APPENDIX A

Additional Details and Fortification for Chapter 1

A.1 Matrix Classes and Special Matrices

The matrices can be grouped into several classes based on their operational properties. A short list of various classes of matrices is given in Tables A.1 and A.2. Some of these have already been described earlier, for example, elementary, symmetric, hermitian, othogonal, unitary, positive definite/semidefinite, negative definite/semidefinite, real, imaginary, and reducible/irreducible.

Some of the matrix classes are defined based on the existence of associated matrices. For instance, A is a diagonalizable matrix if there exists nonsingular matrices T such that $TAT^{-1} = D$ results in a diagonal matrix D. Connected with diagonalizable matrices are normal matrices. A matrix B is a normal matrix if $BB^* = B^*B$. Normal matrices are guaranteed to be diagonalizable matrices. However, defective matrices are not diagonalizable. Once a matrix has been identified to be diagonalizable, then the following fact can be used for easier computation of integral powers of the matrix:

$$A = T^{-1}DT \quad \rightarrow \quad A^k = \left(T^{-1}DT\right)\left(T^{-1}DT\right)\cdots\left(T^{-1}DT\right) = T^{-1}D^kT$$

and then take advantage of the fact that

$$D^K = \begin{pmatrix} d_1^k & & 0 \\ & \ddots & \\ 0 & & d_N^K \end{pmatrix}$$

Another set of related classes of matrices are the idempotent, projection, involutory, nilpotent, and convergent matrices. These classes are based on the results of integral powers. Matrix A is idempotent if $A^2 = A$, and if, in addition, A is hermitian, then A is known as a projection matrix. Projection matrices are used to partition an N-dimensional space into two subspaces that are orthogonal to each other. A matrix B is involutory if it is its own inverse, that is, if $B^2 = I$. For example, a reflection matrix such as the Householder matrix is given by

$$H = I - \frac{2}{\mathbf{v}^*\mathbf{v}}\mathbf{v}\mathbf{v}^*$$

where $\mathbf{v}$ is a nonzero vector, and then $H = H^{-1}$. A convergent matrix (also known as stable matrix) C is a matrix for which $\lim_{k\to\infty} C^k = 0$. These matrices are important

Table A.1. *Matrix classes (based on operational properties)*

Class	Definition	Remarks
Convergent (Stable)	$\lim_{k\to\infty} A^k = 0$	
Defective (Deficient)	$\sum_{k=0}^{r} \alpha_k A^k = 0$ $\alpha_r \neq 0\,;\ r < N$	
Diagonalizable	$T^{-1}AT$ is diagonal for some nonsingular T	
Elementary	Any matrix that scales, interchanges, or adds multiples of rows or columns of another matrix B	• Used in Gaussian elimination
Gram	$A = B^*B$ for some B	• Are Hermitian
Hermitian	$A^* = A$	• $(B + B^*)/2$ is the hermitian part of B. • B^*B and BB^* are hermitian
Idempotent	$A^2 = A$	• $\mathbf{det}(A) = 1$ or $\mathbf{det}(A) = 0$
Involutory	$A^2 = I$, i.e. $A = A^{-1}$	• Examples: identity matrix reverse unit matrices, symmetric orthogonal matrices
Negative definite	$\mathbf{x}^*A\mathbf{x} < 0 \quad \mathbf{x} \neq 0$	
Negative semidefinite	$\mathbf{x}^*A\mathbf{x} \leq 0 \quad \mathbf{x} \neq 0$	
Nilpotent (of degree k)	$A^k = 0\,;\ k > 0$	• $\mathbf{det}(A) = 0$
Normal	$AA^* = A^*A$	• Are diagonalizable
Nonsingular (Invertible)	$\lvert A\rvert \neq 0$	

for procedures that implement iterative computations. If, in addition, $k < \infty$ for $C^k = 0$, then the stable matrix will belong to the subclass of nilpotent matrices.

Aside from the classifications given in Tables A.1 and A.2, we also list some special matrices based on the structure and composition of the matrices. These are given in Table A.3. Some of the items in this table serve as a glossary of terms for the special matrices already described in this chapter. Some of the matrices refer to matrix structures based on the positions of zero and nonzero elements such as banded, sparse, triangular, tridiagonal, diagonal, bidiagonal, anti-diagonal, and Hessenberg. Some involve additional specifications on the elements themselves. These include identity, reverse identity, shift, real, complex, polynomial, rational, positive/negative, or nonpositive/nonnegative matrices. For instance, positive (or nonnegative) matrices

Table A.2. *Matrix classes (based on operations)*

Class	Definition	Remarks
Orthogonal	$A^T = A^{-1}$	
Positive definite	$\mathbf{x}^* A\mathbf{x} > 0\,;\mathbf{x} \neq 0$	
Positive semidefinite	$\mathbf{x}^* A\mathbf{x} \geq 0\,;\mathbf{x} \neq 0$	
Projection	Idempotent and Hermitian	
Reducible	There exists permutation P such that $\widehat{A} = PAP^T$ is block triangular	
Skew-symmetric	$A^T = -A$	• **det**$(A) = 0$ if N is odd • $a_{ii} = 0$, thus trace$(A) = 0$ • $(B - B^T)/2$ is the **skew-symmetric** part of B
Skew-hermitian	$A^* = -A$	• $a_{ii} = 0$ or pure imaginary • $(B - B^*)/2$ is the **skew-hermitian** part of B
Symmetric	$A = A^T$	• $B^T B$ and BB^T are both symmetric but generally not equal • $(B + B^T)/2$ is the **symmetric part** of B
Unitary	$A^* = A^{-1}$	

are matrices having only positive (or nonnegative) elements.[1] Some special matrices depend on specifications on the pattern of the nonzero elements. For instance, we have Jordan, Toeplitz, Shift, Hankel, and circulant matrices, as well as their block matrix versions, that is, block-Jordan, block-Toeplitz, and so forth. There are also special matrices that depend on collective properties of the rows or columns. For instance, stochastic matrices are positive matrices in which the sum of the elements within each row should sum up to unity. Another example are diagonally dominant matrices, where for the elements of any fixed row, the sum of the magnitudes of off-diagonal elements should be less than the magnitude of the diagonal element in that row. Finally, there are matrices whose entries depend on their row and column indices, such as Fourier, Haddamard, Hilbert, and Cauchy matrices. Fourier and Haddamard matrices are used in signal-processing applications.

As can be expected, these tables are not exhaustive. Instead, the collection shows that there are several classes and special matrices found in the literature. They often contain interesting patterns and properties such as analytical formulas for determinants, trace, inverses, and so forth, that could be taken advantage of during analysis and computations.

[1] Note that positive matrices are not the same as positive definite matrices. For instance, with

$$A = \begin{pmatrix} 1 & 5 \\ 5 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & -2 \\ 0 & 2 \end{pmatrix}$$

A is positive but not positive definite, whereas B is positive definite but not positive.

Table A.3. *Matrices classes (based on structure and composition)*

Name	Definition	Remarks
Antidiagonal	$A = \begin{pmatrix} 0 & & \alpha_1 \\ & \cdots & \\ \alpha_N & & 0 \end{pmatrix}$	• AB (or BA) will reverse sequence of rows (columns) of B, scaled by α_i • $\mathbf{det}(A) = (-1)^N \prod \alpha_i$ • MATLAB: `A=flipud(diag(alpha))` where `alpha`=$(\alpha_1, \ldots, \alpha_N)$
Band (or banded)	$a_{ij} = 0$ if $\begin{cases} i > j + p \\ \text{or} \\ j > i + q \end{cases}$	• p is the right-bandwidth • q is the left-bandwidth
Bidiagonal (Stieltjes)	$A = \begin{pmatrix} \alpha_1 & & & 0 \\ \beta_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ 0 & & \beta_{N-1} & \alpha_N \end{pmatrix}$	• $\mathbf{det}(A) = \prod_{i=1}^{N} \alpha_i$ • Let $B = A^{-1}$ then if $j > i$, $b_{ij} = 0$ if $j = i$, $b_{ii} = \dfrac{1}{\alpha_i}$ if $i > j$, $b_{ij} = \dfrac{1}{\alpha_i} \prod_{k=j}^{i-1} \left(-\dfrac{\beta_k}{\alpha_k}\right)$ • MATLAB: `A=diag(v)+diag(w,-1)` where `v`= $(\alpha_1, \ldots, \alpha_N)$ `w`= $(\beta_1, \ldots, \beta_{N-1})$
Binary	$a_{ij} = 0$ or 1	• Often used to indicate incidence relationship between i and j
Cauchy	For given $\mathbf{x}$ and $\mathbf{y}$ $a_{ij} = \dfrac{1}{x_i + y_j}$; $x_i + y_j \neq 0$ and elements of $\mathbf{x}$ and $\mathbf{y}$ are distinct	• Are nonsingular (but often ill-conditioned for large N) • $\mathbf{det}(A) = \dfrac{\prod_{i=2}^{N} \prod_{j=1}^{i-1} f_{ij}}{\prod_{i=1}^{N} \prod_{j=1}^{N} (x_i + y_j)}$ where $f_{ij} = (x_i - x_j)(y_i - y_j)$ • MATLAB: `A=gallery('cauchy',x,y)`
Circulant	$A = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \alpha_N & \alpha_1 & \cdots & \alpha_{N-1} \\ & \cdots & & \\ \alpha_2 & \alpha_3 & \cdots & \alpha_1 \end{pmatrix}$	• Are normal matrices • Are special case of Toeplitz • MATLAB: `A=gallery('circul',alpha)` where `alpha`= $(\alpha_1, \cdots, \alpha_N)$
Companion	$A = \left(\begin{array}{ccc\|c} -p_{n-1} & \cdots & -p_1 & -p_0 \\ \hline 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & 1 & 0 \end{array}\right)$	• p_k are coefficients of a polynomial: $s^N + p_{N-1}s^{n-1} + p_1 s + p_0$ • MATLAB: `A=compan(p)` where `p`= $(1, p_{n-1}, \ldots, p_1, p_0)$
Complex	a_{ij} are complex-valued	

Name	Definition	Remarks
Diagonal	$A = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_N \end{pmatrix}$	• $\mathbf{det}(A) = \prod_i \alpha_i$ • MATLAB: `A=diag(alpha)` where `alpha`$= (\alpha_1, \ldots, \alpha_N)$
Diagonally dominant	$\lvert a_{ii}\rvert > \sum_{i \neq j} \lvert a_{ij}\rvert$ $i = 1, 2, \ldots, N$	• Nonsingular (based on Gersgorin's theorem)

Fourier	$a_{ij} = (1/\sqrt{N}) W^{(i-1)(j-1)}$
	$W = \exp\left(-\sqrt{-1}\frac{2\pi}{N}\right)$
		• Are orthogonal
		• Used in Fourier transforms
		• MATLAB:
		`h=ones(N,1)*[0:N-1];`
		`W=exp(-2*pi/N*1i);`
		`A=W.^(h.*h')/sqrt(N)`

Givens (Rotation)	Identity matrix with 4 elements replaced based on given p and q: $a_{pp} = a_{qq} = \cos(\theta)$ $a_{pq} = -a_{qp} = \sin(\theta)$	• Used to rotate points in hyperplane • Useful in matrix reduction to Hessenberg form • Are orthogonal
Hadamard	$H_k[=]2^k \times 2^k$ $H_k = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes H_{k-1}$ $H_0 = 1$	• Elements either 1 or -1 • Are orthogonal • MATLAB: `A=hadamard(2^k)`
Hankel	$A = \begin{pmatrix} \cdots & & \beta & \alpha \\ & \cdots & \cdots & \gamma \\ \beta & \cdots & \cdots & \\ \alpha & \gamma & & \cdots \end{pmatrix}$	• Each anti-diagonal has the same value • MATLAB: `A=hankel([v,w])` where `v`$= (\ldots, \beta, \alpha)$ `w`$= (\alpha, \gamma, \ldots)$
Hessenberg	$a_{j+k,j} = 0$ $2 \leq k \leq (N - j)$	• Useful in finding eigenvalues • For square B, there is unitary Q such that $A = Q^* B Q$ is upper hessenberg • MATLAB: `[Q,A]=hess(B);` where `A=(Q')(B)(Q)`

Hilbert	$a_{ij} = \frac{1}{i + j - 1}$	• Symmetric and positive definite
		• MATLAB:
		`h=[1:N];`
		`A=gallery('cauchy',h,h-1)`

Identity	$A = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$	• Often denoted by I_N • $\mathbf{det}(A) = 1$ • $AB = BA = B$ • MATLAB: `A=eye(N)`

(*continued*)

Table A.3 *(continued)*

Name	Definition	Remarks
Imaginary	$A = iB$ where B is real and $i = \sqrt{-1}$	
Jordan block	$A = \begin{pmatrix} s & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & s \end{pmatrix}$	• Are bidiagonal • $\mathbf{det}(A) = s^N$ • MATLAB: `A=gallery('jordbloc',N,s)`
Lower Triangular	$a_{i,j} = 0\,;\ j > i$	• $\mathbf{det}(A) = \prod_{i=1}^{N} a_{ii}$ • Let $D = diag(A)$ and $K = D - A$ $A^{-1} = D^{-1}\left[I + \sum_{\ell=1}^{N-1} \left(KD^{-1}\right)^{\ell}\right]$ • MATLAB: `A=tril(B)` extracts the lower triangle of B
Negative	$a_{ij} < 0$	
Non-negative	$a_{ij} \geq 0$	
Non-positive	$a_{ij} \leq 0$	
Permutation	$P = \left(\mathbf{e}_{k_1} \middle\vert \cdots \middle\vert \mathbf{e}_{k_N} \right)^T$ $k_1 \neq \cdots \neq k_N$ $\mathbf{e}_j$ is the j^{th} unit vector	• PA (or AP^T) rearranges columns (or rows) of A based on sequence K • MATLAB: `B=eye(N);P=B(K,:)`
Persymmetric	$A[=]N \times N$ $a_{i,j} = a_{(N+1-j),(N+1-i)}$	$A = RH$ for reverse identity R and symmetric H
Positive	$a_{ij} > 0$	
Polynomial	a_{ij} are polynomial functions	
Real	a_{ij} are real valued	
Rational	a_{ij} are rational functions	
Rectangular (non-square)	$A[=]N \times M; N \neq M$	• if $N > M$ then A is **tall** • if $N < M$ then A is **wide**
Reverse identity	$A = \begin{pmatrix} 0 & & 1 \\ & \cdots & \\ 1 & & 0 \end{pmatrix}$	• AB (or BA) will reverse the order of the rows (or columns) of B • $\mathbf{det}(A) = (-1)^{(N/2)}$ if N is even $\mathbf{det}(A) = (-1)^{(N-1)/2}$ if N is odd • MATLAB: `A=flipud(eye(N))`
Sparse	Significant number of elements are zero	(see Section 1.6)

Name	Definition	Remarks
Stochastic (Probability, transition)	A is real, nonnegative and $\sum_{j=1}^{N} a_{ij} = 1$ for $i = 1, 2, \ldots N$	• aka **Right-Stochastic** • **Left-Stochastic** if $\sum_{i=1}^{N} a_{ij} = 1, \forall j$ • **Doubly-Stochastic** if both right- and left- stochastic
Shift	$A = \left(\begin{array}{c\|ccc} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \\ \hline 1 & 0 & \cdots & 0 \end{array}\right)$	• Are circulant, permuation and Toeplitz • $A^N = I_N$ • MATLAB: `A=circshift(eye(N),-1)`
Toeplitz	$A = \begin{pmatrix} \alpha & \beta & & \cdots \\ \gamma & \alpha & \ddots & \\ & \ddots & \ddots & \beta \\ \cdots & & \gamma & \alpha \end{pmatrix}$	• Each diagonal has the same value • $A = BH$ with reverse identity B and hankel H • MATLAB: `A=toeplitz(v,w)` where `v`$= (\alpha, \gamma, \cdots)$ `w`$= (\alpha, \beta, \cdots)$
Tridiagonal	$A = \begin{pmatrix} \alpha_1 & \beta_1 & & 0 \\ \gamma_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{N-1} \\ 0 & & \gamma_{N-1} & \alpha_N \end{pmatrix}$	• Are Hessenberg matrices • Solution of $A\mathbf{x} = \mathbf{b}$ can be solved using the Thomas algorithm • MATLAB: `A=diag(v)+diag(w,1)...` `+ diag(z,-1)` where `v`$= (\alpha_1, \cdots, \alpha_N)$ `w`$= (\beta_1, \cdots, \beta_{N-1})$ `z`$= (\gamma_1, \cdots, \gamma_{N-1})$
Unit	$a_{ij} = 1$	• MATLAB: `A=ones(N,M)`
Unitriangular	A is (lower or upper) triangular and $a_{ii} = 1$	$\mathbf{det}(A) = 1$
Upper Triangular	$a_{i,j} = 0\,;\ j < i$	• $\mathbf{det}(A) = \prod_{i=1}^{N} a_{ii}$ • Let $D = diag(A)$ and $K = D - A$ $A^{-1} = D^{-1}\left[I + \sum_{\ell=1}^{N-1} \left(KD^{-1}\right)^{\ell}\right]$ • MATLAB: `A=triu(B)` extracts the upper triangle portion of B
Vandermonde	$A = \left(\begin{array}{c\|c\|cc} \alpha_1^{M-1} & & \alpha_1 & 1 \\ \vdots & \cdots & \vdots & \vdots \\ \alpha_N^{M-1} & & \alpha_N & 1 \end{array}\right)$	• If square, $\mathbf{det}(A) = \prod_{i<j}(\alpha_i - \alpha_j)$ • Becomes ill-conditioned for large N • MATLAB: `A=vander(v)` where `v`$= (\alpha_1, \ldots, \alpha_N)$
Zero	$a_{ij} = 0$	• MATLAB: `A=zeros(N,M)`

A.2 Motivation for Matrix Operations from Solution of Equations

Instead of simply taking the various matrix operations at face value with fixed rules, it might be instructive to motivate the development of the matrix algebraic operations through the use of matrix representation of equations' origination from using indexed variables. The aim of this exposition is to illustrate how the various operations, such as matrix products, determinants, adjugates, and inverses, appear to be natural consequences of the operations involved in linear equations.

A.2.1 Matrix Sums, Scalar Products, and Matrix Products

We facilitate the definition of matrix operations by framing it in terms of equations that contain indexed variables. We start with the representation of a set of N linear equations relating M variables $x_1, x_2, \ldots, x_M$ to N variables $y_1, y_2, \ldots, y_N$ given by

$$\begin{aligned} y_1 &= a_{11}x_1 + \cdots + a_{1M}x_M \\ &\vdots \\ y_N &= a_{N1}x_1 + \cdots + a_{NM}x_M \end{aligned}$$

The indexed notation for these equations are given by

$$y_i = \sum_{j=1}^{M} a_{ij}x_j \qquad i = 1, 2, \ldots, N \tag{A.1}$$

By collecting the variables to form matrices:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_M \end{pmatrix} \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1M} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NM} \end{pmatrix}$$

we postulate the matrix representation of (A.1) as

$$\mathbf{y} = A\mathbf{x} \tag{A.2}$$

For instance, consider the set of equations

$$\begin{aligned} y_1 &= x_1 + 3x_2 \\ y_2 &= -x_1 - 2x_2 \end{aligned}$$

then

$$\mathbf{y} = A\mathbf{x} \quad \text{where} \quad A = \begin{pmatrix} 1 & 3 \\ -1 & -2 \end{pmatrix};\ \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix};\ \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

As we proceed from here, we show that the postulated form in (A.2) to represent (A.1) will ultimately result in a definition of matrix products $C = AB$, which is a generalization of (A.2), that is, with $\mathbf{y} = C$ and $\mathbf{x} = B$.

Now let $y_1, \ldots, y_N$ and $z_1, \ldots, z_N$ be related to $x_1, \ldots, x_M$ as follows:

$$y_i = \sum_{j=1}^{M} a_{kj}x_j \qquad \text{and} \qquad z_i = \sum_{j=1}^{M} b_{kj}x_j \qquad i = 1, \ldots, N$$

where a_{ij} and b_{ij} are elements of matrix A and B, respectively.

Let $u_i = y_i + z_i, i = 1, \ldots, N$, then

$$u_i = \sum_{j=1}^{M} a_{ij}x_j + \sum_{j=1}^{M} b_{ij}x_j = \sum_{j=1}^{M} (a_{ki} + b_{ki})\, x_j = \sum_{j=1}^{M} g_{ij}x_j$$

where g_{ij} are the elements of a matrix G. From the rightmost equality, we can then define the **matrix sum** by the following operation:

$$G = A + B \qquad \longleftrightarrow \qquad g_{ij} = a_{ij} + b_{ij} \qquad \begin{matrix} i = 1, \ldots, N \\ j = 1, \ldots, M \end{matrix} \tag{A.3}$$

Next, let $v_i = \alpha y_i, i = 1, \ldots, N$, and $y_i = \sum_{j=1}^{M} a_{kj}x_j$, where α is a scalar multiplier, then

$$v_i = \alpha \sum_{j=1}^{M} a_{ij}x_j = \sum_{j=1}^{M} \alpha a_{ij}x_j = \sum_{j=1}^{M} h_{ij}x_j$$

where h_{ij} are the elements of a matrix H. From the rightmost equality, we can then define the **scalar product** by the following operation:

$$H = \alpha A \qquad \longleftrightarrow \qquad h_{ij} = \alpha a_{ij} \qquad \begin{matrix} i = 1, \ldots, N \\ j = 1, \ldots, M \end{matrix} \tag{A.4}$$

Next, let $w_k = \sum_{i=1}^{N} c_{ki}y_i, k = 1, \ldots, K$, and $y_i = \sum_{j=1}^{M} a_{ij}x_j, i = 1, \ldots, N$, where c_{ki} and a_{ij} are elements of matrices C and A, respectively, then

$$w_k = \sum_{i=1}^{N} c_{ki} \left(\sum_{j=1}^{M} a_{ij}x_j \right) = \sum_{j=1}^{M} \left(\sum_{i=1}^{N} c_{ki}a_{ij} \right) x_j = \sum_{j=1}^{M} f_{kj}x_j$$

where f_{kj} are the elements of a matrix F. From the rightmost equality, we can then define the **matrix product** by the following operation:

$$F = CA \qquad \longleftrightarrow \qquad f_{kj} = \sum_{i=1}^{N} c_{ki}a_{ij} \qquad \begin{matrix} k = 1, \ldots, K \\ j = 1, \ldots, M \end{matrix} \tag{A.5}$$

A.2.2 Determinants, Cofactors, and Adjugates

Let us begin with the case involving two linear equation with two unknowns,

$$\begin{matrix} a_{11}x_1 + a_{12}x_2 & = & b_1 \\ a_{21}x_1 + a_{22}x_2 & = & b_2 \end{matrix} \tag{A.6}$$

One of the unknowns (e.g., x_2) can be eliminated by multiplying the first equation by a_{22} and the second equation by $-a_{12}$, and then adding adding both results. Doing so, we obtain

$$(a_{11}a_{22} - a_{12}a_{21})\, x_1 = a_{22}b_1 - a_{12}b_2 \tag{A.7}$$

We could also eliminate x_1 using a similar procedure. Alternatively, we could simply exchange indices 1 and 2 in (A.7) to obtain

$$(a_{22}a_{11} - a_{21}a_{12})\, x_2 = a_{11}b_2 - a_{21}b_1 \tag{A.8}$$

The coefficients of x_1 and x_2 in (A.7) and (A.8) are essentially the same, which we now define the **determinant** function of a 2×2 matrix,

$$\mathbf{det}\,(M) = \mathbf{det}\begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} = m_{11}m_{22} - m_{12}m_{21} \tag{A.9}$$

Equations (A.7) and (A.8) can be then be combined to yield a matrix equation,

$$\Big(\mathbf{det}(A)\Big)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{A.10}$$

If $\mathbf{det}\,(A) \neq 0$, then we have

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{\mathbf{det}\,(A)}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

and find that the inverse matrix of a 2×2 matrix is given by

$$A^{-1} = \frac{1}{\mathbf{det}\,(A)}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

Next, we look at the case of three equations with three unknowns,

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 & = & b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 & = & b_3 \end{array} \tag{A.11}$$

We can rearrange the first two equations in (A.11) and move terms with x_3 to the other side to mimic (A.6), that is,

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 & = & b_1 - a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 & = & b_2 - a_{23}x_3 \end{array}$$

then using (A.10), we obtain

$$\alpha_{\langle 3\rangle}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}\begin{pmatrix} b_1 - a_{13}x_3 \\ b_2 - a_{23}x_3 \end{pmatrix} \tag{A.12}$$

where

$$\alpha_{\langle 3\rangle} = \mathbf{det}\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

Returning to the third equation in (A.11), we could multiply it by the scalar $\alpha_{\langle 3\rangle}$ to obtain

$$\begin{pmatrix} a_{31} & a_{32} \end{pmatrix}\alpha_{\langle 3\rangle}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \alpha_{\langle 3\rangle}b_3 - a_{33}\alpha_{\langle 3\rangle}x_3 \tag{A.13}$$

We can then substitute (A.12) into (A.13) to obtain

$$\begin{pmatrix} a_{31} & a_{32} \end{pmatrix}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}\begin{pmatrix} b_1 - a_{13}x_3 \\ b_2 - a_{23}x_3 \end{pmatrix} = \alpha_{\langle 3\rangle}b_3 - a_{33}\alpha_{\langle 3\rangle}x_3 \tag{A.14}$$

Next, we note that

$$\begin{array}{rcl} \begin{pmatrix} a_{31} & a_{32} \end{pmatrix}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix} & = & \begin{pmatrix} (a_{31}a_{22} - a_{32}a_{21}) & (-a_{31}a_{12} + a_{32}a_{11}) \end{pmatrix} \\ & = & \begin{pmatrix} -\beta_{\langle 3\rangle} & \gamma_{\langle 3\rangle} \end{pmatrix} \end{array} \tag{A.15}$$

where

$$\beta_{\langle 3\rangle} = \mathbf{det}\begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \quad \text{and} \quad \gamma_{\langle 3\rangle} = \mathbf{det}\begin{pmatrix} a_{11} & a_{31} \\ a_{12} & a_{32} \end{pmatrix}$$

Substituting (A.15) into (A.14) and rearranging to solve for unknown x_3, we obtain

$$\Big(a_{13}\beta_{\langle 3\rangle} - a_{23}\gamma_{\langle 3\rangle} + a_{33}\alpha_{\langle 3\rangle}\Big)x_3 = \beta_{\langle 3\rangle}b_1 - \gamma_{\langle 3\rangle}b_2 + \alpha_{\langle 3\rangle}b_3 \tag{A.16}$$

Looking closer at $\beta_{\langle 3\rangle}$, $\gamma_{\langle 3\rangle}$, and $\alpha_{\langle 3\rangle}$, they are just determinants of three matrix redacts $A_{13\downarrow}$, $A_{23\downarrow}$, and $A_{33\downarrow}$, respectively (where $A_{ij\downarrow}$ are the matrices obtained by removing the i^{th} row and j^{th} column, cf. (1.5)). The determinants of $A_{ij\downarrow}$ are also known as the ij^{th} **minor** of A. We can further incorporate the positive or negative signs appearing in (A.16) with the minors and define them as the **cofactor of** a_{ij}, denoted by $\mathbf{cof}\,(a_{ij})$, and given by

$$\mathbf{cof}(a_{ij}) = (-1)^{i+j}\,\mathbf{det}\,(A_{ij\downarrow}) \tag{A.17}$$

Then we can rewrite (A.16) as

$$\left(\sum_{i=1}^{3} a_{i3}\,\mathbf{cof}\,(a_{i3})\right)x_3 = \begin{pmatrix} \mathbf{cof}\,(a_{13}) & \mathbf{cof}\,(a_{23}) & \mathbf{cof}\,(a_{33}) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{A.18}$$

Instead of applying the same sequence of steps to solve for x_1 and x_2, we just switch indices. Thus to find the equation for x_1, we can exchange the roles of indices 1 and 3 in (A.18). Likewise, for x_2, we can exchange the roles of indices 2 and 3 in (A.18). Doing so, we obtain

$$\left(\sum_{i=1}^{3} a_{i1}\,\mathbf{cof}\,(a_{i1})\right)x_1 = \begin{pmatrix} \mathbf{cof}\,(a_{11}) & \mathbf{cof}\,(a_{21}) & \mathbf{cof}\,(a_{31}) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{A.19}$$

$$\left(\sum_{i=1}^{3} a_{i2}\,\mathbf{cof}\,(a_{i2})\right)x_2 = \begin{pmatrix} \mathbf{cof}\,(a_{12}) & \mathbf{cof}\,(a_{22}) & \mathbf{cof}\,(a_{32}) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{A.20}$$

If we expand the calculations of the coefficients of x_3, x_1 and x_2 in (A.18), (A.19) and (A.20), respectively, they all yield the same sum of six terms, that is,

$$\begin{aligned} \sum_{i=1}^{3} a_{i1}\,\mathbf{cof}\,(a_{i1}) &= \sum_{i=1}^{3} a_{i2}\,\mathbf{cof}\,(a_{i2}) = \sum_{i=1}^{3} a_{i3}\,\mathbf{cof}\,(a_{i3}) \\ &= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} \\ &\quad + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \end{aligned} \tag{A.21}$$

The sum of the six terms in (A.21) can now be defined as the determinant of a 3×3 matrix. By comparing it with the determinant of a 2×2 matrix given in (A.9), we can inductively define the determinants and cofactors for any size matrix as given in Definitions 1.4 and 1.5, respectively.

Based on the definition of determinants and cofactors, we can rewrite (A.10) that was needed for the solution of the size 2 problem as

$$\mathbf{det}(A)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \mathbf{cof}\,(a_{11}) & \mathbf{cof}\,(a_{21}) \\ \mathbf{cof}\,(a_{12}) & \mathbf{cof}\,(a_{22}) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{A.22}$$

Likewise, if we combine (A.18), (A.20), and (A.19) for a size 3 problem, we obtain

$$\mathbf{det}(A)\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} \mathbf{cof}\,(a_{11}) & \mathbf{cof}\,(a_{21}) & \mathbf{cof}\,(a_{31}) \\ \mathbf{cof}\,(a_{12}) & \mathbf{cof}\,(a_{22}) & \mathbf{cof}\,(a_{32}) \\ \mathbf{cof}\,(a_{13}) & \mathbf{cof}\,(a_{23}) & \mathbf{cof}\,(a_{33}) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \tag{A.23}$$

We see that the solution of either case is guaranteed if $\mathbf{det}(A) \neq 0$. From (A.22) and (A.23), we can take the matrix at the right-hand side that premultiplies vector $\mathbf{b}$ in each equation and define them as **adjugates**. They can then be induced to yield definition 1.6 for matrix adjugates.

A.3 Taylor Series Expansion

One key tool in numerical computations is the application of matrix calculus in providing approximations based on the process of **linearization**. The approximation process starts with the **Taylor series expansion** of a multivariable function.

Definition A.1. *Let $f(\mathbf{x})$ be a multivariable function that is sufficiently differentiable; then the* **Taylor series expansion** *of f around a fixed vector $\widehat{\mathbf{x}}$, denoted by* $\text{Taylor}(f, \mathbf{x}, \widehat{\mathbf{x}})$*, is given by*

$$\text{Taylor}(f, \mathbf{x}, \widehat{\mathbf{x}}) = f(\widehat{\mathbf{x}}) + \sum_{K=1}^{\infty} \mathcal{F}_K(f, \mathbf{x}, \widehat{\mathbf{x}}) \tag{A.24}$$

where

$$\mathcal{F}_K(f, \mathbf{x}, \widehat{\mathbf{x}}) = \underbrace{\sum_{k_1,\ldots,k_N \geq 0}}_{\sum_i^N k_i = K} \frac{1}{k_1! \cdots k_N!} \left(\frac{\partial^K f}{\partial x_1^{k_1} \cdots \partial x_N^{k_N}} \right)\Bigg|_{(\mathbf{x}=\widehat{\mathbf{x}})} \prod_{i=1}^{N} (x_i - \widehat{x}_i)^{k_i} \tag{A.25}$$

For $K = 1, 2$, $\mathcal{F}_1$ and $\mathcal{F}_2$ are given by

$$\begin{aligned} \mathcal{F}_1 &= \left(\frac{df}{d\mathbf{x}}\bigg|_{\mathbf{x}=\widehat{\mathbf{x}}} \right)(\mathbf{x} - \widehat{\mathbf{x}}) \\ \mathcal{F}_2 &= (\mathbf{x} - \widehat{\mathbf{x}})^T \left(\frac{1}{2}\,\frac{d^2 f}{d\mathbf{x}^2}\bigg|_{\mathbf{x}=\widehat{\mathbf{x}}} \right)(\mathbf{x} - \widehat{\mathbf{x}}) \end{aligned}$$

THEOREM A.1. *If the series* $\text{Taylor}(f, \mathbf{x}, \widehat{\mathbf{x}})$ *converges for a given* $\mathbf{x}$ *and* $\widehat{\mathbf{x}}$ *then*

$$f(\mathbf{x}) = \text{Taylor}(f, \mathbf{x}, \widehat{\mathbf{x}}) \tag{A.26}$$

PROOF. (See Section A.4.8)

If the series Taylor $(f, \mathbf{x}, \widehat{\mathbf{x}})$ is convergent inside a region $\mathcal{R} = \{\mathbf{x} \mid |\mathbf{x} - \widehat{\mathbf{x}}| < r\}$, where r is called the **radius of convergence**, then $f(\mathbf{x})$ is said to be **analytic in** $\mathcal{R}$.

When $\mathbf{x}$ is equal to $\widehat{\mathbf{x}}$, the Taylor series yields the identity $f(\mathbf{x}) = f(\widehat{\mathbf{x}})$. We expect that as we perturb $\mathbf{x}$ away from $\widehat{\mathbf{x}}$, the terms with $(x_i - \widehat{x}_i)^{k_i}$ will become increasingly significant. However, if we keep $(x_i - \widehat{x}_i)$ sufficiently small, then the terms involving $(x_i - \widehat{x}_i)^{k_i}$ can be made negligible for larger values of $k_i > 0$. Thus a multivariable function can be approximated "locally" by keeping only a finite number of lower order terms of the Taylor series, as long as $\mathbf{x}$ is close to $\widehat{\mathbf{x}}$. We measure "closeness" of two vectors $\mathbf{x}$ and $\widehat{\mathbf{x}}$ by the **Euclidean norm** $\rho(\mathbf{x} - \widehat{\mathbf{x}})$, where

$$\rho(\mathbf{x} - \widehat{\mathbf{x}}) = \sqrt{\sum_{i=1}^{N} (x_i - \widehat{x}_i)^2}$$

The **first-order approximation** of a function $f(\mathbf{x})$ around a small neighborhood of $\widehat{\mathbf{x}}$, that is, $\rho(\mathbf{x} - \widehat{\mathbf{x}}) < \epsilon$, is given by

$$[f_{\text{Lin},\widehat{\mathbf{x}}}(\mathbf{x})] = [f(\widehat{\mathbf{x}})] + \frac{d}{d\mathbf{x}} f\Big|_{\mathbf{x}=\widehat{\mathbf{x}}} (\mathbf{x} - \widehat{\mathbf{x}}) \tag{A.27}$$

Because the right-hand side is a linear function of x_i, the first-order approximation is usually called the **linearized approximation** of $f(\mathbf{x})$, and the approximation process is called the **linearization** of $f(\mathbf{x})$.

The **second-order approximation** of $f(\mathbf{x})$ is given by

$$[f_{\text{Quad},\widehat{\mathbf{x}}}(\mathbf{x})] = [f(\widehat{\mathbf{x}})] + \frac{d}{d\mathbf{x}} f\Big|_{\mathbf{x}=\widehat{\mathbf{x}}} (\mathbf{x} - \widehat{\mathbf{x}}) + \frac{1}{2} (\mathbf{x} - \widehat{\mathbf{x}})^T \left(\frac{d^2}{d\mathbf{x}^2} f \right)\Big|_{\mathbf{x}=\widehat{\mathbf{x}}} (\mathbf{x} - \widehat{\mathbf{x}}) \tag{A.28}$$

where the right-hand side is a quadratic form for x_i. Higher-order approximations are of course possible, but the matrix representations of orders > 2 are much more difficult.

EXAMPLE A.1. Consider the function

$$f(x_1, x_2) = 1 - e^{g(x_1, x_2)}$$

where,

$$g(x_1, x_2) = -4 \left((x_1 - 0.5)^2 + (x_2 + 0.5)^2 \right)$$

A plot of $f(x_1, x_2)$ is shown in Figure A.1.

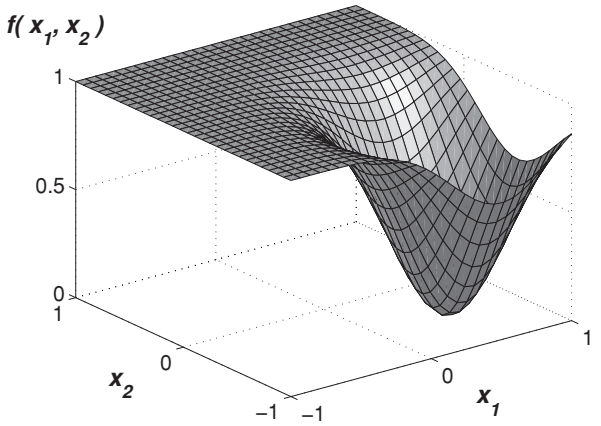


Figure A.1. A plot of $f(\mathbf{x})$ for Example A.1.

The partial derivatives are given by

$$\frac{\partial f}{\partial x_1} = 8e^g\,(x_1 - 0.5) \qquad \frac{\partial f}{\partial x_2} = 8e^g\,(x_2 + 0.5)$$

$$\frac{\partial^2 f}{\partial {x_1}^2} = e^g\left(8 - 64\,(x_1 - 0.5)^2\right) \qquad \frac{\partial^2 f}{\partial {x_2}^2} = e^g\left(8 - 64\,(x_2 + 0.5)^2\right)$$

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1} = -64e^g\,(x_1 - 0.5)\,(x_2 + 0.5)$$

Choosing $\widehat{\mathbf{x}} = (0, 0)$, we have the first-order approximation given by

$$\left[f_{\mathrm{Lin},(0,0)^T}(\mathbf{x})\right] = \left[1 - e^{-2}\right] + 4e^{-2}\begin{pmatrix} -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

or

$$f_{\mathrm{Lin},(0,0)^T}(\mathbf{x}) = \left(1 - e^{-2}\right) + 4e^{-2}\left(-x_1 + x_2\right)$$

and the second-order approximation given by

$$\begin{aligned}\left[f_{\mathrm{Quad},(0,0)^T}(\mathbf{x})\right] \;=\; & \left[1 - e^{-2}\right] + 4e^{-2}\begin{pmatrix} -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ & + 4e^{-2}\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\end{aligned}$$

or

$$f_{\mathrm{Quad},(0,0)^T}(\mathbf{x}) = \left(1 - e^{-2}\right) + 4e^{-2}\left(-x_1 + x_2 - x_1^2 + 4x_1x_2 - x_2^2\right)$$

The first-order approximation is a 2D plane that has the same value as f at $\mathbf{x} = \widehat{\mathbf{x}}$. Conversely, the second-order approximation is a curved surface, which also has the same value as f at $\mathbf{x} = \widehat{\mathbf{x}}$. A plot of the errors resulting from the first-order and second-order approximations are shown in Figure A.2 in a circular region centered at $\widehat{\mathbf{x}} = (0, 0)$. As shown in the plots, the errors present in the second-order approximation are much smaller than the errors present in the first-order approximation.

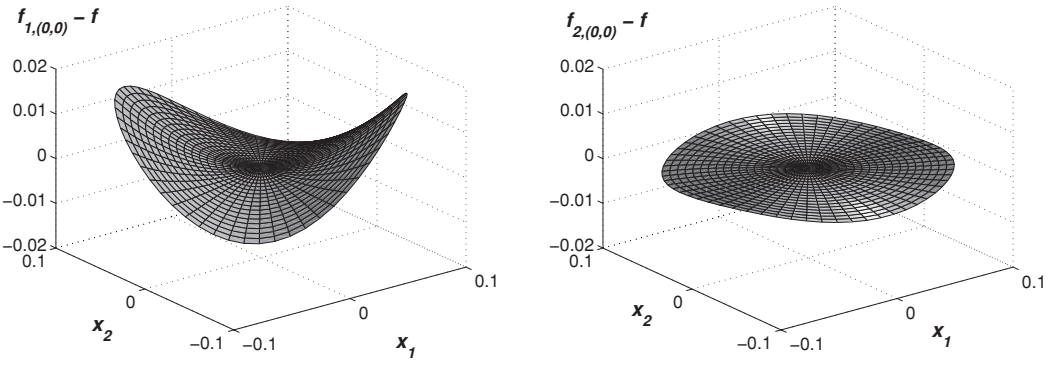


Figure A.2. The errors from f of the first-order approximation (left) and the second-order approximation (right) at $\widehat{\mathbf{x}} = (0, 0)^T$.

From Figure A.1, we can see that minimum value of $f(\mathbf{x})$ occurs at $x_1 = 0.5$ and $x_2 = -0.5$. If we had chosen to expand the Taylor series around the point $\widehat{\mathbf{x}} = (0.5, -0.5)^T$, the gradient will be $df/d\mathbf{x} = (0, 0)$. The Hessian will be given by

$$\frac{d^2}{d\mathbf{x}^2} f = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix}$$

and the second-order approximation is

$$f_{\text{Quad},(0.5,-0.5)^T}(\mathbf{x}) = 4\left((x_1 - 0.5)^2 + (x_2 + 0.5)^2\right)$$

A plot of $f_{\text{Quad},(0.5,-0.5)^T}(\mathbf{x})$ for a region centered at $\widehat{\mathbf{x}} = (0.5, -0.5)^T$ is shown in Figure A.3. Second-order approximations are useful in locating the value of $\mathbf{x}$ that would yield a local minimum for a given scalar function. At the local minimum, the gradient must be a row vector of zeros. Second, if the shape of the curve is strictly concave at a small neighborhood around the minimum point, then a minimum is present. The concavity will depend on whether the Hessian, $d^2 f/d\mathbf{x}^2$, are positive or negative definite.

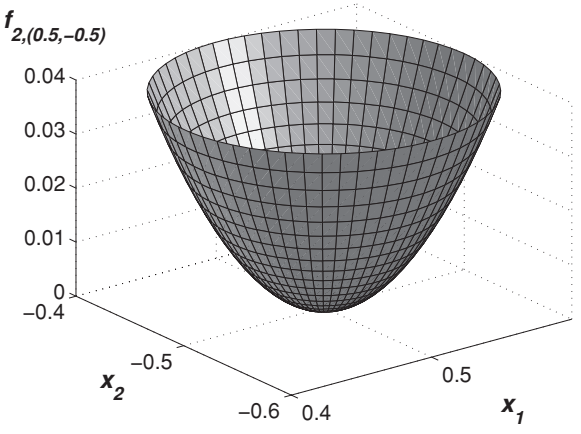


Figure A.3. The second-order approximation at $\widehat{\mathbf{x}} = (0.5, -0.5)^T$.

A.4 Proofs for Lemma and Theorems of Chapter 1

A.4.1 Proof of Properties of Matrix Operations

1. Associative and Distributive Properties.

The proofs are based on the operations given in Table 1.3 plus the associativity of the elements under multiplication or addition. For example,

$$\begin{aligned}(A+(B+C))_{ij} &= a_{ij}+(b_{ij}+c_{ij}) \\ &= (a_{ij}+b_{ij})+c_{ij} = ((A+B)+C)_{ij} \\ \longrightarrow \quad & A+(B+C)=(A+B)+C\end{aligned}$$

For the identity $(AB)\otimes(CD)=(A\otimes C)(B\otimes D)$, let $A[=]m\times p$ and $B[=]p\times n$ and then expand the right-hand side,

$$(A\otimes C)(B\otimes D) = \left(\begin{array}{c|c|c} a_{11}C & \dots & a_{1p}C \\ \hline \vdots & \ddots & \vdots \\ \hline a_{m1}C & \dots & a_{mp}C \end{array}\right)\left(\begin{array}{c|c|c} b_{11}D & \dots & b_{1n}D \\ \hline \vdots & \ddots & \vdots \\ \hline b_{p1}D & \dots & b_{pn}D \end{array}\right)$$

$$= \left(\begin{array}{c|c|c} \sum_{i=1}^{p} a_{1i}b_{i1}CD & \dots & \sum_{i=1}^{p} a_{1i}b_{in}CD \\ \hline \vdots & \ddots & \vdots \\ \hline \sum_{i=1}^{p} a_{mi}b_{i1}CD & \dots & \sum_{i=1}^{p} a_{mi}b_{in}CD \end{array}\right)$$

$$= (AB)\otimes(CD)$$

2. Transposes of Products.

Let $A[=]N\times M$, $B[=]M\times L$, then

$$\begin{aligned}\left((AB)^T\right)_{ij} &= \sum_{m=1}^{M} a_{jm}b_{mi} = \sum_{m=1}^{M} b_{mi}a_{jm} = \left(B^TA^T\right)_{ij} \\ \longrightarrow \quad & (AB)^T = B^TA^T\end{aligned}$$

Let $A[=]N\times M$, $B[=]L\times P$, then

$$\begin{aligned}\left((A\circ B)^T\right)_{ij} &= a_{ji}b_{ji} = \left(A^T\circ B^T\right)_{ij} \\ \longrightarrow \quad & (A\circ B)^T = A^T\circ B^T\end{aligned}$$

$$(A\otimes B)^T = \left(\begin{array}{c|c|c} a_{11}B & \cdots & a_{1M}B \\ \hline \vdots & \ddots & \vdots \\ \hline a_{N1}B & \cdots & a_{NM}B \end{array}\right)^T = \left(\begin{array}{c|c|c} a_{11}B^T & \cdots & a_{N1}B^T \\ \hline \vdots & \ddots & \vdots \\ \hline a_{1M}B^T & \cdots & a_{MN}B^T \end{array}\right)$$

$$= A^T\otimes B^T$$

3. Inverse of Matrix Products and Kronecker Products.

Let $C = B^{-1}A^{-1}$, then

$$\begin{aligned} C(AB) &= \left(B^{-1}A^{-1}\right)(AB) = B^{-1}B = I \\ (AB)C &= (AB)\left(B^{-1}A^{-1}\right) = BB^{-1} = I \end{aligned}$$

Thus $C = B^{-1}A^{-1}$ is the inverse of AB.

For the inverse of Kronecker products use the associativity property,

$$(A \otimes C)(B \otimes D) = (AB) \otimes (CD)$$

then,

$$\begin{aligned} (A \otimes B)(A^{-1} \otimes B^{-1}) &= AA^{-1} \otimes BB^{-1} = I \\ (A^{-1} \otimes B^{-1})(A \otimes B) &= A^{-1}A \otimes B^{-1}B = I \end{aligned}$$

Thus

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

4. Vectorization of Sums and Products.

Let $A, B, C[=]N \times M$ and $C = A + B$

$$\begin{aligned} \text{vec}(C)_{(j-1)N+i} &= c_{ij} = a_{ij} + b_{ij} = \text{vec}(A)_{(j-1)N+i} + \text{vec}(B)_{(j-1)N+i} \\ &\longrightarrow \quad \text{vec}(A+B) = \text{vec}(A) + \text{vec}(B) \end{aligned}$$

Let $M_{(\bullet,j)}$ denote the j^{th} column of any matrix M, then

$$(XC)_{(\bullet,j)} = (X)C_{(\bullet,j)} = \begin{pmatrix} X_{(\bullet,1)} & \cdots & X_{(\bullet,r)} \end{pmatrix} \begin{pmatrix} c_{1j} \\ \vdots \\ c_{rj} \end{pmatrix} = \sum_{i=1}^{r} c_{ij} X_{(\bullet,i)}$$

Extending this to BXC,

$$\begin{aligned} (BXC)_{(\bullet,j)} &= (BX)C_{(\bullet,j)} = B\left(XC_{(\bullet,j)}\right) = \sum_{i=1}^{r} c_{ij} BX_{(\bullet,i)} \\ &= \begin{pmatrix} c_{1j}B & c_{2j}B & \cdots & c_{rj}B \end{pmatrix} \begin{pmatrix} X_{(\bullet,1)} \\ X_{(\bullet,2)} \\ \vdots \\ X_{(\bullet,r)} \end{pmatrix} \\ &= \begin{pmatrix} c_{1j}B & c_{2j}B & \cdots & c_{rj}B \end{pmatrix} \text{vec}(X) \end{aligned}$$

Collecting these into a column,

$$\begin{aligned} \text{vec}(BXC) &= \begin{pmatrix} (BXC)_{(\bullet,1)} \\ (BXC)_{(\bullet,2)} \\ \vdots \\ (BXC)_{(\bullet,s)} \end{pmatrix} = \begin{pmatrix} c_{11}B & c_{21}B & \cdots & c_{r1}B \\ c_{12}B & c_{22}B & \cdots & c_{r2}B \\ \vdots & \vdots & \ddots & \vdots \\ c_{1s}B & c_{21}B & \cdots & c_{rs}B \end{pmatrix} \text{vec}(X) \\ &= \left(C^T \otimes B\right) \text{vec}(X) \end{aligned}$$

5. Reversible Operations.

$$\begin{aligned} \left((A^T)^T\right)_{ij} &= A_{ij} \\ \longrightarrow \qquad & \left(A^T\right)^T = A \end{aligned}$$

Let $C = \left(A^{-1}\right)^{-1}$, then

$$\begin{aligned} CA^{-1} = A^{-1}C &= I \\ \longrightarrow \qquad & C = \left(A^{-1}\right)^{-1} = A \end{aligned}$$

A.4.2 Proof of Properties of Determinants

1. Determinant of Products.

Let $C = AB$, then $c_{ik_i} = \sum_{\ell_i=1}^{n} a_{i\ell_i} b_{\ell_i k_i}$. Using (1.10),

$$\begin{aligned} \mathbf{det}(C) &= \sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) \prod_{i=1}^{n} c_{i,k_i} \\ &= \sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) \left(\sum_{\ell_1=1}^{n} a_{1\ell_1} b_{\ell_1 k_1}\right) \cdots \left(\sum_{\ell_n=1}^{n} a_{n\ell_n} b_{\ell_n k_n}\right) \\ &= \sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \sum_{\ell_1=1}^{n} \cdots \sum_{\ell_n=1}^{n} \epsilon\left(k_1, \ldots, k_n\right) \prod_{i=1}^{n} a_{i\ell_i} \prod_{j=1}^{n} b_{\ell_i k_i} \\ &= \sum_{\ell_1=1}^{n} \cdots \sum_{\ell_n=1}^{n} (a_{1\ell_1} \cdots a_{n\ell_n}) \sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) (b_{\ell_1 k_1} \cdots b_{\ell_n k_n}) \end{aligned}$$

but

$$\sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) (b_{\ell_1 k_1} \cdots b_{\ell_n k_n}) = 0 \quad \text{if } \ell_i = \ell_j$$

so

$$\mathbf{det}(C) = \sum_{\ell_1 \neq \ell_2 \neq \cdots \neq \ell_n} (a_{1\ell_1} \cdots a_{n\ell_n}) \sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) (b_{\ell_1 k_1} \cdots b_{\ell_n k_n})$$

The inner summation can be further reindexed as

$$\sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) \epsilon\left(\ell_1, \ldots, \ell_n\right) (b_{1k_1} \cdots b_{nk_n})$$

and the determinant of C then becomes

$$\begin{aligned} \mathbf{det}(C) &= \left(\sum_{\ell_1 \neq \ell_2 \neq \cdots \neq \ell_n} \epsilon\left(\ell_1, \ldots, \ell_n\right) \prod_{i=1}^{n} a_{i\ell_i}\right) \\ &\qquad \times \left(\sum_{k_1 \neq k_2 \neq \cdots \neq k_n} \epsilon\left(k_1, \ldots, k_n\right) \prod_{j=1} b_{jk_j}\right) \\ &= \mathbf{det}(A)\, \mathbf{det}(B) \end{aligned}$$

2. Determinant of Triangular Matrices.

   For 2×2 triangular matrices,

$$\mathbf{det}\begin{pmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{pmatrix} = u_{11}u_{22} \quad ; \quad \mathbf{det}\begin{pmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{pmatrix} = \ell_{11}\ell_{22}$$

$$\mathbf{det}\begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} = d_{11}d_{22}$$

   Then using induction and row expansion formula (1.12), the result can be proved for any size N.

3. Determinant of Transposes.

   For a 2×2 matrix, that is,

$$\mathbf{det}\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21} = \mathbf{det}\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}$$

   By induction, and using (1.12), the result can be shown to be true for matrices of size N.

4. Determinant of Inverses.

   Because $A^{-1}A = I$, we can take the determinant of both sides, and use the property of determinant of products. Thus

$$\mathbf{det}\left(A^{-1}A\right) = \mathbf{det}\left(A^{-1}\right)\mathbf{det}\left(A\right) = 1 \;\rightarrow\; \mathbf{det}\left(A^{-1}\right) = \frac{1}{\mathbf{det}\left(A\right)}$$

5. Determinant of Matrices with Permuted Columns.

   Let

$$A_K = \begin{pmatrix} A_{\bullet,k_1} & \cdots & A_{\bullet,k_N} \end{pmatrix} = A\begin{pmatrix} \mathbf{e}_{k_1} & \cdots & \mathbf{e}_{k_N} \end{pmatrix}$$

   Using (1.10),

$$\mathbf{det}\begin{pmatrix} \mathbf{e}_{k_1} & \cdots & \mathbf{e}_{k_N} \end{pmatrix} = \epsilon\left(K\right)$$

   Then using the property of determinant of products,

$$\mathbf{det}\left(A_K\right) = \mathbf{det}\left(A\right)\mathbf{det}\begin{pmatrix} \mathbf{e}_{k_1} & \cdots & \mathbf{e}_{k_N} \end{pmatrix} = \epsilon\left(K\right)\mathbf{det}\left(A\right)$$

6. Determinant of Scaled Columns.

$$\begin{pmatrix} \beta_1\, a_{11} \\ \vdots \\ \beta_1\, a_{N1} \end{pmatrix} \cdots \begin{pmatrix} \beta_N\, a_{1N} \\ \vdots \\ \beta_N\, a_{NN} \end{pmatrix} = \begin{pmatrix} a_{11} \\ \vdots \\ a_{N1} \end{pmatrix} \cdots \begin{pmatrix} a_{1N} \\ \vdots \\ a_{NN} \end{pmatrix} \begin{pmatrix} \beta_1 & & 0 \\ & \ddots & \\ 0 & & \beta_N \end{pmatrix}$$

Using the properties of determinant of products and determinant of diagonal matrices,

$$\mathbf{det}\begin{pmatrix} \beta_1\, a_{11} & \left|\right. & \beta_N\, a_{1N} \\ \vdots & \cdots & \vdots \\ \beta_1\, a_{N1} & \left|\right. & \beta_N\, a_{NN} \end{pmatrix} = \left(\prod_{i=1}^{N} \beta_i\right) \mathbf{det}\begin{pmatrix} a_{11} & \left|\right. & a_{1N} \\ \vdots & \cdots & \vdots \\ a_{N1} & \left|\right. & a_{NN} \end{pmatrix}$$

7. Multilinearity Property.

Let $a_{ij} = v_{ij} = w_{ij}$, $j \neq k$ and $v_{ik} - x_i$, $w_{ik} = y_i, a_{ik} = x_i + y_i$ for $i, j = 1, 2, \ldots, n$, By expanding along the k^{th} column,

$$\begin{aligned} \mathbf{det}\Big(\, A \,\Big) &= \sum_{i=1}^{n} (x_i + y_i)\ \mathbf{cof}\,(a_{ik}) \\ &= \sum_{i=1}^{n} x_i\ \mathbf{cof}\,(v_{ik}) + \sum_{i=1}^{n} y_i\ \mathbf{cof}\,(w_{ik}) \\ &= \mathbf{det}\Big(\, V \,\Big) + \mathbf{det}\Big(\, W \,\Big) \end{aligned}$$

8. Determinant when $\sum_{i=1}^{n} \gamma_i A_{i,\bullet} = 0$, for some $\gamma_k \neq 0$.

Let the elements of matrix $V(k, \gamma)$ be the same as the identity matrix I except for the k^{th} row replaced by $\gamma_1, \ldots \gamma_N$, that is,

$$V(k,\gamma) = \begin{pmatrix} 1 & & & \gamma_1 & & 0 & \\ & \ddots & & \vdots & & & \\ & & 1 & \gamma_{k-1} & & & \\ & & & \gamma_k & & & \\ & & & \gamma_{k+1} & 1 & & \\ & 0 & & \vdots & & \ddots & \\ & & & \gamma_N & & & 1 \end{pmatrix}$$

where $\gamma_k \neq 0$. Then evaluating the determinant by expanding along the k^{th} row,

$$\mathbf{det}\Big(\, V(k,\gamma) \,\Big) = \gamma_k$$

Postmultiplying A by $V(k, \gamma)$, we have

$$\begin{aligned} A\, V(k,\gamma) &= \Big(A_{\bullet,1} \;\Big|\; \cdots \;\Big|\; A_{\bullet,(k-1)} \;\Big|\; \Big(\textstyle\sum_{j=1}^{N} \gamma_i A_{\bullet,j}\Big) \;\Big|\; A_{\bullet,(k+1)} \;\Big|\; \cdots \;\Big|\; A_{\bullet,N}\Big) \\ &= \big(A_{\bullet,1} \;\big|\; \cdots \;\big|\; A_{\bullet,(k-1)} \;\big|\; \mathbf{0} \;\big|\; A_{\bullet,(k+1)} \;\big|\; \cdots \;\big|\; A_{\bullet,N}\big) \end{aligned}$$

Taking the determinant of both sides, we get $\mathbf{det}(A)\gamma_k = 0$. Because $\gamma_k \neq 0$, it must be that $\mathbf{det}(A) = 0$.

A.4.3 Proof of Matrix Inverse Formula (1.16)

Let $B = A\ \mathbf{adj}(A)$, then

$$b_{ij} = \sum_{\ell=1}^{N} a_{i\ell}\ \mathbf{cof}(a_{j\ell})$$

Using (1.13), b_{ij} is the determinant of a matrix formed from A except that the j^{th} row is replaced by the i^{th} row of A, that is,

$$b_{ij} = \mathbf{det}\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ & \vdots & \\ a_{i1} & \cdots & a_{iN} \\ & \vdots & \\ a_{i1} & \cdots & a_{iN} \\ & \vdots & \\ a_{N1} & \cdots & a_{NN} \end{pmatrix} \longleftarrow j^{\text{th}} \text{ row}$$

We can use the property of determinant of matrices with linearly dependent rows to conclude that $b_{ij} = 0$ when $i \neq j$, and $b_{ii} = \mathbf{det}\,(A)$, that is,

$$A\,\mathbf{adj}(A) = B = \begin{pmatrix} \mathbf{det}\left(A\right) & & 0 \\ & \ddots & \\ 0 & & \mathbf{det}\left(A\right) \end{pmatrix} = \mathbf{det}\left(A\right)\, I$$

or

$$A\left(\frac{1}{\mathbf{det}\left(A\right)}\,\mathbf{adj}(A)\right) = I$$

Using a similar approach, one can show that

$$\left(\frac{1}{\mathbf{det}\left(A\right)}\,\mathbf{adj}(A)\right) A = I$$

Thus

$$A^{-1} = \frac{1}{\mathbf{det}\left(A\right)}\,\mathbf{adj}(A)$$

A.4.4 Proof of Cramer's Rule

Using $A^{-1} = \mathbf{adj}(A)/\mathbf{det}(A)$,

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \frac{1}{\mathbf{det}(A)} \begin{pmatrix} \mathbf{cof}(a_{11}) & \cdots & \mathbf{cof}(a_{N1}) \\ \vdots & \ddots & \vdots \\ \mathbf{cof}(a_{1N}) & \cdots & \mathbf{cof}(a_{NN}) \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{pmatrix}$$

Thus, for the k^{th} entry in x,

$$x_k = \frac{\sum_{j=1}^{n} b_j \mathbf{cof}(a_{kj})}{\mathbf{det}(A)}$$

The numerator is just the determinant of a matrix, $A_{[k,\mathbf{b}]}$, which is obtained from A with k^{th} column replaced by $\mathbf{b}$.

A.4.5 Proof of Block Matrix Properties

1. Block Matrix Multiplication (Equation (1.30)).

   The result can be shown directly by using the definition of matrix multiplication as given in (A.5).

2. Block Matrix Determinants (Equations (1.31), (1.32) and (1.33)).

   (a) Proof of (1.31):

$$\left[\ \mathbf{det}\left(\begin{array}{c|c} A & 0 \\ \hline C & D \end{array}\right) = \mathbf{det}(A)\mathbf{det}(D)\ \right]$$

   Equation (1.31) is true for $A = (a)\,[=]1 \times 1$, that is,

$$\mathbf{det}\left(\begin{array}{c|c} a & 0 \\ \hline C & B \end{array}\right) = a\ \mathbf{det}(B)$$

   Next, assume that (1.31) is true for $A[=](n-1)\times(n-1)$. Let

$$Z = \left(\begin{array}{c|c} G & 0 \\ \hline C & B \end{array}\right)$$

   with $G[=]n \times n$. By expanding along the first row,

$$\mathbf{det}\left(Z\right) = \sum_{j=1}^{n+m} z_{1j}\mathbf{cof}(z_{1j})$$

   where

$$\begin{aligned} z_{1j} &= \begin{cases} g_{1j} & \text{if } j \le n \\ 0 & j > n \end{cases} \\ \mathbf{cof}(z_{1j}) &= (-1)^{1+j}\mathbf{det}\left(\begin{array}{c|c} G_{1j\downarrow} & 0 \\ \hline C_{\bullet j\downarrow} & B \end{array}\right) \quad , \ j \le n \end{aligned}$$

   then

$$\mathbf{det}\left(Z\right) = \sum_{j=1}^{n} g_{1j}\mathbf{cof}(g_{1j})\mathbf{det}\left(B\right) = \mathbf{det}\left(G\right)\mathbf{det}\left(B\right)$$

   Then (1.31) is proved by induction.

   (b) Proof of (1.32): (assuming A nonsingular)

   Using (1.30), with A nonsingular,

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)\left(\begin{array}{c|c} I & -A^{-1}B \\ \hline 0 & I \end{array}\right) = \left(\begin{array}{c|c} A & 0 \\ \hline C & D - CA^{-1}B \end{array}\right)$$

   Applying property of determinant of products and (1.31),

$$\mathbf{det}\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)\mathbf{det}\left(I\right) = \mathbf{det}(A)\mathbf{det}\left(D - CA^{-1}B\right)$$

(c) Proof of (1.33): (assuming D nonsingular)
Using (1.30), with D nonsingular,

$$\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)\left(\begin{array}{c|c} I & 0 \\ \hline -D^{-1}C & I \end{array}\right) = \left(\begin{array}{c|c} A - BD^{-1}C & B \\ \hline 0 & D \end{array}\right)$$

Applying property of determinants of products and transposes and (1.31),

$$\mathbf{det}\left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right)\mathbf{det}(I) = \mathbf{det}D\ \mathbf{det}\left(A - BD^{-1}C\right)$$

3. Block Matrix Inverse (Equation (1.34)).

$$\begin{aligned} AW + BY &= AA^{-1}\left(I + B\Gamma^{-1}CA^{-1}\right) + B\left(-\Gamma^{-1}CA^{-1}\right) \\ &= I + B\left(\Gamma^{-1}CA^{-1}\right) - B\left(\Gamma^{-1}CA^{-1}\right) = I \end{aligned}$$

$$AX + BZ = -AA^{-1}B\Gamma^{-1} + B\Gamma^{-1} = 0$$

$$\begin{aligned} CW + DY &= CA^{-1}\left(I + B\Gamma^{-1}CA^{-1}\right) + D\left(-\Gamma^{-1}CA^{-1}\right) \\ &= CA^{-1} + CA^{-1}B\Gamma^{-1}CA^{-1} - D\Gamma^{-1}CA^{-1} \\ &= CA^{-1} + \left(CA^{-1}B - D\right)\Gamma^{-1}CA^{-1} \\ &= CA^{-1} - \Gamma\Gamma^{-1}CA^{-1} = 0 \end{aligned}$$

$$\begin{aligned} CX + DZ &= -CA^{-1}B\Gamma^{-1} + D\Gamma^{-1} \\ &= \left(D - CA^{-1}B\right)\Gamma^{-1} \\ &= \Gamma\Gamma^{-1} = I \end{aligned}$$

A.4.6 Proof of Derivative of Determinants

To show the formula for the derivative of a determinant, we can use the definition of a determinant (1.10),

$$\begin{aligned} \frac{d}{dt}\left(\mathbf{det}\,(A)\right) &= \frac{d}{dt}\sum_{k_1\neq k_2\neq\cdots\neq k_N} \epsilon\left(k_1,\ldots,k_N\right) a_{1,k_1}\, a_{2,k_2}\, \ldots\, a_{N,k_N} \\ &= \sum_{k_1\neq k_2\neq\cdots\neq k_N} \epsilon\left(k_1,\ldots,k_N\right)\left(\frac{d}{dt}a_{1,k_1}\right)\, a_{2,k_2}\, \ldots\, a_{N,k_N} \\ &\quad + \ldots + \sum_{k_1\neq k_2\neq\cdots\neq k_N} \epsilon\left(k_1,\ldots,k_N\right)\, a_{1,k_1}\, a_{2,k_2}\, \ldots\, \left(\frac{d}{dt}a_{N,k_N}\right) \\ &= \sum_{n=1}^{N} \mathbf{det}\left(\widehat{A}_{\langle n\rangle}\right) \end{aligned}$$

where

$$\mathbf{det}\left(\widehat{A}_{\langle n\rangle}\right) = \sum_{k_1\neq k_2\neq\cdots\neq k_N} \epsilon\left(k_1,\ldots,k_N\right)\, a_{1,k_1}\, \ldots\, \left(\frac{d}{dt}a_{n,k_n}\right)\, \ldots\, a_{N,k_N}$$

A.4.7 Proofs of Matrix Derivative Formulas (Lemma 1.6)

1. Proof of (1.49): [$d(A\mathbf{x})/d\mathbf{x} = A$]

   Let $N = 1$. Then with $\mathbf{x} = (x_1)$ and $A = (a_{11}, \ldots, a_{M1})^T$,

   $$\frac{d}{d\mathbf{x}} A\mathbf{x} = \begin{pmatrix} d(a_{11}x_1)/dx_1 \\ \vdots \\ d(a_{m1}x_1)/dx_1 \end{pmatrix} = A$$

   Assume (1.49) is true for $\widehat{A}[=]M \times (N-1)$ and $\widehat{\mathbf{x}}[=](N-1) \times 1$. Let

   $$A = \left(\begin{array}{c|c} \widehat{A} & \mathbf{v} \end{array} \right) \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} \widehat{\mathbf{x}} \\ \hline \alpha \end{pmatrix}$$

   where $\mathbf{v}[=]N \times 1$ and α is a scalar. Then

   $$\begin{aligned} \frac{d}{d\mathbf{x}} A\mathbf{x} &= \frac{d}{d\mathbf{x}} \left(\begin{array}{c|c} \widehat{A} & \mathbf{v} \end{array} \right) \begin{pmatrix} \widehat{\mathbf{x}} \\ \hline \alpha \end{pmatrix} \\ &= \frac{d}{d\mathbf{x}} \left(\widehat{A}\widehat{\mathbf{x}} + \mathbf{v}\alpha \right) \\ &= \left(\begin{array}{c|c} \frac{d}{d\widehat{\mathbf{x}}} \left(\widehat{A}\widehat{\mathbf{x}} + \mathbf{v}\alpha \right) & \frac{\partial}{\partial \alpha} \left(\widehat{A}\widehat{\mathbf{x}} + \mathbf{v}\alpha \right) \end{array} \right) \\ &= \left(\begin{array}{c|c} \widehat{A} & \mathbf{v} \end{array} \right) = A \end{aligned}$$

   Thus equation (1.49) can be shown to be true for any N by induction.

2. Proof of (1.50): [$d(\mathbf{x}^T A\mathbf{x})/d\mathbf{x} = \mathbf{x}^T \left(A^T + A \right)$]

   Let $N = 1$, then with $\mathbf{x} = (x_1)$ and $A = (a_{11})$,

   $$\frac{d}{d\mathbf{x}} \mathbf{x}^T A\mathbf{x} = 2x_1 a_{11} = \mathbf{x}^T \left(A^T + A \right)$$

   Assume that (1.50) is true for $\widehat{A}[=](N-1) \times (N-1)$ and $\widehat{\mathbf{x}}[=](N-1) \times 1$. Let

   $$A = \left(\begin{array}{c|c} \widehat{A} & \mathbf{v} \\ \hline \mathbf{w}^T & \beta \end{array} \right) \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} \widehat{\mathbf{x}} \\ \hline \alpha \end{pmatrix}$$

   where $\mathbf{v}[=](N-1) \times 1$, $\mathbf{w}[=](N-1) \times 1$, and α, β are scalars. Then

   $$\begin{aligned} \frac{d}{d\mathbf{x}} \mathbf{x}^T A\mathbf{x} &= \frac{d}{d\mathbf{x}} \left(\widehat{\mathbf{x}}^T \widehat{A}\widehat{\mathbf{x}} + \alpha \left(\mathbf{w} + \mathbf{v} \right)^T \widehat{\mathbf{x}} + \alpha^2 \beta \right) \\ &= \left(\begin{array}{c|c} \widehat{\mathbf{x}}^T \left(\widehat{A}^T + \widehat{A} \right) + \alpha(\mathbf{w}^T + \mathbf{v}^T) & \widehat{\mathbf{x}}^T \left(\mathbf{w} + \mathbf{v} \right) + 2\alpha\beta \end{array} \right) \\ &= \left(\begin{array}{c|c} \widehat{\mathbf{x}}^T & \alpha \end{array} \right) \left(\begin{array}{c|c} \widehat{A}^T + \widehat{A} & \mathbf{w} + \mathbf{v} \\ \hline \mathbf{w}^T + \mathbf{v}^T & 2\beta \end{array} \right) \\ &= \mathbf{x}^T \left(\left(\begin{array}{c|c} \widehat{A}^T & \mathbf{w} \\ \hline \mathbf{v}^T & \beta \end{array} \right) + \left(\begin{array}{c|c} \widehat{A} & \mathbf{v} \\ \hline \mathbf{w}^T & \beta \end{array} \right) \right) \\ &= \mathbf{x}^T \left(A^T + A \right) \end{aligned}$$

where we used the fact that $\mathbf{x}^T\mathbf{v}$ and $\mathbf{x}^T\mathbf{w}$ are symmetric. Thus equation (1.50) can be shown to be true for any N by induction.

3. Proof of (1.51): $\left[\ d^2(\mathbf{x}^T A\mathbf{x})/d\mathbf{x}^2 = A + A^T\ \right]$

$$\frac{d^2}{d\mathbf{x}^2}\left(\mathbf{x}^T A\mathbf{x}\right) = \frac{d}{d\mathbf{x}}\left[\frac{d}{d\mathbf{x}}\mathbf{x}^T A\mathbf{x}\right]^T = \frac{d}{d\mathbf{x}}\left[\mathbf{x}^T\left(A^T + A\right)\right]^T = A + A^T$$

A.4.8 Proof of Taylor Series Expansion (theorem A.1)

Let $f(\mathbf{x})$ be set equal to a power series given by

$$f(\mathbf{x}) = a_0 + \sum_{K=1}^{\infty} \mathcal{S}_K(f, \mathbf{x}, \widehat{\mathbf{x}})$$

where

$$\mathcal{S}_K(f, \mathbf{x}, \widehat{\mathbf{x}}) = \underbrace{\sum_{k_1 \geq 0} \cdots \sum_{k_N \geq 0}}_{\sum_i^N k_i = K} a_{k_1,\cdots,k_N} \prod_{i=1}^{N} (x_i - \widehat{x}_i)^{k_i}$$

At $\mathbf{x} = \widehat{\mathbf{x}}$, we see that $\mathcal{S}_K(f, \widehat{\mathbf{x}}, \widehat{\mathbf{x}}) = 0$ for $K > 0$, or

$$f(\widehat{\mathbf{x}}) = a_0$$

Then, for a fixed value of $k_1, \ldots, k_N$,

$$\frac{\partial^K f}{\partial x_1^{k_1} \cdots \partial x_N^{k_N}} = (k_1! \cdots k_N!)\, a_{k_1,\cdots,k_N} + \text{ terms involving } \prod_{i=1}^{N} (x_i - \widehat{x}_i)^{\beta_i}\Bigg|_{\sum_{i=1}^N \beta_i > 0}$$

After setting $\mathbf{x} = \widehat{\mathbf{x}}$ and rearranging, we have

$$a_{k_1,\cdots,k_N} = \frac{1}{k_1! \cdots k_N!}\left(\frac{\partial^K f}{\partial x_1^{k_1} \cdots \partial x_N^{k_N}}\right)\Bigg|_{(\mathbf{x}=\widehat{\mathbf{x}})}$$

Thus we find that

$$\mathcal{S}_K(f, \mathbf{x}, \widehat{\mathbf{x}}) = \mathcal{F}_K(f, \mathbf{x}, \widehat{\mathbf{x}})$$

with $\mathcal{F}_K$ given in (A.25)

A.4.9 Proof of Sufficient Conditions for Local Minimum (Theorem 1.1)

Let $df/d\mathbf{x} = 0$ at $\mathbf{x} = \mathbf{x}^*$. Then using the second-order Taylor approximation around a perturbation point $(\mathbf{x}^* + \Delta\mathbf{x})$,

$$f(\mathbf{x}^* + \Delta\mathbf{x}) = f(\mathbf{x}^*) + \frac{d}{d\mathbf{x}}f\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \widehat{\mathbf{x}}) + \frac{1}{2}(\Delta\mathbf{x})^T\left(\frac{d^2}{d\mathbf{x}^2}f\right)\Big|_{\mathbf{x}=\mathbf{x}^*}(\Delta\mathbf{x})$$

becomes

$$f(\mathbf{x}^* + \Delta\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}(\Delta\mathbf{x})^T \left(\frac{d^2}{d\mathbf{x}^2} f\right)\Big|_{\mathbf{x}=\mathbf{x}^*} (\Delta\mathbf{x})$$

With the additional condition that the Hessian is positive definite, that is,

$$(\Delta\mathbf{x})^T \left(\frac{d^2}{d\mathbf{x}^2} f\right)\Big|_{\mathbf{x}=\mathbf{x}^*} (\Delta\mathbf{x}) > 0 \qquad \Delta\mathbf{x} \neq \mathbf{0}$$

then

$$f(\mathbf{x}^* + \Delta\mathbf{x}) > f(\mathbf{x}^*) \qquad \text{for all } \Delta\mathbf{x} \neq \mathbf{0}$$

which means that $\mathbf{x}^*$ satisfying both (1.43) and (1.44) are sufficient conditions for $\mathbf{x}^*$ to be a local minimum.

A.5 Positive Definite Matrices

We have seen in Section 1.5.2 that the Hessian of a multivariable function is crucial to the determination of the presence of a local minima or maxima. In this section, we establish an important property of square matrices called **positive definiteness**.

Definition A.2. *Let $f(\mathbf{x})$ be a real-valued multivariable function such that $f(0) = 0$. Then $f(\mathbf{x})$ is* **positive definite** *if*

$$f(\mathbf{x}) > 0 \quad \textit{for all } \mathbf{x} \neq 0 \tag{A.29}$$

and $f(\mathbf{x})$ is **positive semi-definite** *if*

$$f(\mathbf{x}) \geq 0 \quad \textit{for all } \mathbf{x} \tag{A.30}$$

For the special case in which $f(\mathbf{x})$ is a real-valued function given by

$$f(\mathbf{x}) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} \overline{x}_i x_j \tag{A.31}$$

where $\overline{x}_i$ is the complex conjugate of x_i, (A.31) can be represented by

$$[f(\mathbf{x})] = \mathbf{x}^* A \mathbf{x} \tag{A.32}$$

or

$$[f(\mathbf{x})] = \mathbf{x}^* Q \mathbf{x} \tag{A.33}$$

where Q is the Hermitian component of A, that is, $Q = (A + A^*)/2$. To see that (A.32) and (A.33) are equivalent, note that $[f]$ is a real-valued 1×1 matrix that is equal to its conjugate transpose, that is,

$$\begin{aligned} \mathbf{x}^* A \mathbf{x} &= (\mathbf{x}^* A \mathbf{x})^* \\ &= \mathbf{x}^* A^* \mathbf{x} \end{aligned}$$

Then adding $\mathbf{x}^* A \mathbf{x}$ to both sides and dividing by two,

$$\mathbf{x}^* A \mathbf{x} = \frac{1}{2} \mathbf{x}^* (A + A^*) \mathbf{x} = \mathbf{x}^* Q \mathbf{x}$$

Definition A.3. *An $N \times N$ matrix A is* **positive definite**, *denoted* $(A > 0)$, *if*

$$\mathbf{x}^* A\mathbf{x} > 0 \quad \textit{for all } \mathbf{x} \neq 0 \tag{A.34}$$

and A is **positive semi-definite** *if*

$$\mathbf{x}^* A\mathbf{x} \geq 0 \quad \textit{for all } \mathbf{x} \tag{A.35}$$

EXAMPLE A.2. Let $N = 2$, and

$$[f(\mathbf{x})] = \mathbf{x}^* Q\mathbf{x}$$

where $Q = Q^* = (A + A^*)/2$: Expanding the quadratic form in terms of Q and complete the squares,

$$\begin{aligned}
\mathbf{x}^* Q\mathbf{x} &= \begin{pmatrix} \bar{x}_1 & \bar{x}_2 \end{pmatrix} \begin{pmatrix} q_{11} & q_{12} \\ \bar{q}_{12} & q_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&= q_{11}\bar{x}_1 x_1 + \bar{q}_{12}\bar{x}_2 x_1 + q_{12}\bar{x}_1 x_2 + q_{22}\bar{x}_2 x_2 \\
&= q_{11}\left(\bar{x}_1 x_1 + \frac{\bar{q}_{12}}{q_{11}}\bar{x}_2 x_1 + \frac{q_{12}}{q_{11}}\bar{x}_1 x_2 + \frac{q_{12}\bar{q}_{12}}{q_{11}^2}\bar{x}_2 x_2\right) - \frac{q_{12}\bar{q}_{12}}{q_{11}^2}\bar{x}_2 x_2 + q_{22}\bar{x}_2 x_2 \\
&= q_{11}\left(\overline{x_1 + \frac{q_{12}}{q_{11}}x_2}\right)\left(x_1 + \frac{q_{12}}{q_{11}}x_2\right) + \frac{q_{11}q_{22} - \bar{q}_{12}q_{12}}{q_{11}}\bar{x}_2 x_2 \\
&= q_{11}\bar{y}y + \frac{\mathbf{det}(Q)}{q_{11}}\bar{x}_2 x_2
\end{aligned}$$

where

$$y = x_1 + \frac{q_{12}}{q_{11}}x_2$$

Because $(\bar{y}y)$ and $(\bar{x}_2 x_2)$ are positive real values, a set of sufficient conditions for $\mathbf{x}^* Q\mathbf{x} > 0$ is to have $q_{11} > 0$ and $\mathbf{det}(Q) > 0$. These conditions turn out to also be necessary conditions for A to be positive definite.

For instance, consider

$$A = \begin{pmatrix} 5 & 0 \\ 1 & 3 \end{pmatrix} \quad \rightarrow \quad Q = \begin{pmatrix} 5 & 1/2 \\ 1/2 & 3 \end{pmatrix}$$

Because $q_{11} = 5$ and $\mathbf{det}(Q) = 14.75$, the quadratic form is given by

$$\mathbf{x}^* A\mathbf{x} = 5\overline{\left(x_1 + \frac{1}{10}x_2\right)}\left(x_1 + \frac{1}{10}x_2\right) + \frac{14.75}{5}\bar{x}_2 x_2$$

which we can see will always have a positive value if $\mathbf{x} \neq 0$. Thus A is positive definite.

Note that A does not have to be symmetric or Hermitian to be positive definite. However, for the purpose of determining positive definiteness of a square matrix A, one can simply analyze the **Hermitian component** $Q = (A + A^*)/2$, which is what the theorem below will be focused on. We can generalize the procedure shown in Example A.2 to $N > 2$. The same process of completing the square will produce

the following theorem, known as **Sylvester's criterion** for establishing whether a Hermitian matrix is positive definite.

THEOREM A.2. *An $N \times N$ Hermitian matrix H is positive definite if and only if the determinants d_k, $k = 1, \ldots, N$, are all positive, where*

$$d_k = \mathbf{det} \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1k} \\ h_{21} & h_{22} & \cdots & h_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ h_{k1} & h_{k2} & \cdots & h_{kk} \end{pmatrix} \tag{A.36}$$

EXAMPLE A.3. Let H be a symmetric matrix given by

$$H = \begin{pmatrix} 2 & 3 & 3 \\ 3 & 6 & 0.1 \\ 3 & 0.1 & 2 \end{pmatrix}$$

Using Sylvester's criterion, we take the determinants of the principal submatrices of increasing size starting from the upper left corner:

$$\mathbf{det}\,(2) = 2 \quad , \quad \mathbf{det} \begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix} = 3 \quad , \quad \mathbf{det} \begin{pmatrix} 2 & 3 & 3 \\ 3 & 6 & 0.1 \\ 3 & 0.1 & 2 \end{pmatrix} = -40.52$$

Because one of the determinants is not positive, we conclude that H is not positive definite.

Now consider matrix Q given by

$$Q = \begin{pmatrix} 3 & -1 & 0.1 \\ -1 & 4 & 2 \\ 0.1 & 2 & 3 \end{pmatrix}$$

Using Sylvester's criterion on Q, the determinants are:

$$\mathbf{det}\,(3) = 3, \quad \mathbf{det} \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} = 11, \quad \mathbf{det} \begin{pmatrix} 3 & -1 & 0.1 \\ -1 & 4 & 2 \\ 0.1 & 2 & 3 \end{pmatrix} = 20.56$$

Because all the determinants are positive, we conclude that Q is positive definite. Note that matrices that contain only positive elements are known as **positive matrices**. Thus H given previously is a positive matrix. However, as we just showed, H is not positive definite. Conversely, Q given previously is not a positive matrix because it contains some negative elements. However, Q is positive definite. Therefore, it is crucial to distinguish between the definitions of positive definite matrices and positive matrices.

APPENDIX B

Additional Details and Fortification for Chapter 2

B.1 Gauss Jordan Elimination Algorithm

To obtain Q and W, we use a sequence of elementary row and column matrices to obtain (2.3). Each step has the objective of "eliminating" nonzero terms in the off-diagonal positions. This method is generally known as the **Gauss-Jordan elimination method**.

We begin with the **pivoting step**. This step is to find two permutation matrices that would move a chosen element of matrix $A[=]N \times M$, known as the **pivot**, to the upper-left corner, or the $(1, 1)$-position. A practical choice for the pivot is to select, among the elements of A, the element that has the largest absolute value. Suppose the pivot element is located at the ξ^{th} row and η^{th} column; then the required permutation matrices are $P_{(\xi)}$ and $P_{(\eta)}$, where $P_{(\xi)}$ is obtained by taking an $N \times N$ identity matrix and interchanging the first row and the ξ^{th} row, and $P_{(\eta)}$ is obtained by taking an $M \times M$ identity matrix and interchanging the first row and the η^{th} row. Applying these matrices on A will yield

$$P_{(\xi)} A P_{(\eta)}^T = B$$

where B is a matrix that contains the pivot element in the upper-left corner.

By choosing the element with the largest absolute value as the pivot, the pivot is 0 only when $A = 0$. This can then be used as a stopping criterion for the elimination process. Thus if $A \neq 0$, matrix B will have a nonzero value in the upper-left corner, yielding the following partitioned matrix:

$$P_{(\xi)} A P_{(\eta)}^T = B = \left(\begin{array}{c|c} b_{11} & \mathbf{w}^T \\ \hline \mathbf{v} & \Psi \end{array}\right) \tag{B.1}$$

The **elimination process** takes the values of b_{11}, $\mathbf{v}$ and $\mathbf{w}^T$ to form an elementary row operator matrix $G_L[=]N \times N$ and a column elementary operator matrix $G_R[=]M \times M$ given by

$$G_L = \left(\begin{array}{c|ccc} \frac{1}{b_{11}} & 0 & \cdots & 0 \\ \hline & 1 & & 0 \\ -\frac{1}{b_{11}}\,\mathbf{v} & & \ddots & \\ & 0 & & 1 \end{array}\right) \quad \text{and} \quad G_R = \left(\begin{array}{c|ccc} 1 & & -\frac{1}{b_{11}}\,\mathbf{w}^T & \\ \hline 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{array}\right) \tag{B.2}$$

These matrices can now eliminate, or "zero-out", the nondiagonal elements in the first row and first column, while normalizing the $(1,1)^{\text{th}}$ element, that is,

$$
\begin{aligned}
G_L B G_R &= \left(\begin{array}{c|ccc} \dfrac{1}{b_{11}} & 0 & \cdots & 0 \\ \hline & 1 & & 0 \\ -\dfrac{1}{b_{11}}\,\mathbf{v} & & \ddots & \\ & 0 & & 1 \end{array}\right) \left(\begin{array}{c|c} b_{11} & \mathbf{w}^T \\ \hline \mathbf{v} & \Psi \end{array}\right) \left(\begin{array}{c|ccc} 1 & & -\dfrac{1}{b_{11}}\,\mathbf{w}^T & \\ \hline 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{array}\right) \\
&= \left(\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & \Psi - \dfrac{1}{b_{11}}\,\mathbf{v}\mathbf{w}^T & \\ 0 & & & \end{array}\right)
\end{aligned}
$$

Let $\alpha = a_{\xi,\eta}$ be the pivot of A. For computational convenience, we could combine the required matrices, that is, let $E_L = G_L P_{(\xi)}$ and $E_R = P_{(\eta)}^T G_R$, then
If $\xi = 1$,

$$
E_L = \left(\begin{array}{c|ccc} 1/\alpha & 0 & \cdots & 0 \\ \hline -a_{2,\eta}/\alpha & 1 & & 0 \\ \vdots & & \ddots & \\ -a_{m,\eta}/\alpha & 0 & & 1 \end{array}\right) \tag{B.3}
$$

otherwise, if $\xi > 1$,

$$
E_L = \left(\begin{array}{ccc|c|ccc} 0 & \cdots & 0 & 1/\alpha & 0 & \cdots & 0 \\ \hline 0 & & 0 & -a_{2,\eta}/\alpha & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots & \vdots & & \vdots \\ 0 & & 1 & -a_{\xi-1,\eta}/\alpha & 0 & \cdots & 0 \\ \hline 1 & \cdots & 0 & -a_{1,\eta}/\alpha & 0 & \cdots & 0 \\ \hline 0 & \cdots & 0 & -a_{\xi+1,\eta}/\alpha & 1 & & 0 \\ \vdots & & \vdots & \vdots & & \ddots & \\ 0 & \cdots & 0 & -a_{m,\eta}/\alpha & 0 & & 1 \end{array}\right) \begin{array}{l} \\ \\ \\ \leftarrow \xi^{\text{th}} \text{ row} \\ \\ \\ \\ \end{array}
$$

$$
\uparrow
$$

$$
\xi^{\text{th}} \text{ column} \tag{B.4}
$$

If $\eta = 1$

$$
E_R = \left(\begin{array}{c|ccc} 1 & -a_{\xi,2}/\alpha & \cdots & -a_{\xi,n}/\alpha \\ \hline 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{array}\right) \tag{B.5}
$$

otherwise, if $\eta > 1$

$$E_R = \left(\begin{array}{c|ccc|c|ccc} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \hline 0 & 1 & & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots & \vdots & & \vdots \\ 0 & 0 & & 1 & 0 & 0 & \cdots & 0 \\ \hline 1 & -\frac{a_{\xi,2}}{\alpha} & \cdots & -\frac{a_{\xi,\eta-1}}{\alpha} & -\frac{a_{\xi,1}}{\alpha} & -\frac{a_{\xi,\eta+1}}{\alpha} & \cdots & -\frac{a_{\xi,n}}{\alpha} \\ \hline 0 & 0 & \cdots & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \ddots & \\ 0 & 0 & \cdots & 0 & 0 & 0 & & 1 \end{array}\right) \leftarrow \eta^{\text{th}} \text{ row}$$

$$\uparrow$$

$$\eta^{\text{th}} \text{ column} \qquad \text{(B.6)}$$

EXAMPLE B.1. Let A be given by

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

The pivot is $\alpha = a_{3,2} = 4$; thus $\xi = 3$ and $\eta = 2$. Using (B.4) and (B.6),

$$E_L = \begin{pmatrix} 0 & 0 & 1/4 \\ 0 & 1 & -1/2 \\ 1 & 0 & -1/4 \end{pmatrix} \qquad E_R = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1/2 & -3/4 \\ 0 & 0 & 1 \end{pmatrix}$$

from which we get

$$E_L A E_R = \left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & -2 & 3/2 \\ 0 & 1/2 & 1/4 \end{array}\right)$$

The complete Gauss-Jordan elimination method proceeds by applying the same elimination process on the lower-right block matrix to eliminate the off-diagonal elements in the second row and second column, and so on. To summarize, we have the following Gauss-Jordan elimination algorithm:

Gauss-Jordan Elimination Algorithm:

Objective: Given $A[=]N \times M$, find Q and W such that QAW satisfies (2.3) and the rank r.
Initialize: $r \leftarrow 0$, $Q \leftarrow I_M$, $W \leftarrow I_N$ and $\Omega \leftarrow A$.
Iteration: While $r < \min(N, M)$ and $\Omega \neq 0$,

1. Determine the pivot $\alpha = \max_{i,j}\left(\left|\Omega_{ij}\right|\right)$. If $\alpha = 0$, then stop; otherwise, continue.
2. Construct E_L and E_R using (B.3)-(B.6) and extract Γ

$$E_L \Omega E_R = \left(\begin{array}{c|c} 1 & 0 \cdots 0 \\ \hline 0 & \\ \vdots & \Gamma \\ 0 & \end{array}\right)$$

3. Update Q, W, and Ω as follows:

$$Q \leftarrow \begin{cases} E_L & \text{if } r = 0 \\ \left(\begin{array}{c|c} I_r & 0_{[r\times(M-r)]} \\ \hline 0_{[(N-r)\times r]} & E_L \end{array}\right) Q & \text{otherwise} \end{cases}$$

$$W \leftarrow \begin{cases} E_R & \text{if } r = 0 \\ W\left(\begin{array}{c|c} I_r & 0_{[r\times(M-r)]} \\ \hline 0_{[(N-r)\times r]} & E_R \end{array}\right) & \text{otherwise} \end{cases}$$

$$\Omega \leftarrow \Gamma$$

4. Increment r by 1.

EXAMPLE B.2. For the matrix A given by

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

the algorithm will yield the following calculations:

Iteration	Ω	α	ξ	η	E_L	E_R
1	$\begin{pmatrix} 1 & 1 & 1 \\ -1 & 2 & 3 \\ 2 & 4 & 3 \end{pmatrix}$	4	3	2	$\begin{pmatrix} 0 & 0 & 1/4 \\ 0 & 1 & -1/2 \\ 1 & 0 & -1/4 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -1/2 & -3/4 \\ 0 & 0 & 1 \end{pmatrix}$
2	$\begin{pmatrix} -2 & 3/2 \\ 1/2 & 1/4 \end{pmatrix}$	-2	1	1	$\begin{pmatrix} -1/2 & 0 \\ 1/4 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 3/4 \\ 0 & 1 \end{pmatrix}$
3	$\begin{pmatrix} 5/8 \end{pmatrix}$	$\frac{5}{8}$	1	1	$\begin{pmatrix} 8/5 \end{pmatrix}$	$\begin{pmatrix} 1 \end{pmatrix}$

from which Q and W can be obtained as

$$
\begin{aligned}
Q &= \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 8/5 \end{array}\right)\left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & -1/2 & 0 \\ 0 & 1/4 & 1 \end{array}\right)\begin{pmatrix} 0 & 0 & 1/4 \\ 0 & 1 & -1/2 \\ 1 & 0 & -1/4 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & 1/4 \\ 0 & -1/2 & 1/4 \\ 8/5 & 2/5 & -3/5 \end{pmatrix} \\
W &= \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1/2 & -3/4 \\ 0 & 0 & 1 \end{pmatrix}\left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & 1 & 3/4 \\ 0 & 0 & 1 \end{array}\right)\left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \end{array}\right) \\
&= \begin{pmatrix} 0 & 1 & 3/4 \\ 1 & -1/2 & -9/8 \\ 0 & 0 & 1 \end{pmatrix}
\end{aligned}
$$

and the rank $r = 3$

Remarks:

1. By choosing the pivot α to be the element having the largest absolute value, accuracy is also improved because division by small values can lead to larger roundoff errors.
2. The value of rank r is an important property of a matrix. If the matrix is square, $r = M = N$ implies a nonsingular matrix; otherwise it is singular. For a non-square $M \times N$ matrix, if $r = \min(M, N)$, then the matrix is called a matrix of **full rank**; otherwise we refer to them as **partial rank** matrices.[1]
3. Because roundoff errors resulting from the divisions by the pivot tend to propagate with each iteration, the Gauss-Jordan elimination method is often used for medium-sized problems only. This means that in some cases, the value of zero may need to be relaxed to within a specified tolerance.
4. The Gauss-Jordan elimination algorithm can also be used to find the determinant of A. Assuming $r = M = N$, the determinant can be found by taking the products of the pivots and (-1) raised to the number of instances where $\xi \neq 1$ plus the number of instances where $\eta \neq 1$. For example, using the calculations performed in Example B.2, there is one instance of $\xi \neq 1$ and one instance of $\eta \neq 1$ while the pivots are 4, -2, and $5/8$. Thus the determinant is given by

$$\mathbf{det}\left(A\right) = (-1)^{1+1}\,(4)\,(-2)\left(\frac{5}{8}\right) = -5$$

5. A MATLAB file `gauss_jordan.m` is available on the book's webpage that finds the matrices Q and W, as well as inverses Q^{-1} and W^{-1}. The program allows one to prescribe the tolerance level while taking advantage of the sparsity of E_L and E_R.

[1] As discussed later, the rank r determines how many columns or rows are **linearly independent**.

B.2 SVD to Determine Gauss-Jordan Matrices Q and W

In this section, we show an alternative approach to find matrices Q and W. The Gauss-Jordan elimination procedure is known to propagate roundoff errors through each iteration; thus it may be inappropriate to use for large systems. An approach based on a method known as the **singular value decomposition** can be used to find matrices nonsingular Q and W to satisfy (2.3) with improved accuracy but often at some additional computational costs. For any matrix A, there exist unitary matrices U and V (i.e., $U^* = U^{-1}$ and $V^* = V^{-1}$) and a matrix Σ such that

$$U\Sigma V^* = A \tag{B.7}$$

where Σ contains r non-negative real values in the diagonal arranged in decreasing values and where r is the rank of A, that is,

$$\Sigma = \begin{pmatrix} \sigma_1 & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ 0 & & & & & 0 \end{pmatrix} \quad \text{where} \quad \sigma_i > 0,\ i = 1, \ldots, r \tag{B.8}$$

The details for obtaining U, V, and Σ can be found in Section 3.9. Based on (B.7), Q and W can be found as follows:

$$Q = \Sigma^{\langle -1 \rangle} U^* \qquad \text{and} \qquad W = V \tag{B.9}$$

where,

$$\Sigma^{\langle -1 \rangle} = \begin{pmatrix} \sigma_1^{-1} & & & & & 0 \\ & \ddots & & & & \\ & & \sigma_r^{-1} & & & \\ & & & 1 & & \\ & & & & \ddots & \\ 0 & & & & & 1 \end{pmatrix}$$

Alternatively, we can have $Q = U^*$ and $W = V\Sigma^{\langle -1 \rangle}$.

For non-square $A[=]N \times M$, let $k = \min(N, M)$; then we can set $\Sigma^{\langle -1 \rangle}[=]k \times k$. If $N > M$, we can then have $Q = U^*$ and $W = V\Sigma^{\langle -1 \rangle}$. Otherwise, we can set $Q = \Sigma^{\langle -1 \rangle} U^*$ and $W = V$.

Remarks: In MATLAB, one can find the matrices U, V, and $S = \Sigma$ using the statement: `[U,S,V] = svd(A)`. A function `gauss_svd.m` is available on the book's webpage that obtains Q and W using the SVD approach.

EXAMPLE B.3. Let A be given by

$$A = \begin{pmatrix} 12 & -32 & 28 \\ 0 & -4 & 2 \\ 10 & -24 & 22 \\ 3 & -8 & 7 \end{pmatrix}$$

Then the singular value decomposition can be obtained using MATLAB's `svd` command to be

$$U = \begin{pmatrix} -0.7749 & 0.2179 & 0.0626 & -0.5900 \\ -0.0728 & 0.8848 & -0.2314 & 0.3978 \\ -0.5972 & -0.4083 & -0.3471 & 0.5968 \\ -0.1937 & 0.0545 & 0.9066 & 0.3708 \end{pmatrix}$$

$$V = \begin{pmatrix} -0.2781 & -0.6916 & 0.6667 \\ 0.7186 & -0.6103 & -0.3333 \\ -0.6374 & -0.3864 & -0.6667 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 57.0127 & 0 & 0 \\ 0 & 1.8855 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \longrightarrow \Sigma^{\langle -1 \rangle} = \begin{pmatrix} 0.0175 & 0 & 0 \\ 0 & 0.5304 & 0 \\ 0 & 0 & 1.0000 \end{pmatrix}$$

Finally,

$$Q = U^* \qquad \text{and} \qquad W = V\Sigma^{\langle -1 \rangle} = \begin{pmatrix} -0.0049 & -0.3668 & 0.6667 \\ 0.0126 & -0.3237 & -0.3333 \\ -0.0112 & -0.2049 & -0.6667 \end{pmatrix}$$

B.3 Boolean Matrices and Reducible Matrices

Boolean matrices are matrices whose elements are boolean types, that is, `TRUE` and `FALSE`, which are often represented by the integers 1 and 0, respectively. They are strongly associated with graph theory. Because the elements of these matrices are boolean, the operations will involve logical disjunction ("`or`") and logical conjunction ("`and`"). One important application of boolean matrices is to represent the structure of a **directed graph** (or **digraph** for short).

A **digraph** is a collection of **vertices** v_i connected to each other by **directed arcs** denoted by (v_i, v_j) to represent an arc from v_i to v_j. A symbolic representation of a digraph is often obtained by drawing open circles for vertices v_i and connecting vertices v_i and v_j by an arrow for arcs (v_i, v_j). For instance, a graph

$$G = \left(\{v_1, v_2, v_3\} \,\middle|\, \{(v_1, v_2), (v_3, v_1), (v_3, v_2)\} \right) \tag{B.10}$$

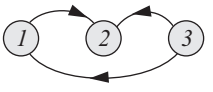
is shown in Figure B.1.

A boolean matrix representation of a digraph is given by a square matrix, say G_{B}, whose elements $g_{ji} = 1$ (`TRUE`) if an arc (v_i, v_j) exists. Thus the boolean matrix for digraph G specified in (B.10) is given by

$$G_{\mathsf{B}} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}_{\mathsf{B}}$$

(We use the subscript B to indicate that the elements are boolean.)

Figure B.1. The digraph G given in (B.10).



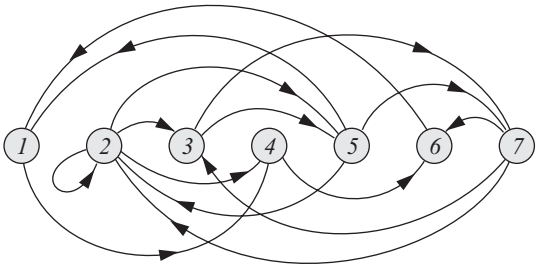


Figure B.2. The influence digraph corresponding to (B.11).

One particular application of boolean matrices (and the corresponding digraphs) is to find a partitioning (and reordering) of simultaneous equations that could improve the efficiency of solving for the unknowns. For a given nonlinear equation such as $x_3 = f(x_1, x_5)$, we say that x_1 and x_5 will influence the value of x_3. Thus we can build an **influence digraph** that will contain vertices v_1, v_3, and v_5, among others, plus the directed arcs (v_1, v_3) and (v_5, v_3).

EXAMPLE B.4. Consider the following set of simultaneous equations:

$$\begin{array}{lcllcl} x_1 &=& f_1(x_5, x_6) & x_5 &=& f_5(x_2, x_3) \\ x_2 &=& f_2(x_2, x_5, x_7) & x_6 &=& f_6(x_4, x_7) \\ x_3 &=& f_3(x_2, x_7) & x_7 &=& f_7(x_3, x_5) \\ x_4 &=& f_4(x_1, x_2) & & & \end{array} \tag{B.11}$$

The influence digraph of (B.11) is shown in Figure B.2.
The boolean matrix representation of digraph G is given by

$$G_{\mathsf{B}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}_{\mathsf{B}} \tag{B.12}$$

The vertices of Figure B.2 can be moved around to show a clearer structure and a partitioning into two subgraphs G_1 and G_2, as shown in Figure B.3, where $G_1 = \{x_2, x_3, x_5, x_7\}$ and $G_2 = \{x_1, x_4, x_6\}$. Under this partitioning, any of the vertices in G_1 can link to vertices in G_2, but none of the vertices of G_2 can reach the nodes of G_1. This decomposition implies that functions $\{f_2, f_3, f_5, f_7\}$ in (B.11) can be used first to solve for $\{x_2, x_3, x_5, x_7\}$ as a group because they are not influenced by either x_1, x_4, or x_6. Afterward, the results can be substituted to functions $\{f_1, f_4, f_6\}$ to solve for $\{x_1, x_4, x_6\}$.[2]

The sub-digraphs G_1 and G_2 in Figure B.3 are described as **strongly connected**. We say that a collection of vertices together with their arcs (only among the vertices in the same collection) are **strongly connected** if any vertex can reach any other

[2] The process in which a set of nonlinear equations are sequenced prior to actual solving of the unknowns is known as **precedence ordering**.

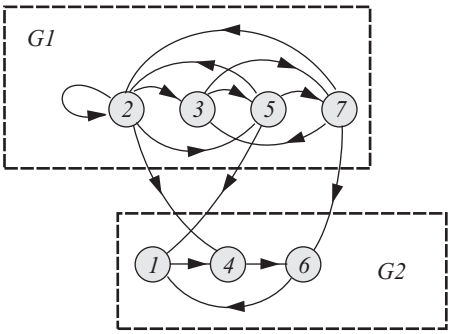


Figure B.3. The influence digraph corresponding to (B.11) after repositioning and partitioning.

vertex in the same collection. Obviously, as the number of vertices increases, the complexity will likely increase such that the decomposition to strongly connected subgraphs will be very difficult to ascertain by simple inspection alone. Instead, we can use the boolean matrix representation of the influence digraph to find the desired partitions.

Because the elements of the boolean matrices are boolean (or logic) variables, the logical "OR" and logical "AND" operations will replace the product ("·") and sum ("+") operations, respectively, during the matrix product operations. This means that we have the following rules[3]:

$$
\begin{gathered}
(0+0)_{\mathsf{B}} = 0 \\
(0+1)_{\mathsf{B}} = (1+0)_{\mathsf{B}} = (1+1)_{\mathsf{B}} = 1 \\
(0\cdot 0)_{\mathsf{B}} = (1\cdot 0)_{\mathsf{B}} = (0\cdot 0)_{\mathsf{B}} = 0 \\
(1\cdot 1)_{\mathsf{B}} = 1 \\
(A\cdot B)_{\mathsf{B}} = C \quad \longleftrightarrow \quad c_{ij} = \Big((a_{i1}\cdot b_{1j})_{\mathsf{B}} + \cdots + (a_{iK}\cdot b_{Kj})_{\mathsf{B}} \Big)_{\mathsf{B}} \\
\left(A^k\right)_{\mathsf{B}} = (A\cdot A\cdots A)_{\mathsf{B}}
\end{gathered}
\tag{B.13}
$$

For instance, we have

$$
\left(\begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \right)_{\mathsf{B}} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}_{\mathsf{B}}
$$

Using the rules in (B.13), we note that for a digraph $A[=]N \times N$, the result of $\left(A^k\right)_{\mathsf{B}}$ with $k \leq N$ will be to add new arcs (v_i, v_j) to the original digraph if there exists a path consisting of at most k arcs that would link v_i to v_j. Thus to find the strongly connected components, we could simply perform the boolean matrix conjunctions enough times until the resulting digraph has settled to a fixed boolean matrix, that is, find $k \leq N$ such that $\left(A^k\right)_{\mathsf{B}} = \left(A^{k+1}\right)_{\mathsf{B}}$.

[3] For clarity, we include a subscript B to denote boolean operation.

EXAMPLE B.5. Using the boolean matrix representation of digraph (B.11) given by (B.12), we can show that

$$(G^3)_{\mathsf{B}} = \begin{pmatrix} 0&0&0&0&1&1&0 \\ 0&1&0&0&1&0&1 \\ 0&1&0&0&0&0&1 \\ 1&1&0&0&0&0&0 \\ 0&1&1&0&0&0&0 \\ 0&0&0&1&0&0&1 \\ 0&0&1&0&1&0&0 \end{pmatrix}^3_{\mathsf{B}} = \begin{pmatrix} 1&1&1&1&1&1&1 \\ 0&1&1&0&1&0&1 \\ 0&1&1&0&1&0&1 \\ 1&1&1&1&1&1&1 \\ 0&1&1&0&1&0&1 \\ 1&1&1&1&1&1&1 \\ 0&1&1&0&1&0&1 \end{pmatrix}_{\mathsf{B}} = (G^4)_{\mathsf{B}}$$

From the result of $(G^3)_{\mathsf{B}}$, we see that columns $\{2, 3, 5, 7\}$ have the same entries, whereas columns $\{1, 4, 6\}$ have the same entries. These two groups of indices determine the subgraphs G_1 and G_2 obtained in Example B.4.

For the special case of linear equations, we could use the results just obtained to determine whether a matrix is **reducible** or not, and if it is, we could also find the required permutation matrix P that is needed to find the reduced form.

Definition B.1. *A square matrix A is a* **reducible matrix** *if there exists a permutation matrix P such that*

$$PAP^T = B = \left(\begin{array}{c|c} B_{11} & 0 \\ \hline B_{12} & B_{22} \end{array}\right)$$

A matrix that is not reducible is simply known as an **irreducible matrix**.

Algorithm for Determination of Reducible Matrices

Given matrix $A[=]N \times N$

1. Replace A by a boolean matrix G, where $g_{ij} = 1_{\mathsf{B}}$ if $a_{ij} \neq 0$ and $g_{ij} = 0_{\mathsf{B}}$ otherwise.
2. Perform matrix conjunctions $(G^k)_{\mathsf{B}}$ until $(G^k)_{\mathsf{B}} = (G^{k-1})_{\mathsf{B}}$, $k \leq N$.
3. Let $\kappa(\ell)$ be the number of logical TRUE entries in column ℓ. Sort the columns of $(G^k)_{\mathsf{B}}$ in descending sequence, $\{j_1, \ldots, j_N\}$, where $j_i \in \{1, \ldots, N\}$ and $b > a$ if $\kappa(j_b) \leq \kappa(j_a)$.
4. Set the permutation matrix to be

$$P = \left(\begin{array}{c|c|c|c} \mathbf{e}_{j_1} & \cdots & \mathbf{e}j_N \end{array}\right)^T$$

5. Evaluate the reduced block triangular matrix $B = PAP^T$ given by

$$B = \begin{pmatrix} B_{11} & & & 0 \\ B_{21} & B_{22} & & \\ \vdots & \vdots & \ddots & \\ B_{M1} & B_{M2} & \cdots & B_{MM} \end{pmatrix}$$

where the block matrices are $B_{ii}[=]\ell_1 \times \ell_i$ and $\sum_{i=1} M\ell_i = N$.

Remarks: A MATLAB code that implements the algorithm for finding the reduced form is available on the book's webpage as `matrixreduce.m`.

