor of k in the denominator). □

Next, suppose that our periodic function u is sampled on the grid of points $x_m = mh$, $m = 0, \pm 1, \pm 2, \ldots$ where $h = L/M$ (as illustrated in Fig. E.2), so that

$$u(x_m) = \sum_{k=-\infty}^{\infty} c_k\, \mathrm{e}^{2\pi \mathrm{i}km/M}. \tag{E.11}$$

The periodicity is reflected in the fact that $\exp(2\pi \mathrm{i}k'm/M) = \exp(2\pi \mathrm{i}km/M)$ for $k' = k + \ell M$ where ℓ is an integer. In this situation we say that the wave numbers $2\pi k$ and $2\pi k'$ are *aliases* of each other. If we now collect all the like coefficients together, then (E.11) can be written as a finite sum of distinct Fourier modes

$$u(x_m) = \sum_{k=0}^{M-1} \tilde{c}_k\, \mathrm{e}^{2\pi \mathrm{i}km/M}, \quad \tilde{c}_k = \sum_{\ell=-\infty}^{\infty} c_{k+\ell M}, \tag{E.12}$$

with modified coefficients $\tilde{c}_k$ that are periodic with period M: $\tilde{c}_{k+M} = \tilde{c}_k$ for all k. Note that if M is an even number (as illustrated above) then the function u can be defined over the interval $(-\frac{M}{2}h, \frac{M}{2}h)$ by simply summing over $M+1$ Fourier modes (instead of M modes) and halving the contribution of the first and last terms (since they are equal). This leads to the alternative representation[5]

$$u(x_m) = \sum_{|k|\le M/2}^{*} \tilde{c}_k\, \mathrm{e}^{2\pi \mathrm{i}km/M}. \tag{E.13}$$

[5] The first and last term adjustment is indicated by the asterisk in the summation.

Note that if M is odd then the range of $|k|$ will be $\lfloor M/2 \rfloor$ and no adjustment of the two end contributions is needed.

The exact representation of the periodic function in (E.12) (or equivalently (E.13)) involves a summation of an infinite number of the coefficients c_k. In a practical computation the exact coefficients $\tilde{c}_k$ will be approximated by (discrete) coefficients C_k and the associated *discrete* Fourier series is constructed

$$U_m = \sum_{k=0}^{M-1} C_k \mathrm{e}^{2\pi \mathrm{i} km/M}, \tag{E.14}$$

where $\{U_m\}$ are a set of periodic grid values ($U_{m+M} = U_m$ for all integers m) that approximate the exact grid values $u(x_m)$. The construction of the discrete Fourier coefficients C_k mirrors the construction used in the continuous case. All that is needed is a suitable inner product.

Theorem E.3 *The discrete Fourier modes* $\{\mathrm{e}^{2\pi \mathrm{i} km/M}\}_{m=0}^{M-1}$ *associated with distinct wave numbers k are mutually orthogonal with respect to the discrete inner product*[6]

$$\langle U, V \rangle_h = \frac{1}{M} \sum_{m=0}^{M}{}'' \, U_m V_m^*. \tag{E.15}$$

Proof By construction

$$\langle \mathrm{e}^{2\pi \mathrm{i} km/M}, \mathrm{e}^{2\pi \mathrm{i} \ell m/M} \rangle_h = \frac{1}{M} \sum_{m=0}^{M-1} \mathrm{e}^{2\pi \mathrm{i}(k-\ell)m/M} = \frac{1}{M} \sum_{m=0}^{M-1} z^m,$$

where $z = \mathrm{e}^{2\pi \mathrm{i}(k-\ell)/M}$ is one of the Mth roots of unity, so $z^M = 1$. The formula for the sum of a geometric progression then gives

$$\langle \mathrm{e}^{2\pi \mathrm{i} km/M}, \mathrm{e}^{2\pi \mathrm{i} \ell m/M} \rangle_h = \begin{cases} M & \text{when } k = \ell \\ \frac{1}{M} \frac{1-z^M}{1-z} = 0 & \text{when } k \neq \ell. \end{cases}$$

□

The general coefficient C_k can thus be determined by multiplying (E.14) by $\exp(-2\pi \mathrm{i} km/M)$ and summing over m. This gives

$$C_k = \frac{1}{M} \sum_{m=0}^{M-1} U_m \mathrm{e}^{-2\pi \mathrm{i} km/M}. \tag{E.16}$$

[6] The associated ℓ_2 norm $\|U\|_{h,2}$ is also defined in (11.41). The primes on the summation symbol signify that the first and last terms of the sum are halved.

The discrete Fourier coefficients will be periodic: $C_{k+M} = C_k$ for all k, and mirroring the exact coefficients c_k they satisfy $C_{-k} = C_k^*$ and $C_{M-k} = C_k^*$: thus we only need half of the coefficients in order to represent a real-valued grid function U. A classic algorithm that can be used to efficiently compute the coefficients $\{C_k\}$ (in both real- and complex-valued cases) is the *Fast Fourier Transform*. The algorithm is especially effective when M is a power of 2. Further details can be found in the book by Briggs [2].