EXAMPLE B.6. Consider the matrix given by

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

then the influence graph is given by the same boolean matrix G_{B} given in example B.5. The algorithm then suggests the following sequence: [2, 3, 5, 7, 1, 4, 6] for P, that is,

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

which then reduces A to a lower block triangular matrix according to the following transformation:

$$PAP^T = \widehat{A} = \left(\begin{array}{cccc|ccc} 1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 3 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 2 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 4 & 0 \end{array}\right)$$

Which means that A is reducible.

Once the block triangular structure has been achieved, a special case of (1.34) can be used, that is,

$$\left(\begin{array}{c|c} B_{11} & 0 \\ \hline B_{21} & B_{22} \end{array}\right)^{-1} = \left(\begin{array}{c|c} B_{11}^{-1} & 0 \\ \hline -B_{22}^{-1} B_{21} B_{11}^{-1} & B_{22}^{-1} \end{array}\right)$$

assuming both B_{11} and B_{22} are nonsingular.

There are several classes of matrices that are known to be irreducible and do not need to be processed by boolean matrices. One example of a class of irreducible matrices is the tri-diagonal matrices with nonzeros entries above and below the main

diagonal. Tri-diagonal matrices are a particular example of sparse matrices, which are matrices that contain a large number of zero entries. For sparse matrices, instead of the search for reduced forms, it is often more useful to find transformations that would reduce the "bandwidth" of the matrices. These issues are discussed in the next section.

B.4 Reduction of Matrix Bandwidth

One of the most often used methods for finding the reordering permutation P to reduce matrix bandwidth is the **Cuthill-Mckee** algorithm. This algorithm does not guarantee a minimal bandwidth, but it often yields an acceptable bandwidth while implementing a reasonable amount of computation. (To simplify the discussion, we assume that the matrix under consideration will already be irreducible. Otherwise, the techniques found in Section B.3 can be used to separate the matrix into strongly connected components, i.e., irreducible submatrices.)

We first introduce a few terms with their corresponding notations:[4]

1. Available nodes: $U = \{u_1, u_2, \ldots\}$. This is a set of indices that has not been processed. It is initialized to contain all the indices, that is, $U = \{1, 2, \ldots, N\}$. The members are removed after each iteration of the algorithm. The algorithm ends once U becomes empty.
2. Current sequence: $V = [v_1, v_2, \ldots]$. This is the set of indices that indicates the current sequence of the permutation P taken from collection U but arranged according to the algorithm.
3. Degree function: $\rho(k)$ = number of nonzero off-diagonal entries in column k. It determines the number of neighbors of index k.
4. Neighbors of index k: $Ne(k) = \{ne_1, ne_2, \ldots\}$ where ne_i are the row indices of column k in A that are nonzero. We could also arrange the elements of $Ne(k)$ as the ordered neighbors $Ne^*(k) = \left[ne_1^*, ne_2^*, \ldots\right]$ sequenced in increasing orders, that is, $\rho(ne_{i+1}^*) > \rho(ne_i^*)$.
5. Entering nodes: $Ent(k, V) = [Ne^*(k) \setminus V]$ where $k \in V$. This is the set of ordered neighbors of index k that has not yet been processed, that is, excluding indices that are already in V.

For instance, consider the matrix

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

The degrees of each index are given by $\Big(\rho(1), \rho(2), \rho(3), \rho(4), \rho(5)\Big) = (2, 1, 1, 2, 2)$. Suppose that the current sequences U and V are given by

$$U = \{3, 4, 5\} \quad \text{and} \quad V = [2, 1]$$

[4] We use a pair of curly brackets to indicate a collection in which the order is not relevant and a pair of square brackets to indicate that the order sequence is important.

then

$$Ne(1) = \{2, 5\} \quad Ne^*(1) = [2, 5] \quad \text{and} \quad Ent\,(1, \{2, 1\}) = [5]$$

For a given initial index s, the Cuthill-McKee sequencing algorithm is given by the following algorithm:

Cuthill-McKee Sequencing Algorithm:

Given matrix $A[=]N \times N$ and starting index s,

1. Evaluate the degrees $\rho\,(i), i = 1, \ldots, N$.
2. Initialize $U = \Big\{\{1, 2, \ldots, N\} \setminus s\Big\}$, $V = [s]$ and $k = 1$.
3. While U is not empty,
   (a) Determine entering indices, $Q = Ent\,(v_k, V)$.
   (If Q is empty, then skip the next two steps and continue the loop iteration.)
   (b) Update U and V: $U \leftarrow \{U \setminus Q\}$ and $V \leftarrow [V, Q]$.
   (c) Increment, $k \leftarrow k + 1$.

Different choices of the starting index s will yield different sequences V and could result in different bandwidths. One choice is to start with the index having the lowest degree $\rho(s)$, but this may not necessarily yield the minimal bandwidth. However, exploring all the indices as starting indices is not desirable either, especially for large matrices. Different methods have been developed to choose the appropriate starting index that would yield a sequence that produces close to, if not the exactly, the minimum bandwidth. We discuss one approach that is iterative.

Using a starting index s (e.g., initially try the index with the lowest order), the Cuthill-McKee algorithm will yield the sequence V_s and its corresponding permutation matrix P_s. This should generate a transformed matrix $B_s = P_s A P_s^T$ that will have a block tri-diagonal structure known as the **level structure rooted at** s, in which the first block is the 1×1 block containing s:

$$B_s = \left(\begin{array}{c|c|c|c|c} s & R_1 & & & 0 \\ \hline F_1 & D_1 & \ddots & & \\ \hline & \ddots & \ddots & \ddots & \\ \hline & & \ddots & \ddots & R_m \\ \hline 0 & & & F_m & D_m \end{array}\right)$$

and where the diagonal blocks D_i are square. The value m is the maximal value that attains a block tri-diagonal structure for B_s and is known as the **depth** of B_s. Let ℓ be the size of the last diagonal block, D_m. Then we can test the indices determined by the last ℓ entries of V_s as starting indices and apply the Cuthill-McKee algorithm to each of these indices. If any of these test indices, say index w, yield a smaller

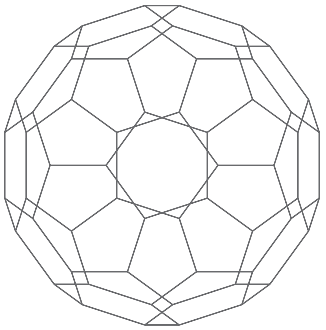


Figure B.4. A graphical representation of the C60 molecule.

bandwidth, then we update s with w and the whole process is repeated.[5] Otherwise, $V = V_s$ is chosen to be the desired sequence.

Remarks:

1. Often, especially in the solution of finite element methods, the reversed ordering has shown a slight computational improvement. Thus a slight modification yields the more popular version known as the **Reverse Cuthill-McKee** reordering, which is to simply reverse the final sequence in V found by the Cuthill-McKee algorithm.
2. The MATLAB command that implements the reverse Cuthill-McKee reordering algorithm of matrix A is: `p=symrcm(A)`, and the permuted matrix can be obtained as: `B=A(p,p)`.
3. A MATLAB function `p=CuthillMcKee(A)` is available on the book's webpage that implements the Cuthill-McKee algorithm.

EXAMPLE B.7. Consider the C60 molecule (or geodesic dome popularized by Buckminster Fuller), which is a form of pure carbon with 60 atoms in a nearly spherical configuration. A graphical figure is shown in Figure B.4. An adjacency (boolean) matrix describing the linkage among the atoms is shown in Figure B.5 in which the dots are `TRUE` and the unmarked positions are `FALSE`. The bandwidth of the original indexing is 34. Note that each node is connected to three other nodes; thus the degrees of each node is 3 for this case. After applying the Cuthill-McKee reordering algorithm, the atoms are relabeled and yield the adjacency matrix shown in Figure B.6. The bandwidth of the reordered matrix is 10.

B.5 Block LU Decomposition

When matrix A is large, taking advantage of inherent block partitions can yield efficient methods for the solution of $A\mathbf{x} = \mathbf{b}$. The block structure could come directly

[5] The method of choosing new starting indices based on the last block of the level structure is based partially on the method developed of Gibbs, Poole and Stockmeyer (1976) for choosing the initial index. Unlike their method, the one discussed here continues with using the Cuthill-McKee algorithm to generate the rest of the permutation.

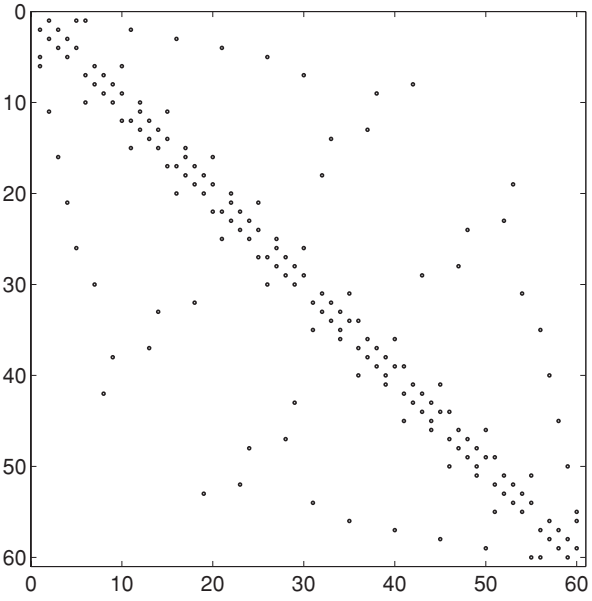


Figure B.5. The initial adjacency matrix for the C60 molecule.

from modular units of connected subsystems, for example, from physical processes composed of different parts. In some cases, it results from the geometry of the problem (e.g., from the finite difference solutions of elliptic partial differential equations). In other cases, the block structure results from reordering of equations and re-indexing of the variables.

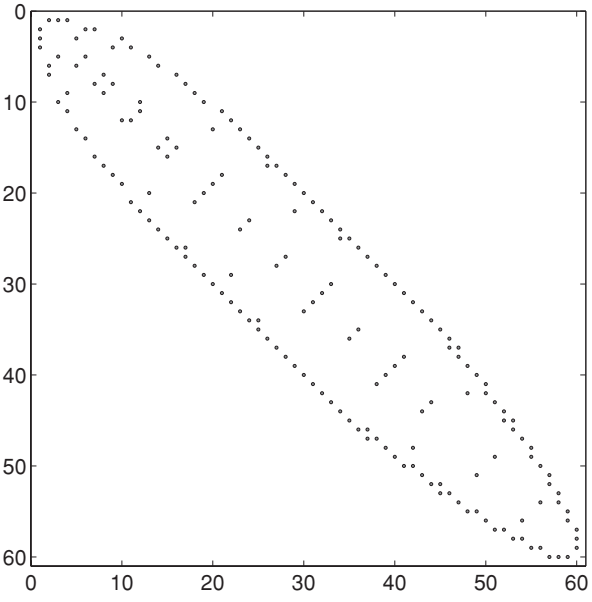


Figure B.6. The adjacency matrix for the C60 molecule based on Cuthill-McKee reordering.

One of the simplest case is when A is lower block triangular.

$$A = \left(\begin{array}{c|c|c} L_{11} & & 0 \\ \hline \vdots & \ddots & \\ \hline L_{n1} & \cdots & L_{nn} \end{array}\right) \tag{B.14}$$

where $L_{i,j}[=]N_i \times N_j, i \geq j$. This induces a partitioning of vectors $\mathbf{x}$ and $\mathbf{b}$ as follows

$$\mathbf{x} = \left(\begin{array}{c} \mathbf{x}_1 \\ \hline \vdots \\ \hline \mathbf{x}_n \end{array}\right) \qquad \text{and} \qquad \mathbf{b} = \left(\begin{array}{c} \mathbf{b}_1 \\ \hline \vdots \\ \hline \mathbf{b}_n \end{array}\right) \tag{B.15}$$

where $\mathbf{x}_k$ and $\mathbf{b}_k$ are column vectors of length N_k.

It is possible that even though the original matrix A does not have the lower block triangular structure of (B.14), one might still be able to find permutation matrices P such that $\widehat{A} = PAP^T$ attains a lower block triangular structure. If so, then A is known as **reducible**. One could use boolean matrices to find the required P, and details of this method are given in Section B.3 as an appendix. Furthermore, a MATLAB code that implements the algorithm for finding the reduced form is available on the book's webpage as `matrixreduce.m`.

Assuming that the block diagonal matrices L_{kk} are square and nonsingular, the solution can be obtained by the block matrix version of forward substitution, that is,

$$\mathbf{x}_1 = L_{11}^{-1}\,\mathbf{b}_1 \qquad \text{and} \qquad \mathbf{x}_k = L_{kk}^{-1}\left(\mathbf{b}_k - \sum_{\ell=1}^{k-1} L_{k,\ell}\,\mathbf{x}_\ell\right) \;;\; k = 2, \ldots, n \tag{B.16}$$

Likewise, when A is upper block triangular, that is,

$$A = \left(\begin{array}{c|c|c} U_{11} & \cdots & U_{1n} \\ \hline & \ddots & \vdots \\ \hline 0 & & U_{nn} \end{array}\right) \tag{B.17}$$

where $U_{i,j}[=]N_i \times N_j, i \leq j$, and assuming that the block diagonal matrices U_{kk} are square and nonsingular, the solution can be obtained by the block matrix version of backward substitution, that is,

$$\mathbf{x}_n = U_{nn}^{-1}\mathbf{b}_n \quad \text{and} \quad \mathbf{x}_k = U_{kk}^{-1}\left(\mathbf{b}_k - \sum_{\ell=k+1}^{n} U_{k,\ell}\,\mathbf{x}_\ell\right) \;;\; k = n-1, \ldots, 1 \tag{B.18}$$

Let A be partitioned as

$$A = \left(\begin{array}{c|c|c} A_{11} & \cdots & A_{1n} \\ \hline \vdots & \ddots & \vdots \\ \hline A_{n1} & \cdots & A_{nn} \end{array}\right) \tag{B.19}$$

where $A_{ij}[=]N_i \times N_j$ with A_{kk} square. Then block matrix computation can be extended to yield block LU decompositions. The block-Crout's method and the block-Doolittle's method are given in Table B.1. Note that L_{ij} and U_{ij} are matrices of size $N_i \times N_j$ and are not triangular in general. Furthermore, when A is block tri-diagonal, a block version of the Thomas algorithm becomes another natural extension. (See Exercise **E2.16** for the block-Thomas algorithm).

Table B.1. *Block matrix LU decompositions*

Name	Algorithm (For $p = 1, \ldots, N$)	
Block Crout's Method	$U_{pp} = I_{N_P}$	
	$L_{ip} = \left(A_{ip} - \sum_{k=1}^{p-1} L_{ik}U_{kp}\right)$	for $i = p, \ldots, n$
	$U_{pj} = L_{pp}^{-1}\left(A_{pj} - \sum_{k=1}^{p-1} L_{pk}U_{kj}\right)$	for $j = p+1, \ldots, n$
Block Doolittle's Method	$L_{pp} = I_{N_p}$	
	$U_{pj} = \left(A_{pj} - \sum_{k=1}^{p-1} L_{pk}U_{kj}\right)$	for $j = p, \ldots, n$
	$L_{ip} = \left(A_{ip} - \sum_{k=1}^{p-1} L_{ik}U_{kp}\right)U_{pp}^{-1}$	for $i = p+1, \ldots, n$

B.6 Matrix Splitting: Diakoptic Method and Schur Complement Method

B.6.1 Diakoptic Method

Let P_R and P_C be row permutation and column permutation matrices, respectively, that will move nonzero elements of $S = A - M$ to the top rows and left columns, leaving a partitioned matrix that has a large zero matrix in the lower-right corner.

$$\widehat{S} = P_R S P_C^T = \left(\begin{array}{c|c} \widehat{S}_{11} & \widehat{S}_{12} \\ \hline \widehat{S}_{21} & 0 \end{array}\right) \tag{B.20}$$

Assume that the size of $\widehat{S}_{11}$ is significantly smaller than the full matrix. If either $\widehat{S}_{12} = 0$ or $\widehat{S}_{21} = 0$, then an efficient solution method known as the **Diakoptic method** is available.

Case 1. $\underline{\widehat{S}_{12} = 0}$

In this case, $P_R = I$. With $S = A - M$ and $\widehat{S} = SP_C^T$, the problem $A\mathbf{x} = \mathbf{b}$ can be recast as follows:

$$A\mathbf{x} = (M + S)\,\mathbf{x} = \mathbf{b} \quad \rightarrow \quad \left(I + H\widehat{S}\right)\mathbf{y} = \mathbf{z}$$

where $H = P_C M^{-1}$, $\mathbf{y} = P_C\mathbf{x}$ and $\mathbf{z} = P_C M^{-1}\mathbf{b}$. Let $\widehat{S}_{11}[=]r \times r$. With $\widehat{S}_{12} = \widehat{S}_{22} = 0$, $L = (I + HS)$ will be block lower triangular matrix, that is,

$$\left[\left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & I_{N-r} \end{array}\right) + \left(\begin{array}{c|c} H_{11} & H_{12} \\ \hline H_{21} & H_{22} \end{array}\right)\left(\begin{array}{c|c} \widehat{S}_{11} & 0 \\ \hline \widehat{S}_{21} & 0 \end{array}\right)\right]\left(\begin{array}{c} \mathbf{y}_r \\ \hline \mathbf{y}_{N-r} \end{array}\right) = \left(\begin{array}{c} \mathbf{z}_r \\ \hline \mathbf{z}_{N-r} \end{array}\right)$$

$$\left(\begin{array}{c|c} L_{11} & 0 \\ \hline L_{21} & L_{22} \end{array}\right)\left(\begin{array}{c} \mathbf{y}_r \\ \hline \mathbf{y}_{N-r} \end{array}\right) = \left(\begin{array}{c} \mathbf{z}_r \\ \hline \mathbf{z}_{N-r} \end{array}\right)$$

where $L_{11} = I_r + H_{11}\widehat{S}_{11} + H_{12}\widehat{S}_{21}$, $L_{21} = H_{21}\widehat{S}_{11} + H_{22}\widehat{S}_{21}$ and $L_{22} = I_{N-r}$. Note that the blocks of L are obrained by just partitioning $(I + H\widehat{S})$.
Assuming L_{11} is nonsingular,

$$\begin{array}{rcl} \mathbf{y}_r & = & L_{11}^{-1}\mathbf{z}_r \\ \mathbf{y}_{N-r} & = & \mathbf{z}_{N-r} - L_{21}\mathbf{y}_r \end{array} \quad \rightarrow \quad \mathbf{x} = P_C^T \begin{pmatrix} \mathbf{y}_r \\ \hline \mathbf{y}_{N-r} \end{pmatrix} \tag{B.21}$$

EXAMPLE B.8. For the equation $A\mathbf{x} = \mathbf{b}$, let

$$A = \begin{pmatrix} 5 & 0 & 1 & 0 & 1 & 0 \\ 1 & 4 & -1 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 2 & 0 \\ 1 & 0 & 1 & 4 & 0 & 0 \\ 1 & 2 & 0 & 1 & 5 & 0 \\ -1 & 0 & 2 & -1 & 1 & 5 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 9 \\ 6 \\ 16 \\ 0 \\ 9 \\ -3 \end{pmatrix}$$

Choosing M to be lower triangular portion of A, we have

$$M = \begin{pmatrix} 5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 & 0 \\ 1 & 0 & 1 & 4 & 0 & 0 \\ 1 & 2 & 0 & 1 & 5 & 0 \\ -1 & 0 & 2 & -1 & 1 & 5 \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Let P_C be

$$P_C = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

then we obtain

$$(I + H\widehat{S}) = \left(\begin{array}{cc|cccc} 1.075 & 0.513 & 0 & 0 & 0 & 0 \\ 0.094 & 1.016 & 0 & 0 & 0 & 0 \\ \hline 0.2 & 0.2 & 1 & 0 & 0 & 0 \\ -0.3 & -0.05 & 0 & 1 & 0 & 0 \\ -0.069 & -0.178 & 0 & 0 & 1 & 0 \\ -0.023 & -0.204 & 0 & 0 & 0 & 1 \end{array}\right) \quad \text{and} \quad \mathbf{z} = \begin{pmatrix} 3.737 \\ 1.297 \\ 1.8 \\ 1.05 \\ -1.384 \\ -2.271 \end{pmatrix}$$

Finally, we get

$$\mathbf{y}_r = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad ; \quad \mathbf{y}_{N-r} = \begin{pmatrix} 1 \\ 2 \\ -1 \\ -2 \end{pmatrix} \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ -1 \\ 1 \\ -2 \end{pmatrix}$$

Case 2. $\widehat{S}_{21} = 0$

For this case, $P_C = I$. With matrices $S = A - M$ and $\widehat{S} = P_R S P_C^T$, the problem $A\mathbf{x} = \mathbf{b}$ can be recast as follows:

$$A\mathbf{x} = (M+S)\,\mathbf{x} = \mathbf{b} \quad \rightarrow \quad \left(I + \widehat{S}\widehat{H}\right)\widehat{\mathbf{y}} = \widehat{\mathbf{z}}$$

where $\widehat{H} = M^{-1}P_R^T$, $\widehat{\mathbf{y}} = P_R M\mathbf{x}$ and $\widehat{\mathbf{z}} = P_R\mathbf{b}$. Let $\widehat{S}_{11}[=]r \times r$. With $\widehat{S}_{21} = \widehat{S}_{22} = 0$,

$$\left[\left(\begin{array}{c|c} I_r & 0 \\ \hline 0 & I_{N-r} \end{array}\right) + \left(\begin{array}{c|c} \widehat{S}_{11} & \widehat{S}_{12} \\ \hline 0 & 0 \end{array}\right)\left(\begin{array}{c|c} \widehat{H}_{11} & \widehat{H}_{12} \\ \hline \widehat{H}_{21} & \widehat{H}_{22} \end{array}\right)\right]\left(\begin{array}{c} \widehat{\mathbf{y}}_r \\ \hline \widehat{\mathbf{y}}_{N-r} \end{array}\right) = \left(\begin{array}{c} \widehat{\mathbf{z}}_r \\ \hline \widehat{\mathbf{z}}_{N-r} \end{array}\right)$$

$$\left(\begin{array}{c|c} U_{11} & U_{12} \\ \hline 0 & U_{22} \end{array}\right)\left(\begin{array}{c} \widehat{\mathbf{y}}_r \\ \hline \widehat{\mathbf{y}}_{N-r} \end{array}\right) = \left(\begin{array}{c} \widehat{\mathbf{z}}_r \\ \hline \widehat{\mathbf{z}}_{N-r} \end{array}\right)$$

where $U_{11} = I_r + \widehat{S}_{11}\widehat{H}_{11} + \widehat{S}_{12}\widehat{H}_{21}, U_{12} = \widehat{S}_{11}\widehat{H}_{12} + \widehat{S}_{12}\widehat{H}_{22}$ and $U_{22} = I_{N-r}$. The blocks of U are obtained by simple partitioning of $\left(I + \widehat{S}\widehat{H}\right)$. Assuming U_{11} is non-singular,

$$\begin{aligned} \widehat{\mathbf{y}}_{N-r} &= \widehat{\mathbf{z}}_{N-r} \\ \widehat{\mathbf{y}}_r &= U_{11}^{-1}\left(\widehat{\mathbf{z}}_r - U_{12}\widehat{\mathbf{y}}_{N-r}\right) \end{aligned} \quad \rightarrow \quad \mathbf{x} = M^{-1}P_R^T\left(\begin{array}{c} \widehat{\mathbf{y}}_r \\ \hline \widehat{\mathbf{y}}_{N-r} \end{array}\right) \tag{B.22}$$

EXAMPLE B.9. For the equation $A\mathbf{x} = \mathbf{b}$, let

$$A = \begin{pmatrix} 5 & 1 & 0 & 1 & 1 & -1 \\ 0 & 4 & 1 & 0 & 2 & 0 \\ 1 & -1 & 4 & 1 & 0 & 2 \\ 0 & 0 & 0 & 4 & 1 & -1 \\ 1 & 0 & 2 & 0 & 5 & 1 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 9 \\ 13 \\ 6 \\ -1 \\ 10 \\ -10 \end{pmatrix}$$

Choosing M to be upper triangular portion of A, we have

$$M = \begin{pmatrix} 5 & 1 & 0 & 1 & 1 & -1 \\ 0 & 4 & 1 & 0 & 2 & 0 \\ 0 & 0 & 4 & 1 & 0 & 2 \\ 0 & 0 & 0 & 4 & 1 & -1 \\ 0 & 0 & 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & 0 & 0 & 5 \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Let P_R be given by

$$P_R = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

then

$$(I+\widehat{S}\widehat{H}) = \left(\begin{array}{cc|cccc} 1.075 & 0.094 & 0.2 & -0.3 & -0.069 & -0.023 \\ 0.513 & 1.016 & 0.2 & -0.05 & -0.178 & 0.204 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right); \quad \widehat{\mathbf{z}} = \begin{pmatrix} 6 \\ 10 \\ 9 \\ 13 \\ -1 \\ -10 \end{pmatrix}$$

Finally, we obtain

$$\widehat{\mathbf{y}}_{N-r} = \begin{pmatrix} 9 \\ 13 \\ -1 \\ 10 \end{pmatrix} \quad ; \quad \widehat{\mathbf{y}}_r = \begin{pmatrix} 7 \\ 3 \end{pmatrix} \quad \text{and} \quad \mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ -1 \\ 1 \\ -2 \end{pmatrix}$$

B.6.2 Schur Complements

In solving partial differential equations, the problem can sometimes be partitioned into subdomains. The boundaries of each subdomain will either be specified by boundary conditions or interfaced with other subdomains. In these approaches, known as **domain decomposition**, the matrix A can end up with the following block structure:

$$A = \left(\begin{array}{c|c|c|c} A_{11} & & 0 & A_{1n} \\ \hline & \ddots & & \vdots \\ \hline 0 & & A_{n-1,n-1} & A_{n-1,n} \\ \hline A_{n,1} & \cdots & A_{n,n-1} & A_{n,n} \end{array}\right) \tag{B.23}$$

EXAMPLE B.10. Consider the domain given in Figure B.7 in which a larger rectangular region (Subdomain I) is attached to a smaller rectangular region (Subdomain II). We identify points $\{1, 2, 3, \ldots, 10\}$ and $\{11, 12, 13, 14\}$ to be the interior points of Subdomain I and Subdomain II, respectively. The remaining interior points $\{15, 16\}$ are the interface points that link both subdomains.

The partial differential equation that models the steady-state temperature distribution is given by

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

subject to values that are fixed for u at the boundaries. Let the boundary points described by the various points shown in Figure B.7 have the following values:

$$(u_a, u_b, u_c, u_d, u_e, u_f, u_g) = (100, 90, 80, 70, 60, 50, 40)$$

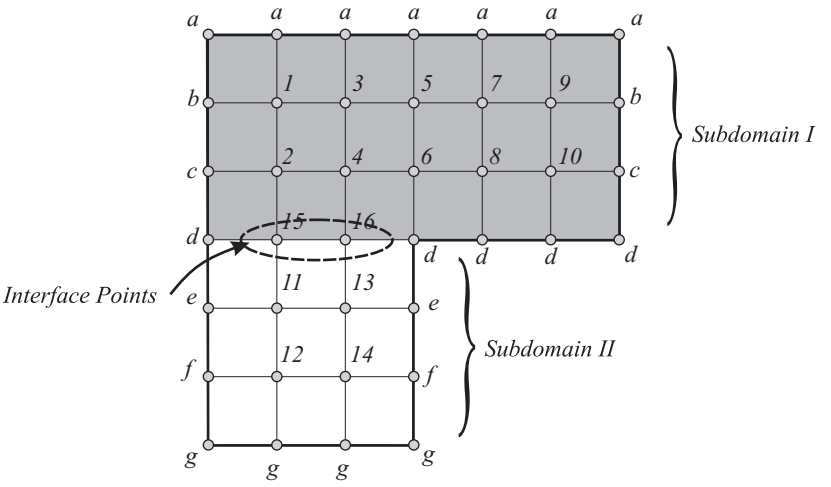


Figure B.7. The labeling of various points for Example B.10.

Using finite difference approximations (cf. Example 1.3) with $\Delta x = \Delta y = 1$, the linear equation will be:

$$\left(\begin{array}{c|c|c} A_{11} & 0 & A_{13} \\ \hline 0 & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \end{array}\right)\left(\begin{array}{c} \mathbf{x}_1 \\ \hline \mathbf{x}_2 \\ \hline \mathbf{x}_3 \end{array}\right) = \left(\begin{array}{c} \mathbf{b}_1 \\ \hline \mathbf{b}_2 \\ \hline \mathbf{b}_3 \end{array}\right)$$

where,

$$A_{11} = \left(\begin{array}{cc|cc|cc|cc|cc} -4 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & -4 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & -4 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & -4 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -4 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & -4 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & -4 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -4 \end{array}\right) ; A_{13} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$A_{22} = \left(\begin{array}{cc|cc} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ \hline 1 & 0 & -4 & 1 \\ 0 & 1 & 1 & -4 \end{array}\right) \quad ; \quad A_{23} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

$$A_{31} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{32} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} ; A_{33} = \begin{pmatrix} -4 & 1 \\ 1 & -4 \end{pmatrix}$$

$$\mathbf{b}_1^T = \begin{pmatrix} 190 & 80 & 100 & 70 & 100 & 70 & 100 & 70 & 190 & 150 \end{pmatrix}$$

$$\mathbf{b}_2^T = \begin{pmatrix} 60 & 90 & 60 & 90 \end{pmatrix}$$

$$\mathbf{b}_3^T = \begin{pmatrix} 70 & 70 \end{pmatrix}$$

$$\begin{aligned}
\mathbf{x}_1^T &= \begin{pmatrix} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10} \end{pmatrix} \\
\mathbf{x}_2^T &= \begin{pmatrix} u_{11} & u_{12} & u_{13} & u_{14} \end{pmatrix} \\
\mathbf{x}_3^T &= \begin{pmatrix} u_{15} & u_{16} \end{pmatrix}
\end{aligned}$$

Note that the structures of A_{11}, A_{22}, and A_{33} are block tri-diagonal, whose inverse can be obtained using the block LU methods (or more specifically, using the block-Thomas algorithm; see Exercise E2.16).

Let us now focus on the solution of $A\mathbf{x} = \mathbf{b}$. With $A = M + S$,

$$A\mathbf{x} = (M+S)\,\mathbf{x} = \mathbf{b} \quad \rightarrow \quad (I+H)\,\mathbf{x} = \mathbf{z}$$

where $H = M^{-1}S$ and $\mathbf{z} = M^{-1}\mathbf{b}$. Choosing M to be

$$M = \left(\begin{array}{c|c|c} A_{11} & & \mathbf{0} \\ \hline & \ddots & \\ \hline \mathbf{0} & & A_{nn} \end{array}\right)$$

we have

$$H = M^{-1}S = \left(\begin{array}{c|c|c|c} A_{11}^{-1} & & & 0 \\ \hline & A_{22}^{-1} & & \\ \hline & & \ddots & \\ \hline 0 & & & A_{nn}^{-1} \end{array}\right)\left(\begin{array}{c|c|c|c} 0 & & & A_{1,n} \\ \hline & \ddots & & \vdots \\ \hline & & 0 & A_{n-1,n} \\ \hline A_{n,1} & \cdots & A_{n,n-1} & 0 \end{array}\right)$$

$$= \left(\begin{array}{c|c|c|c} 0 & & & A_{11}^{-1}A_{1,n} \\ \hline & \ddots & & \vdots \\ \hline & & 0 & A_{n-1,n-1}^{-1}A_{n-1,n} \\ \hline A_{nn}^{-1}A_{n,1} & \cdots & A_{nn}^{-1}A_{n,n-1} & 0 \end{array}\right)$$

Let $B_k = A_{kk}^{-1}A_{kn}$ and $\Omega = \left(A_{nn} - \sum_{k=1}^{n-1} A_{n,k}B_k\right)^{-1}$. Note that the product ΩA_{nn} is the inverse of the Schur complement of A_{nn}. Using the block matrix inverse formula given in (1.36), we obtain

$$\mathbf{x} = (I+H)^{-1}\,\mathbf{z} = \left(\begin{array}{c|c} W & X \\ \hline Y & Z \end{array}\right)\mathbf{z} \tag{B.24}$$

where,

$$Z = \Omega A_{nn} \quad ; \quad Y = -\Omega \left(\begin{array}{c|c|c} A_{n1} & \cdots & A_{n,n-1} \end{array} \right)$$

$$X = -\left(\begin{array}{c} \hline B_1 \\ \hline \vdots \\ \hline B_{n-1} \end{array} \right) \Omega A_{nn} \quad ; \quad W = I + \left(\begin{array}{c} \hline B_1 \\ \hline \vdots \\ \hline B_{n-1} \end{array} \right) \Omega \left(\begin{array}{c|c|c} A_{n1} & \cdots & A_{n,n-1} \end{array} \right)$$

EXAMPLE B.11. We can implement the Schur complement method to the problem in Example B.10. The value of Ω and $\mathbf{z}$ can be found to be

$$\Omega = \begin{pmatrix} -0.3331 & -0.1161 \\ -0.1161 & -0.3362 \end{pmatrix}$$

$$\mathbf{z}^T = \Big(80.357, 52.698, 78.731, 50.436, 84.131, 70.313, 87.481, 76.686,$$
$$89.107, 78.948, 33.75, 41.25, 33.75, 41.25, 23.333, 23.333\Big)$$

and with (B.24), the solution is

$$u^T = \Big(90, 80, 90, 80, 90, 80, 90, 80, 90, 80, 60, 50, 60, 50, 70, 70\Big)$$

which is expected because the given boundary conditions in B.10 show a linear temperature distribution.

B.7 Linear Vector Algebra: Fundamental Concepts

In this section, we give some of the fundamental concepts of linear algebra of vectors. Matrices are treated as collections of column vectors, and thus the product $A\mathbf{x}$ is the process of linearly combining the columns of A scaled by the entries in $\mathbf{x}$.

Let $\mathcal{F}$ be a field of scalars, for example, the fields of real numbers or field of complex numbers. The abstract definition of a **linear vector space** $\mathcal{L}$ (over $\mathcal{F}$) is a collection of objects called **vectors** such that a sum operation is closed; that is, if $\mathbf{v}$ and $\mathbf{w}$ are in $\mathcal{L}$, then so is their sum $\mathbf{v} + \mathbf{w}$. Furthermore, the vector sum operations and the scalar product operations need to obey the conditions given in Table B.2. Some useful definitions and concepts connected with linear vector spaces are given in Table B.3.

To illustrate the idea of span, consider the two vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \\ -1 \end{pmatrix}$$

Based on the definition given in Table B.3, the span of $\mathbf{v}_1$ and $\mathbf{v}_2$ is the collection of all vectors obtained by a linear combination of these two vectors. A representative vector is then given by

$$\mathbf{v} = a\mathbf{v}_1 + b\mathbf{v}_2$$

Table B.2. *Conditions for a linear vector space*

	Conditions for Vector Sums	
1	Associative	$\mathbf{v}+(\mathbf{w}+\mathbf{y})=(\mathbf{v}+\mathbf{w})+\mathbf{y}$
2	Commutative	$\mathbf{v}+\mathbf{w}=\mathbf{w}+\mathbf{v}$
3	Identity is $\mathbf{0}$	$\mathbf{0}+\mathbf{v}=\mathbf{v}$
4	Inverse exist and unique	$\mathbf{v}+(-\mathbf{v})=\mathbf{0}$
	Conditions for Scalar Products	
1	Associative	$\alpha(\beta\mathbf{v})=(\alpha\beta)\mathbf{v}$
2	Identity is 1	$1\mathbf{v}=\mathbf{v}$
3	Vector is distributive over scalar sums	$(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}$
4	Scalar is distributive over vector sums	$\alpha(\mathbf{v}+\mathbf{w})=\alpha\mathbf{v}+\alpha\mathbf{w}$

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 2b \\ a+b \\ a-b \end{pmatrix}$$

or with x and y as independent variables,

$$z = y - x$$

which is the equation of a 2D plane. Next, consider another point

$$\mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

This point is no longer in the span of $\mathbf{v}_1$ and $\mathbf{v}_2$ because the three elements of $\mathbf{v}_3$ do not satisfy $z = y - x$.

Table B.3. *Some important definitions for linear vector spaces*

	Terms and concepts	Conditions
1	$\mathbf{w}$ is a **linear combination** of $\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ based on $\{\alpha_1,\ldots,\alpha_K\}$	$\mathbf{w}=\sum_{i=1}^{K}\alpha_i\mathbf{v}_i$
2	**Span** of $\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ is the space of possible linear combinations	$\mathbf{Span}(\mathbf{v}_1,\ldots,\mathbf{v}_K)=\{\mathbf{w}\}$ such that $\mathbf{w}=\sum_{i=1}^{K}\alpha_i\mathbf{v}_i$ for $\alpha_i\in\mathcal{F}$
3	$\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ are **linearly independent**	$\sum_{i=1}^{k}\alpha_i\mathbf{v}_i=\mathbf{0}$ only if $\alpha_i=0$ for all i
4	$\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ are **linearly dependent**	$\sum_{i=1}^{k}\alpha_i\mathbf{v}_i=\mathbf{0}$ for some $\alpha_i\neq 0$
5	$\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ is the **basis** of subspace $\mathcal{S}$	$\{\mathbf{v}_1,\ldots,\mathbf{v}_K\}$ is linearly independent, and $\mathbf{Span}(\mathbf{v}_1,\ldots,\mathbf{v}_K)=\mathcal{S}$
6	An integer $\mathbf{d}=\mathbf{dim}(\mathcal{S})$ is the **dimension** of subspace $\mathcal{S}$	There exist $\{\mathbf{v}_1,\ldots,\mathbf{v}_\mathbf{d}\}$ that is a basis of S

Table B.4. *Conditions for vector norms*

1	Positivity	$\|\mathbf{v}\| \geq 0$
2	Scaling	$\|\alpha\mathbf{v}\| = \|\alpha\| \, \|\mathbf{v}\|$
3	Triangle Inequality	$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$
4	Unique Zero	$\|\mathbf{v}\| = 0$ only if $\mathbf{v} = 0$

The space of column vectors, using the matrix algebra operations discussed in Section 1.2, satisfy the conditions in Table B.2. Vectors $\mathbf{v}_1, \ldots, \mathbf{v}_M$ (each of length N) can then be linearly combined using scalars x_i, that is,

$$\sum_{i=1}^{M} x_i \mathbf{v}_i$$

or $A\mathbf{x}$, where

$$A = \left(\begin{array}{c|c|c} \mathbf{v}_1 & \cdots & \mathbf{v}_M \end{array} \right)$$

Let $A[=]N \times M$; then an exact solution to $A\mathbf{x} = \mathbf{b}$ written out as

$$x_1 A_{\bullet,1} + \ldots + x_M A_{\bullet,M} = \mathbf{b}$$

means that $\mathbf{b}$ has to be linearly dependent on the columns of A; that is, $\mathbf{b}$ has to reside in the span of the columns of A. The dimension of the span of the columns of A is also the rank of A. This means that the rank of A simply determines how many columns of A are linearly independent. Thus if we augment the columns of A with $\mathbf{b}$ and find an increase in rank, this could only mean that $\mathbf{b}$ is independent of the columns of A.

The evaluation of exact solutions has already been discussed in Chapter 1 and 2. However, if the columns of A and $\mathbf{b}$ have lengths larger than the number of columns in A, that is, $N > M$, then an exact match will not be likely. Instead, the problem becomes the search for a linear combination of the columns of A that match $\mathbf{b}$ as close as possible, based on some chosen measure.

Thus one needs to equip the linear vector space with a measure called the norm. Returning to the abstract linear vector space $\mathcal{L}$, a **norm** is a function that assigns a positive real number to the vectors of $\mathcal{L}$. We denote the norm of $\mathbf{v}$ by $\|\mathbf{v}\|$. Furthermore, this function needs to satisfy the conditions given in Table B.4.

Based on a chosen norm, a vector $\mathbf{v} \neq \mathbf{0}$ can always be **normalized** by scaling $\mathbf{v}$ by the scalar $\alpha = \|\mathbf{v}\|^{-1}$, that is,

$$\left\| \frac{1}{\|\mathbf{v}\|} \mathbf{v} \right\| = \frac{1}{\|\mathbf{v}\|} \|\mathbf{v}\| = 1 \tag{B.25}$$

Among the various possible norms for matrix vectors, we have the **Euclidean norm**, denoted $\|\mathbf{v}\|_2$ defined by

$$\|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^* \mathbf{v}} = \sqrt{\sum_{i=1}^{N} \overline{v}_i v_i} \tag{B.26}$$

In most cases, we default to the Euclidean norms and drop the subscript '2', unless the discussion involves other types of norms. One can show that this definition

satisfies the conditions given in Table B.4 (we include the proof that (B.26) is a norm in Section B.10.1 as an appendix.) If the vectors are represented by points in a N-hyperspace, then the norm is simply the distance of the points from the origin. Note also that only the zero vector will have a zero norm.

B.8 Determination of Linear Independence of Functions

Given a set of multivariable functions, $\{f_1(\mathbf{v}), \ldots, f_M(\mathbf{v})\}$, where $\mathbf{v} = (v_1, \ldots, v_K)$ are independent variables. The functions are linearly independent if and only if the only $\alpha_1 = \cdots = \alpha_M = 0$ is the unique solution of

$$\alpha_1 f_1(\mathbf{v}) + \ldots + \alpha_M f_M(\mathbf{v}) = 0 \tag{B.27}$$

One method to determine whether functions $\{f_1, \ldots, f_M\}$ are linearly independent is the Wronskian approach, extended to multivariable functions. First, take the linear combination

$$\alpha_1 f_1(\mathbf{x}) + \ldots + \alpha_M f_M(\mathbf{x}) = 0 \tag{B.28}$$

Next, generate several partial derivatives of this equation, that is,

$$\begin{pmatrix} f_1 & \cdots & f_M \\ \partial f_1/\partial v_1 & \cdots & \partial f_M/\partial v_1 \\ \partial f_1/\partial v_2 & \cdots & \partial f_M/\partial v_2 \\ & \vdots & \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_M \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \end{pmatrix}$$

Enough equations are generated until a nonsingular submatrix can be obtained. If this occurs, then it would establish that $\{f_1, \ldots, f_m\}$ are linearly independent. However, the Wronskian approach requires the evaluation of partial derivatives and determinants that involve the independent variables. This means that except for small number of functions, the general case will be cumbersome to solve symbolically.

Another method, called the **substitution approach**, first chooses different values for $\mathbf{v}$, say, $\mathbf{v}_i$ with $i = 1, \ldots, M$, and then substitutes them into $f_j(\mathbf{v})$. Matrix $\widehat{A}$ can then be formed as follows:

$$\widehat{A} = \begin{pmatrix} f_1(\mathbf{v}_1) & \cdots & f_M(\mathbf{v}_1) \\ \vdots & \ddots & \vdots \\ f_1(\mathbf{v}_M) & \cdots & f_M(\mathbf{v}_M) \end{pmatrix}$$

If $\widehat{A}$ is nonsingular, we conclude that $\{f_1(\mathbf{v}), \ldots, f_M(\mathbf{v})\}$ are linearly independent.

EXAMPLE B.12. Consider the linear-in-parameter model:

$$y = a_0 + a_1 v + \cdots + a_{M-1} v^{M-1} \tag{B.29}$$

Here, we have one independent variable v. The functions are $f_i(v) = v^{i-1}$.

Using the Wronskian approach, we have

$$W \begin{pmatrix} a_0 \\ \vdots \\ a_{M-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where

$$W = \begin{pmatrix} 1 & v & \cdots & v^{M-1} \\ 0 & 1 & \cdots & (M-1)\, v^{M-2} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & (M-1)! \end{pmatrix}$$

The determinant of W is given by

$$|W| = (M-1)!\,(M-2)! \cdots 1 \neq 0$$

This shows that $\left\{1, v, \cdots v^{M-1}\right\}$ form a linearly independent set of functions.

Using the substitution approach, we could set different constants for v, that is, $v = \lambda_1, \cdots, \lambda_{M-1}$, and substitute each one to obtain $\widehat{A}$,

$$\widehat{A} = \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{M-1} \\ 1 & \lambda_2 & \cdots & \lambda_2^{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{M-1} & \cdots & \lambda_{M-1}^{M-1} \end{pmatrix}$$

which is a Vandermonde matrix. The determinant will be nonzero as long as $\lambda_1, \cdots, \lambda_{M-1}$ are all distinct. Thus using this approach, we obtain the same conclusion about the linear independence of $\left\{1, v, \cdots v^{M-1}\right\}$.

The model given by (B.29) is a very popular empirical nonlinear model known as the **polynomial-fitting model**.

EXAMPLE B.13. Consider another linear-in-parameter model:

$$y = a_0 + a_1 v_1^2 + a_2 (v_1 - v_2)\, v_2 + a_3 v_2^2 \tag{B.30}$$

Here, we have two independent variables v_1 and v_2. The functions are $f_1(\mathbf{v}) = 1$, $f_2(\mathbf{v}) = v_1^2$, $f_3(\mathbf{v}) = (v_1 - v_2)\, v_2$ and $f_4(\mathbf{v}) = v_2^2$.

Using the extended-Wronskian approach, we have

$$W = \begin{pmatrix} 1 & v_1^2 & (v_1 - v_2)\, v_2 & v_2^2 \\ 0 & 2v_2 & v_2 & 0 \\ 0 & 0 & -2v_2 & 2v_2 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 2 \end{pmatrix}$$

We can take two different tracks. The first is to take the determinant of the Grammian $W^T W$. This will mean a 4×4 determinant involving symbolic manipulations. The other method is to choose rows and determine whether a nonsingular submatrix emerges. We show the second track by choosing rows 1, 4, 5, and 6. Doing so, we have

$$W_{[1,4,5,6]} = \begin{pmatrix} 1 & v_1^2 & (v_1 - v_2)\, v_2 & v_2^2 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 2 \end{pmatrix}$$

whose determinant is 4. Thus the functions are linearly independent.

Using the substitution method, we can choose

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Substituting these values to the various functions, we obtain

$$\widehat{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -2 & 1 \end{pmatrix}$$

whose determinant is -2. Thus it also shows that the functions are linearly independent.

B.9 Gram-Schmidt Orthogonalization

Suppose we have a set of linearly independent vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, each of length N, which are basis vectors that span an N-dimensional space. In some cases, the vectors may be too close to each other. The Gram-Schmidt orthogonalization is a simple procedure to obtain a better set of basis vectors with same span, but are perpendicular (or orthogonal) to each other. The Gram-Schmidt algorithm is one procedure to obtain these mutually perpendicular basis vectors.

Definition B.2. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors of the same length. The* **inner product** *of* $\mathbf{a}$ *and* $\mathbf{b}$*, denoted by* $\langle \mathbf{a}, \mathbf{b} \rangle$*, is given by*

$$\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^* \mathbf{b} \tag{B.31}$$

Definition B.3. *Let* $\mathbf{a}$ *and* $\mathbf{b}$ *be two vectors of the same length. Then* $\mathbf{a}$ *and* $\mathbf{b}$ *are* **orthogonal** *to each other if* $\langle \mathbf{a}, \mathbf{b} \rangle = 0$*. A set of vectors* $\mathbf{z}_1, \ldots, \mathbf{z}_N$ *is an* **orthonormal** *set if*

$$\langle \mathbf{z}_i, \mathbf{z}_j \rangle = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{B.32}$$

Gram-Schmidt Algorithm:

Let $\{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ be linearly independent. Set $\mathbf{z}_1 = \dfrac{\mathbf{a}_1}{\|\mathbf{a}_1\|}$ For $k = 2, \ldots, N$,

$$\begin{aligned} \mathbf{y}_k &= \mathbf{a}_k - \sum_{i=1}^{k-1} \langle \mathbf{a}_i, \mathbf{z}_i \rangle \mathbf{z}_i \\ \mathbf{z}_k &= \frac{\mathbf{y}_k}{\|\mathbf{y}_k\|} \end{aligned}$$

Then $\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ is an orthonormal set.

EXAMPLE B.14. Given

$$\mathbf{a}_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad \mathbf{a}_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \qquad \mathbf{a}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Using the Gram-Schmidt method, we obtain

$$\mathbf{z}_1 = \begin{pmatrix} 0.408 \\ 0.816 \\ 0.408 \end{pmatrix} \qquad \mathbf{z}_2 = \begin{pmatrix} -0.436 \\ -0.218 \\ 0.873 \end{pmatrix} \qquad \mathbf{z}_3 = \begin{pmatrix} 0.802 \\ -0.534 \\ 0.267 \end{pmatrix}$$

We can check that $\langle \mathbf{z}_1, \mathbf{z}_1 \rangle = \langle \mathbf{z}_2, \mathbf{z}_2 \rangle = \langle \mathbf{z}_3, \mathbf{z}_3 \rangle = 1$, and $\langle \mathbf{z}_i, \mathbf{z}_j \rangle = 0$ for $i \neq j$.

B.10 Proofs for Lemma and Theorems in Chapter 2

B.10.1 Proof That Euclidean Norm Is a Norm

We need to show that the Euclidean norm defined in (B.26)

$$\|\mathbf{v}\| = \sqrt{\mathbf{v}^*\mathbf{v}} = \sqrt{\sum_{i=1}^{N} \overline{v}_i v_i}$$

satisfies each of the requirements of Table B.4.

1. Positivity is immediate from the definition, because $\overline{v}v = \mathrm{Re}\,(v)^2 + \mathrm{Im}\,(v)^2$.
2. The scaling property is shown as follows:

$$\begin{aligned} \|\alpha\mathbf{v}\| &= \sqrt{\overline{\alpha}\alpha} \sqrt{\sum_{i=1}^{N} \overline{v}_i v_i} \\ &= |\alpha| \, \|\mathbf{v}\| \end{aligned}$$

3. Because $\overline{v}_i v_i = 0$ if and only if $v_i = 0$, the only vector that will yield a zero norm is $\mathbf{v} = \mathbf{0}$.
4. The triangle inequality is more involved. It requires a relationship known as **Cauchy-Schwarz inequality**:

$$\left|\mathbf{v}^*\mathbf{w}\right| \leq \|\mathbf{v}\| \, \|\mathbf{w}\| \tag{B.33}$$

The proof of the Cauchy-Schwarz inequality is given later. For now, we apply (B.33) to prove the triangle inequality of Euclidean norms.

$$\begin{aligned} \|\mathbf{v} + \mathbf{w}\|^2 &= \mathbf{v}^*\mathbf{v} + \mathbf{v}^*\mathbf{w} + \mathbf{w}^*\mathbf{v} + \mathbf{w}^*\mathbf{w} \\ &\leq \|\mathbf{v}\|^2 + \left|\mathbf{v}^*\mathbf{w}\right| + \left|\mathbf{w}^*\mathbf{v}\right| + \|\mathbf{w}\|^2 \\ &\leq \|\mathbf{v}\|^2 + 2\,\|\mathbf{v}\|\,\|\mathbf{w}\| + \|\mathbf{w}\|^2 = (\|\mathbf{v}\| + \|\mathbf{w}\|)^2 \end{aligned}$$

Thus

$$\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$$

PROOF[6] of Cauchy-Schwarz inequality, **Equation (B.33):**
For complex numbers a and b,

$$|\overline{a}b| = \left| |a|e^{-i\,(\arg(a))}|b|e^{i\,(\arg(b))} \right| \leq |a||b| \tag{B.34}$$

and

$$\begin{aligned} 0 &\leq (|a| - |b|)^2 \\ 0 &\leq |a|^2 - 2|a||b| + |b|^2 \\ 2|a||b| &\leq |a|^2 + |b|^2 \\ |a||b| &\leq \frac{1}{2}\left(|a|^2 + |b|^2\right) \end{aligned} \tag{B.35}$$

Combining (B.34) and (B.35),

$$|\overline{a}b| \leq \frac{1}{2}\left(|a|^2 + |b|^2\right) \tag{B.36}$$

Next let $\mathbf{a}$ and $\mathbf{b}$ be normalized vectors defined by

$$\mathbf{a} = \frac{1}{\|\mathbf{v}\|}\mathbf{v} \quad \text{and} \quad \mathbf{a} = \frac{1}{\|\mathbf{w}\|}\mathbf{w}$$

Applying (B.36) plus the fact that $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$,

$$\begin{aligned} \left|\mathbf{a}^*\mathbf{b}\right| &= \left|\sum_{i=1}^{N} \overline{a}_i b_i\right| \\ &\leq \frac{1}{2}\left(\sum_{i=1}^{N} |a_i|^2 + \sum_{i=1}^{N} |b_i|^2\right) = 1 \end{aligned}$$

Then

$$\begin{aligned} \left|\mathbf{a}^*\mathbf{b}\right| &\leq 1 \\ \frac{|\mathbf{v}^*\mathbf{w}|}{\|\mathbf{v}\|\,\|\mathbf{w}\|} &\leq 1 \\ \left|\mathbf{v}^*\mathbf{w}\right| &\leq \|\mathbf{v}\|\,\|\mathbf{w}\| \end{aligned}$$

B.10.2 Proof for Levenberg-Marquardt Update Form (Lemma B.2)

(The proof given here is based on Dennis and Schnabel *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, 1983.)

Let $\varphi\left(\Delta^k\mathbf{x}\right)$ be the function to be minimized in (B.109),

$$\varphi\left(\Delta^k\mathbf{x}\right) = \frac{1}{2}\left\|\mathbf{r}_k + J_k\Delta^k\mathbf{x}\right\|^2 = \frac{1}{2}\mathbf{r}_k^T\mathbf{r}_k + \mathbf{r}_k^T J_k\Delta^k\mathbf{x} + \left(\Delta^k\mathbf{x}\right)^T J_k^T J_k\Delta^k\mathbf{x}$$

[6] We limit the proof only for the case of Euclidean norms, although the Cauchy-Schwarz can be applied to different norms and inner products.

If the minimum of φ lies inside the trust region, then the problem resembles the unconstrained problem, whose solution is immediately given by

$$\Delta^k \mathbf{x}^* = -\left(J_k^T J_k\right)^{-1} J_k^T \mathbf{r}_k$$

that is, $\mu = 0$.

However, if the trust region is smaller, then the minima will be on the boundary of the trust region, that is, $\left\|\Delta^k \mathbf{x}^*\right\| = M_k$. This means that the solution to the constrained minimization problem given in (B.109) will be a step $\Delta^k \mathbf{x}^*$ such that when we perturb it by another vector, say $\mathbf{v}$, the value of φ can be decreased only at the expense of moving it outside the trust region.

A perturbation $\mathbf{v}$ will minimize $\varphi\left(\Delta^k \mathbf{x}^*\right)$ further only if

$$0 \;<\; \varphi\left(\Delta^k \mathbf{x}^* + \mathbf{v}\right) - \varphi\left(\Delta^k \mathbf{x}^*\right) \quad = \quad \left(\mathbf{r}^T J_k + \left(\Delta^k \mathbf{x}^*\right)^T J_k^T J_k\right)\mathbf{v} + \mathbf{v}^T J_k^T J_k \mathbf{v}$$

or, because $\mathbf{v}^T J_k^T J_k \mathbf{v} \geq 0$,

$$\left(\mathbf{r}^T J_k + \left(\Delta^k \mathbf{x}^*\right)^T J_k^T J_k\right)\mathbf{v} > 0 \tag{B.37}$$

The other requirement for $\mathbf{v}$ is that the perturbed step $\Delta^k \mathbf{x}^* + \mathbf{v}$ will have a norm greater than M_k, that is,

$$\left\|\Delta^k \mathbf{x}^* + \mathbf{v}\right\| > \left\|\Delta^k \mathbf{x}^*\right\| \qquad \rightarrow \qquad \left(\Delta^k \mathbf{x}^*\right)^T \mathbf{v} > 0 \tag{B.38}$$

The implication is that the vectors premultiplying $\mathbf{v}$ in (B.37) and (B.38) must point in the opposite directions, or

$$J_k^T \mathbf{r} + J_k^T J_k \Delta^k \mathbf{x}^* = -\mu\left(\Delta^k \mathbf{x}^*\right)$$

for some $\mu > 0$. Thus $\Delta_k \mathbf{x}^*$ is given by the form

$$\Delta^k \mathbf{x}^* = -\left(J_k^T J_k + \mu I\right)^{-1} J_k^T \mathbf{r}_k$$

To show uniqueness, let

$$s(\mu) = \left(J_k^T J_k + \mu I\right)^{-1} J_k^T \mathbf{r}$$

and let $q(\mu)$ be the difference

$$q(\mu) = \left\|s(\mu)\right\| - M_k$$

whose derivative is given by

$$\frac{dq}{d\mu} = -\frac{\mathbf{r}^T J_k \left(J_k^T J_k + \mu I\right)^{-3} J_k^T \mathbf{r}}{\left\|s(\mu)\right\|}$$

The derivative $dq/d\mu$ is always negative for $\mu > 0$, and equal to zero only when $\mathbf{r}_k^T J_k = 0$ (which occurs only when $\mathbf{x}^{[k]}$ is already the minimum of φ). This implies that $q(\mu)$ is zero only for a unique value of μ.

B.11 Conjugate Gradient Algorithm

B.11.1 The Algorithm

We begin with some additional terms and notations. The error vector is the difference of $\mathbf{x}^{(i)}$ from the exact solution $\widehat{\mathbf{x}}$, denoted by $\mathbf{err}^{(i)}$,

$$\mathbf{err}^{(i)} = \mathbf{x}^{(i)} - \widehat{\mathbf{x}} \tag{B.39}$$

The mismatch between $\mathbf{b}$ and $A\mathbf{x}^{(i)}$ is called the i^{th} residual vector, denoted by $\mathbf{r}^{(i)}$, that is,

$$\mathbf{r}^{(i)} = \mathbf{b} - A\mathbf{x}^{(i)} \tag{B.40}$$

By taking the gradient of $f(\mathbf{x})$, we can see that the residual vector is the transpose of the negative gradient of $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^{(i)}$, that is,

$$\frac{d}{d\mathbf{x}} f(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}^{(i)}} = \left(\mathbf{x}^T A - \mathbf{b}^T\right)\big|_{\mathbf{x}=\mathbf{x}^{(i)}} = -(\mathbf{r}^{(i)})^T \tag{B.41}$$

The relationship between the residual vectors and error vectors can be obtained by adding $A\widehat{\mathbf{x}} - \mathbf{b} = 0$ to (B.40),

$$\begin{aligned} \mathbf{r}^{(i)} &= \mathbf{b} - A\mathbf{x}^{(i)} + (A\widehat{\mathbf{x}} - \mathbf{b}) \\ &= -A\left(\mathbf{x}^{(i)} - \widehat{\mathbf{x}}\right) = -A\,\mathbf{err}^{(i)} \end{aligned} \tag{B.42}$$

Returning to the main problem, we formulate the following update equation,

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)} \tag{B.43}$$

where $\mathbf{d}^{(i)}$ is the i^{th} correction vector and $\alpha^{(i)}$ is a factor that will scale the correction vector optimally for the i^{th} update. (The choice for $\mathbf{d}^{(i)}$ and $\alpha^{(i)}$ is discussed later). The residual vector is

$$\mathbf{r}^{(i+1)} = \mathbf{b} - A\mathbf{x}^{(i+1)} \tag{B.44}$$

A more efficient calculation for $\mathbf{r}^{(i+1)}$ can be used. Taking (B.43), we can subtract $\widehat{\mathbf{x}}$ from both sides, multiply by A, and then use (B.42),

$$\begin{aligned} \mathbf{x}^{(i+1)} - \widehat{\mathbf{x}} &= \mathbf{x}^{(i)} - \widehat{\mathbf{x}} + \alpha^{(i)}\mathbf{d}^{(i)} \\ \mathbf{err}^{(i+1)} &= \mathbf{err}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)} \\ A\,\mathbf{err}^{(i+1)} &= A\,\mathbf{err}^{(i)} + \alpha^{(i)}A\mathbf{d}^{(i)} \\ \mathbf{r}^{(i+1)} &= \mathbf{r}^{(i)} - \alpha^{(i)}A\mathbf{d}^{(i)} \end{aligned} \tag{B.45}$$

Although (B.45) is the preferred update equation, it can sometimes accumulate round-off errors. For very large problems, most implementations of the conjugate gradient method include an occasional switch to (B.44) once every K iterations (e.g., $K \leq 50$) and then switch back to (B.45).

The initial direction vector is usually chosen as the initial residual vector,[7] that is,

$$\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} \tag{B.46}$$

[7] This means that the conjugate gradient method begins with the same search direction as a gradient descent method, because $\mathbf{r}^{(0)}$ is the negative gradient of $f(\mathbf{x})$ at $\mathbf{x}^{(0)}$.

Afterward, the next correction vector will be a combination of the previous correction vectors and the most recent residual vector, that is,

$$\mathbf{d}^{(i+1)} = \gamma^{(i+1)}\mathbf{r}^{(i+1)} + \beta^{(i+1)}\mathbf{d}^{(i)} \tag{B.47}$$

where $\gamma^{(i+1)}$ and $\beta^{(i+1)}$ are weighing factors.

All that remains is to choose the three factors: $\alpha^{(i)}$, $\beta^{(i)}$, and $\gamma^{(i)}$. These values are obtained such that:

1. Each i^{th} direction vector $\mathbf{d}^{(i)}$ is independent from the previous direction vectors $\mathbf{d}^{(j)}$ with $j < i$. Specifically, they are chosen to be conjugate (i.e., A-orthogonal) to previous direction vectors,

$$(\mathbf{d}^{(i)})^T A\mathbf{d}^{(j)} = 0 \qquad \text{for } j < i \tag{B.48}$$

2. The i^{th} residual vector is orthogonal to previous residual vectors, and it is also orthogonal to previous direction vectors, that is,

$$(\mathbf{r}^{(i)})^T\mathbf{r}^{(j)} = 0\,;\ (\mathbf{r}^{(i)})^T\mathbf{d}^{(j)} = 0 \qquad \text{for } j < i \tag{B.49}$$

As is shown later in Lemma B.1, these criteria are achieved by using the following values for the scaling factors:

$$\gamma^{(i+1)} = 1\ ;\ \alpha^{(i)} = \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}}\ ;\ \beta^{(i+1)} = \frac{(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}} \tag{B.50}$$

Putting all these pieces together, we have the following algorithm:

Algorithm of Conjugate Gradient.

1. **Initialize:** For a given symmetric matrix $A[=]N \times N$, vector $\mathbf{b}$, and initial guess $\mathbf{x}^{(0)}$, set

$$\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)} \tag{B.51}$$

2. **Update:** For a specified maximum number of iterations, $i_{\max} \geq N$ and specified tolerance $\epsilon \ll 1$, perform the following steps:

   Although $i < i_{\max}$, $\left|(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}\right| > 0$ and $\beta^{(i)} > \epsilon$,

$$\alpha^{(i)} = \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}} \tag{B.52}$$

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)} \tag{B.53}$$

$$\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \alpha^{(i)}A\mathbf{d}^{(i)} \tag{B.54}$$

$$\beta^{(i+1)} = -\frac{(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}} \tag{B.55}$$

$$\mathbf{d}^{(i+1)} = \mathbf{r}^{(i+1)} + \beta^{(i+1)}\mathbf{d}^{(i)} \tag{B.56}$$

The relationships among the various residual vectors and direction vectors are outlined in the following lemma:

LEMMA B.1. *For $i \geq 1$ and $j < i$, using the conjugate gradient algorithm (B.51) to (B.56), we have the following identities for $\mathbf{r}^{(i)}$ and $\mathbf{d}^{(i)}$:*

$$(\mathbf{r}^{(i)})^T \mathbf{r}^{(j)} = 0 \tag{B.57}$$

$$(\mathbf{r}^{(i)})^T \mathbf{d}^{(j)} = 0 \tag{B.58}$$

$$(\mathbf{r}^{(i)})^T \mathbf{r}^{(i)} = (\mathbf{r}^{(i)})^T \mathbf{d}^{(i)} \tag{B.59}$$

$$(\mathbf{d}^{(i)})^T A\mathbf{d}^{(j)} = 0 \tag{B.60}$$

$$(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)} = (\mathbf{r}^{(i)})^T A\mathbf{d}^{(i)} \tag{B.61}$$

$$\frac{(\mathbf{r}^{(i)})^T A\mathbf{d}^{(i-1)}}{(\mathbf{d}^{(i-1)})^T A\mathbf{d}^{(i-1)}} = -\frac{(\mathbf{r}^{(i)})^T \mathbf{r}^{(i)}}{(\mathbf{r}^{(i-1)})^T \mathbf{r}^{(i-1)}} \tag{B.62}$$

$$(\mathbf{r}^{(i)})^T A\mathbf{r}^{(i-1)} = (\mathbf{r}^{(i)})^T A\mathbf{d}^{(i-1)} \tag{B.63}$$

$$(\mathbf{d}^{(i)})^T A\mathbf{r}^{(j)} = 0 \tag{B.64}$$

$$(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(j)} = 0 \tag{B.65}$$

$$(\mathbf{r}^{(i+1)})^T A\mathbf{r}^{(j)} = 0 \tag{B.66}$$

PROOF. (See Section B.11.2.)

The properties given in Lemma B.1 have the following implications:

1. Equation (B.62) show that $\beta^{(i+1)}$ in (B.55) of the algorithm can be replaced by

$$\beta^{(i+1)} = \frac{(\mathbf{r}^{(i+1)})^T \mathbf{r}^{(i+1)}}{(\mathbf{r}^{(i)})^T \mathbf{r}^{(i)}} \tag{B.67}$$

   Because this equation is simpler to calculate, it is implemented in most conjugate gradient methods instead of (B.55).
2. Equations (B.57) and (B.58) show that the residual vectors are orthogonal to past residual vectors and past direction vectors.
3. Equation (B.60) shows that the direction vectors are A-orthogonal to past direction vectors.
4. Equations (B.64) and (B.66) shows that $\mathbf{r}^{(i+1)}$ and $\mathbf{d}^{(i)}$ are A-orthogonal to $\mathbf{r}^{(j)}$ with $j < i$.[8]
5. Equation (B.65), together with (B.60), (B.64), and (B.65), shows that both $\mathbf{r}^{(i+1)}$ and $\mathbf{d}^{(i)}$ are orthogonal to the subspace

$$\mathcal{S} = \left\{A\mathbf{r}^{(0)}, \ldots, A\mathbf{r}^{(i-1)}\right\} = \left\{A\mathbf{d}^{(0)}, \ldots, A\mathbf{d}^{(i-1)}\right\}$$

[8] Based on (B.64), we see that the updated direction vectors $\mathbf{d}^{(i+1)}$ are chosen to be A-orthogonal, or **conjugate**, to current residual vectors, $\mathbf{r}^{(i)}$, which is the gradient of $f(\mathbf{x})$ at $\mathbf{x}^{(i)}$. This is the reason why the method is called the *conjugate gradient method*.

6. Equation (B.63) underlines the fact that although $\mathbf{r}^{(i+1)}$ is A-orthogonal to $\mathbf{r}^{(j)}$ with $j < i$, $\mathbf{r}^{(i+1)}$ is *not* A-orthogonal to $\mathbf{r}^{(i)}$.

One of the more important implications is that if the round-off errors are not present, the solution can be found in a maximum of N moves, that is,

THEOREM B.1. *Let $A[=]N \times N$ be symmetric positive-definite. Then, as long as there are no round-off errors, the conjugate gradient algorithm as given by (B.51) to (B.56) will have a zero error vector after at most N iterations.*

PROOF. (See Section B.11.3.)

We now give a simple example to illustrate the inner workings of the conjugate gradient method for a 2D case.

EXAMPLE B.15. Consider the problem $A\mathbf{x} = \mathbf{b}$ where

$$A = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1.25 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} -1.25 \\ 0.875 \end{pmatrix}$$

Using the initial guess,

$$\mathbf{x}^{(0)} = \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}$$

the conjugate gradient method evaluates the following vectors:

$$\mathbf{r}^{(0)} = \mathbf{d}^{(0)} = \begin{pmatrix} -3.5 \\ -0.25 \end{pmatrix}$$

$$\longrightarrow \mathbf{x}^{(1)} = \begin{pmatrix} -0.3181 \\ 1.3701 \end{pmatrix}; \mathbf{r}^{(1)} = \begin{pmatrix} 0.0712 \\ -0.9967 \end{pmatrix}; \mathbf{d}^{(1)} = \begin{pmatrix} -3.5 \\ -0.25 \end{pmatrix}$$

$$\longrightarrow \mathbf{r}^{(2)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \mathbf{x}^{(2)} = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix} \qquad \mathbf{d}^{(2)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The method terminated after two iterations, and we see that $\mathbf{x}^{(2)}$ solves the linear equation.

To illustrate how the method proceeds, we can plot the three iterations of $\mathbf{x}$ as shown in Figure B.8. Attached to points $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$ are concentric ellipses that are the equipotential contours of $f(\mathbf{x})$, where

$$[f(\mathbf{x})] = \left[\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T \mathbf{b}\right]$$

Because A is a symmetric positive definite matrix, we could factor A to be equal to $S^T S$,

$$S = \begin{pmatrix} 1.4142 & -0.3536 \\ 0 & 1.0607 \end{pmatrix}$$

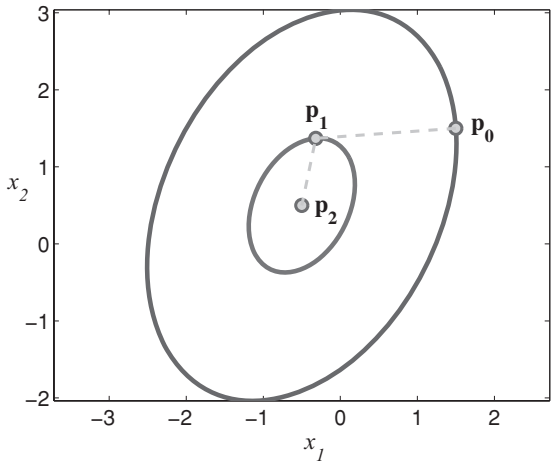


Figure B.8. A plot of the iterated solutions using the conjugate gradient method. The ellipses containing $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(1)}$ are the contours where $f(\mathbf{x})$ =constant.

Then we could plot the same points $\mathbf{x}^{(0)}$, $\mathbf{x}^{(1)}$, and $\mathbf{x}^{(2)}$ using a new coordinate system,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = S\mathbf{x}$$

which yields,

$$\mathbf{y}^{(0)} = \begin{pmatrix} 1.5910 \\ 1.5910 \end{pmatrix} \qquad \mathbf{y}^{(1)} = \begin{pmatrix} -0.9342 \\ 1.4533 \end{pmatrix} \qquad \mathbf{y}^{(2)} = \begin{pmatrix} -0.8839 \\ 0.5303 \end{pmatrix}$$

The scalar function $f(\mathbf{x})$ in terms of $\mathbf{y}$ yields

$$[f]\left[\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}\right] = \left[\mathbf{x}^T S^T S\mathbf{x} - \mathbf{x}^T\mathbf{b}\right] = \left[\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\left(S^T\mathbf{b}\right)\right]$$

We can plot the same iteration points in terms of the new coordinate system shown in Figure B.9. This time the equipotential contours of f attached to the iterated points are concentric circles instead of ellipses.

Because the first direction vector was chosen to be the residual vector, that is, $\mathbf{d}^{(0)} = \mathbf{r}^{(0)}$, where $\mathbf{r}^{(0)}$ is also equal to the gradient of $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^{(0)}$, we see from Figure B.8 that the direction vector is perpendicular to the contour $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^{(0)}$. Afterward, the succeeding direction vectors are chosen A-orthogonal to the previous direction vector. We see in Figure B.8 that $\mathbf{d}^{(0)}$ is not perpendicular to $\mathbf{d}^{(1)}$. Instead, A-orthogonality between $\mathbf{d}^{(0)}$ and $\mathbf{d}^{(1)}$appears as orthogonality in Figure B.9 because

$$\left(S\mathbf{d}^{(1)}\right)^T\left(S\mathbf{d}^{(0)}\right) = (\mathbf{d}^{(1)})^T S^T S\mathbf{d}^{(0)} = (\mathbf{d}^{(1)})^T A\mathbf{d}^{(0)} = 0$$

Thus a geometric interpretation of the conjugate gradient method is that aside from the first direction vector, the next iterations will have, under the coordinate system $S\mathbf{x}$, direction vectors that are perpendicular to increasingly smaller concentric, spherical, equipotential-contours of f. However, these steps are achieved without having to solve for S or transformation to new

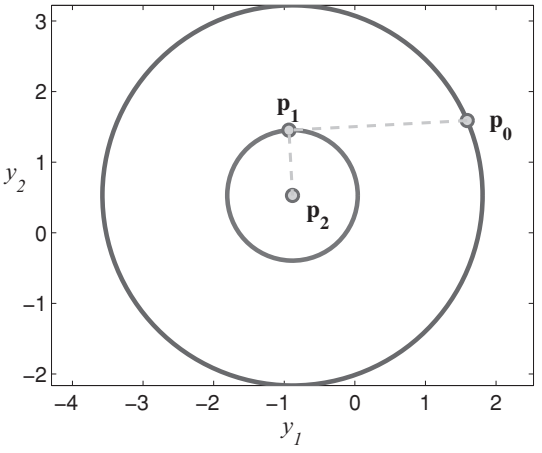


Figure B.9. A plot of the iterated solutions using the conjugate gradient method but under the new coordinates $\mathbf{y} = S\mathbf{x}$. The circles containing $\mathbf{p}^{(0)}$ and $\mathbf{p}^{(1)}$ are the contours where $f(\mathbf{y}) =$ constant.

coordinates $\mathbf{y}$.[9] Instead, the conditions of A-orthogonality of the direction vectors are achieved efficiently by the conjugate gradient method by working only in the original coordinate system of $\mathbf{x}$.

B.11.2 Proof of Properties of Conjugate Gradient method (Lemma B.1)

Based on the initial value of $r_0 = d_0$, we can show that applying (B.52) to (B.56) will satisfy (B.57) to (B.66), that is,

$$\begin{array}{ccc} (\mathbf{r}^{(1)})^T\mathbf{d}_0 = 0 & (\mathbf{r}^{(1)})^T\mathbf{r}_0 = 0 & (\mathbf{r}^{(1)})^T\mathbf{r}^{(1)} = (\mathbf{r}^{(1)})^T\mathbf{d}^{(1)} \\ (\mathbf{d}^{(1)})^TA\mathbf{d}_0 = 0 & (\mathbf{d}^{(1)})^TA\mathbf{d}^{(1)} = (\mathbf{r}^{(1)})^TA\mathbf{d}^{(1)} & (\mathbf{r}^{(1)})^TA\mathbf{r}_0 = (\mathbf{r}^{(1)})^TA\mathbf{d}_0 \\ (\mathbf{d}^{(1)})^TA\mathbf{r}_0 = 0 & (\mathbf{r}^{(2)})^TA\mathbf{r}_0 = 0 & (\mathbf{r}^{(2)})^TA\mathbf{d}_0 = 0 \end{array}$$

$$\frac{(\mathbf{r}^{(1)})^TA\mathbf{d}_0}{(\mathbf{d}_0)^TA\mathbf{d}_0} = -\frac{(\mathbf{r}^{(1)})^T\mathbf{r}^{(1)}}{(\mathbf{r}_0)^T\mathbf{r}_0} \tag{B.68}$$

Assume that the lemma is true for i and $j < i$. Then

1. Using (B.54) and (B.61),

$$(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i)} = (\mathbf{r}^{(i)})^T\mathbf{r}^{(i)} - \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{r}^{(i)} = (\mathbf{r}^{(i)})^T\mathbf{r}^{(i)} - (\mathbf{r}^{(i)})^T\mathbf{r}^{(i)} = 0$$

Whereas using (B.54), (B.57), and (B.64),

$$(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(j)} = (\mathbf{r}^{(i)})^T\mathbf{r}^{(j)} - \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{r}^{(j)} = 0$$

Taken together, this shows that

$$(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(j+1)} = 0 \tag{B.69}$$

[9] In fact, there are several possible values for S such that $S^TS = A$.

2. Using (B.54) and (B.59)

$$(\mathbf{r}^{(i+1)})^T\mathbf{d}^{(i)} = (\mathbf{r}^{(i)})^T\mathbf{d}^{(i)} - \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)} = (\mathbf{r}^{(i)})^T\mathbf{d}^{(i)} - (\mathbf{r}^{(i)})^T\mathbf{r}^{(i)} = 0$$

Whereas using (B.54), (B.58), and (B.60)

$$(\mathbf{r}^{(i+1)})^T\mathbf{d}^{(j)} = (\mathbf{r}^{(i)})^T\mathbf{d}^{(j)} - \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{d}^{(j)} = 0$$

Taken together, this shows that

$$(\mathbf{r}^{(i+1)})^T\mathbf{d}^{(j+1)} = 0 \tag{B.70}$$

3. Using (B.56) and (B.70),

$$\begin{aligned}(\mathbf{r}^{(i+1)})^T\mathbf{d}^{(i+1)} &= (\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)} - \frac{(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{r}^{(i+1)})^T\mathbf{d}^{(i)} \\ &= (\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)}\end{aligned} \tag{B.71}$$

4. Using (B.56),

$$\begin{aligned}(\mathbf{d}^{(i+1)})^TA\mathbf{d}^{(i)} &= (\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)} - \frac{(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)} \\ &= (\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)} - (\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)} = 0\end{aligned}$$

Whereas using (B.56), (B.60), and (B.65)

$$(\mathbf{d}^{(i+1)})^TA\mathbf{d}^{(j)} = (\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(j)} - \frac{(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{d}^{(j)} = 0$$

Taken together, this shows that

$$(\mathbf{d}^{(i+1)})^TA\mathbf{d}^{(j+1)} = 0 \tag{B.72}$$

5. Using (B.56) and (B.72)

$$\begin{aligned}(\mathbf{d}^{(i+1)})^TA\mathbf{d}^{(i+1)} &= (\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i+1)} - \frac{(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i+1)} \\ &= (\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i+1)}\end{aligned} \tag{B.73}$$

6. Using (B.54) and (B.57),

$$\begin{aligned}(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)} &= (\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i)} - \frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)} \\ &= -\frac{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}}(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)}\end{aligned}$$

or rearranging

$$\frac{(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)}}{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}} = -\frac{(\mathbf{r}^{(i+1)})^TA\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^TA\mathbf{d}^{(i)}} \tag{B.74}$$

7. Using (B.56) and (B.65),

$$\begin{aligned}(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)} &= (\mathbf{r}^{(i+1)})^T A\mathbf{r}^{(i)} - \frac{(\mathbf{r}^{(i)})^T A\mathbf{d}^{(i-1)}}{(\mathbf{d}^{(i-1)})^T A\mathbf{d}^{(i-1)}}(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i-1)} \\ &= (\mathbf{r}^{(i+1)})^T A\mathbf{r}^{(i)} \end{aligned} \tag{B.75}$$

8. Using (B.56), (B.61), and (B.75),

$$\begin{aligned}(\mathbf{d}^{(i+1)})^T A\mathbf{r}^{(i)} &= (\mathbf{r}^{(i+1)})^T A\mathbf{r}^{(i)} - \frac{(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^T A\mathbf{r}^{(i)} \\ &= (\mathbf{r}^{(i+1)})^T A\mathbf{r}^{(i)} - (\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)} = 0\end{aligned}$$

Whereas using (B.56), (B.64), and (B.66),

$$(\mathbf{d}^{(i+1)})^T A\mathbf{r}^{(j)} = (\mathbf{r}^{(i+1)})^T A\mathbf{r}^{(j)} - \frac{(\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)}}{(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}}(\mathbf{d}^{(i)})^T A\mathbf{r}^{(j)} = 0$$

Taken together, this shows that

$$(\mathbf{d}^{(i+1)})^T A\mathbf{r}^{(j+1)} = 0 \tag{B.76}$$

9. Using (B.54) for $\mathbf{r}_{i+2}^T$, (B.61), (B.62), and (B.76),

$$\begin{aligned}(\mathbf{r}^{i+2})^T A\mathbf{d}^{(i)} &= (\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(i)} - \alpha^{(i+1)}(\mathbf{d}^{(i+1)})^T A^2\mathbf{d}^{(i)} \\ &= \left(-\frac{(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)}}{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}\right) \\ &\quad + \frac{\alpha^{(i+1)}}{\alpha^{(1)}}(\mathbf{d}^{(i+1)})^T A\left(\mathbf{r}^{(i+1)} - \mathbf{r}^{(i)}\right) \\ &= \left(-\frac{(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)}}{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}\right) \\ &\quad + \left(\frac{(\mathbf{r}^{(i+1)})^T\mathbf{r}^{(i+1)}}{(\mathbf{r}^{(i)})^T\mathbf{r}^{(i)}}(\mathbf{d}^{(i)})^T A\mathbf{d}^{(i)}\right) = 0\end{aligned}$$

Whereas using (B.54) for $\mathbf{r}_{i+2}^T$, multiplying by $A\mathbf{d}^{(j)}$, and then using (B.65) and (B.76),

$$\begin{aligned}(\mathbf{r}^{(i+2)})^T A\mathbf{d}^{(j)} &= (\mathbf{r}^{(i+1)})^T A\mathbf{d}^{(j)} - \alpha^{(i+1)}(\mathbf{d}^{(i+1)})^T A^2\mathbf{d}^{(j)} \\ &= -\alpha^{(i+1)}(\mathbf{d}^{(i+1)})^T A\left(\frac{1}{\alpha^{(j)}}\left(\mathbf{r}^{(j+1)} - \mathbf{r}^{(j)}\right)\right) = 0\end{aligned}$$

Taken together, this shows that

$$(\mathbf{r}^{(i+2)})^T A\mathbf{d}^{(j+1)} = 0 \tag{B.77}$$

10. Using (B.56) for $\mathbf{d}^{(j+1)}$, multiplying by $(\mathbf{r}^{(i+2)})^T$, and then using (B.77)

$$(\mathbf{r}^{(i+2)})^T A\mathbf{r}^{(j+1)} = (\mathbf{r}^{(i+2)})^T A\left(\mathbf{d}^{(j+1)} - \beta^{(j+1)}\mathbf{d}^{(j)}\right) = 0 \tag{B.78}$$

Thus (B.69) through (B.78) show that if (B.57) to (B.66) apply to i with $j < i$, then the same equations should also apply to $i + 1$. The lemma then follows by induction from $i = 1$.

B.11.3 Proof That $\mathbf{err}^{(N)} = 0$ Using CG Method (Theorem B.1)

If the initial guess $\mathbf{x}^{(0)}$ was chosen fortuitously such that $\mathbf{d}^{(\ell)} = 0$ for $\ell < (N-1)$, then the conjugate gradient algorithm would have terminated at less than N iterations with $\mathbf{r}^{(\ell+1)} = 0$.

More generally, with an arbitrary initial guess $\mathbf{x}^{(0)}$, we expect to have the set $\mathcal{D} = \left(\mathbf{d}^{(0)}, \ldots, \mathbf{d}^{(N-1)}\right)$ to be a linearly independent set of vectors that, according to (B.60), is an A-orthogonal set, that is,

$$(\mathbf{d}^{(i)})^T A \mathbf{d}^{(j)} \qquad j < i$$

We can then represent $\mathbf{err}^{(0)}$ using $\mathcal{D}$ as the basis set,

$$\mathbf{err}^{(0)} = \sum_{k=0}^{N-1} \mu_k \mathbf{d}^{(k)} \tag{B.79}$$

To identify the coefficients μ_k, multiply (B.79) by $(\mathbf{d}^{(\ell)})^T A$ while using the A-orthogonal properties of $\mathbf{d}^{(\ell)}$,

$$\begin{aligned} (\mathbf{d}^{(\ell)})^T A\, \mathbf{err}^{(0)} &= \mu_\ell (\mathbf{d}^{(\ell)})^T A \mathbf{d}^{(\ell)} \\ \mu_\ell &= \frac{(\mathbf{d}^{(\ell)})^T A\, \mathbf{err}^{(0)}}{(\mathbf{d}^{(\ell)})^T A \mathbf{d}^{(\ell)}} = -\frac{\mathbf{d}^{(\ell)} \mathbf{r}^{(0)}}{(\mathbf{d}^{(\ell)})^T A \mathbf{d}^{(\ell)}} \end{aligned} \tag{B.80}$$

From (B.53), we have

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{x}^{(0)} + \alpha_0 \mathbf{d}^{(0)} \\ \mathbf{x}_2 &= \mathbf{x}_1 + \alpha_1 \mathbf{d}_1 = \mathbf{x}^{(0)} + \alpha_0 \mathbf{d}^{(0)} + \alpha_1 \mathbf{d}^{(1)} \\ &\vdots \\ \mathbf{x}^{(i)} &= \mathbf{x}^{(0)} + \sum_{m=0}^{i-1} \alpha_m \mathbf{d}^{(m)} \end{aligned}$$

which, when we subtract $\mathbf{x}^*$ on both sides, will yield

$$\mathbf{err}^{(i)} = \mathbf{err}^{(0)} + \sum_{m=0}^{i-1} \alpha_m \mathbf{d}^{(m)} \tag{B.81}$$

or after multiplying both sides by $-A$,

$$\mathbf{r}^{(i)} = \mathbf{r}^{(0)} - \sum_{m=0}^{i-1} \alpha_m A \mathbf{d}^{(m)} \tag{B.82}$$

Premultiplying (B.82) (with $i = \ell$) by $(\mathbf{d}^{(\ell)})^T$, we have

$$(\mathbf{d}^{(\ell)})^T \mathbf{r}^{(\ell)} = (\mathbf{d}^{(\ell)})^T \mathbf{r}^{(0)}$$

which, after applying (B.59), yields

$$(\mathbf{d}^{(\ell)})^T \mathbf{r}^{(0)} = (\mathbf{r}^{(\ell)})^T \mathbf{r}^{(\ell)} \tag{B.83}$$

Applying (B.83) to (B.80) and recalling (B.52), we find that

$$\mu_\ell = -\frac{(\mathbf{r}^{(\ell)})^T \mathbf{r}^{(\ell)}}{(\mathbf{d}^{(\ell)})^T A \mathbf{d}^{(\ell)}} = -\alpha_\ell$$

or going back to (B.79),

$$\mathbf{err}^{(0)} = -\sum_{k=0}^{N-1} \alpha_k \mathbf{d}^{(k)} \tag{B.84}$$

Now take (B.81) and substitute (B.84),

$$\mathbf{err}^{(i)} = \left(-\sum_{k=0}^{N-1} \alpha_k \mathbf{d}^{(k)} \right) + \left(\sum_{m=0}^{i-1} \alpha_m \mathbf{d}^{(m)} \right)$$

Thus when $i = N$,

$$\mathbf{err}^{(N)} = \left(-\sum_{k=0}^{N-1} \alpha_k \mathbf{d}^{(k)} \right) + \left(\sum_{m=0}^{N-1} \alpha_m \mathbf{d}^{(m)} \right) = 0$$

B.12 GMRES Algorithm

B.12.1 Basic Algorithm

To simplify our discussion, we restrict the method to apply only to nonsingular A. Let $\mathbf{x}^{(k)}$ be the k^{th} update for the solution of $A\mathbf{x} = \mathbf{b}$, with $\mathbf{x}^{(0)}$ being the initial guess, and let $\mathbf{r}^{(k)} = \mathbf{b} - A\mathbf{x}^{(k)}$ be the k^{th} residual error based on these updates. Beginning with a normalized vector $\mathbf{u}_1$,

$$\mathbf{u}_1 = \frac{\mathbf{r}^{(0)}}{\left\| \mathbf{r}^{(0)} \right\|} \tag{B.85}$$

a matrix $U_k[=]N \times k$ can be constructed using an orthonormal sequence $\{\mathbf{u}_1, \mathbf{u}_2, \ldots\}$ as

$$U_k = \left(\mathbf{u}_1 \mid \mathbf{u}_2 \mid \cdots \mid \mathbf{u}_k \right) \tag{B.86}$$

that is, $U_k^* U_k = I$, where $\mathbf{u}_k$ are obtained sequentially using **Arnoldi's method** given by

$$\mathbf{p}_k = (I - U_k U_k^*) A\mathbf{u}_k \qquad ; \qquad \mathbf{u}_{k+1} = \frac{\mathbf{p}_k}{\left\| \mathbf{p}_k \right\|} \tag{B.87}$$

One can show that Arnoldi's method will yield the following property of U_k and U_{k+1}:

$$U_{k+1}^* A U_k = \widehat{H}_k \tag{B.88}$$

where $\widehat{H}_k[=](k+1) \times k$ has the form of a truncated Hessenberg matrix, that is, a Hessenberg matrix with the last column removed,

$$\widehat{H}_k = \begin{pmatrix} \times & \times & \cdots & \times \\ \times & \times & \cdots & \times \\ & \times & \cdots & \times \\ & & \ddots & \vdots \\ 0 & & & \times \end{pmatrix}$$

Suppose the length of $\mathbf{x}$ is N. Using the matrices U_k generated by Arnoldi's method, GMRES is able to transform the problem of minimizing the residual $\mathbf{r}^{(k)}$ to an associated least-squares problem

$$\left(U_{k+1}^* A U_k\right)\mathbf{y}_k = \widehat{H}_k \mathbf{y}_k =_{\text{lsq}} \begin{pmatrix} \left\|\mathbf{r}^{(0)}\right\| \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{B.89}$$

where $\mathbf{y}_k$ has a length k, which is presumably much smaller than N. Because of the special Hessenberg structure of $\widehat{H}_k$, efficient approaches for the least-squares solution of (B.89) is also available. Thus, with $\mathbf{y}_k$, the k^{th} solution update is given by

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + U_k \mathbf{y}_k \tag{B.90}$$

To show that (B.89) together with update (B.90) is equivalent to minimizing the k^{th} residual, we simply apply the properties of U_k obtained using Arnoldi's method as follows:

$$\begin{aligned} \min \left\| \mathbf{r}^{(k)} \right\| &= \min \left\| \mathbf{b} - A\left(\mathbf{x}^{(0)} + U_k \mathbf{y}\right) \right\| \\ &= \min \left\| \mathbf{r}^{(0)} - A U_k \mathbf{y} \right\| \\ &= \min \left\| (\|\mathbf{r}^{(0)}\|)\mathbf{u}_1 - U_{k+1}\left(U_{k+1}^* A U_k\right)\mathbf{y} \right\| \\ &= \min \left\| U_{k+1}\left(\mathbf{c}_k - \widehat{H}_k \mathbf{y}\right) \right\| \end{aligned} \tag{B.91}$$

where $\mathbf{c}_k = \left(\|\mathbf{r}^{(0)}\|, 0, \ldots, 0\right)^T$ has a length $(k+1)$. Because $U_{k+1}^* U_{k+1} = I$, (B.91) reduces to (B.89).

If the norm of $\mathbf{r}^{(k)}$ is within acceptable bounds, the process can be terminated. Indeed, GMRES often reaches acceptable solutions of large systems for k much less than the size N of matrix A. The vectors $\mathbf{u}_k$ obtained using Arnoldi's method introduce updates that are similar to the directions used by conjugate gradient method.

Further areas of improvement to the basic GMRES approach just described are usually implemented. These include:

1. Restarting the GMRES method every $m \ll N$ steps to reduce the storage requirements.
2. Taking advantage of the structure of $\widehat{H}_K$ to solve the least-squares problem.
3. Incorporating the evaluation of $\left\|\mathbf{r}^{(k)}\right\|$ inside the iteration loops of the Arnoldi method.

A practical limitation of GMRES is that the size of U_k keeps getting larger as k increases, and U_k is generally not sparse. However, if k is small, the k^{th} residuals may not be sufficiently small at that point. One solution is then to "restart" the GMRES method using the last update after m steps as the new initial guess for another batch of m iterations of GMRES. These computation batches are performed until the desired tolerance is achieved. As expected, small values of m would lead to a slower convergence, whereas a large value of m would mean a larger storage requirement.

Details that address the other two other improvements, that is, special least-squares solution of Hessenberg matrices and the enhanced Arnoldi steps, are

included in Section B.12.2, where the GMRES algorithm is also outlined in that section with the improvements already incorporated.

Note that for both the conjugate gradient method and GMRES method, we have

$$\mathbf{r}^{(k)} = \sum_{i=0}^{k-1} c_k A^k \mathbf{r}^{(0)} \tag{B.92}$$

where c_i are constant coefficients. For the conjugate gradient method, this results directly from (B.45).

For the GMRES method, we have from Arnoldi's method,

$$\mathbf{u}_{j+1} = \frac{1}{\alpha_j}\left(A\mathbf{u}_j - \begin{pmatrix} \mathbf{u}_1 & | & \cdots & | & \mathbf{u}_j \end{pmatrix} \begin{pmatrix} \mathbf{u}_1^* A u_j \\ \vdots \\ \mathbf{u}_j^* A u_j \end{pmatrix}\right) = b_j A\mathbf{u}_j + \left(\sum_{i=1}^{j} a_i \mathbf{u}_i\right)$$

for some coefficients $a_i, i = 1, \ldots, j$ and b_j. When applied to $j = 2, 3, \ldots, k$, together with $\mathbf{u}_1 = \mathbf{r}^{(0)} / \left\|\mathbf{r}^{(0)}\right\|$, we can recursively reduce the last relationship to

$$\mathbf{u}_k = \sum_{i=1}^{k-1} q_i A^{i-1} \mathbf{r}^{(0)}$$

which when applied to the k^{th} update, $\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + U_k \mathbf{y}$,

$$A\left(\mathbf{x}^{(k)} - \mathbf{x}^{(0)}\right) = \mathbf{r}^{(0)} - \mathbf{r}^{(k)} = \sum_{i=1}^{k-1} y_i q_i A^i \mathbf{r}^{(0)} \quad \rightarrow \quad \mathbf{r}^{(k)} = \sum_{i=0}^{k-1} c_i A^i \mathbf{r}^{(0)}$$

When seen in this space, the critical difference between the conjugate gradient and GMRES methods lies in how each method determines the coefficients c_i for the linear combination of $A^i \mathbf{r}^{(0)}$. Otherwise, they both contain updates that resides in a subspace known as the **Krylov subspace**.

Definition B.4. *A k^{th}-order* **Krylov subspace** *based on square matrix $A[=]N \times N$ and vector $\mathbf{v}[=]N \times 1$ is the subspace spanned by vectors $A^i\mathbf{v}$, with $i = 1, \ldots, N$, $k \leq N$, that is,*

$$\mathcal{K}_k(A, \mathbf{v}) = \textbf{Span}\left(\mathbf{v}, A\mathbf{v}, \cdots, A^{k-1}\mathbf{v}\right)$$

There are several other methods that fall under the class known as **Krylov subspace methods** including Lanczos, QMR, and BiCG methods. By restricting the updates to fall within the Krylov subspace, the immediate advantage is that the components of Krylov subspace involve repeated matrix-vector products of the form $A\mathbf{v}$. When A is dense, Krylov methods may just be comparable to other methods, direct or iterative. However, when A is large and sparse, the computations of $A\mathbf{v}$ can be significantly reduced by focusing only on the nonzero components of A.

More importantly, for nonsingular A, it can be shown that the solution of $A\mathbf{x} = \mathbf{b}$ lies in the Krylov subspace, $\mathcal{K}_k(A, \mathbf{b})$ for some $k \leq N$. Thus as we had noted earlier, both the conjugate gradient method and the GMRES method are guaranteed to reach the exact solution in at most N iterations, assuming no round-off errors. In some cases, the specified tolerance of the error vectors may even be reached at k iterations that are much fewer than the maximal N iterations. However, the operative word

here is still "nonsingular." Thus it is possible that the convergence will still be slow if A is nearly singular or ill-conditioned. Several methods are available that choose matrix multipliers C, called **preconditioners**, such that the new matrix $\widehat{A} = CA$ has an improved condition number, but this has to be done without losing much of the advantages of sparse matrices.

B.12.2 Enhancements

We address the two improvements of the basic GMRES. One improvement centers on taking advantage of the truncated Hesseberg structure of $\widehat{H}_k$ during the least-squares solution. The other improvement is to enhance the Arnoldi method for calculating $\mathbf{u}_k$ by incorporating the values of the residuals $\mathbf{r}^{(k)}$.

We begin with an explicit algorithm for Arnoldi's method:

Arnoldi Method:

Given: $A[=]n \times n$, $1 < k \leq n$, and $\mathbf{p}_1[=]n \times 1$.
Initialize:

$$\mathbf{u}_1 = \frac{\mathbf{p}_1}{\|\mathbf{p}_1\|} \qquad \text{and} \qquad U_1 = \left(\begin{array}{c} \mathbf{u}_1 \end{array} \right)$$

Iterate: Loop for $i = 1, \ldots, k$,

$$\mathbf{w}_i = A\mathbf{u}_i \qquad ; \qquad \mathbf{h}_i = U_i^* \mathbf{w}_i$$
$$\mathbf{p}_i = \mathbf{w}_i - U_i \mathbf{h}_i \qquad ; \qquad \alpha_i = \|\mathbf{p}_i\|$$

If $\alpha_i > 0$ or $i < n$

$$\mathbf{u}_{i+1} = \frac{\mathbf{p}_i}{\alpha_i} \qquad \text{and} \qquad U_{i+1} = \left(\begin{array}{c|c} U_i & \mathbf{u}_{i+1} \end{array} \right)$$

Else
Exit and report the value of i as the maximum number of orthogonal vectors found.
End If
End Loop

At the termination of the Arnoldi algorithm, we can generate matrices $\widehat{H}_i = U_{i+1}^* A U_i$ or $H_i = U_i^* A U_i$ depending on whether $\alpha_i > 0$ or not. Alternatively, we could set the nonzero elements of H_k and $\widehat{H}_k$ directly at each iteration of the method by using $\mathbf{h}_j$ and α_j as follows:

$$H_k(i,j) = \begin{cases} \mathbf{h}_j(i) & \text{for } k \geq j \geq i, \\ \alpha_j & \text{for } k \geq j = i-1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \widehat{H}_k = \left(\begin{array}{c|c} H_k & \\ \hline \mathbf{0}_{1\times(k-1)} & \alpha_k \end{array} \right) \tag{B.93}$$

Using the QR decomposition of $H_k = Q_k R_k$, where Q_k is unitary and R_k is upper triangular, we can form an orthogonal matrix

$$\widehat{Q}_{k+1} = \left(\begin{array}{c|c} Q_k & \mathbf{0}_{k\times 1} \\ \hline \mathbf{0}_{1\times k} & 1 \end{array} \right)$$

such that

$$\widehat{Q}_{k+1}^{*}\widehat{H}_{k} = \left(\begin{array}{c|c} R_k & \\ \hline \mathbf{0}_{1\times k} & \alpha_k \end{array}\right)$$

If α_k is nonzero, we can use another orthogonal matrix $G_{k+1}[=](k+1)\times(k+1)$ given by

$$G_{k+1} = \left(\begin{array}{c|c} I_{[k-1]} & \mathbf{0}_{((k-1)\times 2)} \\ \hline \mathbf{0}_{(2\times(k-1))} & \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \end{array}\right)$$

where $c = R_k(k,k)/\rho_k$, $s = \alpha_k/\rho_k$ and $\rho_k = \sqrt{R_k(k,k)^2 + \alpha_k^2}$. Then

$$G_{k+1}\widehat{Q}_{k+1}^{*}\widehat{H}_{k} = \left(\begin{array}{c} \widehat{R}_k \\ \hline \mathbf{0}_{1\times(k+1)} \end{array}\right)$$

where $\widehat{R}_k[=]k\times k$ is an upper triangular matrix that is equal to R_k except for the lower corner element, $\widehat{R}_k(k,k) = \rho_k$.

Let $\Omega_{k+1} = G_{k+1}\widehat{Q}_{k+1}^{*}$ be the combined orthognal matrix. Premultiplying both sides of (B.89) by Ω_{k+1}, the least-squares problem reduces to

$$\widehat{R}_k\mathbf{y} = \left\|\mathbf{r}^{(0)}\right\| \begin{pmatrix} \Omega_{k+1}(1,1) \\ \vdots \\ \Omega_{k+1}(k,1) \end{pmatrix} \tag{B.94}$$

Because $\widehat{R}_k$ is upper triangular, the value of $\mathbf{y}$ can be found using the back-substitution process.

Recursion formulas for Ω_ℓ and R_k are given by

$$\Omega_{\ell+1} = G_{\ell+1}\left(\begin{array}{c|c} \Omega_\ell & \mathbf{0}_{(\ell)\times 1} \\ \hline \mathbf{0}_{1\times(\ell)} & 1 \end{array}\right) \tag{B.95}$$

$$R_\ell = \left(\begin{array}{c|c} \widehat{R}_{\ell-1} & \Omega_\ell\mathbf{h}_\ell \end{array}\right) \tag{B.96}$$

Using $\Omega_1 = [1]$ and R_0 as a null matrix, the recursions (B.95) and (B.96) can be incorporated inside the Arnoldi iterations without having to explicitly solve for Q_k.

Furthermore, when the equality in (B.94) is satisfied, the norm of the k^{th} residual is given by

$$\begin{aligned} \left\|\mathbf{r}^{(k)}\right\| &= \left\|\Omega_{k+1}\left(\mathbf{c}_k - \widehat{H}_k y\right)\right\| \\ &= \left\| \left\|\mathbf{r}^{(0)}\right\| \begin{pmatrix} \Omega_{k+1}(1,1) \\ \vdots \\ \Omega_{k+1}(k+1,1) \end{pmatrix} - \begin{pmatrix} \widehat{R}_k\mathbf{y} \\ \hline 0 \end{pmatrix} \right\| \\ &= \left\|\mathbf{r}^{(0)}\right\| \left| \Omega_{k+1}(k+1,1) \right| \end{aligned} \tag{B.97}$$

This means that the norm of the k^{th} residual can be incorporated inside the iterations of Arnoldi's method, without having to explicitly solve for $\mathbf{x}^{(k)}$.

When $\Omega_{k+1}(k+1,1) = 0$, (B.97) implies that $\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + U_k\mathbf{y}$ is an exact solution. Note that the Arnoldi method will stall at the i^{th} iteration if $\alpha_i = 0$ because

$\mathbf{u}_{i+1}$ requires a division by α_i. However, this does not prevent the formation of $\widehat{H}_i$ as given in (B.93). This implies that $(\widehat{Q}^*_{k+1}\widehat{H}_k)$ is already in the form required by $\widehat{R}_k$ and that $\Omega_{k+1} = \widehat{Q}^*_{k+1}$, or $\Omega_{k+1}(k+1,1) = 0$. Thus when the Arnoldi process in GMRES stalls at a value i, the update $\mathbf{x}_i$ at that point is already an exact solution to $A\mathbf{x} = \mathbf{b}$. This is also the situation when $i = n$, assuming no roundoff errors.

In summary, we have the GMRES algorithm given below:

GMRES Algorithm:

Given: $A[=]n \times n$, $\mathbf{b}[=]n \times 1$, initial guess $\mathbf{x}^{(0)}[=]n \times 1$ and tolerance *tol*.
Initialize:

$$\mathbf{r}^{(0)} = \mathbf{b} - A * \mathbf{x}^{(0)} \quad ; \quad \beta = \left\|\mathbf{r}^{(0)}\right\| \quad ; \quad \mathbf{u} = \mathbf{r}^{(0)}/\beta$$

$$U = \begin{pmatrix} \mathbf{u} \end{pmatrix} \quad ; \quad Q = \begin{pmatrix} 1 \end{pmatrix} \quad ; \quad R = [\,]$$

$$\gamma = \beta \quad ; \quad i = 0 \quad ; \quad \alpha = \beta$$

Iterate: $i \leftarrow i + 1$

While $\gamma > tol$ and $\alpha > tol$

$$\mathbf{w} = A\mathbf{u}; \quad \mathbf{h} = U^*\mathbf{w}; \quad \mathbf{p} = \mathbf{w} - U\mathbf{h}; \quad \mathbf{r} = Q\mathbf{h}; \quad \alpha = \|\mathbf{p}\|$$

if $\alpha > tol$

$$U \leftarrow \left(\begin{array}{c|c} U & \frac{\mathbf{p}}{\|\mathbf{p}\|} \end{array}\right)$$

$$\rho = \sqrt{\mathbf{r}(i)^2 + \alpha^2} \quad ; \quad c = \frac{\mathbf{r}(i)}{\rho}; \quad s = \frac{\alpha}{\rho}$$

$$r_i \leftarrow \rho; \quad R \leftarrow \left(\begin{array}{c|c} R & \mathbf{r} \\ \hline \mathbf{0} & \end{array}\right)$$

$$Q \leftarrow \left(\begin{array}{c|c} I_{[i-1]} & 0 \\ \hline 0 & \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \end{array}\right) \left(\begin{array}{c|c} Q & 0 \\ \hline 0 & 1 \end{array}\right)$$

$$\gamma \leftarrow Q_{i+1,1}$$

end if
End While Loop

Solve for $\mathbf{y}$ using back-substitution:

$$R\mathbf{y} = \beta \begin{pmatrix} Q_{1,1} \\ \vdots \\ Q_{i,1} \end{pmatrix}$$

Evaluate the final solution:

$$\mathbf{x} = \mathbf{x}^{(0)} + U\mathbf{y}$$

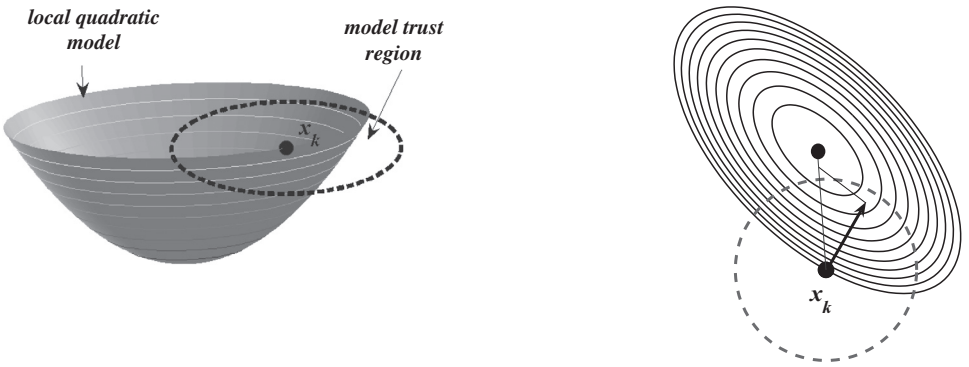


Figure B.10. The trust region and the local quadratic model based on $\mathbf{x}^{(k)}$. The right figure shows the contour plot and the double-dogleg step.

B.13 Enhanced-Newton Using Double-Dogleg Method

Like the line search approach, the double-dogleg method is used only when a full Newton update is not acceptable. The method will use a combination of two types of updates:

1. **Gradient Descent Update**

$$\delta_k^G = -J_k^T \mathbf{F}\left(\mathbf{x}^{(k)}\right) \tag{B.98}$$

2. **Newton Update**

$$\delta_k^N = -J_k^{-1} \mathbf{F}\left(\mathbf{x}^{(k)}\right) \tag{B.99}$$

Because the Newton update was based on a local model derived from a truncated Taylor's series, we could limit the update step to be inside a sphere centered around $\mathbf{x}^{(k)}$ known as the **model-trust region** approach, that is, with $M_k > 0$

$$\|\Delta_k \mathbf{x}\| \leq M_k \tag{B.100}$$

Assuming the Newton step is the optimum local step, the local problem is that of minimizing a scalar function φ_k given by

$$\begin{aligned} \varphi_k(\Delta\mathbf{x}) &= \frac{1}{2}(\mathbf{F}_k + J_k\Delta\mathbf{x})^T(\mathbf{F}_k + J_k\Delta\mathbf{x}) \\ &= \frac{1}{2}\mathbf{F}_k^T\mathbf{F}_k + \left(\mathbf{F}_k^T J_k\right)\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T\left(J_k^T J_k\right)\Delta\mathbf{x} \end{aligned} \tag{B.101}$$

Note that the minimum of $\varphi(\Delta\mathbf{x})$ occurs at $\Delta\mathbf{x} = -J_k^{-1}\mathbf{F}_k$, the Newton step.

The local model is shown in Figure B.10 as a concave surface attached to the point $\mathbf{x}^{(k)}$, whereas the trust region is the circle centered around $\mathbf{x}^{(k)}$.

The double-dogleg procedure starts with the direction along the gradient, that is, a path determined by $\Delta\mathbf{x} = \sigma\delta_k^G$. This will trace a parabola along the surface of the quadratic model as σ increases from 0:

$$\mathcal{P}_G(\sigma) = \varphi\left(\sigma\delta_k^G\right) = \frac{1}{2}\mathbf{F}_k^T\mathbf{F}_k - \sigma\left(\mathbf{F}_k^T J_k J_k^T \mathbf{F}_k\right) + \frac{\sigma^2}{2}\left(\mathbf{F}_k^T\left(J_k J_k^T\right)^2\mathbf{F}_k\right)$$

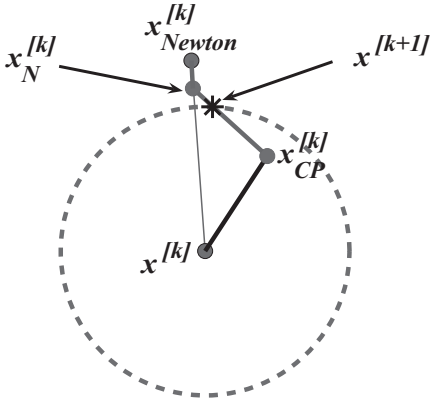


Figure B.11. The double-dogleg method for obtaining the update $\mathbf{x}_{k+1}$.

The minimum of this parabola occurs at

$$\sigma^* = \frac{\mathbf{F}_k^T J_k J_k^T \mathbf{F}_k}{\mathbf{F}_k^T \left(J_k J_k^T\right)^2 \mathbf{F}_k}$$

This yields the point known as the **Cauchy point**,

$$\mathbf{x}_{\text{CP}}^{(k)} = \mathbf{x}^{(k)} + \sigma^* \delta_k^G \tag{B.102}$$

Note that if $\mathbf{x}_{\text{CP}}^{(k)}$ is outside the trust region, $\mathbf{x}_{\text{CP}}^{(k)}$ will need to be set as the intersection of the line along the gradient descent with the boundary of the trust region. In Figure B.10, the contour plot is shown with an arrow originating from $\mathbf{x}^{(k)}$ but terminates at the Cauchy point.

The full Newton step will take $\mathbf{x}^{(k)}$ to the point denoted by $\mathbf{x}_{\text{Newton}}^{(k)}$, which is the minimum point located at the center of the elliptical contours. The Cauchy point, full-Newton update point, and other relevant points, together with the important line segments, are blown up and shown in Figure B.11.

One approach is to draw a line segment from $\mathbf{x}_{\text{Newton}}^{(k)}$ to the Cauchy point $\mathbf{x}_{\text{CP}}^{(k)}$. Then the next update can be set as the intersection of this line segment with the boundary of the trust region. This approach is known as the **Powell update**, or the **single-dogleg step**. However, it has been found that convergence can be further improved by taking another point along the Newton step direction, which we denote by $\mathbf{x}_{\mathcal{N}}^{(k)}$. The **Dennis-Mei approach** suggests that $\mathbf{x}_{\mathcal{N}}^{(k)}$ is evaluated as follows:

$$\mathbf{x}_{\mathcal{N}}^{(k)} = \mathbf{x}^{(k)} + \eta \delta_k^N = \mathbf{x}^{(k)} - \eta J_k^{-1} \mathbf{F}\left(\mathbf{x}^{(k)}\right) \tag{B.103}$$

where

$$\eta = 0.2 + 0.8\,\sigma^* \left[\frac{\mathbf{F}_k^T \left(J_k J_k^T\right) \mathbf{F}_k}{\mathbf{F}_k^T \mathbf{F}_k} \right]$$

The double-dogleg update can then be obtained by finding the intersection between the boundary of the trust region and the line segment from $\mathbf{x}_N^{(k)}$ to $\mathbf{x}_{\text{CP}}^{(k)}$ as shown in Figure B.11, that is,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + (1-\rho)\,\mathbf{x}_{\text{CP}}^{(k)} + \rho \mathbf{x}_{\mathcal{N}}^{(k)} \tag{B.104}$$

where

$$\begin{aligned}\rho &= \frac{-b+\sqrt{b^2-ac}}{a}\\ a &= \left\|\mathbf{x}_{\mathcal{N}}^{(k)}-\mathbf{x}_{\mathrm{CP}}^{(k)}\right\|^2\\ b &= \left(\mathbf{x}_{\mathcal{N}}^{(k)}-\mathbf{x}_{\mathrm{CP}}^{(k)}\right)^T\mathbf{x}_{\mathrm{CP}}^{(k)}\\ c &= \left\|\mathbf{x}_{\mathrm{CP}}^{(k)}\right\|^2-M_k^2\end{aligned}$$

and M_k is the radius of the trust region. In case the update does not produce satisfactory results, then the radius will need to be reduced using an approach similar to the line search method.

To summarize, we have the following enhanced Newton with double-dogleg procedure:

Algorithm of the Enhanced Newton's Method with Double-Dogleg Search.

1. **Initialize.** Choose an initial guess: $\mathbf{x}^{(0)}$
2. **Update.** Repeat the following steps until either $\left\|\mathbf{F}\left(\mathbf{x}^{(k)}\right)\right\| \le \epsilon$ or the number of iterations have been exceeded
   (a) Calculate J_k.
   (If J_k is singular, then stop the method and declare "Singular Jacobian.")
   (b) Calculate the δ_k^G and δ_k^N. (cf. (B.98) and (B.99), respectively).
   (c) Evaluate points $\mathbf{x}_{\mathrm{CP}}^{(k)}$ and $\mathbf{x}_{\mathcal{N}}^{(k)}$: (cf. (B.102) and (B.103))
   (d) Evaluate the step change $\Delta_k\mathbf{x}$:

$$\Delta_k\mathbf{x} = (1-\rho)\,\mathbf{x}_{\mathrm{CP}}^{(k)} + \rho\mathbf{x}_{\mathcal{N}}^{(k)}$$

   where ρ is obtained by (B.104).
   (e) Check if $\Delta_k\mathbf{x}$ is acceptable. If

$$\left\|\mathbf{F}\left(\mathbf{x}^{(k)}+\Delta_k\mathbf{x}\right)\right\|^2 > \left\|\mathbf{F}\left(\mathbf{x}^{(k)}\right)\right\|^2 + 2\alpha\mathbf{F}_k^T J_k\Delta_k\mathbf{x}$$

   with $\alpha \in (0, 0.5)$ (typically $\alpha = 10^{-4}$), then update is unacceptable. Modify the trust region:

$$M_k \leftarrow \max\Big(\,0.1M_k,\ \ \min\left(\,0.5M_k, \lambda\,\|\Delta_k\mathbf{x}\|\,\right)\Big)$$

   where

$$\lambda = -\frac{\mathbf{F}_k^T J_k\Delta_k\mathbf{x}}{\left(\,\left\|\mathbf{F}\left(\mathbf{x}^{(k)}+\Delta_k\mathbf{x}\right)\right\|^2 - \left\|\mathbf{F}\left(\mathbf{x}^{(k)}\right)\right\|^2 - 2\mathbf{F}_k^T J_k\Delta_k\mathbf{x}\right)}$$

   and repeat from step 2c above.
   Otherwise, if acceptable, continue to next step.

   (f) Update $\mathbf{x}^{(k)}$: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta_k\mathbf{x}$

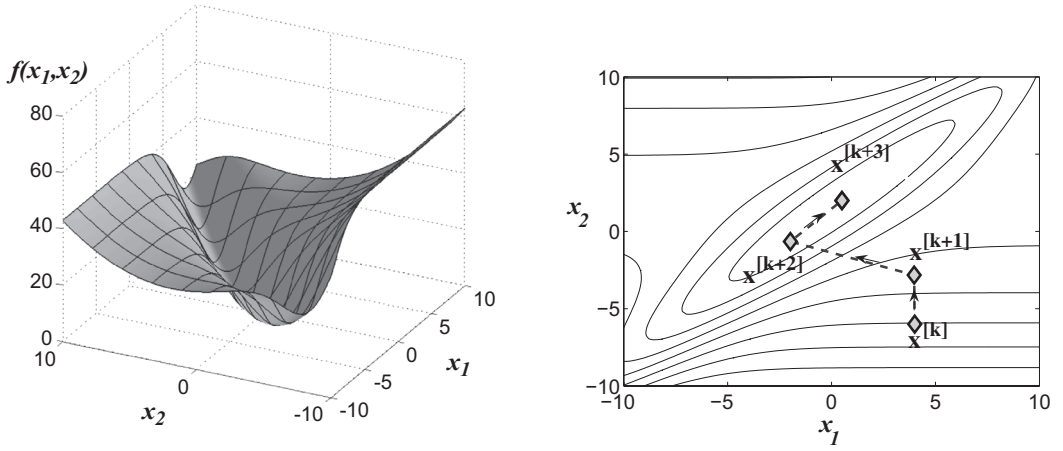


Figure B.12. A surface plot of $f(x_1, x_2)$ of (B.105) is shown in the left figure. The right figure shows the contour plot and the performance of the enhanced Newton with double-dogleg method with the initial guess at $(x_1, x_2) = (4, -6)$.

Remarks: A MATLAB code `nsolve.m` is available on the book's webpage that implements the enhanced Newton method, where the line-search method is implemented when the parameter "`type`" is set to 2. Also, another MATLAB code that uses the enhanced Newton method for minimization of a scalar function is available on the book's webpage as `NewtonMin.m`, where the line-search method is implemented when the parameter `type` is set to 2.

EXAMPLE B.16. Consider the multivariable function

$$f(x_1, x_2) = \zeta_1^2 + \zeta_2^2 + 2 \tag{B.105}$$

where

$$\begin{aligned} \zeta_1(x_1, x_2) &= 5 \tanh\left(-\frac{x_1}{3} + \frac{x_2}{3} - \frac{1}{2}\right) \\ \zeta_2(x_1, x_2) &= 1 - \frac{x_2}{2} \end{aligned}$$

A surface plot of $f(x_1, x_2)$ is shown in Figure B.12. When the enhanced Newton with double-dogleg method was used to find the minimum of $f(x_1, x_2)$, we see in Figure B.12 that starting with $(x_1, x_2)_0 = (4, -6)$, it took only three iterations to settle at the minimum point of $(x_1, x_2)^* = (0.5, 2)$ which yields the value $f = 2$. Conversely, applying the line-search method, in this case with the same initial point, will converge to a different point $(x_1, x_2) = (-432, 2)$ with $f = 27$.

A particular property of the function $f(x_1, x_2)$ in (B.105) is that the minimum is located in a narrow trough. When the line-search approach was used, starting at $(x_1, x_2)_0 = (4, -6)$, the first Newton step pointed away from $(x_1, x_2)^* = (0.5, 2)$. However, the double-dogleg method constrained the search to a local model-trust region while mixing the gradient search direction with the Newton direction. This allowed the double-dogleg method a better chance of locating the minima that is close to the initial guess.

B.14 Nonlinear Least Squares via Levenberg-Marquardt

There are several cases in which the linear least-squares methods given in Section 2.5 are not applicable. In those cases, Newton's methods can be used to find the least-squares solution when the unknown parameters are in nonlinear form. We can formulate the nonlinear least squares as follows:

$$\min_{\mathbf{x}} \frac{1}{2} \left\| \mathbf{r}(\mathbf{x}) \right\|^2 \tag{B.106}$$

where $\mathbf{r}$ is the vector of residuals

$$\mathbf{r}(\mathbf{x}) = \begin{pmatrix} r_1(x_1, \ldots, x_n) \\ \vdots \\ r_m(x_1, \ldots, x_n) \end{pmatrix}$$

with $m \geq n$. For instance, suppose we wish to estimate parameters $\mathbf{x} = (x_1, \ldots, x_n)^T$ of a nonlinear equation

$$f(\mathbf{x}, \mathbf{w}) = 0$$

where $\mathbf{w}$ are measured variables, for example, from experiments. Assuming we have m sets of data given by $\mathbf{w}_1, \ldots, \mathbf{w}_m$, the residual functions are

$$r_i(\mathbf{x}) = f(\mathbf{x}, \mathbf{w}_i) \qquad i = 1, \ldots, m$$

One could apply Newton's method directly to (B.106). However, doing so would involve the calculation of $d^2\mathbf{r}/d\mathbf{x}^2$,

$$\frac{d^2}{d\mathbf{x}^2}\mathbf{r} = \left(\frac{d\mathbf{r}}{d\mathbf{x}}\right)^T \left(\frac{d\mathbf{r}}{d\mathbf{x}}\right) + \sum_{i=1}^{m} r_i \frac{d^2 r_i}{d\mathbf{x}^2}$$

which is cumbersome when m is large.

Another approach is to first linearize $\mathbf{r}$ around $\mathbf{x}_0$, that is,

$$\mathbf{r}_{(\mathbf{x})} = \mathbf{r}_{(\mathbf{x}_0)} + \left(\frac{d}{d\mathbf{x}}\mathbf{r}\right)\Big|_{\mathbf{x}=\mathbf{x}_0} (\mathbf{x} - \mathbf{x}_0) = \mathbf{r}_{(\mathbf{x}_0)} + J_{(\mathbf{x}_0)} (\mathbf{x} - \mathbf{x}_0)$$

where J is the Jacobian matrix given by

$$J_{(\mathbf{x}_0)} = \begin{pmatrix} \frac{\partial r_1}{\partial x_1} & \cdots & \frac{\partial r_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_m}{\partial x_1} & \cdots & \frac{\partial r_m}{\partial x_n} \end{pmatrix}\Bigg|_{\mathbf{x}=\mathbf{x}_0}$$

This transforms the nonlinear least-squares problem (B.106) back to a linear least-squares problem (cf. Section 2.5), that is,

$$\min_{\mathbf{x}-\mathbf{x}_0} \frac{1}{2} \left\| \mathbf{r}_{(\mathbf{x}_0)} + J_{(\mathbf{x}_0)} (\mathbf{x} - \mathbf{x}_0) \right\|^2 \tag{B.107}$$

whose solution is given by the normal equation,

$$\mathbf{x} - \mathbf{x}_0 = -\left(J_{(\mathbf{x}_0)}^T J_{(\mathbf{x}_0)}\right)^{-1} J_{(\mathbf{x}_0)}^T \mathbf{r}_{(\mathbf{x}_0)}$$

We obtain an iterative procedure by letting $\mathbf{x}^{(k)} = \mathbf{x}_0$ be the current estimate and letting $\mathbf{x}^{(k+1)} = \mathbf{x}$ be the next update. This approach is known as the **Gauss-Newton method** for nonlinear least-squares problem:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left(J_k^T J_k\right)^{-1} J_k^T \mathbf{r}_k \tag{B.108}$$

where

$$J_k = J\left(\mathbf{x}^{(k)}\right); \qquad \mathbf{r}_k = \mathbf{r}\left(\mathbf{x}^{(k)}\right)$$

As it was with Newton methods, the convergence of the Gauss-Newton method may need to be enhanced either by the line-search method or by a model-trust region method. However, instead of the line search or the double-dogleg approach, we discuss another model-trust region method known as the **Levenberg-Marquardt method**.

Recall, from Section B.13, that the model trust region is a sphere centered around the current value $\mathbf{x}^{(k)}$. The minimization problem can then be modified to be the constrained form of (B.107):

$$\min_{\Delta^k \mathbf{x}} \frac{1}{2} \left\| \mathbf{r}_k + J_k \Delta^k \mathbf{x} \right\|^2 \qquad \text{subject to} \quad \left\| \Delta^k \mathbf{x} \right\| \leq M_k \tag{B.109}$$

where $\Delta^k \mathbf{x} = \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ is the update step and M_k is the radius of the trust region.

From Figure B.10, we see that there is a unique point in the boundary of the trust region where the value of function in the convex surface is minimized. This observation can be formalized by the following lemma:

LEMMA B.2. **Levenberg-Marquardt Update Form**

The solution to the minimization problem (B.109) is given by

$$\Delta^k \mathbf{x}^* = -\left(J_k^T J_k + \mu I\right)^{-1} J_k^T \mathbf{r}_k \tag{B.110}$$

for some unique value $\mu \geq 0$.

PROOF. (See Section B.10.2.)

Lemma B.2 redirects the minimization problem of (B.109) to the identification of μ such that

$$q(\mu) = \|s_\mu\| - M_k = 0 \tag{B.111}$$

where

$$s_\mu = -\left(J_k^T J_k + \mu I\right)^{-1} J_k^T \mathbf{r}_k$$

Note that we set $\mu = 0$ if $\|s_0\| < M_k$. Also, the derivative of $q(\mu)$ is given by

$$q'(\mu) = \frac{dq}{d\mu} = -\frac{s_\mu^T \left(J_k^T J_k + \mu I\right)^{-1} s_\mu}{\|s_\mu\|} \tag{B.112}$$

Although the Newton method can be used to solve (B.111), the **Moré method** has been shown to have improved convergence. Details of the Moré algorithm are included in Section B.14.1.

To summarize, we have the Levenberg-Marquardt method:

Algorithm of Levenberg-Marquardt Method for Nonlinear Least Squares.

1. **Initialize.** Choose an initial guess: $\mathbf{x}^{(0)}$
2. **Update.** Repeat the following steps until either $\left\|\mathbf{r}\left(\mathbf{x}^{(k)}\right)\right\| \leq \epsilon$ or the number of iterations have been exceeded
   (a) Calculate J_k.
   (b) Calculate μ and s_μ using the Moré algorithm.
   (c) Set $\Delta_k\mathbf{x} = s_\mu$ and check if $\Delta_k\mathbf{x}$ is acceptable. If

$$\left\|\mathbf{r}\left(\mathbf{x}^{(k)} + \Delta_k\mathbf{x}\right)\right\|^2 > \left\|\mathbf{r}\left(\mathbf{x}^{(k)}\right)\right\|^2 + 2\alpha\mathbf{r}_k^T J_k \Delta_k\mathbf{x}$$

with $\alpha \in (0, 0.5)$ (typically $\alpha = 10^{-4}$), then update is unacceptable. Modify the trust region:

$$M_k \leftarrow \max\Big(\ 0.1M_k,\ \ \min\left(\ 0.5M_k, \lambda\left\|\Delta_k\mathbf{x}\right\|\ \right)\ \Big)$$

where

$$\lambda = -\frac{\mathbf{r}_k^T J_k \Delta_k\mathbf{x}}{\left(\ \left\|\mathbf{r}\left(\mathbf{x}^{(k)} + \Delta_k\mathbf{x}\right)\right\|^2 - \left\|\mathbf{r}\left(\mathbf{x}^{(k)}\right)\right\|^2 - 2\mathbf{r}_k^T J_k \Delta_k\mathbf{x}\right)}$$

and repeat from step 2b above.

Otherwise, if acceptable, continue to next step.

   (d) Update $\mathbf{x}^{(k)}$: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \Delta_k\mathbf{x}$

Remarks: A MATLAB code for the Levenberg-Marquardt method (using the Morè algorithm) for solving nonlinear least squares is available on the book's webpage as `levenmarq.m`.

EXAMPLE B.17. Suppose we want to estimate the parameters a, b, c, d, and e of the function:

$$y = d\ \exp\left(ax^2 + bx + c\right) + e$$

to fit the data given in Table B.5. Applying the Levenberg-Marquardt method with the initial guess: $(a, b, c, d, e) = (0, 0, 0, 0, 0)$, we obtain the estimates: $(a, b, c, d, e) = (-0.0519, 0.9355, -1.1346, 0.0399, 2.0055)$. A plot of the model, together with data points, is shown in Figure B.13. We also show, in the right plot of same figure, the number of iterations used and the final value of the residual norm.

B.14.1 Appendix: Moré Method

Algorithm of Moré Method to obtain μ:

1. Generate initial guess.

$$\mu^{(0)} = \begin{cases} 0 & \text{if } k = 0 \\ \left(\mu|_{\mathbf{x}=\mathbf{x}^{(k-1)}}\right)\dfrac{M_{k-1}}{M_k} & \text{otherwise} \end{cases}$$

Table B.5. *Data for example B.17*

x	y	x	y	x	y
0.1152	2.0224	6.5207	2.6118	12.6037	2.4487
0.7604	2.0303	7.0276	2.7355	13.1106	2.3750
1.2673	2.0408	7.5346	2.7855	13.7097	2.2855
2.7419	2.1197	8.3180	2.8539	14.3548	2.2013
3.4332	2.1803	9.4700	2.8645	15.0461	2.1355
4.3088	2.2776	10.3917	2.7934	16.2442	2.0671
4.8618	2.3645	11.2212	2.6645	17.5806	2.0250
5.3687	2.4382	12.0507	2.5487	19.7465	2.0039
6.4747	2.6303				

2. Update.

$$\mu^{(j)} = \mu^{(j-1)} - \frac{\left\| s_{\mu^{(j-1)}} \right\|}{M_k} \left(\frac{q}{q'} \bigg|_{\mu=\mu^{(j-1)}} \right)$$

3. Clip μ between minimum and maximum values

$$\mu^{(j)} \leftarrow \begin{cases} \mu^{(j)} & \text{if } \mathrm{Lo}_j \le \mu^{(j)} \le \mathrm{Hi}_j \\ \max\left(\sqrt{\mathrm{Lo}_j \cdot \mathrm{Hi}_j}\ ,\ 10^{-3}\mathrm{Hi}_j \right) & \text{otherwise} \end{cases}$$

where

$$\mathrm{Lo}_j \leftarrow \begin{cases} -\dfrac{q(0)}{q'(0)} & \text{if } j = 0 \\ \max\left(\left(\mu - \dfrac{q}{q'} \right) \bigg|_{\mu=\mu^{(j-1)}}\ ,\ \mathrm{Lo}_{j-1} \right) & \text{otherwise} \end{cases}$$

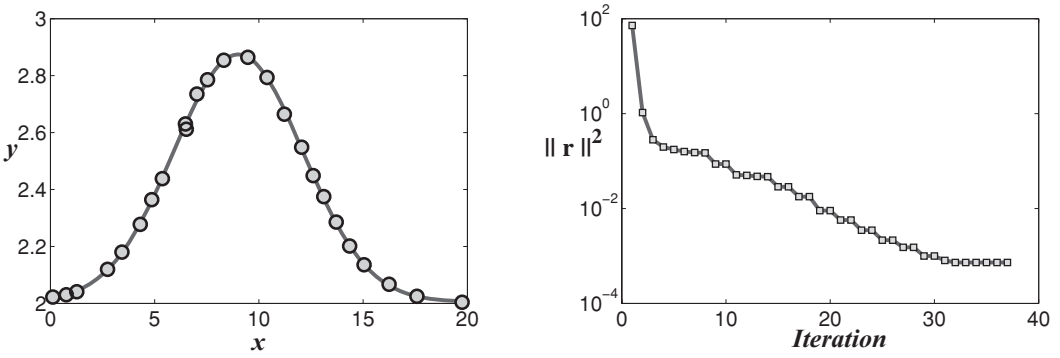


Figure B.13. The model together with the data given in Table B.5. On the right plot, we have the number of iterations performed and the corresponding norm of the residuals.

$$\mathrm{Hi}_j \leftarrow \begin{cases} \dfrac{\left\| J_k^T \mathbf{r}_k \right\|}{M_k} & \text{if } j = 0 \\ \min\left(\mathrm{Hi}_{j-1}, \mu^{(j-1)}\right) & \text{if } q\left(\mu^{(j-1)}\right) < 0 \\ \mathrm{Hi}_{j-1} & \text{otherwise} \end{cases}$$

4. Repeat until:

$$\left\| s_{\mu^{(j)}} \right\| \in \left[0.9M_k \,,\ 1.1M_k \right]$$

APPENDIX C

Additional Details and Fortification for Chapter 3

C.1 Proofs of Lemmas and Theorems of Chapter 3

C.1.1 Proof of Eigenvalue Properties

- **Property 1**: Eigenvalues of triangular matrices are the diagonal elements. Let A be triangular then

$$\det(A - \lambda I) = \prod_{i=1}^{N} (a_{ii} - \lambda) = 0$$

Thus the roots are: $a_{11}, \ldots, a_{NN}$.

For diagonal matrices,

$$A\mathbf{e}_i = a_{ii}\mathbf{e}_i = \lambda_i \mathbf{e}_i$$

Thus the eigenvectors of diagonal matrices are the columns of the identity matrix.

- **Property 2**: Eigenvalues of block triangular matrices are the eigenvalues of the block diagonals.

Let A_{ii} be i^{th} block diagonal of a block triangular matrix $\mathbf{A}$, then

$$\det(\mathbf{A} - \lambda I) = \prod_{i=1}^{N} (A_{ii} - \lambda I) = 0$$

or

$$\det(A_{ii} - \lambda I) = 0$$

- **Property 3**: Eigenvalues of αA is $\alpha\lambda$.

$$(\alpha A)\,\mathbf{v} = (\alpha\lambda)\,\mathbf{v}$$

- **Property 4**: Eigenvalues of A and A^T are the same.

Because $\det(B) = \det\left(B^T\right)$,

$$\det(A - \lambda I) = \det(A - \lambda I)^T = \det\left(A^T - \lambda I\right) = 0$$

Thus the characteristic equation for A and A^T is the same, yielding the same eigenvalues.

- **Property 5**: Eigenvalues of A^k are λ^k.
  For $k = 0$, $A^0 = I$ and the eigenvalues are all 1's.
  For $k > 0$,

$$A^k\mathbf{v} = A^{k-1}(A\mathbf{v}) = \lambda A^{k-1}\mathbf{v} = \cdots = \lambda^k\mathbf{v}$$

  For $k = -1$, assuming A is nonsingular,

$$\mathbf{v} = A^{-1}A\mathbf{v} = \lambda A^{-1}\mathbf{v} \quad \rightarrow \quad A^{-1}\mathbf{v} = \frac{1}{\lambda}\mathbf{v}$$

  (Note: Property 7 implies that eigenvalues are nonzero for nonsingular matrices.) Then for $k < -1$,

$$A^k\mathbf{v} = A^{k+1}\left(A^{-1}\mathbf{v}\right) = \cdots = \lambda^k\mathbf{v}$$

- **Property 6**: Eigenvalues are preserved by similarity transformations.
  Using the eigvenvalue equation for $T^{-1}AT$,

$$\begin{aligned}\det\left(T^{-1}AT - \lambda I\right) &= \det\left(T^{-1}\right)\det(A - \lambda I)\det(T)\\ &= \det(A - \lambda I)\end{aligned}$$

  Because the characteristic polynomials for both A and $T^{-1}AT$ are the same, the eigenvalues will also be the same.
  If $\mathbf{v}$ is an eigenvector of A corresponding to λ and $B = T^{-1}AT$ then

$$A\mathbf{v} = \lambda\mathbf{v} \;\rightarrow\; \begin{aligned} TBT^{-1}\mathbf{v} &= \lambda\mathbf{v}\\ B\left(T^{-1}\mathbf{v}\right) &= \lambda\left(T^{-1}\mathbf{v}\right)\end{aligned}$$

  that is, $T^{-1}\mathbf{v}$ is a eigenvector of B.
- **Property 7**: $\prod\lambda_i = |A|$.
  Using the Schur triangularization,

$$U^*AU = \begin{pmatrix} \lambda_1 & \times & \cdots & \times\\ 0 & \lambda_2 & \cdots & \times\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_N \end{pmatrix}$$

  where U is unitary and '$\times$' represent possible nonzero entries. After taking the determinant of both sides,

$$\left|U^*\right||A||U| = |A| = \prod_{i=1}^{N}\lambda_i$$

- **Property 8**: $\sum\lambda_i = \mathrm{tr}(A)$.
  Using Schur triangularization,

$$U^*AU = \begin{pmatrix} \lambda_1 & \times & \cdots & \times\\ 0 & \lambda_2 & \cdots & \times\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_N \end{pmatrix}$$

  After taking the trace of both sides.

$$\mathrm{tr}(U^*AU) = \mathrm{tr}(AUU^*) = \mathrm{tr}(A) = \sum_{i=1}^{N}\lambda_i$$

- **Property 9**: Eigenvalues of Hermitian matrices are real, and eigenvalues of skew-Hermitian matrices are pure imaginary.

  Let H be Hermitian, then

$$(\mathbf{v}^* H\mathbf{v})^* = \mathbf{v}^* H^*\mathbf{v} = \mathbf{v}^* H\mathbf{v}$$

  which means $\mathbf{v}^* H\mathbf{v}$ is real. Now let λ be an eigenvalue of H. Then

$$\begin{aligned} H\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{v}^* H\mathbf{v} &= \lambda\mathbf{v}^*\mathbf{v} \end{aligned}$$

  Because $\mathbf{v}^*\mathbf{v}$ and $\mathbf{v}^* H\mathbf{v}$ are real, λ has to be real.

  Similarly, let $\widehat{H}$ be skew-Hermitian, then

$$\left(\widehat{\mathbf{v}}^*\widehat{H}\widehat{\mathbf{v}}\right)^* = \widehat{\mathbf{v}}^*\widehat{H}^*\widehat{\mathbf{v}} = -\left(\widehat{\mathbf{v}}^*\widehat{H}^*\widehat{\mathbf{v}}\right)$$

  which means $\widehat{\mathbf{v}}^*\widehat{H}\widehat{\mathbf{v}}$ is pure imaginary. Let $\widehat{\lambda}$ be an eigenvalue of $\widehat{H}$, then

$$\begin{aligned} \widehat{H}\widehat{\mathbf{v}} &= \widehat{\lambda}\widehat{\mathbf{v}} \\ \widehat{\mathbf{v}}^*\widehat{H}\widehat{\mathbf{v}} &= \widehat{\lambda}\widehat{\mathbf{v}}^*\widehat{\mathbf{v}} \end{aligned}$$

  Because $\widehat{\mathbf{v}}^*\widehat{\mathbf{v}}$ is real and $\widehat{\mathbf{v}}^*\widehat{H}\widehat{\mathbf{v}}$ is pure imaginary, $\widehat{\lambda}$ has to be pure imaginary.
- **Property 10**: Eigenvalues of positive definite Hermitian matrices are positive.

  Because H is positive definite, $\mathbf{v}^* H\mathbf{v} > 0$, where $\mathbf{v}$ is an eigenvector of H. However, $\mathbf{v}^* H\mathbf{v} = \lambda\,|\mathbf{v}|^2$. Because $\mathbf{v} > 0$, we must have $\lambda > 0$.
- **Property 11**: Eigenvectors of Hermitian matrices are orthogonal.

  If H is Hermitian, $H^* H = H^2 = HH^*$. Thus, according to Definition 3.5, H is a normal matrix. Then the orthogonality of the eigenvectors of H follows as a corollary to Theorem 3.1.
- **Property 12**: Distinct eigenvalues yield linearly independent eigenvectors.

  Let $\lambda_1, \ldots, \lambda_M$ be a set of distinct eigenvalues of $A[=]N \times N$, with $M \leq N$, and let $\mathbf{v}_1, \ldots, \mathbf{v}_M$ be the corresponding eigenvectors. Then

$$A^k\mathbf{v}_i = \lambda_i A^{k-1}\mathbf{v}_i = \cdots = \lambda_i^k\mathbf{v}_i$$

  We want to find a linear combination of the eigenvector that would equal the zero vector,

$$\alpha_1\mathbf{v}_1 + \cdots + \alpha_n\mathbf{v}_n = 0$$

  After premultiplication by $A,\ A^2, \ldots, A^{M-1}$,

$$\begin{aligned} \alpha_1\lambda_1\mathbf{v}_1 + \cdots + \alpha_M\lambda_M\mathbf{v}_M &= 0 \\ &\vdots \\ \alpha_1\lambda_1^{M-1}\mathbf{v}_1 + \cdots + \alpha_M\lambda_M^{M-1}\mathbf{v}_M &= 0 \end{aligned}$$

  Combining these equations,

$$\left(\begin{array}{c|c|c|c} \alpha_1\mathbf{v}_1 & \cdots & \alpha_M\mathbf{v}_M \end{array}\right) \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{M-1} \\ 1 & \lambda_2 & \cdots & \lambda_2^{M-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_M & \cdots & \lambda_M^{M-1} \end{pmatrix} = 0_{[N\times M]}$$

The Vandermonde matrix is nonsingular if $\lambda_1 \neq \cdots \neq \lambda_M$ (cf. Exercise **E1.14**). Thus

$$\alpha_1 \mathbf{v}_1 = \cdots = \alpha_n \mathbf{v}_M = 0$$

Because none of the eigenvectors are zero vectors, we must have

$$\alpha_1 = \cdots = \alpha_M = 0$$

Thus $\{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ is a linearly independent set of eigenvectors.

C.1.2 Proof for Properties of Normal Matrices (Theorem 3.1)

Applying Schur triangularization to A,

$$U^*AU = B = \begin{pmatrix} \lambda_1 & b_{12} & \cdots & b_{1,N} \\ & \ddots & \ddots & \vdots \\ & & \ddots & b_{N-1,N} \\ 0 & & & \lambda_N \end{pmatrix}$$

If A is normal, then $B = U^*AU$ will also be normal, that is,

$$B^*B = (U^*AU)^*(U^*AU) = U^*A^*AU = U^*AA^*U = (U^*AU)(U^*AU)^* = BB^*$$

Because B is normal, we can equate the first diagonal element of B^*B to the first diagonal element of BB^* as follows:

$$|\lambda_1|^2 = |\lambda_1|^2 + \sum_{k=2}^{N} |b_{1k}|^2$$

This is possible only if $b_{1k} = 0$, for $k = 2, \ldots, N$. Having established this, we can now equate the second diagonal element of B^*B to the second diagonal element of BB^* as follows:

$$|\lambda_2|^2 = |\lambda_2|^2 + \sum_{k=3}^{N} |b_{2k}|^2$$

and conclude that $b_{2k} = 0$, for $k = 3, \ldots, N$. We can continue this logic until the $(N-1)^{\text{th}}$ diagonal of B^*B. At the end of this process, we will have shown that B is diagonal.

We have just established that as long as A is normal, then $U^*AU = \Lambda$, where Λ contains all the eigenvalues of A, including the case of repeated roots. Next, we can show that the columns of U are the eigenvectors of A,

$$\begin{aligned} U^*AU &= \Lambda \\ AU &= U\Lambda \\ \left(AU_{\bullet,1} \mid \cdots \mid AU_{\bullet,N} \right) &= \left(\lambda_1 U_{\bullet,1} \mid \cdots \mid \lambda_N U_{\bullet,N} \right) \end{aligned}$$

or

$$AU_{\bullet,i} = \lambda_i U_{\bullet,i}$$

Now assume that a given matrix, say $C[=]N \times N$, has orthonormal eigenvectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_N\}$ corresponding to eigenvalues $\left\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_N\right\}$, that is, $V^*CV = \widehat{\Lambda}$, where $\widehat{\Lambda} = \text{diag}\left(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_N\right)$.

$$\begin{aligned} \widehat{\Lambda}^*\widehat{\Lambda} &= (V^*CV)^*(V^*CV) = V^*C^*CV \\ \widehat{\Lambda}\widehat{\Lambda}^* &= (V^*CV)(V^*CV)^* = V^*CC^*V \end{aligned}$$

Because $\widehat{\Lambda}^*\widehat{\Lambda} = \widehat{\Lambda}\widehat{\Lambda}^*$, we have

$$C^*C = CC^*$$

This means that when all the eigenvectors are orthogonal, the matrix is guaranteed to be a normal matrix.

C.1.3 Proof That Under Rank Conditions, Matrix Is Diagonalizable (Theorem 3.2)

Suppose λ_1 is repeated k_1 times. From the rank assumption,

$$\text{rank}(\lambda_1 I - A) = N - k_1$$

means that solving

$$(\lambda_1 I - A)\,\mathbf{v} = 0$$

for the eigenvectors contain k_1 arbitrary constants. Thus there are k_1 linearly independent eigenvectors that can be obtained for λ_1. Likewise, there are k_2 linearly independent eigenvectors that can be obtained for λ_2, and so forth. Let the first set of k_1 eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k_1}$ correspond to λ_1 while the subsequent set of k_2 eigenvectors $\mathbf{v}_{k_1+1}, \ldots, \mathbf{v}_{k_1+k_2}$ correspond to eigenvalue λ_2, and so forth. Each eigenvector from the first set is linearly independent from the other set of eigenvectors. And the same can be said of the eigenvectors of the other sets. In the end, all the N eigenvectors obtained will form a linearly independent set.

C.1.4 Proof of Cayley Hamilton Theorem (Theorem 3.3)

Using the Jordan canonical decomposition, $A = TJT^{-1}$, where T is the modal matrix, and J is a matrix in Jordan canonical form with M Jordan blocks,

$$\begin{aligned} a_0 I + a_1 A + \cdots + a_n A^N &= T(a_0 I + a_1 J + \cdots + a_n J^N)T^{-1} \\ &= T \begin{pmatrix} \text{charpoly}(J_1) & 0 & \cdots & 0 \\ 0 & \text{charpoly}(J_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{charpoly}(J_M) \end{pmatrix} T^{-1} \end{aligned} \tag{C.1}$$

The elements of charpoly(J_i) are either 0, charpoly(λ_i), or derivatives of charpoly(λ_i), multiplied by finite scalars. Thus charpoly(J_i) are zero matrices, and the right-hand side of Equation (C.1) is a zero matrix.

C.2 QR Method for Eigenvalue Calculations

For large systems, the determination of eigenvalues and eigenvectors can become susceptible to numerical errors, especially because the roots of polynomials are very sensitive to small perturbations in the polynomial coefficients. A more reliable method is available that uses the QR decomposition method. First, we have present the QR algorithm. Then we describe the **power method**, which is the basis for the QR method for finding eigenvalues. Finally, we apply the QR method.

C.2.1 QR Algorithm

QR Decomposition Algorithm (using Householder operators):

Given $A[=]N \times M$.

1. Initialize. $K = A$, $\widehat{Q} = I_N$
2. Iterate.
   For $j = 1, \ldots, \min(N, M) - 1$
   (a) Extract first column of K: $\mathbf{u} = K_{\bullet,1}$
   (b) Construct a Householder matrix:

$$\begin{aligned} u_1 &\leftarrow u_1 - \|\mathbf{u}\| \\ H &= I - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\mathbf{u}^* \end{aligned}$$

   (c) Update K: $K \leftarrow HK$
   (d) Update last $(N - j)$ rows of $\widehat{Q}$: $\widehat{Q}_{[j,\ldots,N],\bullet} \leftarrow H\widehat{Q}_{[j,\ldots,N],\bullet}$
   (e) Remove the first row and first column of K: $K \leftarrow K_{1,1\downarrow}$
3. Trim the last $(N - M)$ rows of $\widehat{Q}$ if $N > M$: $\widehat{Q} \leftarrow \widehat{Q}_{[M+1,\ldots,N],\bullet\downarrow}$
4. Obtain Q and R:

$$\begin{aligned} Q &= \widehat{Q}^* \\ R &= \widehat{Q}A \end{aligned}$$

C.2.2 Power Method

Let square matrix A have a **dominant eigenvalue**, that is, $|\lambda_1| > |\lambda_j|$, $j > 1$. An iterative approach known as the **power method** can be used to find λ_1 and its corresponding eigenvector $\mathbf{v}_1$.

Power Method Algorithm:

Given matrix $A[=]N \times N$ and tolerance $\epsilon > 0$.

1. Initialize. Set $\mathbf{w} = 0$ and select a random vector for $\mathbf{v}$
2. Iterate. While $\|\mathbf{v} - \mathbf{w}\| > \epsilon$

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{v} \\ \mathbf{v} &\leftarrow A\mathbf{w} \\ \mathbf{v} &\leftarrow \frac{\mathbf{v}}{\|\mathbf{v}\|} \end{aligned}$$

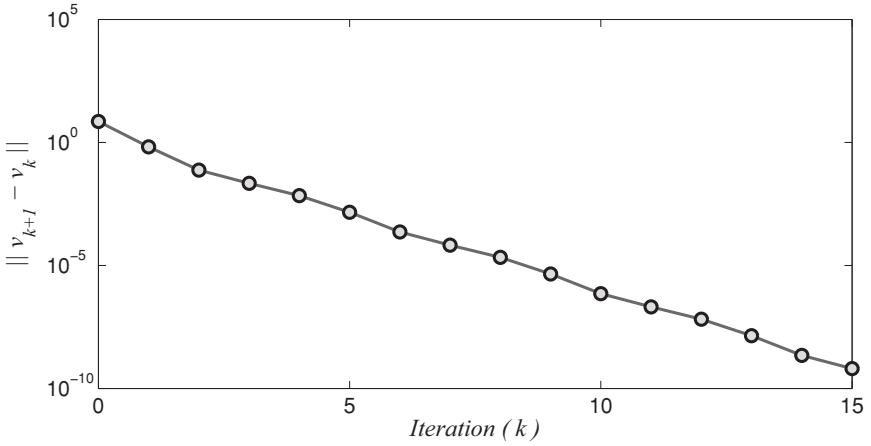


Figure C.1. Convergence of the eigenvector estimation using the power method.

3. Obtain eigenvalue: $\lambda = \dfrac{1}{\mathbf{v}^*\mathbf{v}}\mathbf{v}^* A\mathbf{v}$

A short proof for the validity of the power method is left as an exercise (cf. **E3.24**). The power method is simple but is limited to finding only the dominant eigenvalue and its eigenvector. Also, if the eigenvalue with the largest magnitude is close in magnitude to the second largest, then convergence is very slow. This means that convergence may even suffer for those with complex eigenvalues that happen to have the largest magnitude. In those cases, there are block versions of the power method.

EXAMPLE C.1. Let A be given by

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$$

the power method found the largest eigenvalue $\lambda = 6$ and its corresponding eigenvector $\mathbf{v} = (0.5774, 0.5774, 0.5774)^T$ in a few iterations. The norm $\left\|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\right\|$ is shown in Figure C.1.

C.2.3 QR Method for Finding Eigenvalues

As discussed in Section 2.6, matrix A can be factored into a product, $A = QR$ where Q is unitary and R is upper triangular. If we let $A^{[\langle 1\rangle]}$ be a similarity transformation of A based on Q,

$$A^{[\langle 1\rangle]} = Q^* A Q = RQ \tag{C.2}$$

then $A^{[\langle 1\rangle]}$ simply has reversed the order of Q and R. Because the eigenvalues are preserved under similarity transformations (cf. Section 3.3), A and $A^{[\langle 1\rangle]}$ will have the same set of eigenvalues. One could repeat this process k times and obtain

$$\begin{aligned} A^{[\langle k\rangle]} &= Q^{[\langle k\rangle]} R^{[\langle k\rangle]} \\ A^{[\langle k+1\rangle]} &= R^{[\langle k\rangle]} Q^{[\langle k\rangle]} \end{aligned}$$

where the eigenvalues of $A^{[\langle k \rangle]}$ will be the same as those of A. Because $R^{[\langle k \rangle]}$ is upper triangular, one can show[1] that $A^{[\langle k \rangle]}$ will converge to a matrix that can be partitioned as follows:

$$\lim_{k \to \infty} A^{[\langle k \rangle]} = \left(\begin{array}{c|c} B & C \\ \hline 0 & F \end{array} \right) \tag{C.3}$$

where F is either a 1×1 or a 2×2 submatrix. Because the last matrix is block triangular, the eigenvalues of A will be the union of the eigenvalues of B and the eigenvalues of F. If $F[=]1 \times 1$, then F is a real eigenvalue of A; otherwise, two eigenvalues of A can be found using (3.21).

The same process can now be applied on B. The process continues with QR iterations applied to increasingly smaller matrices until all the eigenvalues of A are found.

EXAMPLE C.2. Consider the matrix

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

After approximately 33 iterations using the QR method described, we obtain

$$A^{[\langle 33 \rangle]} = \begin{pmatrix} -1.3333 & 1.1785 & -0.4083 \\ 0.9428 & -0.6667 & 0.5774 \\ 0.0000 & 0.0000 & 1.0000 \end{pmatrix}$$

which means one eigenvalue can be found as $\lambda_1 = 1$. For the remaining two eigenvalues, we can extract the upper left 2×2 submatrix and use (3.21) to obtain $\lambda_2 = 1 + i$ and $\lambda_3 = 1 - i$.

Although the QR method will converge to the required eigenvalues, the convergence can also be slow sometimes, as shown in preceding example. Two enhancements significantly help in accelerating the convergence. The first enhancement is called the **shifted QR method**. The second enhancement is the **Hessenberg formulation**. Both of these enhancements combine to form the **modified QR method**, which will find the eigenvalues of A with reasonable accuracy. The details of the modified QR method are included in Section C.2.4.

C.2.4 Modified QR Method

In this section, we discuss the two enhancements that will accelerate the convergence of the QR methods for evaluation of the eigenvalues. The first enhancement is to shift the matrix $A^{\langle k \rangle}$ by a scaled identity matrix. Then second is to use Householder transformations to achieve a Hesseberg matrix, which is an upper triangular matrix, but with an additional subdiagonal next to the principal diagonal.

[1] For a detailed proof, refer to G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd Edition, 1996, John Hopkins University Press.

C.2.5 Shifted QR Method

Instead of taking the QR decomposition of $A^{\langle k\rangle}$, one can first shift it as follows:

$$\widetilde{A}^{\langle k\rangle} = A^{\langle k\rangle} - \sigma^{\langle k\rangle} I \tag{C.4}$$

where $\sigma^{\langle k\rangle}$ is the $(N,N)^{\text{th}}$ element of $A^{\langle k\rangle}$.

We now take the QR decomposition of $\widetilde{A}^{\langle k\rangle}$,

$$\widetilde{A}^{\langle k\rangle} = \widetilde{Q}^{\langle k\rangle}\widetilde{R}^{\langle k\rangle} \tag{C.5}$$

which we use to form $A^{\langle k+1\rangle}$ by

$$A^{\langle k+1\rangle} = \widetilde{R}^{\langle k\rangle}\widetilde{Q}^{\langle k\rangle} + \sigma^{\langle k\rangle} I \tag{C.6}$$

Even with the modifications given by (C.4), (C.5), and (C.6), $A^{\langle k+1\rangle}$ will still be a similarity transformation of $A^{\langle k\rangle}$ starting with $A^{\langle 0\rangle} = A$. To see this,

$$\begin{aligned} A^{\langle k+1\rangle} &= \widetilde{R}^{\langle k\rangle}\widetilde{Q}^{\langle k\rangle} + \sigma^{\langle k\rangle} I \\ &= \left(\widetilde{Q}^{\langle k\rangle}\right)^{-1}\left(A^{\langle k\rangle} - \sigma^{\langle k\rangle} I\right)\widetilde{Q}^{\langle k\rangle} + \sigma^{\langle k\rangle} I \\ &= \left(\widetilde{Q}^{\langle k\rangle}\right)^{-1} A^{\langle k\rangle}\widetilde{Q}^{\langle k\rangle} \end{aligned}$$

Note that these modifications introduce only $2N$ extra operations: the subtraction of $\sigma^{\langle k\rangle} I$ from the diagonal of $A^{\langle k\rangle}$, and the addition of $\sigma^{\langle k\rangle} I$ to the diagonal of $\widetilde{R}^{\langle k\rangle}\widetilde{Q}^{\langle k\rangle}$. Nonetheless, the improvements in convergence toward attaining the form given in (C.3) will be significant.

C.2.6 Hessenberg Forms

The second enhancement to the QR method is the use of Householder operators to transform A into an **upper Hessenberg form**. A matrix is said to have the upper Hessenberg form if all elements below the first subdiagonal are zero,

$$H = \begin{pmatrix} \times & \times & \cdots & \times \\ \times & \ddots & \ddots & \vdots \\ & \ddots & \ddots & \times \\ 0 & & \times & \times \end{pmatrix} \tag{C.7}$$

where "$\times$" denotes arbitrary values.

To obtain the upper Hessenberg form, we use the Householder operators $U_{\mathbf{x}-\mathbf{y}}$ given in (3.7),

$$U_{\mathbf{x}-\mathbf{y}} = I - \frac{2}{(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})}(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^*$$

which will transform $\mathbf{x}$ to $\mathbf{y}$, as long as $\|\mathbf{x}\| = \|\mathbf{y}\|$. With the aim of introducing zeros, we will choose $\mathbf{y}$ to be

$$\mathbf{y} = \begin{pmatrix} \|\mathbf{x}\| \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Two properties of Householder operators are noteworthy: they are unitary and Hermitian. The following algorithm will generate a Householder matrix H such that HAH will have an upper Hessenberg form. Also, because HAH is a similarity transformation of A, both A and HAH will have the same set of eigenvalues.

Algorithm for Householder Transformations of A to Upper Hessenberg Form:

Start with $G \leftarrow A$.

For $k = 1, \ldots, (N-2)$

1. Extract vector $\mathbf{w}$: $\quad w_i = G_{k+i,k} \quad ; \quad i = 1, \ldots, (N-k)$
2. Evaluate H:

$$H = \begin{cases} I_{[N]} & \text{if } \|\mathbf{w}-\mathbf{y}\| = 0 \\ \left(\begin{array}{c|c} I_{[k]} & 0 \\ \hline 0 & U_{\mathbf{w}-\mathbf{y}} \end{array}\right) & \text{otherwise} \end{cases}$$

$$\text{where,} \qquad \mathbf{y} = \begin{pmatrix} \|\mathbf{w}\| & 0 & \cdots & 0 \end{pmatrix}^T$$

$$U_{\mathbf{w}-\mathbf{y}} = I - \frac{2}{\|\mathbf{w}-\mathbf{y}\|^2}(\mathbf{w}-\mathbf{y})(\mathbf{w}-\mathbf{y})^*$$

3. **Update G:** $\quad G \leftarrow H\,G\,H$

End loop for k

Because the Householder operators $U_{\mathbf{v}}$ will be applied on matrices, we note the following improvements:

Let $\beta = 2/(\mathbf{v}^*\mathbf{v})$, $\mathbf{w}_1 = A^*\mathbf{v}$, $\mathbf{w}_2 = A\mathbf{v}$ and $\gamma = \mathbf{v}^*A\mathbf{v}$,

1. Instead of multiplication $U_{\mathbf{v}}A$, we use $U_{\mathbf{v}}A = A - \beta\mathbf{v}\mathbf{w}_1^*$.
2. Instead of multiplication $AU_{\mathbf{v}}$, we use $AU_{\mathbf{v}} = A - \beta\mathbf{w}_2\mathbf{v}$.
3. Instead of multiplication $U_{\mathbf{v}}AU_{\mathbf{v}}$, we use $U_{\mathbf{v}}AU_{\mathbf{v}} = A - \beta\mathbf{v}\mathbf{w}_1^* + (\gamma\mathbf{v} - \beta\mathbf{w}_2)\mathbf{v}^*$.

The improvement comes from matrix-vector products and vector-vector products replacing the matrix-matrix multiplications.

Remarks: In Matlab, the command `H=hess(A)` will obtain the Hessenberg matrix H from A.

EXAMPLE C.3. Let

$$A = \begin{pmatrix} 3 & -4 & 0 & 12 & 12 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & -2 & 3 & 3 \\ 0 & 2 & 0 & -5 & -6 \\ 0 & -2 & 0 & 6 & 7 \end{pmatrix}$$

Using the algorithm, the resulting Hessenberg form is

$$G = HAH = \begin{pmatrix} 3 & -4 & 0 & 12 & 12 \\ 0 & 1 & -1.4142 & 1 & 1 \\ 0 & 2.8284 & 1 & -8.4853 & -8.4853 \\ 0 & 0 & 0 & 1.6213 & 3.6213 \\ 0 & 0 & 0 & -0.6213 & -2.6213 \end{pmatrix}$$

One can check that the eigenvalues of G and A will be the same.

Note that for this example, the resulting Hessenberg form is already in the desired block-triangular forms, even before applying the QR or shifted QR algorithms. In general, this will not be the case. Nonetheless, it does suggest that starting off with the upper Hessenberg forms will reduce the number of QR iterations needed for obtaining the eigenvalues of A.

C.2.7 Modified QR Method

We can now combine both enhancements to the QR approach to determine the eigenvalues of A.

Enhanced QR Algorithm for Evaluating Eigenvalues of A:

- **Initialize:**
  1. Set $k = N$.
  2. Specify tolerance ϵ.
  3. Obtain G, a matrix in upper Hessenberg form that is similar to A.
- **Reduce G:** While $k > 2$
  Case 1: $(|G_{k,k-1}| \leq \epsilon)$.
  1. Add $G_{k,k}$ to the list of eigenvalues.
  2. Update G by removing the last row and last column.

  Case 2: $(|G_{k,k-1}| > \epsilon)$ and $(|G_{k-1,k-2}| \leq \epsilon)$.
  1. Add μ_1 and μ_2 to the list of eigenvalues, where

$$\mu_1 = \frac{-b + \sqrt{b^2 - 4c}}{2} \quad ; \quad \mu_2 = \frac{-b - \sqrt{b^2 - 4c}}{2}$$

$$\text{and} \qquad b = -(G_{k-1,k-1} + G_{k,k})$$

$$c = G_{k-1,k-1}G_{k,k} - G_{k,k-1}G_{k-1,k}$$

  2. Update G by removing the last two rows and last two columns.

  Case 3: $(|G_{k,k-1}| > \epsilon)$ and $(|G_{k-1,k-2}| > \epsilon)$.
  Iterate until either Case 2 or Case 3 results:
  Let $\sigma = G_{k,k}$,
  1. Find Q and R such that: $QR = G - \sigma I$
  2. Update G: $G \leftarrow RQ + \sigma I$

  End While-loop
- **Termination:**
  Case 1: $G = [\lambda]$, then add λ to eigenvalue list.
  Case 2: $G[=]2 \times 2$, then add μ_1 and μ_2 to the list of eigenvalues, where

$$\mu_1 = \frac{-b + \sqrt{b^2 - 4c}}{2} \quad ; \quad \mu_2 = \frac{-b - \sqrt{b^2 - 4c}}{2}$$

  and

$$b = -(G_{11} + G_{22}) \quad ; \quad c = G_{11}G_{22} - G_{21}G_{12}$$

EXAMPLE C.4. Let

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

After applying Householder transformations H, we obtain $G = HAH$ that has the upper Hessenberg form

$$G = \begin{pmatrix} 1 & -2 & -0.1060 & -1.3072 & -0.5293 \\ -3 & 1.8889 & -1.1542 & 2.3544 & 1.7642 \\ 0 & -1.0482 & 0.8190 & 0.6139 & -0.4563 \\ 0 & 0 & -1.2738 & 0.0036 & 2.9704 \\ 0 & 0 & 0 & -0.8456 & -1.7115 \end{pmatrix}$$

After ten iterations of the shifted-QR method, G is updated to be

$$\begin{pmatrix} 4.2768 & 0.2485 & -2.2646 & 2.2331 & -5.7024 \\ 0 & -1.8547 & 2.3670 & -1.3323 & 0.2085 \\ 0 & -1.5436 & 0.4876 & 1.0912 & -0.0094 \\ 0 & 0 & -0.2087 & 0.3759 & -0.0265 \\ 0 & 0 & 0 & 0 & -1.2856 \end{pmatrix}$$

and we could extract -1.2856 as one of the eigenvalues. Then the size of G is reduced by deleting the last row and column, that is,

$$G \leftarrow \begin{pmatrix} 4.2768 & 0.2485 & -2.2646 & 2.2331 \\ 0 & -1.8547 & 2.3670 & -1.3323 \\ 0 & -1.5436 & 0.4876 & 1.0912 \\ 0 & 0 & -0.2087 & 0.3759 \end{pmatrix}$$

Note that along the process, even though G will be modified and shrunk, it will still have an upper Hessenberg form.

The process is repeated until all the eigenvalues of A are obtained: -1.2856, 0.0716, $-0.5314 \pm 1.5023i$, and 4.2768.

C.3 Calculations for the Jordan Decomposition

In this section, we develop an algorithm for the construction of a modal matrix T that would obtain the Jordan decomposition of a square matrix A. The canonical basis, that is, the columns of T, is composed of vectors derived from **eigenvector chains** of different orders.

Definition C.1. *Given matrix A and eigenvalues λ, then an* **eigenvector chain** *with respect to λ, of* **order** *r is*

$$\text{chain}(A, \lambda, r) = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r) \tag{C.8}$$

where

$$(A - \lambda I)^r \mathbf{v}_r = 0 \qquad (A - \lambda I)^{r-1} \mathbf{v}_r \neq 0$$

$$\mathbf{v}_j = (A - \lambda I)\mathbf{v}_{j+1} \qquad j = (r-1), \ldots, 1$$

Note: If the order of the chain is 1, then the chain is composed of only one eigenvector.

Algorithm for Obtaining Chain (A,λ,r).

1. **Obtain vector $\mathbf{v}_r$ to begin the chain.**
   (a) Construct matrix M,
   $$M(\lambda, r) = \left(\begin{array}{c|c} (A-\lambda I)^{r-1} & -I \\ \hline (A-\lambda I)^{r} & 0 \end{array}\right)$$
   (b) Use Gauss-Jordan elimination to obtain Q, W, and q such that
   $$QMW = \left(\begin{array}{c|c} I_{[q]} & 0 \\ \hline 0 & 0 \end{array}\right)$$
   (c) Construct vector $\mathbf{h}$
   $$h_j = \begin{cases} 0 & j = 1, 2, \ldots, q \\ \text{a randomly generated number} & j = q+1, \ldots, 2n \end{cases}$$
   (d) Obtain $\mathbf{v}_r$ by extracting the first N elements of $\mathbf{z} = W\mathbf{h}$.
2. **Calculate the rest of the chain.**
   $$\mathbf{v}_j = (A - \lambda I)\mathbf{v}_{j+1} \qquad j = (r-1), \ldots, 1$$

Note that as mentioned in Section B.2, the matrices Q and W can also be found based on the singular value decomposition. This means that with $U\Sigma V^* = M$, we can replace W above by V of the singular value decomposition. Furthermore, the rationale for introducing randomly generated numbers in the preceding algorithm is to find a vector that spans the last $(2n - q)$ columns of W without having to determine which vectors are independent.

EXAMPLE C.5. Let
$$A = \begin{pmatrix} 3 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 1 & 1 \\ 1 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$
Using the algorithm, we can find the chain of order 3 for $\lambda = 3$,
$$\text{chain}(A, 3, 3) = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{pmatrix} = \begin{pmatrix} 0 & -1.2992 & 0.8892 \\ 0 & -1.2992 & 1.1826 \\ -1.2992 & 0.8892 & 1.8175 \\ 0 & 0 & 0 \\ 0 & 0 & -1.2992 \end{pmatrix}$$
we can directly check that
$$(A - \lambda I)^3 \mathbf{v}_3 = \mathbf{0} \qquad (A - \lambda I)^2 \mathbf{v}_3 \neq \mathbf{0}$$
and
$$\mathbf{v}_2 = (A - \lambda I)\,\mathbf{v}_3 \qquad \mathbf{v}_1 = (A - \lambda I)\,\mathbf{v}_2$$

To obtain the canonical basis, we still need to determine the required eigenvector chains. To do so, we need to calculate the orders of matrix degeneracy with respect to an eigenvalue λ_i, to be denoted by $N_{i,k}$, which is just the difference in ranks of succeeding orders, that is,

$$N_{i,k} = \text{rank}(\text{A} - \lambda_\text{i}\text{I})^{\text{k}-1} - \text{rank}(\text{A} - \lambda_\text{i}\text{I})^{\text{k}} \tag{C.9}$$

Using these orders of degeneracy, one can calculate the required orders for the eigenvector chains. The algorithm that follows describes in more detail the procedure for obtaining the canonical basis.

Algorithm for Obtaining Canonical Basis.

Given $A[=]N \times N$.
For each distinct λ_i:

1. **Determine multiplicity m_i.**
2. **Calculate order of required eigenvector chains.**
   Let

$$p_i = \arg\left(\min_{1\le p\le n}\left[\text{rank}(A - \lambda_1 I)^p = (N - m_i)\right]\right)$$

   then obtain $\text{ord}_i = (\gamma_{i,1}, \ldots, \gamma_{i,p_i})$, where

$$\gamma_{i,k} = \begin{cases} N_{i,k} & \text{if } k = p_i \\ \max(0, [N_{i,k} - \sum_{j=k+1}^{p_i} \gamma_{i,j}]) & \text{if } k < p_i \end{cases}$$

   where,

$$N_{i,k} = \text{rank}(\text{A} - \lambda_\text{i}\text{I})^{\text{k}-1} - \text{rank}(\text{A} - \lambda_\text{i}\text{I})^{\text{k}}$$

3. **Obtain the required eigenvector chains.**
   For each $\gamma_{i,k} > 0$, find $\gamma_{i,k}$ sets of chain(A, λ_i, k) and add to the collection of canonical basis.

One can show that the eigenvector chains found will be linearly independent. This means that T is nonsingular. The Jordan canonical form can then be obtained by evaluating $T^{-1}AT = J$.

Although Jordan decomposition is not reliable for large systems, it remains very useful for generating theorems that are needed to handle both diagonalizable and non-diagonalizable matrices. For example, the proof of Cayley-Hamilton theorem uses Jordan block decompositions without necessarily having to evaluate the decompositions.

EXAMPLE C.6. Consider the matrix A,

$$A = \begin{pmatrix} 3 & 0 & 0 & 0 & 1 \\ 0 & 3 & 0 & 1 & 1 \\ 1 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

then

λ_i	m_i	p_i	$N_{i,k}$	ord_i
2	1	1	[1]	[1]
3	4	3	[2, 1, 1]	[1, 0, 1]

Next, calculating the required chains:

$$\text{chain}(A, 2, 1) = \begin{pmatrix} 0 \\ -0.707 \\ 0 \\ 0.707 \\ 0 \end{pmatrix} \qquad \text{chain}(A, 3, 1) = \begin{pmatrix} 0 \\ -0.5843 \\ -1.0107 \\ 0 \\ 0 \end{pmatrix}$$

$$\text{chain}(A, 3, 3) = \begin{pmatrix} 0 & -1.2992 & 0.8892 \\ 0 & -1.2992 & 1.1826 \\ -1.2992 & 0.8892 & 1.8175 \\ 0 & 0 & 0 \\ 0 & 0 & -1.2992 \end{pmatrix}$$

The modal matrix T can then be constructed as,

$$T = \begin{pmatrix} 0 & 0 & 0 & -1.2992 & 0.8892 \\ -0.7071 & -0.5843 & 0 & -1.2992 & 1.1826 \\ 0 & -1.0107 & -1.2992 & 0.8892 & 1.8175 \\ 0.7071 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.2992 \end{pmatrix}$$

The Jordan canonical form is

$$J = T^{-1}AT = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

C.4 Schur Triangularization and SVD

C.4.1 Schur Triangularization Algorithm

Given: $A[=]N \times N$
Initialization: Set $G_N = A$.
For $m = N, N-1, \ldots, 2$

Obtain λ, an eigenvalue of G_m, and its corresponding orthonormal eigenvector $\mathbf{v}$.

Using Gram-Schmidt algorithm (cf. Section 2.6), obtain an orthonormal set of $(m-1)$ vectors $\{\mathbf{w}_1, \ldots, \mathbf{w}_{m-1}\}$ that is also orthonormal to $\mathbf{v}$.

Let $H_m = \left(\begin{array}{c|c|c|c} \mathbf{v} & \mathbf{w}_1 & \cdots & \mathbf{w}_{m-1} \end{array} \right)$; then use

$$H_m^* G_m H_m = \left(\begin{array}{c|c} \lambda & \mathbf{b}^T \\ \hline 0 & G_{m-1} \end{array} \right)$$

to extract G_{m-1} and construct U_m as

$$U_m = \left(\begin{array}{c|c} I_{[N-m]} & 0 \\ \hline 0 & H_m \end{array} \right)$$

Calculate the product:

$$U = U_N U_{N-1} \cdots U_2$$

C.4.2 SVD Algorithm

1. Apply the QR algorithm on A^*A:
   (a) Initialize: $D = A^*A$, $V = I_{[M]}$, ϵ
   (b) Iterate: While $\left\| \text{vec}\left(D - \text{diag}(D)\right) \right\| > \epsilon$
      i. $D = QR$ via QR algorithm
      ii. $D \leftarrow RQ$
      iii. $V \leftarrow VQ$
   (Note: Re-index D and V such that $d_{k+1} > d_k$.)
2. Calculate singular values: $\sigma_i = \sqrt{d_{ii}}$, $i = 1, \ldots, M$.
3. Obtain U:
   Let r be the number of nonzero singular values.
   (a) Extract V_r as the first r columns of V.
   (b) Set $\Sigma_r = \text{diag}\,(\sigma_1, \ldots, \sigma_r)$.
   (c) Calculate: $U_r = AV_r\Sigma_r^{-1}$.
   (d) Find $U_q[=]N \times (M - r)$ such that U_q is orthogonal to U_r.
   (e) Set $U = \left(\begin{array}{c|c} U_r & U_q \end{array} \right)$.
4. Form $\Sigma[=]N \times M$: $\Sigma_{ij} = \begin{cases} \sigma_i & \text{if } i = j \leq r \\ 0 & \text{otherwise} \end{cases}$

C.5 Sylvester's Matrix Theorem

THEOREM C.1. *Let A have all distinct eigenvalues. Let $\mathbf{v}_k$ and $\mathbf{w}_k^*$ be the right and left eigenvectors A, respectively, corresponding to the same k^{th} eigenvalue λ_k, such that $\mathbf{w}_k^*\mathbf{v}_k = 1$. Then any well-defined matrix function $f(A)$ is given by*

$$f(A) = \sum_{k=1}^{N} f(\lambda_k)\, \mathbf{v}_k \mathbf{w}_k^* \tag{C.10}$$

The classic version of Sylvester's matrix theorem gives equivalent formulations of (C.10), two of which are the following:

$$f(A) = \sum_{k=1}^{N} f(\lambda_k) \frac{\prod_{\ell \neq k} (\lambda_\ell I - A)}{\prod_{\ell \neq k} (\lambda_\ell - \lambda_k)} \tag{C.11}$$

and

$$f(A) = \sum_{k=1}^{N} f(\lambda_k) \frac{\mathbf{adj}(\lambda_\ell I - A)}{\prod_{\ell \neq k}(\lambda_\ell - \lambda_k)} \tag{C.12}$$

The advantage of (C.11) is that it does not require the computation of eigenvectors. However, there are some disadvantages to both (C.11) and (C.12). One is that all the eigenvalues have to be distinct; otherwise, a problem arises in the denominator.

To show that (C.10) can be derived from (3.35), we need to first show that the rows of V^{-1} are left eigenvectors of A. Let $\mathbf{w}_k^*$ be the k^{th} row of V^{-1}, then

$$\begin{aligned} AV &= V\Lambda \\ V^{-1}A &= \Lambda V^{-1} \\ \left(\begin{array}{c} \mathbf{w}_1^* \\ \hline \vdots \\ \hline \mathbf{w}_N^* \end{array}\right) A &= \left(\begin{array}{ccc} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{array}\right) \left(\begin{array}{c} \mathbf{w}_1^* \\ \hline \vdots \\ \hline \mathbf{w}_N^* \end{array}\right) \end{aligned}$$

or

$$\mathbf{w}_k^* A = \lambda_k \mathbf{w}_N^*$$

Thus $\mathbf{w}_k^*$ is a left eigenvector of A. Using this partitioning of V^{-1}, (3.35) becomes

$$\begin{aligned} f(A) &= \left(\begin{array}{c|c|c} \mathbf{v}_1 & \cdots & \mathbf{v}_N \end{array}\right) \left(\begin{array}{ccc} f(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & f(\lambda_N) \end{array}\right) \left(\begin{array}{c} \mathbf{w}_1^* \\ \hline \vdots \\ \hline \mathbf{w}_N^* \end{array}\right) \\ &= \left(\begin{array}{c|c|c} \mathbf{v}_1 & \cdots & \mathbf{v}_N \end{array}\right) \left(\sum_{k=1}^{N} f(\lambda_k)\, \mathbf{e}_k \mathbf{e}_k^T\right) \left(\begin{array}{c} \mathbf{w}_1^* \\ \hline \vdots \\ \hline \mathbf{w}_N^* \end{array}\right) \\ &= f(\lambda_1)\, \mathbf{v}_1 \mathbf{w}_1^* + \ldots + f(\lambda_N)\, \mathbf{v}_n \mathbf{w}_N^* \end{aligned}$$

C.6 Danilevskii Method for Characteristic Polynomial

There are several methods for the evaluation of eigenvalues. For smaller matrices, the characteristic polynomials are first determined, and then the roots are then calculated to be the eigenvalues. For larger cases, other methods can be used that bypass the determination of the characteristic polynomial. Nonetheless, there are situations in which the determination of characteristic polynomials becomes the primary goal, such as problems in which the Cayley-Hamilton theorems are used.

One highly effective approach to finding the characteristic polynomial is the **Danilevskii method**. The main idea is to find sequences of elementary matrix operators (e.g., those used in Gaussian elimination) such that a nonsingular matrix S can be used to transform a square matrix A into a lower block triangular matrix in which the block diagonal matrices are in the form of **companion matrices**.

Definition C.2. *A square matrix C is said to be a* **companion matrix** *to a monic polynomial*

$$p(s) = s^n + \alpha_{n-1}s^{n-1} + \ldots + \alpha_1 s + \alpha_0$$

if it has the form

$$C = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \cdots & -\alpha_{n-1} \end{pmatrix} \tag{C.13}$$

It is left as an exercise (cf. **E3.8**) to show that the characteristic equation of C defined in (C.13) will be

$$p(s) = s^n + \alpha_{n-1}s^{n-1} + \ldots + \alpha_1 s + \alpha_0 = 0 \tag{C.14}$$

Furthermore, each distinct eigenvalue λ of C has a corresponding vector given by

$$\mathbf{v} = \begin{pmatrix} 1 \\ \lambda \\ \vdots \\ \lambda^{n-1} \end{pmatrix} \tag{C.15}$$

Thus with a similarity transformation of A based on a similarity transformation by S

$$S^{-1}AS = \left(\begin{array}{c|c|c|c} C_1 & & \cdots & \mathbf{0} \\ \hline Q_{21} & C_2 & & \\ \hline \vdots & \ddots & \ddots & \\ \hline Q_{r1} & Q_{r2} & \cdots & C_r \end{array}\right) \tag{C.16}$$

where C_i are $n_i \times n_i$ companion matrices to polynomials

$$p_i(s) = s^{n_i} + \alpha_{n_i-1}^{[i]} s^{n_i-1} + \cdots + \alpha_1^{[i]} s + \alpha_0^{[i]}$$

the characteristic polynomial of A is then given by

$$\text{charpoly}(A) = \prod_{i=1}^{r} p_i(s) \tag{C.17}$$

To find S, we have the following recursive algorithm:

Danilevski Algorithm:

Let $A[=]N \times N$; then **Danilevski**(A) should yield matrix S such that (C.16) is satisfied.

Initialize $k = 0$ and $S = I_N$

While $k < N$,

$k \leftarrow k + 1$

If $N = 1$,

$$S = 1$$

else

Let $j_{\max} = \arg\left(\max_{j\in\{i+1,\ldots,N\}} \left| a_{ij} \right|\right)$ and $q = a_{i,j_{\max}}$

If $q \neq 0$

Interchange rows $i+1$ and $j_{\max}$ of A
Interchange columns $i+1$ and $j_{\max}$ of A
Interchange columns $i+1$ and $j_{\max}$ of S

$$X = (x_i j) \quad ; \quad x_{ij} = \begin{cases} -a_{k,j}/a_{k,k+1} & \text{if } i = k+1,\ j \neq k+1 \\ -1/a_{k,k+1} & \text{if } i = k+1,\ j = k+1 \\ 1 & \text{if } i = j \neq k+1 \\ 0 & \text{otherwise} \end{cases}$$

$$Y = (y_i j) \quad ; \quad y_{ij} = \begin{cases} a_{k,j} & \text{if } i = k+1 \\ 1 & \text{if } i = j \neq k+1 \\ 0 & \text{otherwise} \end{cases}$$

$$A \leftarrow YAX$$
$$S \leftarrow SX$$

else

Extract the submatrix formed by rows and columns $i+1$ to N of A as H, then solve for $G = \text{Danilevkii}(H)$

$$S \leftarrow S\left(\begin{array}{c|c} I_i & 0 \\ \hline 0 & G \end{array}\right)$$

$$k \leftarrow N$$

end If
End while

The Danilevskii algorithm is known to be among one of the more precise methods for determination of characteristic polynomials and is relatively efficient compared with Leverier's approach, although the latter is still considered very accurate but slow.

A MATLAB function `charpoly` is available on the book's webpage for the evaluation of the characteristic polynomial via the Danilevskii method. The program obtains the matrix S such that $S^{-1}AS$ is in the form of a block triangular matrix given in (C.16). It also yields a set of polynomial coefficients $p_{n_k}^{[k]}$ saved in a cell array. Finally, the set of eigenvalues is also available by solving for the roots of the polynomials. A function `poly(A)` is also available in MATLAB, which is calculated in reverse; that is, the eigenvalues are obtained first, and then the characteristic polynomial is formed.

EXAMPLE C.7.

Given

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

then applying the Danilveskii method, we find

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -2.75 & -0.25 & 0.25 & 0 & 0 \\ 1.5 & 0.5 & -0.1667 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -0.5 & 0.5 \end{pmatrix}$$

$$S^{-1}AS = \left(\begin{array}{ccc|cc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -39 & 6 & 6 & 0 & 0 \\ \hline -0.75 & -0.25 & 0.25 & 0 & 1 \\ 5.75 & 1.25 & -.5833 & -1 & 2 \end{array}\right)$$

and the characteristic polynomial is given by

$$\begin{array}{lcl} p_1(s) & = & s^3 - 6s^2 - 6s + 39 \\ p_2(s) & = & s^2 - 2s + 1 \end{array} \rightarrow \begin{array}{lcl} p(s) & = & p_1(s)p_2(s) \\ & = & s^5 - 8s^4 + 7s^3 + 45s^2 - 84s + 39 \end{array}$$

APPENDIX D

Additional Details and Fortification for Chapter 4

D.1 Proofs of Identities of Differential Operators

The proofs for the identities of differential operations of orthogonal curvilinear coordinates are given as follows:

1. **Gradient** (4.90): apply (4.89) on ψ.
2. **Divergence** (4.91): Using (4.55),

$$\nabla \cdot (w_a \underline{\boldsymbol{\delta}}_a) = \underline{\boldsymbol{\delta}}_a \cdot \nabla w_a + w_a \nabla \cdot \underline{\boldsymbol{\delta}}_a \tag{D.1}$$

The first term in (D.1) can be expanded using (4.90)as follows:

$$\underline{\boldsymbol{\delta}}_a \cdot \nabla w_a = \underline{\boldsymbol{\delta}}_a \cdot \left(\frac{1}{\alpha_a} \underline{\boldsymbol{\delta}}_a \frac{\partial w_a}{\partial a} + \frac{1}{\alpha_b} \underline{\boldsymbol{\delta}}_b \frac{\partial w_a}{\partial b} + \frac{1}{\alpha_c} \underline{\boldsymbol{\delta}}_c \frac{\partial w_a}{\partial c} \right) = \frac{1}{\alpha_a} \frac{\partial w_a}{\partial a} \tag{D.2}$$

From (4.87) and (4.88),

$$\underline{\boldsymbol{\delta}}_a = \underline{\boldsymbol{\delta}}_b \times \underline{\boldsymbol{\delta}}_c = (\alpha_b \underline{\widehat{\nu}}_b) \times (\alpha_c \underline{\widehat{\nu}}_c) = \alpha_b \alpha_c \nabla b \times \nabla c$$

Then the second term in (D.1) becomes,

$$\begin{aligned} w_a \nabla \cdot \underline{\boldsymbol{\delta}}_a &= w_a \nabla \cdot (\alpha_b \alpha_c \nabla b \times \nabla c) \\ &= w_a (\alpha_b \alpha_c \nabla \cdot (\nabla b \times \nabla c)) + w_a (\nabla b \times \nabla c) \cdot \nabla (\alpha_b \alpha_c) \\ &= w_a \frac{1}{\alpha_b \alpha_c} \underline{\boldsymbol{\delta}}_a \cdot \left(\frac{1}{\alpha_a} \underline{\boldsymbol{\delta}}_a \frac{\partial (\alpha_b \alpha_c)}{\partial a} + \frac{1}{\alpha_b} \underline{\boldsymbol{\delta}}_b \frac{\partial (\alpha_b \alpha_c)}{\partial b} + \frac{1}{\alpha_c} \underline{\boldsymbol{\delta}}_c \frac{\partial (\alpha_b \alpha_c)}{\partial c} \right) \\ &= w_a \frac{1}{\alpha_a \alpha_b \alpha_c} \frac{\partial (\alpha_b \alpha_c)}{\partial a} \end{aligned} \tag{D.3}$$

where we used the fact that $\nabla \cdot (\nabla b \times \nabla c) = 0$ (see Exercise **E4.17**).

Substituting (D.2) and (D.3) into (D.1),

$$\nabla \cdot (w_a \underline{\boldsymbol{\delta}}_a) = \frac{1}{\alpha_a} \frac{\partial w_a}{\partial a} + w_a \frac{1}{\alpha_a \alpha_b \alpha_c} \frac{\partial (\alpha_b \alpha_c)}{\partial a} = \frac{1}{\alpha_a \alpha_b \alpha_c} \frac{\partial (w_a \alpha_b \alpha_c)}{\partial a}$$

Similarly, we can obtain

$$\nabla \cdot (w_b \underline{\boldsymbol{\delta}}_b) = \frac{1}{\alpha_a \alpha_b \alpha_c} \frac{\partial (\alpha_a w_b \alpha_c)}{\partial b} \quad ; \quad \nabla \cdot (w_c \underline{\boldsymbol{\delta}}_c) = \frac{1}{\alpha_a \alpha_b \alpha_c} \frac{\partial (\alpha_a \alpha_b w_c)}{\partial c}$$

Combining,

$$\nabla \cdot \underline{\mathbf{w}} = \frac{1}{\alpha_a \alpha_b \alpha_c} \left(\frac{\partial(w_a \alpha_b \alpha_c)}{\partial a} + \frac{\partial(\alpha_a w_b \alpha_c)}{\partial b} + \frac{\partial(\alpha_a \alpha_b w_c)}{\partial c} \right)$$

3. **Curl** (4.92): Using (4.56) and (4.61), the curl of $w_a \underline{\boldsymbol{\delta}}_a$ can be expanded as follows:

$$\begin{aligned}
\nabla \times (w_a \underline{\boldsymbol{\delta}}_a) &= \nabla \times (w_a \alpha_a \nabla a) \\
&= w_a \alpha_a \underbrace{(\nabla \times \nabla a)}_{=0} + \nabla (w_a \alpha_a) \times \nabla a \\
&= \left(\frac{1}{\alpha_a} \underline{\boldsymbol{\delta}}_a \frac{\partial(w_a \alpha_a)}{\partial a} + \frac{1}{\alpha_b} \underline{\boldsymbol{\delta}}_b \frac{\partial(w_a \alpha_a)}{\partial b} + \frac{1}{\alpha_c} \underline{\boldsymbol{\delta}}_c \frac{\partial(w_a \alpha_a)}{\partial c} \right) \left(\frac{1}{\alpha_a} \underline{\boldsymbol{\delta}}_a \right) \\
&= -\frac{1}{\alpha_a \alpha_b} \underline{\boldsymbol{\delta}}_c \frac{\partial(w_a \alpha_a)}{\partial b} + \frac{1}{\alpha_a \alpha_c} \underline{\boldsymbol{\delta}}_b \frac{\partial(w_a \alpha_a)}{\partial c}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\nabla \times (w_b \underline{\boldsymbol{\delta}}_b) &= \frac{1}{\alpha_b \alpha_a} \underline{\boldsymbol{\delta}}_c \frac{\partial(w_b \alpha_b)}{\partial a} - \frac{1}{\alpha_b \alpha_c} \underline{\boldsymbol{\delta}}_a \frac{\partial(w_b \alpha_b)}{\partial c} \\
\nabla \times (w_c \underline{\boldsymbol{\delta}}_c) &= \frac{1}{\alpha_c \alpha_b} \underline{\boldsymbol{\delta}}_a \frac{\partial(w_c \alpha_c)}{\partial b} - \frac{1}{\alpha_c \alpha_a} \underline{\boldsymbol{\delta}}_b \frac{\partial(w_c \alpha_c)}{\partial a}
\end{aligned}$$

Combining all three curls,

$$\begin{aligned}
\nabla \times \underline{\mathbf{w}} = \frac{1}{\alpha_a \alpha_b \alpha_c} \Bigg[& \alpha_a \underline{\boldsymbol{\delta}}_a \left(\frac{\partial(\alpha_c w_c)}{\partial b} - \frac{\partial(\alpha_b w_b)}{\partial c} \right) \\
& + \alpha_b \underline{\boldsymbol{\delta}}_b \left(\frac{\partial(\alpha_a w_a)}{\partial c} - \frac{\partial(\alpha_c w_c)}{\partial a} \right) \\
& + \alpha_c \underline{\boldsymbol{\delta}}_c \left(\frac{\partial(\alpha_b w_b)}{\partial a} - \frac{\partial(\alpha_a w_a)}{\partial b} \right) \Bigg]
\end{aligned}$$

4. **Laplacian of scalar fields** (4.93): Substituting

$$\underline{\mathbf{w}} = \nabla \psi = \frac{1}{\alpha_a} \underline{\boldsymbol{\delta}}_a \frac{\partial \psi}{\partial a} + \frac{1}{\alpha_b} \underline{\boldsymbol{\delta}}_b \frac{\partial \psi}{\partial b} + \frac{1}{\alpha_c} \underline{\boldsymbol{\delta}}_c \frac{\partial \psi}{\partial c}$$

into (4.91),

$$\nabla \cdot \nabla \psi = \frac{1}{\alpha_a \alpha_b \alpha_c} \left(\frac{\partial}{\partial a} \left[\left(\frac{\alpha_b \alpha_c}{\alpha_a} \right) \frac{\partial \psi}{\partial a} \right] + \frac{\partial}{\partial b} \left[\left(\frac{\alpha_a \alpha_c}{\alpha_b} \right) \frac{\partial \psi}{\partial b} \right] + \frac{\partial}{\partial c} \left[\left(\frac{\alpha_a \alpha_b}{\alpha_c} \right) \frac{\partial \psi}{\partial c} \right] \right)$$

5. **Gradient-Vector Dyad** (4.94):

$$\begin{aligned}\nabla\underline{\mathbf{w}} &= \nabla\left(\sum_{k=a,b,c} w_k\underline{\boldsymbol{\delta}}_k\right)\\ &= \sum_{k=a,b,c}\left((\nabla w_k)\,\underline{\boldsymbol{\delta}}_k + w_k\nabla\underline{\boldsymbol{\delta}}_k\right)\\ &= \sum_{k=a,b,c}\sum_{m=a,b,c}\frac{1}{\alpha_m}\frac{\partial w_k}{\partial m}\underline{\boldsymbol{\delta}}_m\underline{\boldsymbol{\delta}}_k + \sum_{k=a,b,c}\sum_{m=a,b,c}\frac{w_k}{\alpha_m}\underline{\boldsymbol{\delta}}_m\frac{\partial\underline{\boldsymbol{\delta}}_k}{\partial m}\end{aligned}$$

D.2 Derivation of Formulas in Cylindrical Coordinates

At a point (r, θ, z), the pair of unit vectors $\underline{\delta}_r$ and $\underline{\boldsymbol{\delta}}_\theta$ is just the pair of unit vectors $\underline{\boldsymbol{\delta}}_x$ and $\underline{\boldsymbol{\delta}}_y$ rotated counter-clockwise by an angle θ, which could be achieved using a rotation operator,[1]

$$R_{r\to c} = \begin{pmatrix}\cos\theta & \sin\theta & 0\\ -\sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{pmatrix} \tag{D.4}$$

Because $R_{r\to c}$ is an orthogonal matrix,

$$\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} = R_{r\to c}\begin{pmatrix}\underline{\boldsymbol{\delta}}_x\\ \underline{\boldsymbol{\delta}}_y\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} \Longleftrightarrow \begin{pmatrix}\underline{\boldsymbol{\delta}}_x\\ \underline{\boldsymbol{\delta}}_y\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} = R_{r\to c}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} \tag{D.5}$$

which is relationship 1 in Table 4.6. We can then apply (D.5) for vector $\underline{\mathbf{v}}$,

$$\begin{pmatrix}v_r & v_\theta & v_z\end{pmatrix}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} = \underline{\mathbf{v}} = \begin{pmatrix}v_x & v_y & v_z\end{pmatrix}\begin{pmatrix}\underline{\boldsymbol{\delta}}_x\\ \underline{\boldsymbol{\delta}}_y\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} = \begin{pmatrix}v_x & v_y & v_z\end{pmatrix}R_{r\to c}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix}$$

Comparing both ends of the equations, we have

$$\begin{pmatrix}v_r\\ v_\theta\\ v_z\end{pmatrix} = R_{r\to c}\begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix} \Longleftrightarrow \begin{pmatrix}v_x\\ v_y\\ v_z\end{pmatrix} = R_{r\to c}^T\begin{pmatrix}v_r\\ v_\theta\\ v_z\end{pmatrix} \tag{D.6}$$

which is relationship 2 in Table 4.6.

For the relationship between the partial differential operators of the rectangular and the cylindrical coordinate system, the chain rule has to be applied. This yields,

$$\begin{pmatrix}\dfrac{\partial}{\partial r}\\ \dfrac{\partial}{\partial\theta}\\ \dfrac{\partial}{\partial z}\end{pmatrix} = \begin{pmatrix}\dfrac{\partial x}{\partial r} & \dfrac{\partial y}{\partial r} & \dfrac{\partial z}{\partial r}\\ \dfrac{\partial x}{\partial\theta} & \dfrac{\partial y}{\partial\theta} & \dfrac{\partial z}{\partial\theta}\\ \dfrac{\partial x}{\partial z} & \dfrac{\partial y}{\partial z} & \dfrac{\partial z}{\partial z}\end{pmatrix}\begin{pmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{pmatrix} = \begin{pmatrix}\cos\theta & \sin\theta & 0\\ -r\sin\theta & r\cos\theta & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\dfrac{\partial}{\partial x}\\ \dfrac{\partial}{\partial y}\\ \dfrac{\partial}{\partial z}\end{pmatrix}$$

[1] Note that the operator in (D.4) will rotate an input vector clockwise by an angle θ. However, because we are rotating the reference axes, the operator would do the reverse; that is, it rotates the axes counterclockwise.

Let $D_{r\to c} = \text{diag}\left(1, r, 1\right)$. Then,

$$\begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \\ \frac{\partial}{\partial z} \end{pmatrix} = D_{r\to c} R_{r\to c} \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \quad \Longleftrightarrow \quad \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} = R_{r\to c}^T D_{r\to c}^{-1} \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \\ \frac{\partial}{\partial z} \end{pmatrix} \tag{D.7}$$

which is relationship 3 in Table 4.6.

To obtain the relationship of the gradient operator ∇ between the rectangular and the cylindrical coordinates, we can apply both (D.5) and (D.7),

$$\begin{aligned} \nabla &= \begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \underline{\boldsymbol{\delta}}_y & \underline{\boldsymbol{\delta}}_z \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \underline{\boldsymbol{\delta}}_r & \underline{\boldsymbol{\delta}}_\theta & \underline{\boldsymbol{\delta}}_z \end{pmatrix} R_{r\to c} \end{pmatrix} \begin{pmatrix} R_{r\to c}^T D_{r\to c}^{-1} \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \\ \frac{\partial}{\partial z} \end{pmatrix} \end{pmatrix} \\ &= \begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \frac{1}{r}\underline{\boldsymbol{\delta}}_y & \underline{\boldsymbol{\delta}}_z \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial r} \\ \frac{\partial}{\partial \theta} \\ \frac{\partial}{\partial z} \end{pmatrix} \end{aligned} \tag{D.8}$$

which is relationship 4 in Table 4.6.

To obtain the partial derivatives of unit vectors in the cylindrical coordinate systems, note that:

1. The direction and magnitude of $\underline{\boldsymbol{\delta}}_r$, $\underline{\boldsymbol{\delta}}_\theta$, and $\underline{\boldsymbol{\delta}}_z$ will not change if we just modify the r position. Thus
$$\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial r} = \frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial r} = \frac{\partial \underline{\boldsymbol{\delta}}_z}{\partial r} = 0$$
2. Likewise, the direction and magnitude of $\underline{\boldsymbol{\delta}}_r$, $\underline{\boldsymbol{\delta}}_\theta$, and $\underline{\boldsymbol{\delta}}_z$ will not change if we just modify the z position. Thus
$$\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial z} = \frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial z} = \frac{\partial \underline{\boldsymbol{\delta}}_z}{\partial z} = 0$$
3. If we just change the θ position, the direction or magnitude of $\underline{\boldsymbol{\delta}}_z$ will also not change. Thus
$$\frac{\partial \underline{\boldsymbol{\delta}}_z}{\partial \theta} = 0$$

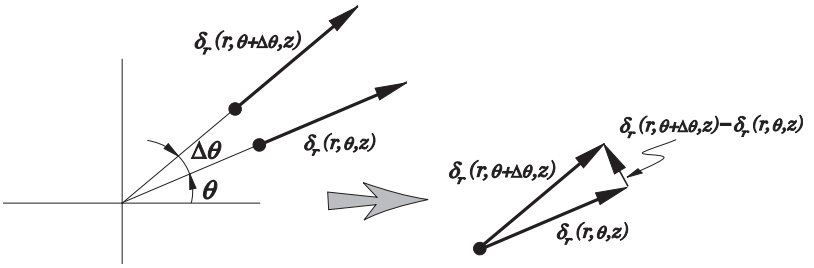


Figure D.1. Unit vectors along r at different θ positions.

What remains is the behavior of $\underline{\boldsymbol{\delta}}_r$ and $\underline{\boldsymbol{\delta}}_\theta$ as we change the θ position. In both cases, the directions do change. Let us first look at how $\underline{\boldsymbol{\delta}}_r$ changes with θ. The partial derivative of $\underline{\boldsymbol{\delta}}_r$ with respect to θ is given by

$$\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial \theta} = \lim_{\Delta\theta \to 0} \frac{\underline{\boldsymbol{\delta}}_r\,(r, \theta + \Delta\theta, z) - \underline{\boldsymbol{\delta}}_r\,(r, \theta, z)}{\Delta\theta}$$

where the subtraction is a vector subtraction. This is shown in (the right side of) Figure D.1. As $\Delta\theta \to 0$, we can see that the vector difference will be pointing perpendicular to $\underline{\boldsymbol{\delta}}_r\,(r, \theta, z)$. Thus

$$\text{direction}\left(\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial \theta}\right) = \text{direction}\,(\underline{\boldsymbol{\delta}}_\theta)$$

For the magnitude,

$$\left|\lim_{\Delta\theta \to 0} \frac{\underline{\delta}_r(r, \theta + \Delta\theta) - \underline{\delta}_r(r, \theta)}{\Delta\theta}\right| = \lim_{\Delta\theta \to 0} \frac{2|\underline{\delta}_r| \sin \Delta\theta/2}{\Delta\theta} = 1$$

Because the direction and magnitude matches $\underline{\boldsymbol{\delta}}_\theta$,

$$\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial \theta} = \underline{\boldsymbol{\delta}}_\theta \tag{D.9}$$

Using a similar argument for $\underline{\boldsymbol{\delta}}_\theta$,

$$\frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial \theta} = \lim_{\Delta\theta \to 0} \frac{\underline{\boldsymbol{\delta}}_\theta\,(r, \theta + \Delta\theta, z) - \underline{\boldsymbol{\delta}}_\theta\,(r, \theta, z)}{\Delta\theta}$$

The vector subtraction is shown in Figure D.2, where the limit yields a vector that is pointing in opposite direction of $\underline{\delta}_r$. The magnitude of the limit is also 1. Thus

$$\frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial \theta} = -\underline{\boldsymbol{\delta}}_r \tag{D.10}$$

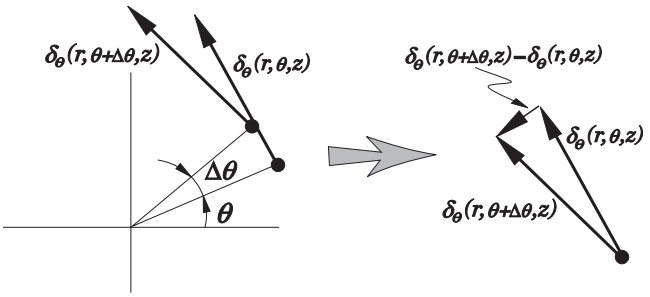


Figure D.2. Unit vectors along θ at different θ positions.

Alternatively, to find the derivatives of the unit vectors of cylindrical coordinates, we could use the fact that $\underline{\boldsymbol{\delta}}_x$, $\underline{\boldsymbol{\delta}}_y$, and $\underline{\boldsymbol{\delta}}_z$ have fixed magnitudes and direction. Then using (D.4) and (D.5),

$$\begin{aligned}
\frac{\partial}{\partial r}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} &= \left(\frac{\partial}{\partial r}R_{r\to c}\right)R_{r\to c}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix}\\
\frac{\partial}{\partial \theta}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} &= \left(\frac{\partial}{\partial \theta}R_{r\to c}\right)R_{r\to c}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix}\\
&= \begin{pmatrix}-\sin\theta & \cos\theta & 0\\ -\cos\theta & -\sin\theta & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\cos\theta & -\sin\theta & 0\\ \sin\theta & \cos\theta & 0\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix}\\
&= \begin{pmatrix}\underline{\boldsymbol{\delta}}_\theta\\ -\underline{\boldsymbol{\delta}}_r\\ 0\end{pmatrix}\\
\frac{\partial}{\partial z}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} &= \left(\frac{\partial}{\partial z}R_{r\to c}\right)R_{r\to c}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix}
\end{aligned}$$

D.3 Derivation of Formulas in Spherical Coordinates

To transform the unit vectors in rectangular coordinates to spherical coordinates at a point $(x, y, z) \leftrightarrow (r, \theta, \phi)$, we need the following sequence of operations:

1. A rotation of ϕ radians counterclockwise along the $\left(\underline{\boldsymbol{\delta}}_x, \underline{\boldsymbol{\delta}}_y\right)$ plane using the rotation operator R_{rs1}.
2. A rotation of θ radians clockwise along the $\left(\underline{\boldsymbol{\delta}}_x, \underline{\boldsymbol{\delta}}_z\right)$ plane using the rotation operator R_{rs2}.
3. A reordering of the unit vectors using the permutation operator E_{rs}.

where,

$$R_{rs1} = \begin{pmatrix}\cos\phi & \sin\phi & 0\\ -\sin\phi & \cos\phi & 0\\ 0 & 0 & 1\end{pmatrix} \quad R_{rs2} = \begin{pmatrix}\cos\theta & 0 & -\sin\theta\\ 0 & 1 & 0\\ \sin\theta & 0 & \cos\phi\end{pmatrix} \quad E_{rs} = \begin{pmatrix}0 & 0 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{pmatrix}$$

Combining all three orthogonal operators in the prescribed sequence will yield an orthogonal operator used to transform $\left(\underline{\boldsymbol{\delta}}_x, \underline{\boldsymbol{\delta}}_y, \underline{\boldsymbol{\delta}}_z\right)$ to $\left(\underline{\boldsymbol{\delta}}_r, \underline{\boldsymbol{\delta}}_\theta, \underline{\boldsymbol{\delta}}_\phi\right)$:

$$R_{r\to s} = E_{rs}R_{rs2}R_{rs1} = \begin{pmatrix}\sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta\\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta\\ -\sin\phi & \cos\phi & 0\end{pmatrix} \tag{D.11}$$

Then, following the same approach used during transformations between rectangular and cylindrical coordinates, we have

$$\begin{pmatrix} \underline{\boldsymbol{\delta}}_r \\ \underline{\boldsymbol{\delta}}_\theta \\ \underline{\boldsymbol{\delta}}_\phi \end{pmatrix} = R_{r\to s} \begin{pmatrix} \underline{\boldsymbol{\delta}}_x \\ \underline{\boldsymbol{\delta}}_y \\ \underline{\boldsymbol{\delta}}_z \end{pmatrix} \quad \Longleftrightarrow \quad \begin{pmatrix} \underline{\boldsymbol{\delta}}_x \\ \underline{\boldsymbol{\delta}}_y \\ \underline{\boldsymbol{\delta}}_z \end{pmatrix} = R_{r\to s}^T \begin{pmatrix} \underline{\boldsymbol{\delta}}_r \\ \underline{\boldsymbol{\delta}}_\theta \\ \underline{\boldsymbol{\delta}}_\phi \end{pmatrix} \tag{D.12}$$

$$\begin{pmatrix} v_r \\ v_\theta \\ v_\phi \end{pmatrix} = R_{r\to s} \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \quad \Longleftrightarrow \quad \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} = R_{r\to s}^T \begin{pmatrix} v_r \\ v_\theta \\ v_\phi \end{pmatrix} \tag{D.13}$$

The partial differential operators between the rectangular and spherical coordinate system are obtained by using the chain rule,

$$\begin{pmatrix} \dfrac{\partial}{\partial r} \\ \dfrac{\partial}{\partial \theta} \\ \dfrac{\partial}{\partial \phi} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial y}{\partial r} & \dfrac{\partial z}{\partial r} \\ \dfrac{\partial x}{\partial \theta} & \dfrac{\partial y}{\partial \theta} & \dfrac{\partial z}{\partial \theta} \\ \dfrac{\partial x}{\partial \phi} & \dfrac{\partial y}{\partial \phi} & \dfrac{\partial z}{\partial \phi} \end{pmatrix} \begin{pmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{pmatrix} = \begin{pmatrix} s_\theta c_\phi & s_\theta s_\phi & c_\theta \\ r c_\theta c_\phi & r c_\theta s_\phi & -r s_\theta \\ -r s_\theta s_\phi & r s_\theta c_\phi & 0 \end{pmatrix} \begin{pmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{pmatrix}$$

Let $D_{r\to s} = \text{diag}\left(1, r, r\sin\theta\right)$. Then,

$$\begin{pmatrix} \dfrac{\partial}{\partial r} \\ \dfrac{\partial}{\partial \theta} \\ \dfrac{\partial}{\partial \phi} \end{pmatrix} = D_{r\to s} R_{r\to s} \begin{pmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{pmatrix} \Longleftrightarrow \begin{pmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{pmatrix} = R_{r\to s}^T D_{r\to s}^{-1} \begin{pmatrix} \dfrac{\partial}{\partial r} \\ \dfrac{\partial}{\partial \theta} \\ \dfrac{\partial}{\partial \phi} \end{pmatrix} \tag{D.14}$$

To obtain the relationship of the gradient operator ∇ between the rectangular and the spherical coordinates, we can apply both (D.12) and (D.14),

$$\begin{aligned} \nabla &= \begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \underline{\boldsymbol{\delta}}_y & \underline{\boldsymbol{\delta}}_z \end{pmatrix} \begin{pmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial z} \end{pmatrix} = \begin{pmatrix} \underline{\boldsymbol{\delta}}_r & \underline{\boldsymbol{\delta}}_\theta & \underline{\boldsymbol{\delta}}_\phi \end{pmatrix} R_{r\to s} \left(R_{r\to s}^T D_{r\to s}^{-1} \begin{pmatrix} \dfrac{\partial}{\partial r} \\ \dfrac{\partial}{\partial \theta} \\ \dfrac{\partial}{\partial \phi} \end{pmatrix} \right) \\ &= \begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \dfrac{1}{r}\underline{\boldsymbol{\delta}}_y & \dfrac{1}{r\sin\theta}\underline{\boldsymbol{\delta}}_z \end{pmatrix} \begin{pmatrix} \dfrac{\partial}{\partial r} \\ \dfrac{\partial}{\partial \theta} \\ \dfrac{\partial}{\partial \phi} \end{pmatrix} \end{aligned} \tag{D.15}$$

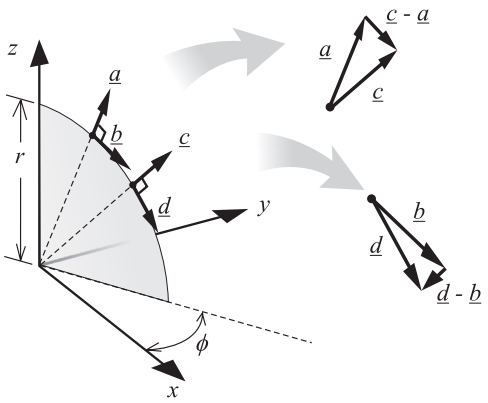


Figure D.3. Unit vectors at fixed r and ϕ. The unit vectors are represented by: $\underline{\mathbf{a}} = \underline{\boldsymbol{\delta}}_r(r, \theta, \phi)$, $\underline{\mathbf{b}} = \underline{\boldsymbol{\delta}}_\theta(r, \theta, \phi)$, $\underline{\mathbf{c}} = \underline{\boldsymbol{\delta}}_r(r, \theta + \Delta\theta, \phi)$, $\underline{\mathbf{d}} = \underline{\boldsymbol{\delta}}_\theta(r, \theta + \Delta\theta, \phi)$.

To obtain the partial derivatives of unit vectors in the spherical coordinate systems, note that:

1. The direction and magnitude of $\underline{\boldsymbol{\delta}}_r$, $\underline{\boldsymbol{\delta}}_\theta$, and $\underline{\boldsymbol{\delta}}_\phi$ will not change if we just modify the r position. Thus

$$\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial r} = \frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial r} = \frac{\partial \underline{\boldsymbol{\delta}}_\phi}{\partial r} = 0$$

2. The direction and magnitude of $\underline{\boldsymbol{\delta}}_\phi$ will not change if we just modify the θ position. Thus

$$\frac{\partial \underline{\boldsymbol{\delta}}_\phi}{\partial \theta} = 0$$

The remaining partial derivatives of unit vectors will change their direction based on their position in space. For a fixed r and ϕ, the vector subtractions are shown in Figure D.3, and the partial derivatives are then given by

$$\frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial \theta} = \underline{\boldsymbol{\delta}}_\theta \qquad \frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial \theta} = -\underline{\boldsymbol{\delta}}_r \tag{D.16}$$

For a fixed r and θ, the vector subtractions are shown in Figure D.4. Note that four of the unit vectors are first projected into the horizontal plane prior to taking limits. The partial derivatives are then given by:

$$\frac{\partial \underline{\boldsymbol{\delta}}_\phi}{\partial \phi} = -\cos\theta \underline{\boldsymbol{\delta}}_\theta - \sin\theta \underline{\boldsymbol{\delta}}_r \;;\; \frac{\partial \underline{\boldsymbol{\delta}}_r}{\partial \phi} = \sin\theta \underline{\boldsymbol{\delta}}_\phi \;;\; \frac{\partial \underline{\boldsymbol{\delta}}_\theta}{\partial \phi} = \cos\theta \underline{\boldsymbol{\delta}}_\phi \tag{D.17}$$

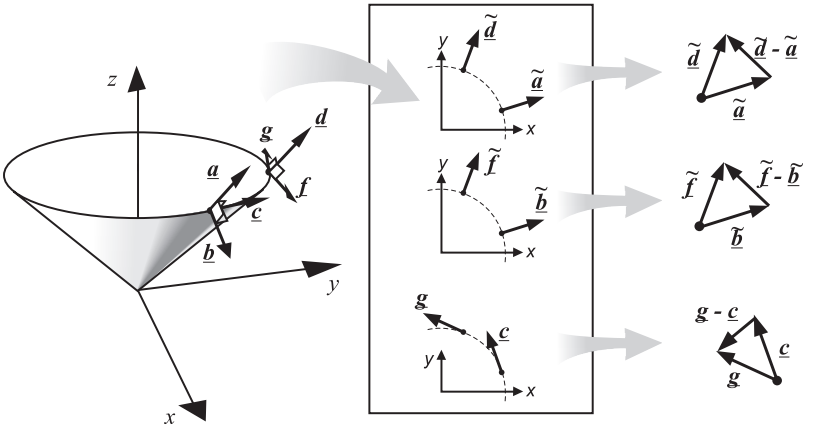


Figure D.4. Unit vectors at fixed r and θ. The unit vectors are represented by: $\underline{\mathbf{a}} = \underline{\boldsymbol{\delta}}_r(r,\theta,\phi)$, $\underline{\mathbf{b}} = \underline{\boldsymbol{\delta}}_\theta(r,\theta,\phi)$, $\underline{\mathbf{c}} = \underline{\boldsymbol{\delta}}_\phi(r,\phi)$, $\underline{\mathbf{d}} = \underline{\boldsymbol{\delta}}_r(r,\theta,\phi+\Delta\phi)$, $\underline{\mathbf{f}} = \underline{\boldsymbol{\delta}}_\theta(r,\theta,\phi+\Delta\phi)$, $\underline{\mathbf{g}} = \underline{\boldsymbol{\delta}}_\phi(r,\theta,\phi+\Delta\phi)$. The unit vectors projected into the horizontal planes are: $\widetilde{\underline{\mathbf{a}}} = \underline{\boldsymbol{\delta}}_r(r,\theta,\phi)\sin\theta$, $\widetilde{\underline{\mathbf{b}}} = \underline{\boldsymbol{\delta}}_\theta(r,\theta,\phi)\cos\theta$, $\widetilde{\underline{\mathbf{d}}} = \underline{\boldsymbol{\delta}}_r(r,\theta,\phi+\Delta\phi)\sin\theta$, $\widetilde{\underline{\mathbf{f}}} = \underline{\boldsymbol{\delta}}_\theta(r,\theta,\phi+\Delta\phi)\cos\theta$.

Alternatively, to find the derivatives of the unit vectors of spherical coordinates, we could use the fact that $\underline{\boldsymbol{\delta}}_x$, $\underline{\boldsymbol{\delta}}_y$, and $\underline{\boldsymbol{\delta}}_z$ have fixed magnitudes and direction. Then using (D.11) and (D.12),

$$\frac{\partial}{\partial r}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix} = \left(\frac{\partial}{\partial r}R_{r\to s}\right)R_{r\to s}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix}$$

$$\begin{aligned}\frac{\partial}{\partial\theta}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix} &= \left(\frac{\partial}{\partial\theta}R_{r\to s}\right)R_{r\to s}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix}\\ &= \begin{pmatrix}c_\theta c_\phi & c_\theta s_\phi & -s_\theta\\ -s_\theta c_\phi & -s_\theta s_\phi & -c_\theta\\ 0 & 0 & 0\end{pmatrix}\begin{pmatrix}s_\theta c_\phi & c_\theta c_\phi & -s_\phi\\ s_\theta s_\phi & c_\theta s_\phi & c_\phi\\ c_\theta & -s_\theta & 0\end{pmatrix}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix}\\ &= \begin{pmatrix}\underline{\boldsymbol{\delta}}_\theta\\ -\underline{\boldsymbol{\delta}}_r\\ 0\end{pmatrix}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\phi}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix} &= \left(\frac{\partial}{\partial r}R_{r\to s}\right)R_{r\to s}^T\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix}\\ &= \begin{pmatrix}-s_\theta s_\phi & s_\theta c_\phi & 0\\ -c_\theta s_\phi & c_\theta c_\phi & 0\\ -c_\phi & -s_\phi & 0\end{pmatrix}\begin{pmatrix}s_\theta c_\phi & c_\theta c_\phi & -s_\phi\\ s_\theta s_\phi & c_\theta s_\phi & c_\phi\\ c_\theta & -s_\theta & 0\end{pmatrix}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_z\end{pmatrix}\\ &= \begin{pmatrix}0 & 0 & s_\theta\\ 0 & 0 & c_\theta\\ -s_\theta & -c_\theta & 0\end{pmatrix}\begin{pmatrix}\underline{\boldsymbol{\delta}}_r\\ \underline{\boldsymbol{\delta}}_\theta\\ \underline{\boldsymbol{\delta}}_\phi\end{pmatrix}\end{aligned}$$

APPENDIX E

Additional Details and Fortification for Chapter 5

E.1 Line Integrals

Line integrals are generalizations of the ordinary integrals of single-variable functions to handle cases in which variations occur along specified curves in two or three dimensions. The line integrals therefore consists of three components: the **path of integration** $C(x, y, z)$, which is a continuous curve; the **integrand** $F(x, y, z)$, which is a scalar function; and the differential $d\lambda$.