Following the argument in the lead-up to (E.13) shows that we can write

$$U_m = \sum_{|k|\le\lfloor M/2\rfloor}^{*} C_k \mathrm{e}^{2\pi \mathrm{i} km/M}, \tag{E.17}$$

in place of (E.14). We can also establish (see Exercise E.5) a discrete version of Parseval's relation,

$$\sum_{k=0}^{M-1} |C_k|^2 = \frac{1}{M}\sum_{m=0}^{M-1} |U_m|^2 = \|U\|_{h,2}^2, \tag{E.18}$$

which implies that the discrete ℓ_2 norm of a periodic function is intrinsically connected to the sum of squares of the discrete Fourier coefficients.

The book by Strang [20, Chap. 4] is recommended for an overview of Fourier analysis and its role in applied mathematics.

Exercises

E.1 Show also that (E.1) can be written as (E.7) when $c_k = \frac{1}{2}(a_k - \mathrm{i}b_k)$ and $\{a_k\}$, $\{b_k\}$ are real sequences.

E.2 Show that Parseval's relation (E.8) follows from (E.2).

E.3 Verify that the discrete inner product (E.15) satisfies all the properties of the function inner product (5.17) that are listed in Exercise 5.11.

E.4 Show that the coefficients in (E.16) satisfy $C_{-\ell} = C_\ell^*$ when U is a real sequence.

E.5 Show that the discrete version of Parseval's relation (E.18) follows from (E.16) using an argument analogous to that in Exercise E.2.

References

1. U.M. Ascher, *Numerical Methods for Evolutionary Differential Equations* (SIAM, Philadelphia, 2008)
2. W.L. Briggs, V.E. Henson, *The DFT: An Owners' Manual for the Discrete Fourier Transform* (SIAM, Philadelphia, 1987)
3. R. Fletcher, D.F. Griffiths, The generalized eigenvalue problem for certain unsymmetric band matrices. Linear Algebra Appl. **29**, 139–149 (1980)
4. B. Fornberg, A finite difference method for free boundary problems. J. Comput. Appl. Math. **233**, 2831–2840 (2010)
5. N.D. Fowkes, J.J. Mahon, *An Introduction to Mathematical Modelling* (Wiley, New York, 1996)
6. G.H. Golub, C.F.V. Loan, *Matrix Computations*, 4th edn., Johns Hopkins Studies in the Mathematical Sciences (Johns Hopkins University Press, Maryland, 2012)
7. D.F. Griffiths, D.J. Higham, *Numerical Methods for Ordinary Differential Equations*, Springer Undergraduate Mathematics Series (Springer, London, 2010)
8. Iserles A, S.P. Nørsett, Order Stars (Chapman & Hall, London, 1991)
9. R. Jeltsch, J.H. Smit, Accuracy barriers of difference schemes for hyperbolic equations. SIAM J. Numer. Anal. **24**, 1–11 (1987)
10. M. Kac, Can one hear the shape of a drum? Am. Math. Mon. **73**, 1–23 (1966)
11. T.W. Körner, *Fourier Analysis* (Cambridge University Press, Cambridge, 1989)
12. E. Kreyszig, *Advanced Engineering Mathematics: International Student Version*, 10th edn. (Wiley, Hoboken, 2011)
13. B.P. Leonard, A stable and accurate convection modelling procedure based on quadratic upstream differencing. Comput. Methods Appl. Mech. Eng. **19**, 59–98 (1979)
14. R.J. LeVeque, *Finite difference methods for ordinary and partial differential equations* (SIAM, Philadelphia, 2007)
15. K.W. Morton, *Numerical Solution of Convection-Diffusion Problems* (Chapman & Hall, Philadelphia, 1996)
16. J.D. Pryce, *Numerical Solution of Sturm-Liouville Problems, Monographs in Numerical Analysis* (Oxford University Press, New York, 1993)
17. R. Rannacher, Discretization of the heat equation with singular initial data. ZAMM **62**, 346–348 (1982)
18. R.D. Richtmyer, K.W. Morton, *Difference Methods for Initial Value Problems* (Wiley Interscience, New York, 1967)
19. H.-G. Roos, M. Stynes, L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations, Convection Diffusion and Flow Problems*, vol. 24 (Springer Series in Computational Mathematics. Springer, Berlin, 1996)

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

20. G. Strang, *Introduction to Applied Mathematics* (Wellesley-Cambridge Press, Wellesley, 1986)
21. J.C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, 2nd edn. (SIAM, Philadelphia, 2004)
22. M. Stynes, Steady-State Convection-Diffusion Problems, in *Acta Numerica*, ed. by A. Iserles (Cambridge University Press, Cambridge, 2005), pp. 445–508
23. E. Süli, D. Mayers, *An Introduction to Numerical Analysis* (Cambridge University Press, Cambridge, 2003)
24. P.K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws. SIAM J. Numer. Anal. **21**, 995–1011 (1984)
25. L.N. Trefethen, Group velocity in finite difference schemes. SIAM Rev. **24**, 113–136 (1982)
26. L.N. Trefethen, D. Bau III, *Numerical Linear Algebra* (SIAM, Philadelphia, 1997)
27. G.N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd edn. (Cambridge Mathematical Library. Cambridge University Press, Cambridge, 1995)

Index

D.F. Griffiths et al., *Essential Partial Differential Equations*, Springer Undergraduate Mathematics Series, DOI 10.1007/978-3-319-22569-2

D

E

F

G

H