Definition E.1. *A* **line integral** *of* $F(x, y, z)$*, with respect to variable* λ *and path* $C(x, y, z)$*, is defined by*

$$\int_C F(x, y, z)\, d\lambda = \lim_{\Delta\lambda_i \to 0, N \to \infty} \sum_{i=0}^{N} F(x_i, y_i, z_i)\, \Delta\lambda_i \tag{E.1}$$

In most applications, the differential $d\lambda$ is set to either dx, dy, dz or ds, where

$$ds = \sqrt{dx^2 + dy^2 + dz^2} \tag{E.2}$$

For the 2D case, $F = F(x, y)$ and the path $C = C(x, y)$. Figure E.1 gives the area interpretation of the line integrals. The integral $\int_C F(x, y)ds$ is the area under the curve $F(x, y)$ as the point travels along curve C. Conversely, the line integral with respect to x, $\int_C F(x, y)dx$ is the area projected onto the plane $y = 0$. The projected integral $\int_C Fdx$ is with respect to segments where $C(x, y)$ has to be single-valued with respect to x. Otherwise, the integration path will have to be partitioned into segments such that it is single-valued with respect to x. For example, the integration path from A to B in Figure E.2 will have to be partitioned into segment ADE, segment EF, and segment FGB. Thus for the integration path shown in Figure E.2, the line integral with respect to x is given by

$$\int_C F(x, y)dx = \int_{[ADE]} F(x, y)dx + \int_{[EF]} F(x, y)dx + \int_{[FGB]} F(x, y)dx \tag{E.3}$$

For the 3D case, another interpretation is more appropriate. One could visualize a mining activity that accumulates substance, say, Q, along path C in the ground containing a concentration distribution of Q. Let $F(x, y, z)$ be the amount of Q

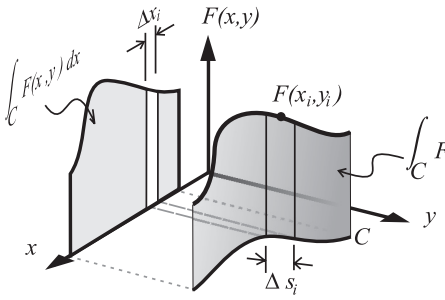


Figure E.1. Area interpretation of line integrals.

gathered per unit length traveled. Then, along the differential path ds, an amount $F(x, y, z)ds$ will have been accumulated, and the total amount gathered along the path C becomes $\int_C F(x, y, z)ds$. Conversely, the integral $\int_C F(x, y, z)dx$ is the amount of Q gathered along the projected path in the x-direction. In this mining scenario, the line integral $\int_C F(x, y, z)dx$ does not appear to be as relevant compared with the line integral with respect to s. However, these line integrals are quite useful during the computation of surface integrals and volume integrals because differential surfaces s are often described by $dx\,dy$, $dx\,dz$, or $dy\,dz$, and differential volumes are often described by the product $dx\,dy\,dz$.[1] Another example is when the integral involves the position vector $\vec{\mathbf{r}}$ of the form

$$\int_C \underline{\mathbf{f}} \cdot d\underline{\vec{\mathbf{r}}} = \int_C \left(f_x dx + f_y dy + f_z dz\right)$$

E.1.1 The Path of Integration

The path of integration will be assumed to be a continuous and sectionally smooth curve. The curve can either be **open** or **closed**. A path is **closed** if the starting point of the path coincides with the end point of the path. Otherwise, the path is said to be **open**. In either case, the direction of the path is crucial during integration. If the path is not self-intersecting at points other than the terminal points, then we say that the curve is a **simple** curve. Non-simple curves can be treated as the direct sum of simple curves, as shown in Figure E.3.

When the path is closed and non-intersecting, we often indicate a closed path by the following notation:

$$\oint_C F ds \qquad C \text{ is a closed, sectionally smooth, nonintersecting path}$$

A 3D path can be described generally by $C = \left(x(t), y(t), z(t)\right) = \vec{\underline{\mathbf{r}}}(t)$, where $\vec{\mathbf{r}}$ is the position vector and t is a parameter going from $t = 0$ to $t = 1$.[2] In some cases, the curve can be parameterized by either $x = t$, $y = t$ or $z = t$. In these cases, the other variables are said to possess an **explicit form**, for example, for $x = t$, we can use $y = y(x)$ and $z = z(x)$.[3]

[1] One could then expect that in other coordinate systems, $d\lambda$ may need involve those coordinates, for example, dr, $d\theta$, $d\phi$, and so forth.

[2] A more general formulation would be to let the parameter start at $t = a$ and end with $t = b$, where $b > a$. Using translation and scaling, this case could be reduced back to $a = 0$ and $b = 1$.

[3] The parameterizations can also originate from coordinate transformations such as polar, cylindrical, or spherical coordinates.

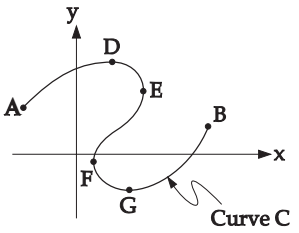


Figure E.2. A curve in which the projection of C onto x or y is not single valued.

EXAMPLE E.1. Consider the closed elliptical path described by

$$\left(\frac{x+3}{2}\right)^2 + (y+2)^2 = 4 \tag{E.4}$$

traversed in the counterclockwise direction as shown in Figure E.4. Let the path start at point $a : (-7, -2)$ and pass through points $b : (-3, -4)$, $c : (1, -2)$, $d : (-3, 0)$, and then back to a. The path can then be described in three equivalent ways:

1. **Parameterized Form.**

$$\begin{aligned} \text{Path } C_{abcda} : x &= -3 - 4\cos(2\pi t) \\ y &= -2 - 2\sin(2\pi t) \end{aligned}$$

from $t = 0$ to $t = 1$

2. **Explicit function of x.**

$$\text{Path } C_{abcda} = C_{abc} + C_{cda}$$

where

$$C_{abc} : y = -2 - \sqrt{4 - \left(\frac{x+3}{2}\right)^2} \qquad \text{from } x = -7 \text{ to } x = 1$$

$$C_{cda} : y = -2 + \sqrt{4 - \left(\frac{x+3}{2}\right)^2} \qquad \text{from } x = 1 \text{ to } x = -7$$

3. **Explicit function of y.**

$$\text{Path } C_{abcda} = C_{ab} + C_{bcd} + C_{da}$$

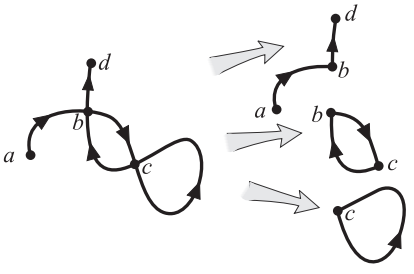


Figure E.3. Separation into simple curves.

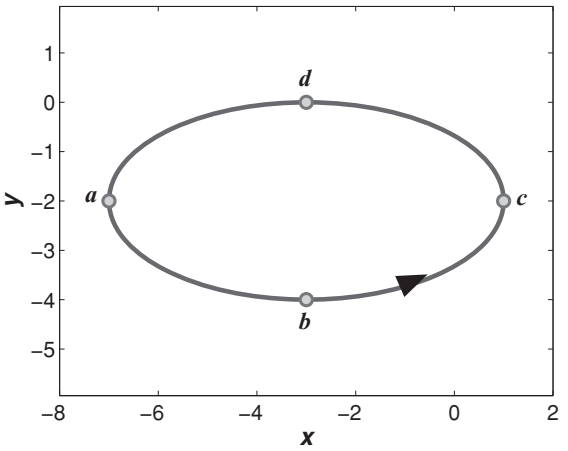


Figure E.4. A close path of integration in counterclockwise direction.

where

$$C_{ab} : x = -3 - 2\sqrt{4 - (y+2)^2} \qquad \text{from } y = -2 \text{ to } y = -4$$

$$C_{bcd} : x = -3 + 2\sqrt{4 - (y+2)^2} \qquad \text{from } y = -4 \text{ to } y = 0$$

$$C_{da} : x = -3 - 2\sqrt{4 - (y+2)^2} \qquad \text{from } y = 0 \text{ to } y = -2$$

E.1.2 Computation of Line Integrals

With the parameterized form of path C based on t, the integrand also becomes a function of t, that is,

$$F\Big(x(t), y(t), z(t)\Big) = g(t) \tag{E.5}$$

Using the chain rule, the line integrals become

$$\begin{aligned}
\int_C F(x,y,z)dx &= \int_0^1 \left(g(t)\frac{dx}{dt}\right) dt \\
\int_C F(x,y,z)dy &= \int_0^1 \left(g(t)\frac{dy}{dt}\right) dt \\
\int_C F(x,y,z)dz &= \int_0^1 \left(g(t)\frac{dz}{dt}\right) dt \\
\int_C F(x,y,z)ds &= \int_0^1 \left(g(t)\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\right) dt
\end{aligned} \tag{E.6}$$

However, if an explicit form is possible, these should be attempted in case they yield simpler calculations. For instance, suppose $y = y(x)$ and $z = z(x)$; then setting

$x = t$, (E.6) are modified by replacing $dx/dt = 1$, $dy/dt = dy/dx$ and $dz/dt = dz/dx$ with the lower limit x_{start} and upper limit x_{end}. For example,

$$\int_C F(x, y, z)dx = \int_{x_{\text{start}}}^{x_{\text{end}}} F(x, y(x), z(x))dx$$

EXAMPLE E.2. Consider the scalar function given by

$$F(x, y) = 2x + y + 3$$

and the counter-clockwise elliptical path of integration given in Example E.1.

Using the parameterized form based on t,

$$\begin{aligned} x(t) &= -3 - 4\cos(2\pi t) \\ y(t) &= -2 - 2\sin(2\pi t) \\ g(t) &= F\Big(x(t), y(t)\Big) = 2(-3 - 4\cos(2\pi t)) + (-2 - 2\sin(2\pi t)) + 3 \end{aligned}$$

and

$$\begin{aligned} dx &= 8\pi \sin(2\pi t)\, dt \\ dy &= -4\pi \cos(2\pi t)\, dt \\ ds &= 4\pi\sqrt{4\sin^2(2\pi t) + \cos^2(2\pi t)}dt \end{aligned}$$

Thus

$$\begin{aligned} \int_C F(x, y)dx &= \int_0^1 g(t)\,(8\pi \sin(2\pi t))\, dt = -8\pi \\ \int_C F(x, y)dy &= \int_0^1 g(t)\,(-4\pi \cos(2\pi t))\, dt = 16\pi \\ \int_C F(x, y)ds &= \int_0^1 g(t)\left(4\pi\sqrt{4\sin^2(2\pi t) + \cos^2(2\pi t)}\right)\, dt = -96.885 \end{aligned}$$

Using the explicit form $y = y(x)$ for the integration path

$$C = C_{abc} + C_{cda}$$

$$C_{abc} : y = y_{abc} = -2 - \sqrt{4 - \left(\frac{x+3}{2}\right)^2} \qquad \text{from } x = -7 \text{ to } x = 1$$

$$C_{cda} : y = y_{cda} = -2 + \sqrt{4 - \left(\frac{x+3}{2}\right)^2} \qquad \text{from } x = 1 \text{ to } x = -7$$

The integrand and differentials for the subpaths are

$$\begin{aligned} F(x, y)_{abc} &= 2x + 3 + \left(-2 - \sqrt{4 - \left(\frac{x+3}{2}\right)^2}\right) \\ F(x, y)_{cda} &= 2x + 3 + \left(-2 + \sqrt{4 - \left(\frac{x+3}{2}\right)^2}\right) \end{aligned}$$

$$\left(\frac{dy}{dx}\right)_{abc} = -\frac{x+3}{2\sqrt{(1-x)(x+7)}}$$
$$\left(\frac{dy}{dx}\right)_{cda} = +\frac{x+3}{2\sqrt{(1-x)(x+7)}}$$
$$\left(\frac{ds}{dx}\right)_{abc} = \sqrt{1+dy_{abc}^2}$$
$$\left(\frac{ds}{dx}\right)_{cda} = -\sqrt{1+dy_{cda}^2}$$

Note that ds has a negative sign for the subpath $[cda]$. This is because the direction of ds is opposite that of dx in this region.

The line integrals are then given by

$$\int_C F(x,y)dx = \int_{-7}^{1} F(x,y)_{abc}dx + \int_{1}^{-7} F(x,y)_{cda}dx$$
$$= -8\pi$$
$$\int_C F(x,y)dy = \int_{-7}^{1} F(x,y)_{abc}\left(\frac{dy}{dx}\right)_{abc}dx + \int_{1}^{-7} F(x,y)_{cda}\left(\frac{dy}{dx}\right)_{cda}dx$$
$$= 16\pi$$
$$\int_C F(x,y)ds = \int_{-7}^{1} F(x,y)_{abc}\left(\frac{ds}{dx}\right)_{abc}dx + \int_{1}^{-7} F(x,y)_{cda}\left(\frac{ds}{dx}\right)_{cda}dx$$
$$= -96.885$$

This shows that either the parameterized form or the explicit form approach can be used to obtain the same values. The choice is usually determined by the tradeoffs between the complexity of the parameterization procedure and the complexity of the resulting integral.

E.2 Surface Integrals

Definition E.2. *A* **surface integral** *of $F(x,y,z)$, with respect to area A and surface of integration $S(x,y,z)$, is defined by*

$$\int_S F(x,y,z)\,dA = \lim_{\Delta A_i \to 0, N\to\infty} \sum_{i=0}^{N} F(x_i,y_i,z_i)\,\Delta A_i \tag{E.7}$$

In most applications, the differential area is specified for either $dA = dx\,dy$, $dy\,dz$, $dx\,dz$, or dS, where dS is the differential area of the surface of integration

To visualize surface integrals, we could go back to the mining scenario for the substance Q, except now the accumulation is obtained by traversing a surface instead of a path. Thus the surface integral $\int_S f(x,y,z)dS$ can be thought of as the total amount mined by sweeping the total surface area S.

E.2.1 Surface of Integration

A general parametric description of surface is based on two independent parameters, u and v,

$$S \;:\; \Big(x(u,v), y(u,v), z(u,v)\Big) \quad \text{as } u \text{ and } v \text{ vary independently in a closed domain.} \tag{E.8}$$

If the parameterization can be done by letting $u = x$ and $v = y$, then the surface is given by the **explicit form** for z

$$S \;:\; z = z(x,y) \quad \text{as } x \text{ and } y \text{ vary independently in a closed domain} \tag{E.9}$$

Other explicit forms are possible, for example, $y = y(x,z)$ and $x = x(y,z)$.

Two important variables are needed during the calculation of surface integrals: the unit normal vector $\vec{\mathbf{n}}$, and the differential area dS at the point (x,y,z). As discussed in Section 4.6, the unit normal to a surface is given by (4.30), that is,

$$\vec{\mathbf{n}} = \frac{\vec{\mathbf{t}}_u \times \vec{\mathbf{t}}_v}{\left\| \vec{\mathbf{t}}_u \times \vec{\mathbf{t}}_v \right\|} \tag{E.10}$$

where

$$\vec{\mathbf{t}}_u = \frac{\partial \vec{\mathbf{r}}}{\partial u} \quad \text{and} \quad \vec{\mathbf{t}}_v = \frac{\partial \vec{\mathbf{r}}}{\partial v}$$

Specifically, we have

$$\vec{\mathbf{t}}_u \times \vec{\mathbf{t}}_v = \left(\frac{\partial(y,z)}{\partial(u,v)} \boldsymbol{\delta}_x + \frac{\partial(z,x)}{\partial(u,v)} \boldsymbol{\delta}_y + \frac{\partial(x,y)}{\partial(u,v)} \boldsymbol{\delta}_z \right) \tag{E.11}$$

where we used the shorthand notation for the Jacobian determinants given by

$$\frac{\partial(a,b)}{\partial(c,d)} = \mathbf{det} \begin{pmatrix} \dfrac{\partial a}{\partial c} & \dfrac{\partial a}{\partial d} \\ \dfrac{\partial b}{\partial c} & \dfrac{\partial b}{\partial d} \end{pmatrix}$$

However, the differential surface area is given by the area of the parallelogram formed by differential arcs form by movement along constant v and u, respectively, that is, the area formed by $\vec{\mathbf{t}}_u\, du$ and $\vec{\mathbf{t}}_v\, dv$. Thus

$$\begin{aligned} dS &= \left\| \left(\vec{\mathbf{t}}_u\, du\right) \times \left(\vec{\mathbf{t}}_v\, dv\right) \right\| = \left\| \vec{\mathbf{t}}_u \times \vec{\mathbf{t}}_v \right\| du\, dv \\ &= \sqrt{\left(\frac{\partial(y,z)}{\partial(u,v)} \right)^2 + \left(\frac{\partial(z,x)}{\partial(u,v)} \right)^2 + \left(\frac{\partial(x,y)}{\partial(u,v)} \right)^2}\, du\, dv \end{aligned} \tag{E.12}$$

If the explicit form $z = z(x,y)$ is possible, that is, with $x = u$ and $y = v$, the formulas reduce to the more familiar ones, that is,

$$\mathbf{n} = \frac{\left(\dfrac{\partial z}{\partial x} \boldsymbol{\delta}_x - \dfrac{\partial z}{\partial y} \boldsymbol{\delta}_y + \boldsymbol{\delta}_z \right)}{\sqrt{1 + \left(\dfrac{\partial z}{\partial x} \right)^2 + \left(\dfrac{\partial z}{\partial y} \right)^2}} \tag{E.13}$$

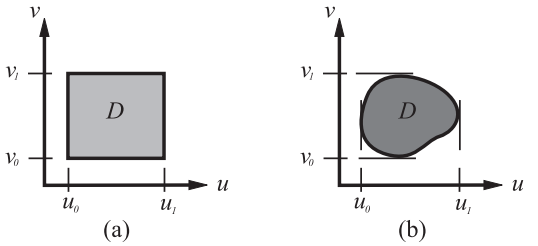


Figure E.5. The boundary of domain D for parameter space can either have (a) an independent range, or (b) an interdependent range.

$$dS = \left(\sqrt{1+\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2}\right) dxdy \tag{E.14}$$

Note that with the square root, the choice for sign depends on the interpretation of the surface direction. In most application, for a surface that encloses a region of 3D space, the surface outward of the enclosed region is often given a positive sign.

EXAMPLE E.3. Consider a circular cylinder of radius R of height h with the bottom base centered at the origin. The differential area at the top and the bottom can be parameterized in terms of r and θ; that is, $x = r\cos\theta$ and $y = r\sin\theta$. At the top, we have $z = 0$ and set $u = r$ and $v = \theta$ as the parameterization. At the bottom, we have $z = h$ but will need to set $u = r$ and $v = \theta$ as the parameterization to obtain the expected outward normal direction. Thus, for the top,

$$\vec{\underline{\mathbf{t}}}_u \times \vec{\underline{\mathbf{t}}}_v = \mathbf{det}\begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \underline{\boldsymbol{\delta}}_y & \underline{\boldsymbol{\delta}}_z \\ \cos\theta & \sin\theta & 0 \\ -r\sin\theta & r\cos\theta & 0 \end{pmatrix} = r\underline{\boldsymbol{\delta}}_z \quad \rightarrow \quad \begin{aligned} \vec{\underline{\mathbf{n}}}_{\text{top}} &= \underline{\boldsymbol{\delta}}_z \\ dS_{\text{top}} &= r\,dr\,d\theta \end{aligned}$$

For the bottom, we have

$$\vec{\underline{\mathbf{t}}}_u \times \vec{\underline{\mathbf{t}}}_v = \mathbf{det}\begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \underline{\boldsymbol{\delta}}_y & \underline{\boldsymbol{\delta}}_z \\ -r\sin\theta & r\cos\theta & 0 \\ \cos\theta & \sin\theta & 0 \end{pmatrix} = -r\underline{\boldsymbol{\delta}}_z \quad \rightarrow \quad \begin{aligned} \vec{\underline{\mathbf{n}}}_{\text{bottom}} &= -\underline{\boldsymbol{\delta}}_z \\ dS_{\text{bottom}} &= r\,dr\,d\theta \end{aligned}$$

For the side of the cylinder, we let $u = \theta$ and $v = z$ and $r = R$. Then

$$\vec{\underline{\mathbf{t}}}_u \times \vec{\underline{\mathbf{t}}}_v = \mathbf{det}\begin{pmatrix} \underline{\boldsymbol{\delta}}_x & \underline{\boldsymbol{\delta}}_y & \underline{\boldsymbol{\delta}}_z \\ -R\sin\theta & R\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = R\left(\cos\theta\underline{\boldsymbol{\delta}}_x + \sin\theta\underline{\boldsymbol{\delta}}_y\right) = R\underline{\boldsymbol{\delta}}_r$$

which the yields

$$\vec{\underline{\mathbf{n}}}_{\text{side}} = \underline{\boldsymbol{\delta}}_r \quad \text{and} \quad dS_{\text{side}} = Rd\theta dz$$

E.2.2 Computation of Surface Integrals

Under the parameterized form of the surface of integration, the domain of the parameter space is a closed 2D plane in the (u, v) space. The boundary may either be defined independently by fixed ranges for u and v, or the boundary has to described by giving explicit dependencies of u on v or vice versa (see Figure E.5).

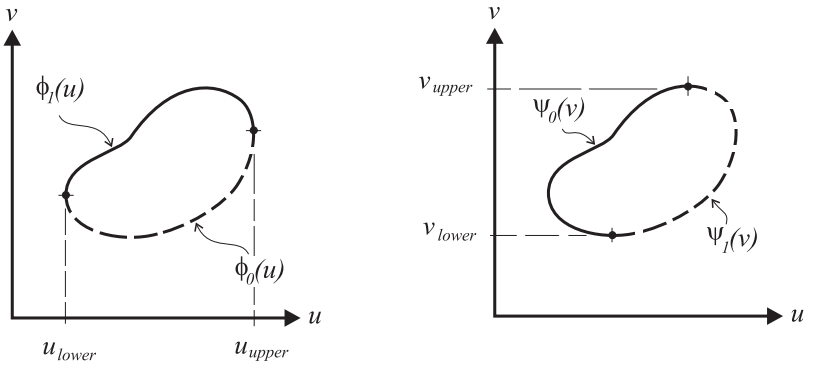


Figure E.6. The two possible domain descriptions: (a) boundary is partitioned into two segments such that $v = \phi(u)$, and (b) boundary is partitioned into two segments such that $u = \psi(v)$.

If the ranges of u and v are independent, then domain D can, without loss of generality, be given as

$$D \; : \; 0 \leq u \leq 1 \quad ; \quad 0 \leq v \leq 1$$

The surface integral becomes

$$\int_S F(x, y, z) dS = \int_0^1 \int_0^1 g(u, v) du \, dv$$

where

$$g(u, v) = f(x(u, v), y(u, v), z(u, v)) \sqrt{\left(\frac{\partial(y, z)}{\partial(u, v)}\right)^2 + \left(\frac{\partial(z, x)}{\partial(u, v)}\right)^2 + \left(\frac{\partial(x, y)}{\partial(u, v)}\right)^2} \tag{E.15}$$

Thus

$$\begin{aligned} h(v) &= \int_0^1 g(u, v) du \qquad \text{holding } v \text{ constant} \\ \int_S F(x, y, z) dS &= \int_0^1 h(v) dv \end{aligned} \tag{E.16}$$

If u and v are interdependent at the boundary of the parameter space, then two domain descriptions are possible:

$$D_u \; : \; u_{\text{lower}} \leq u \leq u_{\text{upper}} \quad ; \quad \phi_0(u) \leq v \leq \phi_1(u) \tag{E.17}$$

or

$$D_v \; : \; v_{\text{lower}} \leq v \leq v_{\text{upper}} \quad ; \quad \psi_0(v) \leq u \leq \psi_1(v) \tag{E.18}$$

where u_{lower}, u_{upper}, v_{lower} and v_{upper} are constants. Both domain descriptions are shown in Figure E.6, and both are equally valid.

With the first description given by (E.17), the surface integral is given by

$$\begin{aligned} h(u) &= \int_{\phi_0(u)}^{\phi_1(u)} g(u,v)dv \qquad \text{holding } u \text{ constant} \\ \int_S F(x,y,z)dS &= \int_{u_{\text{lower}}}^{u_{\text{upper}}} h(u)du \end{aligned} \tag{E.19}$$

where $g(u, v)$ is the same function as in (E.15). Similarly, using the second description given in (E.18),

$$\begin{aligned} h(v) &= \int_{\psi_0(v)}^{\psi_1(v)} g(u,v)du \qquad \text{holding } u \text{ constant} \\ \int_S F(x,y,z)dS &= \int_{v_{\text{lower}}}^{v_{\text{upper}}} h(v)dv \end{aligned} \tag{E.20}$$

For the special case in which the surface is given by $z = z(x, y)$,

$$\begin{aligned} u &= x & v &= y \\ g(u,v) &= g(x,y) = f(x,y,z(x,y))\sqrt{1+\left(\frac{\partial z}{\partial x}\right)^2+\left(\frac{\partial z}{\partial y}\right)^2} \\ u_{\text{lower}} &= x_{\text{lower}} & u_{\text{upper}} &= x_{\text{upper}} \\ \phi_1(u) &= \phi_1(x) & \phi_0(u) &= \phi_0(x) \\ v_{\text{lower}} &= y_{\text{lower}} & v_{\text{upper}} &= y_{\text{upper}} \\ \psi_1(v) &= \psi_1(y) & \psi_0(v) &= \psi_0(y) \end{aligned}$$

EXAMPLE E.4. Consider the integrand given by

$$F(x,y,z) = 2x + y - z + 3$$

and the surface of integration provided by the ellipsoid,

$$x^2 + \left(\frac{y}{2}\right)^2 + z^2 = 1$$

A parameterized form is given by

$$x = \sin(u)\cos(v) \quad ; \quad y = \sin(u)\sin(v) \quad ; \quad z = \cos(u)$$

where the parameter domain is described by

$$0 \le u \le 2\pi \qquad 0 \le v \le \pi$$

The Jacobian determinants can be evaluated as

$$\begin{aligned} \frac{\partial(x,y)}{\partial(u,v)} &= -2\sin(u)\cos(u) \\ \frac{\partial(y,z)}{\partial(u,v)} &= -2\sin^2(u)\cos(v) \\ \frac{\partial(x,z)}{\partial(u,v)} &= \sin^2(u)\cos(v) \end{aligned}$$

which then gives

$$\begin{aligned} g(u,v) &= F(x,y,z)\sqrt{\left(\frac{\partial(y,z)}{\partial(u,v)}\right)^2+\left(\frac{\partial(z,x)}{\partial(u,v)}\right)^2+\left(\frac{\partial(x,y)}{\partial(u,v)}\right)^2} \\ &= \alpha(u,v)\beta(u,v) \end{aligned}$$

where

$$\begin{aligned} \alpha(u,v) &= 2\sin(u)\cos(v)+2\sin(u)\sin(v)-\cos(u)+3 \\ \beta(u.v) &= \sqrt{3\cos^2(v)\left((\cos(u)-1)^2(\cos(u)+1)^2\right)+(1+2\cos^2(u)-3\cos^4(u))} \end{aligned}$$

The surface integral can then be solved numerically to be

$$\int_0^{\pi}\int_0^{2\pi} g(u,v)du\,dv = 64.4$$

As an alternative, we can partition the elliptical surface into two halves. The upper half and lower half can be described by z_u and z_ℓ, respectively, where

$$z_u = \sqrt{1-x^2-\left(\frac{y^2}{2}\right)} \qquad z_\ell = -\sqrt{1-x^2-\left(\frac{y^2}{2}\right)}$$

In either half, the (x,y)-domain can be described by

$$D: \; -1 \le x \le 1 \qquad -2\sqrt{1-x^2} \le y \le 2\sqrt{1-x^2}$$

For the upper half,

$$\frac{dz_u}{dx} = \frac{-2x}{\sqrt{4-4x^2-y^2}} \qquad \frac{dz_u}{dy} = \frac{-y/2}{\sqrt{4-4x^2-y^2}}$$

with an integrand

$$g_u(x,y) = \left(2x+y-\sqrt{1-x^2-\left(\frac{y^2}{2}\right)}+3\right)\left(\frac{1}{2}\sqrt{\frac{3y^2-16}{-4+4x^2+y^2}}\right)$$

For the lower half,

$$\frac{dz_\ell}{dx} = \frac{2x}{\sqrt{4-4x^2-y^2}} \qquad \frac{dz_\ell}{dy} = \frac{y/2}{\sqrt{4-4x^2-y^2}}$$

with an integrand

$$g_\ell(x,y) = \left(2x+y+\sqrt{1-x^2-\left(\frac{y^2}{2}\right)}+3\right)\left(\frac{1}{2}\sqrt{\frac{3y^2-16}{-4+4x^2+y^2}}\right)$$

Combining everything, we can calculate the surface integral via numerical integration to be

$$\begin{aligned} \mathcal{I}_u &= \int_{-1}^{+1} \int_{-2\sqrt{1-x^2}}^{2\sqrt{1-x^2}} g_u(x,y)dy\,dx = 26.6 \\ \mathcal{I}_\ell &= \int_{-1}^{+1} \int_{-2\sqrt{1-x^2}}^{2\sqrt{1-x^2}} g_\ell(x,y)dy\,dx = 37.8 \\ \int_S f dS &= \mathcal{I}_u + \mathcal{I}_\ell = 64.4 \end{aligned}$$

which is the same value as the previous answer using the parameterized description.

Remark: In the example just shown, we have used numerical integration. This is usually the preferred route when the integrand becomes too complicated to integrate analytically. There are several ways in which the numerical approximation can be achieved, including the rectangular or trapezoidal approximations or Simpson's methods. We have also included another efficient numerical integration technique called the **Gauss-Legendre quadrature method** in the appendix as Section E.4.

E.3 Volume Integrals

Definition E.3. *A* **volume integral** *of $F(x, y, z)$, with respect to W and volume of integration $V(x, y, z)$, is defined by*

$$\int_V F(x,y,z)\,dW = \lim_{\Delta W_i \to 0, N \to \infty} \sum_{i=0}^{N} F(x_i, y_i, z_i)\,\Delta W_i \tag{E.21}$$

In most applications, the differential volume is specified by $dW = dx\,dy\,dz$.

To continue the visual interpretation via mining used earlier for both the line and surface integrals, the mining activity now accumulates substance Q indicated by $\int_V F(x, y, z)dV$ by carving out a volume V specified by the boundary.

E.3.1 Volume of Integration

In most cases, the rectangular coordinate system is sufficient to describe the surface of the volume, and thus the differential volume is given by $dV = dx\,dy\,dz$. However, in other cases, another set of coordinates allow for easier computation, for example, cylindrical or spherical coordinates. Let this set of new coordinates be given by parameters (u, v, w). Let $\vec{\mathbf{r}}$ be the position vector. At a point $\mathbf{p}$, we can trace paths C_1, C_2, and C_3 that pass through point $\mathbf{p}$, each path formed by holding the other two parameters fixed. This is shown in Figure E.7, where the differential arcs along each of each curve are given by $\underline{\mathbf{a}}$, $\underline{\mathbf{b}}$, and $\underline{\mathbf{c}}$ where

$$\underline{\mathbf{a}} = \frac{\partial \vec{\mathbf{r}}}{\partial u} du \quad ; \quad \underline{\mathbf{b}} = \frac{\partial \vec{\mathbf{r}}}{\partial v} dv \quad ; \quad \underline{\mathbf{c}} = \frac{\partial \vec{\mathbf{r}}}{\partial w} dw$$

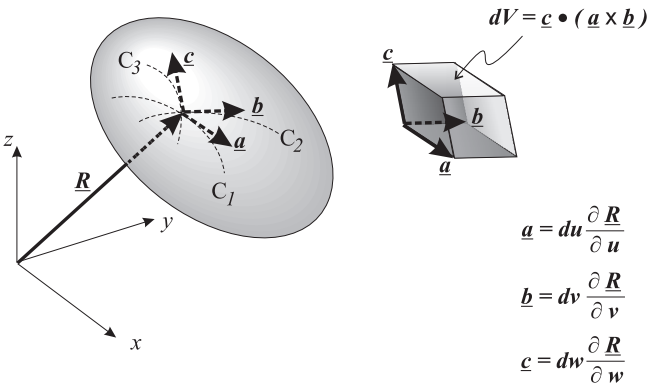


Figure E.7. Graphical representation of differential volume, dV, as function of u, v, and w. Note that the position described by $\vec{\mathbf{r}}$ can be anywhere on or inside V.

The differential volume is then formed by the absolute value of the triple product formed by $\underline{\mathbf{a}}$, $\underline{\mathbf{b}}$, and $\underline{\mathbf{c}}$, that is,

$$dV = \left| \underline{\mathbf{c}} \cdot (\underline{\mathbf{a}} \times \underline{\mathbf{b}}) \right| = \left| \mathbf{det} \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{pmatrix} \right| du\, dv\, dw \qquad \text{(E.22)}$$

EXAMPLE E.5. For the spherical coordinates, using $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, and $z = r\cos\theta$ with the parameters $u = r$, $v = \theta$, and $w = \phi$, we have

$$dV = \mathbf{det} \begin{pmatrix} \sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ \cos\theta & -r\sin\theta & 0 \end{pmatrix} dr\, d\theta\, d\phi = r^2 \sin\theta\, dr\, d\theta\, d\phi$$

E.3.2 Computation of Volume Integrals

Having determined the differential volume and the integrand, one needs to identify the limits of integration in each of the variables x, y, and z, or of parameters u, v, and w.

If the limits are independent,

$$u_{\min} \le u \le u_{\max} \;;\; v_{\min} \le v \le v_{\max} \;;\; w_{\min} \le w \le w_{\max}$$

the volume integral can be integrated in a nested fashion,

$$\int_V F dV = \int_{w_{\min}}^{w_{\max}} \left(\int_{v_{\min}}^{v_{\max}} \left(\int_{u_{\min}}^{u_{\max}} G(u, v, w) \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du \right) dv \right) dw \qquad \text{(E.23)}$$

$$w_{\min} \leq w \leq w_{\max}$$

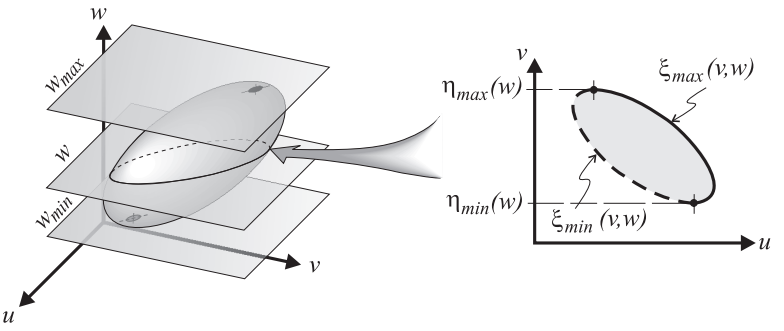


Figure E.8. A nested description of volume boundaries.

where

$$G(u, v, w) = F\left(x(u, v, w), y(u, v, w), z(u, v, w)\right) \tag{E.24}$$

If the surface of the volume space is represented by a set of interdependent parameters, there are six possible descriptions that can be used based on the sequence of dependencies. We only describe the sequence $w \to v \to u$. As shown in Figure E.8, we can first identify the maximum and minimum value of w, that is,

$$w_{\min} \leq w \leq w_{\max}$$

Taking a slice of the volume at a fixed w, a closed region whose boundary can be identified by

$$\eta_{\min}(w) \leq v \leq \eta_{\max}(w)$$

Finally, as shown in Figure E.8, the limits of v for this slice will divide the closed curve into two segments. Each of these segments can then be described by functions of v and w, where the value of w was that used to obtain the slice,

$$\xi_{\min}(v, w) \leq u \leq \xi_{\max}(v, w)$$

Thus we end up with a slightly different nested integration given by

$$\int_V F dV = \int_{w_{\min}}^{w_{\max}} \int_{\eta_{\min}(w)}^{\eta_{\max}(w)} \int_{\xi_{\min}(v,w)}^{\xi_{\max}(v,w)} G(u, v, w) \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du\, dv\, dw \tag{E.25}$$

where $G(u, v, w)$ is the same function as in (E.24).

EXAMPLE E.6. Consider the integrand given by

$$F(x, y, z) = 2x + y - z + 3$$

and the volume of integration given by the ellipsoid

$$x^2 + \left(\frac{y}{2}\right)^2 + z^2 \leq 1$$

Using the parameterization

$$x = u \sin(v) \cos(w) \;\; ; \;\; y = 2u \sin(v) \sin(w) \;\; ; \;\; z = u \cos(v)$$

with boundaries,

$$0 \le u \le 1 \;;\; 0 \le v \le 2\pi \;;\; 0 \le w \le \pi$$

Let $s_w = \sin(w)$, $c_w = \cos(w)$, $s_v = \sin(v)$, and $c_v = \cos(v)$. Then the differential volume is

$$\begin{aligned} dV &= \left| \mathbf{det} \begin{pmatrix} s_v c_w & u c_v c_w & -u s_v s_w \\ 2 s_v s_w & 2u c_v s_w & 2u s_v c_w \\ c_v & -u s_v & 0 \end{pmatrix} \right| du\,dv\,dw \\ &= 2u^2 \left| s_v \right| du\,dv\,dw \end{aligned}$$

while the integrand becomes

$$G = 2u s_v \left(c_w + s_w\right) - u c_v + 3$$

Combining all the elements together, we can compute the volume integral as

$$\int_0^{\pi} \int_0^{2\pi} \int_0^1 G(u, v, w) \left(2u^2 \left| s_v \right| du\,dv\,dw\right) = 8\pi$$

Alternatively, we could use the original variables x, y and z. Doing so, the differential volume is $dV = dx\,dy\,dz$, whereas the boundary of the volume of integration is given by

$$\begin{aligned} \text{Surface boundary:} \qquad -\sqrt{1 - z^2 - \left(\frac{y}{2}\right)^2} \le\ & x \ \le \sqrt{1 - z^2 - \left(\frac{y}{2}\right)^2} \\ -2\sqrt{1 - z^2} \le\ & y \ \le 2\sqrt{1 - z^2} \\ -1 \le\ & z \ \le 1 \end{aligned}$$

Thus the volume integral is given by

$$\int_{-1}^{1} \int_{-2\sqrt{1-z^2}}^{2\sqrt{1-z^2}} \int_{-\sqrt{1-z^2-(y/2)^2}}^{\sqrt{1-z^2-(y/2)^2}} (2x + y - z)\, dx\,dy\,dz = 8\pi$$

which is the same answer obtained by using the parameterized description.

E.4 Gauss-Legendre Quadrature

The n-point **Gauss-Legendre quadrature** is a numerical approximation of the integral $\int_{-1}^{+1} F(x)dx$ that satisfies two conditions:

1. The integral is approximated by a linear combination of n values of $F(x)$, each evaluated at $-1 \le x_i \le 1$, that is,

$$\int_{-1}^{1} F(x)dx \approx \sum_{i=1}^{n} W_i F(x_i) \tag{E.26}$$

and

2. When $F(x)$ is a $(2n-1)^{\text{th}}$ order polynomial, the approximation becomes an equality, that is, if $F(x) = \sum_{m=0}^{2n-1} a_m x^m$,

$$\int_{-1}^{1} \left(\sum_{m=0}^{2n-1} a_m x^m \right) dx = \sum_{i=1}^{n} W_i \left(\sum_{m=0}^{2n-1} a_m x_i^m \right) \tag{E.27}$$

Approximations having the form given in (E.26) are generally called **quadrature formulas**. Other quadrature formulas include Newton-Cotes' formulas, Simpson's formulas, and trapezoidal formulas. The conditions given in (E.27) distinguish the values found for W_i and x_i as being Gauss-Legendre quadrature parameters.

A direct approach to determine W_i and x_i is obtained by generating the required equations using (E.27):

$$\begin{aligned} \int_{-1}^{1} \left(\sum_{m=0}^{2n-1} a_m x^m \right) dx &= \sum_{i=1}^{n} W_i \left(\sum_{m=0}^{2n-1} a_m x_i^m \right) \\ \sum_{m=0}^{2n-1} a_m x^m \int_{-1}^{1} x^m dx &= \sum_{m=0}^{2n-1} a_m \sum_{i=1}^{n} W_i x_i^m \\ \sum_{m=0}^{2n-1} a_m \left(\int_{-1}^{1} x^m dx - \sum_{i=1}^{n} W_i x_i^m \right) &= 0 \end{aligned} \tag{E.28}$$

Because the condition in (E.27) should be true for any polynomial of order $(2n-1)$, (E.28) should be true for arbitrary values of a_m, $m = 0, 1, \ldots, (2n-1)$. This yields

$$\sum_{i=1}^{n} W_i x_i^m = \gamma_m \qquad \text{for } m = 0, 1, \ldots, (2n-1) \tag{E.29}$$

where

$$\gamma_m = \int_{-1}^{1} x^m dx = \begin{cases} 2/(m+1) & \text{if } m \text{ is even} \\ 0 & \text{if } m \text{ is odd} \end{cases} \tag{E.30}$$

This means that we have $2n$ independent equations that can be used to solve the $2n$ unknowns: x_i and W_i. Unfortunately, the equation becomes increasingly difficult to solve as n gets larger. This is due to the nonlinear terms such as $W_i x_i^m$ appearing in (E.29).

An alternative approach is to separate the problem of identifying the x_i values from the problem of identifying the W_i values. To do so, we use Legendre polynomials and take advantage of their orthogonality properties.

We first present some preliminary formulas:

1. Any polynomial of finite order can be represented in terms of Legendre polynomials, that is,

$$\sum_{i=0}^{q} c_i x^i = \sum_{j=0}^{q} b_j \mathcal{P}_j(x) \tag{E.31}$$

where $\mathcal{P}_j(x)$ is the Legendre polynomial of order j. (To obtain a Legendre polynomial, one can either use definition given in (I.31) or use Rodriguez's formula given in (9.46).)

2. Let $R_{(2n-1)}(x)$ be a polynomial of order $(2n-1)$ formed by the product of a polynomial of order $(n-1)$ and a Legendre polynomial of order n, that is,

$$R_{(2n-1)}(x) = \left(\sum_{i=0}^{n-1} c_i x^i\right)(\mathcal{P}_n(x)) \tag{E.32}$$

With this definition, the integral of $R_{(2n-1)}(x)$, with limits from -1 to 1, is guaranteed to be zero. To see this, we apply (E.31) to the first polynomial on the right-hand side of (E.32), integrate both sides, and then apply the orthogonality properties of Legendre polynomials (cf. (9.48)), that is,

$$\begin{aligned}\int_{-1}^{1} R_{(2n-1)}(x)dx &= \int_{-1}^{1}\left[\left(\sum_{i=0}^{n-1} b_i\mathcal{P}_i(x)\right)(\mathcal{P}_n(x))\right]dx \\ &= \sum_{i=0}^{n-1} b_i\left[\int_{-1}^{1}\mathcal{P}_i(x)\mathcal{P}_n(x)dx\right] \\ &= 0\end{aligned} \tag{E.33}$$

3. One can always decompose a $(2n-1)^{\text{th}}$ order polynomial, say, $\psi_{(2n-1)}(x)$, into a sum of two polynomials

$$\psi_{(2n-1)}(x) = \zeta_{(n-1)}(x) + R_{(2n-1)}(x) \tag{E.34}$$

where $\zeta_{(n-1)}(x)$ is an $(n-1)^{\text{th}}$ order polynomial and $R_{(2n-1)}(x)$ is a $(2n-1)^{\text{th}}$ order polynomial that satisfies the form given in (E.32).

To show this fact constructively, let $r_1, \ldots, r_n$ be the roots of the n^{th}-order Legendre polynomial, $\mathcal{P}_n(x)$. By virtue of the definition given in (E.32), we see that $R_{(2n-1)}(r_i) = 0$ also. Using this result, we can apply each of the n roots to (E.34) and obtain

$$\psi_{(2n-1)}(r_i) = \zeta_{(n-1)}(r_i) \qquad i = 1, 2, \ldots, n \tag{E.35}$$

One can then obtain $\zeta_{(n-1)}(x)$ to be the unique $(n-1)^{\text{th}}$ order polynomial that passes through n points given by $\left(r_i, \psi_{(2n-1)}(r_i)\right)$. Subsequently, $R_{(2n-1)}(x)$ can be found by subtracting $\zeta_{(n-1)}(x)$ from $\psi_{(2n-1)}(x)$.

4. Using the decomposition given in (E.34) and the integral identity given in (E.33), an immediate consequence is the following identity:

$$\int_{-1}^{1}\psi_{(2n-1)}(x)dx = \int_{-1}^{1}\zeta_{(n-1)}(x)dx \tag{E.36}$$

This means the integral of an $(2n-1)^{\text{th}}$ order polynomial can always be replaced by the integral of a corresponding $(n-1)^{\text{th}}$ order polynomial.

We now use the last two results, namely (E.35) and (E.36), to determine the Gauss-Legendre parameters. Recall (E.27), which is the condition for a Gauss-Legendre quadrature, and apply it to $\psi_{(2n-1)}(x)$,

$$\int_{-1}^{1}\psi_{(2n-1)}(x)dx = \sum_{i=1}^{n} W_i\psi_{(2n-1)}(x_i) \tag{E.37}$$

Now set $x_i = r_i$, that is, the roots of the n^{th} order Legendre polynomial. Next, apply (E.35) on the right-hand side, and apply (E.36) on the left-hand side of the equation:

$$\begin{aligned} \int_{-1}^{1} \psi_{(2n-1)}(x)dx &= \sum_{i=1}^{n} W_i \psi_{(2n-1)}(r_i) \\ \int_{-1}^{1} \zeta_{(n-1)}(x)dx &= \sum_{i=1}^{n} W_i \zeta_{(n-1)}(r_i) \end{aligned} \tag{E.38}$$

Let $\zeta_{(n-1)}(x) = \sum_{k=0}^{n-1} b_k x^k$. Then (E.38) becomes

$$\begin{aligned} \int_{-1}^{1} \sum_{k=0}^{n-1} b_k x^k dx &= \sum_{i=1}^{n} W_i \left(\sum_{k=0}^{n-1} b_k r_i^k \right) \\ \sum_{k=0}^{n-1} b_k \int_{-1}^{1} x^k dx &= \sum_{k=0}^{n-1} b_k \sum_{i=1}^{n} W_i \, r_i^k \\ \sum_{k=0}^{n-1} b_k \left(\sum_{i=1}^{n} W_i \, r_i^k \; - \; \gamma_k \right) &= 0 \end{aligned} \tag{E.39}$$

where

$$\gamma_k = \int_{-1}^{1} x^k dx = \begin{cases} 2/(k+1) & \text{if } k \text{ is even} \\ 0 & \text{if } k \text{ is odd} \end{cases} \tag{E.40}$$

The b_k value should be left arbitrary because it corresponds to a general polynomial $\psi_{(2n-1)}$, as required by the second condition for a Gauss-Legendre quadrature. This then yields n equations. In matrix form, we have

$$\begin{pmatrix} 1 & 1 & \dots & 1 \\ r_1 & r_2 & \dots & r_n \\ \vdots & \vdots & \ddots & \vdots \\ r_1^{n-1} & r_2^{n-1} & \dots & r_n^{n-1} \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{pmatrix} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{n-1} \end{pmatrix} \tag{E.41}$$

In summary, to obtain the parameters for an n-point Gauss-Legendre quadrature, first solve for the roots r_i of the n^{th}-order Legendre polynomial, $i = 1, \dots, n$. After substituting these values into (E.41), we can solve for $W_i, i = 1, \dots, n$.[4]

[4] The first equation in (E.41), $\sum_{i=1}^{n} W_i = 2$, can be viewed as a partition of the domain $-1 \le x \le 1$ into n segments having widths W_i. As each of these partitions are given the corresponding heights of $F(x_i = r_i)$, the integral approximation is seen as a sum of rectangular areas. This means that the process replaces the original shape of the integration area into a set of quadrilaterals. Hence, the general term "quadrature." For integrals of function in two dimensions, a similar process is called "cubature."

EXAMPLE E.7. For $n = 3$, we have

$$\mathcal{P}_3(x) = \frac{x}{2}\left(5x^2 - 3\right)$$

whose roots are, arranged in increasing order, $r_1 = -\sqrt{0.6}, r_2 = 0$ and $r_3 = \sqrt{0.6}$. Substituting these values in (E.41),

$$\begin{pmatrix} 1 & 1 & 1 \\ -\sqrt{0.6} & 0 & \sqrt{0.6} \\ 0.6 & 0 & 0.6 \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 2/3 \end{pmatrix}$$

whose solution is given by $W_1 = W_3 = 5/9$ and $W_2 = 8/9$.

Note that $r_1 = -r_3$. This is not a coincidence but a property of Legendre polynomials. In general, for an n^{th}-order Legendre polynomial: (1) for n odd, one of the roots will always be zero, and (2) each positive root will have a corresponding negative root of the same magnitude.

Extending the results to p-dimensional box domains represented by mutually orthogonal coordinates: $\{x_1, \ldots, x_p\}$, the Gauss-Legendre formulas can be applied one dimension at a time, that is,

$$\begin{aligned} &\int_{-1}^{1} \cdots \int_{-1}^{1} \left[f\left(x_1, \ldots, x_n\right)\right] dx_1 \cdots dx_p \\ &\qquad = \int_{-1}^{1} \cdots \int_{-1}^{1} \left[\sum_{i_p=1}^{n} W_{i_p} f\left(x_1, \ldots, x_{p-1}, r_{i_p}\right)\right] dx_1 \cdots dx_{p-1} \\ &\qquad \vdots \\ &\qquad = \sum_{i_1=1}^{n} \cdots \sum_{i_p=1}^{n} \left(W_{i_1} \cdots W_{i_p}\right) f(r_{i_1}, \ldots, r_{i_p}) \end{aligned} \tag{E.42}$$

where W_i and r_i are the same values obtained earlier for the one-dimensional case.

E.5 Proofs of Integral Theorems

E.5.1 Proof of Green's Lemma (Lemma 5.1)

To prove the lemma, we make use of two possible descriptions of the boundary as given in (E.17) and (E.18).

Recalling (E.17), the domain of the surface of integration S is given by

$$D: \; u_{\text{lower}} \le u \le u_{\text{upper}}; \quad \phi_0(u) \le v \le \phi_1(u)$$

where the closed contour C is the sum given by

$$C = C_{0,v} - C_{1,v}$$

where $C_{0,v}$ and $C_{1,v}$, the curves described by $\phi_0(u)$ and $\phi_1(u)$, respectively, are positive with increasing values of u.

Applying this description to the second surface integral in (5.1),

$$\begin{aligned}\int_S \frac{\partial F(u,v)}{\partial v}du\,dv &= \int_{u_{\text{lower}}}^{u_{\text{upper}}}\left(\int_{\phi_0(u)}^{\phi_1(u)}\frac{\partial F(u,v)}{\partial v}dv\right)du \\ &= \int_{u_{\text{lower}}}^{u_{\text{upper}}}\left(F(u,\phi_1(u))-F(u,\phi_0(u))\right)du \\ &= -\oint_C F(u,v)du \end{aligned} \tag{E.43}$$

Likewise, using (E.18), the domain of the surface of integration S is given by

$$D:\ v_{\text{lower}} \le v \le v_{\text{upper}};\quad \psi_0(v) \le u \le \psi_1(v)$$

where the closed contour C is now equal to the sum given by

$$C = C_{1,u} - C_{0,u}$$

where $C_{0,v}$ and $C_{1,v}$, the curves described by $\psi_0(u)$ and $\psi_1(u)$, respectively, are positive with increasing values of v.

Applying this domain description to the first surface integral in (5.1),

$$\begin{aligned}\int_S \frac{\partial G(u,v)}{\partial u}du\,dv &= \int_{v_{\text{lower}}}^{v_{\text{upper}}}\left(\int_{\psi_0(v)}^{\psi_1(v)}\frac{\partial G(u,v)}{\partial u}dv\right)du \\ &= \int_{v_{\text{lower}}}^{v_{\text{upper}}}\left(G(\psi_1(v),v)-G(\psi_0(v),v)\right)du \\ &= \oint_C G(u,v)dv \end{aligned} \tag{E.44}$$

Combining (E.43) and (E.44), we arrive at the formula given in Green's lemma,

$$\oint_C G(u,v)dv + \oint_C F(u,v)du = \int_S \frac{\partial G(u,v)}{\partial u}du\,dv - \int_S \frac{\partial F(u,v)}{\partial v}du\,dv$$

E.5.2 Proof of Divergence Theorem (Theorem 5.1)

In rectangular coordinates, let $\underline{\mathbf{f}}$ be given by

$$\underline{\mathbf{f}} = f_x\,\underline{\boldsymbol{\delta}}_x + f_y\,\underline{\boldsymbol{\delta}}_y + f_z\,\underline{\boldsymbol{\delta}}_z$$

The volume integral in (5.5) can be expanded to be the sum of three terms

$$\int_V \nabla\cdot\underline{\mathbf{f}}\,dV = \int_V \frac{\partial f_x}{\partial x}\,dV + \int_V \frac{\partial f_y}{\partial y}\,dV + \int_V \frac{\partial f_z}{\partial z}\,dV \tag{E.45}$$

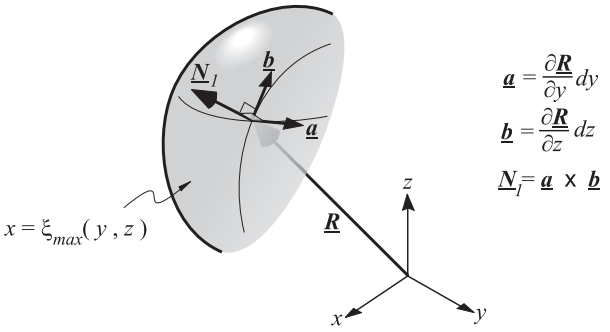


Figure E.9. The normal vector to the surface $x = \xi_{\max}(y, z)$ is given by $\underline{\mathbf{N}}_1$ which has a magnitude equal to the differential surface, dS_1.

For the first term in (E.45), we can use the following description of the volume of integration: [5]

$$\begin{aligned} V: \quad z_{\min} &\le z \le z_{\max} \\ \eta_{\min}(z) &\le y \le \eta_{\max}(z) \\ \xi_{\min}(y, z) &\le x \le \xi_{\max}(y, z) \end{aligned}$$

to obtain the following triple integral formulation

$$\int_V \frac{\partial f_x}{\partial x}\, dV = \int_{z_{\max}} \int_{\eta_{\min}(z)}^{\eta_{\max}(z)} \int_{\xi_{\min}(y,z)}^{\xi_{\max}(y,z)} \frac{\partial f_x}{\partial x}\, dx\, dy\, dz$$

After performing the inner integration with respect to x, the result is a difference of two surface integrals

$$\begin{aligned} \int_V \frac{\partial f_x}{\partial x}\, dV \quad = \quad & \int_{z_{\min}}^{z_{\max}} \int_{\eta_{\min}(z)}^{\eta_{\max}(z)} f_x(\xi_{\max}(y, z), y, z)\, dy\, dz \\ & - \int_{z_{\max}}^{z_{\max}} \int_{\eta_{\min}(z)}^{\eta_{\max}(z)} f_x(\xi_{\min}(y, z), y, z)\, dy\, dz \end{aligned} \tag{E.46}$$

The first surface integral in (E.46) is with respect to the surface: S_1: $x = \xi_{\max}(y, z)$. To determine the differential area of the surface, dS_1 at a point in the surface, we can use the position vector $\vec{\underline{\mathbf{r}}}$ of the point in surface S_1. Along the curve in the surface, in which z is fixed, we have a tangent vector given by $(\partial \vec{\underline{\mathbf{r}}}/\partial y)\, dy$. Likewise, along the curve in the surface, in which y is fixed, we have a tangent vector given by $(\partial \vec{\underline{\mathbf{r}}}/\partial z) dz$. This is shown in Figure E.9. By taking the cross product of these two tangent vectors, we obtain a vector $\underline{\mathbf{N}}_1$ which is normal to surface S_1 whose magnitude is the area of the parallelogram formed by the two tangent vectors, that is,

$$\underline{\mathbf{N}}_1 = dS_1\, \vec{\underline{\mathbf{n}}}_1$$

where $\vec{\underline{\mathbf{n}}}_1$ is the unit normal vector.

Thus, with the position vector $\vec{\underline{\mathbf{r}}}$ along the surface given by

$$\vec{\underline{\mathbf{r}}} = \xi_{\max}(y, z)\, \underline{\boldsymbol{\delta}}_x + y\, \underline{\boldsymbol{\delta}}_y + z\, \underline{\boldsymbol{\delta}}_z$$

[5] This assumes that any line that is parallel to the x axis will intersect the surface boundary of region V at two points, except at the edges of the boundary, where it touches at one point. If this assumption is not true for V, it can always be divided into subsections for which this assumption can hold. After applying the divergence theorem to these smaller regions, they can be added up later, and the resulting sum can be shown to satisfy the divergence theorem.

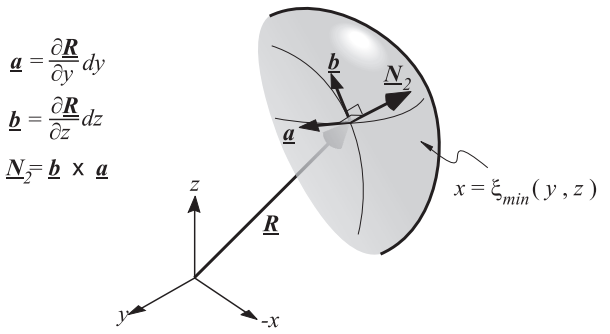


Figure E.10. The normal vector to the surface $x = \xi_{\min}(y, z)$ is given by $\underline{\mathbf{N}}_2$ which has a magnitude equal to the differential surface, dS_2.

we have

$$\begin{aligned} dS_1\, \underline{\vec{\mathbf{n}}}_1 &= \left(\frac{\partial \underline{\vec{\mathbf{r}}}}{\partial y} \times \frac{\partial \underline{\vec{\mathbf{r}}}}{\partial z}\right) dy\, dz \\ &= \left(\frac{\partial \xi_{\max}}{\partial y}\, \underline{\boldsymbol{\delta}}_x + \underline{\boldsymbol{\delta}}_y\right) \times \left(\frac{\partial \xi_{\max}}{\partial z}\, \underline{\boldsymbol{\delta}}_x + \underline{\boldsymbol{\delta}}_z\right) dy\, dz \\ &= \left(\underline{\boldsymbol{\delta}}_x - \frac{\partial \xi_{\max}}{\partial z}\, \underline{\boldsymbol{\delta}}_y - \frac{\partial \xi_{\max}}{\partial y}\, \underline{\boldsymbol{\delta}}_z\right) dy\, dz \end{aligned}$$

By taking the dot product of both sides with $\underline{\boldsymbol{\delta}}_x$,

$$(\underline{\vec{\mathbf{n}}}_1 \cdot \underline{\boldsymbol{\delta}}_x)\; dS_1 = \; dydz \tag{E.47}$$

The same arguments can be used for the other surface given by $x = \xi_{\min}(y, z)$. The difference is that, as shown in Figure E.10, the normal vector $\underline{\mathbf{N}}_2 = (\partial \underline{\vec{\mathbf{r}}}/\partial z) \times (\partial \underline{\vec{\mathbf{r}}}/\partial y)$, and thus

$$(\underline{\vec{\mathbf{n}}}_2 \cdot \underline{\boldsymbol{\delta}}_x)\; dS_2 = -\, dydz \tag{E.48}$$

Returning to equation (E.46), we can now use the results in (E.47) and (E.48) to obtain,

$$\int_V \frac{\partial f_x}{\partial x}\, dV = \int_{S_1} f_x\, (\underline{\vec{\mathbf{n}}}_1 \cdot \underline{\boldsymbol{\delta}}_x) + \int_{S_2} f_x\, (\underline{\vec{\mathbf{n}}}_2 \cdot \underline{\boldsymbol{\delta}}_x) = \int_S f_x\, \underline{\boldsymbol{\delta}}_x \cdot \underline{\vec{\mathbf{n}}} \tag{E.49}$$

Following the same procedure, we could show that the other two terms in (E.45) can be evaluated to be

$$\int_V \frac{\partial f_y}{\partial y}\, dV = \int_S f_y\, \underline{\boldsymbol{\delta}}_y \cdot \underline{\vec{\mathbf{n}}} \tag{E.50}$$

$$\int_V \frac{\partial f_z}{\partial z}\, dV = \int_S f_z\, \underline{\boldsymbol{\delta}}_z \cdot \underline{\vec{\mathbf{n}}} \tag{E.51}$$

Adding the three equations: (E.49), (E.50) and (E.51), we end up with the divergence theorem, that is,

$$\int_V \left(\frac{\partial f_x}{\partial x} + \frac{\partial f_y}{\partial y} + \frac{\partial f_y}{\partial z}\right) dV = \int_S \left(f_x\, \underline{\boldsymbol{\delta}}_x + f_y\, \underline{\boldsymbol{\delta}}_y + f_z\, \underline{\boldsymbol{\delta}}_z\right) \cdot \underline{\vec{\mathbf{n}}}\, dS \tag{E.52}$$

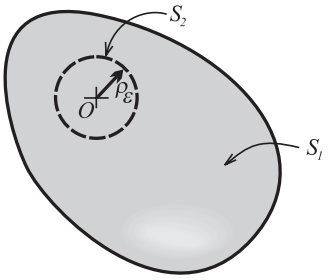


Figure E.11. A small sphere of radius ρ_ϵ removed from V, yielding surface S_1 and S_2.

E.5.3 Proof of Green's Theorem (Theorem 5.2)

First, we have

$$\begin{aligned}\nabla \cdot (\phi \nabla \psi) &= (\nabla \phi \cdot \nabla \psi) + \phi \nabla^2 \psi \\ \nabla \cdot (\psi \nabla \phi) &= (\nabla \psi \cdot \nabla \phi) + \psi \nabla^2 \phi\end{aligned}$$

Subtracting both equations,

$$\nabla \cdot (\phi \nabla \psi - \psi \nabla \phi) = \phi \nabla^2 \psi - \psi \nabla^2 \phi$$

Then taking the volume integral of both sides, and applying the divergence theorem,

$$\int_S (\phi \nabla \psi - \psi \nabla \phi) \cdot \vec{\underline{\mathbf{n}}}\, dS = \int_V \left(\phi \nabla^2 \psi - \psi \nabla^2 \phi\right)\, dV$$

E.5.4 Proof of Gauss' Theorem (Theorem 5.3)

Suppose the origin is not in the region bounded by S. Then,

$$\begin{aligned}\nabla \cdot \frac{1}{r^2} \underline{\boldsymbol{\delta}}_r &= \nabla \left(\frac{1}{r^2}\right) \cdot \underline{\boldsymbol{\delta}}_r + \frac{1}{r^2} \nabla \cdot \underline{\boldsymbol{\delta}}_r \\ &= -\frac{2}{r^3} \underline{\boldsymbol{\delta}}_r \cdot \underline{\boldsymbol{\delta}}_r + \frac{1}{r^2} \left(\frac{2}{r}\right) \\ &= 0\end{aligned}$$

Thus with the divergence theorem,

$$\int_S \left(\frac{1}{r^2}\right) \underline{\boldsymbol{\delta}}_r \cdot \vec{\underline{\mathbf{n}}}\, dS = \int_V \nabla \cdot \left(\frac{1}{r^2}\right) \underline{\boldsymbol{\delta}}_r\, dV = 0$$

Next, suppose the origin is inside S. We remove a small sphere of radius ρ_ϵ, which leaves a region having two surfaces: the original surface S_1 and a spherical surface inside given by S_2 (see Figure E.11).

The reduced volume $\widetilde{V}$ is now bounded by S_1 and S_2. Because the region in $\widetilde{V}$ satisfies the condition that the origin is not inside, we conclude that

$$\int_{\widetilde{V}} \nabla \cdot \frac{1}{r^2} \underline{\boldsymbol{\delta}}_r dV = \int_{S_1} \frac{1}{r^2} \underline{\boldsymbol{\delta}}_r \cdot \vec{\underline{\mathbf{n}}}\, dS + \int_{S_2} \frac{1}{r^2} \underline{\boldsymbol{\delta}}_r \cdot \vec{\underline{\mathbf{n}}}\, dS = 0$$

Focusing on S_2, the unit normal is given by $-\underline{\boldsymbol{\delta}}_r$, and

$$\frac{1}{r^2} \underline{\boldsymbol{\delta}}_r \cdot \vec{\underline{\mathbf{n}}} = -\frac{1}{r^2} \quad \rightarrow \quad \int_{S_2} \frac{1}{r^2} \underline{\boldsymbol{\delta}}_r \cdot \vec{\underline{\mathbf{n}}}\, dS = 4\pi$$

Thus if the origin is inside $S = S_1$,

$$\int_S \frac{1}{r^2} \underline{\boldsymbol{\delta}}_r \cdot \vec{\underline{\mathbf{n}}}\, dS = 4\pi$$

E.5.5 Proof of Stokes' Theorem (Theorem 5.4)

Let S be parameterized by u and v, then,

$$\begin{aligned}
\oint_C \underline{\mathbf{f}} \cdot d\underline{\mathbf{r}} &= \oint_C f_x dx + f_y dy + f_z dz \\
&= \oint_C f_x \left(\frac{\partial x}{\partial u} du + \frac{\partial x}{\partial v} dv \right) + \oint_C f_y \left(\frac{\partial y}{\partial u} du + \frac{\partial y}{\partial v} dv \right) \\
&\quad + \oint_C f_z \left(\frac{\partial z}{\partial u} du + \frac{\partial z}{\partial v} dv \right) \\
&= \oint_C f(u,v) du + g(u,v) dv \qquad \text{(E.53)}
\end{aligned}$$

where,

$$\begin{aligned}
f(u,v) &= f_x \frac{\partial x}{\partial u} + f_y \frac{\partial y}{\partial u} + f_z \frac{\partial z}{\partial u} \\
g(u,v) &= f_x \frac{\partial x}{\partial v} + f_y \frac{\partial y}{\partial v} + f_z \frac{\partial z}{\partial v}
\end{aligned}$$

Applying Green's lemma, (5.1), to (E.53)

$$\oint_C (f(u,v)du + g(u,v)dv) = \int_S \left(\frac{\partial g}{\partial u} - \frac{\partial f}{\partial v} \right) du\, dv \qquad \text{(E.54)}$$

The integrand of the surface integral in (E.54) can be put in terms of the curl of $\underline{\mathbf{f}}$ as follows:

$$\begin{aligned}
\frac{\partial g}{\partial u} - \frac{\partial f}{\partial v} &= \left(\frac{\partial f_x}{\partial u}\frac{\partial x}{\partial v} + f_x \frac{\partial^2 x}{\partial v \partial u} + \frac{\partial f_y}{\partial u}\frac{\partial y}{\partial v} + f_y \frac{\partial^2 y}{\partial v \partial u} + \frac{\partial f_z}{\partial u}\frac{\partial z}{\partial v} + f_z \frac{\partial^2 z}{\partial v \partial u} \right) \\
&\quad - \left(\frac{\partial f_x}{\partial v}\frac{\partial x}{\partial u} + f_x \frac{\partial^2 x}{\partial u \partial v} + \frac{\partial f_y}{\partial v}\frac{\partial y}{\partial u} + f_y \frac{\partial^2 y}{\partial u \partial v} + \frac{\partial f_z}{\partial v}\frac{\partial z}{\partial u} + f_z \frac{\partial^2 z}{\partial u \partial v} \right) \\
&= \sum_{m=x,y,z} \sum_{k=x,y,z} \frac{\partial F_k}{\partial m}\frac{\partial m}{\partial u}\frac{\partial k}{\partial v} - \sum_{m=x,y,z} \sum_{k=x,y,z} \frac{\partial F_k}{\partial m}\frac{\partial m}{\partial v}\frac{\partial k}{\partial u} \\
&= \left(\frac{\partial f_x}{\partial y}\frac{\partial(y,x)}{\partial(u,v)} + \frac{\partial f_x}{\partial z}\frac{\partial(z,x)}{\partial(u,v)} \right) + \left(\frac{\partial f_y}{\partial x}\frac{\partial(x,y)}{\partial(u,v)} + \frac{\partial f_y}{\partial z}\frac{\partial(z,y)}{\partial(u,v)} \right) \\
&\quad + \left(\frac{\partial f_z}{\partial x}\frac{\partial(x,z)}{\partial(u,v)} + \frac{\partial f_z}{\partial y}\frac{\partial(y,z)}{\partial(u,v)} \right) \\
&= \left(\frac{\partial f_y}{\partial x} - \frac{\partial f_x}{\partial y} \right) \frac{\partial(x,y)}{\partial(u,v)} + \left(\frac{\partial f_z}{\partial y} - \frac{\partial f_y}{\partial z} \right) \frac{\partial(y,z)}{\partial(u,v)} \\
&\quad + \left(\frac{\partial f_x}{\partial z} - \frac{\partial f_z}{\partial x} \right) \frac{\partial(z,x)}{\partial(u,v)} \\
&= (\nabla \times \underline{\mathbf{f}}) \cdot \left(\frac{\partial(y,z)}{\partial(u,v)} \underline{\boldsymbol{\delta}}_x + \frac{\partial(z,x)}{\partial(u,v)} \underline{\boldsymbol{\delta}}_y + \frac{\partial(x,y)}{\partial(u,v)} \underline{\boldsymbol{\delta}}_z \right) \qquad \text{(E.55)}
\end{aligned}$$

Recall that $\vec{\underline{\mathbf{n}}}\, dS$ is given by

$$\vec{\underline{\mathbf{n}}}\, dS = \left(\frac{\partial(y,z)}{\partial(u,v)}\, \underline{\boldsymbol{\delta}}_x + \frac{\partial(z,x)}{\partial(u,v)}\, \underline{\boldsymbol{\delta}}_y + \frac{\partial(x,y)}{\partial(u,v)}\, \underline{\boldsymbol{\delta}}_z \right) du\, dv \tag{E.56}$$

Combining (E.53), (E.54), (E.55) and (E.56), will yield

$$\oint_C \underline{\mathbf{f}} \cdot d\underline{\mathbf{r}} = \int_S (\nabla \times \underline{\mathbf{f}}) \cdot \vec{\underline{\mathbf{n}}}\, dS$$

which is Stokes' theorem.

E.5.6 Proof of Leibnitz formulas

1. **One-Dimensional Case (Theorem 5.6).** Using the definition of a derivative:

$$\frac{d}{d\alpha}\left(\int_{g(\alpha)}^{h(\alpha)} F(\alpha, x)\, dx \right) = \lim_{\Delta\alpha \to 0} \frac{1}{\Delta\alpha} \left(\int_{g(\alpha+\Delta\alpha)}^{h(\alpha+\Delta\alpha)} F(\alpha+\Delta\alpha, x) dx - \int_{g(\alpha)}^{h(\alpha)} F(\alpha, x) dx \right) \tag{E.57}$$

The first integral in the left-hand side of (E.57) can be divided into three parts,

$$\int_{g(\alpha+\Delta\alpha)}^{h(\alpha+\Delta\alpha)} F(\alpha+\Delta\alpha, x)\, dx = \int_{h(\alpha)}^{h(\alpha+\Delta\alpha)} F(\alpha+\Delta\alpha, x)\, dx + \int_{g(\alpha)}^{h(\alpha)} F(\alpha+\Delta\alpha, x)\, dx + \int_{g(\alpha+\Delta\alpha)}^{g(\alpha)} F(\alpha+\Delta\alpha, x)\, dx \tag{E.58}$$

Furthermore, the first integral in the left-hand side of (E.58) can be approximated by the trapezoidal rule,

$$\int_{h(\alpha)}^{h(\alpha+\Delta\alpha)} F(\alpha+\Delta\alpha, x)\, dx \approx \frac{1}{2}\left[F\left(\alpha+\Delta\alpha, h_{(\alpha+\Delta\alpha)}\right) + F\left(\alpha+\Delta\alpha, h_{(\alpha)}\right) \right] \left(h_{(\alpha+\Delta\alpha)} - h_{(\alpha)} \right) \tag{E.59}$$

Likewise, we can also approximate the third integral in the left-hand side of (E.58) as

$$\int_{g(\alpha+\Delta\alpha)}^{g(\alpha)} F(\alpha+\Delta\alpha, x)\, dx \approx \frac{1}{2}\left[F\left(\alpha+\Delta\alpha, g_{(\alpha+\Delta\alpha)}\right) + F\left(\alpha+\Delta\alpha, g_{(\alpha)}\right) \right] \left(g_{(\alpha)} - g_{(\alpha+\Delta\alpha)} \right) \tag{E.60}$$

Substituting (E.59) and (E.60) into (E.58), and then into (E.57),

$$\begin{aligned}\frac{d}{d\alpha}\int_{g(\alpha)}^{h(\alpha)} F(\alpha,x)\,dx = \lim_{\Delta\alpha\to 0}\Bigg[&\int_{g(\alpha))}^{h(\alpha)}\left(\frac{F(\alpha+\Delta\alpha,x)-F(\alpha,x)}{\Delta\alpha}\right)dx\\ &+\frac{F\left(\alpha+\Delta\alpha,h_{(\alpha+\Delta\alpha)}\right)+F\left(\alpha+\Delta\alpha,h_{(\alpha)}\right)}{2\Delta\alpha}\left(h_{(\alpha+\Delta\alpha)}-h_{(\alpha)}\right)\\ &+\frac{F\left(\alpha+\Delta\alpha,g_{(\alpha+\Delta\alpha)}\right)+F\left(\alpha+\Delta\alpha,g_{(\alpha)}\right)}{2\Delta\alpha}\left(g_{(\alpha)}-g_{(\alpha+\Delta\alpha)}\right)\Bigg]\\ =\int_{g(\alpha))}^{h(\alpha)}&\frac{\partial}{\partial\alpha}F(\alpha,x)\,dx+F(\alpha,h(\alpha))\frac{dh}{d\alpha}-F(\alpha,g(\alpha))\frac{dg}{d\alpha}\end{aligned}$$

2. **Three-Dimensional Case (Theorem 5.7).** From the definition of the derivative,

$$\begin{aligned}&\frac{d}{d\alpha}\int_{V(\alpha)} f(x,y,z,\alpha)dV\\ &=\lim_{\Delta\alpha\to 0}\left[\frac{1}{\Delta\alpha}\int_{V(\alpha+\Delta\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV-\int_{V(\alpha)} f(x,y,z,\alpha)dV\right]\end{aligned}$$

By adding and subtracting the term $\int_{V(\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV$ in the right-hand side,

$$\begin{aligned}&\frac{d}{d\alpha}\int_{V(\alpha)} f(x,y,z,\alpha)dV\\ &=\lim_{\Delta\alpha\to 0}\frac{1}{\Delta\alpha}\left[\int_{V(\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV-\int_{V(\alpha)} f(x,y,z,\alpha)dV\right]\\ &\quad+\lim_{\Delta\alpha\to 0}\frac{1}{\Delta\alpha}\left[\int_{V(\alpha+\Delta\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV-\int_{V(\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV\right]\\ &=\int_{V(\alpha)}\frac{\partial f}{\partial\alpha}dV+\lim_{\Delta\alpha\to 0}\frac{1}{\Delta\alpha}\left[\int_{V(\alpha+\Delta\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV\right.\\ &\qquad\left.-\int_{V(\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV\right]\end{aligned}\tag{E.61}$$

The last group of terms in the right-hand side (E.61) is the difference of two volume integrals involving the same integrand. We can combine these integrals by changing the volume of integration to be the region between $V(\alpha+\Delta\alpha)$ and $V(\alpha)$.

$$\begin{aligned}&\int_{V(\alpha+\Delta\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV-\int_{V(\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV=\\ &\qquad\int_{V(\alpha+\Delta\alpha)-V(\alpha)} f(x,y,z,\alpha+\Delta\alpha)dV\end{aligned}\tag{E.62}$$

We could approximate the differential volume in (E.62) as the parallelepiped formed by the vectors $(\partial\vec{\mathbf{r}}/\partial u)du$, $(\partial\vec{\mathbf{r}}/\partial v)dv$ and $(\partial\vec{\mathbf{r}}/\partial\alpha)d\alpha$, where u and v are parameters used to describe surface $S(\alpha)$. This is shown in Figure E.12.

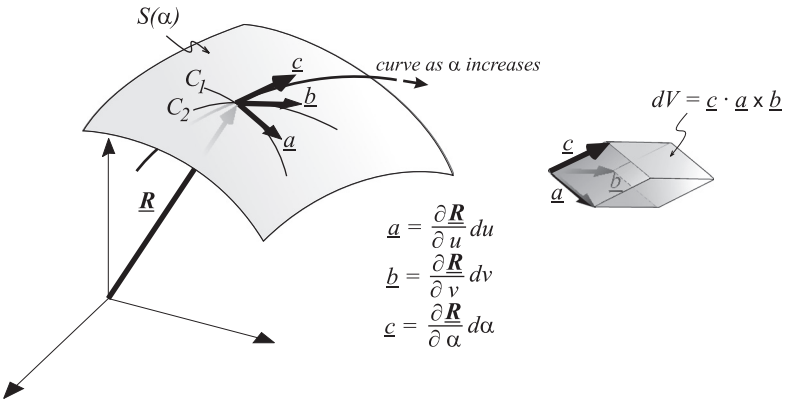


Figure E.12. Graphical representation of differential volume emanating from points in $S(\alpha)$ towards $S(\alpha + \Delta\alpha)$.

Recall that

$$\left(\frac{\partial \vec{\mathbf{r}}}{\partial u}du\right) \times \left(\frac{\partial \vec{\mathbf{r}}}{\partial v}dv\right) = \vec{\mathbf{n}}\, dS$$

which then gives a differential volume attached to $S(\alpha)$

$$\begin{aligned} dV|_{(x,y,z)\in V(\alpha+\Delta\alpha)-V(\alpha)} &= \frac{\partial \vec{\mathbf{r}}}{\partial \alpha} \cdot \left(\frac{\partial \vec{\mathbf{r}}}{\partial u} \times \frac{\partial \vec{\mathbf{r}}}{\partial u}\right) d\alpha\, du\, dv \\ &= \frac{\partial \vec{\mathbf{r}}}{\partial \alpha} \cdot \vec{\mathbf{n}}\, d\alpha\, dS \end{aligned}$$

The volume integral for points bounded between the surfaces of $V(\alpha)$ and $V(\alpha + \Delta\alpha)$ can now be approximated as follows:

$$\int_{V_{(\alpha+\Delta\alpha)}-V_{(\alpha)}} f(x, y, z, \alpha + \Delta\alpha)dV \approx \int_{S(\alpha)} f(x, y, z, \alpha + \Delta\alpha)\frac{\partial \vec{\mathbf{r}}}{\partial \alpha} \cdot \vec{\mathbf{n}}\, \Delta\alpha\, dS \tag{E.63}$$

Substituting (E.63) into (E.62) and then to (E.61),

$$\begin{aligned} \frac{d}{d\alpha}\int_{V(\alpha)} f(x, y, z, \alpha)dV &= \int_{V(\alpha)} \frac{\partial f}{\partial \alpha}dV \\ &\quad + \lim_{\Delta\alpha\to 0} \frac{1}{\Delta\alpha}\int_{S(\alpha)} f(x, y, z, \alpha + \Delta\alpha)\frac{\partial \vec{\mathbf{r}}}{\partial \alpha} \cdot \vec{\mathbf{n}}\Delta\alpha dS \\ &= \int_{V(\alpha)} \frac{\partial f}{\partial \alpha}dV + \int_{S(\alpha)} f(x, y, z, \alpha)\frac{\partial \vec{\mathbf{r}}}{\partial \alpha} \cdot \vec{\mathbf{n}}\, dS \end{aligned}$$

which is the Leibnitz rule for differentiation of volume integrals.

APPENDIX F

Additional Details and Fortification for Chapter 6

F.1 Supplemental Methods for Solving First-Order ODEs

F.1.1 General Ricatti Equation

In some cases, the solution of a first-order differential equation is aided by increasing the order to a second-order differential equation. One such case is the **generalized Ricatti differential equation** given by the following general form:

$$\frac{dy}{dx} = P(x)y^2 + Q(x)y + R(x) \tag{F.1}$$

Note that when $P(x) = 0$, we have a first-order linear differential equation, and when $R(x) = 0$, we have the Bernouli differential equation.

Using a method known as the **Ricatti transformation**,

$$y(x) = -\frac{1}{P(x)w}\frac{dw}{dx}$$

we obtain

$$\begin{aligned}
\frac{dy}{dx} &= -\frac{1}{Pw}\frac{d^2w}{dx^2} + \frac{1}{Pw^2}\left(\frac{dw}{dx}\right)^2 + \frac{1}{P^2w}\frac{dP}{dx}\left(\frac{dw}{dx}\right) \\
Py^2 &= \frac{1}{Pw^2}\left(\frac{dw}{dx}\right)^2 \\
Qy &= -\frac{Q}{Pw}\frac{dw}{dx}
\end{aligned}$$

which then reduces (F.1) to be

$$\frac{d^2w}{dx^2} - \left(\frac{dP(x)/dx}{P(x)} + Q(x)\right)\frac{dw}{dx} + P(x)R(x)w = 0 \tag{F.2}$$

Note that (F.2) is a second-order ordinary differential equation. Nonetheless, it is a linear differential equation, which is often easier to solve than the original nonlinear first-order equation.

EXAMPLE F.1. Consider the following differential equation:

$$\frac{dy}{dt} = xy^2 - \frac{2}{x}y - \frac{1}{x^3}$$

Noting that $P(x) = x$, $Q(x) = -2/x$ and $R(x) = -1/x^3$, the Ricatti transformation $y = -(dw/dx)/(xw)$ converts it to a linear second-order differential equation given by

$$x^2\frac{d^2w}{dx^2} + x\frac{dw}{dx} - w = 0$$

which is a Euler-Cauchy equation (cf. Section 6.4.3). Thus we need another transformation $z = \ln(x)$, which would transform the differential equation further to be

$$\frac{d^2w}{dz^2} = w$$

whose solution becomes,

$$w(z) = Ae^{-z} + Be^{z} \quad \rightarrow \quad w(x) = A\frac{1}{x} + Bx$$

Putting it back in terms of y,

$$y = -\frac{1}{xw}\frac{dw}{dx} = \frac{\frac{A}{x^2} - B}{x\left(\frac{A}{x} + Bx\right)} = \frac{1}{x^2}\frac{C - x^2}{C + x^2}$$

where $C = A/B$ is an arbitrary constant.

F.1.2 Legendre Transformations

Usually, methods that introduce a change of variables involve only transformations from the original independent and dependent variables, say, x and y, to new independent and dependent variables, say, p and q. In some cases, however, we need to consider the derivatives as separate variables in the transformations, for example,

$$\begin{aligned} p &= p\left(x, y, \frac{dy}{dx}\right) \\ q &= q\left(x, y, \frac{dy}{dx}\right) \\ \frac{dq}{dp} &= \frac{dq}{dp}\left(x, y, \frac{dy}{dx}\right) \end{aligned} \tag{F.3}$$

These types of transformations are called **contact transformations**.

One particular type of contact transformation is the **Legendre transformation**. This type of transformation is very useful in the field of thermodynamics for obtaining equations in which the roles of intensive and extensive variables need to be switched in a way that conserves the information content of the original fundamental equations. In the case here, the Legendre transformation is used to solve differential equations.

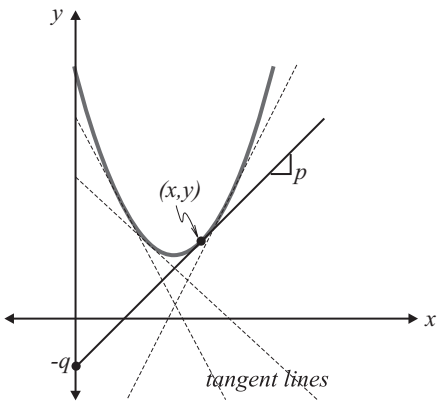


Figure F.1. Description of a curve as an envelope of tangent lines used for Legendre transformation rules.

The Legendre transformation takes a curve $y = y(x)$ and obtains an equivalent description by using an envelope generated by a family of tangent lines to the curve at the point (x, y), that is,

$$y = p\ x + (-q) \tag{F.4}$$

where p is the slope and $-q$ is the y-intercept. This is illustrated in Figure F.1.

The Legendre transformation uses the following transformations:

$$p = \frac{dy}{dx}\ ; \quad q = x\frac{dy}{dx} - y \quad \text{and} \quad \frac{dq}{dp} = x \tag{F.5}$$

where p is the new independent variable and q is the new dependent variable. The inverse Legendre transformations are given by

$$x = \frac{dq}{dp}\ ; \quad y = p\,\frac{dq}{dp} - q \quad \text{and} \quad \frac{dy}{dx} = p \tag{F.6}$$

Now consider the differential equation

$$f\left(x, y, \frac{dy}{dx}\right) = 0 \tag{F.7}$$

In terms of the new variables, we have

$$f\left(\frac{dq}{dp}, p\,\frac{dq}{dp} - q, p\right) \tag{F.8}$$

It is hoped that (F.8) will be easier to solve than (F.7), such as when the derivative dy/dx appears in nonlinear form while x and y are in linear or affine forms. If this is the case, one should be able to solve (F.8) to yield a solution of the form given by: $S(p, q) = 0$. To return to the original variables, we observe that

$$\frac{\partial S}{\partial p} + \left(\frac{\partial S}{\partial q}\right)\frac{dq}{dp} = 0 \rightarrow g(p, xp - y) + h(p, xp - y)x = 0$$

where g and h are functions resulting from the partial derivatives. Together with $f(x, y, p) = 0$, one needs to remove the presence of p to obtain a general solution $s(x, y) = 0$. In some cases, if this is not possible, p would have to be left as a parameter, and the solution will be given by curves described by $x = x(p)$ and $y = y(p)$.

In particular, Legendre transformations can be applied to a differential equations given by

$$y = x\psi(p) + \eta(p) \tag{F.9}$$

where $\psi(p) \neq p$.[1] For instance, one may have a situation in which the dependent variable y is modeled empirically as a function of $p = dy/dx$ in the form given by (F.9). After using the Legendre transformation, we arrive at

$$\frac{dq}{dp} + \left(\frac{1}{\psi(p) - p}\right) q = \left(\frac{\eta(p)}{p - \psi(p)}\right)$$

which is a linear differential equation in variables p and q.

EXAMPLE F.2. Consider the differential equation given by

$$\left(\frac{dy}{dx}\right)^2 = x\frac{dy}{dx} + y$$

then after the Legendre transformation, we obtain

$$\frac{dq}{dp} - \frac{1}{2p}q = \frac{p}{2}$$

whose solution is given by

$$q = \frac{p^2}{3} + C\sqrt{p}$$

After taking the derivative dq/dp,

$$\frac{dq}{dp} = x = \frac{2}{3}p + \frac{C}{2}\sqrt{p}$$

Unfortunately, $p(x)$ is not easily found. Instead, we could treat p as a parameter, that is, $x = x(p)$, and insert this back to the given equation to obtain

$$y = -x(\alpha)\,\alpha + \alpha^2\,;\ \text{subject to } x(\alpha) = \frac{2}{3}\alpha + \frac{C}{2}\sqrt{\alpha}$$

where α is a parameter for the solution $(y(\alpha), x(\alpha))$, and C is an arbitrary constant.

F.2 Singular Solutions

For some differential equations, a solution may exist that does not have arbitrary constants of integration. These solutions are called **singular solutions**. Singular solutions, if they exist for a differential equation, have a special property that it is the envelope of the general solutions. Thus their utility is often in the determination of the bounds of the solution domain.

For a first-order differential equation,

$$f\left(x, y, \frac{dy}{dx}\right) = 0 \tag{F.10}$$

the general solution is given by

$$\phi(x, y, C) = 0 \tag{F.11}$$

[1] If $\psi(p) = p$, an algebraic equation results, that is, $q = -\eta(p)$.

where C is an arbitrary constant. For ϕ to be a singular solution, it should not be a function of the arbitrary constant C. Thus

$$\frac{\partial \phi}{\partial C} = S(x, y) = 0 \tag{F.12}$$

where $S(x, y)$ is obtained with the aid of (F.11). To determine whether $S(x, y) = 0$ is indeed a singular solution, one needs to check if

$$\frac{\partial S}{\partial x} + \frac{\partial S}{\partial y}\frac{dy}{dx} = 0 \tag{F.13}$$

will satisfy the original differential equation (F.10). If it does, then it is a singular solution.

EXAMPLE F.3. **Clairaut's equation** is given by

$$y = x\frac{dy}{dx} + \left(\frac{dy}{dx}\right)^2 \tag{F.14}$$

Using the quadratic equation to find dy/dx as an explicit function of x and y, this can be rearranged to give

$$\frac{dy}{dx} = \frac{x}{2}\left(-1 \pm \sqrt{1 + 4\frac{y}{x^2}}\right)$$

which is an isobaric equation (cf. (6.13)). By introducing a new variable, $u = y/x^2$, the original differential equation can be reduced to a separable equation, that is,

$$\frac{du}{4u + 1 \pm \sqrt{1 + 4u}} = -\frac{1}{2}\frac{dx}{x}$$

whose solution is given by

$$\ln(\sqrt{4u + 1} \pm 1) = -\ln(x) + k \quad \rightarrow \quad y = Cx + C^2$$

where C is an arbitrary constant.

To search for the singular function, following (F.11) yields

$$\phi(x, y, C) = y - Cx - C^2 = 0 \tag{F.15}$$

then

$$\frac{\partial \phi}{\partial C} = -x - 2C = 0 \tag{F.16}$$

where C can be eliminated from (F.16) using (F.15) to obtain

$$S(x, y) = \pm\sqrt{x^2 + 4y} = 0 \quad \rightarrow \quad y = -\frac{x^2}{4} \tag{F.17}$$

Finally, we can check that (F.17) satisfies (F.14), thus showing that (F.17) is indeed a singular solution of (F.14).

A simpler alternative approach to solving Clairaut's equation is to take the derivative of (F.14) with respect to x while letting $p = dy/dx$, that is,

$$\begin{aligned} p &= p + x\frac{dp}{dx} + 2p\frac{dp}{dx} \\ 0 &= \frac{dp}{dx}(x + 2p) \end{aligned}$$

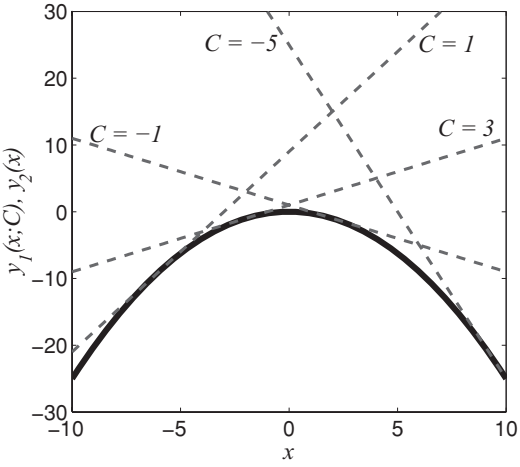


Figure F.2. Plot of the general solution $y_1(x) = Cx + C^2$ (dotted lines) and the singular solution $y_2(x) = -x^2/4$ (solid curve).

then

$$\frac{dp}{dx} = 0 \qquad \text{and} \qquad p = -\frac{x}{2}$$

yielding two solutions of different forms

$$y_1 = c_1 x + c_2 \quad \text{and} \quad y_2 = -\frac{x^2}{4} + c_3$$

Substituting both solutions into (F.14) will result in $c_3 = 0$ and $c_2 = c_1^2$. Thus the general solution is given by

$$y_1 = cx + c^2$$

while the singular solution is given by

$$y_2 = -\frac{x^2}{4}$$

If we plot the general solution $y_1(x) = cx + c^2$ and the singular solution, $y_2 = -x^2/4$, we see in Figure F.2 that the singular solution is an envelope for the general solution.

F.3 Finite Series Solution of $d\mathbf{x}/dt = A\mathbf{x} + \mathbf{b}(t)$

The method shown here solves the linear equation with constant coefficient A given by

$$\frac{d}{dt}\mathbf{x} = A\mathbf{x} + \mathbf{b}(t)$$

subject to $\mathbf{x}(0) = \mathbf{x}_0$. It is applicable also to matrices A that are not diagonalizable. The steps of the procedure are given as follows:

1. Let the vector of eigenvalues of $A[=]n \times n$ be grouped into p distinct sets of repeated eigenvalues, that is,

$$\underline{\widehat{\lambda}} = \left(\underline{\lambda}_1 \;\middle|\; \cdots \;\middle|\; \underline{\lambda}_p\right) \qquad \text{with} \qquad \underline{\lambda_i} = \begin{pmatrix}\lambda_i & \cdots & \lambda_i\end{pmatrix}\,[=]\,1 \times m_i$$

where $\lambda_i \neq \lambda_k$ when $i \neq k$, and $\sum_{i=1}^{p} m_i = n$.

2. Next, define the matrix $Q[=]n \times n$,

$$Q = \left(\begin{array}{c} Q_1 \\ \hline \vdots \\ \hline Q_p \end{array}\right) \quad Q_i = \begin{pmatrix} q_{0,0}(\lambda_i) & \cdots & q_{0,n-1}(\lambda_i) \\ \vdots & \ddots & \vdots \\ q_{m_i-1,0}(\lambda_i) & \cdots & q_{m_i-1,n-1}(\lambda_i) \end{pmatrix} [=] m_i \times n \quad \text{(F.18)}$$

where,

$$q_{j,\ell}(\lambda) = \begin{cases} 0 & \text{if} \quad \ell < j \\ \dfrac{\ell!}{(\ell-j)!}\lambda^{(\ell-j)} & \end{cases}$$

and define the vector $\mathbf{g}(t)[=]n \times 1$ as

$$\mathbf{g}(t) = \left(\begin{array}{c} \mathbf{g}_1 \\ \hline \vdots \\ \hline \mathbf{g}_p \end{array}\right) \quad \mathbf{g}_i = \begin{pmatrix} 1 \\ t \\ \vdots \\ t^{m_i-1} \end{pmatrix} e^{\lambda_i t} [=] m_i \times 1 \quad \text{(F.19)}$$

3. Combining the results, we have

$$\mathbf{v}(t) = \begin{pmatrix} c_0 \\ \vdots \\ c_{n-1}t^{n-1} \end{pmatrix} = Q^{-1}\mathbf{g}(t)$$

where Q, as given in (F.18), is a matrix of constants. The matrix exponential is then given by

$$e^{At} = \sum_{\ell=0}^{n-1} v_{\ell+1}(t)A^{\ell} \quad \text{(F.20)}$$

We can now apply (F.20) to solve the general linear equation,

$$\frac{d}{dt}\mathbf{x} = A\mathbf{x} + \mathbf{b}(t) \qquad A \text{ is constant}$$

In terms of Q and $\mathbf{g}(t)$ given in (F.18) and (F.19), respectively, we have

$$\mathbf{x}(t) = H_1\,\mathbf{g}(t) \;+\; H_2\,\mathbf{w}(t) \quad \text{(F.21)}$$

where

$$\begin{aligned} H_1 &= \left(\mathbf{x}_0 \;\middle|\; A\mathbf{x}_0 \;\middle|\; \cdots \;\middle|\; A^{n-1}\mathbf{x}_0\right) Q^{-1} \\ H_2 &= \left(I_{[n]} \;\middle|\; A \;\middle|\; \cdots \;\middle|\; A^{n-1}\right)\left(Q^{-1} \otimes I_{[n]}\right) \\ \mathbf{w}(t) &= \int_0^t \mathbf{g}\left(t-\tau\right) \otimes \mathbf{b}\left(\tau\right) d\tau \;\; [=] \;\; n^2 \times 1 \end{aligned}$$

The advantage of (F.21) is the clear separation of constant matrices H_1 and H_2 from $\mathbf{g}(t)$ and $\mathbf{w}(t)$, respectively.[2] This allows for the evaluation of integrals given in each element of $\mathbf{w}(t)$ one term at a time. For instance, one could use the following

[2] The span of the columns of H_1 is also known as the **Krylov subspace** of A based on $\mathbf{x}_0$.

convolution formula for the special case of $b_i = e^{\sigma t}$:

$$\int_0^t (t-\tau)^m e^{\lambda(t-\tau)} e^{\sigma\tau} d\tau = \begin{cases} \dfrac{m!}{(\sigma-\lambda)^{m+1}} \left(e^{\sigma t} - e^{\lambda t} \sum_{k=0}^{m} \dfrac{1}{k!} (\sigma-\lambda)^k t^k \right) & \text{if } \lambda \neq \sigma \\ \dfrac{t^{m+1}}{m+1} e^{\lambda t} & \text{if } \lambda = \sigma \end{cases} \tag{F.22}$$

Because Q is formed by the eigenvalues of A, both H_1 and H_2 are completely determined by A and $\mathbf{x}_0$ alone, that is, they are both independent of $\mathbf{b}(t)$. A MATLAB function `linode_mats.m` is available on the book's webpage. This function can be used to evaluate H_1 and H_2, with A, $\mathbf{x}_0$ as inputs.

EXAMPLE F.4. Consider the linear system

$$\frac{d}{dt}\mathbf{x} = \begin{pmatrix} -2 & 0 & 0 \\ -\frac{1}{2} & -2 & -\frac{1}{2} \\ 1 & 2 & 0 \end{pmatrix} \mathbf{x} + \begin{pmatrix} 1 \\ 2 \\ -e^{-3t} \end{pmatrix} \quad ; \quad \mathbf{x}(0) = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

The eigenvalues of A are $\{-2, -1, -1\}$. Then, we can evaluate Q and $\mathbf{g}$ as

$$Q = \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 0 & 1 & -2 \end{pmatrix} \quad ; \quad \mathbf{g} = \begin{pmatrix} e^{-2t} \\ e^{-t} \\ te^{-t} \end{pmatrix}$$

Matrices H_1 and H_2 are

$$H_1 = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ -1 & 2 & 1 \end{pmatrix} \quad H_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & -\frac{1}{2} & 1 & 0 & 0 & -1 & -\frac{1}{2} \\ -1 & 0 & 0 & 1 & 0 & 1 & 0 & 2 & 1 \end{pmatrix}$$

and the integrals can be evaluated to be

$$\mathbf{w}(t) = \int_0^t \mathbf{g}\left(t-\tau\right) \otimes \mathbf{b}\left(\tau\right) d\tau = \begin{pmatrix} \left(1-e^{-2t}\right)/2 \\ 1-e^{-2t} \\ e^{-3t}-e^{-2t} \\ 1-e^{-t} \\ 2-2e^{-t} \\ \left(e^{-3t}-e^{-t}\right)/2 \\ 1-(1+t)\,e^{-t} \\ 2-2\,(1+t)\,e^{-t} \\ \left(-e^{-3t}+(1-2t)\,e^{-t}\right)/4 \end{pmatrix}$$

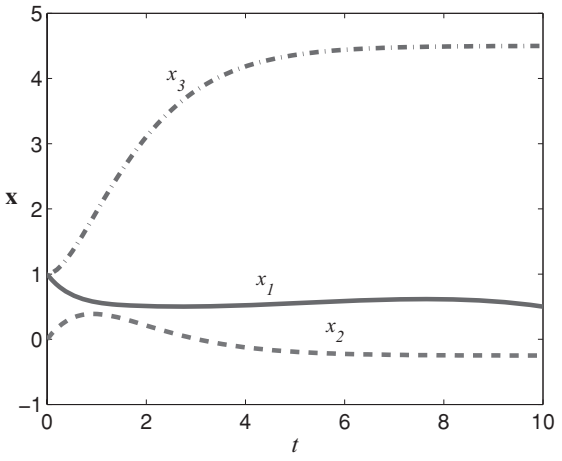


Figure F.3. A plot of the solution of the system given in Example F.4.

Note that $\mathbf{w}(0) = \mathbf{0}$ as it should, because the initial conditions are contained only in H_1. Furthermore, note that columns 2, 3, and 7 in H_2 are all zero, which implies that the corresponding elements in $\mathbf{w}(t)$ are not relevant to the solution of $\mathbf{x}(t)$.

Combining the results using (F.21), a plot of the solutions is shown in Figure F.3.

F.4 Proof for Lemmas and Theorems in Chapter 6

F.4.1 Proof of Theorem 6.1: Similarity Transformations Yield Separable-Variables Forms

Using the conditions of symmetry,

$$\lambda^{\beta-\alpha} F(x, y) = F\left(\lambda^{\alpha} x, \lambda^{\beta} y\right) \tag{F.23}$$

where $F(x, y) = -M(x, y)/N(x, y)$. Taking the partial derivative of (F.23) with respect to λ,

$$(\beta - \alpha)\,\lambda^{\beta-\alpha-1} F(x, y) = \alpha\lambda^{\alpha-1} x \frac{\partial F\left(\lambda^{\alpha} x, \lambda^{\beta} y\right)}{\partial\left(\lambda^{\alpha} x\right)} + \beta\lambda^{\beta-1} y \frac{\partial F\left(\lambda^{\alpha} x, \lambda^{\beta} y\right)}{\partial\left(\lambda^{\beta} y\right)}$$

Next, fix $\lambda = 1$ to obtain

$$\alpha x \frac{\partial F}{\partial x} + \beta y \frac{\partial F}{\partial y} = (\beta - \alpha) F$$

which is a linear first-order partial differential equation that is solvable using the method of characteristics.[3] The characteristic equations are given by,

$$\frac{dx}{\alpha x} = \frac{dy}{\beta y} = \frac{dF}{(\beta - \alpha) F}$$

[3] See Section 10.1 for details on the method of characteristics.

which yield two invariants, ϕ_1 and ϕ_2, given by

$$\phi_1 = \frac{y^\alpha}{x^\beta} = u \qquad \text{and} \qquad \phi_2 = \frac{F}{x^{(\beta-\alpha)/\alpha}}$$

from which general solution is obtained as

$$\phi_2 = G(\phi_1) \qquad \rightarrow \qquad F = \frac{dy}{dx} = x^{(\beta-\alpha)/\alpha} G(u)$$

Taking the derivative of u with respect to x,

$$\frac{du}{dx} = \frac{d}{dx}\left(\frac{y^\alpha}{x^\beta}\right) = \alpha\frac{u}{y}\frac{dy}{dx} - \beta\frac{u}{x} = \alpha\frac{u}{y}x^{(\beta-\alpha)/\alpha}G(u) - \beta\frac{u}{x}$$

and with $y = u^{1/\alpha}x^{\beta/\alpha}$,

$$\frac{du}{dx} = \left(\alpha u^{(\alpha-1)/\alpha}G(u) - \beta u\right)\frac{1}{x}$$

F.4.2 Proof of Similarity Reduction of Second-Order Equations, Theorem 6.2

Using the similarity transformations,

$$\begin{aligned}
\frac{d^2\widetilde{y}}{d^2\widetilde{x}} &= f\left(\widetilde{x}, \widetilde{y}, \frac{d\widetilde{y}}{d\widetilde{x}}\right) \\
\lambda^{\beta-2}\frac{d^2y}{dx^2} &= f\left(\lambda x, \lambda^\beta y \lambda^{\beta-1}\frac{dy}{dx}\right) \\
\lambda^{\beta-2} f\left(x, y, \frac{dy}{dx}\right) &= f\left(\lambda x, \lambda^\beta y, \lambda^{\beta-1}\frac{dy}{dx}\right)
\end{aligned}$$

Next, we take the partial derivative of this equation with respect to λ and then set $\lambda = 1$,

$$x\frac{\partial f}{\partial x} + \beta y\frac{\partial f}{\partial y} + (\beta-1)\left(\frac{dy}{dx}\right)\frac{\partial f}{\partial(dy/dx)} = (\beta-2)\,f$$

which is a linear first-order partial differential equation. The characteristic equations[4] are given by

$$\frac{dx}{x} = \frac{dy}{\beta y} = \frac{d(dy/dx)}{(\beta-1)\,(dy/dx)} = \frac{df}{(\beta-2)\,f}$$

which yields three invariants, ϕ_1, ϕ_2, and ϕ_3,

$$\begin{aligned}
\phi_1 &= \frac{y}{x^\beta} = u \\
\phi_2 &= \frac{(dy/dx)}{x^{\beta-1}} = v \\
\phi_3 &= \frac{f}{x^{\beta-2}}
\end{aligned}$$

[4] See Section 10.1 for details on the method of characteristics.

The general solution for the partial differential equation is then given by

$$\phi_3 = G(\phi_1, \phi_2) \quad \rightarrow \quad \frac{f}{x^{\beta-2}} = G(u, v)$$

We can evaluate the derivatives of u and v,

$$\begin{aligned} x\frac{du}{dx} &= v - \beta u \\ x\frac{dv}{dx} &= \frac{d^2y}{dx^2}\frac{1}{x^{\beta-2}} + (1-\beta)\,v = \frac{f}{x^{\beta-2}} + (1-\beta)\,v \\ &= G(u, v) + (1-\beta)\,v \end{aligned}$$

Dividing the last equation by the one before it,

$$\frac{dv}{du} = \frac{G(u, v) + (1-\beta)\,v}{v - \beta u}$$

F.4.3 Proof of Properties of Exponentials, Theorem 6.3

Let matrices G_i and H_j be defined as

$$G_i = \frac{t^i}{i!}\mathbf{A}^i, \qquad H_j = \frac{s^j}{j!}\mathbf{A}^j$$

then $e^{\mathbf{A}t}$ and $e^{\mathbf{A}s}$ can be expanded to be

$$\begin{aligned} e^{\mathbf{A}t} &= G_0 + G_1 + G_2 + G_3 + \cdots \\ e^{\mathbf{A}s} &= H_0 + H_1 + H_2 + H_3 + \cdots \end{aligned}$$

taking the matrix product,

$$\begin{aligned} e^{\mathbf{A}t}e^{\mathbf{A}s} &= (G_0 + G_1 + G_2 + G_3 + \cdots)(H_0 + H_1 + H_2 + H_3 + \cdots) \\ &= G_0H_0 + G_1H_0 + G_2H_0 + G_3H_0 + \cdots \\ &\quad + G_0H_1 + G_1H_1 + G_2H_1 + G_3H_1 + \cdots \\ &\quad + G_0H_2 + G_1H_2 + G_2H_2 + G_3H_2 + \cdots \\ &\quad + \cdots \\ &= Q_0 + Q_1 + Q_2 + \cdots \end{aligned}$$

where

$$\begin{aligned} Q_k &= \sum_{i=0}^{k} G_iH_{k-i} \\ &= \sum_{i=0}^{k}\left(\frac{t^i}{i!}\mathbf{A}^i\right)\left(\frac{s^{k-i}}{(k-i)!}\mathbf{A}^{k-i}\right) \\ &= \frac{1}{k!}\mathbf{A}^k\sum_{i=0}^{k}\frac{k!}{i!\,(k-i)!}t^is^{k-i} \\ &= \frac{1}{k!}\mathbf{A}^k(s+t)^k \end{aligned}$$

Thus

$$\begin{aligned} e^{\mathbf{A}t}e^{\mathbf{A}s} &= I + (s+t)\mathbf{A} + \frac{(s+t)^2}{2!}\mathbf{A}^2 + \cdots \\ &= e^{\mathbf{A}(s+t)} \end{aligned}$$

which proves (6.51). Note also that matrices $e^{\mathbf{A}t}$ and $e^{\mathbf{A}s}$ commute.
By letting $s = -t$,

$$e^{\mathbf{A}t}e^{-\mathbf{A}t} = e^{-\mathbf{A}t}e^{\mathbf{A}t} = I$$

Thus $e^{-\mathbf{A}t}$ is the inverse of $e^{\mathbf{A}t}$
Now let matrices Ω_i and Ψ_j be defined as

$$\Omega_i = \frac{t^i}{i!}\mathbf{A}^i, \qquad \Psi_j = \frac{t^j}{j!}\mathbf{W}^j$$

then $e^{\mathbf{A}t}$ and $e^{\mathbf{W}t}$ can be expanded to be

$$\begin{aligned} e^{\mathbf{A}t} &= \Omega_0 + \Omega_1 + \Omega_2 + \Omega_3 + \cdots \\ e^{\mathbf{W}t} &= \Psi_0 + \Psi_1 + \Psi_2 + \Psi_3 + \cdots \end{aligned}$$

taking the matrix product,

$$\begin{aligned} e^{\mathbf{A}t}e^{\mathbf{W}t} &= (\Omega_0 + \Omega_1 + \Omega_2 + \Omega_3 + \cdots)(\Psi_0 + \Psi_1 + \Psi_2 + \Psi_3 + \cdots) \\ &= \Omega_0\Psi_0 + \Omega_1\Psi_0 + \Omega_2\Psi_0 + \Omega_3\Psi_0 + \cdots \\ &\quad + \Omega_0\Psi_1 + \Omega_1\Psi_1 + \Omega_2\Psi_1 + \Omega_3\Psi_1 + \cdots \\ &\quad + \Omega_0\Psi_2 + \Omega_1\Psi_2 + \Omega_2\Psi_2 + \Omega_3\Psi_2 + \cdots \\ &\quad + \cdots \\ &= R_0 + R_1 + R_2 + \cdots \end{aligned}$$

where

$$\begin{aligned} R_k &= \sum_{i=0}^{k} \Omega_i\Psi_{k-i} \\ &= \sum_{i=0}^{k} \left(\frac{t^i}{i!}\mathbf{A}^i\right)\left(\frac{t^{k-i}}{(k-i)!}\mathbf{W}^{k-i}\right) \\ &= \frac{1}{k!}t^k \sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!}\mathbf{A}^i\mathbf{W}^{k-i} \end{aligned} \tag{F.24}$$

Suppose $\mathbf{A}$ and $\mathbf{W}$ commute, then

$$\begin{aligned} (\mathbf{A}+\mathbf{W})^2 &= (\mathbf{A}+\mathbf{W})(\mathbf{A}+\mathbf{W}) \\ &= \mathbf{A}^2 + \mathbf{W}\mathbf{A} + \mathbf{A}\mathbf{W} + \mathbf{W}^2 \\ &= \mathbf{A}^2 + 2\mathbf{A}\mathbf{W} + \mathbf{W}^2 \\ (\mathbf{A}+\mathbf{W})^3 &= (\mathbf{A}+\mathbf{W})^2(\mathbf{A}+\mathbf{W}) \\ &= \mathbf{A}^3 + 2\mathbf{A}\mathbf{W}\mathbf{A} + \mathbf{W}^2\mathbf{A} \end{aligned}$$

$$
\begin{aligned}
& & & +\mathbf{A}^2\mathbf{W} + 2\mathbf{A}\mathbf{W}^2 + \mathbf{W}^3 \\
& & = & \quad \mathbf{A}^3 + 3\mathbf{A}^2\mathbf{W} + 3\mathbf{A}\mathbf{W}^2 + \mathbf{W}^3 \\
& & \vdots & \\
(\mathbf{A}+\mathbf{W})^k & & = & \quad \sum_{i=0}^{k} \frac{k!}{i!\,(k-i)!} \mathbf{A}^i \mathbf{W}^{k-i}
\end{aligned}
$$

which will not be true in general unless $\mathbf{A}$ and $\mathbf{W}$ commute. Thus, if and only if $\mathbf{A}$ and $\mathbf{W}$ commutes, (F.24) becomes

$$R_k = \frac{t^k}{k!} (\mathbf{A}+\mathbf{W})^k$$

and

$$
\begin{aligned}
e^{\mathbf{A}t} e^{\mathbf{W}t} &= I + t(\mathbf{A}+\mathbf{W}) + \frac{t^2}{2!}(\mathbf{A}+\mathbf{W})^2 + \cdots && \text{(F.25)} \\
&= e^{(\mathbf{A}+\mathbf{W})t} && \text{(F.26)}
\end{aligned}
$$

Lastly, for (6.54),

$$
\begin{aligned}
\frac{d}{dt} e^{\mathbf{A}t} &= \frac{d}{dt}\left(I + \mathbf{A}t + \frac{t^2}{2!}\mathbf{A}^2 + \frac{t^3}{3!}\mathbf{A}^3 + \cdots \right) \\
&= \mathbf{A} + \mathbf{A}^2 t + \frac{t^2}{2!}\mathbf{A}^3 + \frac{t^3}{3!}\mathbf{A}^4 + \cdots \\
&= \mathbf{A}\left(I + \mathbf{A}t + \frac{t^2}{2!}\mathbf{A}^2 + \frac{t^3}{3!}\mathbf{A}^3 + \cdots \right) \\
&= \mathbf{A} e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}
\end{aligned}
$$

which implies that $\mathbf{A}$ and $e^{\mathbf{A}t}$ commutes.

F.4.4 Proof That Matrizants Are Invertible, Theorem 6.4

Using property 9 of Table 1.6,

$$\frac{d}{dt}\Big(\det(\mathbf{M})\Big) = \sum_{k=1}^{n} \det\left(\widehat{\mathbf{M}}_k\right)$$

where,

$$\widehat{\mathbf{M}}_k = \left(\widehat{m}_{ij}^{(k)}\right) \qquad ; \qquad \widehat{m}_{ij}^{(k)} = \begin{cases} m_{ij} & \text{if } i \neq k \\ \dfrac{dm_{ij}}{dt} & \text{if } i = k \end{cases}$$

Recalling the property of $\mathbf{M}$, (cf. (6.65)),

$$\frac{d\mathbf{M}}{dt} = \mathbf{A}\mathbf{M} \qquad \rightarrow \qquad \frac{dm_{ij}}{dt} = \sum_{\ell=1}^{n} a_{i,\ell}\, m_{\ell,j}$$

where a_{ij} and m_{ij} are the $(i, j)^{\text{th}}$ element of $\mathbf{A}$ and $\mathbf{M}$, respectively. Then

$$\widehat{\mathbf{M}}_k = \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & & \vdots \\ \left(\sum_{\ell=1}^{n} a_{k,\ell}\, m_{\ell,1}\right) & \cdots & \left(\sum_{\ell=1}^{n} a_{k,\ell}\, m_{\ell,n}\right) \\ \vdots & & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix}$$

and

$$\det\left(\widehat{\mathbf{M}}_k\right) = \sum_{\ell=1}^{n} a_{k,\ell}\ \det \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & & \vdots \\ m_{\ell,1} & \cdots & m_{\ell,n} \\ \vdots & & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix}$$

Thus

$$\frac{d}{dt}\Big(\det(\mathbf{M})\Big) = a_{11}\det(\mathbf{M}) + \cdots + a_{nn}\det(\mathbf{M}) = \text{trace}(\mathbf{A})\det(\mathbf{M})$$

Integrating,

$$\det(\mathbf{M}) = e^{\int \text{trace}(\mathbf{A})dt}$$

Because the trace of $\mathbf{A}$ is bounded, the determinant of $\mathbf{M}$ will never be zero, that is, $\mathbf{M}^{-1}$ exists.

F.4.5 Proof of Instability Theorem, Theorem 6.5.

For the general case, including nondiagonalizable matrices, we use the modal matrices that transforms $\mathbf{A}$ to a canonical Jordan block form,

$$\mathbf{A} = TJT^{-1}$$

where

$$J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_m \end{pmatrix} \quad ; \quad J_k = \begin{pmatrix} \lambda_k & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_k \end{pmatrix}$$

Let $\mathbf{z} = T^{-1}\mathbf{x}$ and $Q(t) = T^{-1}\mathbf{b}(t)$, then

$$\frac{d}{dt}\mathbf{z} = J\,\mathbf{z} + Q \quad \rightarrow \quad \frac{d}{dt}\mathbf{z}_k = J_k\mathbf{z}_k + \mathbf{q}_k$$

If a Jordan block is a 1×1 matrix, then the corresponding differential equation is a scalar first-order differential equation. However, for larger sizes, the solution is given by

$$\mathbf{z}_k(t) = e^{J_k t}\mathbf{z}_k(0) + \int_0^t e^{J_k(t-\tau)}\mathbf{q}_k(t)d\tau$$

where

$$e^{J_k t} = \begin{pmatrix} e^{\lambda_k t} & te^{\lambda_k t} & \cdots & (t^{\ell-1}/(\ell-1)!)e^{\lambda_k t} \\ 0 & e^{\lambda_k t} & \cdots & (t^{\ell-2}/(\ell-2)!)e^{\lambda_k t} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_k t} \end{pmatrix}$$

If any of the eigenvalues have a positive real part, then some elements of $\mathbf{z}$ will grow unbounded as t increases. Because $\mathbf{x} = T\mathbf{z}$, the system will be unstable under this condition.

APPENDIX G

Additional Details and Fortification for Chapter 7

G.1 Differential Equation Solvers in MATLAB

G.1.1 IVP Solvers

As a quick example, consider the following system:

$$\begin{aligned} \frac{dy_1}{dt} &= \left(2e^{-t}+1\right) y_2 - 3y_1 \\ \frac{dy_2}{dt} &= -2y_2 \end{aligned} \tag{G.1}$$

with $y_1(0) = 2$ and $y_2(0) = 1$. Then the following steps are needed:

1. Build a file, say `derfunc.m`, to evaluate derivatives of the state space model:

```
function dy = derfunc(t,y)
    y1  = y(1);   y2 = y(2)          ;
    dy1 = (2*exp(-t)+1)*y2 -3*y1     ;
    dy2 = -2*y2                      ;
    dy  = [dy1;dy2]                  ;
```

2. Run the initial value solver

```
>> [t,y]=ode45(@derfunc,[0,2],[2;1]);
```

   where `[t,y]` are the output time and states, respectively, `derfunc` is the file name of the derivative function, `[0,2]` is the time span, and `[2;1]` is the vector of initial values. A partial list of solvers that are possible alternatives to `ode45` is given in Table G.1. It is often suggested to first try `ode45`. If the program takes too long, then it could be due to the system being stiff. In those cases, one can attempt to use `ode15s`.

There are more advanced options available for these solvers in MATLAB. In addition to the ability to set relative errors or absolute errors, one can also include "event handling" (e.g., modeling a bouncing ball), allow passing of model parameters, or solving equations in mass-matrix formulations, that is,

$$M(t, \mathbf{y})\frac{d}{dt}\mathbf{y} = \mathbf{f}(t, \mathbf{y})$$

Table G.1. *Some initial value solvers for MATLAB*

Solver	Description	Remarks
`ode23`	$(2, 3)^{\text{th}}$ Bogacki-Shampine Embedded Runge-Kutta	
`ode45`	$(4, 5)^{\text{th}}$ Dormand-Prince Embedded Runge-Kutta	Efficient for most non-stiff problems.
`ode113`	Adams-Bashforth-Moulton Predictor-Corrector	Also for non-stiff problems. Used when state-space model is more computationally intensive.
`ode15s`	Variable-order BDF (Gear's method)	For stiff problems. May not be as accurate as `ode45`. Allows settings/definition of Jacobians. Can be used to solve DAE problems with index-1.
`ode23s`	Order-2 Rosenbrock method	For stiff problems. May solve problems where `ode15s` fails.
`ode23t`	Trapezoid method	For stiff problems. Implements some numerical damping. Used also to solve DAE problems with index-1.
`ode23tb`	Trapezoid method stage followed by BDF stage	For stiff problems. May be more efficient than `ode15s` at crude tolerances.

where $M(t, \mathbf{y})$ is either singular (as it would be for DAE problems) and/or has preferable sparsity patterns.

G.1.2 BVP Solver

As a quick example, consider the same system as (G.1), but instead of the initial conditions, we wish to satisfy the following two-point boundary conditions: $y_1(1) = 0.3$ and $y_2(0) = 1$. Then the following steps are needed to solve this boundary value problem in MATLAB:

1. Build the model file, say, `derfunc.m`, as done in the previous section.
2. Let $\mathbf{r}$ be the vector of residuals from the boundary conditions; that is, reformulate the boundary conditions in a form where the the right hand side is made equal to zero,

$$\mathbf{r} = \begin{pmatrix} y_1(1) - 0.3 \\ y_2(0) - 1 \end{pmatrix}$$

Now build another file, say, `bconds.m`, that generates $\mathbf{r}$,

```
function r = bconds(yinit,yfinal)
    r1 = yfinal(1)-0.3  ;
    r2 = yinit(2)-1     ;
    r  = [r1;r2]        ;
```

Note that this file does not know that the final point is at $t = 1$. That information will have to come from a structured data, `trialSoln`, that is formed in the next step.

3. Create a trial solution data, `trialSoln`,

```
>> trialSoln.x = linspace(0,1,10);
>> trialSoln.y = [0.5;0.2]*ones(1,10);
```

The data in `trialSoln.x` give the initial point $t = 0$, final point $t = 1$, and 10 mesh points. One can vary the mesh points so that finer mesh sizes can be focused around certain regions. The data in `trialSoln.y` just give the intial conditions repeated at each mesh point. This could also be altered to be closer to the final solution. (Another MATLAB command `bvpinit` is available to create the same initial data and has other advanced options.)

4. Run the BVP solver,

```
>> soln = bvp4c(@derfunc,@bconds,trialSoln);
```

The output, `soln`, is also a structured data. Thus for plotting or other postprocessing of the output data, one may need to extract the t variable and y variables as follows:

```
>> t=soln.x;     y=soln.y;
```

There are several advanced options for `bvp4c`, including the solution of multipoint BVPs, some singular BVPs, and BVPs containing unknown parameters. The solver used in `bvp4c` is said to be finite difference method coupled with a three-stage implicit Runge-Kutta method known as Lobatto III-a.

G.1.3 DAE Solver

Consider the van der Pol equation in Lienard coordinates given by

$$\begin{aligned} \frac{dy_1}{dt} &= -y_2 \\ 0 &= y_1 - \left(\frac{y_2^3}{3} - y_2\right) \end{aligned}$$

which could be put into the mass matrix form as

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \frac{d}{dt} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} -y_2 \\ y_1 - \left(\frac{y_2^3}{3} - y_2\right) \end{pmatrix}$$

The following steps are needed to solve this DAE problem using MATLAB.

1. Build the model file, say, `daevdpol.m`,

```
function  dy = daevdpol( t, y )
    y1  =  y(1)                 ;
    y2  =  y(2)                 ;
    dy1 = -y2                   ;
    dy2 =  y1 - (y2^3/3 - y2)   ;
    dy  =  [dy1;dy2]            ;
```

2. Make sure the initial conditions are consistent. For instance, the algebraic condition is satisfied for $\mathbf{y} = (-0.0997, 0.1)^T$.

3. Set the parameter, `options`, to include the mass matrix information using the command

```
>> options=odeset('Mass',[1,0;0,0]);
```

4. Run the DAE solver

```
>>[t,y]=ode15s(@daevdpol,[0,2],[-0.0997;0.1],options)
```

where `[0,2]` is the time span.

G.2 Derivation of Fourth-Order Runge Kutta Method

G.2.1 Fourth-Order Explicit RK Method

To obtain a fourth-order approximation, we truncate the Taylor series expansion as follows:

$$y_{k+1} \approx y_k + h \left.\frac{dy}{dx}\right|_{t_k,y_k} + \frac{h^2}{2!} \left.\frac{d^2y}{dx^2}\right|_{t_k,y_k} + \frac{h^3}{3!} \left.\frac{d^3y}{dx^3}\right|_{t_k,y_k} + \frac{h^4}{4!} \left.\frac{d^4y}{dx^4}\right|_{t_k,y_k} \tag{G.2}$$

The coefficient of h^i in (G.2) can then be matched with the coefficients of h in (7.13). This approach is very long and complicated. For instance, by expanding the derivatives of y in terms of f and its partial derivatives,

$$\begin{aligned}
\frac{dy}{dt} &= f \\
\frac{d^2y}{dt^2} &= \frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f \\
\frac{d^3y}{dt^3} &= \frac{\partial^2 f}{\partial t^2} + 2f\frac{\partial^2 f}{\partial t \partial y} + \frac{\partial^2 f}{\partial y^2} f^2 + \frac{\partial f}{\partial t}\frac{\partial f}{\partial y} + \left(\frac{\partial f}{\partial y}\right)^2 f \\
&\vdots
\end{aligned}$$

The number of terms increases exponentially with increases in the order of differentiation. These equations, including those for higher orders, can be made more tractable using an elegant formulation using labeled trees (see, e.g., Hairer and Wanner [1993]).

As an alternative, we simplify the process by picking specific forms for $f(t, y)$. The first choice is to let $f(t, y) = t^3$. The analytical solution from y_k to y_{k+1} is given by

$$\begin{aligned}
\frac{d}{dt}y &= t^3 \\
y_{k+1} - y_k &= \frac{(t_k + h)^4 - t_k^4}{4} \\
y_{k+1} &= y_k + t_k^3 h + \frac{3}{2} t_k^2 h^2 + t_k h^3 + \frac{1}{4} h^4
\end{aligned} \tag{G.3}$$

Applying a four-stage Runge-Kutta method using (7.12) and (7.13),

$$\begin{aligned}
\delta_{k1} &= h\, t_k^3 \\
\delta_{k2} &= h\, (t_k + a_2 h)^3 \\
\delta_{k3} &= h\, (t_k + a_3 h)^3
\end{aligned}$$

$$\begin{aligned}
\delta_{k4} &= h\ (t_k + a_4 h)^3 \\
y_{k+1} &= y_k + c_1\delta_{k1} + c_2\delta_{k2} + c_3\delta_{k3} + c_4\delta_{k4} \\
&= y_k + (c_1 + c_2 + c_3 + c_4)\, t_k^3 h \\
&\quad + 3\,(c_2 a_2 + c_3 a_3 + c_4 a_4)\, t_k^2 h^2 \\
&\quad + 3\,\left(c_2 a_2^2 + c_3 a_3^2 + c_4 a_4^2\right) t_k h^3 \\
&\quad + \left(c_2 a_2^3 + c_3 a_3^3 + c_4 a_4^3\right) t_k h^4 \qquad \text{(G.4)}
\end{aligned}$$

Comparing (G.3) and (G.4),

$$\begin{aligned}
c_1 + c_2 + c_3 + c_4 &= 1 \\
c_2 a_2 + c_3 a_3 + c_4 a_4 &= \frac{1}{2} \\
c_2 a_2^2 + c_3 a_3^2 + c_4 a_4^2 &= \frac{1}{3} \\
c_2 a_2^3 + c_3 a_3^3 + c_4 a_4^3 &= \frac{1}{4} \qquad \text{(G.5)}
\end{aligned}$$

Next, we choose $f(t, y) = ty$. The analytical solution is given by

$$\begin{aligned}
\frac{d}{dt}y &= ty \\
\ln\left(\frac{y_{k+1}}{y_k}\right) &= \frac{(t_k + h)^2 - t_k^2}{2} \\
y_{k+1} &= y_k\ \exp\left(\frac{(t_k + h)^2 - t_k^2}{2}\right) \qquad \text{(G.6)}
\end{aligned}$$

The Taylor series expansion is given by,

$$\begin{aligned}
y_{k+1} &= y_k \left[1 + t_k h + \frac{1}{2}\left(1 + t_k^2\right) h^2 \right. \\
&\quad \left(\frac{1}{2} t_k + \frac{1}{6} t_k^3\right) h^3 \\
&\quad \left. \left(\frac{1}{8} + \frac{1}{4} t_k^2 + \frac{1}{24} t_k^4\right) h^4 + O(h^5)\right] \qquad \text{(G.7)}
\end{aligned}$$

Applying the four-stage Runge-Kutta method using (7.12) and (7.13),

$$\begin{aligned}
\delta_{k1} &= h\, t_k y_k \\
\delta_{kj} &= h\ (t_k + a_j h)\left(y_k + \sum_{\ell=1}^{j-1} b_{j\ell}\delta_{k\ell}\right) \qquad j = 2, 3, 4 \\
y_{k+1} &= y_k + c_1\delta_{k1} + c_2\delta_{k2} + c_3\delta_{k3} + c_4\delta_{k4} \\
&= y_k \left[1 + \sigma_{1,1} t_k h + \left(\sigma_{2,0} + \sigma_{2,2} t_k^2\right) h^2 \right. \\
&\quad \left(\sigma_{3,1} t_k + \sigma_{3,3} t_k^3\right) h^3 \\
&\quad \left. \left(\sigma_{4,0} + \sigma_{4,2} t_k^2 + \sigma_{4,4} t_k^4\right) h^4 + O(h^5)\right] \qquad \text{(G.8)}
\end{aligned}$$

where,

$$
\begin{aligned}
\sigma_{1,1} &= \sum_{i=1}^{4} c_i \\
\sigma_{2,0} &= \sum_{i=2}^{4} c_i a_i \\
\sigma_{2,2} &= \sum_{i=2}^{4} c_i \sum_{j=1}^{i-1} b_{ij} \\
\sigma_{3,1} &= \sum_{i=3}^{4} c_i \sum_{\ell=2}^{i-1} a_\ell b_{i,\ell} + \sum_{i=2}^{4} c_i a_i \sum_{j=1}^{i-1} b_{ij} \\
\sigma_{3,3} &= \sum_{i=3}^{4} c_i \sum_{\ell=1}^{i-2} \sum_{j=\ell+1}^{i-1} b_{ij} b_{j\ell} \\
\sigma_{4,0} &= \sum_{i=3}^{4} c_i a_i \sum_{j=2}^{i-1} b_{ij} a_j \\
\sigma_{4,2} &= \sum_{i=3}^{4} c_i a_i \sum_{\ell=1}^{i-2} \sum_{j=\ell+1}^{i-1} b_{ij} b_{j\ell} + \sum_{i=3}^{4} c_i \sum_{\ell=2}^{i-1} b_{i\ell} a_\ell \sum_{j=1}^{\ell-1} b_{\ell j} + c_4 b_{43} b_{32} a_2 \\
\sigma_{4,4} &= c_4 b_{43} b_{32} b_{21}
\end{aligned}
$$

Now compare the coefficients of (G.8) and (G.7). Using (7.17) and including (G.5), we end up with the eight equations necessary for the fourth-order approximation:

$$
\begin{aligned}
c_1 + c_2 + c_3 + c_4 &= 1 & \qquad c_3 b_{32} a_2 + c_4 (b_{43} a_3 + b_{42} a_2) &= \frac{1}{6} \\
c_2 a_2 + c_3 a_3 + c_4 a_4 &= \frac{1}{2} & \qquad c_3 a_3 b_{32} a_2 + c_4 a_4 (b_{43} a_3 + b_{42} a_2) &= \frac{1}{8} \\
c_2 a_2^2 + c_3 a_3^2 + c_4 a_4^2 &= \frac{1}{3} & \qquad c_3 b_{32} a_2^2 + c_4 \left(b_{43} a_3^2 + b_{42} a_2^2\right) &= \frac{1}{12} \\
c_2 a_2^3 + c_3 a_3^3 + c_4 a_4^3 &= \frac{1}{4} & \qquad c_4 b_{43} b_{32} a_2 &= \frac{1}{24}
\end{aligned}
\tag{G.9}
$$

After replacing a_j with $\sum_\ell b_{j\ell}$, there are ten unknowns (b_{ij} and $c_j, i < j, j = 1, 2, 3, 4$) with only eight equations, yielding two degrees of freedom. One choice is to set $b_{31} = b_{41} = 0$. This will result in the coefficients given in the tableau shown in (7.14).

Another set of coefficients that satisfies the eight conditions given in (G.9) is the Runge-Kutta tableau given by

$$\left(\begin{array}{c|c} \mathbf{a} & B \\ \hline & \mathbf{c}^T \end{array}\right) = \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{2}{3} & -\frac{1}{3} & 1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{array} \tag{G.10}$$

G.2.2 Fourth-Order Implicit Runge Kutta (Gauss-Legendre)

Let us now obtain the two-stage implicit Runge Kutta method that yields a fourth-order approximation.[1] We begin by choosing $f(t, y) = y$ from which the full implicit formulation becomes

$$\begin{aligned} \delta_{k1} &= h\left(y_k + b_{11}\delta_{k1} + b_{12}\delta_{k2}\right) \\ \delta_{k2} &= h\left(y_k + b_{21}\delta_{k1} + b_{22}\delta_{k2}\right) \end{aligned}$$

or

$$\begin{pmatrix} \delta_{k1} \\ \delta_{k2} \end{pmatrix} = \begin{pmatrix} (1/h) - b_{11} & -b_{12} \\ -b_{21} & (1/h) - b_{22} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} y_k$$

Substituting into (7.13),

$$\begin{aligned} y_{k+1} &= y_k + \begin{pmatrix} c_1 & c_2 \end{pmatrix} \begin{pmatrix} (1/h) - b_{11} & -b_{12} \\ -b_{21} & (1/h) - b_{22} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} y_k \\ &= y_k \left(\frac{1 + p_1 h + p_2 h^2}{1 + q_1 h + q_2 h^2}\right) \end{aligned} \tag{G.11}$$

where

$$\begin{aligned} p_1 &= c_1 + c_2 - b_{11} - b_{22} \\ p_2 &= c_1\left(b_{12} - b_{22}\right) + c_2\left(b_{21} - b_{11}\right) + b_{22}b_{11} - b_{12}b_{21} \\ q_1 &= -b_{11} - b_{22} \\ q_2 &= b_{11}b_{22} - b_{12}b_{21} \end{aligned}$$

The analytical solution of $dy/dt = y$ is $y_{k+1} = y_k e^h$. In light of the rational form given in (G.11), we can use a fourth-order Pade' approximation of e^h instead of the Taylor series expansion, that is,

$$y_{k+1} = y_k \left(\frac{1 + (h/2) + (h^2/12)}{1 - (h/2) + (h^2/12)}\right) \tag{G.12}$$

[1] The usual development of the Gauss-Legendre method is through the use of collocation theory, in which a set of interpolating Lagrange polynomials is used to approximate the differential equation at the collocation points. Then the roots of the s-degree Legendre polynomials are used to provide the collocation points. See, e.g., Hairer, Norsett and Wanner (1993).

Matching the coefficients of (G.11) and (G.12), we obtain

$$\begin{aligned}
\frac{1}{2} &= c_1 + c_2 - b_{11} - b_{22} \\
\frac{1}{12} &= c_1\,(b_{12} - b_{22}) + c_2\,(b_{21} - b_{11}) + b_{22}b_{11} - b_{12}b_{21} \\
-\frac{1}{2} &= -b_{11} - b_{22} \\
\frac{1}{12} &= b_{11}b_{22} - b_{12}b_{21}
\end{aligned} \tag{G.13}$$

which still leaves two degrees of freedom. A standard choice is to use the roots of the second-degree Legendre polynomial to fix the values of a_1 and a_2,[2] that is,

$$\mathcal{P}_2(t) = t^2 - t + (1/6)$$

yielding the roots

$$a_1 = \frac{1}{2} - \frac{\sqrt{3}}{6} \qquad \text{and} \qquad a_2 = \frac{1}{2} + \frac{\sqrt{3}}{6} \tag{G.14}$$

Also, recall the consistency condition (7.17),

$$\frac{1}{2} - \frac{\sqrt{3}}{6} = b_{11} + b_{12} \qquad \text{and} \qquad \frac{1}{2} + \frac{\sqrt{3}}{6} = b_{21} + b_{22} \tag{G.15}$$

From (G.13) and (G.15), we find that: $c_1 = c_2 = 1/2$, $b_{11} = b_{22} = 1/4$, $b_{12} = 1/4 - \sqrt{3}/6$ and $b_{21} = 1/4 + \sqrt{3}/6$.

G.3 Adams-Bashforth Parameters

To determine the values of b_j for the Adams-Bashforth method, we choose $f(y)$ that facilitates the determination of the coefficients. The simplest choice is $f(y) = y$. Doing so, the n^{th}- order Adams-Bashforth method becomes

$$y_{k+1} = y_k + h\sum_{j=0}^{m} b_j f\left(y_{k-j}\right) = y_k + h\sum_{j=0}^{m} b_j y_{k-j} \tag{G.16}$$

where $m = n - 1$. With $f(y) = y$, the analytical solution of $y_{k+\ell}$ starting at y_k is given by

$$y_{k+\ell} = e^{\ell h}\, y_k \tag{G.17}$$

Substituting this relationship to (G.16) results in

$$e^{h} = 1 + h\sum_{j=0}^{m} b_j\, e^{-jh} \tag{G.18}$$

which when expanded using Taylor's series will yield

$$1 + h + \frac{h^2}{2!} + \frac{h^3}{3!} + \cdots \quad = \quad 1 + h\sum_{j=0}^{m} b_j \left(1 - jh + \frac{(jh)^2}{2!} - \frac{(jh)^3}{3!} + \cdots\right)$$

[2] See Section 9.2 for a discussion on Legendre polynomials.

$$
\begin{aligned}
1+\frac{h}{2!}+\frac{h^2}{3!}+\cdots &= \sum_{j=0}^{m} b_j\left(1-jh+\frac{(jh)^2}{2!}-\frac{(jh)^3}{3!}+\cdots\right)\\
&= \left(\sum_{j=0}^{m} b_j\right)-h\left(\sum_{j=0}^{m} j\,b_j\right)+\frac{h^2}{2!}\left(\sum_{j=0}^{m} j^2 b_j\right)+\cdots
\end{aligned}
$$

By comparing the different coefficients of h^ℓ on both sides we get

$$
\frac{(-1)^\ell}{\ell+1}=\begin{cases}\sum_{j=1}^{m} j^\ell\, b_j & \text{if } \ell>0\\ \sum_{j=0}^{m} b_j & \text{if } \ell=0\end{cases}
$$

or in matrix form,

$$
\begin{pmatrix}
1 & 1 & 1 & \cdots & 1\\
0 & 1 & 2 & \cdots & m\\
\vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 1 & 2^m & \cdots & m^m
\end{pmatrix}
\begin{pmatrix} b_0\\ b_1\\ \vdots\\ b_m\end{pmatrix}
=
\begin{pmatrix} 1\\ -\frac{1}{2}\\ \vdots\\ \frac{(-1)^m}{m+1}\end{pmatrix}
\tag{G.19}
$$

G.4 Variable Step Sizes for BDF

For variable step sizes, the coefficients of the multistep methods will no longer be constant. In this section, we treat only the BDF formulas. The approach should generally be similar for the other multistep methods.

Let h_k be the step size at t_k and put the BDF equation (7.38) into an equivalent form,[3]

$$
\sum_{i=-1}^{m} \gamma_{(i|k)}\, y_{k-i} = h_k f\left(y_{k+1}\right) \tag{G.20}
$$

Using the same technique of finding the necessary conditions by the simple application of the approximation to $dy/dt = y$, that is, $f(y) = y$ and $y = e^t y_0$, we note that

$$
y_{k-j} = e^{(t_{k-j}-t_{k+1})} y_{k+1}
$$

Then (G.20) reduces to

$$
\begin{aligned}
\sum_{i=-1}^{m} \gamma_{(i|k)}\, e^{(t_{k-i}-t_{k+1})}\, y_{k+1} &= h_k\, y_{k+1}\\
\sum_{i=-1}^{m} \gamma_{(i|k)} \left(1+(t_{k-i}-t_{k+1})+\frac{(t_{k-i}-t_{k+1})^2}{2!}+\cdots\right) &= h_k
\end{aligned}
$$

[3] The form (G.20), in which the derivative function f is kept on one side without unknown coefficients, is often preferred when solving differential algebraic equations (DAE).

For the p^{th}-order approximation, we again let $m = p - 1$, and the equation will yield

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & (t_{k+1} - t_k) & (t_{k+1} - t_{k-1}) & \dots & (t_{k+1} - t_{k-p+1}) \\ 0 & (t_{k+1} - t_k)^2 & (t_{k+1} - t_{k-1})^2 & \dots & (t_{k+1} - t_{k-p+1})^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & (t_{k+1} - t_k)^p & (t_{k+1} - t_{k-1})^p & \dots & (t_{k+1} - t_{k-p+1})^p \end{pmatrix} \begin{pmatrix} \gamma_{(-1|k)} \\ \gamma_{(0|k)} \\ \gamma_{(1|k)} \\ \vdots \\ \gamma_{(p-1|k)} \end{pmatrix} = \begin{pmatrix} 0 \\ -h_k \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{G.21}$$

Because the right-hand side is just $-h_k \mathbf{e}_2$, this equation can be solved directly using Cramer's rule and using the determinant formulas of Vandermonde matrices. The results are

$$\gamma_{(\ell|k)} = \begin{cases} \displaystyle\sum_{j=0}^{p} \frac{t_{k+1} - t_k}{t_{k+1} - t_{k-j}} & \text{if } \ell = -1 \\ \displaystyle -\left(\frac{t_{k+1} - t_k}{t_{k+1} - t_{k-\ell}}\right) \prod_{j \geq 0, j \neq \ell}^{p} \left(\frac{t_{k+1} - t_{k-j}}{t_{k-\ell} - t_{k-j}}\right) & \text{if } \ell \geq 0 \end{cases} \tag{G.22}$$

Note that this formula involves product terms that are the Lagrange formulas used in polynomial interpolation, which is how most textbooks derive the formulas for BDF coefficients. The approach taken here, however, has the advantage that it automatically fixes the order of approximation when we truncated the Taylor series of the exponential functions based on the chosen order.

When the step sizes are constant, that is, $h_k = h$, then $(t_{k+1} - t_{k-j}) = (j+1)h$, and (G.22) can be used to find γ_ℓ independent of k. For instance, for the sixth-order BDF method, that is, $p = 6$, the coefficient of y_{k-3} becomes

$$\gamma_3 = -\frac{1}{4}\left(\frac{1}{1-4} \cdot \frac{2}{2-4} \cdot \frac{3}{3-4} \cdot \frac{5}{5-4} \cdot \frac{6}{6-4}\right) = \frac{15}{4}$$

To determine the appropriate value for h_k, we can first set $h_k = h_{k-1}$ and then use either of the error-control methods given in Section G.5 to modify it. The step-doubling approach might be simpler for the general nonlinear case.

G.5 Error Control by Varying Step Size

To improve accuracy, one could include more terms from the Taylor series expansion. Another way is to decrease the value of h. However, decreasing h will increase the number of points to be solved, thereby increasing the length of computation and storage. Thus the step size has to be chosen by balancing accuracy requirements with computational loads. In addition, the step sizes h_k do not need to be uniform at each step

G.5.1 Estimation of Local Truncation Error

First, we need to estimate the truncation error at the kth step. Consider two integration methods: one that obtains an nth order approximation, and another that obtains an $(n+1)$th order approximation. Starting from the same value of y_k, let w_{k+1} and z_{k+1} be the update value for y_k using the $(n+1)$th and the nth order approximation methods, respectively, that is, for one-step methods

$$w_{k+1} = y_k + h_k \Phi(t_k, y_k)|_{(n+1)th \text{ order}} \quad \text{and} \quad z_{k+1} = y_k + h_k \Phi(t_k, y_k)|_{(n)th \text{ order}}$$

where $\Phi(t_k, y_k)$ is a transition formula based on the particular method chosen.

Subtracting z_{k+1} from w_{k+1}, we obtain an estimate of the truncation error of $f(t, y)$, that is,

$$\tau_{k+1}(h_k) = \frac{|w_{k+1} - z_{k+1}|}{h_k} \tag{G.23}$$

In addition, we expect that the truncation error, $\tau_{k+1}(h_k)$, is of the order of h_k^n, that is, for some constant C,

$$\tau_{k+1}(h_k) = Ch_k^n \qquad \rightarrow \qquad Ch_k^n = \frac{|w_{k+1} - z_{k+1}|}{h_k} \tag{G.24}$$

We want to find a different step size, $h_k^{\text{revised}} = \theta h_k$ ($\theta > 0$), such that the truncation error using the revised step size will be less than a prescribed tolerance ϵ, that is, $\tau_{k+1}(\theta h_k) \leq \epsilon$. Using (G.24),

$$\tau_{k+1}(\theta h_k) = (C)\,(\theta h_k)^n = \theta^n C h_k^n = \theta^n \frac{|w_{k+1} - z_{k+1}|}{h_k} \leq \epsilon$$

Rearranging,

$$\theta \leq \left(\frac{\epsilon h_k}{|w_{k+1} - z_{k+1}|} \right)^{1/n} \tag{G.25}$$

To incorporate (G.25), we can set θ to be equal to the right hand side of (G.25) with ϵ divided by 2, that is,

$$\hat{\theta} = \left(\frac{\epsilon h_k}{2|w_{k+1} - z_{k+1}|} \right)^{1/n} \tag{G.26}$$

This would guarantee a strict inequality in (G.25).

The implementation of (G.26) is shown in the flowchart given in Figure G.1. In the flowchart, we see that if the truncation error, τ_{k+1}, is less than the tolerance ϵ, we can set y_{k+1} to be w_{k+1}. Otherwise, we choose θ to be

$$\theta = \min \left(\theta_{\max},\ \max \left(\theta_{\min}, \hat{\theta} \right) \right) \tag{G.27}$$

If τ_{k+1} happens to be much less than ϵ, the scaling factor θ will be greater than unity, which means the previous step size was unnecessarily small. Thus the step size could be increased. However, if τ_{k+1} is greater than ϵ, θ will be less than unity, which means the step size has to be reduced to satisfy the accuracy requirements. As shown in the flowchart, we also need to constrain the step size h_k to be within a preset maximum bound, $h_{\max}$, and minimum bound, $h_{\min}$.

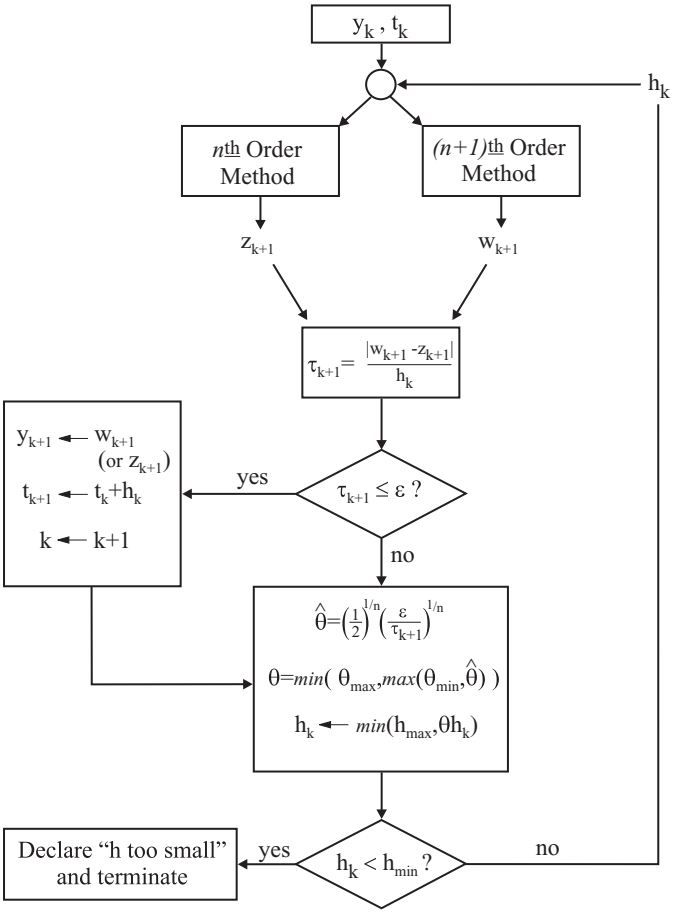


Figure G.1. Flowchart for error control.

G.5.2 Embedded Runge-Kutta Formulas

The error control procedure shown in the flowsheet given in Figure G.1 requires two Runge-Kutta computations for the update of y_k. One is an nth order method, whereas the other is an $(n + 1)$th order method. Normally, this would mean using two Runge-Kutta tableaus, one for each method. However, to improve efficiency, one could find a different tableau such that both updates can share some of the intermediate calculations of δ_{ij} given in (7.13). This is done usually at a cost of increasing more terms in (7.13). Conversely, because both tableaus are merged into one tableau, the net change would usually mean fewer function evaluations. When two or more tableaus are merged to share the same function evaluations, we refer to these as **embedded Runge-Kutta formulas**, and the corresponding tableaus are called **embedded Runge-Kutta tableaus**.

Two of the more popular embedded Runge-Kutta methods are the Fehlberg-4-5 method and the Dormand-Prince-5-4 method. The Fehlberg tableau is given in (G.28). The row for z_{k+1} (second from the bottom) is used to determine the fourth-order update, whereas the row for w_{k+1} (last row) is used to determine the fifth-order update. However, the Fehlberg method uses z_{k+1} (the lower order result) as the update for y_{k+1} because the parameter values of the embedded

tableau were determined to minimize errors in the fourth-order estimates. The Dormand-Prince tableau is given in (G.29). The Dormand-Prince has a few more additional terms than the Fehlberg tableau. It was optimized for the fifth-order estimate instead. This means that the last row is the fourth-order estimate, whereas the second to the last row is the fifth-order estimate. So the Dormand-Prince tableau shown in (G.29) will use w_{k+1}, a fifth-order result, as the update for y_{k+1}. A MATLAB code for the Fehlberg 4/5 embedded Runge-Kutta method together with the error control algorithm shown in Figure G.1 is available on the book's webpage as `fehlberg45.m`.

$$\textbf{F45}: \left[\begin{array}{c|cccccc} 0 & & & & & & \\ \frac{1}{4} & \frac{1}{4} & & & & & \\ \frac{3}{8} & \frac{3}{32} & \frac{9}{32} & & & & \\ \frac{12}{13} & \frac{1932}{2197} & -\frac{7200}{2197} & \frac{7296}{2197} & & & \\ 1 & \frac{439}{216} & -8 & \frac{3680}{513} & -\frac{845}{4104} & & \\ \frac{1}{2} & -\frac{8}{27} & 2 & -\frac{3544}{2565} & \frac{1859}{4104} & -\frac{11}{40} & \\ \hline z_{k+1} & \frac{25}{216} & 0 & \frac{1408}{2565} & \frac{2197}{4104} & -\frac{1}{5} & 0 \\ \hline w_{k+1} & \frac{16}{135} & 0 & \frac{6656}{12825} & \frac{28561}{56430} & -\frac{9}{50} & \frac{2}{55} \end{array}\right] \tag{G.28}$$

$$\textbf{DP54}: \left[\begin{array}{c|ccccccc} 0 & & & & & & & \\ \frac{1}{5} & \frac{1}{5} & & & & & & \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & & & & & \\ \frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & & & & \\ \frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & & & \\ 1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & & \\ 1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & \\ \hline z_{k+1} & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \hline w_{k+1} & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & -\frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \end{array}\right] \tag{G.29}$$

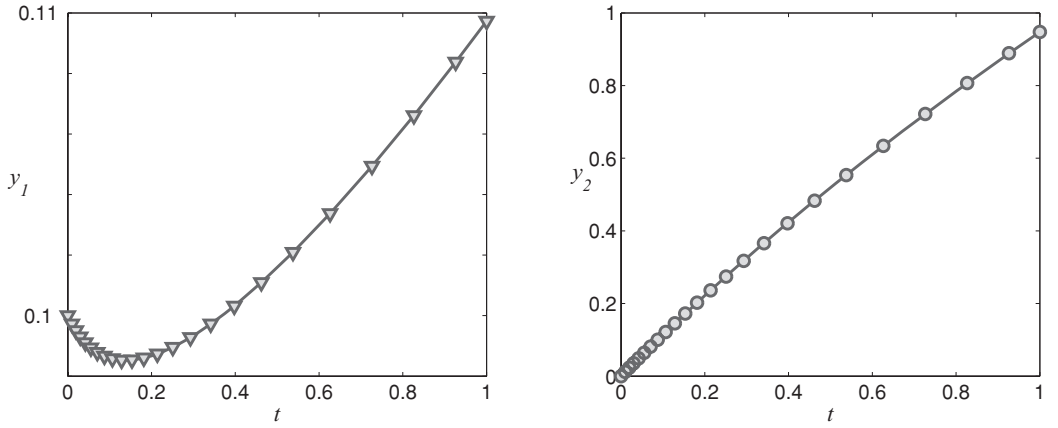


Figure G.2. Numerical solution for Example G.1 showing varying step sizes based on error-control strategy for tolerance $\epsilon = 10^{-8}$.

EXAMPLE G.1. Consider the following set of differential equations to model the production of enzyme

$$\begin{aligned} \frac{dy_1}{dt} &= (\mu - D)\, y_1 \\ \frac{dy_2}{dt} &= D\,(y_{2f} - y_2) - \frac{\mu y_1}{Y} \\ \mu &= \frac{\mu_{\max} y_2}{k_m + y_2} \end{aligned}$$

where $Y = 0.4$, $D = 0.3$, $y_{2f} = 4.0$, $\mu_{\max} = 0.53$, and $k_m = 0.12$ are the yield, dilution rate, feed composition, maximum rate, and Michaelis-Menten parameter, respectively. Assuming an initial condition of $\mathbf{y}(0) = (0.1, 0)^T$, we have the plots shown in Figure G.2 after applying the Fehlberg 4/5 embedded Runge-Kutta method using error control with tolerance $\epsilon = 10^{-8}$. We see that the step sizes are smaller near $t = 0$ but increased as necessary.

G.5.3 Step Doubling

For implicit methods, such as the fourth-order Gauss-Legendre IRK method given in Section 7.2.2, there are no embedded methods. One approach is to use a higher order version and, together with the fourth-order result, obtain an estimate of the local error to be used for step-size control.

Another method is the **step-doubling** approach. In this approach, one approximation, $z_{k+2} \approx y_{k+2}$, is obtained by using the chosen implicit method twice with a step-size of h_k. Another approximation, $w_{k+2} \approx y_{k+2}$, is obtained by applying the chosen implicit method once, but with a step-size of $2h_k$. Let $Err\,(h_k)$ be the local error using a step-size of h_k, which will be proportional to h_k^{n+1}, where n is the order of accuracy of the solver, then

$$Err\,(h_k) = Ch_k^{n+1} \qquad \rightarrow \qquad Err\,(2h_k) = 2^{n+1} Ch_k^{n+1}$$

and

$$\begin{aligned}\left|w_{k+2}-z_{k+2}\right| &= 2^{n+1}Ch_k^{n+1}-2Ch_k^{n+1}\\ &= \left(2^{n+1}-2\right)Err\left(h_k\right)\end{aligned}$$

or

$$Err\left(h_k\right)=\frac{\left|w_{k+2}-z_{k+2}\right|}{2^{n+1}-2}$$

To control the error within a tolerance, ϵ, we need to change the step-size by a factor θ, that is,

$$\begin{aligned}Err\left(\theta h_k\right) &\leq \epsilon\\ C\left(\theta h_k\right)^{n+1} &\leq\\ \theta^{n+1}Err\left(h_k\right) &\leq\\ \theta^{n+1}\frac{\left|w_{k+2}-z_{k+2}\right|}{2^{n+1}-2} &= \gamma\epsilon\end{aligned}$$

where $\gamma<1$, for example, $\gamma=0.9$. This yields the formula for θ based on the step-doubling approach:

$$\theta=\left(\frac{\gamma\epsilon\left(2^{n+1}-2\right)}{\left|w_{k+2}-z_{k+2}\right|}\right)^{1/(n+1)} \tag{G.30}$$

The MATLAB code for the Gauss-Legendre IRK is available on the book's webpage as `glirk.m`, and it incorporates the error control based on the step-doubling method.

EXAMPLE G.2. Consider the van der Pol oscillator described by the following equation:

$$\begin{aligned}\frac{dy_1}{dt} &= y_2\\ \frac{dy_2}{dt} &= \mu\left(1-y_1^2\right)y_2-y_1\end{aligned}$$

subject to the initial condition $\mathbf{y}=(1,1)^T$. When $\mu=500$, the system becomes practically stiff. Specifically, for the range $t=0$ to $t=800$, the Fehlberg 4/5 Runge Kutta will appear to "hang." Instead, we could apply the Gauss-Legendre Implicit Runge Kutta, together with error-control based on the step-doubling approach using tolerance $\epsilon=10^{-6}$. This results in the plot shown in Figure G.3, which shows that small step sizes are needed where the slopes are nearly vertical.

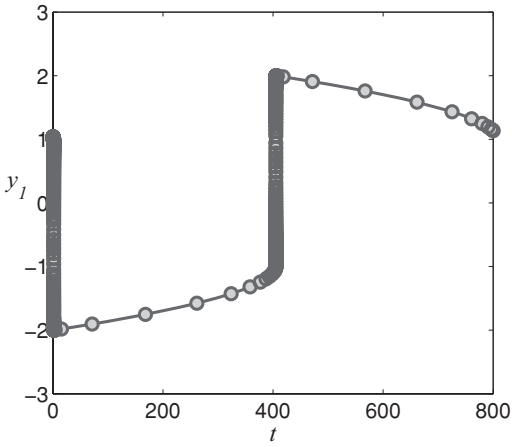


Figure G.3. The response for a van der Pol oscillator when $\mu = 500$ using Gauss-Legendre IRK method with error control based on step-doubling procedure.

G.6 Proof of Solution of Difference Equation, Theorem 7.1

First, we can rewrite the (7.46) in terms of constants $\beta_{j,\ell}$ instead of $c_{j,\ell}$ as follows:

$$\begin{aligned}\mathcal{S}(j,n) &= \left(\sum_{\ell=0}^{k_j-1} c_{j,\ell}\, n^{\ell}\right)(\sigma_j)^n = \left(\sum_{\ell=0}^{k_j-1} \beta_{j,\ell}\frac{n!}{(n-\ell)!}\right)(\sigma_j)^n \\ &= \sum_{\ell=0}^{k_j-1} \beta_{j,\ell}\,(\sigma_j)^{\ell}\; D_{\sigma_j}^{\ell}\left((\sigma_j)^n\right)\end{aligned}$$

where $\left(D_{\sigma_j}^{\ell} = \dfrac{d^{\ell}}{d\,(\sigma_j)^{\ell}}\right)$. Next, apply the difference operators of (7.43) on $\mathcal{S}(j,n)$ in place of y, with $\chi_{(\sigma_j)} = \sum_{i=0}^{p}\alpha_i\,(\sigma_j)^i$,

$$\begin{aligned}\sum_{i=0}^{p}\alpha_i\mathcal{Q}^i\left(\mathcal{S}(j,n)\right) &= \sum_{\ell=0}^{k_j-1}\beta_{j,\ell}\,(\sigma_j)^{\ell}\,D_{\sigma_j}^{\ell}\left[\sum_{i=0}^{p}\alpha_i\,(\sigma_j)^{n+i}\right] \\ &= \sum_{\ell=0}^{k_j-1}\beta_{j,\ell}\,(\sigma_j)^{\ell}\,D_{\sigma_j}^{\ell}\left[\chi_{(\sigma_j)}\,(\sigma_j)^n\right] \\ &= \sum_{\ell=0}^{k_j-1}\beta_{j,\ell}\,(\sigma_j)^{\ell}\sum_{m=0}^{\ell}\frac{\ell!}{m!(\ell-m)!}D_{\sigma_j}^{m}\left[\chi_{(\sigma_j)}\right]D_{\sigma_j}^{\ell-m}\left[(\sigma_j)^n\right]\end{aligned}$$

Because σ_j is a k_j-fold root of $\chi_{(\sigma)} = 0$,

$$D_{\sigma_j}^{\ell}\left[\chi_{(\sigma_j)}\right] = 0 \qquad ; \qquad \ell = 0, 1, \ldots, k_j - 1$$

$$\rightarrow \qquad \sum_{i=0}^{p}\alpha_i\mathcal{Q}^i\left(\mathcal{S}(j,n)\right) = 0$$

Combining all the results,

$$\sum_{i=0}^{p}\alpha_i\mathcal{Q}^i(y_n) = \sum_{j=1}^{M}\left(\sum_{i=0}^{p}\alpha_i\mathcal{Q}^i\left(\mathcal{S}(j,n)\right)\right) = 0$$

G.7 Nonlinear Boundary Value Problems

Consider the nonlinear boundary value problems given by

$$\frac{d}{dt}\mathbf{x} = \mathbf{F}(t, \mathbf{x}) \tag{G.31}$$

subject to the nonlinear boundary conditions,

$$\mathbf{q}(\mathbf{x}(0), \mathbf{x}(T)) = 0 \tag{G.32}$$

First, let us define the following vectors:

1. Let $\mathbf{x}_0$ be any initial value of $\mathbf{x}$ for the system given in (G.31).
2. Let $\mathbf{x}_T$ be the value of $\mathbf{x}$ at $t = T$ corresponding to $\mathbf{x}_0$. Thus

$$\mathbf{x}_T = \mathbf{x}_T(\mathbf{x}_0) \tag{G.33}$$

and these vectors could be evaluated by using any initial value solvers such as Runge-Kutta method after setting $\mathbf{x}_0$ as the initial condition.

The main idea of the shooting method is to find the appropriate value for $\mathbf{x}_0$ such that the boundary conditions given in (G.32) are satisfied, that is,

$$\text{Find } \mathbf{x}_0 \quad \text{such that} \quad \mathbf{q}(\mathbf{x}_0, \mathbf{x}_T(\mathbf{x}_0)) = 0$$

For some small problems, a trial-and-error approach may be sufficient. However, as the number of variables and the level of complexity increase, a systematic method such as Newton's method is preferable.[4]

Newton's method uses an initial guess, $\mathbf{x}_0^{(0)}$, and improves the value of $\mathbf{x}_0$ iteratively using the following update equation:

$$\mathbf{x}_0^{(k+1)} = \mathbf{x}_0^{(k)} + \Delta\mathbf{x}_0^{(k)} \tag{G.34}$$

where,

$$\Delta\mathbf{x}_0^{(k)} = -J^{-1}\mathbf{q}\left(\mathbf{x}_0^{(k)}, \mathbf{x}_T\left(\mathbf{x}_0^{(k)}\right)\right) \tag{G.35}$$

$$J = \left.\frac{d\mathbf{q}}{d\mathbf{x}_0}\right|_{\mathbf{x}_0 = \mathbf{x}_0^{(k)}} \tag{G.36}$$

Once $\mathbf{q}\left(\mathbf{x}_0^{(k)}, \mathbf{x}_T\left(\mathbf{x}_0^{(k)}\right)\right)$ is close to zero, we can set $\mathbf{x}_0 = \mathbf{x}_0^{(k)}$ and solve for $\mathbf{x}(t)$ from $t = 0$ to $t = T$. If the number of iterations exceeds a maximum, then either a better initial guess is required or a different method needs to be explored.

The terms in (G.36) generate a companion set of initial value problem. Specifically, J is the square Jacobian matrix of $\mathbf{q}$. The added complexity stems from the dependence of $\mathbf{q}$ on $\mathbf{x}_T$, which in turn depends on $\mathbf{x}_0$ through the integration process of (G.31).

Let the boundary conditions be given as

$$\mathbf{q}(\mathbf{x}_0, \mathbf{x}_T) = \begin{pmatrix} q_1(x_{01}, \ldots, x_{0n}, x_{T1}, \ldots, x_{Tn}) \\ \vdots \\ q_n(x_{01}, \ldots, x_{0n}, x_{T1}, \ldots, x_{Tn}) \end{pmatrix} = \mathbf{0} \tag{G.37}$$

[4] The Newton-search approach is not guaranteed to converge for all systems. It is a local scheme and thus requires a good initial guess.

then

$$\begin{aligned}\frac{d\mathbf{q}}{d\mathbf{x}_0} &= \left(\frac{\partial \mathbf{q}(\mathbf{a},\mathbf{b})}{\partial \mathbf{a}}\frac{\partial \mathbf{a}}{\partial \mathbf{x}_0}\right)_{\mathbf{a}=\mathbf{x}_0,\mathbf{b}=\mathbf{x}_T} + \left(\frac{\partial \mathbf{q}(\mathbf{a},\mathbf{b})}{\partial \mathbf{b}}\frac{d\mathbf{b}}{d\mathbf{x}_0}\right)_{\mathbf{a}=\mathbf{x}_0,\mathbf{b}=\mathbf{x}_T} \\ &= Q_a(\mathbf{x}_0,\mathbf{x}_T) + Q_b(\mathbf{x}_0,\mathbf{x}_T)\,\mathbf{M}(T) \end{aligned} \tag{G.38}$$

where,

$$Q_a = \begin{pmatrix} \eta_{11} & \cdots & \eta_{1n} \\ \vdots & \ddots & \vdots \\ \eta_{n1} & \cdots & \eta_{nn} \end{pmatrix} \tag{G.39}$$

$$\eta_{ij} = \left.\frac{\partial q_i(a_1,\ldots,a_n,b_1,\ldots,b_n)}{\partial a_j}\right|_{a_k=\mathbf{x}_{0k},b_\ell=\mathbf{x}_{T\ell}} \tag{G.40}$$

$$Q_b = \begin{pmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & \ddots & \vdots \\ \omega_{n1} & \cdots & \omega_{nn} \end{pmatrix} \tag{G.41}$$

$$\omega_{ij} = \left.\frac{\partial q_i(a_1,\ldots,a_n,b_1,\ldots,b_n)}{\partial b_j}\right|_{a_k=\mathbf{x}_{0k},b_\ell=\mathbf{x}_{T\ell}} \tag{G.42}$$

$$\mathbf{M}(T) = \frac{d\mathbf{x}_T}{d\mathbf{x}_0} \tag{G.43}$$

To determine $\mathbf{M}(T)$, take the derivative of the original differential equation (G.31) with respect to $\mathbf{x}_0$,

$$\begin{aligned}\frac{d}{d\mathbf{x}_0}\left(\frac{d}{dt}\mathbf{x}\right) &= \frac{d}{d\mathbf{x}_0}\mathbf{F}(t,\mathbf{x}) \\ \frac{d}{dt}\left(\frac{d\mathbf{x}}{d\mathbf{x}_0}\right) &= \frac{\partial \mathbf{F}}{\partial \mathbf{x}}\frac{d\mathbf{x}}{d\mathbf{x}_0} \\ \frac{d}{dt}\mathbf{M}(t) &= \mathbf{A}(t,\mathbf{x})\,\mathbf{M}(t) \end{aligned} \tag{G.44}$$

where

$$\mathbf{M}(t) = \frac{d\mathbf{x}}{d\mathbf{x}_0} \tag{G.45}$$

$$\mathbf{A}(t,\mathbf{x}) = \frac{\partial \mathbf{F}}{\partial \mathbf{x}} \tag{G.46}$$

and

$$\mathbf{M}(0) = I \tag{G.47}$$

Note that $\mathbf{A}(t,\mathbf{x})$ depends on the $\mathbf{x}$ consistent with the $\mathbf{x}_0$ used. Thus the following integration needs to be performed simultaneously:

$$\frac{d}{dt}\mathbf{x} = \mathbf{F}(t,\mathbf{x}) \qquad \mathbf{x}(0) = \mathbf{x}_0 \tag{G.48}$$

$$\frac{d}{dt}\mathbf{M} = \mathbf{A}(t,\mathbf{x})\,\mathbf{M} \qquad \mathbf{M}(0) = I \tag{G.49}$$

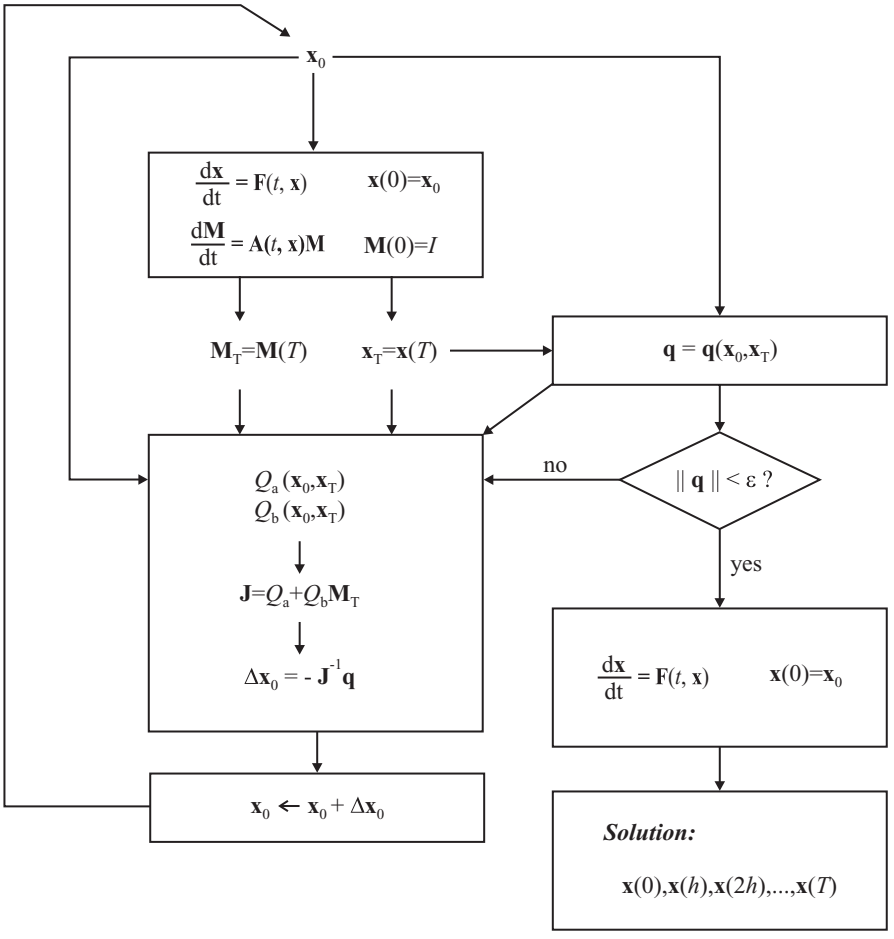


Figure G.4. A flowchart for nonlinear shooting implemented with Newton's method.

Having calculated $\mathbf{x}_T = \mathbf{x}(T)$ and $\mathbf{M}(T)$, we can then substitute these values together with $\mathbf{x}_0$ to determine Q_a, Q_b and $d\mathbf{q}/d\mathbf{x}_0$. Thereafter, the update to $\mathbf{x}_0$ can be determined, and the iteration continues until the desired tolerance on $\|\mathbf{q}\|$ is obtained. A flowchart showing the calculation sequences is given in Figure G.4.

EXAMPLE G.3. Consider the following set of differential equations:

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -k_1 e^{-2t} x_1 x_2 + k_3 \\ -k_1 e^{-2t} x_1 x_2 + k_2 x_3 \\ k_1 e^{-2t} x_1 x_2 - k_2 x_3 \end{pmatrix}$$

subject to the following boundary conditions:

$$\mathbf{q}(\mathbf{x}(0), \mathbf{x}(T)) = \begin{pmatrix} x_1(0) - x_2(T) - 0.164 \\ x_2(0) x_2(T) - 0.682 \\ x_3(0) + x_3(T) - 1.136 \end{pmatrix} = \mathbf{0}$$

with $T = 2$, $k_1 = 10$, $k_2 = 3$ and $k_3 = 1$.

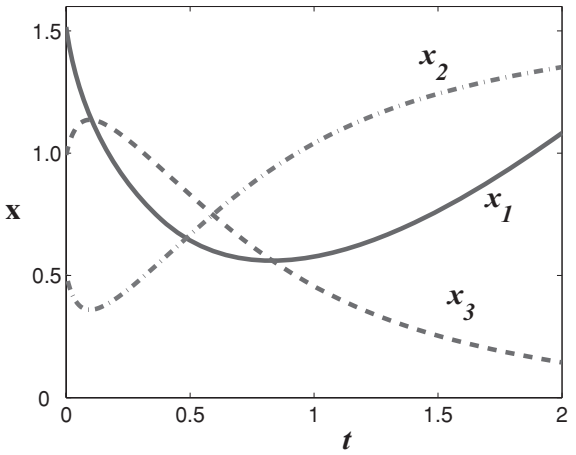


Figure G.5. Solution for boundary value problem given in Example G.3 .

We can calculate Q_a and Q_b to be in a form that can be evaluated readily based on values of $\mathbf{x}_0$ and $\mathbf{x}_T$,

$$Q_a = \begin{pmatrix} 1 & 0 & 0 \\ 0 & x_2(T) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$Q_b = \begin{pmatrix} 0 & -1 & 0 \\ 0 & x_2(0) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Similarly, we can calculate $\mathbf{A} = \partial \mathbf{F}/\partial \mathbf{x}$,

$$\mathbf{A}(t, \mathbf{x}) = \begin{pmatrix} \left(-k_1 e^{-2t} x_2\right) & \left(-k_1 e^{-2t} x_1\right) & 0 \\ \left(-k_1 e^{-2t} x_2\right) & \left(-k_1 e^{-2t} x_1\right) & k_2 \\ \left(k_1 e^{-2t} x_2\right) & \left(k_1 e^{-2t} x_1\right) & -k_2 \end{pmatrix}$$

Using an initial guess of $\mathbf{x}_0^{(0)} = (1, 1, 1)^T$ and a tolerance of $\epsilon = 1 \times 10^{-10}$, it took five iterations to converge to the following initial and final conditions:

$$\mathbf{x}_0 = \begin{pmatrix} 1.516 \\ 0.504 \\ 0.992 \end{pmatrix} \qquad \mathbf{x}_T = \begin{pmatrix} 1.083 \\ 1.352 \\ 0.144 \end{pmatrix}$$

Plots of the solutions are shown in Figure G.5. (A MATLAB file `nbvp.m` is available on the book's webpage and solves this specific example. The code contains sections that are customizable to apply to different nonlinear boundary value problems.)

G.8 Ricatti Equation Method

Consider the linear differential equation,

$$\frac{d}{dt}\mathbf{x} = \mathbf{A}(t)\mathbf{x} + \mathbf{b}(t) \tag{G.50}$$

with separated boundary conditions such that k conditions are specified at $t = 0$ and $(n - k)$ conditions are specified at $t = T$,

$$Q_0\,\mathbf{x}(0) = \beta_0 \tag{G.51}$$

$$Q_T\,\mathbf{x}(T) = \beta_T \tag{G.52}$$

where Q_0 is a $k \times n$ matrix of constants and Q_T is an $(n - k) \times n$ matrix of constants.

As an alternative to the shooting method, we look for a transformation of the original state variable given by

$$\mathbf{x}(t) = \mathbf{S}(t)\mathbf{z}(t) \tag{G.53}$$

where $\mathbf{S}(t)$ is an $n \times n$ transformation matrix and $\mathbf{z}(t)$ is the new state vector. The aim of the transformation is to recast the original problem into a partially decoupled problem such that the solution of first k values of $\mathbf{z}$ can be solved independently of the last $(n - k)$ values of $\mathbf{z}$, that is,

$$\frac{d}{dt}\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} H_{11}(t) & 0 \\ H_{21}(t) & H_{22}(t) \end{pmatrix}\begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{q}_1(t) \\ \mathbf{q}_2(t) \end{pmatrix} \tag{G.54}$$

where $\mathbf{z}_1(t)[=]k \times 1$ and $\mathbf{z}_2(t)[=](n - k) \times 1$. In addition, the transformation will be done such that the $\mathbf{z}_1(0)$ can be specified from (G.51), whereas $\mathbf{z}_2(T)$ can be specified from (G.52).

Thus $\mathbf{z}_1$ is first solved using initial value solvers to determine $\mathbf{z}_1(t = T)$. Afterward, $\mathbf{z}_1(T)$ is combined with $\mathbf{z}_2(T)$, after using (G.52), to form $\mathbf{z}(t = T)$. The terminal condition for $\mathbf{x}(t)$ at $t = T$ can then be found from (G.53). Having $\mathbf{x}(T)$, the trajectory of $\mathbf{x}(t)$ can be evaluated by integrating backward from $t = T$ to $t = 0$.

To obtain the form in (G.54), we first apply (G.53) to the original equation, (G.50),

$$\begin{aligned} \left(\frac{d}{dt}\mathbf{S}\right)\mathbf{z} + \mathbf{S}\frac{d}{dt}\mathbf{z} &= \mathbf{A}\mathbf{S}\mathbf{z} + \mathbf{b} \\ \frac{d}{dt}\mathbf{z} &= \mathbf{S}^{-1}(t)\left(\mathbf{A}(t)\mathbf{S}(t) - \frac{d}{dt}\mathbf{S}\right)\mathbf{z} + \mathbf{S}^{-1}(t)\mathbf{b}(t) \\ &= H(t)\mathbf{z} + \mathbf{q}(t) \end{aligned}$$

where

$$\begin{aligned} \mathbf{S}^{-1}(t)\left(\mathbf{A}(t)\mathbf{S}(t) - \frac{d}{dt}\mathbf{S}\right) &= H(t) \\ \frac{d}{dt}\mathbf{S} &= \mathbf{A}(t)\mathbf{S}(t) - \mathbf{S}(t)H(t) \end{aligned} \tag{G.55}$$

and

$$\mathbf{q}(t) = \mathbf{S}^{-1}\mathbf{b}(t) \tag{G.56}$$

We can choose $\mathbf{S}$ to be an upper triangular matrix given by

$$\mathbf{S} = \begin{pmatrix} I_k & R(t) \\ 0 & I_{n-k} \end{pmatrix} \tag{G.57}$$

whose inverse is given by

$$\mathbf{S}^{-1} = \begin{pmatrix} I_k & -R(t) \\ 0 & I_{n-k} \end{pmatrix} \tag{G.58}$$

After substitution of (G.57) into (G.55),

$$\begin{pmatrix} 0 & \frac{d}{dt}R \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} I & R(t) \\ 0 & I \end{pmatrix} - \begin{pmatrix} I & R(t) \\ 0 & I \end{pmatrix}\begin{pmatrix} H_{11} & 0 \\ H_{21} & H_{22} \end{pmatrix}$$
$$= \begin{pmatrix} A_{11} - (H_{11} + RH_{21}) & A_{11}R + A_{12} - RH_{22} \\ A_{21} - H_{21} & A_{21}R + A_{22} - H_{22} \end{pmatrix}$$

By comparing elements on both sides, we have the following equations

$$\begin{aligned} H_{21} &= A_{21} \\ H_{22} &= A_{21}R + A_{22} \\ H_{11} &= A_{11} - RH_{21} = A_{11} - RA_{21} \\ \frac{d}{dt}R &= A_{11}R + A_{12} - RA_{21}R - RA_{22} \end{aligned} \tag{G.59}$$

where the last equation is a matrix Ricatti equation.

Because H_{11} depends on $R(t)$, we need to solve for $\mathbf{z}_1$ and R using the first k equations of (G.54) and (G.59) as initial value problems, that is,

$$\begin{aligned} \frac{d}{dt}\mathbf{z}_1 &= (A_{11} - RA_{21})\,\mathbf{z}_1 + \mathbf{q}_1 \\ \frac{d}{dt}R &= A_{11}R + A_{12} - RA_{21}R - RA_{22} \end{aligned} \tag{G.60}$$

Note that $\mathbf{z}_1$ is a vector, whereas R is a matrix. To determine the required initial conditions, we can find $\mathbf{z}_1(0)$ in terms of $R(0)$ using (G.53) and (G.58),

$$\mathbf{z}_1(0) = \left(\begin{array}{c|c} I_k & -R \end{array}\right)\mathbf{x}(0) \tag{G.61}$$

Assume that the first k columns of Q_0 in (G.51) are linearly independent[5]; that is, let C be the nonsingular matrix consisting of the first k columns of Q_0, then

$$Q_0\mathbf{x}(0) = C\left(\begin{array}{c|c} I & C^{-1}D \end{array}\right)\mathbf{x}(0) = \beta_0$$

Next, choose $R(0) = -C^{-1}D$ and premultiply $\mathbf{z}_1(0)$ (in (G.61)) by C,

$$C\mathbf{z}_1(0) = C\left(\begin{array}{c|c} I & C^{-1}D \end{array}\right)\mathbf{x}(0) = \beta_0 \qquad \rightarrow \qquad \mathbf{z}_1(0) = C^{-1}\beta_0 \tag{G.62}$$

In summary, the first phase, known as the **forward-sweep phase** of the Ricatti equation method, is to solve for $R(t)$ and $\mathbf{z}_1(t)$, that is,

$$\frac{d}{dt}R = A_{11}R + A_{12} - RA_{21}R - RA_{22} \quad ; \quad R(0) = -C^{-1}D \tag{G.63}$$

where $Q_0 = \left(\begin{array}{c|c} C & D \end{array}\right)$, followed by

$$\frac{d}{dt}\mathbf{z}_1 = (A_{11} - RA_{21})\,\mathbf{z}_1 + \left(\begin{array}{c|c} I & -R \end{array}\right)\mathbf{b} \quad ; \quad \mathbf{z}_1(0) = C^{-1}\beta_0 \tag{G.64}$$

and integrate until $t = T$ to obtain the values of $\mathbf{z}_1(T)$ and $R(T)$.

[5] If the first k columns of Q_0 are not invertible, a reordering of $\mathbf{x}$ may be required.

The second phase of the method is to find the conditions for $\mathbf{x}(T)$ by combining the results from the first phase with the other set of boundary conditions given by (G.52). By partitioning Q_T as

$$Q_T = \left(\begin{array}{c|c} F & G \end{array} \right)$$

where F is $(n-k) \times n$ and G is $(n-k) \times (n-k)$, we get

$$\begin{aligned} Q_T \mathbf{x}(T) = \left(\begin{array}{c|c} F & G \end{array} \right) \begin{pmatrix} I & R(T) \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbf{z}_1(T) \\ \mathbf{z}_2(T) \end{pmatrix} &= \beta_T \\ F\mathbf{z}_1 + FR(T)\mathbf{z}_2(T) + G\mathbf{z}_2(T) &= \beta_T \\ \mathbf{z}_2(T) = (FR(T) + G)^{-1} (\beta_T - F\mathbf{z}_1(T)) & \qquad \text{(G.65)} \end{aligned}$$

which can be used to form $\mathbf{x}(T)$, that is,

$$\mathbf{x}(T) = \begin{pmatrix} I & R(T) \\ 0 & I \end{pmatrix} \begin{pmatrix} \mathbf{z}(T) \end{pmatrix} \qquad \text{(G.66)}$$

Having evaluated $\mathbf{x}(T)$ means we now have all the information at one boundary. We could then use the original differential equations given in (G.50) and integrate backward starting from $t = T$ until $t = 0$. This second phase is also known as the **backward-sweep phase** of the Ricatti equation method.

The Ricatti equation method (which is also sometimes called the **invariant embedding method**) is sometimes more stable than the shooting method, especially when the process (7.55) is unstable. However, there are also situations when the shooting methods turn out to be more stable. Thus both methods may need to be explored in case one or the other does not yield good results. Note also that our development of the Ricatti equation method is limited to cases with separated boundary conditions.

APPENDIX H

Additional Details and Fortification for Chapter 8

H.1 Bifurcation Analysis

The behavior around a non-hyperbolic equilibrium point can change under slight modifications of the process parameters. Under these conditions, the system is classified as **structurally unstable**. By perturbing the parameter slightly, the characteristics can sometimes yield additional equilibrium points and can change the stability of equilibrium points. **Bifuraction analysis** is the study of how the structural behaviors of the system are affected by variations in the key parameters.

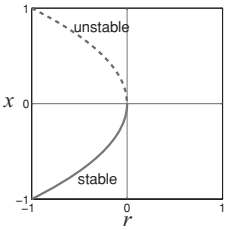
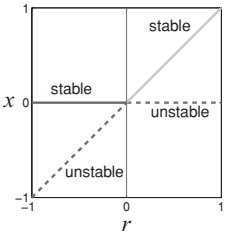
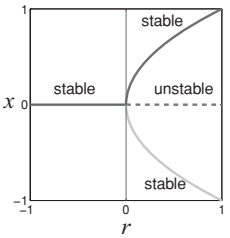
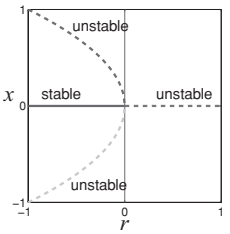
For the one-dimensional case, there are three main types of bifurcations. A summary of the different types of bifurcations for one-dimensional systems is given in Table H.1. Included in the table are the normal forms and the corresponding bifurcation diagram. The bifurcation diagrams show the locus of equilibrium points, if they exist, at different values of parameter r. We use the convention that represents the locus of stable equilibrium points by solid curves and the locus of unstable equilibrium points by dashed curves.

The first type of bifurcation is the saddle-node. **Saddle-node bifurcations** are characterized by the absence of equilibrium points to one side of the non-hyperbolic equilibrium point, and saddle-node bifurcations are also known as **blue-sky bifurcations** to highlight the sudden appearance of equilibrium points as if they appeared "out of the sky." The term "saddle-node" is more appropriate for the 2D case. The second type of bifurcation is the transcritical bifurctation. **Transcritical bifurcations** are characterized by the intersection of two locus of equilibrium points at a non-hyperbolic point. After both curves cross each other, their stability switch from stable to unstable and vice versa. The third type of bifurcation is the pitchfork bifurcation. **Pitchfork bifurcations** are characterized by additional equilibrium points as they cross the non-hyperbolic equilibrium point from a single locus curve of stable (supercritical) or unstable (subcritical) equilibrium points. The name of this bifurcation comes from the bifurcation diagram (as shown in Table H.1) resembling a pitchfork.

For cases that are more general than the given normal forms, let $\dot{x} = f(x, r)$ where $x = 0$ is a non-hyperbolic equilibrium point at $r = 0$. A Taylor series expansion around $(x, r) = (0, 0)$ is given by

$$f(x, r) = f(0, 0) + x\frac{\partial f}{\partial x} + r\frac{\partial f}{\partial r} + \frac{x^2}{2}\frac{\partial^2 f}{\partial x^2} + \frac{r^2}{2}\frac{\partial^2 f}{\partial r^2} + rx\frac{\partial^2 f}{\partial r \partial x} + \cdots$$

Table H.1. *Types of bifurcations for one-dimensional systems*

	Type	Normal form	Bifurcation diagram
1	Saddle-Node	$\dot{x} = r + x^2$	
2	Transcritical	$\dot{x} = rx - x^2$	
3	Pitchfork	$\dot{x} = rx - x^3$	(Supercritical) 
		$\dot{x} = rx + x^3$	(Subcritical) 

where all the various partial derivatives are evaluated at $(x, r) = (0, 0)$. Because $(x, r) = (0, 0)$ is a non-hyperbolic equilibrium point, the first two terms are zero, that is, $f(0, 0) = 0$ and $\partial f/\partial x(0, 0) = 0$.

We will truncate the series after the second-order derivatives to yield bifurcation analysis of saddle-node bifurcations and transcritical bifurcations. This means that equilibrium points near $(x, r) = (0, 0)$ will be given by the roots of the second-order polynomial in x,

$$\alpha_2(r)x^2 + \alpha_1(r)x + \alpha_0(r) = 0 \tag{H.1}$$

where

$$\begin{aligned}\alpha_2(r) &= \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\\ \alpha_1(r) &= r\frac{\partial^2 f}{\partial r \partial x}\\ \alpha_0(r) &= r\frac{\partial f}{\partial r} + \frac{r^2}{2}\frac{\partial^2 f}{\partial r^2}\end{aligned}$$

which was obtained by setting the right-hand side of the Taylor series expansion to zero. Solving for the roots of (H.1), we find the neighboring equilibrium points around $(x, r) = (0, 0)$,

$$x_{eq} = \frac{-r\dfrac{\partial^2 f}{\partial r \partial x} \pm \sqrt{r^2\left(\dfrac{\partial^2 f}{\partial r \partial x}\right)^2 - 4\left(\dfrac{1}{2}\dfrac{\partial^2 f}{\partial x^2}\right)\left(r\dfrac{\partial f}{\partial r} + \dfrac{r^2}{2}\dfrac{\partial^2 f}{\partial r^2}\right)}}{\dfrac{\partial^2 f}{\partial x^2}} \tag{H.2}$$

For saddle-node bifurcations, consider $|r| \ll 1$. Then (H.2) will reduce to

$$x_{eq}\big|_{\text{saddle-node}} = \pm\sqrt{-2r\left(\frac{\partial f}{\partial r}\right)\left(\frac{\partial^2 f}{\partial x^2}\right)^{-1}} \tag{H.3}$$

which then requires

$$r\left(\frac{\partial f}{\partial r}\right)\left(\frac{\partial^2 f}{\partial x^2}\right)^{-1} < 0 \tag{H.4}$$

for equilibrium points to exist.

For transcritical bifurcations, we set an additional condition that $\partial f/\partial r(0, 0) = 0$. Then (H.2) reduces to

$$x_{eq} = r\frac{-\dfrac{\partial^2 f}{\partial r \partial x} \pm \sqrt{\left(\dfrac{\partial^2 f}{\partial r \partial x}\right)^2 - \left(\dfrac{\partial^2 f}{\partial x^2}\right)\left(\dfrac{\partial^2 f}{\partial r^2}\right)}}{\dfrac{\partial^2 f}{\partial x^2}} \tag{H.5}$$

A pair of equilibrium points will then exist if the value inside the square root is positive, plus $\partial^2 f/\partial x^2 \neq 0$, that is,

$$\left(\frac{\partial^2 f}{\partial r \partial x}\right)^2 - \left(\frac{\partial^2 f}{\partial x^2}\right)\left(\frac{\partial^2 f}{\partial r^2}\right) > 0 \qquad \text{and} \qquad \frac{\partial^2 f}{\partial x^2} \neq 0 \tag{H.6}$$

As r changes sign, the stability of the equilibrium points will switch, thereby giving the character of transcritical bifurcations.

For both saddle-node and transcritical bifurcations, the stability can be assessed by regrouping the Taylor series approximation as

$$\dot{x} \approx \alpha_0(r) + \Big(\alpha_1(r) + \alpha_2(r)x\Big)\,x = \alpha_0(r) + \beta_{(x,r)}x$$

where

$$\beta_{(x,r)} = \alpha_1(r) + \alpha_2(r)x$$

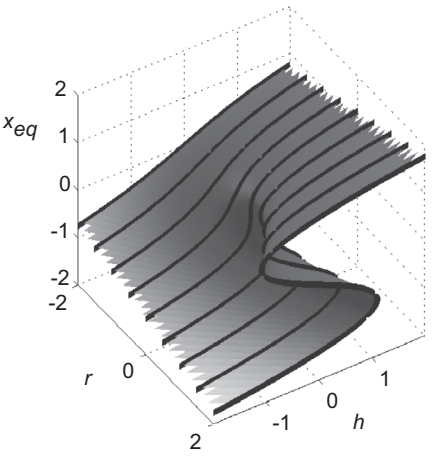


Figure H.1. Two-parameter bifurcation diagram.

Then applying the formulas for x_{eq} (Equation (H.3) for saddle-node bifurcations and Equation (H.5) for transcritical bifurcations), we find that

$$x_{eq,i} \quad \text{is stable if} \quad \beta_{(x_{eq,i},r)} < 0 \quad \text{for } i = 1, 2$$

For pitchfork bifurcations, the Taylor series will need to include third-order derivatives such that a third-order polynomial can be obtained for the equilibrium points. The computations are lengthier, but with the additional condition that $\partial f/\partial r(0, 0) = 0$ and $\partial^2 f/\partial x^2(0, 0) = 0$, the conditions simplify to the following conditions

$$r\frac{\partial^2 f}{\partial x \partial r}\frac{\partial^3 f}{\partial x^3} \quad \begin{cases} > 0 & \text{for single equilibrium points} \\ < 0 & \text{for three equilibrium points} \end{cases} \tag{H.7}$$

It is important to remember that all the partial derivatives given in the conditions (H.4), (H.6), and (H.7) are evaluated at $(x, r) = (0, 0)$.

Aside from the three type of bifurcations discussed thus far, the introduction of one more parameter can also make the bifurcations change, including the addition or removal of non-hyperbolic equilibrium points. This situation is known as **codimension two bifurcations**. An example of these types of bifurcation is the catastrophe model given by

$$\dot{x} = f(x, r, h) = x^3 - rx - h \tag{H.8}$$

where r and h are parameters. A surface locus of equilibrium points is shown in Figure H.1. In the figure, we see that the surface has a continuous fold, and thus, dependent on the values of r and h, there can be either one, two, or three equilibrium points. These regions can be separated by two intersecting curves as shown in the (r, h) plane as shown in Figure H.2. The point where the two separating curves intersect is known as the **cusp point**. Many physical phenomenon, such as phase changes of material, that is, vapor liquid equilibria, are described by these types of bifurcations or catastrophe models.

Next, consider the bifurcation diagram for x_{eq} at $r = 2$ as shown in Figure H.3. When $r = 2$, there are two non-hyperbolic equilibrium points: one at $(x, h) = (0.816, -1.089)$ and another at $(x, h) = (-0.816, 1.089)$, both of which yield

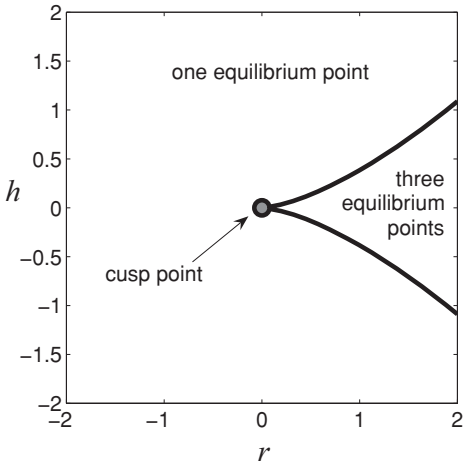


Figure H.2. Phase diagram in the (r, h)-plane.

saddle-node bifurcations. When $h > -1.089$ and gradually decreased, the equilibrium points following the top curve in Figure H.3 will also decrease continuously. However, as h moves past the critical value $h = -1.089$, the equilibrium point will jump to follow the values of the lower curve. The opposite thing happens for the lower curve; that is, as h gradually increases until it passes the value of 1.089, the equilibrium point jumps to follow the upper locus of equilibrium points. This characteristic of having the behavior depend on the direction of parameter change is known as **hysteresis**.

The bifurcations of second-order systems include all three types of the first-order cases, namely saddle-node, transcritical, and pitchfork bifurcations. These three types of bifurcations are extended by means of simply adding one more differential equation. The canonical forms are given in Table H.2. These types of bifurcations are centered at non-hyperbolic equilibrium points that have zero eigenvalues.

The **Hopf bifurcation** is a type of bifurcation that is not available to one-dimensional systems because it involves pure imaginary eigenvalues. These bifurcations yield the appearance or disappearance of limit cycles. A supercritical Hopf bifurcation occurs when a stable focus can shift to a stable limit cycle. Conversely, a subcritical Hopf bifurcation occurs when an unstable limit cycle changes to an

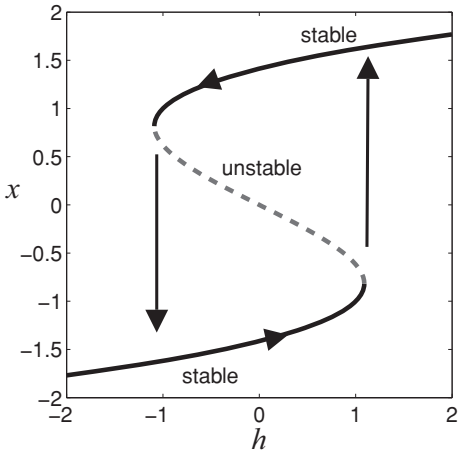


Figure H.3. Bifurcation diagram for (x, h) when $r = 2$.

Table H.2. *Normal forms for Bifurcations of 2D Systems*

	Type	Normal form
1	Saddle-Node	$\dot{y} = -y$ $\dot{x} = r + x^2$
2	Transcritical	$\dot{y} = -y$ $\dot{x} = rx - x^2$
3	Pitchfork	Supercritical: $\dot{y} = -y$ $\dot{x} = rx - x^3$ Subcritical: $\dot{y} = -y$ $\dot{x} = rx + x^3$
4	Hopf	$\dot{\theta} = \omega$ $\dot{\rho} = \mu\rho + a\rho^3$ Supercritical: $a < 0$ Subcritical: $a > 0$

unstable focus. The canonical form of a Hopf bifucation given in terms of polar coordinates (ρ, θ),

$$\frac{d\theta}{dt} = \omega \qquad \text{and} \qquad \frac{d\rho}{dt} = \mu\rho + a\rho^3$$

where $\rho = \sqrt{x^2 + y^2}$ and $\theta = \tan^{-1}(y/x)$. It can be shown that when $a < 0$, the system exhibits a supercritical Hopf bifurcation. However, when $a > 0$, the system exhibits a subcritical Hopf bifurcation. These are shown in Figures H.4.

It turns out that Hopf bifurcations can occur for orders ≥ 2. A general theorem is available that prescribes a set of sufficient conditions for the existence of a Hopf bifurcation.

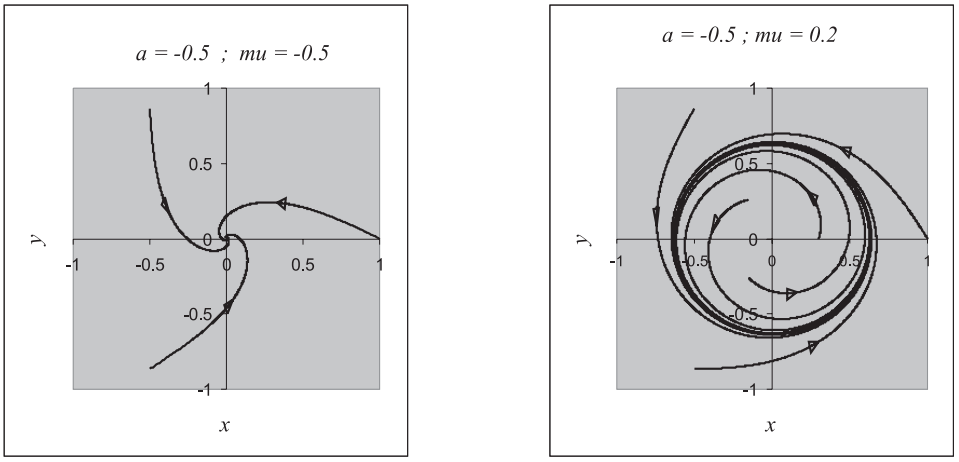
THEOREM H.1. *Let λ_h be a value of parameter λ such that the system $d\mathbf{x}/dt = \mathbf{f}(\mathbf{x}; \lambda)$ has an equilibrium point $\mathbf{x}_{eq}(\lambda_h)$ with the Jacobian matrix $J = d\mathbf{f}/d\mathbf{x}$ at $\mathbf{x} = \mathbf{x}_{eq}(\lambda_h)$ having a pair of pure imaginary eigenvalues, $\pm i\,\mu(\lambda_h)$ $(i = \sqrt{-1})$, whereas the rest of the eigenvalues have nonzero real parts. In addition, let the real and imaginary parts of the eigenvalues $\mu(\lambda)$ be smooth functions of parameter λ in which*

$$\frac{d}{d\lambda}\Big(Re(\mu(\lambda))\Big) \neq 0$$

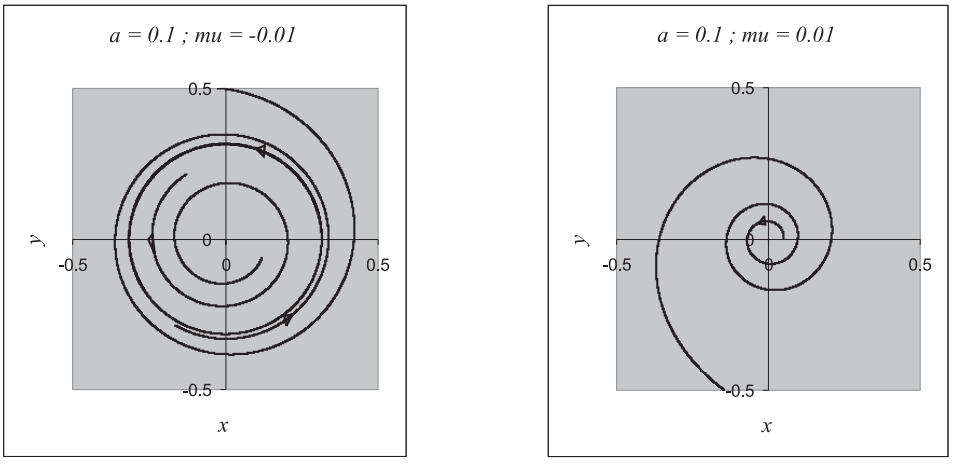
in a neighborhood around λ_h. Under these conditions, the system will have a Hopf bifurcation at $\lambda = \lambda_h$.

There are several physical systems that exhibit Hopf bifurcations, such as in the fields of biomedical science, aeronautics, fluid mechanics, and chemistry.[1] In

[1] A good elementary treatment of Hopf bifurcations, including several examples and exercises, can be found in S. Strogatz, *Nonlinear Dynamics and Chaos*, Perseus Book Publishing, Massachusetts, 1994.



(a) Supercritical Hopf bifurcations.



(b) Subcritical Hopf bifurcations.

Figure H.4. Phase plane plots showing supercritical and subcritical Hopf bifurcations. ($\omega = 0.5$).

chemistry, there are several well-known reaction systems, such as the Belousov-Zhabotinsky (BZ) system, known collectively as **oscillating chemical reactions**. Depending on the critical conditions, the systems can oscillate spontaneously. One of the well-known examples of a Hopf bifurcation is the Brusselator reaction, which is given in Exercise **E8.19**. Although it is strictly fictitious, its simplification still allows one to understand the onset of Hopf bifurcations in real systems.

APPENDIX I

Additional Details and Fortification for Chapter 9

I.1 Details on Series Solution of Second-Order Systems

For $N = 2$, the differential equation for which $x = 0$ is a regular singular point is given by

$$x^2 \tilde{P}_2(x) \frac{d^2y}{dx^2} + x\tilde{P}_1(x) \frac{dy}{dx} + \tilde{P}_0(x)y = 0 \tag{I.1}$$

where

$$\begin{aligned} \widetilde{P}_2(x) &= \widetilde{\rho}_{2,0} + \widetilde{\rho}_{2,1}x + \widetilde{\rho}_{2,2}x^2 + \cdots \\ \widetilde{P}_1(x) &= \widetilde{\rho}_{1,0} + \widetilde{\rho}_{1,1}x + \widetilde{\rho}_{1,2}x^2 + \cdots \\ \widetilde{P}_0(x) &= \widetilde{\rho}_{0,0} + \widetilde{\rho}_{0,1}x + \widetilde{\rho}_{0,2}x^2 + \cdots \end{aligned} \tag{I.2}$$

and $\widetilde{\rho}_{2,0} \neq 0$.

The indicial equation (9.28) becomes

$$\begin{aligned} \widetilde{\rho}_{0,0} + \widetilde{\rho}_{1,0}r + \widetilde{\rho}_{2,0}(r)(r-1) &= 0 \\ \widetilde{\rho}_{2,0}r^2 + (\widetilde{\rho}_{1,0} - \widetilde{\rho}_{2,0})\, r + \widetilde{\rho}_{0,0} &= 0 \end{aligned} \tag{I.3}$$

and the indicial roots are

$$r = \frac{(\widetilde{\rho}_{2,0} - \widetilde{\rho}_{1,0}) \pm \sqrt{(\widetilde{\rho}_{2,0} - \widetilde{\rho}_{1,0})^2 - 4\widetilde{\rho}_{0,0}\widetilde{\rho}_{2,0}}}{2\widetilde{\rho}_{2,0}} \tag{I.4}$$

We denote the larger root (if real) by r_a and the other root by r_b.

When the roots differ by an integer, say $r_a - r_b = m \geq 0$,

$$\begin{aligned} r_a + r_b &= 1 - \frac{\widetilde{\rho}_{1,0}}{\widetilde{\rho}_{2,0}} \\ 2r_a - m &= \\ r_a &= \frac{1}{2}\left(m + 1 - \frac{\widetilde{\rho}_{1,0}}{\widetilde{\rho}_{2,0}}\right) \end{aligned} \tag{I.5}$$

When the roots are equal, $m = 0$,

$$r_a = \frac{1}{2}\left(1 - \frac{\widetilde{\rho}_{1,0}}{\widetilde{\rho}_{2,0}}\right) \tag{I.6}$$

Using r_a, we are guaranteed one solution, which we will denote by $u(x)$,

$$u(x) = \sum_{n=0}^{\infty} \widetilde{\phi}_n(r_a) x^{r_a+n} \tag{I.7}$$

where

$$\widetilde{\phi}_n(r_a) = \begin{cases} 1 & \text{if } n = 0 \\ -\dfrac{\sum_{k=0}^{n-1} Q_{n,k}(r_a)\widetilde{\phi}_k(r_a)}{Q_{n,n}(r_a)} & \text{if } n > 0 \end{cases}$$

$$Q_{n,k}(r_a) = \widetilde{\rho}_{0,n-k} + \widetilde{\rho}_{1,n-k}(k + r_a) + \widetilde{\rho}_{2,n-k}(k + r_a)(k + r_a - 1)$$

If $(r_a - r_b)$ is not an integer, the second solution, $v(x)$, is immediately given by

$$v(x) = \sum_{n=0}^{\infty} \widetilde{\phi}_n(r_b) x^{r_b+n} \tag{I.8}$$

where

$$\widetilde{\phi}_n(r_b) = \begin{cases} 1 & \text{if } n = 0 \\ -\dfrac{\sum_{k=0}^{n-1} Q_{n,k}(r_b)\widetilde{\phi}_k(r_b)}{Q_{n,n}(r_b)} & \text{if } n > 0 \end{cases}$$

$$Q_{n,k}(r_b) = \widetilde{\rho}_{0,n-k} + \widetilde{\rho}_{1,n-k}(k + r_b) + \widetilde{\rho}_{2,n-k}(k + r_b)(k + r_b - 1)$$

If the indicial roots differ by an integer, that is, $m \geq 0$, we can use the d'Alembert method of order reduction (cf. Lemma I.1 in Section I.2) to find the other solution. For $N = 2$, this means the second solution is given by

$$v(x) = u(x) \int z(x) dx \tag{I.9}$$

where $z(x)$ is an intermediate function that solves a first-order differential equation resulting from the d'Alembert order reduction method. Using $u(x)$ as obtained in (I.7), $z(x)$ can be obtained by solving

$$\begin{aligned} x^2 \widetilde{P}_2(x) u \frac{dz}{dx} + \left(2x^2 \widetilde{P}_2(x) \frac{du}{dx} + x \widetilde{P}_1(x) u \right) z &= 0 \\ \frac{1}{z}\frac{dz}{dx} &= -\left(\frac{\widetilde{P}_1(x)}{x\widetilde{P}_2(x)} + 2\frac{1}{u}\frac{du}{dx} \right) \end{aligned} \tag{I.10}$$

With u, $\widetilde{P}_2(x)$ and $\widetilde{P}_1(x)$ defined by equations (I.7) and (I.2), respectively, the left-hand side of (I.10) can be replaced by an infinite series,

$$\frac{1}{z}\frac{dz}{dx} = -\sum_{n=-1}^{\infty} (\alpha_n + \beta_n) x^n \tag{I.11}$$

where the terms α_n and β_n are defined as

$$\alpha_n = \begin{cases} 2r_a & \text{if } n = -1 \\ \left(2(r_a + n + 1)\widetilde{\phi}_{n+1}(r_a) - \sum_{k=-1}^{n-1} \alpha_k \widetilde{\phi}_{n-k}(r_a)\right) & \text{if } n \geq 0 \end{cases} \tag{I.12}$$

$$\beta_n = \begin{cases} \widetilde{\rho}_{1,0}/\widetilde{\rho}_{2,0} & \text{if } n = -1 \\ \left(\widetilde{\rho}_{1,n+1} - \sum_{k=-1}^{n-1} \beta_k \widetilde{\rho}_{2,n-k}\right)/\widetilde{\rho}_{2,0} & \text{if } n \geq 0 \end{cases} \tag{I.13}$$

For (I.12), we used the fact that $\widetilde{\phi}_0(r_a) = 1$.

For indicial roots differing by an integer, we can use (I.5), and the coefficient for first term involving $(1/x)$ in (I.11) becomes

$$\alpha_{-1} + \beta_{-1} = \frac{\widetilde{\rho}_{1,0}}{\widetilde{\rho}_{2,0}} + 2r_a = m + 1$$

Then returning to (I.11), z can be evaluated as follows:

$$\begin{aligned} \frac{1}{z}\frac{dz}{dx} &= -\left(\frac{m+1}{x} + \sum_{n=0}^{\infty} (\alpha_n + \beta_n)\, x^n\right) \\ \ln(z) &= -\left(\ln\left(x^{m+1}\right) + \sum_{n=0}^{\infty} \frac{(\alpha_n + \beta_n)}{n+1} x^{n+1}\right) \\ z &= x^{-(m+1)} \exp\left[-\sum_{n=0}^{\infty} \frac{(\alpha_n + \beta_n)}{n+1} x^{n+1}\right] \end{aligned}$$

We can also expand the exponential function as a Taylor series,

$$\exp\left[-\sum_{n=0}^{\infty} \frac{(\alpha_n - \beta_n)}{n+1} x^{n+1}\right] = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \cdots$$

Due to the complexity of the definitions of γ_i, $i = 1, 2, \ldots$, we just treat the γ_i's as constants for now. The Taylor series expansion is being used at this point only to find the form needed for the second independent solution. Once the solution forms are set, a direct substitution is used later to find the unknown coefficients. Thus we can rewrite z as

$$z = \begin{cases} \sum_{k=0}^{m-1} \gamma_k x^{k-m-1} \\ \quad + \gamma_m x^{-1} + \sum_{n=m+1}^{\infty} \gamma_n x^{n-m-1} & \text{if } m > 0 \\ \gamma_0 x^{-1} + \sum_{n=1}^{\infty} \gamma_n x^{n-1} & \text{if } m = 0 \end{cases}$$

and

$$\int z dx = \begin{cases} \sum_{k=0}^{m-1} (\gamma_k/(k-m))\, x^{k-m} \\ \quad + \gamma_m \ln|x| + \sum_{n=m+1}^{\infty} (\gamma_n/n - m)\, x^{n-m} & \text{if } m > 0 \\ \gamma_0 \ln|x| + \sum_{n=1}^{\infty} (\gamma_n/n)\, x^n & \text{if } m = 0 \end{cases}$$

This integral can now be combined with u to yield the form for the second independent solution, that is,

$$\begin{aligned} v(x) &= u(x)\int z\,dx \\ &= \left(\sum_{n=0}^{\infty}\widetilde{\phi}_n x^{r_a+n}\right)\int z\,dx \\ v(x) &= \begin{cases} \eta u\ln|x| + \sum_{n=0}^{\infty} b_n x^{r_b+n} & \text{if } m>0 \\ u\ln|x| + \sum_{n=1}^{\infty} b_n x^{r_b+n} & \text{if } m=0 \end{cases} \end{aligned} \tag{I.14}$$

Note that for $m = 0$, the infinite series starts at $n = 1$ and the coefficient of $(u \ln |x|)$ is one. The parameter η is set equal to 1 when $m = 0$ because η will later be combined with a constant of integration. However, when $m > 0$, η should not be fixed to 1, because $\eta = 0$ in some cases. Instead, we will set $b_0 = 1$ in anticipation of merging with the arbitrary constant of integration.

Having found the necessary forms of the second solution, Theorem 9.2 summarizes the general solution of a second-order linear differential equation that includes the recurrence formulas needed for the coefficients of the power series based on the Frobenius method.

I.2 Method of Order Reduction

For an N^{th}-order homogenous linear differential equation given by

$$\sum_{i=0}^{N}\Phi_i(x)\frac{d^i y}{dx^i} = 0 \tag{I.15}$$

Suppose we know one solution, say, $u(x)$, that solves (I.15). By introducing another function, $q(x)$, as a multiplier to $u(x)$, we can obtain

$$y = q(x)u(x) \tag{I.16}$$

as another solution to (I.15) that is linearly independent from u. To evaluate $q(x)$, we will need to solve another linear differential equation of reduced order as given in the following lemma:

LEMMA I.1. **d'Alembert's Method of Order Reduction**
Let $q(x)$ be given by

$$q(x) = \int z(x)dx \tag{I.17}$$

where $z(x)$ is the solution of an $(N-1)^{\text{th}}$ order differential equation given by

$$\sum_{i=1}^{N}F_i(x)\frac{d^{i-1}z}{dx^{i-1}} = 0 \tag{I.18}$$

with

$$F_i(x) = \sum_{k=i}^{N}\frac{k!}{(k-i)!i!}\Phi_k(x)\frac{d^{(k-i)}u}{dx^{(k-i)}} \tag{I.19}$$

and $u(x)$ is a known solution of (I.15). Then $y = q(x)u(x)$ is also a solution of (I.15).

PROOF. First, applying Leibnitz's rule (9.6) to the n^{th} derivative of the product $y = qu$,

$$\frac{d^i y}{dx^i} = \sum_{j=0}^{i} \left(\begin{array}{c} i \\ j \end{array} \right) \frac{d^j q}{dx^j} \frac{d^{(i-j)} u}{dx^{(i-j)}}$$

where

$$\left(\begin{array}{c} i \\ j \end{array} \right) = \frac{i!}{j!(i-j)!}$$

Substituting these derivatives into (I.15),

$$\begin{aligned} \sum_{i=0}^{N} \Phi_i(x) \sum_{j=0}^{i} \left(\begin{array}{c} i \\ j \end{array} \right) \frac{d^j q}{dx^j} \frac{d^{(i-j)} u}{dx^{(i-j)}} &= 0 \\ \left(q \sum_{i=0}^{N} \Phi_i(x) \frac{d^i u}{dx^i} \right) + \sum_{i=1}^{N} \Phi_i(x) \sum_{j=1}^{i} \left(\begin{array}{c} i \\ j \end{array} \right) \frac{d^j q}{dx^j} \frac{d^{(i-j)} u}{dx^{(i-j)}} &= 0 \end{aligned}$$

Because u satisfies (I.15), the first group of terms vanishes. The remaining terms can then be reindexed to yield

$$\sum_{i=1}^{N} \left(\sum_{k=i}^{N} \left(\begin{array}{c} k \\ i \end{array} \right) \Phi_k(x) \frac{d^{(k-i)} u}{dx^{(k-i)}} \right) \frac{d^i q}{dx^i} = 0$$

Letting $z = dq/dx$, we end up with an $(N-1)^{\text{th}}$ order linear differential equation in z.

This method can be used repeatedly for the reduced order differential equations. However, in doing so, we require that at least one solution is available at each stage of the order reductions. Fortunately, from the results of the previous section, it is always possible to find at least one solution for the differential equations using the Frobenius method. For instance, with $N = 3$, the Frobenius series method will generate one solution, say, u. Then via d'Alembert's method, another solution given by $y = qu$ produces a second-order differential equation for $z = dq/dt$. The Frobenius series method can generate one solution for this second-order equation, say, v. Applying the order reduction method one more time for $z = wv$, we end up with having to solve a first-order differential equation for w.[1]

Having solved for w, we can go backward:

$$\begin{aligned} z &= \alpha_1 v + \alpha_2 wv \\ q &= \alpha_1 \int v dx + \alpha_2 \int wv dx \\ y &= \beta_1 u + \beta_2 qu \\ &= \beta_1 u + \beta_2 \alpha_1 u \int v dx + \beta_2 \alpha_2 u \int wv dx \end{aligned}$$

[1] The resulting first-order differential equation is always of the separable type.

with α_1, α_2, β_1, and β_2 as arbitrary coefficients. Thus the approach of recursive order reduction can be used to generate the general solution for homogenous linear differential equation. One disclaimer to this solution approach is that, although the general solutions can be found in principle, the evaluation of the integrals via quadrature may still be difficult. This means that in case another simpler method is available, such as when all the indicial roots are distinct, those approaches should be attempted first.

I.3 Examples of Solution of Regular Singular Points

In this section, we have three examples to show how Theorem 9.2, which is the Frobenius series solution to linear second-order equations, is applied to the cases where $r_a - r_b$ is not an integer, $r_a - r_b = 0$, and $r_a - r_b = m$ is a positive integer.

EXAMPLE I.1. Given the equation

$$2x^2\frac{d^2y}{dx^2} + x(1-x)\frac{dy}{dx} - y = 0$$

The terms for $\widetilde{\rho}_{i,j}$ are $\widetilde{\rho}_{2,0} = 2$, $\widetilde{\rho}_{1,0} = 1$, $\widetilde{\rho}_{1,1} = -1$, and $\widetilde{\rho}_{0,0} = -1$, whereas the rest are zero. The indicial roots become $r_a = 1$ and $r_b = -0.5$. Because the difference is not an integer, $\eta = 0$ and $b_n = \widetilde{\phi}_n(r_b)$. The only nonzero values of $Q_{n,k}$ are

$$Q_{n,n}(r) = n\,(2n + 4r - 1) \qquad \text{and} \qquad Q_{n,n-1}(r) = -\,(n - 1 + r)$$

The recurrence formulas are then given by

$$\widetilde{\phi}_n(r) = \frac{n - 1 + r}{n\,(2n + 4r - 1)}\widetilde{\phi}_{n-1}(r) \qquad \text{where} \qquad n > 0$$

Thus

$$\begin{aligned}
\widetilde{\phi}_n(r_a) &= \frac{1}{2n+3}\widetilde{\phi}_{n-1}(r_a) = \left(\frac{1}{2n+3}\right)\left(\frac{1}{2n+1}\right)\cdots\left(\frac{1}{5}\right)\widetilde{\phi}_0 \\
&= \frac{(2n+2)(2n)\cdots 6\cdot 4!}{(2n+3)(2n+2)(2n+1)\cdots 5\cdot 4!} = 3\frac{2^{n+1}(n+1)!}{(2n+3)!} \\
\widetilde{\phi}_n(r_b) &= \frac{1}{2n}\phi_{n-1}(r_b) = \left(\frac{1}{2n}\right)\left(\frac{1}{2(n-1)}\right)\cdots\left(\frac{1}{2}\right)\widetilde{\phi}_0(r_b) = \frac{1}{2^n n!}
\end{aligned}$$

and the complete solution is given by

$$y(x) = A\sum_{n=0}^{\infty} 3\frac{2^{n+1}(n+1)!}{(2n+3)!}x^{n+1} + B\sum_{n=0}^{\infty}\frac{1}{2^n n!}x^{n-(1/2)}$$

This can be put in closed form as follows:

$$y(x) = A\left(-3 + \frac{3}{2}\sqrt{\frac{2\pi}{x}}\,e^{x/2}\,\mathrm{erf}\left(\frac{x}{2}\right)\right) + B\frac{e^{x/2}}{\sqrt{x}}$$

EXAMPLE I.2. Given the equation

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + xy = 0$$

The terms for $\widetilde{\rho}_{i,j}$ are $\widetilde{\rho}_{2,0} = 1$, $\widetilde{\rho}_{1,0} = 1$, and $\widetilde{\rho}_{0,1} = 1$, whereas the rest are zero. The indicial roots are $r_a = r_b = 0$. Because the difference is an integer with $m = 0$, we have $\eta = 1$. The only nonzero values of $Q_{n,k}$ are

$$Q_{n,n}(0) = n^2 \qquad \text{and} \qquad Q_{n,n-1}(0) = 1$$

Thus $\widetilde{\phi}_0(0) = 1$ and for $n > 0$,

$$\widetilde{\phi}_n(0) = -\frac{1}{n^2}\widetilde{\phi}_{n-1}(0) = \left(\frac{-1}{n^2}\right)\left(\frac{-1}{(n-1)^2}\right)\cdots(-1) = \frac{(-1)^n}{(n!)^2}$$

which yields the first solution $u(x)$

$$u(x) = \sum_{n=0}^{\infty}\frac{(-1)^n}{(n!)^2}x^n$$

which could also be cast in terms of the hypergeometric function ${}_1F_2$ as

$$u(x) = 1 - x\,({}_1F_2\,[1;2,2;-x])$$

For the second solution, we need $\sigma_n(0)$,

$$\sigma_n(0) = (2n)\,\widetilde{\phi}_n(0) = \frac{(-1)^n 2n}{(n!)^2}$$

Because $m = r_b - r_a = 0$, we set $b_0 = 0$, and the other coefficients are given by

$$\begin{aligned} b_n &= -\frac{Q_{n,n-1}(0)b_{n-1} + \sigma_n(0)}{Q_{n,n}(0)} \\ &= -\frac{1}{n^2}b_{n-1} - 2\frac{(-1)^n}{n(n!)^2} = \frac{(-1)^n}{(n!)^2}b_0 - 2\frac{(-1)^n}{(n!)^2} + \cdots - 2\frac{(-1)^n}{n(n!)^2} \\ &= -2\frac{(-1)^n}{(n!)^2}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right) \end{aligned}$$

Thus the second solution is given by

$$v(x) = u(x)\ln(x) - 2\sum_{n=1}^{\infty}\frac{(-x)^n}{(n!)^2}\left(\sum_{k=1}^{n}\frac{1}{k}\right)$$

and the complete solution is $y = Au(x) + Bv(x)$.

EXAMPLE I.3. Given the equation

$$9x^2\frac{d^2y}{dx^2} + 3x\frac{dy}{dx} + (2x - 8)y = 0$$

The terms for $\widetilde{\rho}_{i,j}$ are $\widetilde{\rho}_{2,0} = 9$, $\widetilde{\rho}_{1,0} = 3$, $\widetilde{\rho}_{0,0} = -8$, and $\widetilde{\rho}_{0,1} = 2$. The indicial roots are $r_a = 4/3$ and $r_b = -2/3$. The difference is an integer; then $m = 2$.

The only nonzero values of $Q_{n,k}$ are

$$Q_{n,n}(r) = n\left(9n + 9(2r-1) + 3\right) \qquad \text{and} \qquad Q_{n,n-1}(r) = 2$$

Thus $\widetilde{\phi}_0(r) = 1$ and

$$\widetilde{\phi}_n(r) = -\frac{2}{n\,(9n + 9(2r-1) + 3)}\widetilde{\phi}_{n-1}(r)$$

Using the larger root, $r_a = 4/3$,

$$\begin{aligned}
\widetilde{\phi}_n\left(\frac{4}{3}\right) &= \frac{-2}{9n(n+2)}\widetilde{\phi}_{n-1}\left(\frac{4}{3}\right) \\
&= \left(\frac{-2}{9n(n+2)}\right)\left(\frac{-2}{9(n-1)(n+1)}\right)\cdots\left(\frac{-2}{9\cdot 1\cdot 3}\right)\widetilde{\phi}_0\left(\frac{4}{3}\right) \\
&= \frac{(-1)^n 2^{n+1}}{9^n (n!)(n+2)!}
\end{aligned}$$

The first solution is then given by

$$u(x) = \sum_{n=0}^{\infty} \frac{(-1)^n 2^{n+1}}{9^n (n!)(n+2)!} x^{n+(4/3)}$$

or in terms of hypergeometric functions,

$$u(x) = x^{4/3}\left(1 - \frac{2x}{27}\left({}_1F_2\left[1; 2, 4; -\frac{2x}{9}\right]\right)\right)$$

Because $m = 2$, we only need $\widetilde{\phi}_1(r_b)$ for the second solution,

$$\widetilde{\phi}_1\left(-\frac{2}{3}\right) = \frac{-2}{9(-1)} = \frac{2}{9}$$

Next, we need $\sigma_n(r_a)$ and η,

$$\begin{aligned}
\sigma_n(r_a) &= [9(2r_a + 2n - 1) + 3]\,\widetilde{\phi}_n\,(r_a) = \frac{(-1)^n 2^{n+1} 18(n+1)}{9^n (n!)(n+2)!} \\
\eta &= -\frac{Q_{m,m-1}\widetilde{\phi}_{m-1}(r_b)}{\sigma_0(r_a)} = -\frac{2}{9^2}
\end{aligned}$$

For the coefficients b_n, we have $b_0 = 1$, $b_1 = 2/9$, $b_2 = 0$ and the rest are found by recurrence, that is,

$$\begin{aligned}
b_n &= -\frac{Q_{n,n-1}(r_b)}{Q_{n,n}(r_b)} b_{n-1} - \eta\frac{\sigma_{n-m}(r_a)}{Q_{n,n}(r_b)} \\
&= \frac{-2}{9n(n-2)} b_{n-1} + \left(\frac{(-1)^n 2^{n+1}}{9^n (n-2)!n!}\right)\left(\frac{(n-1)}{n(n-2)}\right) \\
&= \frac{(-2)^{n-2}2}{9^{n-2} n!(n-2)!} b_2 + \left(\frac{(-1)^n 2^{n+1}}{9^n (n-2)!n!}\right)\left(\frac{2}{(3)(1)} + \cdots + \frac{(n-1)}{(n)(n-2)}\right) \\
&= \left(\frac{(-1)^n 2^{n+1}}{9^n (n-2)!n!}\right)\sum_{k=0}^{n-3} \frac{(n-1-k)}{(n-k)(n-2-k)}, \qquad \text{for } n > 2
\end{aligned}$$

The second solution is then given by

$$\begin{aligned} v(x) &= x^{-2/3} + \frac{2}{9}x^{1/3} - \frac{2}{81}u\ln(x) \\ &\quad + \sum_{n=3}^{\infty} x^{n-(2/3)} \left(\frac{(-1)^n 2^{n+1}}{9^n (n-2)!n!} \right) \left(\sum_{k=0}^{n-3} \frac{(n-1-k)}{(n-k)(n-2-k)} \right) \end{aligned}$$

and the complete solution is $y = Au(x) + Bv(x)$.

I.4 Series Solution of Legendre Equations

I.4.1 Legendre Equations

The **Legendre equation of order** μ is given by the following equation:

$$\left(1 - x^2\right) \frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + \mu\left(\mu + 1\right) y = 0 \tag{I.20}$$

Using the series solution expanded around the ordinary point $x = 0$, we seek a solution of the form

$$y = \sum_{n=0}^{\infty} a_n x^n$$

With $N = 2$, the coefficients $\rho_{i,j}$ are: $\rho_{2,0} = 1$, $\rho_{2,2} = -1$, $\rho_{1,1} = -2$, and $\rho_{0,0} = \mu\left(\mu + 1\right)$. Based on (9.21), the only nonzero values are for $n = k$, that is,

$$\begin{aligned} \phi_{n,n} &= -\frac{\rho_{0,0} + \rho_{1,1}(n) + \rho_{2,2}(n)(n-1)}{\rho_{N,0}(n+1)(n+2)} = -\frac{\mu(\mu+1) - n(n+1)}{(n+1)(n+2)} \\ &= -\frac{(\mu + n + 1)(\mu - n)}{(n+1)(n+2)} \end{aligned}$$

which yields the following recurrence equation:

$$a_{n+2} = -\frac{(\mu + n + 1)(\mu - n)}{(n+1)(n+2)} a_n$$

When separated according to even or odd subscripts, with $n \geq 1$,

$$a_{2n} = \frac{(-1)^n}{(2n)!} \prod_{k=0}^{n-1} \left[\mu - 2(n-k)\right] \left[\mu + 2(n-k) + 1\right] a_0 \tag{I.21}$$

$$a_{2n+1} = \frac{(-1)^n}{(2n+1)!} \prod_{k=0}^{n-1} \left[\mu - 2(n-k) - 1\right] \left[\mu + 2(n-k) + 2\right] a_1 \tag{I.22}$$

where a_0 and a_1 are arbitrary.

Let functions $\Lambda_{2n}(\mu)$ and $\Lambda_{2n+1}(\mu)$ be defined as

$$\Lambda_{2n}(\mu) = \frac{(-1)^n}{(2n)!} \prod_{k=0}^{n-1} \left(\mu - 2(n-k)\right) \left(\mu + 2(n-k) + 1\right) \tag{I.23}$$

$$\Lambda_{2n+1}(\mu) = \frac{(-1)^n}{(2n+1)!} \prod_{k=0}^{n-1} \left(\mu - 2(n-k) - 1\right) \left(\mu + 2(n-k) + 2\right) \tag{I.24}$$

then the solution to the Legendre equation of order μ is

$$y = a_0 \left(1 + \sum_{n=1}^{\infty} \Lambda_{2n}(\mu) x^{2n}\right) + a_1 \left(x + \sum_{n=1}^{\infty} \Lambda_{2n+1}(\mu) x^{2n+1}\right) \tag{I.25}$$

The two infinite series are called the **Legendre functions of the second kind**, namely $L_{\text{even}}(x)$ and $L_{\text{odd}}(x)$, where

$$L_{\text{even}}(x) = 1 + \sum_{n=1}^{\infty} \Lambda_{2n}(\mu) x^{2n} \tag{I.26}$$

$$L_{\text{odd}}(x) = x + \sum_{n=1}^{\infty} \Lambda_{2n+1}(\mu) x^{2n+1} \tag{I.27}$$

For the special case when $\mu = \mu_{\text{even}}$ is an even integer, $\Lambda_{\mu_{\text{even}}+2j}(\mu_{\text{even}}) = 0$, $j = 1, \ldots$, and thus $L_{\text{even}}(x)$ becomes a finite sum. Similarly, when $\mu = \mu_{\text{odd}}$ is an odd integer, $\Lambda_{\mu_{\text{odd}}+2j}(\mu_{\text{odd}}) = 0$, $j = 1, \ldots$, and $L_{\text{odd}}(x)$ becomes a finite sum. In either case, the finite sums will define a set of important polynomials. By carefully choosing the values of a_0 and a_1, either of the finite polynomials can be normalized to be 1 at x=1. If $\mu = \mu_{\text{even}} = 2\ell$, we need

$$a_0 = A(-1)^{\ell} \frac{(2\ell)!}{2^{\ell} (\ell!)^2} \tag{I.28}$$

Conversely, if $\mu = \mu_{\text{odd}} = 2\ell + 1$,

$$a_1 = A(-1)^{\ell} \frac{(2\ell + 2)!}{2^{\ell} \ell! (\ell + 1)!} \tag{I.29}$$

where A is arbitrary. Thus with these choices for a_0 and a_1, we can rewrite (I.25) to be

$$y = A\mathcal{P}_n(x) + B\mathcal{Q}_n(x) \tag{I.30}$$

where n is an integer. $\mathcal{Q}_n$ is the Legendre function that is an infinite series, whereas $\mathcal{P}_n$ is a finite polynomial referred to as **Legendre polynomial of order** n and given by

$$\mathcal{P}_n(x) = \sum_{k=0}^{\text{Int}(n/2)} \frac{(-1)^k [2n - 2k]!}{2^n k! (n - k)! (n - 2k)!} x^{n-2k} \tag{I.31}$$

where

$$\text{Int}(n/2) = \begin{cases} n/2 & \text{if } n \text{ even} \\ (n-1)/2 & \text{if } n \text{ odd} \end{cases} \tag{I.32}$$

The Legendre functions, $\mathcal{Q}_n(x)$, has a closed form that can be obtained more conveniently by using the method of order reduction. Applying d'Alembert's method of order reduction, we can set $\mathcal{Q}_n(x) = q(x)\mathcal{P}_n(x)$, where $q(x)$ is obtained via Lemma I.1 given in Section I.2. Applying this approach to (I.20),

$$0 = \left(1 - x^2\right) \mathcal{P}_n \frac{dz}{dx} + \left(2\left(1 - x^2\right) \frac{d\mathcal{P}_n}{dx} - 2x\mathcal{P}_n\right) z$$

$$\frac{dz}{z} = \frac{2x\, dx}{1 - x^2} - 2\frac{d\mathcal{P}_n}{\mathcal{P}_n}$$

$$\begin{aligned} z &= \frac{1}{(\mathcal{P}_n)^2} \exp \int \left(\frac{2x}{1-x^2} \right) dx \\ &= \frac{-1}{(1-x^2)(\mathcal{P}_n)^2} \end{aligned}$$

Thus with $q = -\int z dz$,

$$\mathcal{Q}_n(x) = \mathcal{P}_n(x) \int \left[\frac{1}{(1-x^2)(\mathcal{P}_n(x))^2} \right] dx \tag{I.33}$$

where we included a factor of (-1) to make it consistent with (I.26) and (I.27).

I.4.2 Associated Legendre Equation

A generalization of the Legendre equation (I.20) is the **associated Legendre equation** given by

$$(1-x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + \left(n(n+1) - \frac{m^2}{1-x^2} \right) y = 0 \tag{I.34}$$

Note that if $m = 0$, we get back the Legendre equation.

We now consider the situation in which n and m are nonnegative integers. Instead of solving (I.34) by series solution, we approach the solution by using a change of variable, namely let

$$w = (1-x^2)^{-m/2} y \tag{I.35}$$

With $y = qw$, where $q = (1-x^2)^{m/2}$, the terms on the right-hand side of (I.34) can each be divided by q and then evaluated to be

$$\begin{aligned} \frac{1}{q}\left(n(n+1) - \frac{m^2}{1-x^2} \right) y &= \left(n(n+1) - \frac{m^2}{1-x^2} \right) w \\ -\frac{2x}{q}\frac{dy}{dx} &= \frac{2mx^2}{1-x^2} w - 2x\frac{dw}{dx} \\ \frac{1-x^2}{q}\frac{d^2y}{dx^2} &= \frac{m\left[(m-1)x^2 - 1\right]}{1-x^2} w - 2mx\frac{dw}{dx} + (1-x^2)\frac{d^2w}{dx^2} \end{aligned}$$

Doing so reduces (I.34) to

$$(1-x^2)\frac{d^2w}{dx^2} - 2(m+1)x\frac{dw}{dx} + (n-m)(n+m+1)w = 0 \tag{I.36}$$

Now let S be defined by

$$S(x) = A\mathcal{P}_n(x) + B\mathcal{Q}_n(x) \tag{I.37}$$

Then S satisfies the Legendre equation given by (I.20). With $f(x) = 1 - x^2$, $df/dx = -2x$ and $a = n(n+1)$, (I.20) can be rewritten with S replacing y, as

$$f\frac{d^2S}{dx^2} + \frac{df}{dx}\frac{dS}{dx} + aS = 0 \tag{I.38}$$

Furthermore, with $d^2f/dx^2 = -2$ and $d^kf/dx^k = 0$ for $k > 2$, the mth derivative of each term in (I.38) is, using the Leibnitz rule (9.6),

$$\begin{aligned}
\frac{d^m}{dx^m}\left(f\frac{d^2S}{dx^2}\right) &= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{d^kf}{dx^k}\right)\left(\frac{d^{(2+m-k)}S}{dx^{(2+m-k)}}\right) \\
&= f\left(\frac{d^{(2+m)}S}{dx^{(2+m)}}\right) + m\left(\frac{df}{dx}\right)\left(\frac{d^{(1+m)}S}{dx^{(1+m)}}\right) \\
&\quad + \frac{m(m-1)}{2}\left(\frac{d^2f}{dx^2}\right)\left(\frac{d^mS}{dx^m}\right) \\
\frac{d^m}{dx^m}\left(\frac{df}{dx}\frac{dS}{dx}\right) &= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{d^{k+1}f}{dx^{k+1}}\right)\left(\frac{d^{(1+m-k)}S}{dx^{(1+m-k)}}\right) \\
&= \frac{df}{dx}\left(\frac{d^{(1+m)}S}{dx^{(1+m)}}\right) + m\left(\frac{d^2f}{dx^2}\right)\left(\frac{d^mS}{dx^m}\right) \\
\frac{d^m}{dx^m}(aS) &= a\frac{d^mS}{dx^m}
\end{aligned}$$

and adding all the terms together, we obtain

$$\left(1-x^2\right)\frac{d^2}{dx^2}\left(\frac{d^mS}{dx^m}\right) - 2(m+1)x\frac{d}{dx}\left(\frac{d^mS}{dx^m}\right) + (n-m)(n+m+1)\left(\frac{d^mS}{dx^m}\right) = 0 \tag{I.39}$$

Comparing (I.39) with (I.36),

$$\begin{aligned}
w &= \frac{d^mS}{dx^m} \\
\left(1-x^2\right)^{-m/2}y &= A\frac{d^m\mathcal{P}_n}{dx^m} + B\frac{d^m\mathcal{Q}_n}{dx^m}
\end{aligned}$$

Thus the solution to the associated Legendre equation (I.34) is

$$y = \widehat{A}\mathcal{P}_{n,m}(x) + \widehat{B}\mathcal{Q}_{n,m}(x) \tag{I.40}$$

where $\mathcal{P}_{n,m}$ and $\mathcal{Q}_{n,m}$ are the **associated Legendre polynomials** and **associated Legendre functions**, respectively, of order n and degree m defined by[2]

$$\begin{aligned}
\mathcal{P}_{n,m} &= (-1)^m\left(1-x^2\right)^{m/2}\frac{d^m}{dx^m}\mathcal{P}_n(x) \\
\mathcal{Q}_{n,m} &= (-1)^m\left(1-x^2\right)^{m/2}\frac{d^m}{dx^m}\mathcal{Q}_n(x)
\end{aligned} \tag{I.41}$$

[2] In some references, the factor $(-1)^m$ is neglected, but we chose to include it here because MATLAB happens to use the definition given in (I.41).

I.5 Series Solution of Bessel Equations

I.5.1

The **Bessel equation of order** ν is given by the following differential equation:

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + \left(x^2 - \nu^2\right)y = 0 \tag{I.42}$$

Using a series expansion around the regular singular point $x = 0$, we can identify the following coefficients: $\widetilde{\rho}_{2,0} = 1$, $\widetilde{\rho}_{1,0} = 1$, $\widetilde{\rho}_{0,0} = -\nu^2$, and $\widetilde{\rho}_{0,2} = 1$. The indicial roots using (9.2) are $r_a = \nu$ and $r_b = -\nu$. Applying the Frobenius method summarized in Theorem 9.2, the only nonzero values of $Q_{n,k}$ are

$$Q_{n,n-2}(r) = 1 \qquad \text{and} \qquad Q_{n,n}(r) = n(n+2r)$$

thus $\widetilde{\phi}_0(r) = 1, \widetilde{\phi}_1(r) = 0, \widetilde{\phi}_n(r) = -\widetilde{\phi}_{n-2}(r)/[n(n+2r)]$, for $n > 1$, and $\sigma_n(r) = (2r + 2n)\widetilde{\phi}_n(r)$. Furthermore, because $\widetilde{\phi}_1(r) = 0$, functions corresponding to odd subscripts will be zero, that is,

$$\widetilde{\phi}_{2n+1}(r) = 0 \qquad \text{for } n = 0, 1, \ldots$$

For those with even subscripts,

$$\begin{aligned}\widetilde{\phi}_{2n}(r) &= \frac{-1}{4n(n+r)}\widetilde{\phi}_{2n-2} = \left(\frac{-1}{4n(n+r)}\right)\cdots\left(\frac{-1}{4(1)(1+r)}\right)\widetilde{\phi}_0 \\ &= \frac{(-1)^n}{4^n n!\prod_{k=0}^{n-1}(n+r-k)}\end{aligned}$$

Depending on the value of the order ν, we have the various cases to consider:

- **Case 1: 2ν is not an integer.** We have $a_{2k+1} = b_{2k+1} = 0, k = 0, 1, \ldots$, and for $n = 1, 2, \ldots$

$$a_{2n} = \frac{(-1)^n}{4^n n!\prod_{k=0}^{n-1}(n+\nu-k)} \qquad \text{and} \qquad b_{2n} = \frac{(-1)^n}{4^n n!\prod_{k=0}^{n-1}(n-\nu-k)}$$

The two independent solutions are then given by

$$u(x) = \sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+\nu}}{4^n n!\prod_{k=0}^{n-1}(n+\nu-k)} \quad \text{and} \quad v(x) = \sum_{n=0}^{\infty}\frac{(-1)^n x^{2n-\nu}}{4^n n!\prod_{k=0}^{n-1}(n-\nu-k)}$$

These results can further be put in terms of Gamma functions (cf. (9.9)), and after extracting constants out of the summations, we obtain

$$u(x) = 2^{\nu}\Gamma(\nu+1)J_{\nu}(x) \qquad \text{and} \qquad v(x) = 2^{-\nu}\Gamma(-\nu+1)J_{-\nu}(x)$$

where $J_{\nu}(x)$ is known as the **Bessel function of the first kind** defined by

$$J_{\nu}(x) = \sum_{n=0}^{\infty}\left(\frac{x}{2}\right)^{2n+\nu}\frac{(-1)^n}{n!\Gamma(n+\nu+1)} \tag{I.43}$$

where the order ν in the definition (I.43) may or may not be an integer. Thus in terms of Bessel functions, the complete solution, for ν not an integer, is given by

$$y = AJ_{\nu}(x) + BJ_{-\nu}(x) \tag{I.44}$$

- **Case 2:** 2ν **is an odd integer.** Let $m = r_a - r_b = 2\nu$ be an odd integer ℓ. Because, $\widetilde{\phi}_k = 0$ when k is odd, the value η in (9.43) will be zero. This means that $b_{2n} = \widetilde{\phi}_{2n}(-\nu)$, and we end up with the same result as that of case 1, that is,

$$y = AJ_{\nu}(x) + BJ_{-\nu}(x) \tag{I.45}$$

- **Case 3:** $2\nu \neq 0$ **is an even integer.** Let $\nu = \ell$ with ℓ an integer. For the first root $r_a = \ell$, we have $a_{2n} = \widetilde{\phi}_{2n}(\ell)$ and the first solution becomes

$$u(x) = 2^{\ell}\ell! J_{\ell}(x) \tag{I.46}$$

For the second solution, we will separate $v(x)$ into three parts: v_1 , v_2, and v_3, where v_1 contains the terms with $b_{2n}(x)$, $n < \ell$, v_2 is the term with $\ln(x)$ and v_3 contains the rest of the terms.

For v_1, we take $n < \ell$, for which $b_{2n} = \widetilde{\phi}_{2n}(-\ell) = \dfrac{(\ell - n - 1)!}{4^n n!(\ell - 1)!}$ and obtain

$$v_1(x) = \sum_{n=0}^{\ell-1} \frac{x^{2n-\ell}(\ell - n - 1)!}{4^n n!(\ell - 1)!} = \left[\frac{1}{2^{\ell}(\ell - 1)!}\right] \sum_{n=0}^{\ell-1} \left(\frac{x}{2}\right)^{2n-\ell} \frac{(\ell - n - 1)!}{n!} \tag{I.47}$$

For v_2, with $m = 2\ell$, we find that

$$\eta = -\frac{Q_{2\ell,2\ell-2}(-\ell)\widetilde{\phi}_{2\ell-2}(-\ell)}{\sigma_0(\ell)} = -\frac{2}{4^{\ell}\ell!(\ell - 1)!}$$

and together with $u(x)$ in (I.46), we obtain

$$v_2(x) = \eta u(x)\ln(x) = -2\left[\frac{1}{2^{\ell}(\ell - 1)!}\right] J_{\ell}(x)\ln(x) \tag{I.48}$$

For v_3, one can first show that

$$\begin{aligned} b_{2(n+\ell)} &= -\frac{Q_{2(n+\ell),2(n-1+\ell)}(-\ell)b_{2(n-1+\ell)} + \eta\sigma_{2n}(\ell)}{Q_{2(n+\ell),2(n+\ell)}(-\ell)} \\ &= \frac{(-1)^n}{4^n n!(n+\ell)!} b_{2\ell} \\ &\quad + \left[\frac{1}{4^{\ell}(\ell - 1)!}\right]\left[\frac{(-1)^n}{4^n n!(n+\ell)!}\right]\left[\sum_{k=0}^{n-1}\left(\frac{1}{n-k} + \frac{1}{n-k+\ell}\right)\right] \end{aligned}$$

Because $b_m = b_{2\ell} = 0$, we obtain $v_3(x)$ to be

$$v_3(x) = \left[\frac{1}{2^{\ell}(\ell - 1)!}\right] \sum_{n=1}^{\infty}\left[\left(\frac{x}{2}\right)^{2n+\ell} \frac{(-1)^n}{n!(n+\ell)!} \sum_{k=1}^{n}\left(\frac{1}{k} + \frac{1}{k+\ell}\right)\right] \tag{I.49}$$

Adding up (I.47), (I.48) and (I.49), we have the second solution $v(x)$ as

$$\begin{aligned} v(x) &= v_1(x) + v_2(x) + v_3(x) \\ &= -\left[\frac{1}{2^\ell(\ell-1)!}\right]\left\{2J_\ell(x)\ln(x) - \sum_{n=0}^{\ell-1}\left(\frac{x}{2}\right)^{2n-\ell}\frac{(\ell-n-1)!}{n!}\right. \\ &\quad \left. -\sum_{n=1}^{\infty}\left(\frac{x}{2}\right)^{2n+\ell}\frac{(-1)^n}{n!(n+\ell)!}\left(\sum_{k=1}^{n}\left(\frac{1}{k}+\frac{1}{k+\ell}\right)\right)\right\} \end{aligned} \tag{I.50}$$

A more standard solution formulation known as the **Weber form** is given by

$$y = AJ_\ell(x) + BY_\ell(x) \tag{I.51}$$

where the function $Y_\ell(x)$ is known as **Bessel function of the second kind** (also known as the **Neumann function**), defined as

$$\begin{aligned} Y_\ell(x) &= \frac{2}{\pi}J_\ell(x)\left[\ln\left(\frac{x}{2}\right)+\gamma\right] - \frac{1}{\pi}\sum_{n=0}^{\ell-1}\left(\frac{x}{2}\right)^{2n-\ell}\frac{(\ell-n-1)!}{n!} \\ &\quad -\frac{1}{\pi}\sum_{n=0}^{\infty}\left(\frac{x}{2}\right)^{2n+\ell}\frac{(-1)^n}{n!(n+\ell)!}\left[\sum_{k=1}^{n+\ell}\frac{1}{k}\right] \end{aligned} \tag{I.52}$$

where γ is known as **Euler's constant**, defined by

$$\gamma = \lim_{n\to\infty}\left[\left(1+\frac{1}{2}+\cdots+\frac{1}{n}\right)-\ln(n)\right] = 0.572215664\ldots \tag{I.53}$$

- **Case 4:** $\nu = 0$**.** With $\eta = 1$, a similar procedure as in Case 3 above will lead to a solution of the same Weber form,

$$y = AJ_0(x) + BY_0(x) \tag{I.54}$$

where

$$Y_0(x) = \frac{2}{\pi}J_0(x)\left(\ln\left(\frac{x}{2}\right)+\gamma\right) - \frac{2}{\pi}\sum_{n=1}^{\infty}\left(\frac{x}{2}\right)^{2n}\frac{(-1)^n}{(n!)^2}\left(\sum_{k=1}^{n}\frac{1}{k}\right) \tag{I.55}$$

An alternative method for computing the Bessel functions is to define the Bessel function of the second kind as

$$Y_\nu(x) = \frac{J_\nu(x)\cos(\nu\pi) - J_{-\nu}(x)}{\sin(\nu\pi)} \tag{I.56}$$

Then for $\nu = n$, an integer, we simply take the limit, that is,

$$Y_n(x) = \lim_{\nu\to n} Y_\nu(x) \tag{I.57}$$

This means we can unify the solutions to both cases of ν being an integer or not, as

$$y(x) = AJ_\nu(x) + BY_\nu(x) \tag{I.58}$$

I.5.2 Bessel Equations of Parameter λ

A simple extension to the Bessel equation is to introduce a parameter λ in the Bessel equation as follows

$$x^2\frac{d^2y}{dx^2}+x\frac{dy}{dx}+\left(\lambda^2x^2-\nu^2\right)y=0 \tag{I.59}$$

Instead of approaching the equation directly with a series solution, we could simply use a change of variable, namely $w=\lambda x$. Then

$$dx=\frac{1}{\lambda}dw \qquad ; \qquad \frac{dy}{dx}=\lambda\frac{dy}{dw} \qquad ; \qquad \frac{d^2y}{dx^2}=\lambda^2\frac{d^2y}{dw^2}$$

Substituting these into (I.59), we get

$$w^2\frac{d^2y}{dw^2}+w\frac{dy}{dw}+\left(w^2-\nu^2\right)y=0$$

whose solution is given by

$$y=AJ_\nu(w)+BY_\nu(w)$$

or

$$y=AJ_\nu(\lambda x)+BY_\nu(\lambda x) \tag{I.60}$$

I.5.3 Modified Bessel Equations and Functions

The **modified Bessel equations of order** ν is given by

$$x^2\frac{d^2y}{dx^2}+x\frac{dy}{dx}-\left(x^2+\nu^2\right)y=0 \tag{I.61}$$

which is just the Bessel equation with parameter $i=\sqrt{-1}$, that is,

$$x^2\frac{d^2y}{dx^2}+x\frac{dy}{dx}+\left((i)^2x^2-\nu^2\right)y=0$$

Then the solution is given by

$$y=AJ_\nu(ix)+BY_\nu(ix)$$

Another form of the solution is given by

$$y=AI_\nu(ix)+BK_\nu(ix) \tag{I.62}$$

where $I_\nu(x)$ is the **modified Bessel equation of the first kind of order** ν defined by

$$I_\nu(x)=\exp\left(-\frac{\nu\pi i}{2}\right)J_\nu(ix) \tag{I.63}$$

and $K_\nu(x)$ is the **modified Bessel equation of the second kind of order** ν defined by

$$K_\nu(x)=\exp\left(\frac{(\nu+1)\pi i}{2}\right)\left[J_\nu(ix)+iY_\nu(ix)\right] \tag{I.64}$$

I.6 Proofs for Lemmas and Theorems in Chapter 9

I.6.1 Proof of Series Expansion Formula, Theorem 9.1

Assuming a series solution of the form

$$y = \sum_{n=0}^{\infty} a_n x^n \tag{I.65}$$

the derivatives are given by

$$\begin{aligned}
\frac{dy}{dx} &= \sum_{n=1}^{\infty} n a_n x^{n-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} x^n \\
\frac{d^2y}{dx^2} &= \sum_{n=1}^{\infty} (n+1)(n) a_{n+1} x^{n-1} = \sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n \\
&\vdots \\
\frac{d^N y}{dx^N} &= \sum_{n=0}^{\infty} \frac{(n+N)!}{n!} a_{n+N} x^n
\end{aligned} \tag{I.66}$$

After substitution of (9.18) and (I.66) into (9.17), while using (9.5),

$$\sum_{n=0}^{\infty} x^n \sum_{k=0}^{n} \sum_{j=0}^{N} \left(a_{k+j} \rho_{j,n-k} \frac{(k+j)!}{k!} \right) = 0$$

Because x is not identically zero, we have

$$\sum_{k=0}^{n} \sum_{j=0}^{N} \left(a_{k+j} \rho_{j,n-k} \frac{(k+j)!}{k!} \right) = 0 \qquad \text{for } n = 0, 1, \ldots, \infty \tag{I.67}$$

For a fixed n, let

$$\mu_{j,k} = \rho_{j,n-k} \frac{(k+j)!}{k!} \qquad \rightarrow \qquad \mu_{j,m-j} = \begin{cases} \rho_{0,n-m} & \text{if } j = 0 \\ \rho_{j,n-m+j} \prod_{i=0}^{j-1} (m-i) & \text{if } j > 0 \end{cases}$$

We can rearrange the summation in (I.67) to have the following structure:

$j=0$	$j=1$	$\cdots$	$j=N$	
$\mu_{0,0}$				a_0
$\mu_{0,1}$	$\mu_{1,0}$			a_1
$\mu_{0,2}$	$\mu_{1,1}$	$\ddots$		
$\vdots$	$\mu_{1,2}$	$\ddots$	$\mu_{N,0}$	
$\mu_{0,n}$	$\vdots$	$\ddots$	$\mu_{N,1}$	$\vdots$
	$\mu_{1,n}$		$\mu_{N,2}$	
		$\ddots$	$\vdots$	
			$\mu_{N,n}$	a_{n+N}

where the group of terms to the left of a_m are summed up as the coefficient of a_m. Note that $\mu_{j,\ell} = 0$ if $\ell < 0$. In addition, we can define $\rho_{j,\ell} = 0$ for $\ell < 0$, and

obtain $\mu_{j,m-j} = 0$ for $m - j > n$. Thus the coefficients of a_m for $m \le (n+N)$ can be formulated as

$$\text{coef}\,(a_m) = \begin{cases} \mu_{0,m} + \sum_{j=1}^{N} \mu_{j,m-j} & \text{if } m < n+N \\ \mu_{N,n} & \text{if } m = n+N \end{cases}$$

Letting $a_0, a_1, \ldots, a_{N-1}$ be arbitrary, we have for $n = 0, 1, \ldots,$

$$\mu_{N,n} a_{n+N} + \sum_{m=0}^{n+N-1} a_m \left(\mu_{0,m} + \sum_{j=1}^{N} \mu_{j,m-j} \right) = 0$$

$$\begin{aligned} a_{n+N} &= -\frac{\sum_{m=0}^{n+N-1} a_m \left(\mu_{0,m} + \sum_{j=1}^{N} \mu_{j,m-j} \right)}{\mu_{N,n}} \\ &= \sum_{m=0}^{n+N-1} \phi_{n,m} a_m \end{aligned}$$

where

$$\phi_{n,m} = (-1) \frac{\rho_{0,n-m} + \sum_{j=1}^{N} \rho_{j,n-m+j} \prod_{i=0}^{j-1} (m-i)}{\rho_{N,0} \prod_{i=1}^{N} (n+i)}$$

and

$$\rho_{j,\ell} = 0 \qquad \ell < 0$$

I.6.2 Proof of Frobenius Series Method, Theorem 9.2

The formula of a_n has already been discussed (cf. (I.7)). The same is true for when $(r_b - r_a)$ is not an integer, where we simply set $\eta = 0$ and $b_n = \widetilde{\phi}_n(r_b)$ (cf. (I.8)). Thus the remaining case to be proved is when $r_a - r_b = m$ is a positive integer.

Based on the forms given in (I.14), consider the case where $m > 0$. Then v, $x(dv/dx)$ and $x^2(d^2v/dx^2)$ becomes

$$\begin{aligned} v &= \eta u \ln(x) + \sum_{n=0}^{\infty} b_n x^{n+r_b} \\ x\frac{dv}{dx} &= \eta \left[u + x \ln(x) \frac{du}{dx} \right] + \sum_{n=0}^{\infty} b_n (n+r_b) x^{n+r_b} \\ x^2 \frac{d^2 v}{dx^2} &= \eta \left[-u + 2x\frac{du}{dx} + x^2 \ln(x) \frac{d^2 u}{dx^2} \right] + \sum_{n=0}^{\infty} b_n (n+r_b)(n+r_b-1) x^{n+r_b} \end{aligned}$$

Substituting into

$$x^2\widetilde{P}_2(x)\frac{d^2v}{dx^2} + x\widetilde{P}_1(x)\frac{dv}{dx} + \widetilde{P}_0(x)v = 0$$

we have

$$\begin{aligned} 0 \quad = \quad & \eta\ln(x)\left(x^2\widetilde{P}_2(x)\frac{d^2u}{dx^2} + x\widetilde{P}_1(x)\frac{du}{dx} + \widetilde{P}_0(x)u\right) \\ & + \eta\left[\widetilde{P}_2(x)\left(-u + 2x\frac{du}{dx}\right) + \widetilde{P}_1(x)u\right] \\ & + \widetilde{P}_2(x)\sum_{n=0}^{\infty} b_n(n+r_b)(n+r_b-1)x^{n+r_b} \\ & + \widetilde{P}_1(x)\sum_{n=0}^{\infty} b_n(n+r_b)x^{n+r_b} \\ & + \widetilde{P}_0(x)\sum_{n=0}^{\infty} b_n x^{n+r_b} \end{aligned}$$

Because u is a solution to the differential equation, the group of terms multiplying $\ln(x)$ is equal to zero. After substitution of $\widetilde{P}_i(x) = \sum_{n=0}^{\infty}\widetilde{\rho}_{i,n}x^n$ and $u(x) = \sum_{n=0}^{\infty}\widetilde{\phi}_n(r_a)x^{n+r_a}$, the equation above becomes

$$\begin{aligned} \sum_{n=0}^{\infty} x^{n+r_a}\eta\sum_{k=0}^{n}\widetilde{\phi}_k(r_a)\left(\widetilde{\rho}_{1,n-k} + (2r_a+2k-1)\,\widetilde{\rho}_{2,n-k}\right) & \\ + \sum_{n=0}^{\infty} x^{n+r_b}\sum_{k=0}^{n} b_k Q_{n,k}(r_b) \quad = \quad & 0 \end{aligned}$$

With $r_a = r_b + m$, the first summation can be reindexed, that is,

$$\begin{aligned} \sum_{n=m}^{\infty} x^{n+r_b}\eta\sum_{k=0}^{n-m}\widetilde{\phi}_k(r_a)\left(\widetilde{\rho}_{1,n-m-k} + (2r_a+2k-1)\,\widetilde{\rho}_{2,n-m-k}\right) & \\ + \sum_{n=0}^{\infty} x^{n+r_b}\sum_{k=0}^{n} b_k Q_{n,k}(r_b) \quad = \quad & 0 \end{aligned}$$

Using the definition of $\sigma_n(r)$ given in (9.40), we arrive at the working equation,

$$\begin{aligned} & \left(\sum_{n=0}^{m-1} x^{n+r_b}\sum_{k=0}^{n} b_k Q_{n,k}(r_b)\right) \\ & + x^{r_b+m}\left(\eta\sigma_0(r_a) + b_m Q_{m,m}(r_b) + \sum_{k=0}^{m-1} b_k Q_{m,k}(r_b)\right) \\ & + \left(\sum_{n=m+1}^{\infty} x^{n+r_b}\left[\eta\sigma_{n-m}(r_a) + \sum_{k=0}^{n} b_k Q_{n,k}(r_b)\right]\right) \quad = \quad 0 \end{aligned}$$

Thus for $n < m$, the formula for b_n becomes those for $\widetilde{\phi}_n(r_b)$. For $n = m$, note that $Q_{m,m}(r_b) = 0$, and we have b_m arbitrary, which we can set to zero. Doing so and making the coefficient of x^{m+r_b} be equal to zero,

$$\eta = -\frac{\sum_{k=0}^{m-1} b_k Q_{m,k}(r_b)}{\sigma_0(r_a)}$$

For $n > m > 0$, each coefficient of x^{n+r_b} can be set to zero, which yields the recurrence formula for b_n,

$$b_n = -\frac{\eta\sigma_{n-m}(r_a) + \sum_{k=0}^{n-1} Q_{n,k}(r_b) b_k}{Q_{n,n}(r_b)}$$

Finally, if $m = 0$, a similar derivation can be followed, except that we can set $\eta = 1$ as discussed before. The working equation is now given by

$$\begin{aligned} & x^{r_b}\left(\sigma_0(r_a) + b_m Q_{m,m}(r_b)\right) \\ + & \left(\sum_{n=1}^{\infty} x^{n+r_b}\left[\sigma_{n-m}(r_a) + \sum_{k=0}^{n} b_k Q_{n,k}(r_b)\right]\right) \quad = \quad 0 \end{aligned}$$

Note that for this case, $r_a = r_b = (1 - \widetilde{\rho}_{1,0}/\widetilde{\rho}_{2,0})/2$, which means $\sigma_0 = 0$. With $Q_{0,0}(r_b) = 0$, b_0 can be arbitrary and thus can be set to be zero. The remaining coefficients then become

$$b_n = -\frac{\sigma_n(r_a) + \sum_{k=0}^{n-1} Q_{n,k}(r_b) b_k}{Q_{n,n}(r_b)}$$

I.6.3 Proof of Bessel Function Identities

1. **Derivatives of $J_\nu(x)$.** Recall the definition of $J_\nu(x)$,

$$J_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\nu+1)}\left(\frac{x}{2}\right)^{2m+\nu}$$

To show (9.63), multiply $J_\nu(x)$ by x^ν and then take the derivative with respect to x,

$$\begin{aligned} \frac{d}{dx}\left(x^\nu J_\nu(x)\right) &= \frac{d}{dx}\left[\sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+2\nu}}{m!\Gamma(m+\nu+1)2^{2m+\nu}}\right] \\ &= \sum_{m=0}^{\infty} \frac{(-1)^m (2m+2\nu) x^{2m+2\nu-1}}{m!\Gamma(m+\nu+1)2^{2m+\nu}} \\ &= x^\nu \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\nu)}\left(\frac{x}{2}\right)^{2m+\nu-1} \\ &= x^\nu J_{\nu-1}(x) \end{aligned}$$

To show (9.64), multiply $J_\nu(x)$ by $x^{-\nu}$ and then take the derivative with respect to x,

$$\begin{aligned}
\frac{d}{dx}\left(x^{-\nu}J_\nu(x)\right) &= \frac{d}{dx}\left[\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m}}{m!\Gamma(m+\nu+1)2^{2m+\nu}}\right] \\
&= \sum_{m=1}^{\infty}\frac{(-1)^m(2m)x^{2m-1}}{m!\Gamma(m+\nu+1)2^{2m+\nu}} \\
&= \sum_{m=1}^{\infty}\frac{(-1)^m x^{2m-1}}{(m-1)!\Gamma(m+\nu+1)2^{2m+\nu-1}} \\
&= -x^{-\nu}\sum_{m=0}^{\infty}\frac{(-1)^m}{m!\Gamma(m+\nu+2)}\left(\frac{x}{2}\right)^{2m+\nu+1} \\
&= -x^{-\nu}J_{\nu+1}(x)
\end{aligned}$$

To show (9.65), expand the derivative operation on $x^\nu J_\nu(x)$

$$\frac{d}{dx}\left(x^\nu J_\nu(x)\right) = \nu x^{\nu-1}J_\nu(x) + x^\nu\frac{d}{dx}J_\nu(x)$$

and equate with (9.63) to obtain

$$\begin{aligned}
\nu x^{\nu-1}J_\nu(x) + x^\nu\frac{d}{dx}J_\nu(x) &= x^\nu J_{\nu-1}(x) \\
\frac{d}{dx}J_\nu(x) &= J_{\nu-1}(x) - \frac{\nu}{x}J_\nu(x)
\end{aligned}$$

To show (9.66), expand the derivative operation on $x^{-\nu}J_\nu(x)$

$$\frac{d}{dx}\left(x^{-\nu}J_\nu(x)\right) = -\nu x^{-\nu-1}J_\nu(x) + x^{-\nu}\frac{d}{dx}J_\nu(x)$$

and equate with (9.64) to obtain

$$\begin{aligned}
-\nu x^{-\nu-1}J_\nu(x) + x^{-\nu}\frac{d}{dx}J_\nu(x) &= -x^{-\nu}J_{\nu+1}(x) \\
\frac{d}{dx}J_\nu(x) &= -J_{\nu+1}(x) + \frac{\nu}{x}J_\nu(x)
\end{aligned}$$

2. **Derivatives of $Y_\nu(x)$.** Recall the definition of $Y_\nu(x)$,

$$\begin{aligned}
Y_\nu(x) &= \frac{2}{\pi}\left[\ln\left(\frac{x}{2}\right)+\gamma\right]J_\nu(x) - \frac{1}{\pi}\sum_{m=0}^{\nu-1}\frac{(\nu-m-1)!}{m!}\left(\frac{x}{2}\right)^{2m-\nu} \\
&\quad - \frac{1}{\pi}\sum_{m=1}^{\infty}\frac{(-1)^m}{m!(m+\nu)!}\left(\frac{x}{2}\right)^{2m+\nu}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad - \frac{1}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m}{m!(m+\nu)!}\left(\frac{x}{2}\right)^{2m+\nu}\left[\sum_{k=1}^{m+\nu}\frac{1}{k}\right]
\end{aligned}$$

To show (9.67), multiply $Y_\nu(x)$ by x^ν and then take the derivative with respect to x, while incorporating (9.63),

$$\begin{aligned}
\frac{d}{dx}\left(x^\nu Y_\nu(x)\right) &= \frac{2}{\pi}\frac{d}{dx}\left(\left[\ln\left(\frac{x}{2}\right)+\gamma\right]x^\nu J_\nu(x)\right) \\
&\quad -\frac{1}{\pi}\frac{d}{dx}\sum_{m=0}^{\nu-1}\frac{(\nu-m-1)!x^{2m}}{m!2^{2m-\nu}} \\
&\quad -\frac{1}{\pi}\frac{d}{dx}\sum_{m=1}^{\infty}\frac{(-1)^m x^{2m+2\nu}}{m!(m+\nu)!2^{2m+\nu}}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad -\frac{1}{\pi}\frac{d}{dx}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+2\nu}}{m!(m+\nu)!2^{2m+\nu}}\left[\sum_{k=1}^{m+\nu}\frac{1}{k}\right] \\
&= \frac{2}{\pi}x^{\nu-1}J_\nu(x)+\frac{2}{\pi}\left[\ln\left(\frac{x}{2}\right)+\gamma\right]x^\nu J_{\nu-1}(x) \\
&\quad -\frac{1}{\pi}\sum_{m=1}^{\nu-1}\frac{(\nu-m-1)!x^{2m-1}}{(m-1)!2^{2m-\nu-1}} \\
&\quad -\frac{1}{\pi}\sum_{m=1}^{\infty}\frac{(-1)^m x^{2m+2\nu-1}}{m!(m+\nu-1)!2^{2m+\nu-1}}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad -\frac{1}{\pi}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+2\nu-1}}{m!(m+\nu-1)!2^{2m+\nu-1}}\left[\sum_{k=1}^{m+\nu-1}\frac{1}{k}\right] \\
&\quad -\frac{2}{\pi}x^{\nu-1}\sum_{m=0}^{\infty}\frac{(-1)^m}{m!(m+\nu)!}\left(\frac{x}{2}\right)^{2m+\nu} \\
&= \frac{2}{\pi}\left[\ln\left(\frac{x}{2}\right)+\gamma\right]x^\nu J_{\nu-1}(x) \\
&\quad -\frac{1}{\pi}x^\nu\sum_{m=0}^{\nu}\frac{(\nu-m)!}{(m)!}\left(\frac{x}{2}\right)^{2m-\nu+1} \\
&\quad -\frac{1}{\pi}x^\nu\sum_{m=1}^{\infty}\frac{(-1)^m}{m!(m+\nu-1)!}\left(\frac{x}{2}\right)^{2m+\nu-1}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad -\frac{1}{\pi}x^\nu\sum_{m=0}^{\infty}\frac{(-1)^m}{m!(m+\nu-1)!}\left(\frac{x}{2}\right)^{2m+\nu-1}\left[\sum_{k=1}^{m+\nu-1}\frac{1}{k}\right] \\
&= x^\nu Y_{\nu-1}(x)
\end{aligned}$$

To show (9.68), multiply $Y_\nu(x)$ by $x^{-\nu}$ and then take the derivative with respect to x, while incorporating (9.64),

$$\begin{aligned}
\frac{d}{dx}\left(x^{-\nu}Y_\nu(x)\right) &= \frac{2}{\pi}\frac{d}{dx}\left(\left[\ln\left(\frac{x}{2}\right)+\gamma\right]x^{-\nu}J_\nu(x)\right) \\
&\quad -\frac{1}{\pi}\frac{d}{dx}\sum_{m=0}^{\nu-1}\frac{(\nu-m-1)!x^{2m-2\nu}}{m!2^{2m-\nu}} \\
&\quad -\frac{1}{\pi}\frac{d}{dx}\sum_{m=1}^{\infty}\frac{(-1)^m x^{2m}}{m!(m+\nu)!2^{2m+\nu}}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad -\frac{1}{\pi}\frac{d}{dx}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m}}{m!(m+\nu)!2^{2m+\nu}}\left[\sum_{k=1}^{m+\nu}\frac{1}{k}\right] \\
&= \frac{2}{\pi}x^{-\nu-1}J_\nu(x)-\frac{2}{\pi}\left[\ln\left(\frac{x}{2}\right)+\gamma\right]x^{-\nu}J_{\nu+1}(x) \\
&\quad +\frac{1}{\pi}\sum_{m=0}^{\nu-1}\frac{(\nu-m)!x^{2m-2\nu-1}}{m!2^{2m-\nu-1}} \\
&\quad -\frac{1}{\pi}\sum_{m=1}^{\infty}\frac{(-1)^m x^{2m-1}}{(m-1)!(m+\nu)!2^{2m+\nu-1}}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad -\frac{1}{\pi}\sum_{m=1}^{\infty}\frac{(-1)^m x^{2m-1}}{(m-1)!(m+\nu)!2^{2m+\nu-1}}\left[\sum_{k=1}^{m+\nu}\frac{1}{k}\right] \\
&= -\frac{2}{\pi}\left[\ln\left(\frac{x}{2}\right)+\gamma\right]x^{-\nu}J_{\nu+1}(x) \\
&\quad +\frac{1}{\pi}x^{-\nu}\sum_{m=0}^{\nu}\frac{(\nu-m)!}{(m)!}\left(\frac{x}{2}\right)^{2m-\nu-1} \\
&\quad +\frac{1}{\pi}x^{-\nu}\sum_{m=1}^{\infty}\frac{(-1)^m}{m!(m+\nu+1)!}\left(\frac{x}{2}\right)^{2m+\nu+1}\left[\sum_{k=1}^{m}\frac{1}{k}\right] \\
&\quad +\frac{1}{\pi}x^{-\nu}\sum_{m=0}^{\infty}\frac{(-1)^m}{m!(m+\nu+1)!}\left(\frac{x}{2}\right)^{2m+\nu+1}\left[\sum_{k=1}^{m+\nu+1}\frac{1}{k}\right] \\
&= -x^{-\nu}Y_{\nu-1}(x)
\end{aligned}$$

To show (9.69), expand the derivative operation on $x^\nu Y_\nu(x)$

$$\frac{d}{dx}\left(x^\nu Y_\nu(x)\right) = \nu x^{\nu-1}Y_\nu(x)+x^\nu\frac{d}{dx}Y_\nu(x)$$

and equate with (9.67) to obtain

$$\begin{aligned}
\nu x^{\nu-1}Y_\nu(x)+x^\nu\frac{d}{dx}Y_\nu(x) &= x^\nu Y_{\nu-1}(x) \\
\frac{d}{dx}Y_\nu(x) &= Y_{\nu-1}(x)-\frac{\nu}{x}Y_\nu(x)
\end{aligned}$$

To show (9.70), expand the derivative operation on $x^{-\nu}Y_\nu(x)$

$$\frac{d}{dx}\left(x^{-\nu}Y_\nu(x)\right) = -\nu x^{-\nu-1}Y_\nu(x) + x^{-\nu}\frac{d}{dx}Y_\nu(x)$$

and equate with (9.68) to obtain

$$\begin{aligned} -\nu x^{-\nu-1}Y_\nu(x) + x^{-\nu}\frac{d}{dx}Y_\nu(x) &= -x^{-\nu}Y_{\nu+1}(x) \\ \frac{d}{dx}Y_\nu(x) &= -Y_{\nu+1}(x) + \frac{\nu}{x}Y_\nu(x) \end{aligned}$$

3. **Derivatives of $I_\nu(x)$.** Recall the definition of $I_\nu(x)$,

$$I_\nu(x) = \exp\left(-\frac{\nu\pi}{2}i\right)J_\nu(ix)$$

To show (9.71), multiply $I_\nu(x)$ by x^ν and then take the derivative with respect to x, while using (9.65),

$$\begin{aligned} \frac{d}{dx}x^\nu I_\nu(x) &= \exp\left(-\frac{\nu\pi}{2}i\right)\left[\nu x^{\nu-1}J_\nu(ix) + x^\nu\left(iJ_{\nu-1}(ix) - \frac{\nu}{x}J_\nu(ix)\right)\right] \\ &= x^\nu \exp\left(-\frac{(\nu-1)\pi}{2}i\right)J_{\nu-1}(ix) \\ &= x^\nu I_{\nu-1}(x) \end{aligned}$$

To show (9.72), multiply $I_\nu(x)$ by x^ν and then take the derivative with respect to x, while using (9.66),

$$\begin{aligned} \frac{d}{dx}x^{-\nu}I_\nu(x) &= \exp\left(-\frac{\nu\pi}{2}i\right)\left[-\nu x^{-\nu-1}J_\nu(ix) + x^\nu\left(-iJ_{\nu+1}(ix) + \frac{\nu}{x}J_\nu(ix)\right)\right] \\ &= x^{-\nu}\exp\left(-\frac{(\nu+1)\pi}{2}i\right)J_{\nu+1}(ix) \\ &= x^{-\nu}I_{\nu+1}(x) \end{aligned}$$

To show (9.73), expand the derivative operation on $x^\nu I_\nu(x)$

$$\frac{d}{dx}\left(x^\nu I_\nu(x)\right) = \nu x^{\nu-1}I_\nu(x) + x^\nu\frac{d}{dx}I_\nu(x)$$

and equate with (9.71) to obtain

$$\begin{aligned} \nu x^{\nu-1}I_\nu(x) + x^\nu\frac{d}{dx}I_\nu(x) &= x^\nu I_{\nu-1}(x) \\ \frac{d}{dx}I_\nu(x) &= I_{\nu-1}(x) - \frac{\nu}{x}I_\nu(x) \end{aligned}$$

To show (9.74), expand the derivative operation on $x^{-\nu}I_\nu(x)$

$$\frac{d}{dx}\left(x^{-\nu}I_\nu(x)\right) = -\nu x^{-\nu-1}I_\nu(x) + x^{-\nu}\frac{d}{dx}I_\nu(x)$$

and equate with (9.72) to obtain

$$\begin{aligned} -\nu x^{-\nu-1} I_\nu(x) + x^{-\nu} \frac{d}{dx} I_\nu(x) &= x^{-\nu} I_{\nu+1}(x) \\ \frac{d}{dx} I_\nu(x) &= I_{\nu+1}(x) + \frac{\nu}{x} I_\nu(x) \end{aligned}$$

4. **Derivatives of $K_\nu(x)$.** Recall the definition of $I_\nu(x)$,

$$K_\nu(x) = \exp\left(\frac{(\nu+1)\pi}{2} i\right) \left(J_\nu(ix) + iY_\nu(ix)\right)$$

To show (9.75), multiply $K_\nu(x)$ by x^ν and then take the derivative with respect to x, while using (9.65) and (9.69),

$$\begin{aligned} \frac{d}{dx} x^\nu K_\nu(x) &= \exp\left(\frac{(\nu+1)\pi}{2} i\right) \Big[\nu x^{\nu-1} \left(J_\nu(ix) + iY_\nu(ix)\right) \\ &\quad + x^\nu \left(iJ_{\nu-1}(ix) - \frac{\nu}{x} J_\nu(ix) - Y_{\nu-1}(ix) - i\frac{\nu}{x} Y_\nu(ix)\right)\Big] \\ &= -x^\nu \exp\left(\frac{(\nu)\pi}{2} i\right) \left(J_{\nu-1}(ix) + iY_{\nu-1}(ix)\right) \\ &= x^\nu K_{\nu-1}(x) \end{aligned}$$

To show (9.72), multiply $I_\nu(x)$ by x^ν and then take the derivative with respect to x, while using (9.66) and (9.70),

$$\begin{aligned} \frac{d}{dx} x^{-\nu} K_\nu(x) &= \exp\left(\frac{(\nu+1)\pi}{2} i\right) \Big[-\nu x^{-\nu-1} \left(J_\nu(ix) + iY_\nu(ix)\right) \\ &\quad + x^{-\nu} \left(-iJ_{\nu+1}(ix) + \frac{\nu}{x} J_\nu(ix) + Y_{\nu+1}(ix) + i\frac{\nu}{x} Y_\nu(ix)\right)\Big] \\ &= -x^{-\nu} \exp\left(\frac{(\nu+2)\pi}{2} i\right) \left(J_{\nu+1}(ix) + iY_{\nu+1}(ix)\right) \\ &= -x^{-\nu} K_{\nu+1}(x) \end{aligned}$$

To show (9.77), expand the derivative operation on $x^\nu K_\nu(x)$

$$\frac{d}{dx}\left(x^\nu K_\nu(x)\right) = \nu x^{\nu-1} K_\nu(x) + x^\nu \frac{d}{dx} K_\nu(x)$$

and equate with (9.75) to obtain

$$\begin{aligned} \nu x^{\nu-1} K_\nu(x) + x^\nu \frac{d}{dx} K_\nu(x) &= -x^\nu K_{\nu-1}(x) \\ \frac{d}{dx} K_\nu(x) &= -K_{\nu-1}(x) - \frac{\nu}{x} I_\nu(x) \end{aligned}$$

To show (9.78), expand the derivative operation on $x^{-\nu} K_\nu(x)$

$$\frac{d}{dx}\left(x^{-\nu} K_\nu(x)\right) = -\nu x^{-\nu-1} K_\nu(x) + x^{-\nu} \frac{d}{dx} K_\nu(x)$$

and equate with (9.72) to obtain

$$\begin{aligned} -\nu x^{-\nu-1}K_\nu(x) + x^{-\nu}\frac{d}{dx}K_\nu(x) &= -x^{-\nu}K_{\nu+1}(x) \\ \frac{d}{dx}K_\nu(x) &= -K_{\nu+1}(x) + \frac{\nu}{x}K_\nu(x) \end{aligned}$$

5. **Bessel functions of negative integral orders.** We use induction to prove the identity.

The recurrence formula yields the following two relationships,

$$\begin{aligned} J_{-n-1}(x) &= -\frac{2n}{x}J_{-n}(x) - J_{-n+1}(x) \\ J_{n-1}(x) &= \frac{2n}{x}J_n(x) - J_{n+1}(x) \end{aligned}$$

Adding and subtracting these equations,

$$\begin{aligned} J_{-n-1}(x) &= -\frac{2n}{x}\left(J_{-n}(x) - J_n(x)\right) \\ &\quad - \left(J_{-n+1}(x) + J_{n-1}(x)\right) - J_{n+1}(x) \qquad \text{(I.68)} \\ J_{-n-1}(x) &= -\frac{2n}{x}\left(J_{-n}(x) + J_n(x)\right) \\ &\quad - \left(J_{-n+1}(x) - J_{n-1}(x)\right) + J_{n+1}(x) \qquad \text{(I.69)} \end{aligned}$$

If n is even, while using the inductive hypothesis, that is, supposing that $J_n(x) = J_{-n}(x)$ and $J_{n-1}(x) = -J_{-n+1}(x)$, we can then use (I.68) and see that

$$J_{-(n+1)}(x) = -J_{n+1}(x)$$

If n is odd, while using the inductive hypothesis, that is, supposing that $J_n(x) = -J_{-n}(x)$ and $J_{n-1}(x) = J_{-n+1}(x)$, we can then use (I.69) and see that

$$J_{-(n+1)}(x) = J_{n+1}(x)$$

To complete the proof, we note that

$$J_0(x) = (-1)^0 J_0(x)$$

and with the recurrence formula,

$$J_{-1}(x) = -J_1(x)$$

We can then continue the induction process to show that the identity is satisfied for $n = 2, 3, \ldots$ and conclude that

$$J_{-n}(x) = (-1)^n J_n(x)$$

Similar approaches can be used to show the identities for $Y_{-n}(x)$, $I_{-n}(x)$ and $K_{-n}(x)$.

APPENDIX J

Additional Details and Fortification for Chapter 10

J.1 Shocks and Rarefaction

For the general quasilinear first-order PDEs, it is possible that the solutions of the characteristic equations will yield a surface that contains folds – resulting in multiple values of u for each point in some region of the space of independent variables. When this occurs, the classic solution (i.e., completely smooth solution) is not possible. Instead, a discontinuous solution that splits the domain into two or more regions with continuous surface solutions will have to suffice. A solution that covers both the classic solution and solutions with discontinuities are called **weak solutions** or **generalized solutions**. The discontinuities are known as **shocks**, and their paths can be traced as curves in the domain of the independent variables known as **shock paths**.

We limit our discussion to PDEs whose independent variables are time $0 \leq t < \infty$ and a space dimension $-\infty < x < \infty$, given by the form

$$\frac{\partial u}{\partial t} + b(x,t,u)\frac{\partial u}{\partial x} = c(x,t,u) \tag{J.1}$$

subject to a Cauchy condition

$$u(x, t=0) = u_0(x) \tag{J.2}$$

The method of characteristics immediately yields the following characteristic equations

$$\frac{dt}{ds} = 1 \quad ; \quad \frac{dx}{ds} = b(x,t,u) \quad ; \quad \frac{du}{ds} = c(x,t,u) \tag{J.3}$$

subject to initial conditions, $t(a, s=0) = 0$, $x(a, s=0) = a$, $u(a, s=0) = u_0(a)$. The solution for t is immediately given by $t = s$. This reduces the problem to

$$\frac{dx}{ds} = b(x,s,u) \quad ; \quad \frac{du}{ds} = c(x,s,u) \tag{J.4}$$

which can be solved either analytically or numerically for fixed values of a, where a is the parameter along the Cauchy condition. Because of the coupling of the equations in (J.4), the solution for x and u is a curve $C(x,u)$ that is parameterized by a and s. Unfortunately, these curves can contain folds, that is, several u values may correspond to a point (x,t).

To illustrate, consider the inviscid Burger equation given by

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \tag{J.5}$$

with the Cauchy initial condition (J.2). Then the solution of (J.4) with $b(x, s, u) = u$, $c(x, s, u) = 0$, $u(a, s = 0) = u_0(a)$, and $x(a, s = 0) = a$, is given by

$$u(a, s) = u_0(a) \qquad \text{and} \qquad x(a, s) = u_0(a)s + a$$

Furthermore, let $u_0(x)$ be given by

$$u_0(x) = \frac{3}{2}\left[\frac{1}{1 + e^{q(x)}} - \frac{1}{2.5 + q(x)} + \frac{1}{2}\right] \quad \text{with} \quad q(x) = \left(\frac{x - 10}{10}\right)^2 \tag{J.6}$$

We can plot $u(a, s)$ versus $x(a, s)$ at different fixed values of s with $-80 \le a \le 100$ as shown in Figure J.1.

From the plots in Figure J.1, we see that as s increases, the initial shape moves to the right and slants more and more to the right. At $s = 29.1$, portions of the curve near $x = 41.0$ will have a vertical slope, and a fold is starting to form. When $s = 80$, three values of u correspond to values in the neighborhood of $x = 78$. At $s = 120$, portions of the curve near $x = 54.8$ will again have a vertical slope. Then at $s = 300$, we see that around $x = 165$ and $x = 235$, three values of u correspond to each of these x values. Finally, we see that at $s = 600$, there are five values of u that correspond to $x = 370$.

J.1.1 Break Times

We refer to the values of $s(= t)$ in which portions of the curves just begin to fold as the **break times**, denoted by s_{break}. From the plots given in Figure J.1, we see that several shocks are possible, each with their respective break times. Assuming that the initial data $u_0(a)$ are continuous, the shock that starts to form at the break time is along a characteristic that starts at a, which intersects with a neighboring characteristic that starts at $a + \epsilon$. This means

$$\frac{\partial x}{\partial a} = 0 \qquad \text{at } s = s_{\text{break}} \tag{J.7}$$

Suppose the shock at the break time will belong to a characteristic starting from a that belongs to a range $[a_{\text{left}}, a_{\text{right}}]$. For instance, one could plot the characteristics based on a uniform distribution of a and then determine adjacent values of a whose characteristics intersect, as shown in Figure J.2. The values of a_{left} and a_{right} can then be chosen to cover this pair of adjacent values of a. The break time s_{break} and the critical point a_{critical} can then be determined by solving the following minimization problem

$$\min_{a \in [a_{\text{left}},\ a_{\text{right}}]} \{s\} \qquad \text{such that} \qquad \left(\frac{\partial x}{\partial a} \le 0\right) \tag{J.8}$$

The values of x at s_{break} along the characteristic corresponding to a_{critical} will be the **break position**, denoted by x_{break},

$$x_{\text{break}} = x\left(a_{\text{critical}}, s_{\text{break}}\right) \tag{J.9}$$

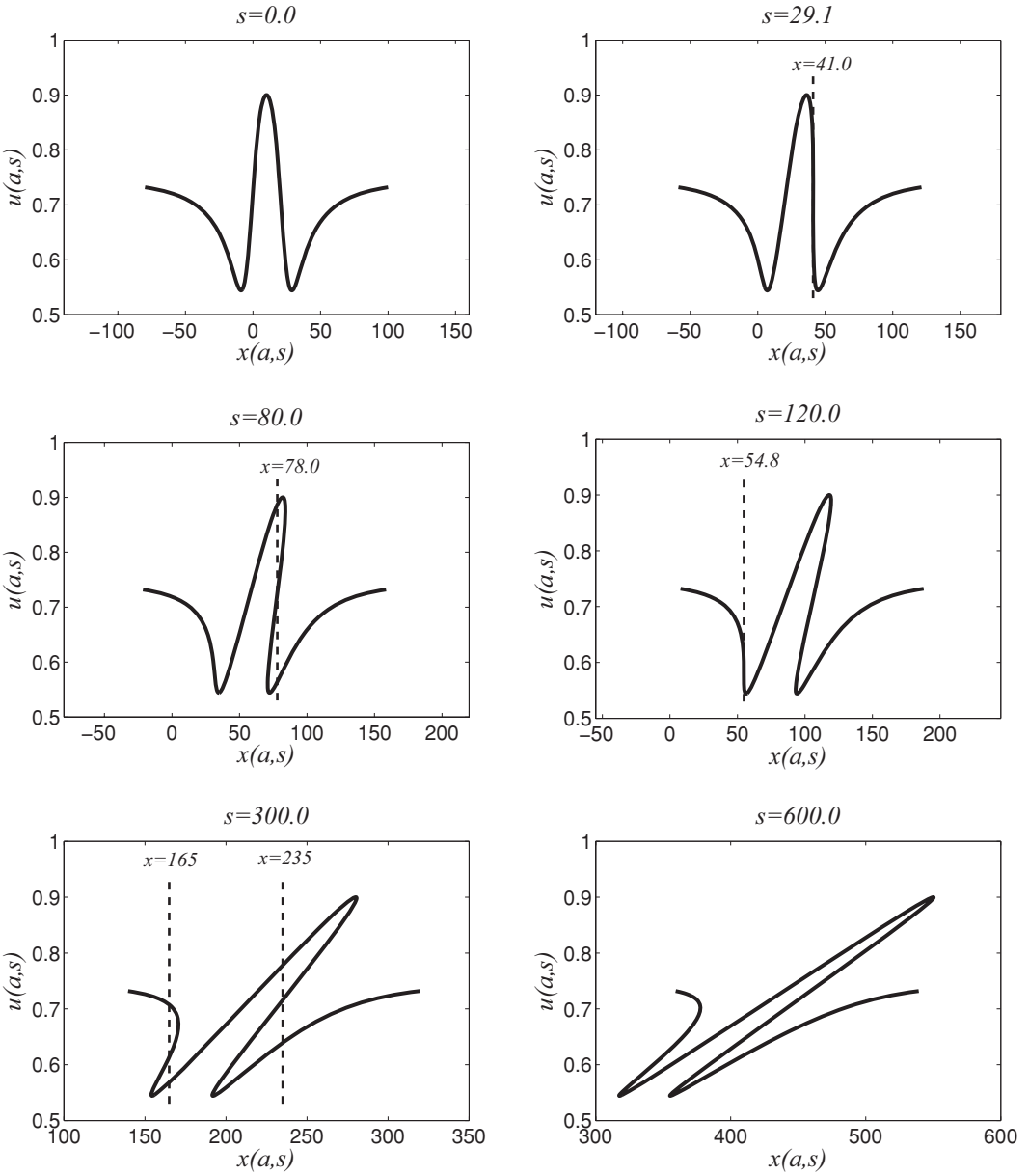


Figure J.1. Plots of u versus x for different values of s, with $-80 \leq a \leq 100$.

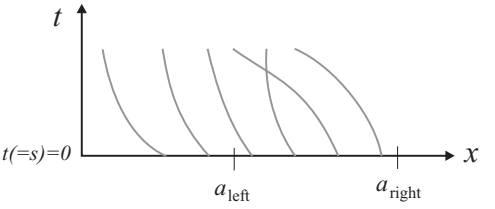


Figure J.2. Determination of a_{left} and a_{right}.

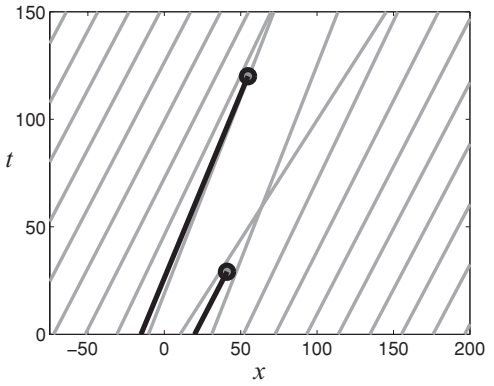


Figure J.3. The characteristics corresponding to uniformly distributed values of a. Also included are two characteristics along a_{critical}. The circles are the break points $(x_{\text{break}}, s_{\text{break}})$.

In particular, the characteristics (x, t) for the inviscid Burger equation (J.5) are given by straight lines

$$t = \frac{x - a}{u_0(a)} \quad \text{if} \quad u_0(a) \neq 0 \tag{J.10}$$

(If $u_0(a) = 0$, the characteristics are vertical lines at a.) For the initial data $u_0(x)$ of (J.6), a set of characteristics corresponding to a set of uniformly distributed a values is shown in Figure J.3. From this figure, we could set $[a_{\text{left}}, a_{\text{right}}] = [0, 50]$ to determine the break time of the first shock point. We could also set $[a_{\text{left}}, a_{\text{right}}] = [-50, 0]$ to determine the break time of the other shock point. Solving the minimization problem of (J.8) for each of these intervals yields the following results:

$$\begin{array}{ccccc} s_{\text{break},1} = 29.1 & ; & a_{\text{critical},1} = 19.84 & ; & x_{\text{break},1} = 41.0 \\ s_{\text{break},2} = 120 & ; & a_{\text{critical},2} = -15.25 & ; & x_{\text{break},2} = 54.8 \end{array}$$

In Figure J.3, this information is indicated by two darker lines starting at $(t, x) = (0, a_{\text{critical}})$ and ending at the points $(t, x) = (s_{\text{break}}, x_{\text{break}})$. These break times and break positions are also shown in Figure J.1 for $s = 29.1$ and $s = 120$ to be the correct values where portions of the curves are starting to fold.

J.1.2 Weak Solutions

Once the break times and positions have been determined, a discontinuity in solution will commence as $t = s$ increases and a weak solution has to be used. A function $\tilde{u}(x, t)$ is a weak solution of a partial differential equation, such as (J.1),

$$\frac{\partial u}{\partial t} + b(x, t, u)\frac{\partial u}{\partial x} = c(x, t, u)$$

if

$$\int_0^\infty \int_{-\infty}^\infty \left(\vartheta(x, t) \left[\frac{\partial \tilde{u}}{\partial t} + b(x, t, u)\frac{\partial \tilde{u}}{\partial x} - c(x, t, \tilde{u}) \right] \right) dx\, dt = 0 \tag{J.11}$$

for *all* smooth functions $\vartheta(x, t)$, which has the property that $\vartheta = 0$ for x outside of some closed interval $[x_{\text{left}}, x_{\text{right}}]$ and for t outside of some closed interval $[t_{\text{low}}, t_{\text{high}}]$ (with $-\infty < x_{\text{left}} < \infty$ and $0 \leq t_{\text{low}} < t_{\text{high}} < \infty$). The main idea of (J.11) is that via integration by parts, partial derivatives of discontinuous $\tilde{u}(x, t)$ can be avoided by transferring the derivative operations instead on continuous functions $\vartheta(x, t)$.

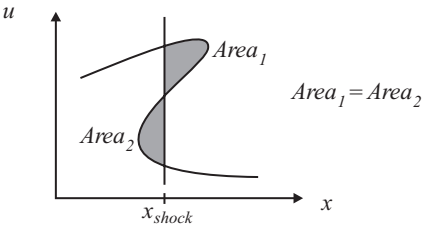


Figure J.4. The location of x_{shock} based on equal area rule.

Another important point is that the function $\vartheta(x, t)$ is kept arbitrary; that is, there is no need to specify this function nor the domain given by x_{right}, x_{left}, t_{low}, or t_{high}. This will keep the number of discontinuities to a minimum. For instance, if a continuous $\tilde{u}$ can satisfy (J.11) for arbitrary ϑ, then no discontinuity need to be introduced, and $\tilde{u} = u$, a classic solution.

For the special case in which $c(x, t, \tilde{u}) = c(x, t)$ is continuous, let the desired discontinuity that satisfies (J.11) occur at $(t = s, x_{\text{shock}}(s))$. The value of x_{shock} will occur when two characteristics, one initiated at $a = a_{(-)}$ and another initiated at $a = a_{(+)}$, intersected to yield x_{shock}. The condition (J.11) implies that x_{shock} is located at a position where the area of the chopped region to right of x_{shock} is equal to the area of the chopped region to the left of x_{shock}, as shown in Figure J.4.

J.1.3 Shock Fitting

Based on the equal area rule, a shock path $x_{\text{shock}}(s)$ with $s \geq s_{\text{break}}$ can be determined by solving the following integral:

$$\int_{a_{(-)}}^{a_{(+)}} \left[u(a, s) \frac{\partial x}{\partial a} \right] da = 0 \tag{J.12}$$

such that $x\left(a_{(-)}, s\right) = x\left(a_{(+)}, s\right) = x_{\text{shock}}(s)$.

Generally, the location of the shock path, especially one that is based on the equal area rule, will require numerical solutions. We outline a scheme to determine the shock path in a region where the folds yield triple values u for some x (i.e., the case shown in Figure J.4). This scheme depends on the following operations that require nonlinear solvers:

1. **Detection of Fold Edges**. Let a_{critical} be the value found at the break time of the shock, then

$$\begin{pmatrix} a_{\text{edge},1} \\ a_{\text{edge},2} \end{pmatrix} = \text{EDGE}\left(a_{\text{critical}}\right) \tag{J.13}$$

where

$$\left.\frac{\partial x}{\partial a}\right|_{a_{\text{edge},1}} = 0 = \left.\frac{\partial x}{\partial a}\right|_{a_{\text{edge},2}} \qquad \text{and} \qquad a_{\text{edge},1} < a_{\text{critical}} < a_{\text{edge},2}$$

2. **Root Finding for** a. Let x_g be in a region where three different values of u correspond to one value of x and s.

$$\begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{pmatrix} = \text{FINDa}\,(x_g, s) \tag{J.14}$$

where

$$a_1 > a_2 > a_3 \quad \text{and} \quad x(a_1, s) = x(a_2, s) = x(a_3, s) = x_g$$

3. **Evaluation of Net Area**.

$$\mathcal{I}(y) = \int_{a_1(y)}^{a_3(y)} \left[u(s, a) \frac{\partial x}{\partial a} \right] da \tag{J.15}$$

where $a_1(y)$ and $a_3(y)$ are found using the operation FINDa(y).

Shock-Fitting Scheme:

- Given: s_{break}, Δs and a_{critical}
- For $s = s_{\text{break}} + \Delta s, s_{\text{break}} + 2\Delta s, \ldots$

1. Calculate x_g as the average of the edge values,

$$x_g = \frac{1}{2} \left[\, x\left(s, a_{\text{edge},1}\right) + x\left(s, a_{\text{edge},2}\right) \, \right]$$

where $a_{\text{edge},1}$ and $a_{\text{edge},2}$ are found using EDGE (a_{critical}).
2. Using x_g as the initial guess, find $\hat{x}$ such that

$$\mathcal{I}\,(\,\hat{x}\,) = 0$$

3. $x_{\text{shock}}(s) \leftarrow \hat{x}$

Using the shock-fitting scheme on the Burger equation (J.5) subject to the initial condition (J.6), we find two shocks paths, one starting at $(t, x) = (29.1, 41)$ and the other one starting at $(t, x) = (120, 54.8)$, as shown in Figure J.5. One can see that the shock paths for this example are approximately straight lines. Furthermore, we also note that the two shock paths do not intersect with each other. Thus even though the curves shown in Figure J.1 for the case of $s = 600$ may contain portions in which u has more that three values corresponding to a specific value of x, it does not immediately imply that the two shocks path would intersect. In the next section, we show that the shock paths will need to satisfy jump conditions and that the path being linear is not due to the initial condition but rather due to the coefficient $b(x, t, u) = u$ for the inviscid Burger equation.

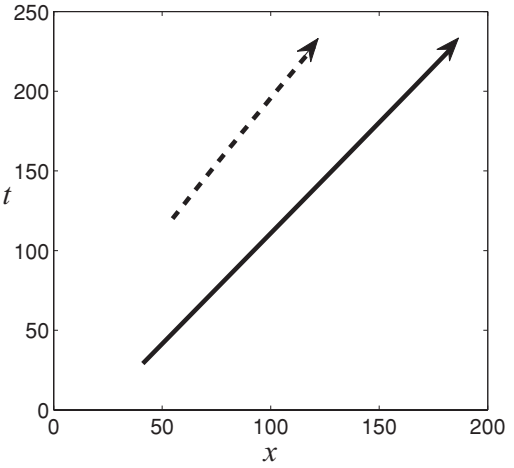


Figure J.5. Two shock paths for the Burger equation under the conditions given by (J.6) using the shock-fitting scheme based on the equal-area principle.

J.1.4 Jump Conditions

We further limit our discussion to the case where $b(x, t, u) = b(u)$ in (J.1). Under this condition, the differential equation (J.1) results from (or can be recast as) a conservation equation given by

$$\frac{\partial}{\partial t}\int_{\alpha}^{\beta} u\,(x,t)\,dx = \left(\text{flux}\left(u_{(\alpha,t)}\right) - \text{flux}\left(u_{(\beta,t)}\right)\right) + \int_{\alpha}^{\beta} c\,(x,t,u)\,dt \qquad \text{(J.16)}$$

where $\text{flux}(u) = \int b(u)du$ and $c(x, t, u)$ is the volumetric rate of generation for u.

Now suppose at t, $\alpha < \beta$ is chosen so that the shock discontinuity is at $x = x_s$ located between α and β. Let x_s^- and x_s^+ be the locations slightly to left of and right of x_s, respectively. Then

$$\frac{\partial}{\partial t}\left(\int_{\alpha}^{x_s^-} u\,(x,t)\,dx + \int_{x_s^+}^{\beta} u\,(x,t)\,dx\right) = \left(\text{flux}\left(u_{(\alpha,t)}\right) - \text{flux}\left(u_{(\beta,t)}\right)\right) + \int_{\alpha}^{\beta} c\,(t,x)\,dt \qquad \text{(J.17)}$$

Applying the Leibnitz rule (5.52) to (J.17), we obtain

$$\int_{\alpha}^{x_s^-} \frac{\partial u}{\partial t}dx + u\left(x_s^-, t\right)\frac{dx_s^-}{dt} + \int_{x_s^+}^{\beta} \frac{\partial u}{\partial t}dx - u\left(x_s^+, t\right)\frac{dx_s^+}{dt}$$
$$= \left(\text{flux}\left(u_{(\alpha,t)}\right) - \text{flux}\left(u_{(\beta,t)}\right)\right) + \int_{\alpha}^{\beta} c\,(t,x)\,dt$$

Next, we take the limit as $\alpha \to x_s^-$ and $\beta \to x_s^+$. This yields

$$u\left(x_s^-, t\right)\frac{dx_s^-}{dt} - u\left(x_s^+, t\right)\frac{dx_s^+}{dt} = \left(\text{flux}\left(u_{(x_s^-,t)}\right) - \text{flux}\left(u_{(x_s^+,t)}\right)\right)$$

where $\int_{x_s^-}^{x_s^+} c dx = 0$ if we assume that $c(x, t, u)$ is piecewise continuous.[1] As the previous section showed, the shock propagation is continuous and implies that

[1] A more complete assumption for c is that it does contain any Dirac delta distribution (i.e., delta impulses).

$dx_s^+/dt = dx_s^-/dt = dx_s/dt$. Using the **jump notation**, $\lfloor\, \xi \,\rceil = \xi|_{u(x_s^-,t)} - \xi|_{u(x_s^+,t)}$, we arrive at

$$\frac{dx_s}{dt} = \frac{\lfloor\, \text{flux}(u) \,\rceil}{\lfloor\, u \,\rceil} \tag{J.18}$$

which is known as the **Rankine-Hugoniot jump conditions**.[2] This condition equates the **shock speed** dx_s/dt to the ratio of jump values of the flux(u) and u. This can be used to help find the next position of the discontinuity for some simple cases; that is, the shock path can be found using (J.18) without using the equal area approach discussed in the previous section. Furthermore, the jump condition can be used to eliminate some shock solutions that satisfy the partial differential equations on the piecewise continuous region, but nonetheless would violate the Rankine-Hugoniot conditions.

EXAMPLE J.1. Consider the inviscid Burger's equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$$

subject to the discontinuous condition

$$u(x,0) = \begin{cases} 1 & \text{if } x \le a \\ 0 & \text{if } x > a \end{cases}$$

For this problem, $b(u) = u$, and the flux is

$$\text{flux}(u) = \int u\, du = \frac{u^2}{2}$$

Because the initial condition is immediately discontinuous, the break time in this case is at $t = 0$. Using the Rankine-Hugoniot jump condition (J.18),

$$\frac{dx_s}{dt} = \frac{\left\lfloor\, \dfrac{u^2}{2} \,\right\rceil}{\lfloor\, u \,\rceil} = \frac{u^+ + u^-}{2}$$

Because $u =$ constant along the characteristics, $u^- = 1$ and $u^+ = 0$, yielding

$$\frac{dx_s}{dt} = \frac{1}{2} \quad \rightarrow \quad x_s = \frac{t}{2} + a$$

Thus the solution is given by

$$u(x,t) = \begin{cases} 1 & \text{if } x \le \dfrac{t}{2} + a \\ \\ 0 & \text{if } x > \dfrac{t}{2} + a \end{cases}$$

[2] If the conservation equation (J.16) is given in a more general form by

$$\frac{\partial}{\partial t}\int_\alpha^\beta \phi(x,t,u)\, dx = \Big(\text{flux}\left(\alpha, t, u_{(\alpha,t)}\right) - \text{flux}\left(\beta, t, u_{(\beta,t)}\right)\Big) + \int_\alpha^\beta c(x,t,u)\, dt$$

the Rankine-Hugoniot condition (J.18) should be replaced instead by

$$\frac{dx_s}{dt} = \frac{\lfloor\, \text{flux}\,(x,t,u) \,\rceil}{\lfloor\, \phi\,(x,t,u) \,\rceil}$$

The jump conditions given in (J.18) will generally not guarantee a unique solution. Instead, additional conditions known as **admissibility conditions**, more popularly known as **Lax entropy conditions**, are needed to achieve physical significance and uniqueness. We now state without proof the following condition known as the **Lax entropy conditions** applicable to the case where flux(u) is convex, that is, $d^2\,\text{flux}/du^2 > 0$:

$$\left.\frac{d\,\text{flux}}{du}\right|_{u=u^-} \geq \frac{dx_s}{dt} \geq \left.\frac{d\,\text{flux}}{du}\right|_{u=u^+} \tag{J.19}$$

Thus these conditions put the necessary bounds on the shock speed, at least for the case of convex fluxes.[3] This condition simply implies that if the characteristics appear to be intersecting in the direction of decreasing t (time reversal), then this solution is not admissible.

EXAMPLE J.2. For the inviscid Burger equation and initial condition given by

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \qquad u(x,0) = \begin{cases} A & \text{for } x \leq 0 \\ B & \text{for } x > 0 \end{cases}$$

where $A < B$. Let $0 < m < 1$, then a solution that contains two shock paths given by

$$u(x,t) = \begin{cases} A & \text{for } x \leq (A+m)t/2 \\ m & \text{for } (A+m)t/2 < x \leq (m+B)t/2 \\ B & \text{for } x > (m+B)/2 \end{cases} \tag{J.20}$$

will satisfy the Rankine-Hugoniot jump conditions at both regions of discontinuities. This means there are an infinite number of possible solutions that will satisfy the differential equation and jump discontinuity conditions.

However, using the entropy conditions given in (J.19), we obtain

$$A > \frac{dx_s}{dt} > B$$

which is not true (because it was a given in the initial condition that $A < B$). This means that the discontinuous solutions in (J.20) are inadmissible based on the entropy conditions. We see in the next section that the rarefaction solution turns out to be the required solution.

J.1.5 Rarefaction

When a first-order quasilinear PDE is coupled with a discontinuous initial condition, we call this problem a **Riemann problem**. We already met these types of problems in previous sections. In Example J.1, we saw that the Riemann problem there resulted in a shock propagated solution for the inviscid Burger equation, where $u(x \leq a, 0) = 1$ and $u(x > a, 0) = 0$. However, if the conditions were switched, that is, with $u(x \leq a, 0) = 0$ and $u(x > a, 0) = 1$, the method of characteristics will leave a domain in the (x, t) plane without specific characteristic curves, as shown in Figure J.6.[4] In contrast

[3] A set of more general conditions are given by **Oleinik entropy conditions**, which are derived using the approach known as the vanishing viscosity methods.

[4] If the initial condition were not discontinuous, this would have been filled in without any problem, especially because the characteristics would not even intersect and no shocks would occur.

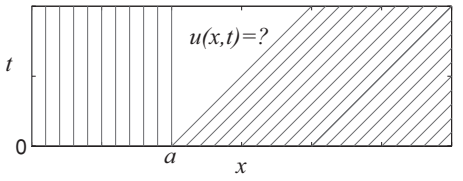


Figure J.6. Rarefaction in a Riemann problem.

to the shock-fitting problem, this case is called the **rarefaction**, a term that originates from the phenomenon involving wave expansion of gases.

We limit our discussion to the case of (J.1), where $b(x, t, u) = b(u)$ and $c(x, t, u) = 0$ with the additional assumption that the inverse function $b^{-1}(\cdot)$ can be obtained. Consider

$$\frac{\partial u}{\partial t} + b(u)\frac{\partial u}{\partial x} = 0 \tag{J.21}$$

subject to

$$u(x, 0) = \begin{cases} u^{\text{left}} & \text{if } x \le a \\ u^{\text{right}} & \text{if } x > a \end{cases} \tag{J.22}$$

where $b\left(u^{\text{left}}\right) < b\left(u^{\text{right}}\right)$. Let the initial data be parameterized by ξ, that is, at $s = 0$, $t(s = 0) = 0$, $x(s = 0) = \xi$ and $u(\xi, 0) = u^{\text{left}}$ or $u(\xi, 0) = u^{\text{right}}$ when $\xi \le a$ or $\xi > a$, respectively. Then the characteristics are given by

$$x = b\left(u(\xi, 0)\right)t + \xi \quad \rightarrow \quad x = \begin{cases} b\left(u^{\text{left}}\right)t + \xi & \text{if } \xi \le a \\ b\left(u^{\text{right}}\right)t + \xi & \text{if } \xi > a \end{cases}$$

Rarefaction will start at $x = a$ when $t = 0$. The characteristics at this point can be rearranged to be

$$u(a, 0) = \lim_{(x,t)\to(a,0)} b^{-1}\left(\frac{x - a}{t}\right)$$

We could pose that the solution in the rarefaction domain to be of the form

$$u(x, t) = b^{-1}\left(\frac{x - a}{t}\right)$$

and see that this will satisfy the differential equation, that is,

$$\frac{\partial u}{\partial t} + b(u)\frac{\partial u}{\partial x} = 0 \quad \rightarrow \quad \frac{1}{t}\left(-\frac{x - a}{t} + \frac{x - a}{t}\right)\left(\frac{d}{d\left((x - a)/t\right)}b^{-1}\left(\frac{x - a}{t}\right)\right) = 0$$

The solution of (J.21) subject to (J.22) is then given by

$$u(x, t) = \begin{cases} u^{\text{left}} & \text{if } x \le b\left(u^{\text{left}}\right)t + a \\ b^{-1}\left(\dfrac{x - a}{t}\right) & \text{if } b\left(u^{\text{left}}\right)t + a < x \le b\left(u^{\text{right}}\right)t + a \\ u^{\text{right}} & \text{if } x > b\left(u^{\text{right}}\right)t + a \end{cases} \tag{J.23}$$

It is left as an exercise (**E10.20**) to show that (J.23) is piecewise continuous.

EXAMPLE J.3. For the inviscid Burger equation and initial conditions given by,

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \quad \text{subject to} \quad u(x,0) = \begin{cases} 0.5 & \text{if } x \le 2 \\ 1.5 & \text{if } x > 2 \end{cases}$$

the rarefaction solution becomes

$$u(x,t) = \begin{cases} 0.5 & \text{if } x \le 0.5t + 2 \\ \dfrac{x-2}{t} & \text{if } 0.5t + 2 < x \le 1.5t + 2 \\ 1.5 & \text{if } x > 1.5t + 2 \end{cases}$$

J.2 Classification of Second-Order Semilinear Equations: $n > 2$

When the number of independent variables is more than two, the principal part of the semilinear equation is given by the following general form:

$$F_{\text{prin}} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{i,j}\,(\mathbf{x})\,\mu_{i,j} \tag{J.24}$$

Just as we did in the previous section, we look for a new set of independent variables $\{\xi_1, \ldots, \xi_n\}$, such that under the new coordinates,

$$\hat{F}_{\text{prin}}\,(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} \epsilon_i \mu_{i,i}^{(\xi)} \qquad \text{where} \quad \epsilon = 0, \ -1 \text{ or } +1 \tag{J.25}$$

where we use the following notation:

$$\begin{aligned} \mu_i^{(\xi)} &= \frac{\partial u}{\partial \xi_i} \\ \mu_{i,j}^{(\xi)} &= \frac{\partial^2 u}{\partial \xi_i \partial \xi_j}, \quad 1 \le i, j \le n \\ \mu_{i,j,k}^{(\xi)} &= \frac{\partial^3 u}{\partial \xi_i \partial \xi_j \partial \xi_k}, \quad 1 \le i, j, k \le n \\ &\vdots \end{aligned} \tag{J.26}$$

The classification of these forms is then given in the following definition:

Definition J.1. *The canonical forms of second-order semilinear equations given by*

$$\sum_{i=1}^{n} \epsilon_i \mu_{i,i}^{(\xi)} = f\left(\xi_1, \ldots, \xi_n, u, \mu_1^{(\xi)}, \ldots, \mu_n^{(\xi)}\right) \tag{J.27}$$

are classified to be **elliptic**, **parabolic**, **hyperbolic**, *and* **ultrahyperbolic** *according to the following conditions:*

Elliptic:	*if* $\epsilon_i \neq 0$ *all have the same sign*
Parabolic:	*if* $\epsilon_i = 0$ *for some* $1 \le i \le n$
Hyperbolic:	*if* $\epsilon_i \neq 0$ *all have the same sign except for one*
Ultra-Hyperbolic:	*if* $\epsilon_i \neq 0$ *and* $\epsilon_a \ge \epsilon_b > 0$, $\epsilon_c \le \epsilon_d < 0$ *for* $a \neq b \neq c \neq d$

Unfortunately, finding a change in coordinates,

$$\xi_i = \xi_i(x_1, x_2, \ldots, x_n) \qquad i = 1, 2, \ldots, n \tag{J.28}$$

that would yield the canonical forms (J.27) may not be always be possible. However, when the coefficients in the principal parts are constants, then the equation can be transformed into the canonical forms given in Definition J.1.

THEOREM J.1. *Consider the second-order semilinear equation given by*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}\mu_{i,j} = f(\mathbf{x}, u, \mu_1, \ldots, \mu_n) \tag{J.29}$$

where $A_{i,j} = A_{j,i}$ are constants. Let $(\xi_1, \xi_2, \ldots, \xi_n)$ be a set of new independent variables defined by

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix} = DU \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{J.30}$$

where, U is an orthogonal matrix such that $UAU^T = \Lambda$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues of A and $D = \text{diag}(d_1, d_2, \ldots, d_n)$, where

$$d_i = \begin{cases} 1/\sqrt{|\lambda_i|} & \text{if } \lambda_1 \neq 0 \\ \\ 0 & \text{if } \lambda_i = 0 \end{cases}$$

and λ_i is the ith eigenvalue of A. Then under the change of coordinates given by (J.30), the partial differential equation (J.29) becomes

$$\sum_{i=1}^{n} \epsilon_i \mu_{i,i}^{(\xi)} = \hat{f}\left(\xi_1, \ldots, \xi_n, u, \mu_1^{(\xi)}, \ldots, \mu_n^{(\xi)}\right) \qquad , \ \epsilon_i = 0, \ -1 \text{ or } 1 \tag{J.31}$$

PROOF. With (J.30), the partial differential operators $\partial/\partial x_i$ are

$$\begin{pmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{pmatrix} = U^T D \begin{pmatrix} \partial/\partial \xi_1 \\ \partial/\partial \xi_2 \\ \vdots \\ \partial/\partial \xi_n \end{pmatrix}$$

Using the partial differential operators, the partial differential equation (J.29) can written as

$$\begin{pmatrix} \partial/\partial x_1 & \partial/\partial x_2 & \cdots & \partial/\partial x_n \end{pmatrix} A \begin{pmatrix} \partial/\partial x_1 \\ \partial/\partial x_2 \\ \vdots \\ \partial/\partial x_n \end{pmatrix} u = f(\mathbf{x}, u, \mu_1, \ldots, \mu_n)$$

or

$$\begin{pmatrix} \partial/\partial\xi_1 & \cdots & \partial/\partial\xi_n \end{pmatrix} DUAU^T D \begin{pmatrix} \partial/\partial\xi_1 \\ \vdots \\ \partial/\partial\xi_n \end{pmatrix} u = f(\mathbf{x}, u, \mu_1, \ldots, \mu_n)$$

which can then be simplified to be

$$\sum_{i=1}^{n} \text{sign}(\lambda_i)\, \mu_{i,i}^{(\xi)} = \hat{f}\left(\xi_1, \ldots, \xi_n, u, \mu_1^{(\xi)}, \ldots, \mu_n^{(\xi)}\right)$$

where

$$\text{sign}(\lambda_i) = \begin{cases} +1 & \text{if } \lambda_i > 0 \\ 0 & \text{if } \lambda_i = 0 \\ -1 & \text{if } \lambda_i < 0 \end{cases}$$

EXAMPLE J.4. Consider the second-order differential equation with three independent variables x, y, and z,

$$3\frac{\partial^2 u}{\partial x^2} + 5\frac{\partial^2 u}{\partial x \partial y} - 2\frac{\partial^2 u}{\partial x \partial z} + \frac{\partial^2 u}{\partial y^2} + 2\frac{\partial^2 u}{\partial y \partial z} + 3\frac{\partial^2 u}{\partial z^2} = ku \tag{J.32}$$

We now look for new coordinates ξ_1, ξ_2, and ξ_3 that would transform (J.32) into the canonical form given in (J.27) for purposes of classification.

Extracting the coefficients into matrix A,

$$A = \begin{pmatrix} 3 & 2.5 & -1 \\ 2.5 & 1 & 1 \\ -1 & 1 & 3 \end{pmatrix}$$

Using Schur triangularization, we can obtain the orthogonal matrix U

$$U = \begin{pmatrix} 0.5436 & -0.7770 & 0.3176 \\ -0.0153 & 0.3692 & 0.9292 \\ 0.8392 & 0.5099 & -0.1888 \end{pmatrix}$$

and the diagonal normalizing matrix D,

$$D = \text{diag}\,(0.9294, 0.5412, 0.4591)$$

The new coordinates are obtained as follows

$$\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = DU \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0.5252x - 0.7221y + 0.5029z \\ -0.0083x + 0.1998y + 9.5029z \\ 0.3853x + 0.2341y - 0.0867z \end{pmatrix}$$

As a check, we can apply the change of coordinates while noting that the second-order derivatives of ξ_i, e.g., $\partial^2 \xi_i / (\partial x \partial y)$, are zero. Thus

$$\frac{\partial^2 u}{\partial p\, \partial q} = \sum_{i=1}^{3}\sum_{j=1}^{3} \left(\frac{\partial \xi_i}{\partial p}\frac{\partial \xi_j}{\partial q}\right) \frac{\partial^2 u}{\partial \xi_i \partial \xi_j} \qquad ; \text{for } p, q = x, y, z$$

When substituted into (J.32), we obtain

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\epsilon_{ij}\frac{\partial^2 u}{\partial\xi_i\partial\xi_j}=ku$$

where

$$\epsilon_{ij}=a_{11}\frac{\partial^2\xi_i}{\partial x^2}+a_{12}\frac{\partial^2\xi_i}{\partial x\partial y}+\cdots+a_{33}\frac{\partial^2\xi_j}{\partial z^2}$$

For instance,

$$\begin{aligned}\epsilon_{12} &= (3)(0.5052)(-0.083)+(2.5)(0.5052)(0.1998)\\ &\quad +(-1)(0.5052)(0.5029)+\ \cdots\ +(3)(0.2952)(0.5029)\end{aligned}$$

After performing the computations, we find $\epsilon_{11}=-1,\epsilon_{22}=\epsilon_{33}=1$ and for $i\neq j$, $\epsilon_{ij}=0$, i.e.

$$-\frac{\partial^2 u}{\partial\xi_1\partial\xi_1}+\frac{\partial^2 u}{\partial\xi_2\partial\xi_2}+\frac{\partial^2 u}{\partial\xi_3\partial\xi_3}=ku$$

Thus we can classify (J.32) to be hyperbolic.

J.3 Classification of High-Order Semilinear Equations

For partial differential equations that have orders higher than two, the canonical forms are more difficult to fix. Instead, the classification is to indicate whether a solution by characteristics is possible or not. We limit our discussion to cases involving two independent variables.

Recall that for the second-order equation with two independent variables given by

$$A(x,y)u_{xx}+B(x,y)u_{x,y}+C(x,y)u_{y,y}=f(x,y,u,u_x,u_y) \tag{J.33}$$

the characteristics were obtained by solving the characteristic form,

$$Q(\xi_x,\xi_y)=A(x,y)\xi_x^2+B(x,y)\xi_x\xi_y+C(x,y)\xi_y^2 \tag{J.34}$$

Prior to the determination of whether the equation can be transformed to the hyperbolic, elliptic, or parabolic canonical forms, the roots of the characteristic form became critical. For the hyperbolic equations, the roots were real. For the parabolic equations, the roots were equal. And for the elliptic equations, the roots were complex. By using the character of the roots, we can then extend the concept of hyperbolic, parabolic, and elliptic to higher orders.

Definition J.2. *For an mth-order semilinear partial differential equation in two independent variables x and y,*

$$\sum_{i=0}^{m}A_i(x,y)\frac{\partial^m u}{\partial^i x\partial^{m-i}y}=f\left(x,y,u,\mu_{[1]},\ldots,\mu_{[m-1]}\right) \tag{J.35}$$

the **characteristic form** *is given by*

$$Q\left(\xi_x, \xi_y\right) = \sum_{i=0}^{m} A_i(x, y)\xi_x^i \xi_y^{m-i} = \prod_{i=0}^{m} \left(\xi_x - r_i(x, y)\xi_y\right) \tag{J.36}$$

where $r_i(x, y)$, $i = 1, 2, \ldots, m$ *are the roots of the polynomial*

$$\sum_{i=0}^{m} A_i\left(x, y\right) r^i = 0 \tag{J.37}$$

Then at a fixed point (x, y)*, equation (J.35) is classified as*

Hyperbolic:	*if all the roots* r_i *are real and distinct*
Parabolic:	*if all the roots* r_i *are equal*
Elliptic:	*if all the roots* r_i *are complex*
Mixed:	*otherwise*

Thus for the hyperbolic case, we can determine m characteristics $\xi_{(i)}(x, y)$ by solving the m characteristic equations given by

$$\xi_{(i),x} - r_i(x, y)\xi_{(i),y} = 0 \qquad i = 1, 2, \ldots, m \tag{J.38}$$

that is, solving

$$\frac{dx}{1} = -\frac{dy}{r_i(x, y)} \qquad \xi_{(i)}(x, y) = \text{constant} \tag{J.39}$$

Furthermore, note that if m is an odd number, then the partial differential equation can not be elliptic.

APPENDIX K

Additional Details and Fortification for Chapter 11

K.1 d'Alembert Solutions

Having the general solution for the one-dimensional wave equation as given in (11.17), we can start fitting them to initial and boundary conditions. We first consider the case with an infinite x domain and only the initial conditions are specified. The solution for this type of problem is given by a form known as the **d'Alembert solution**. Next, we consider the case of semi-infinite domain, that is, $x \geq 0$, where we extend the applicability of d'Alembert solutions for systems with additional boundary conditions. Finally, we consider the case where the spatial domain is a finite segment, for example, $0 \leq x \leq L$.

K.1.1 Infinite-Domain Wave Equation with Only Initial Conditions

The system is described by

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \qquad -\infty \leq x \leq \infty$$

$$\text{subect to} \qquad u(x,0) = f(x) \qquad \text{and} \qquad \frac{\partial u}{\partial t}(x,0) = g(x)$$

Applying the initial conditions to the general solution for u given in (11.17),

$$\begin{aligned} u(x,0) &= f(x) = \phi(x) + \psi(x) \\ \frac{\partial u}{\partial t}(x,0) &= g(x) = c\frac{d\phi}{dx} - c\frac{d\psi}{dx} \end{aligned} \tag{K.1}$$

because at $t = 0$, $\phi(x + ct) = \phi(x)$ and $\psi(x - ct) = \psi(x)$. Taking the derivative of $f(x)$,

$$\frac{df}{dx} = \frac{d\phi}{dx} + \frac{d\psi}{dx} \tag{K.2}$$

Solving (K.1) and (K.2) simultaneously for $d\phi/dx$ and $d\psi/dx$,

$$\frac{d\phi}{dx} = \frac{1}{2}\frac{df}{dx} + \frac{1}{2c}g(x) \quad \text{and} \quad \frac{d\psi}{dx} = \frac{1}{2}\frac{df}{dx} - \frac{1}{2c}g(x)$$

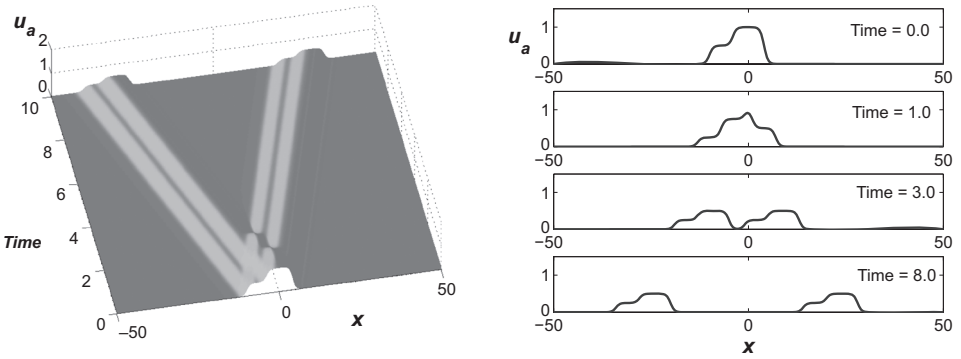


Figure K.1. A surface plot of the trajectories of u_a (left) and a set of four snapshots of u_a at different time instants (right) for the d'Alembert's solution based on zero initial velocity.

and

$$\begin{aligned}\phi(x) &= \frac{1}{2}f(x) + \frac{1}{2c}\int_0^x g(\tau)d\tau + \kappa_1 \\ \psi(x) &= \frac{1}{2}f(x) - \frac{1}{2c}\int_0^x g(\tau)d\tau + \kappa_2\end{aligned}$$

However, $\kappa_1 = -\kappa_2$ because $f(0) = \phi(0) + \psi(0)$. Returning to (11.17),

$$u(x,t) = \frac{1}{2}\left[f\left(x - ct\right) + f\left(x + ct\right)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\tau)d\tau \tag{K.3}$$

Equation (K.3) is known as the **d'Alembert's solution** of the initial value problem.

EXAMPLE K.1. Let $c = 3$, $g(x) = \text{sech}(x)$, and $f(x) = \sum_{i=1}^{4} \zeta(\alpha_i, \beta_i, \gamma_i, x)$, where

$$\zeta(\alpha, \beta, \gamma, x) = \frac{\gamma}{2}\Big[1 + \tanh\left(\alpha x + \beta\right)\Big] \quad \text{and}$$

i	α_i	β_i	γ_i
1	1	4	1
2	1	−4	−1
3	1	4	−0.5
4	1	10	0.5

Let $u_a(x,t) = \frac{1}{2}\left(f(x+ct) + f(x-ct)\right)$ and $u_b(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)ds$. From Figures K.1, we see that the initial distribution given by $f(x)$ is gradually split into two shapes that are both half the height of the original distribution. Both shapes move at constant speed equal to c but travel in the opposite directions. However, for u_b, we see from Figures K.2 that the influence of the initial velocities is propagated within a triangular area determined by speed c. Combining both effects, the solution $u = u_a + u_b$ is shown in Figures K.3.

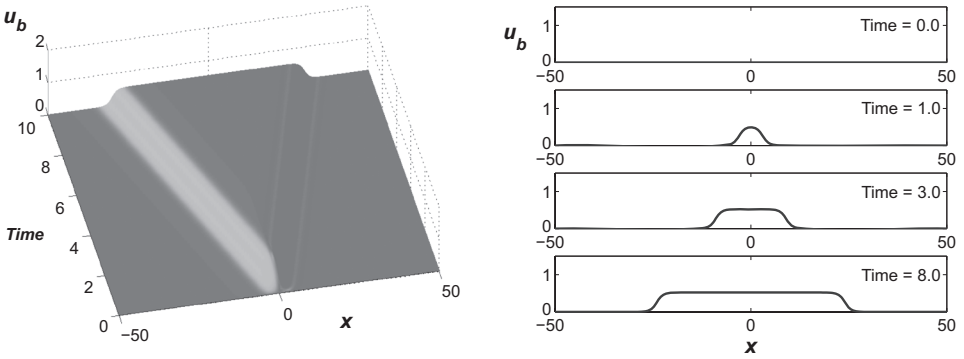


Figure K.2. A surface plot of the trajectories of u_b (left) and a set of four snapshots of u_b at different time instants (right) for the d'Alembert's solution based on zero initial distribution.

K.1.2 Semi-Infinite Domain Wave Equation with Dirichlet Boundary Conditions

The equations are given by

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} &= 0 \quad \text{for} \quad x \geq 0 \\ \left.\begin{aligned} u(x,0) &= f(x) \\ \frac{\partial u}{\partial t}(x,0) &= g(x) \end{aligned}\right\} &\text{ for } x \geq 0 \quad ; \quad u(0,t) = \varsigma(t) \quad t \geq 0 \end{aligned} \tag{K.4}$$

where, for continuity, $\varsigma(0) = f(0)$ and $d\varsigma/dt(0) = g(0)$. We can first find a solution, $v(x,t)$, whose domain is $-\infty < x < \infty$. The desired solution, $u(x,t)$, will be obtained by restricting $v(x,t)$ values at $0 \leq x \leq \infty$, that is,

$$u(x,t) = v(x,t)\big|_{x \geq 0} \tag{K.5}$$

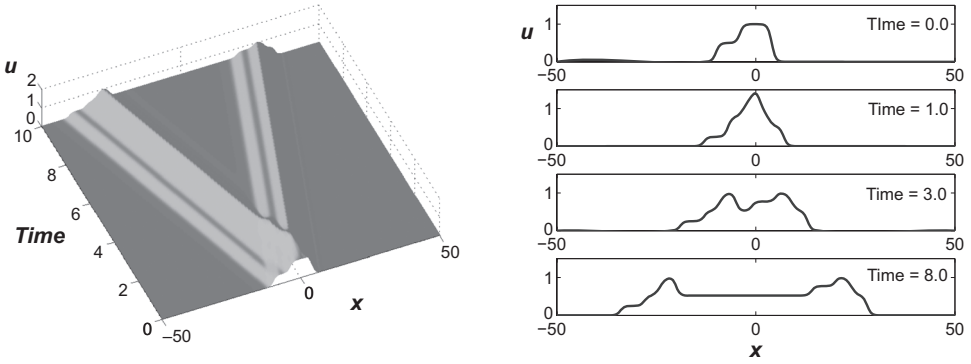


Figure K.3. A surface plot of the trajectories (left) and four snapshots of the distribution at different time instants for $u = u_a + u_b$.

Thus let v be the solution of the extended problem given by

$$\begin{aligned} \frac{\partial^2 v}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2} &= 0 \qquad -\infty \le x \le \infty \\ v(x,0) &= f_e(x) \\ \frac{\partial v}{\partial t}(x,0) &= g_e(x) \end{aligned} \quad ; \quad v(0,t) = \varsigma(t) \qquad t \ge 0$$

where,

$$f_e(x) = f(x) \quad \text{and} \quad g_e(x) = g(x) \qquad \text{for } x \ge 0$$

Note that f_e and g_e have not been defined completely. The solution for $v(x,t)$ is the d'Alembert's solution, given by $v = \phi_e(x+ct) + \psi_e(x-ct)$, where

$$\phi_e(s) = \frac{1}{2}f_e(s) + \frac{1}{2c}\int_0^s g_e(\tau)d\tau \quad \text{and} \quad \psi_e(s) = \frac{1}{2}f_e(s) - \frac{1}{2c}\int_0^s g_e(\tau)d\tau$$

For $x \ge 0$, we have $(x+ct) > 0$ and so $\phi_e(x+ct)$ is immediately given by

$$\phi_e(x+ct) = \frac{1}{2}f(x+ct) + \frac{1}{2c}\int_0^{x+ct} g(\tau)d\tau$$

However, because $(x-ct) < 0$ when $x < ct$, $\psi_e(x-ct)$ has to be handled differently because $f_e\,(s<0)$ and $g_e\,(s<0)$ has not been defined. At $x = 0$, we have

$$v(0,t) = \phi_e(ct) + \psi_e(-ct) = \varsigma(t)$$

or

$$\psi_e(s) = \varsigma\left(-\frac{s}{c}\right) - \phi_e(-s) \quad \rightarrow \quad \psi_e(x-ct) = \varsigma\left(t - \frac{x}{c}\right) - \phi_e(ct-x)$$

Combining the results, and restricting the domain to $x \ge 0$,

$$u(x,t) = \begin{cases} \dfrac{1}{2}\Big[f(ct+x) - f(ct-x)\Big] + \dfrac{1}{2c}\displaystyle\int_{ct-x}^{ct+x} g(\tau)d\tau \\ \qquad\qquad + \varsigma\left(t - \dfrac{x}{c}\right) & \text{for } 0 \le x < ct \\ \dfrac{1}{2}\Big[f(x-ct) + f(x+ct)\Big] + \dfrac{1}{2c}\displaystyle\int_{x-ct}^{x+ct} g(\tau)d\tau & \text{for } x \ge ct \end{cases} \tag{K.6}$$

K.1.3 Semi-Infinite Wave Equation with Nonhomogeneous Neumann Conditions

The equations are given by

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} &= 0 \qquad x \ge 0 \\ u(x,0) &= f(x) \\ \frac{\partial u}{\partial t}(x,0) &= g(x) \end{aligned} \Bigg\} \text{ for } x \ge 0 \quad ; \quad u(0,t) = \varsigma(t) \qquad t \ge 0 \tag{K.7}$$

where, for continuity, $\vartheta(0) = \dfrac{df}{dx}(0)$. Again, we solve the following extended problem but this time with the Neumann boundary condition,

$$\begin{aligned} \frac{\partial^2 v}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2} &= 0 \qquad -\infty \le x \le \infty \\ v(x,0) &= f_e(x) \\ \frac{\partial v}{\partial t}(x,0) &= g_e(x) \qquad ; \qquad \frac{\partial v}{\partial x}(0,t) = \vartheta(t) \end{aligned}$$

with $f_e(x \ge 0) = f(x)$ and $g_e(x \ge 0) = g(x)$. As before, we have $v = \phi_e(x+ct) + \psi_e(x-ct)$, where

$$\phi_e(s) = \frac{1}{2}f_e(s) + \frac{1}{2c}\int_0^s g_e(\tau)d\tau \quad \text{and} \quad \psi_e(s) = \frac{1}{2}f_e(s) - \frac{1}{2c}\int_0^s g_e(\tau)d\tau$$

Because $(x+ct) > 0$,

$$\phi_e(x+ct) = \frac{1}{2}f(x+ct) + \frac{1}{2c}\int_0^{x+ct} g(\tau)d\tau$$

However, for $\psi_e(x-ct)$, we can use the Neumann condition to handle the range $0 \le x < ct$,

$$\frac{\partial v}{\partial x}(0,t) = \vartheta(t) = \phi_e'(ct) + \psi_e'(-ct)$$

from which

$$\psi_e'(s) = \vartheta\left(-\frac{s}{c}\right) - \phi_e'(-s) \rightarrow \begin{aligned} \psi_e(s) &= -c\int_0^{-s/c} \vartheta(\tau)d\tau - \phi_e'(-s) \\ \psi_e(x-ct) &= -c\int_0^{t-(x/c)} \vartheta(\tau)d\tau - \phi_e'(ct-x) \end{aligned}$$

Combining the results, while restricting the solution to $x \ge 0$,

$$u(x,t) = \begin{cases} \dfrac{1}{2}\left[f(ct+x) - f(ct-x)\right] + \dfrac{1}{2c}\displaystyle\int_{ct-x}^{x+ct} g(\tau)d\tau \\ \qquad -c\displaystyle\int_0^{t-(x/c)} \vartheta(\tau)d\tau & \text{for } 0 \le x \le ct \\ \dfrac{1}{2}\left[f(x-ct) + f(x+ct)\right] + \dfrac{1}{2c}\displaystyle\int_{x-ct}^{x+ct} g(\tau)d\tau & \text{for } x \ge ct \end{cases} \tag{K.8}$$

K.1.4 Wave Equation in Finite Domain

We consider only the special homogeneous Dirichlet condition for $0 \le x \le L < \infty$. The equations are given by

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} &= 0 \qquad \text{for} \qquad 0 \le x \le L \\ \left.\begin{aligned} u(x,0) &= f(x) \\ \frac{\partial u}{\partial t}(x,0) &= g(x) \end{aligned}\right\} &\text{ for } 0 \le x \le L \qquad ; \qquad u(0,t) = 0 \qquad t \ge 0 \end{aligned} \tag{K.9}$$

where, for continuity, we need $f(0) = 0 = f(L)$. For this case, we use the method of reflection given by the following extension,

$$\begin{aligned} \frac{\partial^2 v}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 v}{\partial t^2} = 0 \qquad & -\infty \le x \le \infty \\ v(x,0) &= f_e(x) \\ \frac{\partial v}{\partial t}(x,0) &= g_e(x) \qquad v(0,t) = 0 \end{aligned}$$

with f_e and g_e both extended to be odd periodic functions, that is,

$$f_e(x) = \begin{cases} f(x) & \text{for } 0 \le x \le L \\ -f(-x) & \text{for } -L \le x \le 0 \\ f_e(x-2L) & |x| > L \end{cases}$$

The solution can then given by

$$u(x,t) = v(x,t)\Big|_{x\ge 0} \tag{K.10}$$

where

$$v(x,t) = \frac{1}{2}\left(f_e(x+ct) + f_e(x-ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} g_e(\tau)d\tau$$

K.2 Proofs of Lemmas and Theorems in Chapter 11

K.2.1 Proof for Solution of Reducible Linear PDE, Theorem 11.1

First, let $m = 2$. Substituting (11.12), while using the commutativity between L_1 and L_2 given in (11.11),

$$\begin{aligned} Lu &= L_1L_2(\alpha_1u_1 + \alpha_2u_2) = L_1(\alpha_1L_2u_1 + \alpha_2L_2u_2) \\ &= \alpha_1L_1L_2u_1 = \alpha_1L_2(L_1u_1) = 0 \end{aligned}$$

Next, assume the theorem is true for $m = \ell - 1$. Then with $L = L_AL_\ell = L_\ell L_A$ where $L_A = \prod_{i=1}^{\ell-1} L_i$ whose solution is given by $u_A = \sum_{i=1}^{\ell-1}\alpha_iu_i$, and with $u = u_A + \alpha_\ell u_\ell$, we have

$$\begin{aligned} Lu &= L_AL_\ell(u_A + \alpha_\ell u_\ell) = L_A(L_\ell u_A + \alpha_\ell L_\ell u_\ell) \\ &= L_AL_\ell u_A = L_\ell(L_Au_A) = 0 \end{aligned}$$

Then, by induction we have proven that (11.12) is a solution for the case when $L_i \ne L_j$ for $i \ne j, i, j = 1, \ldots, m$.

For the case where L_i is repeated k times, note that

$$\begin{aligned} L_i^k(g_ju_i) &= \sum_{\ell=0}^{k}\frac{k!}{(k-\ell)!\ell!}L_i^\ell g_j\, L_i^{k-\ell}u_i \\ &= u_i\, L_i^k g_j = 0 \end{aligned}$$

Thus

$$L_i^k\left[\left(\sum_{j=1}^{k} g_j\right)u_i\right] = 0$$

K.2.2 Proof of Sturm-Liouville Theorem, Theorem 11.2

We begin with the following identity, where $\phi_n \neq \phi_m$:

$$\frac{d}{dx}\left(p(x)\left[\phi_m\frac{d\phi_n}{dx}-\phi_n\frac{d\phi_m}{dx}\right]\right)=\phi_m\frac{d}{dx}\left[p(x)\frac{d\phi_n}{dx}\right]-\phi_n\frac{d}{dx}\left[p(x)\frac{d\phi_m}{dx}\right]$$

Using (11.51) to substitute for terms on the right-hand side, we get

$$\frac{dz(x)}{dx}=(\lambda_n-\lambda_m)\,r(x)\phi_n\phi_m$$

where,

$$z(x)=p(x)\left[\phi_m\frac{d\phi_n}{dx}-\phi_n\frac{d\phi_m}{dx}\right]$$

Integrating both sides,

$$\frac{z(B)-z(A)}{\lambda_n-\lambda_m}=\int_A^B r(x)\phi_n\phi_m dx$$

Functions ϕ_n and ϕ_m both satisfy the boundary condition at $x = B$, which we could write in matrix form as

$$\mathbf{B}\begin{pmatrix}\beta_B\\ \gamma_B\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix} \tag{K.11}$$

where

$$\mathbf{B}=\begin{pmatrix}\phi_m(B) & d\phi_m/dx(B)\\ \phi_n(B) & d\phi_n/dx(B)\end{pmatrix}$$

Because, in a Sturm-Liouville system, β_B and γ_B are not allowed to both be zero, (K.11) has a solution only if the determinant of matrix **B** is zero. This implies

$$z(B)=p(B)\left[\phi_m(B)\frac{d\phi_n}{dx}(B)-\phi_n(B)\frac{d\phi_m}{dx}\right](B)=p(B)\det(\mathbf{B})=0$$

The same argument follows through with the boundary condition at $x = A$, which implies $z(A) = 0$. Thus we have for $\lambda_m \neq \lambda_n$,

$$\int_A^B r(x)\phi_n(x)\phi_m(x)dx=0 \qquad \text{for } m\neq n$$

K.2.3 Proof of Similarity Transformation Method, Theorem 11.3

Assuming symmetry is admitted based on the similarity transformations $\widetilde{t}=\lambda^{-\alpha}t$, $\widetilde{x}=\lambda^{-\beta}x$ and $\widetilde{u}=\lambda^{-\gamma}u$, we have

$$F\left(\lambda^{\beta}x,\lambda^{\alpha}t,\lambda^{\gamma}u,\ldots,\lambda^{\gamma-\kappa\alpha-(m-\kappa)\beta}\mu^{[\kappa,m-\kappa]}\ldots,\right)=0 \tag{K.12}$$

where

$$\mu^{[\kappa,m-\kappa]}=\frac{\partial^m\widetilde{u}}{(\partial^{\kappa}\widetilde{t})\,(\partial^{m-\kappa}\widetilde{x})}$$

After taking the derivative with respect to λ and then setting $\lambda = 1$, we obtain a quasilinear differential equation given by

$$\beta\widetilde{x}\frac{\partial F}{\partial \widetilde{x}} + \alpha\widetilde{t}\frac{\partial F}{\partial \widetilde{t}} + \gamma\widetilde{u}\frac{\partial F}{\partial \widetilde{u}} + \cdots + \Big(\gamma - \kappa\alpha - (m-\kappa)\beta\Big)\,\mu^{[\kappa,m-\kappa]}\frac{\partial F}{\partial \mu^{[\kappa,m-\kappa]}} + \cdots = 0$$

where the other terms include the partial derivatives of F with respect to the partial derivatives $\partial\widetilde{u}/\partial\widetilde{t}$, $\partial\widetilde{u}/\partial x$, etc. Method of characteristics yields the following equations:

$$\frac{d\widetilde{x}}{\beta\widetilde{x}} = \frac{d\widetilde{t}}{\alpha\widetilde{t}} = \frac{d\widetilde{u}}{\gamma\widetilde{u}} = \cdots \frac{d\mu^{[\kappa,m-\kappa]}}{\Big(\gamma - \kappa\alpha - (m-\kappa)\beta\Big)\,\mu^{[\kappa,m-\kappa]}} = \cdots = \frac{dF}{0}$$

At this point, we assume that $\alpha = 1$ for brevity.[1] Solving the first equations excluding the last term will yield the following invariants

$$\begin{aligned}
\frac{d\widetilde{t}}{\widetilde{t}} = \frac{d\widetilde{x}}{\beta\widetilde{x}} \quad &\rightarrow \quad \zeta = \frac{\widetilde{x}}{\widetilde{t}^{\beta}} \\
\frac{d\widetilde{t}}{\widetilde{t}} = \frac{d\widetilde{u}}{\gamma\widetilde{u}} \quad &\rightarrow \quad \psi = \frac{\widetilde{u}}{\widetilde{t}^{\gamma}} \\
&\vdots \\
\frac{d\widetilde{t}}{\widetilde{t}} = \frac{d\mu^{[\kappa,m-\kappa]}}{\Big(\gamma - \kappa - (m-\kappa)\beta\Big)\,\mu^{[\kappa,m-\kappa]}} \quad &\rightarrow \quad \phi_{\kappa,m} = \frac{\mu^{[\kappa,m-\kappa]}}{\widetilde{t}^{(\gamma-\kappa-(m-\kappa)\beta)}} \\
&\vdots
\end{aligned}$$

plus F, which is another invariant. We also can now use x, t, and u instead of $\widetilde{x}$, $\widetilde{t}$, and $\widetilde{u}$ because the invariants also satisfy the symmetry conditions. The general solution of the quasilinear equation can now be given by

$$F = g\,(\zeta, \psi, \ldots, \phi_{\kappa,m}, \ldots) = 0$$

For the invariants with $\kappa = 0$, that is, the partial derivatives with respect to x only, we have

$$\mu^{[0,m]} = \frac{\partial^m u}{\partial x^m} = \left(t^{\gamma - m\beta}\right)\frac{d^m\psi}{d\zeta^m} \quad \rightarrow \quad \phi_{0,m} = \frac{d^m\psi}{d\zeta^m}$$

With

$$\mu^{[\kappa,m-\kappa]} = \frac{\partial^\kappa}{\partial t^\kappa}\left(\mu^{[0,m-\kappa]}\right)$$

one can show by induction that

$$\mu^{[\kappa,m-\kappa]} = t^{\gamma-\kappa-(m-\kappa)\beta}\sum_{j=0}^{\kappa} c_j\;\zeta^j\;\frac{d^{m-\kappa+j}\psi}{d\zeta^{m-\kappa+j}} \quad \rightarrow \quad \phi_{\kappa,m} = \sum_{j=0}^{\kappa} c_j\;\zeta^j\;\frac{d^{m-\kappa+j}\psi}{d\zeta^{m-\kappa+j}}$$

[1] If $\alpha = 0$, then we could set $\beta = 1$ and proceed with the role of t replaced by x.

where c_j are simply constants that depend on j, m, κ, β, and γ whose complicated forms are not needed for the purpose of this proof. Thus we conclude that because the invariants $\phi_{\kappa,m}$ are just functions $h_{\kappa,m}$ of ζ, ψ and derivatives of $\psi(\zeta)$, we have shown that

$$g\left(\zeta, \psi, \ldots, \phi_{\kappa,m}, \ldots\right) = g\left(\zeta, \psi, \ldots, h_{\kappa,m}\left(\zeta, \psi, \frac{d\psi}{d\zeta}, \ldots\right), \ldots\right) = 0$$

is a nonlinear ordinary differential equation for $\psi(\zeta)$.

APPENDIX L

Additional Details and Fortification for Chapter 12

L.1 The Fast Fourier Transform

In this appendix, we obtain matrix representations of the discrete Fourier transforms, which is often used to find the Fourier series through the use of numerical integration methods.

For a periodic function $g(t)$ with period T, we have the complex form of the Fourier series defined by

$$g_{\mathrm{FS}}(t) = \sum_{k=-\infty}^{\infty} C_k \exp\left(\frac{2\pi i k t}{T}\right) \tag{L.1}$$

where $i = \sqrt{-1}$. The Fourier coefficients C_ℓ can be evaluated by first setting $g_{\mathrm{FS}}(t) = g(t)$ and then multiplying (L.1) by $\exp(-2\pi i \ell / T)$, followed by an integration with respect to t from 0 to T,

$$\int_0^T g(t) \exp\left(-\frac{2\pi i \ell t}{T}\right) dt = \sum_{k=-\infty}^{\infty} C_k \int_0^T \exp\left(\frac{2\pi i (k-\ell) t}{T}\right) dt$$

Because

$$e^{2m\pi i} = \cos(2m\pi) + i \sin(2m\pi) = 1 \qquad \text{with } m \text{ an integer}$$

we have

$$\int_0^T \exp\left(\frac{2\pi i (k-\ell) t}{T}\right) dt = \frac{T}{2\pi i (k-\ell)} \left(e^{2\pi i (k-\ell)} - 1\right) = \begin{cases} T & \text{if } k = \ell \\ 0 & \text{if } k \neq \ell \end{cases}$$

Thus

$$C_\ell = \frac{1}{T} \int_0^T g(t) \exp\left(-\frac{2\pi i \ell}{T} t\right) dt \tag{L.2}$$

Now suppose the function $g(t)$, $t \in [0, T]$, is represented by $(N+1)$ uniformly distributed points, that is, $g_0, \ldots, g_N$, with $g_k = g(k\Delta t)$, $\Delta t = t_{k+1} - t_k$, and $T = N\Delta t$. Using the trapezoidal approximation of the integral in (L.2), we have the discretized version given by

$$C_\ell = \frac{1}{N\Delta t} \left(\frac{g_0 + g_N}{2} \Delta t + \sum_{k=1}^{N-1} g_k \exp\left(-\frac{2\pi \ell k}{N} i\right) \Delta t \right)$$

Now let $y_\ell = NC_\ell$ and

$$x_k = \begin{cases} \dfrac{g_0 + g_N}{2} & \text{for } k = 1 \\ g_{k-1} & \text{for } k = 2, \ldots, N \end{cases} \tag{L.3}$$

then we obtain

$$y_\ell = \sum_{k=1}^{N} x_k \, W_{[N]}^{(k-1)(\ell-1)} \qquad k = \ell, \ldots, N \tag{L.4}$$

where $W_{[N]} = e^{(-2\pi/N)i}$. Equation (L.4) is known as the **discrete Fourier transform** of vector $\mathbf{x} = (x_1, \ldots, x_N)^T$. For the determination of y_ℓ, $\ell = 1, \ldots, N$, a matrix representation of (L.4) is given by

$$\mathbf{y} = F_{[N]}\mathbf{x} \tag{L.5}$$

where

$$F_{[N]} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & W_{[N]} & \cdots & W_{[N]}^{N-1} \\ 1 & W_{[N]}^2 & \cdots & W_{[N]}^{2(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & W_{[N]}^{N-1} & \cdots & W_{[N]}^{(N-1)(N-1)} \end{pmatrix} \tag{L.6}$$

For the special case of $N = 2^m$ for some integer $m \geq 1$, we can obtain the classic algorithm known as the **Radix-2 Fast Fourier Transform**, or often simply called **Fast Fourier Transform FFT**. The FFT algorithm significantly reduces the number of operations in the evaluation of (L.5).

First, note from (L.6) that $F_{[1]} = 1$. For $N = 2^m, m \geq 1$, we can separate the odd and even indices and use the fact that $W_{[N]}^2 = W_{[N/2]}$ to obtain a rearrangement of (L.4) as follows:

$$\begin{aligned} y_\ell &= \sum_{k=1}^{N/2} \left(x_{2k-1} W_{[N]}^{(2k-2)(\ell-1)} \right) + \left(x_{2k} W_{[N]}^{(2k-1)(\ell-1)} \right) \\ &= \sum_{k=1}^{N/2} x_{2k-1} \left(W_{[N]}^2 \right)^{(k-1)(\ell-1)} + \sum_{k=1}^{N/2} x_{2k} \left(W_{[N]}^2 \right)^{(k-1/2)(\ell-1)} \\ &= \sum_{k=1}^{N/2} x_{2k-1} W_{[N/2]}^{(k-1)(\ell-1)} + W_{[N]}^{\ell-1} \sum_{k=1}^{N/2} x_{2k} W_{[N/2]}^{(k-1)(\ell-1)} \end{aligned} \tag{L.7}$$

Equation (L.7) is known as the **Danielson-Lanczos equation**. Let $\mathbf{y} = \left(\begin{array}{c|c} \mathbf{y}_A^T & \mathbf{y}_B^T \end{array} \right)^T$ where $\mathbf{y}_A = (y_1, \ldots, y_{N/2})^T$ and $\mathbf{y}_B = \left(y_{(N/2)+1}, \ldots, y_N\right)^T$. Because $W_{[N]}^N = 1$ and $W_{[N]}^{N/2} = -1$, for $\ell = 1, \ldots, N/2$,

$$\begin{aligned} \mathbf{y}_A &= F_{[N/2]} P_{[N]}^{\text{odd}} \mathbf{x} + \Omega_{[N/2]} F_{[N/2]} P_{[N]}^{\text{even}} \mathbf{x} \\ \mathbf{y}_B &= F_{[N/2]} P_{[N]}^{\text{odd}} \mathbf{x} - \Omega_{[N/2]} F_{[N/2]} P_{[N]}^{\text{even}} \mathbf{x} \end{aligned}$$

where

$$P_{[N]}^{\text{odd}} = \left(\begin{array}{c|c|c|c} \mathbf{e}_1 & \mathbf{e}_3 & \ldots & \mathbf{e}_{N-1} \end{array} \right)^T \quad ; \quad P_{[N]}^{\text{even}} = \left(\begin{array}{c|c|c|c} \mathbf{e}_2 & \mathbf{e}_4 & \ldots & \mathbf{e}_N \end{array} \right)^T$$

$$\Omega_{[N/2]} = \begin{pmatrix} 1 & & & 0 \\ & W_{[N]} & & \\ & & \ddots & \\ 0 & & & W_{[N]}^{(N/2)-1} \end{pmatrix}$$

Comparing with (L.5), we have

$$F_{[N]} = \left(\begin{array}{c|c} F_{[N/2]} & \Omega_{[N/2]} F_{[N/2]} \\ \hline F_{[N/2]} & -\Omega_{[N/2]} F_{[N/2]} \end{array} \right) P_{[N]}^{(\text{o|e})} = Z_{[N]} \left(I_2 \otimes F_{[N/2]} \right) P_{[N]}^{(\text{o|e})} \tag{L.8}$$

where

$$P_{[N]}^{(\text{o|e})} = \left(\begin{array}{c} P_{[N]}^{\text{odd}} \\ \hline P_{[N]}^{\text{even}} \end{array} \right) \quad \text{and} \quad Z_{[N]} = \left(\begin{array}{c|c} I_{N/2} & \Omega_{[N/2]} \\ \hline I_{N/2} & -\Omega_{[N/2]} \end{array} \right)$$

Using the identities $AC \otimes BD = (A \otimes B)(C \otimes D)$ and $A \otimes (B \otimes C) = (A \otimes B) \otimes C$, we have

$$\begin{aligned} I_2 \otimes F_{[N/2]} &= I_2 \otimes \left(Z_{[N/2]} (I_2 \otimes F_{[N/4]}) \right) P_{[N/2]}^{(\text{o|e})} \\ &= \left(I_2 \otimes \left(Z_{[N/2]} \left(I_2 \otimes F_{[N/4]} \right) \right) \right) \left(I_2 \otimes P_{[N/2]}^{(\text{o|e})} \right) \\ &= \left(I_2 \otimes Z_{[N/2]} \right) \left(I_4 \otimes F_{[N/4]} \right) \left(I_2 \otimes P_{[N/2]}^{(\text{o|e})} \right) \end{aligned}$$

Continuing the recursion we obtain, with $F_{[1]} = 1$, $N = 2^m$,

$$F_{[N]} = G_{[N]} P_{[N]}^{\text{bitreverse}} \tag{L.9}$$

where,

$$\begin{aligned} G_{[N]} &= Z_{[N]} \left(I_2 \otimes Z_{[N/2]} \right) \left(I_4 \otimes Z_{[N/4]} \right) \cdots \left(I_{N/2} \otimes Z_{[2]} \right) \\ P_{[N]}^{\text{bitreverse}} &= \left(I_{N/2} \otimes P_{[2]}^{(\text{o|e})} \right) \cdots \left(I_4 \otimes P_{[N/4]}^{(\text{o|e})} \right) \left(I_2 \otimes P_{[N/2]}^{(\text{o|e})} \right) P_{[N]}^{(\text{o|e})} \end{aligned}$$

It can be shown that the effect of $P_{[N]}^{\text{bitreverse}}$ on $\mathbf{x}$ is to rearrange the elements of $\mathbf{x}$ by reversing the bits of the binary number equivalent of the indices. To illustrate, let $N = 8$, then

$$P_{[8]}^{\text{bitreverse}} \mathbf{x} = \left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \left(\begin{array}{cccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \mathbf{x} = \begin{pmatrix} x_1 \\ x_5 \\ x_3 \\ x_7 \\ x_2 \\ x_6 \\ x_4 \\ x_8 \end{pmatrix}$$

Instead of building the permutations, we could look at the bit reversal of the binary equivalents of the indices of $\mathbf{x}$ (beginning with index 0),

$$\begin{pmatrix} 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{pmatrix} \xrightarrow{\text{reverse bits}} \begin{pmatrix} 000 \\ 100 \\ 010 \\ 110 \\ 001 \\ 101 \\ 011 \\ 111 \end{pmatrix} \xrightarrow{\text{decimal}} \begin{pmatrix} 0 \\ 4 \\ 2 \\ 6 \\ 1 \\ 5 \\ 3 \\ 7 \end{pmatrix} \xrightarrow{\text{add 1}} \begin{pmatrix} 1 \\ 5 \\ 3 \\ 7 \\ 2 \\ 6 \\ 4 \\ 8 \end{pmatrix}$$

In summary, we have the following algorithm:

FFT Algorithm:
Given: $\mathbf{x}[=]2^m \times 1$
$\mathbf{y} \leftarrow$ Reverse Bits($\mathbf{x}$)
For $r = 1, \ldots, m$
$\mathbf{y} \leftarrow \left(I_{2^{m-r}} \otimes Z_{[2^r]}\right)\mathbf{y}$
End

Remark: The MATLAB command for the FFT function is `y=fft(x)`.

EXAMPLE L.1. Let

$$g(t) = \begin{cases} 5 & \text{if } t < 20 \\ \dfrac{t}{4}\cos\left(\dfrac{2\pi}{20}t\right) & \text{if } 20 \le t < 80 \\ 28 - \dfrac{t}{10} & \text{if } t \ge 20 \end{cases}$$

Now apply the Fourier series to approximate $g(t)$ for $0 \le t \le 200$ with $T = 200$ and sampling $N = 2^{10} + 1$ uniformly distributed data points for $g(t)$.

Using $\mathbf{x}$ as defined in (L.3) and $\mathbf{y} = FFT(\mathbf{x})$, we can obtain a finite series approximation given by

$$g_{FFT,L} = \sum_{k=-L}^{L} C_k W^{-kt} = \frac{1}{N}\left(y_1 + 2\sum_{k=1}^{L} \text{Real}\left(y_{k+1}W^{-kt}\right)\right)$$

Note that only the first $N/2 = 2^{m-1}$ terms of $\mathbf{y}$ are useful for the purpose of approximation, i.e. $L \le N/2$.

Figure L.1 shows the quality of approximation for $L = 10$ and $L = 25$.

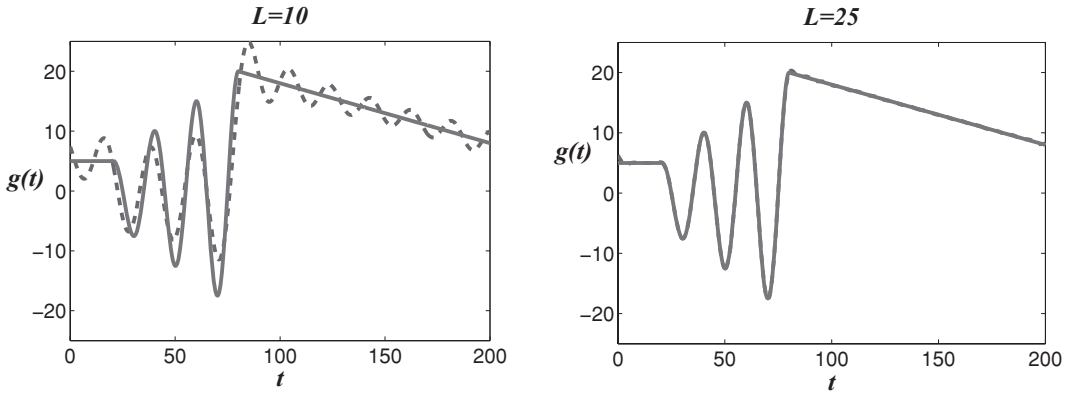


Figure L.1. Fourier series approximation of $g(t)$ using $L = 10$ and $L = 25$.

L.2 Integration of Complex Functions

In this appendix, we briefly describe the notations, definitions, and results from complex function theory. Specifically, we focus on the methods for contour integrations of complex functions.

L.2.1 Analytic Functions and Singular Points

Definition L.1. *Let $z = z_{\text{re}} + iz_{\text{im}}$ be a complex variable, with $z_{\text{re}}, z_{\text{im}} \in \mathbb{R}$. Then a complex function $f(z) = f_{\text{re}}(z_{\text{re}}, z_{\text{im}}) + i\, f_{\text{im}}(z_{\text{re}}, z_{\text{im}})$ is* **analytic** *(or* **holomorphic***) in a domain D, that is, a connected open set, if for every circle centered at $z = z^*$ inside D, $f(z)$ can be represented by a Taylor series expanded around $z = z^*$,*

$$f(z) = \sum_{k=0}^{\infty} \alpha_k (z - z^*)^k \tag{L.10}$$

where,

$$\alpha_k = \frac{1}{k!} \left. \frac{d^k f}{dz^k} \right|_{z=z^*} \tag{L.11}$$

Implicit in the preceding definition is the existence of derivatives, $d^k f/dz^k$, for $k \geq 1$. One necessary and sufficient condition for analyticity of $f(z)$ is given by the following theorem:

THEOREM L.1. *A complex function $f(z) = f_{\text{re}}(z_{\text{re}}, z_{\text{im}}) + i\, f_{\text{im}}(z_{\text{re}}, z_{\text{im}})$ is analytic in D if and only if both real functions $f_{\text{re}}(z_{\text{re}}, z_{\text{im}})$ and $f_{\text{im}}(z_{\text{re}}, z_{\text{im}})$ are continuously differentiable and*

$$\frac{\partial f_{\text{re}}}{\partial z_{\text{re}}} = \frac{\partial f_{\text{im}}}{\partial z_{\text{im}}} \tag{L.12}$$

$$\frac{\partial f_{\text{re}}}{\partial z_{\text{im}}} = -\frac{\partial f_{\text{im}}}{\partial z_{\text{re}}} \tag{L.13}$$

for all $z = z_{\text{re}} + iz_{\text{im}}$ in D.

The pair of equations (L.12) and (L.13) are known as the **Cauchy-Riemann conditions**.

Some Important Properties of Analytic Functions:

Let $f(z)$, $f_1(z)$ and $f_2(z)$ be analytic in the same domain D, then

1. Linear combinations of analytic functions are analytic; that is, $f_{\text{sum}}(z) = \alpha_1 f_1(z) + \alpha_2 f_2(z)$ is analytic.
2. Products of analytic functions are analytic; that is, $f_{\text{prod}}(z) = f_1(z) f_2(z)$ is analytic.
3. Division of analytic functions are analytic except at the zeros of the denominator; that is, $f_{\text{div}}(z) = f_1(z)/f_2(z)$ is analytic except at the zeros of $f_2(z)$.
4. Composition of analytic functions are analytic; that is, $f_{\text{comp}}(z) = f_1(f_2(z))$ is analytic.
5. The inverse function, $f^{-1}(f(z)) = z$, is analytic if $df/dz \neq 0$ in D and $f(z_1) \neq f(z_2)$ when $z_1 \neq z_2$.
6. The chain rule is given by

$$\frac{d}{dz}[f_2(f_1(z))] = \frac{df_2}{df_1}\frac{df_1}{dz} \tag{L.14}$$

Definition L.2. *A point z_o in domain D is called a* **singularity** *or* **singular point** *of a complex function $f(z)$ if it is not analytic at $z = z_o$. If $f(z)$ is analytic at $z = z_o$, then z_o is called a* **regular point**.

The singular points can further be classified as follows:

1. *A point z_o is a* **removable singular point** *if $f(z)$ can be made analytic by defining it at z_o.*

   (If $\lim_{z \to z_o} f(z)$ is bounded, it can be included in the definition of $f(z)$. Then $f(z)$ can be expanded as a Taylor series around z_o. For example, with $f(z) = (z-1)(3z/(z-1))$, the point $z = 1$ is a removable singularity.)
2. *A point z_o is an* **isolated singular point** *if for some $\rho > 0$, $f(z)$ is analytic for $0 < |z - z_o| < \rho$ but not analytic at $z = z_0$.*
3. *A point z_o is a* **pole of order** *k, where k is a positive integer, if $\left(g_1(z) = (z - z_o)^k f(z)\right)$ has a removable singularity at $z = z_o$, but $\left(g_2(z) = (z - z_o)^{k-1} f(z)\right)$ does not have a removable singularity at $z = z_o$. If $k = 1$, then we call it a* **simple pole**.
4. *A point z_o is an* **essential singular point** *if it is an isolated singularity that is not a pole or removable.*

L.2.2 Contour Integration of Complex Functions

In calculating the closed contour integration of $f(z)$, denoted by

$$I_C(f) = \oint_C f(z)dz \tag{L.15}$$

we are assuming that C is a simple-closed curve; that is, C is a curve that begins and ends at the same point without intersecting itself midway. Furthermore, the line

integral will be calculated by traversing the curve C in the counterclockwise manner (or equivalently, the interior of the simple-closed curve is to the left of C during the path of integration). The interior of curve C defines a domain $D(C)$, which is of the type **simply connected**, as defined next.

Definition L.3. *A 2D domain D is called a* **simply-connected domain** *if the interior points of every simple-closed curve C in D are also in D. Otherwise, the domain is called a* **multiply connected domain**.

In short, a simply connected domain is one that does not contain any holes.[1]

Because of the presence of several theorems that follow later, we start with a brief outline of the development of techniques for contour integration in the complex plane.

1. We start with Cauchy's theorem to handle the special case when $f(z)$ is analytic on and inside the closed curve C.
2. In Theorem L.3, we show that even though C is the original contour, a smaller curve C' inside C can yield the same contour integral values, as long as $f(z)$ remains analytic on C, C' and the annular region between C and C'.
3. Having established that the contour used for integration is not unique, we shift the focus instead on specific points and construct small circular contours around these points. This leads to the definition of **residues**. Theorem L.4 then gives a formula to calculate the residues of poles.
4. Using residues, we can then generalize Cauchy's theorem, Theorem L.2, to handle cases when curve C encloses n isolated singularities. The result is the **residue theorem**.

THEOREM L.2. **Cauchy's Theorem.** *Let $f(z)$ be analytic on and inside a simple closed curve C, then*

$$\oint_C f(z)dz = 0 \tag{L.16}$$

PROOF. With z traversing along the curve C,

$$dz = \left(\frac{dz_{\text{re}}}{ds} + i\frac{dz_{\text{im}}}{ds}\right) ds$$

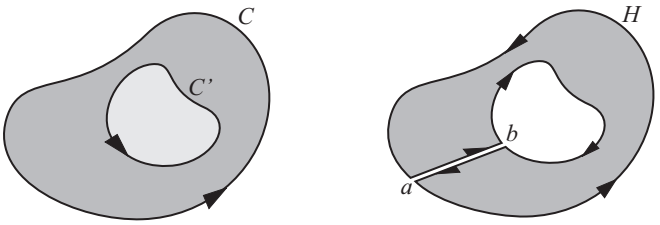
or in terms of the components of the unit outward normal vector, $\mathbf{n} = n_{\text{re}} + i\, n_{\text{im}}$,

$$dz = (-n_{\text{im}} + i\, n_{\text{re}})\, ds$$

because

$$\frac{dz_{\text{re}}}{ds} = -n_{\text{im}} \qquad ; \qquad \frac{dz_{\text{im}}}{ds} = n_{\text{re}}$$

[1] For higher dimensional regions, if any simple closed path in D can be shrunk to a point, then D is simply connected.



a) curves C and C' *b) curve H*

Figure L.2. The curves C, C', and H used for proof of Theorem L.3.

Thus

$$\begin{aligned}\oint_C f(z)dz &= \oint (f_{\text{re}} + if_{\text{im}})\,(-n_{\text{im}} + i\,n_{\text{re}})\,ds \\ &= -\oint_C (f_{\text{im}}n_{\text{re}} + f_{\text{re}}n_{\text{im}})\,ds + i\oint_C (f_{\text{re}}n_{\text{re}} - f_{\text{im}}n_{\text{im}})\,ds\end{aligned}$$

Using the divergence theorem,

$$\oint_C f(z)dz = \int\int \left(\frac{\partial f_{\text{im}}}{\partial z_{\text{re}}} + \frac{\partial f_{\text{re}}}{\partial z_{\text{im}}}\right) dz_{\text{re}}dz_{\text{im}} + i\int\int \left(\frac{\partial f_{\text{re}}}{\partial z_{\text{re}}} - \frac{\partial f_{\text{im}}}{\partial z_{\text{im}}}\right) dz_{\text{re}}dz_{\text{im}}$$

Because analytic functions satisfy the Cauchy-Riemann conditions, the integrands are both zero.

THEOREM L.3. *Let C and C' be two simple closed curves where C' is strictly inside C. Let $f(z)$ be analytic on curves C and C' and in the annular region between C and C'. Then*

$$\oint_C f(z)dz = \oint_{C'} f(z)dz \tag{L.17}$$

PROOF. Based on Figure L.2, we see that the integral based on curve H is given by

$$\oint_H f(z)dz = \oint_C f(z)dz + \int_a^b f(z)dz + \int_b^a f(z)dz - \oint_{C'} f(z)dz$$

However, the path integral from a to b satisfies

$$\int_a^b f(z)dz = -\int_b^a f(z)dz$$

Furthermore, because $f(z)$ is analytic in the interior of H (i.e., the annular region between C and C'), Theorem L.2 implies that $\oint_H f(z)dz = 0$. Thus

$$\oint_C f(z)dz = \oint_{C'} f(z)dz$$

Theorem L.3 does not constrain how the shrinking of curve C to C' occurs except for the conditions given in the theorem. For instance, if $f(z)$ is analytic

throughout the interior of C, then the smaller curve C' can be located anywhere inside C.

We now shift our focus on point z_o and the contours surrounding it.

Definition L.4. *For a given point z_o and function $f(z)$, let $\overline{C}$ be a simple closed curve that encircles z_o such that z_o is the only possible singularity of $f(z)$ inside $\overline{C}$; then the* **residue of** $f(z)$ **at** z_o *is defined as*

$$\mathbf{Res}_{z_o}(f) = \frac{1}{2\pi i}\oint_{\overline{C}} f(z)dz \tag{L.18}$$

Note that if $f(z)$ is analytic at the point z_o, $\mathbf{Res}_{z_o} = 0$. If z_o is a singular point of $f(z)$, the residue at z_o will be nonzero.[2] Using Theorem L.3, we can evaluate residues at the poles of $f(z)$ by choosing $\overline{C}$ to be a small circle centered around z_o.

THEOREM L.4. **Cauchy Integral Representation.**[3] *Let z_o be a pole of order $k \geq 1$ of $f(z)$, then*

$$\mathbf{Res}_{z_o}(f) = \frac{1}{(k-1)!}\lim_{z\to z_o}\frac{d^{k-1}}{dz^{k-1}}\left([z-z_o]^k f(z)\right) \tag{L.19}$$

PROOF. First, consider the function $h(z) = (z-z_o)^\ell$, where ℓ is an integer. Let O_ρ : $|z-z_o| = \rho$, where $\rho > 0$. The points on the circle O_ρ is given by

$$z = z_o + \rho e^{i\theta} \qquad \text{for } 0 \leq \theta \leq 2\pi$$

and

$$(z-z_o)^\ell = \rho^\ell e^{i\ell\,\theta} \qquad ; \qquad dz = i\rho e^{i\theta}d\theta$$

Thus

$$\oint_{O_\rho}(z-z_o)^\ell dz = i\rho^{\ell+1}\int_0^{2\pi} e^{i(\ell+1)\theta}d\theta = \begin{cases} 2\pi i & \text{if } \ell = -1 \\ 0 & \text{if } \ell \neq -1 \end{cases} \tag{L.20}$$

where $\rho > 0$ is bounded.

[2] A result known as Morera's theorem guarantees that if $\mathbf{Res}_{z_o}(f) = 0$, then $f(z)$ is analytic at a small neighborhood around z_o.

[3] Note that Theorems L.2 and L.4 are both associated with Cauchy's name, but the two theorems are not the same. Strictly speaking, Cauchy's integral representation actually refers only to the case of a simple pole, that is, $k = 1$.

Because z_o is a pole of order k of $f(z)$, there exists a curve $\overline{C}$ such that the function

$$g(z) = (z - z_o)^k f(z)$$

is analytic inside and on a curve $\overline{C}$, which includes z_o as an interior point; that is, it could be expanded into a Taylor series around z_o,

$$\begin{aligned} (z - z_o)^k f(z) &= g(z) = \sum_{n=0}^{\infty} \alpha_n (z - z_o)^n \\ f(z) &= \sum_{n=0}^{\infty} \alpha_n (z - z_o)^{n-k} \end{aligned} \tag{L.21}$$

where

$$\alpha_n = \lim_{z \to z_o} \frac{1}{n!} \frac{d^n g}{dz^n} = \lim_{z \to z_o} \frac{1}{n!} \frac{d^n}{dz^n} \left([z - z_o]^k f(z) \right) \tag{L.22}$$

Based on the definition of the residue, choose the curve $\overline{C}$ to be a small circle O_ρ centered at z_o such that $f(z)$ is analytic on and inside the circle O_ρ except at z_o. This means that the radius of the circle, ρ, must be chosen to be small enough such that, inside O_ρ, z_o is the only singular point of $f(z)$. Taking the contour integral of (L.21), with substitutions of (L.20) and (L.22),

$$\begin{aligned} \oint_{O_\rho} f(z) dz &= \sum_{n=0}^{\infty} \alpha_n \oint_{O_\rho} (z - z_o)^{n-k} dz \\ &= 2\pi i \, \alpha_{k-1} \\ &= \frac{2\pi i}{(k-1)!} \lim_{z \to z_o} \frac{d^{k-1}}{dz^{k-1}} \left([z - z_o]^k f(z) \right) \end{aligned}$$

Thus

$$\mathbf{Res}_{z_o}(f) = \frac{1}{(k-1)!} \lim_{z \to z_o} \frac{d^{k-1}}{dz^{k-1}} \left([z - z_o]^k f(z) \right)$$

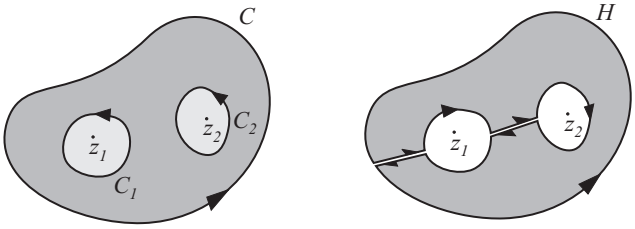
We now state a generalization of Theorem L.3. This theorem is very useful for the evaluation of contour integrals in the complex plane.

THEOREM L.5. **Residue Theorem.** *Let $f(z)$ be analytic on and inside the closed curve C except for isolated singularities: z_ℓ, $\ell = 1, 2, \ldots, n$. Then the contour integral of $f(z)$ along C is given by*

$$\oint_C f(z) dz = 2\pi i \sum_{\ell=1}^{n} \mathbf{Res}_{z_\ell}(f) \tag{L.23}$$

PROOF. We prove the theorem only for $n = 2$, but the same arguments can be generalized easily for $n > 2$.

Let C_1 and C_2 be nonintersecting closed curves inside C such that the pole z_1 is inside C_1 only and the pole z_2 is inside C_2 only. As shown in Figure L.3, $f(z)$ will be analytic in the curve H as well as in the interior points of H.



a) curves C, C_1 and C_2 *b) curve H*

Figure L.3. Curves C, C_1, C_2, and H used in Theorem L.5.

Thus

$$\oint_H f(z)dz = 0 = \oint_C f(z)dz - \oint_{C_1} f(z)dz - \oint_{C_2} f(z)dz$$

or

$$\oint_C f(z)dz = \oint_{C_1} f(z)dz + \oint_{C_2} f(z)dz$$

Using the results of Theorem L.4,

$$\oint_C f(z)dz = 2\pi i \left[\mathbf{Res}_{z_1}(z) + \mathbf{Res}_{z_2}(z) \right]$$

Generalizing the approach to $n \geq 1$,

$$\oint_C f(z)dz = 2\pi i \sum_{\ell=1}^{n} \mathbf{Res}_{z_\ell}(z)$$

Note that Theorem L.5 is true whether the isolated singularities are poles or essential singularities. However, we limit our applications only to singularities involving poles. As such, the formula for calculating residues when singularities are poles (cf. Theorem L.4) is used when invoking the method of residues.

L.2.3 Path Integrals with Infinite Limits

The method of residues can be applied to calculating path integrals in the complex plane,

$$\int_P f(z)dz \tag{L.24}$$

where path P is a curve parameterized by $a \leq t \leq b$, that is,

$$P : z = z_{\text{re}}(t) + iz_{\text{im}}(t) \tag{L.25}$$

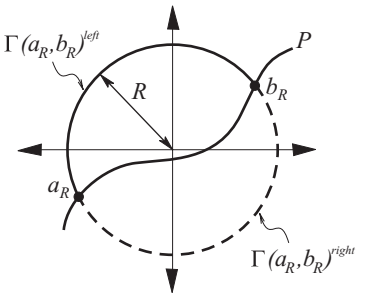


Figure L.4. The circular arcs $\Gamma^{\text{left}}_{a_R,b_R}$ and $\Gamma^{\text{right}}_{a_R,b_R}$.

including the case where

$$|z(t=a)| = \infty \qquad \text{and} \qquad |z(t=b)| = \infty \tag{L.26}$$

We refer to these paths as **infinite paths**.

Some Technical Issues:

1. **Parameterization.** Path P will be parameterized using $a \le t \le b$. We assume that P does not intersect itself and that the path is bounded, with possible exceptions at the end points.

2. **Connecting Arcs.** Let a_R and b_R, with $a \le a_R < b_R \le b$, be values of t such that $|P(a_R)| = |P(b_R)| = R$. Then one can connect both $P(a_R)$ and $P(b_R)$ by a circular arc of radius R. (*We assume the arc does not to intersect P except at $t = a_R$ and $t = b_R$*). We denote the arc by $\Gamma^{\text{left}}_{a_R,b_R}$ if the arc starting from a_R is to the left of path P. Likewise, we denote the arc by $\Gamma^{\text{right}}_{a_R,b_R}$ if the arc starting from a_R is to the right of path P (see Figure L.4).

   The main idea is to combine either the left or right arc with the subpath, $P(a_R, b_R)$, to obtain a simple closed curve from which we can apply the method of residues.

3. **Convergence Assumptions.** In handling the path integration along the left circular arcs, we assume the following condition:

$$\lim_{R\to\infty} R \max_{z\in\Gamma(a_R,b_R)^{\text{left}}} |f(z)| = 0 \tag{L.27}$$

   We refer to (L.27) as the **convergence condition in the left arc**. Together with the following inequality (also known as **Darboux's inequality**),

$$\left|\int_{\Gamma(a_R,b_R)^{\text{left}}} f(z)dz\right| \le \int_{\Gamma(a_R,b_R)^{\text{left}}} \big|f(z)\big|\,|dz| < 2\pi R \max_{z\in\Gamma(a_R,b_R)^{\text{left}}} |f(z)| \tag{L.28}$$

   we obtain

$$\lim_{R\to\infty} \left|\int_{\Gamma(a_R,b_R)^{\text{left}}} f(z)dz\right| = 0 \tag{L.29}$$

   Similarly, we assume the **convergence condition in the right arc** given by

$$\lim_{R\to\infty} R \max_{z\in\Gamma(a_R,b_R)^{\text{right}}} |f(z)| = 0 \tag{L.30}$$

and obtain

$$\lim_{R\to\infty}\left|\int_{\Gamma(a_R,b_R)^{\text{right}}} f(z)dz\right| = 0 \tag{L.31}$$

4. **Cauchy Principal Value.** With finite limits, the following identity is true:

$$\int_{P(a_R,b_r)} f(z)dz = \int_{P(a_R,0)} f(z)dz + \int_{P(0,b_r)} f(z)dz$$

However, the integral $\int_{P(a_R,0)} f(z)dz$ or the integral $\int_{P(0,b_R)} f(z)dz$, or both, may diverge as $a_R, b_R \to \infty$, even though the integral

$$PV(f) = \lim_{a_R,b_R\to\infty}\int_{P(a_R,b_R)} f(z)dz \tag{L.32}$$

converges. In our calculations of $\int_P fdz$ that follow, we mean the limit calculation of (L.32). The integral in (L.32) is known as the **Cauchy principal value** of $f(z)$.

We now state a theorem that shows how the method of residues can be applied to complex integrations along infinite paths.

THEOREM L.6. *Let $P(t)$ be an infinite path that does not pass through any singular points of $f(z)$.*

1. *Let $z_1, z_2, \ldots, z_n$ be the singularities in the region to the left of path $P(t)$, and $f(z)$ satisfies the absolute convergence in the left arc condition given in (L.27), then*

$$\int_P f(z)dz = \sum_{\ell=1}^{n} \mathbf{Res}_{z_\ell}(f) \tag{L.33}$$

2. *Let $\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_m$ be the singularities in the region to the right of path $P(t)$, and $f(z)$ satisfies the absolute convergence in the right arc condition given in (L.30), then*

$$\int_P f(z)dz = -\sum_{\ell=1}^{m} \mathbf{Res}_{\hat{z}_\ell}(f) \tag{L.34}$$

PROOF. Based on Figure L.5, where R is chosen large enough such that the contour formed by the subpath $P(a_R, b_R)$ and -$\Gamma(a_R, b_R)^{\text{left}}$ will contain all the singular points of $f(z)$ that are to the left of P. Then using the theorem of residues,

$$\int_{P(a_R,b_R)} f(z)dz - \int_{\Gamma(a_R,b_R)^{\text{left}}} f(z)dz = \sum_{\ell=1}^{n} \mathbf{Res}_{z_\ell}(f)$$

As $R \to \infty$, (L.29) then implies

$$\int_P f(z)dz = \sum_{\ell=1}^{n} \mathbf{Res}_{z_\ell}(f)$$

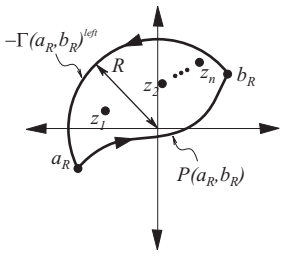


Figure L.5. The contour used to prove (L.33) in Theorem L.6.

Likewise, based on Figure L.6, where R is chosen large enough such that the contour formed by the subpath $-P(a_R, b_R)$ and $\Gamma(a_R, b_R)^{\text{right}}$ will contain all the singular points of $f(z)$ that are to the right of P. Then using the theorem of residues,

$$-\int_{P(a_R,b_R)} f(z)dz + \int_{\Gamma(a_R,b_R)^{\text{right}}} f(z)dz = \sum_{\ell=1}^{m} \mathbf{Res}_{\hat{z}_\ell}(f)$$

As $R \to \infty$, (L.31) then implies

$$\int_P f(z)dz = -\sum_{\ell=1}^{n} \mathbf{Res}_{z_\ell}(f)$$

Note that the convergence conditions, (L.27) and (L.30), are sufficient conditions that may sometimes be too conservative. In some cases, they could be relaxed. In particular, we have the result known as **Jordan's lemma**, which is useful when calculating Fourier transforms and Fourier-Sine/Fourier-Cosine transforms.

THEOREM L.7. *Let $f(z) = g(z)e^{i\omega z}$, where $\omega > 0$, with $\Gamma(a_R, b_R)^{\text{left}}$ and $\Gamma(a_R, b_R)^{\text{right}}$ as the semicircle in the upper half and lower half, respectively, of the complex plane,*

1. *If*

$$\lim_{R\to\infty} \left(\max_{z\in\Gamma(a_R,b_R)^{\text{left}}} |g(z)| \right) = 0 \tag{L.35}$$

then

$$\lim_{R\to\infty} \left| \int_{\Gamma(a_R,b_R)^{\text{left}}} f(z)dz \right| = 0 \tag{L.36}$$

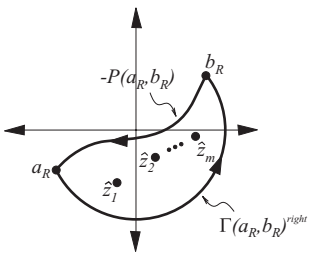


Figure L.6. The contour used to prove (L.34) in Theorem L.6.

2. *If*

$$\lim_{R\to\infty}\left(\max_{z\in\Gamma(a_R,b_R)^{\text{right}}}|g(z)|\right)=0 \tag{L.37}$$

then

$$\lim_{R\to\infty}\left|\int_{\Gamma(a_R,b_R)^{\text{right}}}f(z)dz\right|=0 \tag{L.38}$$

PROOF. We show the theorem only for the left arc, that is, upper half of the complex plane,

On the semicircle, we have $z=Re^{i\theta}$. Thus

$$dz=Re^{i\theta}d\theta \quad\longrightarrow\quad |dz|=R|d\theta|$$

and

$$\begin{aligned} e^{i\omega z} &= e^{i\omega R(\cos\theta+i\sin\theta)} \\ &= e^{-\omega R\sin\theta}e^{i\omega R\cos\theta} \quad\longrightarrow\quad \left|e^{i\omega z}\right|=e^{-\omega R\sin\theta}\end{aligned}$$

Also, note that with $0\le\theta\le\frac{\pi}{2}$,

$$\sin\theta\ge\frac{2\theta}{\pi}$$

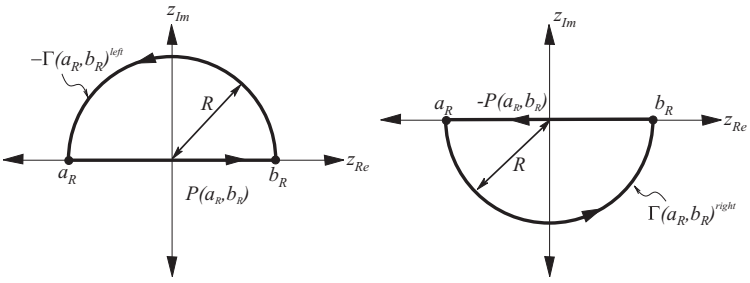
Using these identities and inequality,

$$\begin{aligned}\left|\int_{\Gamma(a_R,b_R)^{\text{left}}}f(z)dz\right| &\le \int_{\Gamma(a_R,b_R)^{\text{left}}}\left|g(z)\right|\left|e^{iwz}\right||dz| \\ &\le \left(\max_{z\in\Gamma(a_R,b_R)^{\text{left}}}|g(z)|\right)\left(\int_{\Gamma(a_R,b_R)^{\text{left}}}\left|e^{iwz}\right||dz|\right) \\ &< \left(\max_{z\in\Gamma(a_R,b_R)^{\text{left}}}|g(z)|\right)\left(2R\int_0^{\pi/2}e^{-\omega R\sin\theta}d\theta\right) \\ &< \left(\max_{z\in\Gamma(a_R,b_R)^{\text{left}}}|g(z)|\right)\left(2R\int_0^{\pi/2}e^{-2\omega R\theta/\pi}d\theta\right) \\ &< \left(\max_{z\in\Gamma(a_R,b_R)^{\text{left}}}|g(z)|\right)\left(\frac{\pi}{\omega}\left[1-e^{-\omega R}\right]\right) \\ &< \left(\max_{z\in\Gamma(a_R,b_R)^{\text{left}}}|g(z)|\right)\left(\frac{\pi}{\omega}\right)\end{aligned}$$

Using condition (L.35), we have

$$\lim_{R\to\infty}\left|\int_{\Gamma(a_R,b_R)^{\text{left}}}f(z)dz\right|=0$$

Theorem L.7 assumed $\omega>0$ and $\omega\theta>0$. For $\omega<0$, we need to traverse the path in the opposite directions; that is, we need to replace θ by $-\theta$.



a) $\omega < 0$ *b)* $\omega > 0$

Figure L.7. The contours used for evaluating a Fourier integral.

EXAMPLE L.2. Consider the Fourier integral,

$$\mathcal{F}\left[\frac{x^3}{1+x^4}\right] = \int_{-\infty}^{\infty} \frac{x^3}{1+x^4} e^{-i\omega x} dx \tag{L.39}$$

Here, the path is $P = t$, with $-\infty \leq t \leq \infty$, that is, the real line. The poles of $g(x) = x^3/(1+x^4)$ are: $(-1 \pm i)/\sqrt{2}$, $(1 \pm i)/\sqrt{2}$.

With the situation of $\omega < 0$, we can use the closed-contour in the upper complex plane, that is, $z_{\text{im}} \geq 0$, see Figure (L.7).

Because

$$\lim_{R\to\infty} \left(\max_{|z|=R, z_{\text{im}}>0} \left| \frac{z^3}{1+z^4} \right| \right) = 0$$

we could use the residue theorem and Theorem L.7 to compute the integral

$$\int_{-\infty}^{\infty} \frac{x^3}{1+x^4} e^{-i\omega x} dx = 2\pi i \left(\mathbf{Res}_{[(1+i)/\sqrt{2}]}(f) + \mathbf{Res}_{[(-1+i)/\sqrt{2}]}(f) \right) \tag{L.40}$$

where

$$f = \frac{x^3}{1+x^4} e^{-i\omega x}$$

For $\omega < 0$,

$$\begin{aligned}
\mathbf{Res}_{[(1+i)/\sqrt{2}]}(f) &= \lim_{z\to(1+i)/\sqrt{2}} \left[\left(z - \frac{1+i}{\sqrt{2}} \right) f(z) \right] = \frac{1}{4} e^{(1-i)\omega/\sqrt{2}} \\
\mathbf{Res}_{[(-1+i)/\sqrt{2}]}(f) &= \lim_{z\to(-1+i)/\sqrt{2}} \left[\left(z - \frac{-1+i}{\sqrt{2}} \right) f(z) \right] = \frac{1}{4} e^{(1+i)\omega/\sqrt{2}} \\
\int_{-\infty}^{\infty} \frac{x^3}{1+x^4} e^{-i\omega x} dx &= \pi i \left[\cos\left(\frac{\omega}{\sqrt{2}} \right) e^{\omega/\sqrt{2}} \right]
\end{aligned}$$

For $\omega > 0$, we can use the closed-contour in the lower region of the complex plane. Doing so, we have

$$\int_{-\infty}^{\infty} \frac{x^3}{1+x^4} e^{-i\omega x} dx = -2\pi i \left(\mathbf{Res}_{[(1-i)/\sqrt{2}]}(f) + \mathbf{Res}_{[(-1-i)/\sqrt{2}]}(f) \right) \tag{L.41}$$

and

$$\mathbf{Res}_{[(1-i)/\sqrt{2}]}(f) = \lim_{z\to(1-i)/\sqrt{2}}\left[\left(z-\frac{1-i}{\sqrt{2}}\right)f(z)\right] = \frac{1}{4}e^{(-1-i)\omega/\sqrt{2}}$$

$$\mathbf{Res}_{[(-1-i)/\sqrt{2}]}(f) = \lim_{z\to(-1-i)/\sqrt{2}}\left[\left(z-\frac{-1-i}{\sqrt{2}}\right)f(z)\right] = \frac{1}{4}e^{(-1+i)\omega/\sqrt{2}}$$

$$\int_{-\infty}^{\infty}\frac{x^3}{1+x^4}e^{i\omega x}dx = -\pi i\left[\cos\left(\frac{\omega}{\sqrt{2}}\right)e^{-\omega/\sqrt{2}}\right]$$

Combining both cases,

$$\mathcal{F}\left[\frac{x^3}{1+x^4}\right] = \int_{-\infty}^{\infty}\frac{x^3}{1+x^4}e^{i\omega x}dx = -i\left[\text{sgn}(\omega)\right]\pi\cos\left(\frac{\omega}{\sqrt{2}}\right)e^{[-|\omega|/\sqrt{2}]}$$

Special Applications and Extensions:

1. **Functions Involving Sines and Cosines.** Let $P(t) = t$, with $-\infty \le t \le \infty$. When the integrand contains $\cos(x)$ or $\sin(x)$ in the numerator, the method of residues cannot be used directly because the arc conditions given in (L.27) or (L.30) are no longer satisfied. (For instance, $\lim_{z_{\text{im}}\to\pm\infty}|\cos(z)| = \lim_{z_{\text{im}}\to\pm\infty}|\sin(z)| = \infty$).

   An alternative approach is to use Jordan's lemma.
   Because

$$g(x)\cos(\alpha x) = \text{Re}\left[g(x)e^{ix}\right] \tag{L.42}$$

   we could apply the method of residues on the integral in the right hand side of the following equation:

$$\int_{-\infty}^{\infty}g(x)\cos(\alpha x)dx = \text{Re}\left[\int_{-\infty}^{\infty}g(x)e^{i\alpha x}dx\right] \tag{L.43}$$

   Similarly, we have

$$\int_{-\infty}^{\infty}g(x)\sin(\alpha x)dx = \text{Im}\left[\int_{-\infty}^{\infty}g(x)e^{i\alpha x}dx\right] \tag{L.44}$$

   Based on Jordan's lemma, that is, Theorem L.7, with $\omega = \alpha > 0$, we need to satisfy only the condition given in (L.35) and apply it to the contour in the upper region of the complex plane,

$$\lim_{R\to\infty}\left(\max_{|z|=R,\, z_{\text{im}}\ge 0}|g(z)|\right) = 0 \tag{L.45}$$

EXAMPLE L.3. Consider the following integral

$$\int_{-\infty}^{\infty}\frac{x^2\cos x}{1+x^4}dx$$

Using a semicircle in the upper region of the complex plane as the contour of integration, we apply Theorem L.7 to obtain

$$\lim_{R\to\infty}\int_{-R}^{R} f(z_{\text{re}})dz_{\text{re}} = 2\pi i\left(\mathbf{Res}_{[1+i]}(f) + \mathbf{Res}_{[-1+i]}(f)\right)$$

where,

$$f(z) = \frac{z^2 e^{iz}}{1+z^4}$$

with

$$\begin{aligned}\mathbf{Res}_{[1+i]/\sqrt{2}}(f) &= \frac{\sqrt{2}(1-i)}{8}e^{(-1+i)/\sqrt{2}}\\ \mathbf{Res}_{[-1+i]\sqrt{2}}(f) &= \frac{-\sqrt{2}(1+i)}{8}e^{(-1-i)/\sqrt{2}}\end{aligned}$$

Then,

$$\begin{aligned}\int_{-\infty}^{\infty}\frac{x^2\cos x}{1+x^4}dx &= \text{Re}\left[\int_{-\infty}^{\infty}\frac{x^2e^{ix}}{1+x^4}dx\right]\\ &= \frac{\pi}{\sqrt{2}}\left[\cos\left(\frac{1}{\sqrt{2}}\right)-\sin\left(\frac{1}{\sqrt{2}}\right)\right]e^{(-1/\sqrt{2})}\end{aligned}$$

2. **Rectangular Contours.** Sometimes the limits involve a line that is shifted parallel to the real axis or the imaginary axis. In these cases, it may often be convenient to use evaluations already determined for the real line or imaginary axis. To do so, we need a rectangular contour. This is best illustrated by an example.

EXAMPLE L.4. Let us evaluate the Fourier transform of a Gaussian function,

$$\mathcal{F}\left[e^{-\alpha x^2}\right] = \int_{-\infty}^{\infty} e^{-\alpha x^2}\, e^{-i\omega x}dx = \int_{-\infty}^{\infty} e^{-\alpha x^2 - i\omega x}dx$$

where $\alpha > 0$.

First, consider $\omega > 0$. We could simplify the integral by first completing the squares,

$$\begin{aligned}-\alpha x^2 - i\omega x &= -\alpha\left(x^2 + \frac{i\omega}{\alpha}x + \left[\left(\frac{i\omega}{2\alpha}\right)^2 - \left(\frac{i\omega}{2\alpha}\right)^2\right]\right)\\ &= -\alpha\left(x+\frac{i\omega}{2\alpha}\right)^2 - \frac{\omega^2}{4\alpha}\end{aligned}$$

thus

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-\alpha x^2 - i\omega x}dx &= e^{-\omega^2/(4\alpha)}\int_{-\infty}^{\infty} e^{-\alpha[x+i\omega/(2\alpha)]^2}dx\\ &= e^{-\omega^2/(4\alpha)}\int_{-\infty+i\omega/(2\alpha)}^{\infty+i\omega/(2\alpha)} e^{-\alpha z^2}dz\end{aligned}$$

Now consider the rectangular contour shown in Figure L.8.

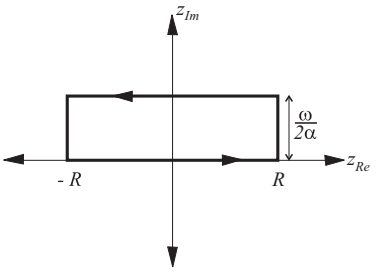


Figure L.8. A rectangular contour used in Example L.4.

Because the function $e^{-\alpha z^2}$ is analytic throughout the region,

$$\int_{-R}^{R} e^{-\alpha z^2} dz + \int_{R}^{R+i\omega/(2\alpha)} e^{-\alpha z^2} dz + \int_{R+i\omega/(2\alpha)}^{-R+i\omega/(2\alpha)} e^{-\alpha z^2} dz + \int_{-R+i\omega/(2\alpha)}^{-R} e^{-\alpha z^2} dz = 0$$

Two of the integrals reduces to zero,

$$\lim_{R\to\infty} \int_{R}^{R+i\omega/(2\alpha)} e^{-\alpha z^2} dz = 0$$

and

$$\lim_{R\to\infty} \int_{-R+i\omega/(2\alpha)}^{-R} e^{-\alpha z^2} dz = 0$$

resulting with

$$\int_{-\infty+i\omega/(2\alpha)}^{\infty+i\omega/(2\alpha)} e^{-\alpha z^2} dz = \int_{-\infty}^{\infty} e^{-\alpha z^2} dz = \sqrt{\frac{\pi}{\alpha}}$$

Using a rectangular contour in the lower region, a similar approach can be used to handle $\omega < 0$. Combining all the results, we obtain

$$\mathcal{F}\left[e^{-\alpha^2 x}\right] = \sqrt{\frac{\pi}{\alpha}}\, e^{-\omega^2/(4\alpha)}$$

This says that the Fourier transform of a Gaussian function is another Gaussian function.

3. **Path P Contains a Finite Number of Simple Poles.** When the path of integration contains simple poles, the path is often modified to avoid the poles using a semicircular indentation having a small radius, ϵ as shown in Figure L.9. Assuming convergence, the calculation for the integral proceeds by taking the limit as $\epsilon \to 0$.

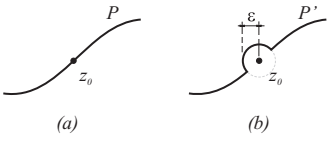


Figure L.9. (a) Pole z_o lies on path P. (b) Path P' avoids z_o.

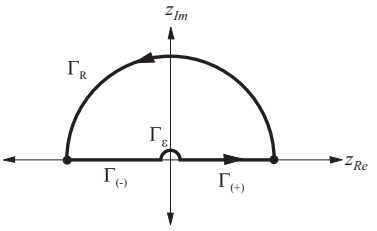


Figure L.10. The contour used to solve $\int_{-\infty}^{\infty}[\sin(x)/x]dx$.

EXAMPLE L.5. Let us determine the integral

$$\int_0^\infty \frac{\sin(x)}{x}dx \tag{L.46}$$

First, we evaluate the integrals with limits from $-\infty$ to ∞. Using the techniques for solving integrals with sinusoids as given in (L.44),

$$\int_{-\infty}^{\infty} \frac{\sin(x)}{x}dx = \text{Im}\left[\int_{-\infty}^{\infty} \frac{e^{ix}}{x}dx\right]$$

Using the path in the real line, $z = 0$ is a pole in the real line. Thus, modifying the path to avoid the origin, we obtain the closed contour shown in Figure L.10 given as $C = \Gamma_{(-)} + \Gamma_\epsilon + \Gamma_{(+)} + \Gamma_R$.

The integral along Γ_ϵ can be evaluated by setting $z = \epsilon e^{i\theta}$. As a consequence,

$$\frac{dz}{z} = id\theta$$

and

$$\int_{\Gamma_\epsilon} \frac{e^{iz}}{z}dz = \int_\pi^0 \exp\left(i\epsilon e^{i\theta}\right)id\theta$$

and taking the limit as $\epsilon \to 0$,

$$\lim_{\epsilon\to 0}\int_{\Gamma_\epsilon} \frac{e^{iz}}{z}dz = -i\pi$$

Conversely, we have

$$\lim_{R\to\infty}\int_{\Gamma_R} \frac{e^{iz}}{z}dz = 0$$

Thus

$$\int_{-\infty}^{\infty} \frac{e^{ix}}{x}dx = i\pi$$

or

$$\int_{-\infty}^{\infty} \frac{\sin(x)}{x}dx = \text{Im}\left[i\pi\right] = \pi$$

Because the function $\sin(x)/x$ is an even function, we could just divide the value by 2 to obtain the integral with limits from 0 to ∞, that is,

$$\int_0^\infty \frac{\sin(x)}{x}dx = \frac{\pi}{2} \tag{L.47}$$

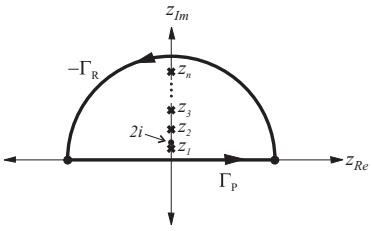


Figure L.11. The contour used to solve $\int_{-\infty}^{\infty}[(x^2+4)\cosh(x)]^{-1}dx$.

4. **Regions Containing Infinite Number of Poles.** In case there is an infinite number of poles in the region inside the contour, we simply extend the summation of the residues to contain all the poles in that region. If the infinite sum of residues converge, then the method of residues will still be valid, that is,

$$\oint_C f(z)dz = \sum_{\ell=1}^{\infty} \mathbf{Res}_{z_\ell}(f) \tag{L.48}$$

EXAMPLE L.6. Let us evaluate the following integral:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{1}{(x^2+4)\cosh(x)}dx \tag{L.49}$$

From the roots of (z^2+4) and the roots of $\cosh(z) = \cos(iz)$, the singularities are all simple poles given by:

$$z_0 = 2i, \quad z_\ell = \frac{2\ell-1}{2}\pi i, \quad \ell = 1, 2, \ldots, \infty$$

and their complex conjugates.

Using the usual semicircular contour to cover the upper region of the complex plane as shown in Figure L.11, the method of residues yields,

$$\lim_{R\to\infty}\left[\int_{\Gamma_P} f(z)dz - \int_{\Gamma_R} f(z)dz\right] = 2\pi i\left(\mathbf{Res}_{(2i)}[f] + \sum_{\ell=1}^{\infty}\mathbf{Res}_{(z_\ell)}[f]\right) \tag{L.50}$$

Along the path of Γ_R, we have $z = Re^{i\theta}$. We find that

$$\lim_{R\to\infty}\left|\frac{1}{(R^2e^{i2\theta}+4)\cosh\left(Re^{i\theta}\right)}\right| < \lim_{R\to\infty} R^{-2}\left|\exp\left(-Re^{i\theta}\right)\right| = 0$$

Thus we have $\lim_{R\to\infty}\int_{\Gamma_R} f(z)dz = 0$.

As for the residues,

$$\mathbf{Res}_{(2i)}[f] = \lim_{z\to 2i}\frac{1}{(z+2i)\cosh(z)} = \frac{1}{4i\ \cos(2)}$$

and with $z_\ell = i(2\ell-1)\pi/2$, together with the application of L'Hospital's rule,

$$\begin{aligned}\mathbf{Res}_{(z_\ell)}[f] &= \lim_{z\to z_\ell}\frac{z-z_\ell}{(z^2+4)\cosh(z)} \\ &= \frac{1}{z_\ell^2+4}\ \frac{-1}{i\sin(iz_\ell)} \\ &= (-1)^\ell\ \frac{4}{i\left(4^2-(2\ell-1)^2\pi^2\right)}\end{aligned}$$

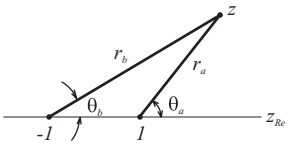


Figure L.12. Geometric interpretation of r_a, r_b, θ_a, and θ_b given in (L.53).

Combining all these results, we have

$$\int_{-\infty}^{\infty} \frac{1}{(x^2+4)\cosh(x)} dx = \frac{\pi}{2\cos(2)} + 8\pi \sum_{\ell=1}^{\infty} (-1)^{\ell} \frac{1}{4^2 - (2\ell-1)^2 \pi^2} \tag{L.51}$$

5. Integrals along Branch Cuts. When the integrand involves multivalued complex functions, a branch cut is necessary to evaluate the integral. This means that a Riemann sheet[4] has to be specified by selecting the range of the arguments of complex variable z. Usually, the ranges for the argument are either $0 < \arg(z) < 2\pi$, $-\pi < \arg(z) < \pi$, $(\pi/2) < \arg(z) < (5\pi/2)$ or $-\pi/2 < \arg(z) < (3\pi/2)$ for branch cuts along the positive real line, negative real line, positive imaginary line, or negative real line, respectively. In other cases, the range of $\arg(z)$ may be a finite segment in the complex plane.

   Once the particular Riemann sheet has been selected, the method of residues can proceed as before.

EXAMPLE L.7. Consider the integral

$$\int_{-1}^{1} \frac{dx}{(x^2+1)\sqrt{1-x^2}} \tag{L.52}$$

This is a finite integral in which the integrand contains a square root in the denominator. One can check that the points $z = 1$ and $z = -1$ are **branch points**[5] of $f(z)$ where

$$f(z) = \frac{1}{(z^2+1)\sqrt{z^2-1}}$$

(Note that we used $z^2 - 1$. The form $1 - x^2$ will show up from the calculations later.)

We could be rewrite the square root terms as

$$\begin{aligned} \sqrt{z^2-1} &= \sqrt{(z-1)(z+1)} \\ &= \sqrt{(r_a e^{i\theta_a})(r_b e^{i\theta_b})} \\ &= \sqrt{r_a r_b}\, e^{i(\theta_a+\theta_b)/2} \end{aligned}$$

where,

$$z - 1 = r_a e^{i\theta_a} \quad \text{and} \quad z + 1 = r_b e^{i\theta_b} \tag{L.53}$$

(see Figure L.12.)

[4] By Riemann sheet, we simply mean a subdomain that is single-valued.

[5] A point z_o is **branch point** of a function $f(z)$ if there exists a closed curve that encircles z_o that would yield different evaluations of $f(z)$ after one encirclement.

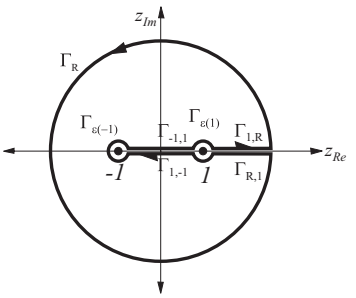


Figure L.13. Contour used for solving the integral in Example L.7.

We can then specify the branch cut by fixing the ranges on θ_a and θ_b to be

$$0 < \theta_a < 2\pi \quad \text{and} \quad 0 < \theta_b < 2\pi$$

Aside from being branch points, the points $z = \pm 1$ are also singular points.

We can then choose the contour shown in Figure L.13 and implement the method of residues. The closed-contour C is given by

$$\begin{aligned} C &= \Gamma_R + \Gamma_{R,1} + \Gamma_{\epsilon(1)^{\text{lower}}} + \Gamma_{1,-1} + \Gamma_{\epsilon(-1)} + \Gamma_{-1,1} + \Gamma_{\epsilon(1)^{\text{upper}}} + \Gamma_{1,R} \\ &= \left(\Gamma_R + \Gamma_{\epsilon(1)} + \Gamma_{\epsilon(-1)}\right) + (\Gamma_{R,1} + \Gamma_{1,R}) + (\Gamma_{1,-1} + \Gamma_{-1,1}) \end{aligned}$$

Following earlier methods, we can evaluate the integrals along the three circular paths: the outer circle Γ_R and the pair of inner circles $\Gamma_{\epsilon(1)}$ $\Gamma_{\epsilon(-1)}$, to yield zero values as the limits of $R \to \infty$ and $\epsilon \to 0$ are approached, respectively. Thus we need to evaluate only the four remaining straight paths. Because $f(z)$ is multivalued, the path along a common segment, but in opposite directions, may not necessarily cancel. We now show that the integrals along $\Gamma_{1,R}$ and $\Gamma_{R,1}$ will cancel, whereas the integrals along $\Gamma_{1,-1}$ and $\Gamma_{-1,1}$ will not.

Along the path $\Gamma_{R,1}$, we have $z_{\text{im}} = 0$, $1 < z_{\text{re}} \leq R$, $\theta_a = 2\pi$ and $\theta_b = 2\pi$, thus

$$f(z)\big|_{\Gamma_{R,1}} = \frac{1}{(1+x^2)\sqrt{r_a r_b}\, e^{2\pi i}} = \frac{1}{(1+x^2)\sqrt{x^2-1}}$$

Similarly, along path $\Gamma_{1,R}$, we have $z_{\text{im}} = 0$, $1 < z_{\text{re}} < R$, $\theta_a = 0$ and $\theta_b = 0$,

$$f(z)\big|_{\Gamma_{1,R}} = \frac{1}{(1+x^2)\sqrt{r_a r_b}} = \frac{1}{(1+x^2)\sqrt{x^2-1}}$$

The sum of integrals along both $\Gamma_{1,R}$ and $\Gamma_{R,1}$ is then given by

$$\begin{aligned} \int_{\Gamma_{1,R}} f(z)dz + \int_{\Gamma_{R,1}} f(z)dz &= \int_1^R \frac{1}{(1+x^2)\sqrt{x^2-1}}dx \\ &\quad + \int_R^1 \frac{1}{(1+x^2)\sqrt{x^2-1}}dx \\ &= 0 \end{aligned}$$

Along the path $\Gamma_{1,-1}$, we have $z_{\text{im}} = 0$, $-1 < z_{\text{re}} \leq 1$, $\theta_a = \pi$ and $\theta_b = 2\pi$, thus

$$f(z)\big|_{\Gamma_{1,-1}} = \frac{1}{(1+x^2)\sqrt{r_a r_b}\, e^{3\pi i/2}} = \frac{-1}{(1+x^2)i\sqrt{1-x^2}}$$

Similarly, along path $\Gamma_{-1,1}$, we have $z_{\text{im}} = 0$, $-1 < z_{\text{re}} < 1$, $\theta_a = \pi$ and $\theta_b = 0$,

$$f(z)\big|_{\Gamma_{-1,1}} = \frac{1}{(1+x^2)\sqrt{r_a r_b}\, e^{\pi i/2}} = \frac{1}{(1+x^2)i\sqrt{1-x^2}}$$

Note that we used $r_a r_b = (1-x^2)$ because $|x| < 1$.

Thus the sum of integrals along both $\Gamma_{1,-1}$ and $\Gamma_{-1,1}$ is given by

$$\begin{aligned}\int_{\Gamma_{1,-1}} f(z)dz + \int_{\Gamma_{-1,1}} f(z)dz &= \int_1^{-1} \frac{-1}{(1+x^2)i\sqrt{1-x^2}}dx \\ &\quad + \int_{-1}^{1} \frac{1}{(1+x^2)i\sqrt{1-x^2}}dx \\ &= \frac{2}{i}\int_{-1}^{1} \frac{1}{(1+x^2)\sqrt{1-x^2}}dx\end{aligned}$$

Next, we need to calculate the residues at the poles $z = \pm i$. Note that because the function is multivalued, we need to be careful when taking the limits of the square root. First, consider the pole $z = -i$. At this point, we have

$$\begin{aligned} z - 1 = -i - 1 &= \sqrt{2}\, e^{5\pi i/4} \\ z + 1 = -1 + i &= \sqrt{2}\, e^{7\pi i/4}\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{Res}_{-i}[f] &= \lim_{z\to -i}\left(\frac{z+i}{(1+z^2)\sqrt{(z-1)(z+1)}}\right) \\ &= \left(\frac{1}{-2i}\right)\left(\frac{1}{\sqrt{2}\, e^{3\pi i/2}}\right) \\ &= \frac{-1}{2\sqrt{2}}\end{aligned}$$

For the other pole, $z = i$,

$$\begin{aligned} z - 1 = i - 1 &= \sqrt{2}\, e^{3\pi i/4} \\ z + 1 = i + 1 &= \sqrt{2}\, e^{\pi i/4}\end{aligned}$$

and

$$\begin{aligned}\mathbf{Res}_{i}[f] &= \lim_{z\to i}\left(\frac{z-i}{(1+z^2)\sqrt{(z-1)(z+1)}}\right) \\ &= \left(\frac{1}{2i}\right)\left(\frac{1}{\sqrt{2}\, e^{\pi i/2}}\right) \\ &= \frac{-1}{2\sqrt{2}}\end{aligned}$$

Finally, we combine all the previous calculations to obtain

$$\begin{aligned}\int_C f(z)dz &= 2\pi i\left(\mathbf{Res}_{-i}[f]+\mathbf{Res}_i[f]\right)\\ \frac{2}{i}\int_{-1}^{1}\frac{1}{(1+x^2)\sqrt{x^2-1}}dx &= 2\pi i\left(\frac{-1}{\sqrt{2}}\right)\\ \int_{-1}^{1}\frac{1}{(1+x^2)\sqrt{x^2-1}}dx &= \frac{\pi}{\sqrt{2}}\end{aligned}$$

L.3 Dirichlet Conditions and the Fourier Integral Theorem

Definition L.5. *A function $f(x)$ is said to satisfy the* **Dirichlet conditions** *in the interval (a, b), if the interval (a, b) can be partitioned into a finite number of subintervals such that $f(x)$ is bounded and monotonic in each of these subintervals. This means:*

1. *There are a finite number of maxima and minima for $f(x)$ in (a, b).*
2. *$f(x)$ has no infinite discontinuities, but it can have a finite number of bounded discontinuities.*

Then we have the following theorem, known as the **Fourier integral theorem.**

THEOREM L.8. *Let $f(x)$ be such that $\int_{-\infty}^{\infty}|f(x)|dx<\infty$, and let $f(x)$ satisfy Dirichlet's conditions given in definition L.5 for $(a, b) = (-\infty, \infty)$, then*

$$\frac{1}{2}\left[f(x^+)+f(x^-)\right]=\frac{1}{\pi}\int_0^{\infty}\int_{-\infty}^{\infty}f(t)\cos\left(\omega(x-t)\right)\,dt\,d\omega \tag{L.54}$$

where

$$f(x^+)=\lim_{\eta\to 0}f\left(x+|\eta|\right)\qquad \textit{and}\qquad f(x^-)=\lim_{\eta\to 0}f\left(x-|\eta|\right)$$

PROOF. As opposed to the prior approach of taking limits on the Fourier series (cf. (12.5)), equation (L.54) given in Theorem L.8 can be more correctly derived from another important result known as Dirichlet's integral theorem,

$$\frac{1}{2}\left[f(x^+)+f(x^-)\right]=\lim_{\theta\to\infty}\frac{1}{\pi}\int_{-\infty}^{\infty}f(x+\eta)\frac{\sin(\theta\eta)}{\eta}d\eta \tag{L.55}$$

as long as $f(x)$ satisfy Dirichlet's conditions. The proof of (L.55) is given in section L.6.1 (page 836).

Let $t = x + \eta$. Also, we use the fact that

$$\frac{\sin(\theta\eta)}{\eta}=\int_0^{\theta}\cos(\eta\omega)\,d\omega \tag{L.56}$$

Substituting (L.56) into (L.55) with x held fixed, we get

$$\frac{1}{2}\left[f(x^+)+f(x^-)\right]=\lim_{\theta\to\infty}\frac{1}{\pi}\int_{-\infty}^{\infty}f(t)\int_0^{\theta}\cos\left((x-t)\omega\right)\,d\omega\,dt \tag{L.57}$$

The last important detail deals with the validity of interchanging the sequence of integration in (L.57). With the assumption in Theorem L.8 that $(\int_{-\infty}^{\infty} |f(t)|dt < \infty)$, we can show that (see Section L.6.2),

$$\lim_{\theta\to\infty} \int_{-\infty}^{\infty} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt = \lim_{\theta\to\infty} \int_0^{\theta} \int_{-\infty}^{\infty} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \tag{L.58}$$

So with (L.58) substituted to (L.57), we obtain the Fourier integral equation given in (L.54)

L.4 Brief Introduction to Distribution Theory and Delta Distributions

In this appendix, we introduce some of the basic theory and tools to generalize the concept of functions, with special attention to the construction of delta distributions. We also include a brief discussion of a very important class of distributions, called **tempered distributions**, that generalizes the theory of Fourier transforms for functions that may not be absolutely integrable.

L.4.1 The Delta Distribution (Delta Function)

The **delta distribution**, denoted by $\delta(t)$ and often known as the **delta function**, is an important operation in applied mathematics. However, it does not satisfy the classical requirements of functions; for example, it is not defined at $t = 0$. Instead, a new concept known as **distributions** (also known as **generalized functions**) had to be constructed to give the necessary mathematical rigor to $\delta(t)$. Once the theory for distribution was built, the constructs allow for the definition of other distributions, including the derivatives of $\delta(t)$ and $\delta(g(t))$, where $g(t)$ is a continuous function.

Consider the Heaviside step function, $\mathcal{H}(t)$, defined as

$$\mathcal{H}(t) = \begin{cases} 0 & \text{if} \quad t < 0 \\ 1 & \text{if} \quad t \geq 0 \end{cases} \tag{L.59}$$

The delta distribution is often defined as the "derivative" of the Heaviside step function. Unfortunately, because of the discontinuity at $t = 0$, the derivative is not defined there. However, the integral

$$\left\langle \mathcal{H}(t), g(t) \right\rangle_{[a,b]} = \int_a^b \mathcal{H}(t)\, g(t) dt \tag{L.60}$$

with $g(t)$ at least piecewise continuous, does not present any computational or conceptual problems. We can use this fact to explore the action of $\delta(t)$ by studying the integral,

$$\left\langle \delta(t), g(t) \right\rangle = \int_{-\infty}^{\infty} \delta(t)\, g(t) dt \tag{L.61}$$

where $g(t)$ is a bounded differentiable function with bounded derivatives.

By having $\delta(t)$ be the derivative of $\mathcal{H}(t)$, (L.61) can be integrated by parts,

$$\begin{aligned}\int_{-\infty}^{\infty}\delta(t)\,g(t)dt &= \int_{-\infty}^{\infty}\frac{d}{dt}\mathcal{H}(t)\,g(t)dt \\ &= -\int_{-\infty}^{\infty}\mathcal{H}(t)\,\frac{dg}{dt}dt \\ &\quad +\mathcal{H}(\infty)\,g(\infty)-\mathcal{H}(-\infty)\,g(-\infty) \\ &= -\int_{0}^{\infty}\frac{dg}{dt}dt+g(\infty) \\ &= g(0) \end{aligned} \tag{L.62}$$

Thus $\delta(t)$ can be defined based on the associated action on $g(t)$, resulting with a number $g(0)$. If $g(t)=1$,

$$\int_{-\infty}^{\infty}\delta(t)\,dt=1 \tag{L.63}$$

The operational definition of $\delta(t)$ given in (L.62) may suffice for some applications. Other applications, however, require extensions of this operation to accommodate algebraic operations and calculus involving $\delta(t)$. To do so, the theory of distributions was developed by L. Schwarz as a framework to define mathematical objects called **distributions** and their operations, of which $\delta(t)$ is one particular example.

L.4.2 Theory of Distributions

Consider the following collection of continuous functions that are used to define distributions:

Definition L.6. *A continuous bounded function $\varphi(t)$ is a* **test function** *if*

1. *$\varphi(t)\in\mathcal{C}^{\infty}$, i.e. $d^k\varphi/dt^k$ is continuous for all integer k*
2. *$\varphi(t)$ has compact support $[a,b]$, i.e. $\varphi(t)=0$ for $(-\infty\le t<a)$ and $(b<t\le\infty)$*

An example of a test function is the smooth-pulse function given by

$$\varphi_{ab}(t)=\begin{cases}0 & \text{if} \quad t\le a\\ \exp\left[1-\frac{ab}{(t-a)(t-b)}\right] & \text{if} \quad a<t<b\\ 0 & \text{if} \quad t\ge b\end{cases} \tag{L.64}$$

A plot of $\varphi_{ab}(t)$ is shown in Figure L.14.

Definition L.7. *A* **distribution**, $\textsc{Dist}(t)$, *is a mapping from the set of test functions, Φ_{test}, to the set of real (or complex) numbers given by*

$$\left\langle \textsc{Dist}(t),\varphi(t)\right\rangle=\int_{-\infty}^{\infty}\textsc{Dist}(t)\,\varphi(t)dt \tag{L.65}$$

for $\varphi\in\Phi_{\text{test}}$, such that the map is

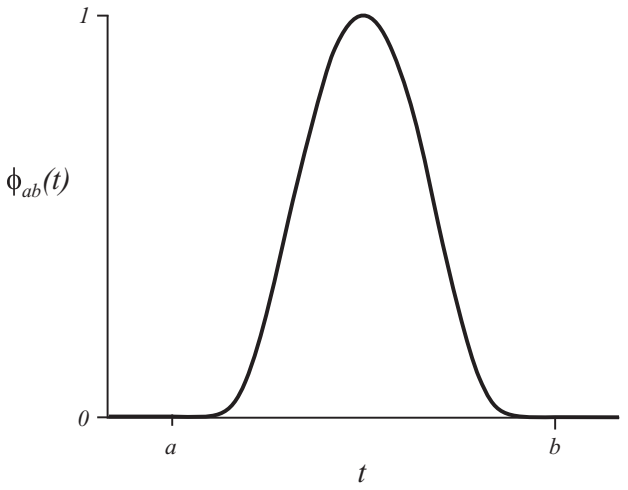


Figure L.14. A plot of the smooth pulse function defined by (L.64).

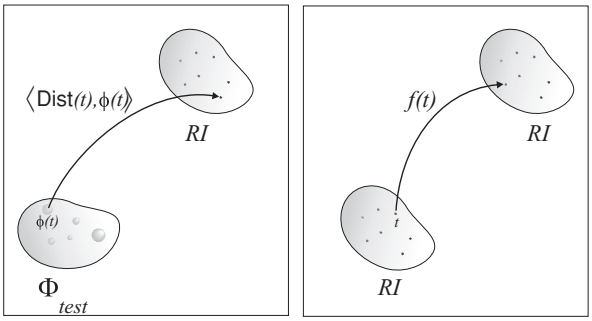
1. Linear: *For* $\varphi, \psi \in \Phi_{\text{test}}$ *and* α, β *constants,*

$$\langle \text{DIST}(t), \alpha\varphi(t) + \beta\psi(t) \rangle = \alpha \langle \text{DIST}(t), \varphi(t) \rangle + \beta \langle \text{DIST}(t), \psi(t) \rangle \quad \text{(L.66)}$$

and

2. Continuous: *For any convergent sequence of test functions* $\varphi_n \to 0$ *then* $\langle \text{DIST}(t), \varphi_n(t) \rangle \to 0$*, where the convergence of sequence of test functions satisfies.*
   (a) All the test functions in the sequence have the same compact support.
   (b) For each k, the kth derivatives of the test functions converges uniformly to zero.

Note that although we denote a distribution by DIST(t), (L.65) shows that the argument t is an integration variable. Distributions are also known as **generalized functions** because functions can also act as distributions. Moreover, using a very narrow smooth-pulse function, for example, $\varphi_{ab}(t)$ in (L.64) centered around t_o with $a \to b$ and under appropriate normalization, the distribution based on a function $f(t)$ reduces to the same evaluation operation of $f(t)$ at $t = t_o$. However, the important difference is that distributions are mappings from test functions to real (or complex) numbers, whereas functions are mappings from real (or complex) numbers to real (or complex) numbers, as shown in Figure L.15.



a) Distributions b) Functions

Figure L.15. A comparison of the mappings of distributions and functions.

Based on the conventional rules of integration, the following operation on distributions also yield distributions:

1. **Linear Combination of Distributions**. Let $g_1(t), g_2(t) \in \mathcal{C}^\infty$, that is, infinitely differentiable functions, then

$$\text{DIST}_{\text{comb}}(t) = [g_1(t)\text{DIST}_1(t) + g_2(t)\text{DIST}_2(t)]$$

is a distribution and

$$\begin{aligned}\langle [g_1(t)\text{DIST}_1(t) + g_2(t)\text{DIST}_2(t)], \varphi(t)\rangle = \\ \langle \text{DIST}_1(t), g_1(t)\varphi(t)\rangle \ + \ \langle \text{DIST}_2(t), g_2(t)\varphi(t)\rangle\end{aligned} \tag{L.67}$$

In particular, if $g_1(t) = \alpha$ and $g_2(t) = \beta$ are constants,

$$\begin{aligned}\langle [\alpha\text{DIST}_1(t) + \beta\text{DIST}_2(t)], \varphi(t)\rangle = \\ \alpha\langle \text{DIST}_1(t), \varphi(t)\rangle \ + \ \beta\langle \text{DIST}_2(t), \varphi(t)\rangle\end{aligned} \tag{L.68}$$

To prove (L.67), we simply evaluate the integral,

$$\begin{aligned}&\langle [g_1(t)\text{DIST}_1(t) + g_2(t)\text{DIST}_2(t)], \varphi(t)\rangle \\ &\quad = \int_{-\infty}^{\infty} [g_1(t)\text{DIST}_1(t) + g_2(t)\text{DIST}_2(t)]\varphi(t)dt \\ &\quad = \int_{-\infty}^{\infty} [g_1(t)\text{DIST}_1(t)\varphi(t)]dt + \int_{-\infty}^{\infty} [g_2(t)\text{DIST}_2(t)\varphi(t)]dt \\ &\quad = \langle \text{DIST}_1(t), g_1(t)\varphi(t)\rangle + \langle \text{DIST}_2(t), g_2(t)\varphi(t)\rangle\end{aligned}$$

2. **Invertible Monotonic Transformation of Argument**. Let $\vartheta(t)$ be an invertible and monotonic transformation of argument t, that is, $(d\vartheta/dt \neq 0)$, then

$$\text{DIST}_\vartheta(t) = \text{DIST}(\vartheta(t))$$

is also a distribution, and

$$\langle \text{DIST}(\vartheta(t)), \varphi(t)\rangle = \left\langle \text{DIST}(z), \frac{\varphi(\vartheta^{-1}(z))}{\varrho(z)} \right\rangle \tag{L.69}$$

where

$$\begin{aligned} z &= \vartheta(t) \\ \zeta(t) &= \left|\frac{d\vartheta}{dt}\right| \\ \varrho(z) &= \zeta(\vartheta^{-1}(z)) \end{aligned} \tag{L.70}$$

In particular, we have for translation, $\vartheta(t) = t - \alpha$, then

$$\begin{aligned}\langle \text{DIST}(t-\alpha), \varphi(t)\rangle &= \langle \text{DIST}(z), \varphi(z+a)\rangle \\ &= \langle \text{DIST}(t), \varphi(t+a)\rangle\end{aligned} \tag{L.71}$$

where we replaced z by t again because these can be considered dummy variables during the integration process.

Another particular example is for scaling of the argument, $\vartheta(t) = \alpha t$, then

$$\begin{aligned} \langle \text{DIST}(\alpha t), \varphi(t) \rangle &= \frac{1}{|\alpha|} \left\langle \text{DIST}(z), \varphi\left(\frac{z}{\alpha}\right) \right\rangle \\ &= \frac{1}{|\alpha|} \left\langle \text{DIST}(t), \varphi\left(\frac{t}{\alpha}\right) \right\rangle \end{aligned} \tag{L.72}$$

To prove (L.69), evaluate the integral,

$$\langle \text{DIST}(\vartheta(t)), \varphi(t) \rangle = \int_{-\infty}^{\infty} \text{DIST}(\vartheta(t)) \varphi(t) dt \tag{L.73}$$

$$= \int_{\vartheta(-\infty)}^{\vartheta(\infty)} \text{DIST}(z) \varphi(\vartheta^{-1}(z)) \frac{1}{d\vartheta/dt} dz \tag{L.74}$$

Recall that $\vartheta(t)$ is an invertible monotonic transformation of t. Suppose $\vartheta(t)$ is strictly monotonically increasing. Then $z \to \infty$ as $t \to \infty$ and $d\vartheta/dt > 0$. However, if $\vartheta(t)$ is strictly monotonically decreasing, $z \to -\infty$ as $t \to \infty$ and $d\vartheta/dt > 0$. For the latter case, the lower limit of integration will be $+\infty$ and the upper limit is $-\infty$. Thus, for either case, by fixing the upper limit to be $+\infty$ and the lower limit to be $-\infty$, we take the absolute value of $d\vartheta/dt$ when defining $\zeta(t)$ in (L.70).

3. **Derivatives of Distributions**. The derivative of distribution $\text{DIST}(t)$, denoted by $\text{DIST}'(t)$, is also a distribution. After applying integration by parts, the operation of $\text{DIST}'(t)$ is given by

$$\begin{aligned} \langle \text{DIST}'(t), \varphi(t) \rangle &= \left\langle \frac{d}{dt} \text{DIST}(t), \varphi(t) \right\rangle \\ &= \int_{-\infty}^{\infty} \frac{d\text{DIST}(t)}{dt} \varphi(t) dt \\ &= -\int_{-\infty}^{\infty} \text{DIST}(t) \frac{d\varphi}{dt} dt \\ &= -\left\langle \text{DIST}(t), \frac{d\varphi(t)}{dt} dt \right\rangle \end{aligned} \tag{L.75}$$

Using the preceding operations of distributions, we have the following theorem that describes the calculus available for distributions.

THEOREM L.9. *Let* $\text{DIST}(t)$, $\text{DIST}_1(t)$, *and* $\text{DIST}_2(t)$ *be distributions,* $g(t)$ *be a* $\mathcal{C}^\infty$ *function, and* α *be a constant, then*

1. *The derivative of sums of distributions are given by*

$$\frac{d}{dt}(\text{DIST}_1(t) + \text{DIST}_2(t)) = \frac{d}{dt}(\text{DIST}_1(t)) + \frac{d}{dt}(\text{DIST}_2(t)) \tag{L.76}$$

2. *The derivative of a scalar product of a distribution with* $g(t)$ *is given by*

$$\frac{d}{dt}[g(t)\text{DIST}(t)] = \frac{dg}{dt}\text{DIST}(t) + g(t)\frac{d}{dt}\text{DIST}(t) \tag{L.77}$$

For the special case of $g(t) = \alpha$,

$$\frac{d}{dt}\left[\alpha \text{DIST}\,(t)\right] = \alpha \frac{d}{dt}\text{DIST}\,(t) \tag{L.78}$$

3. *The derivative of a distribution under argument transformation* $\vartheta(t)$*, where* $\vartheta(t)$ *is an invertible monotonic function, is given by*

$$\frac{d}{dt}\left[\text{DIST}\,(\vartheta(t))\right] = \left[\frac{d\vartheta}{dt}\right]\frac{d}{d\vartheta}\left[\text{DIST}\,(\vartheta)\right] \tag{L.79}$$

PROOF. See Section L.8.

L.4.3 Properties and Identities of Delta Distribution

As a consequence of distribution theory, some of the properties and identities of delta distribution are given by:

1. **Sifting property.**

$$\begin{aligned}\int_{-\infty}^{\infty} \delta\,(t-\alpha)\, f(t)dt &= \int_{-\infty}^{\infty} \delta\,(t)\, f(t+\alpha)dt \\ &= f(\alpha)\end{aligned} \tag{L.80}$$

2. **Rescaling property.** Let $\alpha \neq 0$,

$$\begin{aligned}\int_{-\infty}^{\infty} \delta\,(\alpha t)\, f(t)dt &= \frac{1}{|\alpha|}\int_{-\infty}^{\infty} \delta\,(t)\, f(t/\alpha)dt \\ &= \frac{1}{|\alpha|} f(0)\end{aligned} \tag{L.81}$$

A special case is when $\alpha = -1$, then $\delta\,(-t) = \delta\,(t)$.

3. **Identities Involving Derivatives.**

$$\left\langle \frac{d^n}{dt^n}\delta\,(t)\,, f(t)\right\rangle = (-1)^k \left\langle \frac{d^{(n-k)}}{dt^{(n-k)}}\delta\,(t)\,, \frac{d^k}{dt^k} f(t)\right\rangle\,,\ 0 \leq k \leq n \tag{L.82}$$

$$t^n \frac{d^m}{dt^m}\delta\,(t) = \begin{cases} 0 & \text{if} \quad 0 \leq m < n \\ (-1)^n \frac{m!}{(m-n)!}\frac{d^{(m-n)}}{dt^{(m-n)}}\delta\,(t) & \text{if} \quad 0 \leq n \leq m \end{cases} \tag{L.83}$$

(See Section L.8 for the proof of (L.83).)

Special cases include the following:

$$t\frac{d}{dt}\delta\,(t) = -\delta\,(t) \tag{L.84}$$

$$t^2\frac{d}{dt}\delta\,(t) = 0 \tag{L.85}$$

$$\frac{d}{dt}\delta\,(-t) = -\frac{d}{dt}\delta\,(t) \tag{L.86}$$

4. **Identities under Argument Transformation.** Let $g(t)$ have a finite number of isolated and distinct roots, $r_1 \neq r_2 \neq \cdots \neq r_n$, and $|dg/dt|_{(t=r_k)} \neq 0$ for $k = 1, 2, \ldots, n$

$$\delta(g(t)) = \sum_{k=1}^{n} \frac{1}{|dg/dt|_{(t=r_k)}} \delta(t - r_k) \tag{L.87}$$

(See Section L.8 for the proof of (L.87).)

A special case is when $g(t) = t^2 - a^2$,

$$\delta\left(t^2 - a^2\right) = \frac{\delta(t-a) + \delta(t+a)}{2|a|} \tag{L.88}$$

L.4.4 Limit Identities for Delta Distribution

In the previous section, although we have shown several properties and identities of the delta distribution, it may sometimes be advantageous to base calculations on functions whose limits become the delta distribution. Surprisingly, the approximate functions do not even need to be positive definite, nor do they need to be symmetric with respect to the $t = 0$ axis.

THEOREM L.10. *Let $f(t)$ have the following properties:*

1. *$f(t)$ is piecewise continuous*
2. *$\left|\int_{-\infty}^{\infty} f(t)dt\right| < \infty$ and $\lim_{|t|\to\infty} f(t) = 0$*
3. *$\int_{-\infty}^{\infty} f(t)dt = 1$*

Then extending this function with a parameter α as follows,

$$F(\alpha, t) = \alpha f(\alpha t) \tag{L.89}$$

we have the following identity,

$$\lim_{\alpha\to\infty} F(\alpha, t) = \delta(t) \tag{L.90}$$

PROOF. See Section L.8

This theorem unifies different approaches used in different fields of applied mathematics to define the delta distribution. Some of the most common examples of functions used are:

1. **Gaussian Function.**

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{L.91}$$

and

$$F(\alpha, t) = \frac{\alpha}{\sqrt{2\pi}} e^{-(\alpha x)^2/2} \tag{L.92}$$

A plot of $F(\alpha, t)$ based on the gaussian function is shown in Figure L.16.

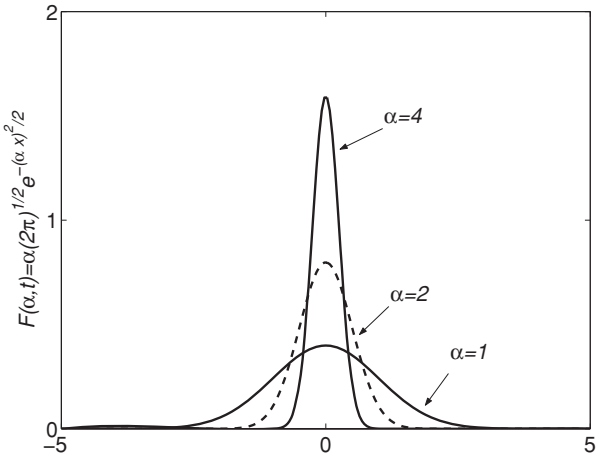


Figure L.16. A plot of $F(\alpha, t)$ based on the Gaussian function.

2. **Rectangular Pulse.** Let $\mathcal{H}(t)$ be the unit Heaviside step function; then the unit rectangular pulse function is given by

$$f(t) = \mathcal{H}\left(t + \frac{1}{2}\right) - \mathcal{H}\left(t - \frac{1}{2}\right) \tag{L.93}$$

and

$$F(\alpha, t) = \alpha\left(\mathcal{H}\left(\alpha t + \frac{1}{2}\right) - \mathcal{H}\left(\alpha t - \frac{1}{2}\right)\right) \tag{L.94}$$

A plot of $F(\alpha, t)$ based on the rectangular pulse function is shown in Figure L.17.

3. **Sinc Function**

$$f(t) = \frac{\sin(x)}{\pi x} \tag{L.95}$$

and

$$F(\alpha, t) = \frac{\sin(\alpha x)}{\pi x} \tag{L.96}$$

A plot of $F(\alpha, t)$ based on the sinc function is shown in Figure L.18.

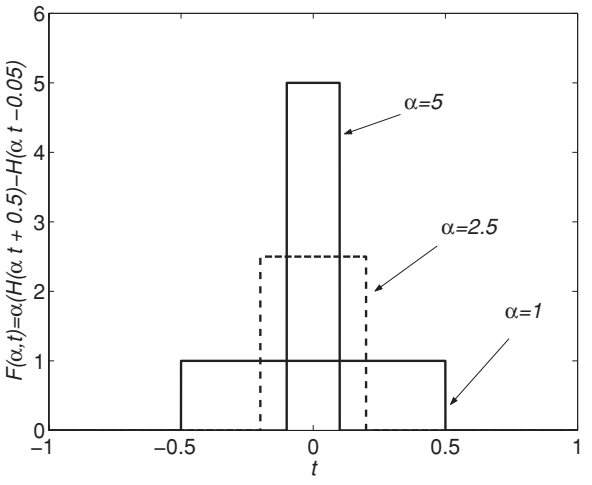


Figure L.17. A plot of $F(\alpha, t)$ based on the rectangular pulse function.

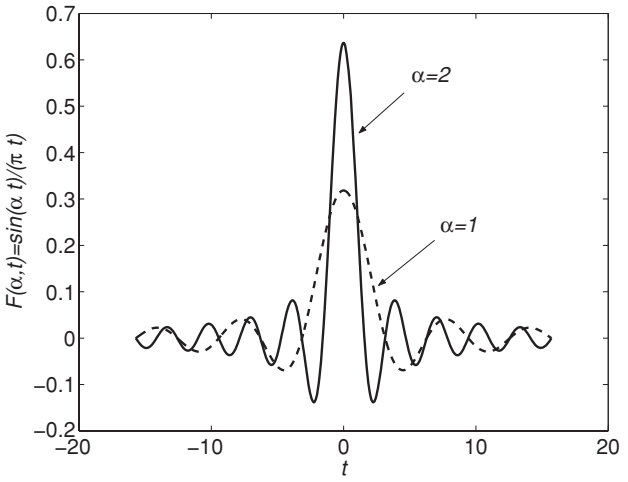


Figure L.18. A plot of $F(\alpha, t)$ based on the sinc function.

L.4.5 Delta Distribution for Higher Dimensions

Definition L.8. *For the Cartesian space of independent variables,* $\mathbf{x} \in \mathbb{R}^n$*,*

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{L.97}$$

the delta distribution of $\mathbf{x}$ *is given by*

$$\delta(\mathbf{x}) = \delta(x_1)\,\delta(x_2)\cdots\delta(x_n) \tag{L.98}$$

Under this definition, the properties of $\delta(t)$ can be used while integrating along each dimension. For instance, the sifting property for $g(\mathbf{x})$ with $\mathbf{p} \in \mathbb{R}^n$ becomes

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(\mathbf{x} - \mathbf{p})\, g(\mathbf{x})\, dx_1 \cdots dx_n = g(\mathbf{p}) \tag{L.99}$$

Note, however, that when dealing with the general curvilinear coordinates, normalization is needed to provide consistency.

Definition L.9. *Let* $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ *be a set of n curvilinear coordinates,*

$$\begin{aligned} \mu_1 &= \mu_1(x_1, x_2, \ldots, x_n) \\ \mu_2 &= \mu_2(x_1, x_2, \ldots, x_n) \\ &\vdots \\ \mu_n &= \mu_1(x_1, x_2, \ldots, x_n) \end{aligned} \tag{L.100}$$

that is invertible with the Cartesian coordinates $\mathbf{x} = (x_1, \ldots, x_n)$*, that is, the Jacobian matrix,*

$$J_{\mathrm{C}\to\mu} = \frac{\partial(\mu_1, \ldots, \mu_n)}{\partial(x_1, \ldots, x_n)} = \begin{pmatrix} \partial\mu_1/\partial x_1 & \cdots & \partial\mu_1/\partial x_n \\ \vdots & \ddots & \vdots \\ \partial\mu_n/\partial x_1 & \cdots & \partial\mu_n/\partial x_n \end{pmatrix} \tag{L.101}$$

is nonsingular.

Then, the delta distribution under the new coordinates of $\boldsymbol{\mu}$ *is given by*

$$\delta(\boldsymbol{\mu}) = \frac{\delta(\mu_1)\,\delta(\mu_2)\cdots\delta(\mu_n)}{\left|\det\left(J_{\mu\to\mathrm{C}}\right)\right|} \tag{L.102}$$

where $J_{\mu\to\mathrm{C}}$ *is the inverse of* $J_{\mathrm{C}\to\mu}$*,*

$$|J_{\mu\to\mathrm{C}}| = \left|\frac{\partial(x_1, \ldots, x_n)}{\partial(\mu_1, \ldots, \mu_n)}\right| \tag{L.103}$$

The inclusion of the denominator term in (L.102) is to maintain consistency, that is,

$$\begin{aligned}
\int_V \delta(\mathbf{x})\,dV &= \int_V \delta(\boldsymbol{\mu})\,dV \\
&= \int_{x_{n,\mathrm{lo}}}^{x_{n,\mathrm{hi}}} \cdots \int_{x_{1,\mathrm{lo}}}^{x_{1,\mathrm{hi}}} \frac{(\delta(\mu_1)\cdots\delta(\mu_n))}{\left|\det\left(J_{\mu\to\mathrm{C}}\right)\right|} dx_1 \cdots dx_n \\
&= \int_{\mu_{n,\mathrm{lo}}}^{\mu_{n,\mathrm{hi}}} \cdots \int_{\mu_{1,\mathrm{lo}}}^{\mu_{1,\mathrm{hi}}} \frac{(\delta(\mu_1)\cdots\delta(\mu_n))}{\left|\det\left(J_{\mu\to\mathrm{C}}\right)\right|} \left|\det\left(J_{\mu\to\mathrm{C}}\right)\right| d\mu_1 \cdots d\mu_n \\
&= \int_{\mu_{n,\mathrm{lo}}}^{\mu_{n,\mathrm{hi}}} \cdots \int_{\mu_{1,\mathrm{lo}}}^{\mu_{1,\mathrm{hi}}} (\delta(\mu_1)\cdots\delta(\mu_n))\,d\mu_1 \cdots d\mu_n \\
1 &= 1
\end{aligned}$$

where we used the relationship of multidimensional volumes in curvilinear coordinates, that is,

$$dV = dx_1 \cdots dx_n = \left|\det\left(J_{\mu\to\mathrm{C}}\right)\right| d\mu_1 \cdots d\mu_n$$

and $\mathbf{x}$ is an interior point in region V.

EXAMPLE L.8. Consider the spherical coordinate system, $\boldsymbol{\mu}_{\text{sphere}} = (r, \theta, \phi)$. With

$$\begin{aligned} x &= r\sin(\theta)\cos(\phi) \\ y &= r\sin(\theta)\sin(\phi) \\ z &= r\cos(\theta) \end{aligned}$$

the Jacobian determinant, $|J_{\text{Sphere}\to\text{C}}|$, is given by

$$\begin{aligned} |J_{\text{Sphere}\to\text{C}}| &= \left|\frac{\partial(x, y, z)}{\partial(r, \theta, \phi)}\right| \\ &= \left|\begin{pmatrix} \partial x/\partial r & \partial x/\partial\theta & \partial x/\partial\phi \\ \partial y/\partial r & \partial y/\partial\theta & \partial y/\partial\phi \\ \partial z/\partial r & \partial z/\partial\theta & \partial z/\partial\phi \end{pmatrix}\right| \\ &= \left|\begin{pmatrix} \sin(\theta)\cos(\phi) & r\cos(\theta)\cos(\phi) & -r\sin(\theta)\sin(\phi) \\ \sin(\theta)\sin(\phi) & r\cos(\theta)\sin(\phi) & r\sin(\theta)\cos(\phi) \\ \cos(\theta) & -r\sin(\theta) & 0 \end{pmatrix}\right| \\ &= r^2\sin(\theta) \end{aligned}$$

Thus

$$\delta(r, \theta, \phi) = \frac{\delta(r)\,\delta(\theta)\,\delta(\phi)}{r^2\sin(\theta)}$$

L.5 Tempered Distributions and Fourier Transforms

The set of test functions defined in Definition L.6 contains infinitely differentiable continuous functions with compact support. If we relax some of these specifications and replace them with functions that are **rapidly decreasing functions** or **Schwartz functions** (to be defined next), we can generate a subset of distributions, called **tempered distributions**. Tempered distributions can then be used to define generalized Fourier transforms that can be applied on functions such as unit step functions, sines, and cosines and on distributions such as the delta distribution.

Definition L.10. *A continuous function $f(t)$ belongs to the* **Schwartz class**, *denoted by $\mathcal{S}$, if $f(t)$ is:*

1. **Infinitely differentiable**, *that is, $f \in \mathcal{C}^\infty$*
   and
2. **Rapidly decreasing**, *that is, there is a constant C_{nm}, such that*

$$\left|t^n \frac{d^m f}{dt^m}\right| < C_{nm}\,, \quad \text{as } t \to \pm\infty \quad \text{for } n, m = 0, 1, 2, \ldots$$

A classic example of a Schwartz function that does not have compact support is given by

$$f(t) = e^{-|t|^2} \tag{L.104}$$

A plot of (L.104) is shown in Figure L.19.

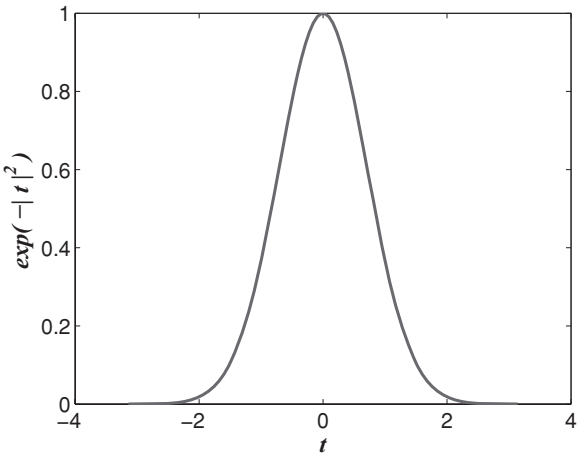


Figure L.19. A plot of $f(t) = exp(-|t|^2)$.

If we now replace test functions defined in Definition L.6 by Schwartz functions, we have the following definition for tempered distributions.

Definition L.11. *A* **tempered distribution***, denoted* $\text{TDIST}(t)$*, is a mapping from the set of Schwartz test functions,* $\mathcal{S}$*, to the set of real (or complex) numbers given by*

$$\langle \text{TDIST}(t), \varphi(t) \rangle = \int_{-\infty}^{\infty} \text{TDIST}(t)\, \varphi(t) dt \tag{L.105}$$

for $\varphi \in \mathcal{S}$*, such that the map is*

1. Linear: *For* $\varphi, \psi \in \mathcal{S}$ *and* α, β *constants,*

$$\langle \text{TDIST}(t), \alpha\varphi(t) + \beta\psi(t) \rangle = \alpha \langle \text{TDIST}(t), \varphi(t) \rangle + \beta \langle \text{TDIST}(t), \psi(t) \rangle \tag{L.106}$$

and

2. Continuous: *For any convergent sequence of Schwartz test functions* $\varphi_n \to 0$ *then* $\langle \text{TDIST}(t), \varphi_n(t) \rangle \to 0$

Because the set of test functions (with compact support), Φ_{test}, are already Schwartz test functions, the set of tempered distributions is automatically included in the set of regular distributions, that is, $\{\text{TDIST}(t)\} \subset \{\text{DIST}(t)\}$, which says that the class of regular distributions is much larger. This means that some distributions are not tempered distributions. The major issue is integrability, because Schwartz functions only decay to zero at $t = \pm\infty$, whereas regular test functions with compact support are zero outside the support. Fortunately, the delta distribution can be shown to be also a tempered distribution.

L.5.1 Generalized Fourier Transforms

Even though the space of tempered distributions is smaller than that of regular distributions, one of the main applications of tempered distributions is the generalization of Fourier transforms. This begins with the fact that Fourier or inverse Fourier transforms of Schwartz functions are again Schwartz functions.

THEOREM L.11. *Let $f \in \mathcal{S}$ then $\mathcal{F}[f] \in \mathcal{S}$, where $\mathcal{S}$ is the class of Schwartz functions and $\mathcal{F}$ is the Fourier transform operator.*

PROOF. First, we note an upper bound on Fourier transforms,

$$\begin{aligned} \left| \mathcal{F}[f] \right| &= \left| \int_{-\infty}^{\infty} e^{-i\omega t} f(t) dt \right| \\ &\leq \left| \int_{-\infty}^{\infty} f(t) dt \right| \end{aligned} \tag{L.107}$$

Next, we need two derivative formulas. The first formula is given by,

$$\begin{aligned} \mathcal{F}\left[(-it)^m f(t) \right] &= \int_{-\infty}^{\infty} e^{-i\omega t} (-it)^m f(t) dt \\ &= \int_{-\infty}^{\infty} \left(\frac{d^m}{d\omega^m} \left(e^{-i\omega t} \right) \right) f(t) dt \\ &= \frac{d^m}{d\omega^m} \int_{-\infty}^{\infty} e^{-i\omega t} f(t) dt \\ &= \frac{d^m}{d\omega^m} \left(\mathcal{F}[f] \right) \end{aligned} \tag{L.108}$$

The second derivative formula is given by,

$$\begin{aligned} \mathcal{F}\left[\frac{d^n f}{dt^n} \right] &= \int_{-\infty}^{\infty} e^{-i\omega t} \frac{d^n f}{dt^n} dt \\ &= (i\omega)^n \mathcal{F}[f] \\ &\quad \text{(after integration by parts)} \end{aligned} \tag{L.109}$$

Combining (L.108) and (L.109),

$$\mathcal{F}\left[\frac{d^n}{dt^n} \left((-it)^m f \right) \right] = (i\omega)^n \frac{d^m}{d\omega^m} \left(\mathcal{F}[f] \right) \tag{L.110}$$

After some rearranging and taking absolute values,

$$\left| \omega^n \frac{d^m}{d\omega^m} \left(\mathcal{F}[f] \right) \right| = \left| \mathcal{F}\left[\frac{d^n}{dt^n} \left((-it)^m f \right) \right] \right| \tag{L.111}$$

Applying the upper bound given by (L.107),

$$\left| \omega^n \frac{d^m}{d\omega^m} \left(\mathcal{F}[f] \right) \right| \leq \left| \int_{-\infty}^{\infty} \frac{d^n}{dt^n} \left(t^m f \right) dt \right| \tag{L.112}$$

Because f is a Schwartz function, the term on the right-hand side can be replaced by a constant C_{nm}. This means that $\mathcal{F}[f]$ is also a Schwartz function.

With this fact, we can define the Fourier transform of tempered distributions.

Definition L.12. *Let* $\text{TDIST}(t)$ *be a tempered distribution and* $\varphi(t)$ *a Schwartz function. Then the* **generalized Fourier transform** *of* $\text{TDIST}(t)$*, denoted by* $\mathcal{F}[\text{TDIST}(t)]$*, is a tempered distribution defined by the following operation*

$$\left\langle \mathcal{F}[\text{TDIST}(t)], \varphi(\omega) \right\rangle = \left\langle \text{TDIST}(\omega), \mathcal{F}[\varphi(t)] \right\rangle \tag{L.113}$$

Note that (L.113) is acceptable because $\text{TDIST}(\omega)$ was already assumed to be a tempered distribution and $\mathcal{F}[\varphi(t)]$ is guaranteed to be a Schwartz function (via Theorem L.11). Also, note the change of independent variable from t to ω, because the Fourier transform yields a function in ω. The tempered distribution $\text{TDIST}()$ will be based on ω.

With this definition, we are able to define Fourier transforms of functions such as cosines and sines and distributions such as delta distributions. Moreover, the Fourier transforms of distributions will yield the same Fourier transform operation if the distribution is a function that allow the classical Fourier transform.

EXAMPLE L.9. Fourier transform of delta distribution. Let $\varphi(\omega)$ be a Schwartz function.

$$\begin{aligned}
\left\langle \mathcal{F}[\delta(t-a)], \varphi(\omega) \right\rangle &= \int_{-\infty}^{\infty} \delta(\omega - a)\mathcal{F}[\varphi(t)]\, d\omega \\
&= \int_{-\infty}^{\infty} \delta(\omega - a) \left(\int_{-\infty}^{\infty} e^{-i\omega t}\varphi(t)dt \right) d\omega \\
&= \int_{-\infty}^{\infty} e^{-iat}\varphi(t)dt \\
&= \left\langle e^{-iat}, \varphi(t) \right\rangle \\
&= \left\langle e^{-ia\omega}, \varphi(\omega) \right\rangle
\end{aligned}$$

where we used the sifting property of delta distribution. Also, in the last line, we substituted ω for t by considering t can as a dummy integration variable.

Comparing both sides, we conclude that

$$\mathcal{F}[\delta(t-a)] = e^{-ia\omega} \tag{L.114}$$

and for the special case of $a = 0$,

$$\mathcal{F}[\delta(t)] = 1 \tag{L.115}$$

with "1" treated as a tempered distribution.

EXAMPLE L.10. Fourier transform of e^{iat}, cosines, and sines. First consider the Fourier transform of e^{iat},

$$\left\langle \mathcal{F}\left[e^{iat}\right], \varphi(\omega) \right\rangle = \int_{-\infty}^{\infty} e^{ia\omega}\mathcal{F}[\varphi(t)]\, d\omega \tag{L.116}$$

where the right-hand side can be seen as 2π times the inverse Fourier transform at $t = a$, that is,

$$\begin{aligned}\int_{-\infty}^{\infty} e^{ia\omega}\mathcal{F}\left[\varphi(t)\right]d\omega &= 2\pi\mathcal{F}^{-1}\mathcal{F}\left[\varphi(t)\right]\Big|_{t=a} \\ &= 2\pi\ \varphi(t)\big|_{t=a} \ = \ 2\pi\varphi(a) \\ &= 2\pi\int_{-\infty}^{\infty}\delta(t-a)\varphi(t)dt \\ &= 2\pi\Big\langle\ \delta(t-a),\varphi(t)\ \Big\rangle \\ &= 2\pi\Big\langle\ \delta(\omega-a),\varphi(\omega)\ \Big\rangle \end{aligned} \tag{L.117}$$

Comparing (L.116) and (L.117), we conclude that

$$\mathcal{F}\left[e^{iat}\right] = 2\pi\delta(\omega - a) \tag{L.118}$$

In particular, we have for $a = 0$,

$$\mathcal{F}\left[1\right] = 2\pi\delta(\omega) \tag{L.119}$$

Using (L.118), Euler's identity, and the linearity property of tempered distributions, we have

$$\begin{aligned}\mathcal{F}\left[\cos\left(at\right)\right] &= \mathcal{F}\left[\frac{e^{iat}+e^{-iat}}{2}\right] \\ &= \frac{1}{2}\left(\mathcal{F}\left[e^{iat}\right]+\mathcal{F}\left[e^{-iat}\right]\right) \\ &= \pi\Big(\ \delta\left(\omega-a\right)+\delta\left(\omega+a\right)\ \Big)\end{aligned} \tag{L.120}$$

Similarly for sine, we obtain

$$\mathcal{F}\left[\sin\left(at\right)\right] = i\pi\Big(\ \delta\left(\omega+a\right)-\delta\left(\omega-a\right)\ \Big) \tag{L.121}$$

Suppose $f(t)$ already possesses a classical Fourier transform; for example, it satisfies Dirichlet conditions and it is integrable; then we end up with the same evaluation. To see this, we have:

$$\begin{aligned}\Big\langle\mathcal{F}\left[f(t)\right],\varphi(\omega)\Big\rangle &= \int_{-\infty}^{\infty}f(\omega)\int_{-\infty}^{\infty}e^{-i\omega t}\varphi(t)dtd\omega \\ &= \int_{-\infty}^{\infty}\varphi(t)\int_{-\infty}^{\infty}e^{-i\omega t}f(\omega)d\omega dt \\ &= \int_{-\infty}^{\infty}\varphi(\omega)\int_{-\infty}^{\infty}e^{-i\omega t}f(t)dt\ d\omega \\ &= \left\langle\ \left[\int_{-\infty}^{\infty}e^{-i\omega t}f(t)dt\right],\varphi(\omega)\ \right\rangle\end{aligned}$$

where we exchanged the roles of ω and t in the last two lines. Thus we have

$$\mathcal{F}\left[f(t)\right] = \int_{-\infty}^{\infty}e^{-i\omega t}f(t)dt$$

This shows that we indeed obtained a generalization of the classic Fourier transform.

L.5.2 Generalized Fourier Transform of Integrals

All the properties of the classic Fourier transforms carry over to the generalized Fourier transforms. One additional property, however, that takes advantage of tempered distribution is the property for generalized Fourier transform of integrals.

THEOREM L.12. *Let $f(t)$ have a generalized Fourier transform. Then*

$$\mathcal{F}\left[\int_{-\infty}^{t} f(\eta)d\eta\right] = \pi\delta(\omega)\left(\mathcal{F}\left[f(t)\right]\Big|_{\omega=0}\right) + \frac{1}{i\omega}\mathcal{F}\left[f(t)\right] \tag{L.122}$$

PROOF. First, we apply the operation of tempered distributions on the generalized Fourier transforms as follows:

$$\begin{aligned}\left\langle \mathcal{F}\left[\int_{-\infty}^{t} f(\eta)d\eta\right], \varphi(\omega)\right\rangle &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\omega} f(\eta)d\eta\right)\int_{-\infty}^{\infty} e^{-i\omega t}\varphi(t)dt\, d\omega \\ &= \int_{-\infty}^{\infty}\varphi(t)\int_{-\infty}^{\infty} e^{-i\omega t}\left(\int_{-\infty}^{\omega} f(\eta)d\eta\right)d\omega\, dt \\ &= \int_{-\infty}^{\infty}\varphi(t)\left[\xi_{\cos}(t) + \xi_{\sin}(t) + \zeta(t)\right]dt \end{aligned} \tag{L.123}$$

where the terms $\xi_{\cos}(t)$, $\xi_{\sin}(t)$ and $\zeta(t)$ are obtained after integration by parts[6] to be

$$\xi_{\cos}(t) = \lim_{\omega\to\infty} -\frac{\cos(\omega t)}{it}\int_{-\infty}^{\infty} f(\omega)d\omega \tag{L.124}$$

$$\begin{aligned}\xi_{\sin}(t) &= \lim_{\omega\to\infty}\frac{\sin(\omega t)}{it}\int_{-\infty}^{\infty} f(\omega)d\omega \\ &= \pi\delta(t)\left(\mathcal{F}\left[f(\omega)\right]\Big|_{t=0}\right)\end{aligned} \tag{L.125}$$

$$\begin{aligned}\zeta(t) &= \frac{1}{it}\int_{-\infty}^{\infty} e^{-i\omega t}f(\omega)d\omega \\ &= \frac{1}{it}\mathcal{F}\left[f(\omega)\right]\end{aligned} \tag{L.126}$$

Next, expand (L.123) to obtain the three additive terms evaluated as,

$$\int_{-\infty}^{\infty}\varphi(t)\xi_{\cos}(t)dt = 0 \quad \text{(treated as a principal value integral)} \tag{L.127}$$

$$\int_{-\infty}^{\infty}\varphi(t)\xi_{\sin}(t)dt = \left\langle\left[\,\pi\delta(\omega)\left(\mathcal{F}\left[f(t)\right]\Big|_{\omega=0}\right)\,\right], \varphi(\omega)\right\rangle \tag{L.128}$$

$$\int_{-\infty}^{\infty}\varphi(t)\zeta(t)dt = \left\langle\left[\,\frac{1}{i\omega}\mathcal{F}\left[f(t)\right]\,\right], \varphi(\omega)\right\rangle \tag{L.129}$$

[6] Let $\left(\, u = \int_{-\infty}^{\omega} f(\eta)d\eta \,\right)$ and $\left(\, dv = \exp(-i\omega t)\,\right)$. Then, $v = -[1/(it)]\exp(-i\omega t)$. And using Leibnitz rule, $du = f(\omega)d\omega$.

We again switched the roles of t and ω in (L.128) and (L.129). Adding these three terms together and then comparing with the right-hand side of (L.123), we obtain (L.122).

EXAMPLE L.11. Fourier transform of the unit step function and signum function. The (dual) definition of the unit step function is that it is the integral to the delta distribution. Using (L.122) and the fact that $\mathcal{F}[\delta(t)] = 1$ (cf. (L.115)), we have

$$\begin{aligned}\mathcal{F}[\mathcal{H}(t)] &= \mathcal{F}\left[\int_{-\infty}^{t} \mathcal{H}(\eta)\, d\eta\right] \\ &= \pi\delta(\omega)\left(\mathcal{F}[\delta(t)]\Big|_{\omega=0}\right) + \frac{1}{i\omega}\mathcal{F}[\delta(t)] \\ &= \pi\delta(\omega) + \frac{1}{i\omega} \end{aligned} \tag{L.130}$$

Furthermore, with the relationship between $\mathcal{H}(t)$ and $\mathrm{sgn}(t)$ given by,

$$\mathrm{sgn}(t) = 2\mathcal{H}(t) - 1 \tag{L.131}$$

we can proceed as before, while using (L.119)

$$\begin{aligned}\left\langle \mathcal{F}[\mathrm{sgn}(t)], \varphi(\omega)\right\rangle &= 2\left\langle \mathcal{F}[\mathcal{H}(t)], \varphi(\omega)\right\rangle - \left\langle 1, \varphi(\omega)\right\rangle \\ &= 2\left\langle \left[\frac{1}{i\omega} + \pi\delta(\omega)\right], \varphi(\omega)\right\rangle - \left\langle 2\pi\delta(\omega), \varphi(\omega)\right\rangle \\ &= 2\left\langle \frac{1}{i\omega}, \varphi(\omega)\right\rangle \end{aligned}$$

Thus

$$\mathcal{F}[\mathrm{sgn}(t)] = \frac{2}{i\omega} \tag{L.132}$$

L.6 Supplemental Lemmas, Theorems, and Proofs

L.6.1 Dirichlet Integral Theorem

Part of this theorem is used for the proof of Fourier's integral theorem (Theorem L.8).

THEOREM L.13. *Let $f(x+\eta)$ satisfy Dirichlet's conditions in the interval (a, b), where $a \geq -\infty$ and $b \leq \infty$, then*

$$\lim_{\theta\to\infty} \frac{2}{\pi}\int_a^b f(x+\eta)\,\frac{\sin(\theta\eta)}{\eta}\,d\eta = \begin{cases} [f(x^+) + f(x^-)] & \text{if} \quad a < 0 < b \\ f(x^+) & \text{if} \quad 0 = a < b \\ f(x^-) & \text{if} \quad a < b = 0 \\ 0 & \text{if} \quad \begin{array}{l} 0 < a < b \\ \text{or} \quad a < b < 0 \end{array} \end{cases} \tag{L.133}$$

PROOF. We start with the fact that

$$\int_0^\infty \frac{\sin(q)}{q}\, dq = \frac{\pi}{2} \tag{L.134}$$

and for $q_2 > q_1$,

$$\lim_{q_1 \to \infty} \int_{q_1}^{q_2} \frac{\sin(q)}{q}\, dq = 0 \tag{L.135}$$

Assume that $f(x+\eta)$ is monotonic in a subinterval (α, β) of (a, b) for the case $a > 0$, the mean value theorem says there exists $a < \xi < b$ such that, with $q = \theta\eta$,

$$\begin{aligned} \int_\alpha^\beta f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\, d\eta &= f(x+\alpha^+)\int_\alpha^\xi \frac{\sin(\theta\eta)}{\eta}\, d\eta + f(x+\beta^-)\int_\xi^\beta \frac{\sin(\theta\eta)}{\eta}\, d\eta \\ &= f(x+\alpha^+)\int_{\theta\alpha}^{\theta\xi} \frac{\sin(q)}{q}\, dq + f(x+\beta^-)\int_{\theta\xi}^{\theta\beta} \frac{\sin(q)}{q}\, dq \end{aligned}$$

and with (L.135),

$$\lim_{\theta \to \infty} \int_\alpha^\beta f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\, d\eta = 0 \tag{L.136}$$

Note that so far, (L.136) has been shown to apply to a subinterval where $f(x+\eta)$ is monotonic. However, because $f(x+\eta)$ satisfies Dirichlet's conditions, the interval (a, b) can be partitioned into n subintervals (a_i, a_{i+1}), with

$$0 < a = a_0 < a_1 < \cdots < a_n = b$$

such that $f(x)$ is monotonic inside each subinterval (e.g., with a_i occurring either at a discontinuity, minima, or maxima of $f(x)$). Thus

$$\lim_{\theta \to \infty} \int_{(a>0)}^b f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\, d\eta = \lim_{\theta \to \infty} \sum_{i=0}^{n-1} \int_{a_i}^{a_{i+1}} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\, d\eta = 0 \tag{L.137}$$

Similarly with $b < 0$, the same approach can be used to show

$$\lim_{\theta \to \infty} \int_a^{(b<0)} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\, d\eta = 0 \tag{L.138}$$

Next, for the case when $a = 0$, we need to focus only on the first interval, $(0, a_1)$, in which $f(x)$ is monotonic because (L.137) says the integral in the interval (a_1, b) is zero. Using the mean value theorem again, there exists $0 < \xi < a_1$ such that

$$\begin{aligned} \int_0^{a_1} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\, d\eta &= f(x^+)\int_0^\xi \frac{\sin(\theta\eta)}{\eta}\, d\eta + f(x+a_1{}^-)\int_\xi^{a_1} \frac{\sin(\theta\eta)}{\eta}\, d\eta \\ &= f(x^+)\int_0^{\theta\xi} \frac{\sin(q)}{q}\, dq + f(x+a_1{}^-)\int_{\theta\xi}^{\theta a_1} \frac{\sin(q)}{q}\, dq \end{aligned}$$

Applying (L.134) and (L.135),

$$\lim_{\theta\to\infty}\int_0^{a_1} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\,d\eta = f(x^+)\frac{\pi}{2}$$

or with $0 = a < b$,

$$\lim_{\theta\to\infty}\frac{2}{\pi}\int_0^{b} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\,d\eta = f(x^+) \tag{L.139}$$

Likewise, for $a < b = 0$, the same approach will yield

$$\lim_{\theta\to\infty}\frac{2}{\pi}\int_a^{0} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\,d\eta = f(x^-) \tag{L.140}$$

For the last case, that is, $a < 0 < b$, we simply add (L.139) and (L.140) to obtain

$$\lim_{\theta\to\infty}\frac{2}{\pi}\int_{(a<0)}^{(b>0)} f(x+\eta)\frac{\sin(\theta\eta)}{\eta}\,d\eta = f(x^+) + f(x^-) \tag{L.141}$$

L.6.2 A Technical Lemma for Fourier Integral Theorem

LEMMA L.1. *Let $f(x)$ be absolutely convergent, that is,*

$$\int_{-\infty}^{\infty}|f(x)|dx < \infty$$

then

$$\lim_{\theta\to\infty}\int_{-\infty}^{\infty} f(t)\int_0^{\theta}\cos\left((x-t)\omega\right)\,d\omega\,dt = \lim_{\theta\to\infty}\int_0^{\theta}\int_{-\infty}^{\infty} f(t)\cos\left((x-t)\omega\right)\,dt\,d\omega \tag{L.142}$$

PROOF. First, we look at the integrals of (L.142)

$$\begin{aligned}\int_0^{\infty} f(t)\int_0^{\theta}\cos\left((x-t)\omega\right)\,d\omega\,dt \quad &= \quad \int_0^{\tau} f(t)\int_0^{\theta}\cos\left((x-t)\omega\right)\,d\omega\,dt \\ &\quad + \int_{\tau}^{\infty} f(t)\int_0^{\theta}\cos\left((x-t)\omega\right)\,d\omega\,dt\end{aligned} \tag{L.143}$$

and

$$\begin{aligned}\int_0^{\theta}\int_0^{\infty} f(t)\cos\left((x-t)\omega\right)\,dt\,d\omega \quad &= \quad \int_0^{\theta}\int_0^{\tau} f(t)\cos\left((x-t)\omega\right)\,dt\,d\omega \\ &\quad + \int_0^{\theta}\int_{\tau}^{\infty} f(t)\cos\left((x-t)\omega\right)\,dt\,d\omega\end{aligned} \tag{L.144}$$

where $0 \le \tau < \infty$.

With $\tau < \infty$ and $\theta < \infty$, the sequence of integrals with finite limits can be interchanged, that is,

$$\int_0^{\tau} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt = \int_0^{\theta} \int_0^{\tau} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \quad \text{(L.145)}$$

With the assumption of absolute convergence, there exist T such that

$$\left| \int_{\tau}^{\infty} |f(t)| dt \right| < \frac{\epsilon}{\theta}$$

with $\epsilon > 0$ and $\tau > T$,

$$\begin{aligned} \left| \int_{\tau}^{\infty} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt \right| &= \left| \int_{\tau}^{\infty} f(t) \frac{\sin((x-t)\omega)}{x-t}\, dt \right| \\ &= \left| \int_{x+\tau}^{\infty} f(x+\eta) \frac{\sin(\eta\omega)}{\eta}\, d\eta \right| \\ &< \frac{1}{\tau} \left| \int_{\tau}^{\infty} |f(x+\eta)|\, d\eta \right| < \frac{\epsilon}{2\tau\theta} \end{aligned}$$

and

$$\begin{aligned} \left| \int_0^{\theta} \int_{\tau}^{\infty} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \right| &< \int_0^{\theta} \int_{\tau}^{\infty} |f(t)|\, dt\, d\omega \\ &< \frac{\epsilon}{2\theta} \int_0^{\theta} d\omega = \frac{\epsilon}{2} \end{aligned}$$

Combining both results,

$$\left| \int_{\tau}^{\infty} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt - \int_0^{\theta} \int_{\tau}^{\infty} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \right| < \frac{\epsilon}{2}\left(1 + \frac{1}{\tau\theta}\right) < \epsilon$$

Thus

$$\int_{\tau}^{\infty} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt = \int_0^{\theta} \int_{\tau}^{\infty} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \quad \text{(L.146)}$$

Taking the difference between (L.143) and (L.144), and then substituting (L.145) and (L.146),

$$\int_0^{\infty} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt = \int_0^{\theta} \int_0^{\infty} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \quad \text{(L.147)}$$

Using a similar approach, we can show

$$\int_{-\infty}^{0} f(t) \int_0^{\theta} \cos\left((x-t)\omega\right)\, d\omega\, dt = \int_0^{\theta} \int_{-\infty}^{0} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega \quad \text{(L.148)}$$

Adding (L.147) and (L.148), and then take the limit as $\omega \to \infty$,

$$\int_{-\infty}^{\infty} f(t) \int_{0}^{\infty} \cos\left((x-t)\omega\right)\, d\omega\, dt = \int_{0}^{\infty} \int_{-\infty}^{\infty} f(t) \cos\left((x-t)\omega\right)\, dt\, d\omega$$

L.7 More Examples of Laplace Transform Solutions

In this section, we solve the partial differential equations using the Laplace transforms. The first example shows the solution for the diffusion equation under boundary conditions that are different from Example 12.13. The second example shows the Laplace transform solution for the diffusion equation that includes a linear source term.

EXAMPLE L.12. Here we extend the results of Example 12.13 to handle a different set of boundary conditions. Thus with

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \tag{L.149}$$

under a constant initial condition, $u(x,0) = C_i$. In Example 12.13, we have already partially found the solution in the Laplace domain to be given by (12.83); this was found to be

$$\widehat{U} = Ae^{\lambda x} + Be^{-\lambda x} + \frac{C_i}{s} \tag{L.150}$$

where $\lambda_i = \sqrt{s}/\alpha$. Now we investigate the solutions for a different set of boundary conditions.

1. **Finite Domain.** Let the boundary conditions be

$$u(0,t) = C_0 \qquad \text{and} \qquad u(L,t) = C_L$$

Applying these to (L.150),

$$\widehat{U}(0,s) = \frac{1}{s}C_0 \quad = \quad A + B + \frac{C_i}{s}$$

$$\widehat{U}(L,s) = \frac{1}{s}C_L \quad = \quad Ae^{\lambda L} + Be^{-\lambda L} + \frac{C_i}{s}$$

or

$$A = \frac{e^{-\lambda L}(C_0 - C_i) - (C_L - C_i)}{s\left(e^{-\lambda L} - e^{\lambda L}\right)} \quad \text{and} \quad B = \frac{-e^{\lambda L}(C_0 - C_i) + (C_L - C_i)}{s\left(e^{-\lambda L} - e^{\lambda L}\right)}$$

Substituting back to (12.83),

$$\widehat{U} = (C_0 - C_i)\,\widehat{U}_a + (C_L - C_i)\,\widehat{U}_b + \frac{C_i}{s}$$

where

$$\widehat{U}_a = \frac{1}{s}\frac{\sinh\left(\lambda(L-x)\right)}{\sinh\left(\lambda L\right)} \qquad \widehat{U}_b = \frac{1}{s}\frac{\sinh\left(\lambda x\right)}{\sinh\left(\lambda L\right)}$$

To evaluate the inverse Laplace transform, we can use the residue theorem for an infinite number of simple poles (cf. Section L.2.3).[7] Fortunately, the poles of both $\widehat{U}_a$ and $\widehat{U}_b$ are all simple poles. They are given by[8]

$$s = 0 \qquad \text{and} \qquad \lambda L = ik\pi \rightarrow s_k = -\left(\frac{\alpha k\pi}{L}\right)^2 \ , \ k = 1, 2, \ldots$$

Note that $s_k = 0$ is a removable singularity of both $(\sinh(\lambda(L-x))/\sinh(\lambda L))$ and $(\sinh(\lambda x)/\sinh(\lambda L))$. Thus s_0 is not included in the sequence of s_k poles, leaving $s = 0$ to be a simple pole of $\widehat{U}_a$ and $\widehat{U}_b$.

Then with $\gamma > 0$,

$$\begin{aligned}
\mathcal{L}^{-1}\left[\widehat{U}_a\right] &= \frac{1}{2\pi i}\int_{\gamma-\infty}^{\gamma+\infty} e^{st}\widehat{U}_a(x,s)ds = \sum_{z=0,s_k} \mathbf{Res}_{\,z}\left(e^{st}\widehat{U}_a\right) \\
\mathbf{Res}_{\,0}\left(e^{st}\widehat{U}_a\right) &= \lim_{s\to 0}\ e^{st}\,\frac{\sinh(\lambda(L-x))}{\sinh(\lambda L)} = \frac{L-x}{L} \\
\mathbf{Res}_{\,s_k}\left(e^{st}\widehat{U}_a\right) &= \frac{e^{s_k t}}{s_k}\sinh\left(\frac{(L-x)\sqrt{s_k}}{\alpha}\right)\lim_{s\to s_k}\left(\frac{s-s_k}{\sinh(L\sqrt{s}/\alpha)}\right) \\
&= \frac{e^{s_k t}}{s_k}\sinh\left(\frac{(L-x)\sqrt{s_k}}{\alpha}\right)\frac{1}{\cosh\left(L\sqrt{s_k}/\alpha\right)}\,\frac{2\alpha\sqrt{s_k}}{L} \\
&= (-1)^k\frac{2}{k\pi}\sin\left(k\pi\frac{L-x}{L}\right)\exp\left(-\left[\frac{\alpha k\pi}{L}\right]^2 t\right)
\end{aligned}$$

Similarly, for $\widehat{U}_b$,

$$\begin{aligned}
\mathcal{L}^{-1}\left[\widehat{U}_b\right] &= \frac{1}{2\pi i}\int_{\gamma-\infty}^{\gamma+\infty} e^{st}\widehat{U}_b(x,s)ds = \sum_{z=0,s_k} \mathbf{Res}_{z}\left(e^{st}\widehat{U}_b\right) \\
\mathbf{Res}_{\,0}\left(e^{st}\widehat{U}_b\right) &= \frac{x}{L} \\
\mathbf{Res}_{\,s_k}\left(e^{st}\widehat{U}_b\right) &= (-1)^k\frac{2}{k\pi}\sin\left(k\pi\frac{x}{L}\right)\exp\left(-\left[\frac{\alpha k\pi}{L}\right]^2 t\right)
\end{aligned}$$

Combining all the results, we have the solution for $u(x,t)$:

$$u = u_{\text{steady-state}} + u_{\text{transient}}$$

[7] The following identities are also useful for the calculations in this example:

$$\sinh(i|z|) = i\sin(|z|)\ , \ \ \cosh(i|z|) = \cos(|z|)$$

and

$$\frac{d}{dz}\sinh(z) = \cosh(z)\ , \ \frac{d}{dz}\cosh(z) = \sinh(z)$$

[8] With $f(x = i|z|) = \sinh(i|z|) = i\sin(|z|)$, the roots of $f(x)$ are then given by $x = i\arcsin(0)$.

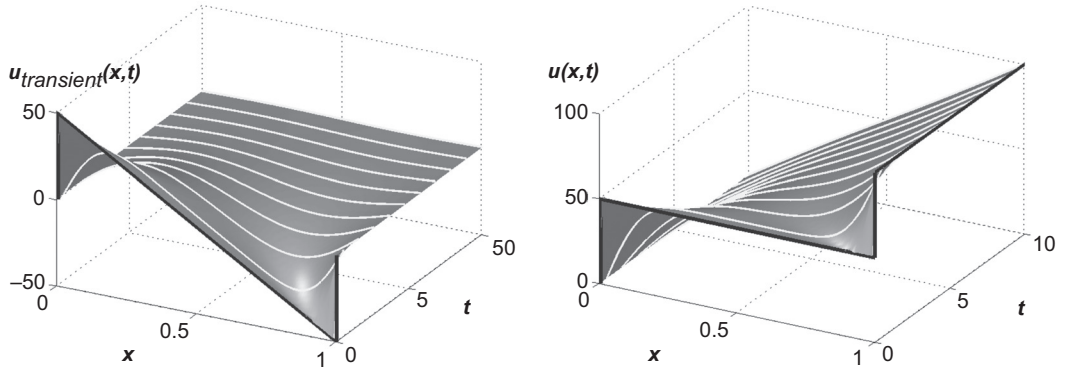


Figure L.20. Plots of $u_{\text{transient}}$ and $u(x,t)$ for example 12.13.

where

$$u_{\text{steady-state}} = C_i + (C_0 - C_i)\left(1 - \frac{x}{L}\right) + (C_L - C_i)\frac{x}{L}$$

$$u_{\text{transient}} = \sum_{k=1}^{\infty} \beta_k(t)\left((C_0 - C_i)\,\zeta_k(L-x) + (C_L - C_i)\,\zeta_k(x)\right)$$

with

$$\beta_k(t) = (-1)^k \frac{2}{k\pi} \exp\left(-\left[\frac{\alpha k\pi}{L}\right]^2 t\right) \qquad \text{and} \qquad \zeta_k(y) = \sin\left(k\pi\frac{y}{L}\right)$$

Plots of $u_{\text{transient}}(x,t)$ and $u(x,t)$ are shown in Figure L.20, for $\alpha = 0.1$, $L = 1$, $C_0 = 0$, $C_i = 50$, and $C_L = 100$, where the summation for $u_{\text{transient}}$ was truncated after $k = 250$.

1. **Dirichlet Conditions and Neumann Conditions, in Finite Domain.** Let

$$u(0,t) = C_0 \;;\; \frac{\partial u}{\partial x}(L,t) = 0 \qquad \text{and} \qquad u(x,0) = 0$$

Then (L.150) becomes

$$\widehat{U} = Ae^{\lambda x} + Be^{-\lambda x}$$

where $\lambda = \sqrt{s}/\alpha$. Using the Laplace transform of the boundary conditions,

$$\frac{C_0}{s} = A + B \qquad \text{and} \qquad 0 = Ae^{\lambda L} - Be^{\lambda L}$$

or

$$A = \frac{e^{-\lambda L}}{e^{\lambda L} + e^{-\lambda L}} \qquad \text{and} \qquad B = \frac{e^{+\lambda L}}{e^{\lambda L} + e^{-\lambda L}}$$

Thus

$$\widehat{U} = \frac{C_0}{s}\,\frac{e^{-\lambda(L-x)} + e^{\lambda(L-x)}}{e^{\lambda L} + e^{-\lambda L}} = \frac{C_0}{s}\,\frac{e^{-\lambda(2L-x)} + e^{-\lambda x}}{1 + e^{-2\lambda L}} \qquad \text{(L.151)}$$

Let $q = e^{-2\lambda L}$. Then using the fact that

$$\frac{1}{1+q} = \sum_{n=0}^{\infty}(-1)^n q^n$$

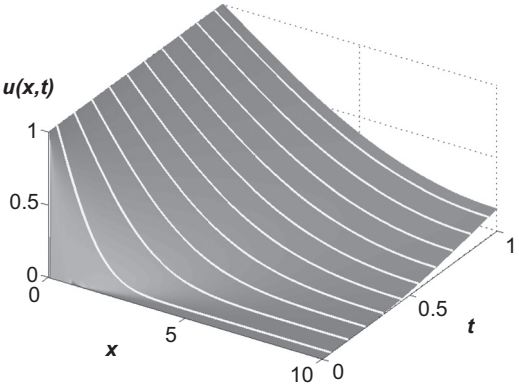


Figure L.21. A plot of the solution given by equation (L.152).

equation (L.151) becomes

$$\widehat{U} = C_0 \left(\sum_{n=0}^{\infty} (-1)^n \frac{1}{s} e^{-\beta_n(x)\sqrt{s}} + \sum_{n=0}^{\infty} (-1)^n \frac{1}{s} e^{-\gamma_n(x)\sqrt{s}} \right)$$

where,

$$\beta_n(x) = \frac{2L(n+1) - x}{\alpha} \qquad \text{and} \qquad \gamma_n(x) = \frac{2Ln + x}{\alpha}$$

Finally, the solution is given by

$$u(x,t) = C_0 \sum_{n=0}^{\infty} (-1)^n \left(\operatorname{erfc}\left(\frac{\beta_n(x)}{2\sqrt{t}} \right) + \operatorname{erfc}\left(\frac{\gamma_n(x)}{2\sqrt{t}} \right) \right) \qquad \text{(L.152)}$$

A plot of (L.152) with $C_0 = 1$, $L = 10$, and $\alpha = 4$ is shown in Figure L.21. Note that, although the plot qualitatively looks similar to Figure 12.3, the main difference is that profiles of $u(x, t)$ at fixed t have a zero slope at $x = L$.

EXAMPLE L.13. Laplace transform solution of diffusion equation with linear source term in a semi-infinite domain.

Consider the equation

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} + \sigma u \qquad \text{(L.153)}$$

with a constant initial condition $u(x, 0) = C_i$ and boundary conditions

$$u(0, t) = f(t) \qquad \text{and} \qquad \lim_{x \to \infty} |u(x, t)| < \infty$$

Taking the Laplace transform, we obtain

$$\alpha^2 \frac{d^2 \widehat{U}}{dx^2} = s\widehat{U} - C_i + \sigma \widehat{U}$$

whose solution is given by

$$\widehat{U} = Ae^{\lambda x} + Be^{-\lambda x} + \frac{C_i}{s}$$

where $\lambda = (\sqrt{s+\sigma})/\alpha$. Applying the boundary conditions, we get

$$A = 0 \qquad \text{and} \qquad B = \mathcal{L}[f] + \frac{C_i}{s}$$

Thus

$$\widehat{U} = \left(\mathcal{L}[f] - \frac{C_i}{s}\right) e^{-(\sqrt{s+\sigma})x/\alpha} + \frac{C_i}{s}$$

Using the convolution theorem, we have

$$u(x,t) = \int_0^t (f(t-\tau) - C_i)\, \mathcal{L}^{-1}\left[e^{-(\sqrt{s+\sigma})x/\alpha}\right] d\tau + C_i$$

To obtain the required inverse Laplace transform of the exponential term, we can start from item 7 in Table 12.4 and apply the derivative theorem,

$$\begin{aligned}\mathcal{L}^{-1}\left[e^{-\sqrt{s}}\right] &= \mathcal{L}^{-1}\left[\, s\left(\frac{1}{s}e^{-\sqrt{s}}\right) - \lim_{t\to 0}\left(\operatorname{erfc}\left(\frac{1}{2\sqrt{t}}\right)\right)\,\right] \\ &= \frac{d}{dt}\left(\operatorname{erfc}\left(\frac{1}{2\sqrt{t}}\right)\right) = \frac{1}{2\sqrt{\pi t^3}}\, e^{-1/(4t)}\end{aligned}$$

Next, applying both shifting and scaling,

$$\mathcal{L}^{-1}\left[e^{-\sqrt{(s+a)/b}}\right] = \frac{1}{2\sqrt{\pi b t^3}} \exp\left(-\frac{1}{4bt} - at\right)$$

Thus with $a = \sigma$ and $b = (\alpha/x)^2$,

$$u(x,t) = \frac{x}{2\alpha\sqrt{\pi}} \int_0^t \left[\frac{f(t-\tau) - C_i}{\sqrt{\tau^3}} \exp\left(-\frac{x^2}{4\alpha^2\tau} - \sigma\tau\right)\right] d\tau \;+\; C_i \qquad \text{(L.154)}$$

The integral in equation (L.154) is difficult to evaluate both analytically and numerically. If the boundary condition $f(t)$ is constant, then a closed-form solution is available. For the more general case, numerical integration is needed to evaluate the solution.

Case 1. $\underline{f(t) = C_0}$ **where** C_0 **is constant.** In this situation, (L.154) becomes

$$u(x,t) = \frac{x\,(C_0 - C_i)}{2\alpha\sqrt{\pi}} \mathcal{I}(x,t) \;+\; C_i$$

where,

$$\mathcal{I}(x,t) = \int_0^t \left[\frac{1}{\tau\sqrt{\tau}} \exp\left(-\frac{x^2}{4\alpha^2\tau} - \sigma\tau\right)\right] d\tau$$

To evaluate $\mathcal{I}(x,t)$, we introduce some auxiliary variables. Let q_1 and q_2 be defined by

$$q_1(\tau) = \frac{a}{\sqrt{\tau}} + b\sqrt{\tau} \qquad \text{and} \qquad q_2(\tau) = \frac{a}{\sqrt{\tau}} - b\sqrt{\tau}$$

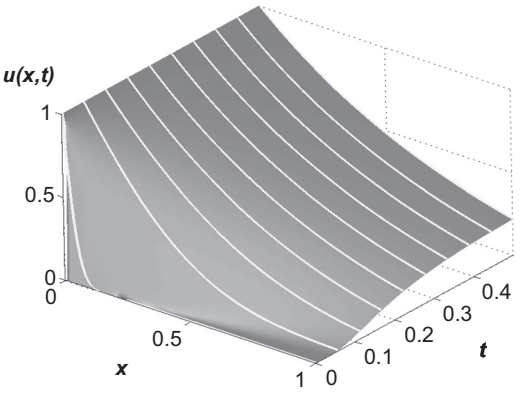


Figure L.22. A plot of the solution given by (L.155).

then

$$dq_1 = \frac{1}{2}\left(-\frac{a}{\tau\sqrt{\tau}} + \frac{b}{\sqrt{\tau}}\right)d\tau \qquad \text{and} \qquad dq_2 = \frac{1}{2}\left(-\frac{a}{\tau\sqrt{\tau}} - \frac{b}{\sqrt{\tau}}\right)d\tau$$

$$\frac{a^2}{\tau} + b^2\tau = q_1^2 - 2ab = q_2^2 + 2ab$$

With $a = x/(2\alpha)$ and $b = \sqrt{\sigma}$ and after some algebraic manipulations, we get

$$\begin{aligned}\mathcal{I}(x,t) &= -\frac{2}{a}\,e^{2ab}\int_{\infty}^{q_1(t)} e^{-q_1^2}\,dq_1 - bg(x,t)\\ \mathcal{I}(x,t) &= -\frac{2}{a}\,e^{-2ab}\int_{\infty}^{q_2(t)} e^{-q_2^2}\,dq_2 + bg(x,t)\end{aligned}$$

where

$$g(x,t) = \int_0^t \frac{1}{\sqrt{\tau}}\exp\left(-\frac{x^2}{4\alpha^2\tau} - \sigma\tau\right)d\tau$$

The integral $g(x,t)$ is just as difficult to integrate as $\mathcal{I}(x,t)$. Fortunately, we avoid this by adding the two forms of $\mathcal{I}(x,t)$ based on q_1 and q_2 to obtain

$$\mathcal{I}(x,y) = \frac{\alpha\sqrt{\pi}}{x}\left[e^{2ab}\text{erfc}\left(\frac{x}{2\alpha\sqrt{\tau}} + \sigma\sqrt{\tau}\right) + e^{-2ab}\text{erfc}\left(\frac{x}{2\alpha\sqrt{\tau}} - \sigma\sqrt{\tau}\right)\right]$$

or

$$\begin{aligned}u(x,y) &= \frac{C_0 - C_i}{2}\left[e^{x\sqrt{\sigma}/\alpha}\text{erfc}\left(\frac{x}{2\alpha\sqrt{t}} + \sqrt{\sigma t}\right)\right.\\ &\qquad \left. + e^{-x\sqrt{\sigma}/\alpha}\text{erfc}\left(\frac{x}{2\alpha\sqrt{t}} - \sqrt{\sigma t}\right)\right] + C_i \qquad \text{(L.155)}\end{aligned}$$

A plot of (L.155) with $C_0 = 1$, $C_i = 0$, $\alpha = 1$, and $\sigma = 2$ is shown in Figure L.22.

Case 2. $f(t)$ **not constant.** In the general case that $f(t)$ is not constant, numerical integration is more appropriate. However, because of the presence of $\sqrt{\tau^3}$ in the denominator of the integrand in (L.154), a removable singularity occurs at $\tau = 0$. The neighborhood around this singularity remains difficult to evaluate with

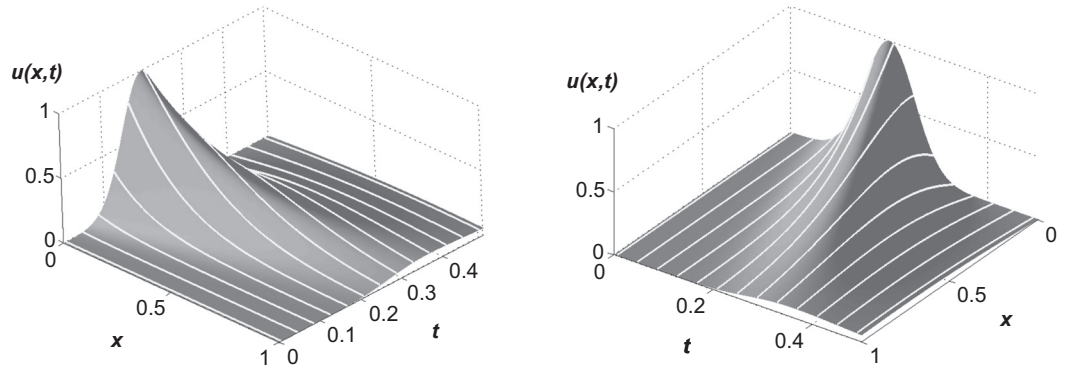


Figure L.23. A plot of the solution given by (L.156) in two perspectives.

acceptable precision. As in the previous case, an auxiliary variable is needed. This time, we introduce p where

$$p(\tau) = \frac{1}{\sqrt{\tau}}$$

whose differential is

$$dp = -\frac{1}{2\tau\sqrt{\tau}} d\tau$$

Then (L.154) becomes

$$u(x,t) = -\frac{x}{\sqrt{\pi}\alpha} \int_{p(t)}^{\infty} \left[f\left(t - \frac{1}{p^2}\right) - C_i \right] \exp\left(-\left(\frac{xp}{2\alpha}\right)^2 - \frac{\sigma}{p^2} \right) dp + C_i \quad \text{(L.156)}$$

Take as an example, $f(t)$ as a Gaussian function given by

$$f(t) = e^{-200(t-0.2)^2}$$

With $C_i = 0$, $\alpha = 1$ and $\sigma = 2$, a plot of (L.156) can be obtained via numerical integration and is shown in Figure L.23.

L.8 Proofs of Theorems Used in Distribution Theory

PROOF OF THEOREM L.9. The first result, (L.76), comes from direct application of the formula for derivatives of distributions, (L.75), on the linear combination operation given in (L.67).

For (L.77),

$$\begin{aligned}
\left\langle g(t)\frac{d}{dt}\text{Dist}(t), \varphi(t) \right\rangle &= \left\langle \frac{d}{dt}\text{Dist}(t), g(t)\varphi(t) \right\rangle \\
&= -\left\langle \text{Dist}(t), \frac{d}{dt}(g(t)\varphi(t)) \right\rangle \\
&= -\left\langle \text{Dist}(t), g(t)\frac{d\varphi}{dt} \right\rangle - \left\langle \text{Dist}(t), \varphi(t)\frac{dg}{dt} \right\rangle \\
&= \left\langle \frac{d}{dt}[g(t)\text{Dist}(t)], \varphi(t) \right\rangle - \left\langle \left[\frac{dg}{dt}\right] \text{Dist}(t), \varphi(t) \right\rangle
\end{aligned}$$

After rearranging the equation, we arrive at (L.77).

To obtain (L.79),

$$\begin{aligned}
\left\langle \frac{d}{dt}\left[\textsc{Dist}\left(\vartheta(t)\right)\right], \varphi(t)\right\rangle &= -\int_{-\infty}^{\infty} \textsc{Dist}\left(\vartheta\right) \frac{d\varphi}{dt} dt \\
&= -\int_{-\infty}^{\infty} \textsc{Dist}\left(\vartheta\right) \frac{d\varphi}{dt}\left[\frac{dt}{d\vartheta}\frac{d\vartheta}{dt}\right] dt \\
&= -\int_{\vartheta(-\infty)}^{\vartheta(\infty)} \textsc{Dist}\left(\vartheta\right) \frac{d\varphi(t)}{d\vartheta} d\vartheta \\
&= \int_{\vartheta(-\infty)}^{\vartheta(\infty)} \left(\frac{d}{d\vartheta}\textsc{Dist}\left(\vartheta\right)\right) \varphi(t) d\vartheta \\
&= \int_{-\infty}^{\infty} \left(\frac{d}{d\vartheta}\textsc{Dist}\left(\vartheta\right)\right) \varphi(t) \frac{d\vartheta}{dt} dt \\
&= \left\langle \left(\frac{d\vartheta}{dt}\right) \frac{d}{d\vartheta}\left[\textsc{Dist}\left(\vartheta\right)\right], \varphi(t)\right\rangle
\end{aligned}$$

PROOF OF EQUATION (L.83). First, we prove the case when $n = m$. Using integration by parts,

$$\begin{aligned}
&\left\langle t^n \frac{d^n}{dt^n}\delta\left(t\right), \varphi(t)\right\rangle \\
&\quad = (-1)^n \int_{-\infty}^{\infty} \delta\left(t\right)\left[\sum_{i=0}^{n}\binom{n}{i}\left(\frac{d^{(n-i)}}{dt^{(n-i)}}t^n\right)\left(\frac{d^i}{dt^i}\varphi\right)\right] dt \\
&\quad = (-1)^n n! \int_{-\infty}^{\infty} \delta\left(t\right)\varphi(t) dt + (-1)^n \int_{-\infty}^{\infty} \delta\left(t\right)\left[\sum_{i=1}^{n}\binom{n}{i}\left(\frac{n!}{i!}t^i\right)\left(\frac{d^i}{dt^i}\varphi\right)\right] dt \\
&\quad = \left\langle (-1)^n n!\delta\left(t\right), \varphi(t)\right\rangle
\end{aligned}$$

Thus

$$t^n \frac{d^n}{dt^n}\delta\left(t\right) = (-1)^n n!\delta\left(t\right)$$

Let $\ell > 0$ then

$$\begin{aligned}
\left\langle t^\ell t^n \frac{d^n}{dt^n}\delta\left(t\right), \varphi(t)\right\rangle &= (-1)^n n!\left\langle \delta\left(t\right), t^\ell \varphi(t)\right\rangle \\
&= 0
\end{aligned}$$

Thus

$$t^n \frac{d^m}{dt^m}\delta\left(t\right) = 0 \quad \text{if} \quad 0 \le m < n$$

Finally, for the case $n \leq m$, we apply the induction process. Let $m = n + k$

$$\left\langle t^n \frac{d^{(n+k)}}{dt^{(n+k)}}\delta(t), \varphi(t)\right\rangle$$

$$= (-1)^n \int_{-\infty}^{\infty} \frac{d^k}{dt^k}\delta(t)\left[\sum_{i=0}^{n}\binom{n}{i}\left(\frac{d^{(n-i)}}{dt^{(n-i)}}x^n\right)\left(\frac{d^i}{dt^i}\varphi\right)\right]dt$$

$$= (-1)^n \int_{-\infty}^{\infty} \sum_{i=0}^{n}\binom{n}{i}\left(\frac{n!}{i!}t^i\right)\left(\frac{d^k}{dt^k}\delta(t)\right)\left(\frac{d^i}{dt^i}\varphi\right)dt$$

(using induction at this point)

$$= (-1)^n \int_{-\infty}^{\infty} \sum_{i=0}^{\min(n,k)} \frac{(-1)^i (n!)^2 k!}{(i!)^2 (n-i)!(k-i)!}\left(\frac{d^{(k-i)}}{dt^{(k-i)}}\delta(t)\right)\left(\frac{d^i}{dt^i}\varphi\right)dt$$

$$= (-1)^n \int_{-\infty}^{\infty} \frac{(n+k)!}{k!}\left(\frac{d^k}{dt^k}\delta(t)\right)\varphi(t)dt$$

$$= (-1)^n \left\langle \frac{(n+k)!}{k!}\frac{d^k}{dt^k}\delta(t), \varphi(t)\right\rangle$$

Thus

$$t^n \frac{d^m}{dt^m}\delta(t) = (-1)^n \frac{m!}{(m-n)!}\frac{d^{(m-n)}}{dt^{(m-n)}}\delta(t) \quad \text{if} \quad 0 \leq n \leq m$$

PROOF OF EQUATION (L.87). In (L.69), we required that the argument transformation $\vartheta(t)$ be monotonic and invertible; thus we cannot immediately apply that result for the more general requirements for $g(t)$. Nonetheless, we can take advantage of the fact that the delta distribution is mostly zero except at the roots of it arguments, that is, $\delta(g(t)) = 0$ when $g(t) \neq 0$. This allows us to partition the path of integration to smaller segments surrounding the roots of $g(t)$,

$$\begin{aligned}\langle \delta(g(t)), \varphi(t)\rangle &= \int_{-\infty}^{\infty} \delta(g(t))\varphi(t)dt \\ &= \sum_{k=1}^{N}\int_{r_k-\epsilon}^{r_k+\epsilon} \delta(g(t))\varphi(t)dt\end{aligned}$$

where $\epsilon > 0$ is small enough such that $g(t)$ is monotonic and invertible in the range $(r_k - \epsilon) \leq t \leq (r_k + \epsilon)$ for all k.

We can now apply an equation similar to (L.69) for each integral term,

$$\begin{aligned}\int_{r_k-\epsilon}^{r_k+\epsilon}\delta\left(g(t)\right)\varphi(t)dt &= \int_{g(r_k-\epsilon)}^{g(r_k+\epsilon)}\delta\left(z\right)\frac{\varphi\left(g^{-1}(z)\right)}{|dg/dt|_{(g^{-1}(z))}}dz\\ &= \frac{\varphi\left(r_k\right)}{|dg/dt|_{t=r_k}}\\ &= \frac{1}{|dg/dt|_{t=r_k}}\left\langle\delta\left(t-r_k\right),\varphi(t)\right\rangle\end{aligned}$$

Combining both results,

$$\begin{aligned}\left\langle\delta\left(g(t)\right),\varphi(t)\right\rangle &= \sum_{k=1}^{N}\frac{1}{|dg/dt|_{t=r_k}}\left\langle\delta\left(t-r_k\right),\varphi(t)\right\rangle\\ &= \left\langle\sum_{k=1}^{N}\left[\frac{1}{|dg/dt|_{t=r_k}}\delta\left(t-r_k\right)\right],\varphi(t)\right\rangle\end{aligned}$$

PROOF OF THEOREM L.10. Using (L.72),

$$\left\langle F(\alpha,t),\varphi(t)\right\rangle=\alpha\left\langle f(t),\varphi\left(\frac{t}{\alpha}\right)\right\rangle$$

then

$$\begin{aligned}\left\langle F(\alpha,t)-\delta\left(t\right),\varphi(t)\right\rangle &= \left\langle f(t),\varphi\left(\frac{t}{\alpha}\right)\right\rangle-\varphi(0)\\ &\quad\left(\text{ then with }\int_{-\infty}^{\infty}f(t)dt=1\right)\\ &= \int_{-\infty}^{\infty}f(t)\Phi\left(\frac{t}{\alpha}\right)dt\end{aligned}$$

where

$$\Phi(t)=\varphi(t)-\varphi(0)$$

Taking absolute values of both sides, we obtain the following inequality,

$$\left|\left\langle F(\alpha,t)-\delta\left(t\right),\varphi(t)\right\rangle\right|\leq A+B$$

where,

$$\begin{aligned}A &= \left|\int_{-\infty}^{-q}f(t)\Phi\left(\frac{t}{\alpha}\right)dt+\int_{q}^{\infty}f(t)\Phi\left(\frac{t}{\alpha}\right)dt\right|\\ B &= \left|\int_{-q}^{q}f(t)\Phi\left(\frac{t}{\alpha}\right)dt\right|\end{aligned}$$

Now choose $\kappa>0$, $q>0$ and $\alpha>q/\kappa$ such that

1. $|\Phi(t)|<\varepsilon$ for $|t|<\kappa$
2. $\int_{-\infty}^{-q}|f(t)|dt+\int_{q}^{\infty}|f(t)|dt<\varepsilon$

then

$$\begin{aligned} A &\leq 2\varepsilon \cdot \max_t |\varphi(t)| \\ B &\leq \varepsilon \cdot \left| \int_{-\infty}^{\infty} f(t)dt \right| \end{aligned}$$

or

$$\left| \langle F(\alpha, t) - \delta(t), \varphi(t) \rangle \right| \leq \varepsilon \left(2 \max_t |\varphi(t)| + \left| \int_{-\infty}^{\infty} f(t)dt \right| \right)$$

Because all the terms on the right-hand side of the inequality is fixed except for ε, we can choose ε arbitrarily small. Hence

$$\lim_{\alpha \to \infty} F(\alpha, t) = \delta(t)$$

APPENDIX M

Additional Details and Fortification for Chapter 13

M.1 Method of Undetermined Coefficients for Finite Difference Approximation of Mixed Partial Derivative

For the case of mixed partial derivatives, we use the general formula of the Taylor series of $u(x, t)$ expanded around (x_k, t_q):

$$u_{k+j}^{(q+i)} = \sum_{m=0}^{\infty} \sum_{\ell=0}^{\infty} f_{m,\ell}\, \widehat{\gamma}_{\ell,j}^{m,i} \tag{M.1}$$

where

$$f_{m,\ell} = \frac{\partial^{m+\ell}}{\partial t^m \partial x^\ell}\, u \bigg|_{(x_k,t_q)} \Delta t^m \Delta x^\ell \qquad \text{and} \qquad \widehat{\gamma}_{\ell,j}^{m,i} = \gamma_{\ell,j}\, \gamma_{m,i}$$

and $\gamma_{\ell,j}$ has been defined in (13.11).

The computation for the general case involves high-order tensorial sums. A simple example is the approximation of $\dfrac{\partial^2 u}{\partial x \partial t}$.

EXAMPLE M.1. Approximation of Mixed Partial Derivatives. Let $D_{1,x,1,y}$ be the approximation of the mixed derivative defined as a linear combination of the values at neighboring points,

$$D_{1,x,1,y} = \frac{1}{\Delta x \Delta y} \sum_{i=-1}^{1} \sum_{j=-1}^{1} u_{k+i,n+j}\, \alpha_{i,j} = \frac{\partial^2 u}{\partial x \partial y}\bigg|_{(x_k,y_n)} + \text{Error} \tag{M.2}$$

Applying (M.1), we obtain

$$\begin{aligned} \left(\sum_{\ell=0}^{2} \sum_{m=0}^{2} f_{m,\ell} \sum_{i=-1}^{1} \sum_{j=-1}^{1} \widehat{\gamma}_{\ell,j}^{m,i}\, \alpha_{i,j} \right) - f_{1,1} \;&=\; \Delta t \Delta x\, (\text{Error}) \\ &\quad - \left(\sum_{\ell=3}^{\infty} \sum_{m=0}^{\infty} f_{m,\ell} \sum_{i=-1}^{1} \sum_{j=-1}^{1} \widehat{\gamma}_{\ell,j}^{m,i}\, \alpha_{i,j} \right) \\ &\quad - \left(\sum_{\ell=0}^{\infty} \sum_{m=3}^{\infty} f_{m,\ell} \sum_{i=-1}^{1} \sum_{j=-1}^{1} \widehat{\gamma}_{\ell,j}^{m,i}\, \alpha_{i,j} \right) \end{aligned} \tag{M.3}$$

Setting the left-hand side (M.3) equal to 0 results in a set of nine independent linear equations:

$$\widehat{\gamma}_{\ell,j}^{m,i}\,\alpha_{i,j} = \begin{cases} 0 & \text{if } (i,j) \neq (1,1) \\ 1 & \text{if } (i,j) = (1,1) \end{cases}$$

Solving these equations, we obtain

$$\alpha_{i,j} = \frac{ij}{4} \qquad i,j = -1,0,1$$

which yields the following finite difference approximation of the mixed partial derivative:

$$\left.\frac{\partial^2 u}{\partial x \partial y}\right|_{(x_k,y_n)} \approx \frac{1}{4\Delta x \Delta y}\left(u_{n+1}^{(k+1)} - u_{n-1}^{(k+1)} - u_{n+1}^{(k-1)} + u_{n-1}^{(k-1)}\right) \tag{M.4}$$

To determine the order of truncation error, note that the coefficients of the lower order terms of $f_{m,\ell}$ are

$$\sum_{i=-1}^{1}\sum_{j=-1}^{1}\widehat{\gamma}_{\ell,j}^{m,i}\,\alpha_{i,j} = \begin{cases} 0 & \text{if} \quad l=0, \quad \text{or } m=0 \\ \dfrac{1}{2m!}\left(1+(-1)^{m+1}\right) & \text{if} \quad l=1 \\ \dfrac{1}{2\ell!}\left(1+(-1)^{\ell+1}\right) & \text{if} \quad m=1 \end{cases}$$

yielding

$$\text{Error} = \frac{\Delta t^2}{3!}\left(\frac{\partial^4}{\partial t^3 \partial x^1}\,u\right)\Big|_{(x,t)} + \frac{\Delta x^2}{3!}\left(\frac{\partial^4}{\partial t^1 \partial x^3}\,u\right)\Big|_{(x,t)} + \cdots$$

or $\text{Error} = \mathcal{O}\left(\Delta t^2, \Delta x^2\right)$.

M.2 Finite Difference Formulas for 3D Cases

For the 3D, time-invariant case, a general second-order linear differential equation is given by

$$\begin{aligned} &\mu_{xx}(x,y,z)\frac{\partial^2 u}{\partial x^2} + \mu_{yy}(x,y,z)\frac{\partial^2 u}{\partial y^2} + \mu_{zz}(x,y,z)\frac{\partial^2 u}{\partial z^2} \\ &+ \mu_{xy}(x,y,z)\frac{\partial^2 u}{\partial x \partial y} + \mu_{yz}(x,y,z)\frac{\partial^2 u}{\partial y \partial z} + \mu_{xz}(x,y,z)\frac{\partial^2 u}{\partial x \partial z} \\ &+ \beta_x(x,y,z)\frac{\partial u}{\partial x} + \beta_y(x,y,z)\frac{\partial u}{\partial y} + \beta_z(x,y,z)\frac{\partial u}{\partial z} \\ &+ \zeta(x,y,z)u + \eta(x,y,z) \quad = \quad 0 \end{aligned} \tag{M.5}$$

Let the superscript "(3b)" denote matrix augmentation that will flatten the 3D tensor into a matrix representation. For instance, for $k = 1,\ldots,K, n = 1,\ldots,N$, and

$m = 1, \ldots, M$, we have the following $K \times NM$ matrix for u_{knm}:

$$U^{(3\flat)} = \left(\begin{array}{ccc|c|ccc} u_{1,1,1} & \cdots & u_{1,N,1} & & u_{1,1,M} & \cdots & u_{1,N,M} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ u_{K,1,1} & \cdots & u_{K,N,1} & & u_{K,1,M} & \cdots & u_{K,N,M} \end{array}\right)$$

$$\underline{\zeta}^{(3\flat)} = \left(\begin{array}{ccc|c|ccc} \zeta_{1,1,1} & \cdots & \zeta_{1,N,1} & & \zeta_{1,1,M} & \cdots & \zeta_{1,N,M} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \zeta_{K,1,1} & \cdots & \zeta_{K,N,1} & & \zeta_{K,1,M} & \cdots & \zeta_{K,N,M} \end{array}\right)$$

etc.

where $u_{k,n,m} = u(k\Delta x, n\Delta y, m\Delta z)$, $\zeta_{k,n,m} = \zeta(k\Delta x, n\Delta y, m\Delta z)$, etc. Likewise, let the superscripts "$(2\flat, x)$" and "$(2\flat, y)$" denote column augmentation, as the indices are incremented along the x and y directions, respectively. For instance,

$$\mathfrak{b}_{(1,z)}^{(2\flat,x)} = \left(\left.\mathfrak{b}_{(1,z)}\right|_{k=1} \;\middle|\; \cdots \;\middle|\; \left.\mathfrak{b}_{(1,z)}\right|_{k=K} \right)$$

The partial derivatives can then be approximated by finite difference approximations in matrix forms as follows:

$$\frac{\partial u}{\partial x} \to \mathcal{D}_{(1,x)} U^{(3\flat)} + \mathcal{B}_{(1,x)}^{(3\flat)} \quad ; \quad \frac{\partial^2 u}{\partial x^2} \to \mathcal{D}_{(2,x)} U^{(3\flat)} + \mathcal{B}_{(2,x)}^{(3\flat)}$$

$$\frac{\partial u}{\partial y} \to U^{(3\flat)} \left(I_M \otimes \mathcal{D}_{(1,y)}^T\right) + \left(\mathcal{B}_{(1,y)}^T\right)^{(3\flat)} \quad ; \quad \frac{\partial^2 u}{\partial y^2} \to U^{(3\flat)} \left(I_M \otimes \mathcal{D}_{(2,y)}^T\right) + \left(\mathcal{B}_{(2,y)}^T\right)^{(3\flat)}$$

$$\frac{\partial u}{\partial z} \to U^{(3\flat)} \left(\mathcal{D}_{(1,z)}^T \otimes I_N\right) + \widehat{\mathcal{B}}_{(1,z)} \quad ; \quad \frac{\partial^2 u}{\partial z^2} \to U^{(3\flat)} \left(\mathcal{D}_{(2,z)}^T \otimes I_N\right) + \widehat{\mathcal{B}}_{(2,z)}$$

$$\begin{aligned} \frac{\partial^2 u}{\partial x \partial y} &\to \mathcal{D}_{(1,x)} U^{(3\flat)} \left(I_M \otimes \mathcal{D}_{(1,y)}^T\right) + \mathcal{B}_{(1,x,1,y)}^{(3\flat)} \\ \frac{\partial^2 u}{\partial y \partial z} &\to U^{(3\flat)} \left(\mathcal{D}_{(1,z)}^T \otimes \mathcal{D}_{(1,y)}^T\right) + \widehat{\mathcal{B}}_{(1,y,1,z)} \\ \frac{\partial^2 u}{\partial x \partial z} &\to \mathcal{D}_{(1,x)} U^{(3\flat)} \left(\mathcal{D}_{(1,z)}^T \otimes I_N\right) + \widehat{\mathcal{B}}_{(1,x,1,z)} \end{aligned}$$

where $\mathcal{D}_{(1,x)}, \mathcal{D}_{(1,y)}, \mathcal{D}_{(1,z)}, \mathcal{D}_{(2,x)}, \mathcal{D}_{(2,y)}$, and $\mathcal{D}_{(2,z)}$ are matrices that can take forms such as those given in Section 13.2.2 depending on order of approximation and boundary conditions. The matrices $\mathcal{B}_{(1,x)}$, $\mathcal{B}_{(2,x)}$, ..., and so forth contain the boundary data. The new matrices $\widehat{\mathcal{B}}_{1,z}$ and $\widehat{\mathcal{B}}_{2,z}$ are given by a sequence of transformations as

$$\widehat{\mathcal{B}}_{(1,z)} = \text{reshape}\left(\left[\left(\mathfrak{b}_{(1,z)}^{(2\flat,x)}\right)^{(2\flat,y)}\right]^T,\ K,\ NM\right) \tag{M.6}$$

$$\widehat{\mathcal{B}}_{(2,z)} = \text{reshape}\left(\left[\left(\mathfrak{b}_{(2,z)}^{(2\flat,x)}\right)^{(2\flat,y)}\right]^T,\ K,\ NM\right) \tag{M.7}$$

(The matrices $\widehat{\mathcal{B}}_{1,x,1,z}$ and $\widehat{\mathcal{B}}_{1,y,1,z}$ are left as exercises.)

Just as in the previous section, we can use the properties of matrix vectorizations to obtain the following linear equation problem corresponding to (M.5):

$$\mathfrak{R}_{3\mathrm{D}}\,\mathrm{vec}\left(U^{(3\flat)}\right) = \mathfrak{f}_{3\mathrm{D}} \tag{M.8}$$

where,

$$\begin{aligned}
\mathfrak{R}_{3\mathrm{D}} \;=\; & \left[\underline{\underline{\mu_{xx}}}^{(3\flat)}\right]^{\mathrm{dv}} I_M \otimes I_N \otimes \mathcal{D}_{(2,x)} \;+\; \left[\underline{\underline{\mu_{yy}}}^{(3\flat)}\right]^{\mathrm{dv}} I_M \otimes \mathcal{D}_{(2,y)} \otimes I_K \\
& \qquad + \left[\underline{\underline{\mu_{zz}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathcal{D}_{(2,z)} \otimes I_N \otimes I_K \\
& + \left[\underline{\underline{\mu_{xy}}}^{(3\flat)}\right]^{\mathrm{dv}} I_M \otimes \mathcal{D}_{(1,y)} \otimes \mathcal{D}_{(1,x)} \;+\; \left[\underline{\underline{\mu_{yz}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathcal{D}_{(1,z)} \otimes \mathcal{D}_{(1,y)} \otimes I_K \\
& \qquad + \left[\underline{\underline{\mu_{xz}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathcal{D}_{(1,z)} \otimes I_N \otimes \mathcal{D}_{(1,x)} \\
& + \left[\underline{\underline{\beta_x}}^{(3\flat)}\right]^{\mathrm{dv}} I_M \otimes I_N \otimes \mathcal{D}_{(1,x)} \;+\; \left[\underline{\underline{\beta_y}}^{(3\flat)}\right]^{\mathrm{dv}} I_M \otimes \mathcal{D}_{(1,y)} \otimes I_K \\
& \qquad + \left[\underline{\underline{\beta_z}}^{(3\flat)}\right]^{\mathrm{dv}} \mathcal{D}_{(1,z)} \otimes I_N \otimes I_K \\
& + \left[\underline{\zeta}^{(3\flat)}\right]^{\mathrm{dv}}
\end{aligned}$$

$$\begin{aligned}
\mathfrak{f}_{3\mathrm{D}} \;=\; & \left[\underline{\underline{\mu_{xx}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\mathcal{B}^{(3\flat)}_{(2,x)}\right) \;+\; \left[\underline{\underline{\mu_{yy}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\left(\mathcal{B}_{(2,y)^T}\right)^{(3\flat)}\right) \\
& \qquad + \left[\underline{\underline{\mu_{zz}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\widehat{\mathcal{B}}_{(2,z)}\right) \\
& + \left[\underline{\underline{\mu_{xy}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\mathcal{B}^{(3\flat)}_{(1,x,1,y)}\right) \;+\; \left[\underline{\underline{\mu_{yz}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\widehat{\mathcal{B}}_{(1,y,1,z)}\right) \\
& + \left[\underline{\underline{\mu_{xz}}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\widehat{\mathcal{B}}_{(1,x,1,z)}\right) \\
& + \left[\underline{\underline{\beta_x}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\mathcal{B}^{(3\flat)}_{(1,x)}\right) \;+\; \left[\underline{\underline{\beta_y}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\left(\mathcal{B}^{T}_{(1,y)}\right)^{(3\flat)}\right) \\
& \qquad + \left[\underline{\underline{\beta_z}}^{(3\flat)}\right]^{\mathrm{dv}} \mathrm{vec}\left(\widehat{\mathcal{B}}_{(1,z)}\right) \\
& + \mathrm{vec}\left(\underline{\eta}^{(3\flat)}\right)
\end{aligned}$$

EXAMPLE M.2. Consider the 3D Poisson equation

$$\nabla^2 u = \eta(x, y, z) \qquad 0 \le x, y, z \le 1 \tag{M.9}$$

where,

$$\eta(x,y,z) = \exp\left(-2\left[z-x\right]^2 - 5\left[1-\frac{4}{5}z-y\right]^2\right)$$
$$\times\left[-\frac{122}{5}+16\left(x-z\right)^2+100\left(1-\frac{4}{5}z-y\right)^2+\left(8-\frac{52}{5}z-8y+4x\right)^2\right]$$

subject to the following six Dirichlet boundary conditions:

$$\begin{aligned}
u\left(0,y,z\right) = \alpha_0(y,z) &= \exp\left(-2z^2 - 5\left[1-\frac{4}{5}z-y\right]^2\right)\\
u\left(1,y,z\right) = \alpha_1(y,z) &= \exp\left(-2\left[z-1\right]^2 - 5\left[1-\frac{4}{5}z-y\right]^2\right)\\
u\left(x,0,z\right) = \beta_0(x,z) &= \exp\left(-2\left[z-x\right]^2 - 5\left[1-\frac{4}{5}z\right]^2\right)\\
u\left(x,1,z\right) = \beta_1(x,z) &= \exp\left(-2\left[z-x\right]^2 - \frac{16}{5}z^2\right)\\
u\left(x,y,0\right) = \gamma_0(x,y) &= \exp\left(-2x^2 - 5\left[1-y\right]^2\right)\\
u\left(x,y,1\right) = \gamma_1(x,y) &= \exp\left(-2\left[1-x\right]^2 - 5\left[\frac{1}{5}-y\right]^2\right) \qquad \text{(M.10)}
\end{aligned}$$

The exact solution is given by

$$u\left(x,y,z\right) = \exp\left(-2\left[z-x\right]^2 - 5\left[1-\frac{4}{5}z-y\right]^2\right) \qquad \text{(M.11)}$$

Using $\Delta x = \Delta y = \Delta z = 0.05$, and central difference formulas for $\mathcal{D}_{(2,x)}$, $\mathcal{D}_{(2,y)}$, $\mathcal{D}_{(2,y)}$, $\mathcal{B}_{(2,x)}$, $\mathcal{B}_{(2,y)}$, and $\widehat{\mathcal{B}}_{(2,z)}$, the linear equation (M.8) can be solved for $\text{vec}\left(U^{(3b)}\right)$. The results are shown in Figure M.1 at different values of z, where the approximations are shown as points, whereas the exact solutions are shown as surface plots. (A MATLAB file `poisson_3d.m` is available on the book's webpage that implements the finite difference solution and obtains the plots shown in this example.) The errors from the exact solution (M.11) are shown in Figure M.2 at different fixed values of z. The errors are in the range $\pm 1.7 \times 10^{-3}$.

M.3 Finite Difference Solutions of Linear Hyperbolic Equations

Consider the following linear hyperbolic equations

$$\frac{\partial}{\partial t}\widehat{\mathbf{u}} + A\frac{\partial}{\partial x}\widehat{\mathbf{u}} = \widehat{\mathbf{c}} \qquad \text{(M.12)}$$

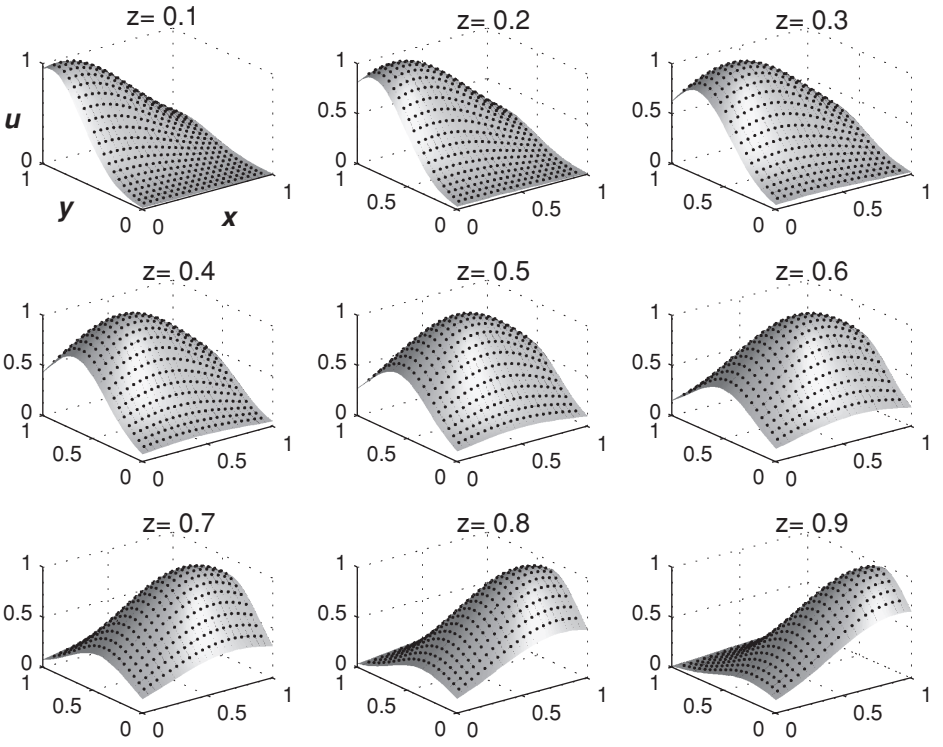


Figure M.1. The finite difference solution to (M.9) at different values of z, subject to conditions (M.10). The approximations are shown as points, whereas the exact solutions, (M.11), at the corresponding z values are shown as surface plots.

where $\widehat{\mathbf{u}} = (\widehat{u}_1, \ldots, \widehat{u}_J)^T$ and A is a constant $J \times J$ matrix. If A is diagonalizable, that is, there exist a nonsingular matrix V and a diagonal matrix Λ such that $A = V\Lambda V^{-1}$, then with

$$\mathbf{u} = V^{-1}\widehat{\mathbf{u}} \qquad \mathbf{c} = V^{-1}\widehat{\mathbf{c}}$$

we can decouple (M.12) into J equations

$$\frac{\partial u_i}{\partial t} + \lambda_i \frac{\partial u_i}{\partial x} = c_i$$

Thus in the discussion that follows, we consider

$$\frac{\partial u}{\partial t} + \lambda \frac{\partial u}{\partial x} = c \tag{M.13}$$

as a representative system for handling a system of first-order hyperbolic equations. However, in our discussion of the scalar case, we allow for $c = c(x, t)$.

M.3.1 Upwind Schemes

We can use either forward, backward, or central difference approximations for $\partial u/\partial x$ toward a semi-discrete approach. Time marching can then be implemented by a forward Euler or backward Euler. This will yield six types of schemes, namely

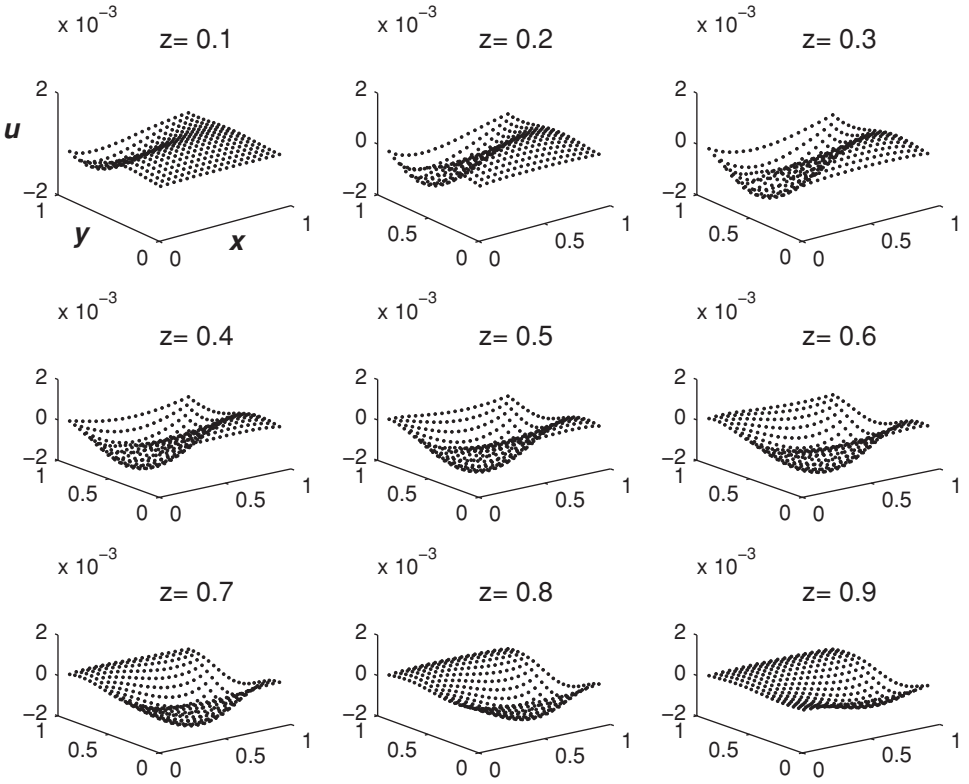


Figure M.2. The error distribution between the finite difference approximation (using central difference formulas) and the exact solutions, (M.11), at different z values.

forward-time-forward-space (FTFS), forward-time-central-space (FTCS), forward-time-backward-space (FTBS), backward-time-forward-space (BTFS), backward-time-central-space (BTCS), and backward-time-backward-space (BTBS). Each scheme will have different stability ranges for Δt in relation to Δx and λ. In Table M.1, we summarize the different upwind schemes and their stability based on another parameter η

$$\eta = \lambda \frac{\Delta t}{\Delta x} \tag{M.14}$$

which is known as the **Courant number**. The stability conditions included in the table are obtained using the von Neumann method and are given as Exercise in **E13.15**.

We can make the following observations:

1. The forward-time schemes: FTFS, FTCS, and FTBS are explicit schemes, whereas the backward-time schemes: BTFS, BTCS and BTBS are implicit schemes.
2. The central-space schemes are given by FTCS and BTCS, with the explicit FTCS being unstable and the implicit BTCS being unconditionally stable.
3. The noncentral space schemes have their stability dependent on the sign of η, or equivalently on the sign of λ. Both forward-space schemes, FTFS and BTFS, are

Table M.1. *Basic finite difference schemes for scalar hyperbolic equations*

Scheme	Approximation equation	Stability region
FTFS	$u_k^{(q+1)} = (1+\eta)\,u_k^{(q)} - \eta u_{k+1}^{(q)} + \Delta t c_k^{(q)}$	$-1 \le \eta \le 0$
FTCS	$u_k^{(q+1)} = u_k^{(q)} - \frac{\eta}{2}\left(u_{k+1}^{(q)} - u_{k-1}^{(q)}\right) + \Delta t c_k^{(q)}$	None
FTBS	$u_k^{(q+1)} = \eta u_{k-1}^{(q)} + (1-\eta)\,u_k^{(q)} + \Delta t c_k^{(q)}$	$0 \le \eta \le 1$
BTFS	$(1-\eta)\,u_k^{(q+1)} + \eta u_{k+1}^{(q+1)} + \Delta t c_k^{(q+1)} = u_k^{(q)}$	$\eta \le 0$
BTCS	$u_k^{(q+1)} + \frac{\eta}{2}\left(u_{k+1}^{(q+1)} - u_{k-1}^{(q+1)}\right) + \Delta t c_k^{(q+1)} = u_k^{(q)}$	All
BTBS	$(1+\eta)\,u_k^{(q+1)} - \eta u_{k-1}^{(q+1)} + \Delta t c_k^{(q+1)} = u_k^{(q)}$	$\eta \ge 0$
Leapfrog	$u_k^{(q+1)} = u_k^{(q-1)} - \eta u_{n+1}^{(q)} + \eta u_{n-1}^{(q)} + 2\Delta t c_k^{(q)}$	$\lvert\eta\rvert \le 1$
Lax-Friedrichs	$u_k^{(q+1)} = \left(\frac{1-\eta}{2}\right) u_{k+1}^{(q)} + \left(\frac{1+\eta}{2}\right) u_{k-1}^{(q)} + \Delta t c_k^{(q)}$	$\lvert\eta\rvert \le 1$
Lax-Wendroff	$u_k^{(q+1)} = \left(1-\eta^2\right) u_k^{(q)} + \frac{1}{2}\left(\eta^2-\eta\right) u_{k+1}^{(q)} + \frac{1}{2}\left(\eta^2+\eta\right) u_{k-1}^{(q)} + c_k^{(q)}\Delta t + \left(\frac{\partial c}{\partial t} - \lambda\frac{\partial c}{\partial x}\right)_k^{(q)} \Delta t^2$	$\lvert\eta\rvert \le 1$
Crank-Nicholson	$\frac{\eta}{4}u_{k+1}^{(q+1)} + u_k^{(q+1)} - \frac{\eta}{4}u_{k-1}^{(q+1)} = -\frac{\eta}{4}u_{k+1}^{(q)} + u_k^{(q)} + \frac{\eta}{4}u_{k-1}^{(q)} + \Delta t c_k^{(q+1/2)}$	All

stable only for negative η values, whereas both backward-space schemes, FTBS and BTBS, are stable only for positive η values.[1]

From the last observation, we can still recover the use of noncentral schemes by switching between forward-space and backward-space schemes depending on the sign of λ. This combination is called the **upwind schemes**, because the direction of space difference is adjusted to be opposite to the wave speed λ. Specifically, with

$$\eta^{(+)} = \frac{\eta + |\eta|}{2} \qquad \text{and} \qquad \eta^{(-)} = \frac{\eta - |\eta|}{2} \tag{M.15}$$

[1] Note that even though BTFS and BTBS are both implicit schemes, neither are unconditionally stable.

we have the explicit upwind scheme, which combines both FTFS and FTBS in one equation,

$$u_k^{(q+1)} = \left(1 + \eta^{(-)} - \eta^{(+)}\right) u_k^{(q)} + \eta^{(+)} u_{k-1}^{(q)} - \eta^{(-)} u_{k+1}^{(q)} \tag{M.16}$$

and whose stability range is given by $0 < |\eta| < 1$. Likewise, we have the implicit upwind scheme, which combines both BTFS and BTBS in one equation,

$$\left(1 - \eta^{(-)} + \eta^{(+)}\right) u_k^{(q+1)} - \eta^{(+)} u_{k-1}^{(q+1)} + \eta^{(-)} u_{k+1}^{(q+1)} = u_k^{(q)} \tag{M.17}$$

whose stability range is given by $0 < |\eta|$.

M.3.2 Other Finite Difference Schemes

There are four more important schemes: the leapfrog (or CTCS) scheme, the Lax-Friedrichs scheme, the Lax-Wendroff scheme, and Crank-Nicholson scheme. The first three are explicit, whereas the last one is an implicit scheme.

The leapfrog and the Lax-Friedrichs schemes are improvements to the FTCS scheme to overcome its unconditional instability. The **leapfrog scheme** uses the central difference approximation for $\partial u/\partial t$. Thus we have

$$\lambda \frac{u_{k+1}^{(q)} - u_{k-1}^{(q)}}{2\Delta x} + \frac{u_k^{(q+1)} - u_k^{(q-1)}}{2\Delta t} = c_k^{(q)} \tag{M.18}$$

Note that the leapfrog scheme needs values at both t_q and t_{q-1} to obtain values at t_{q+1}. Thus the leapfrog schemes often require other one-step marching, such as Lax-Friedrich or Lax-Wendroff to provide it with values at t_1, and then continue with the leapfrog for t_q, $q \geq 2$.

The **Lax-Friedrichs scheme** approximates the time derivative as a forward time difference, but between $u_k^{(q+1)}$ and the average at the current point, $\frac{1}{2}\left(u_{k+1}^{(q)} + u_{k-1}^{(q)}\right)$. Thus the scheme is given by

$$\lambda \frac{u_{k+1}^{(q)} - u_{k-1}^{(q)}}{2\Delta x} + \frac{u_k^{(q+1)} - \frac{1}{2}\left(u_{k+1}^{(q)} + u_{k-1}^{(q)}\right)}{\Delta t} = c_k^{(q)} \tag{M.19}$$

Note that the leapfrog scheme used the values at t_{q-1}, whereas the Lax-Friedrichs continues to stay within t_q.

The third explicit finite difference scheme uses the Taylor series approximation for u,

$$u_k^{(q+1)} = u_k^{(q)} + \left.\frac{\partial u}{\partial t}\right|_{t=q\Delta t, x=k\Delta x} \Delta t + \frac{1}{2} \left.\frac{\partial^2 u}{\partial t^2}\right|_{t=q\Delta t, x=k\Delta x} \Delta t^2 + \mathcal{O}\left(\Delta t^3\right) \tag{M.20}$$

and then substitutes the following identities obtained from the given differential equation

$$\frac{\partial u}{\partial t} = -\lambda \frac{\partial u}{\partial x} + c \quad \text{and} \quad \frac{\partial^2 u}{\partial t^2} = \lambda^2 \frac{\partial^2 u}{\partial x^2} - \lambda c + \frac{\partial c}{\partial t}$$

into (M.20). Afterward, the central difference approximation is used for $\partial u/\partial x$ and $\partial^2 u/\partial x^2$. After truncation of $\mathcal{O}\left(\Delta t^3\right)$ terms, the following scheme, known as the

Lax-Wendroff scheme, results

$$u_k^{(q+1)} = u_k^{(q)} - \lambda\frac{\Delta t}{2\Delta x}\left(u_{k+1}^{(q)} - u_{k-1}^{(q)}\right) + \frac{1}{2}\lambda^2\frac{\Delta t^2}{\Delta x^2}\left(u_{k+1}^{(q)} - 2u_k^{(q)} + u_{k-1}^{(q)}\right) + c_k^{(q)}\Delta t + \left(\frac{\partial c}{\partial t} - \lambda\frac{\partial c}{\partial x}\right)_k^{(q)}\Delta t^2$$

or

$$u_k^{(q+1)} = \left(1-\eta^2\right)u_k^{(q)} + \frac{1}{2}\left(\eta^2-\eta\right)u_{k+1}^{(q)} + \frac{1}{2}\left(\eta^2+\eta\right)u_{k-1}^{(q)} + c_k^{(q)}\Delta t + \left(\frac{\partial c}{\partial t} - \lambda\frac{\partial c}{\partial x}\right)_k^{(q)}\Delta t^2 \tag{M.21}$$

Using the von Neumann method, one can show that the stability range of the three explicit schemes, namely the leapfrog, Lax-Friedrichs, and Lax-Wendroff schemes, given in (M.18), (M.19), and (M.21), respectively, are all given by $|\eta| \leq 1$. The approximation errors for these methods are $\mathcal{O}\left(\Delta x^2, \Delta t^2\right)$, $\mathcal{O}\left(\Delta x, \Delta t\right)$, and $\mathcal{O}\left(\Delta x^2, \Delta t^2\right)$ for leapfrog, Lax-Friedrichs, and Lax-Wendroff schemes, respectively.

The Crank-Nicholson scheme is an implicit scheme that could be seen as an attempt to improve the accuracy of the BTCS scheme, which may be unconditionally stable but only has approximation errors of $\mathcal{O}\left(\Delta x^2, \Delta t\right)$. However, unlike the leapfrog scheme, where values at t_{q-1} are introduced, this method tries to avoid this from occurring by using a central difference approximation at a point between t_{q+1} and t_q, that is, at $t = t_{q+1/2}$, with a time increment $\Delta t/2$. However, by doing so, the spatial derivative at $t = t_{q+1/2}$ must be estimated by averages. Thus the **Crank-Nicholson scheme** uses the following approximation for the time derivative:

$$\frac{\lambda}{2\Delta x}\left[\frac{u_{k+1}^{(q+1)} + u_{k+1}^{(q)}}{2} - \frac{u_{k-1}^{(q+1)} + u_{k-1}^{(q)}}{2}\right] + \left(\frac{u_k^{(q+1)} - u_k^{(q)}}{2(\Delta t/2)}\right) = c_k^{(q+1/2)}$$

or

$$\frac{\eta}{4}u_{k+1}^{(q+1)} + u_k^{(q+1)} - \frac{\eta}{4}u_{k-1}^{(q+1)} = -\frac{\eta}{4}u_{k+1}^{(q)} + u_k^{(q)} + \frac{\eta}{4}u_{k-1}^{(q)} + \Delta t c_k^{(q+1/2)} \tag{M.22}$$

The approximation error of the Crank-Nicholson scheme is $\mathcal{O}\left(\Delta x^2, \Delta t^2\right)$. Using the von Neumann method, we can show that the Crank-Nicholson scheme, like the BTCS scheme, is unconditionally stable.

EXAMPLE M.3. For the scalar hyperbolic partial differential equation given by

$$\frac{\partial u}{\partial t} + 0.5\frac{\partial u}{\partial x} = 0 \tag{M.23}$$

we consider both a continuous initial condition and a discontinuous initial condition.

1. **Continuous initial condition.** Let initial condition be a Gaussian function given by

$$u(0,t) = e^{-8(5x-1)^2} \tag{M.24}$$

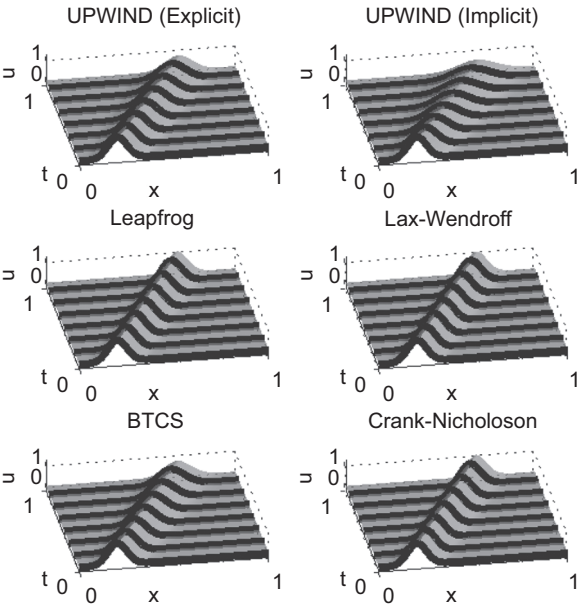


Figure M.3. Numerical solutions for continuous initial condition using the various schemes.

Using the various stable schemes, the finite-difference solutions with $\Delta x = \Delta t = 0.01$ are shown in Figures M.3 and M.4. It appears that the leapfrog, Lax-Wendroff, and Crank-Nicholson schemes yielded good approximations.

2. **Discontinuous initial condition**. Let the initial condition be a square pulse given by

$$u(0,t) = \begin{cases} 1 & \text{if } 0.2 \le x \le 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{M.25}$$

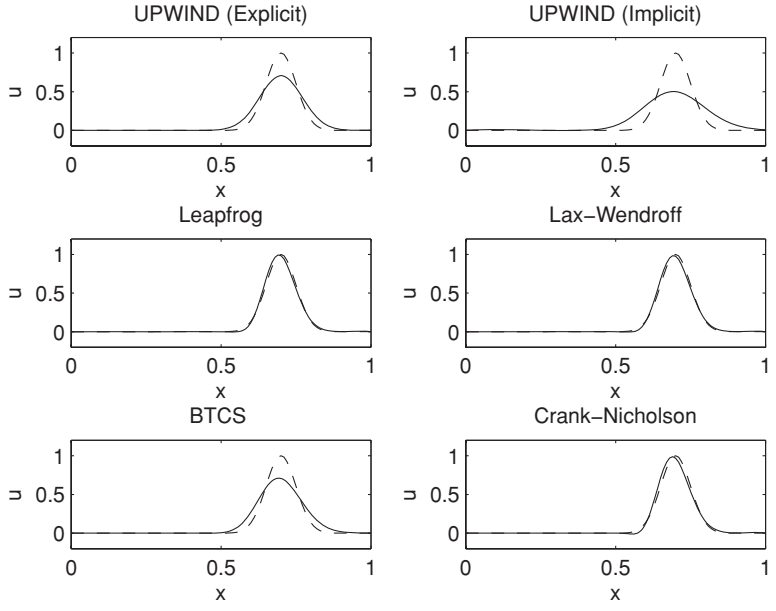


Figure M.4. Comparison with exact solutions for different schemes at $t = 1$. The exact solution is given as dashed lines.

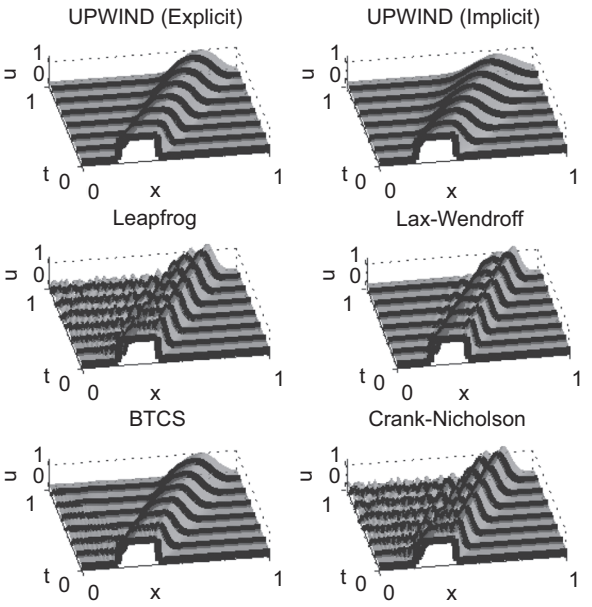


Figure M.5. Numerical solutions for discontinuous initial condition using the various schemes.

Using the various stable schemes, the finite-difference solutions with $\Delta x = \Delta t = 0.01$ are shown in Figures M.5 and M.6. As one can observe from the plots, none of the schemes match the exact solution very well. This is due to numerical dissipation introduce by the schemes. Dissipation was instrumental for stability, but it also smoothed the discontinuity. However, the other schemes had growing amount of oscillations. These are due to the spurious roots of the schemes. Significant amounts of oscillation throughout the spatial domain can be observed in both the leapfrog and Crank-Nicholson schemes. The Lax-Wendroff appears to perform the best; however, a smaller mesh size should improve the approximations.

More importantly, however, is that if one had chosen $|\eta| = 1$, both the Lax-Wendroff and Lax-Friedrich schemes reduce to yield an exact solution as shown in Figure M.7 because the discontinuity will travel along the characteristic; that is, with $c(x, t) = 0$ and $\Delta t = \Delta x \left| \frac{1}{\lambda} \right|$ (or $|\eta| = 1$), both schemes reduce to

$$u_k^{(q+1)} = \begin{cases} u_{k+1}^{(q)} & \text{if } \eta = -1 \\ u_{k-1}^{(q)} & \text{if } \eta = +1 \end{cases}$$

The example shows that the Lax-Wendroff performed quite well, especially when Δt was chosen carefully so that $|\eta| = 1$. Note that the case in which it yielded an exact solution (at the grid points) is limited primarily to a constant η and zero homogeneous case, that is, $c(x, t) = 0$. The other issue remains that Lax-Wendroff and Lax-Friedrich are still explicit time-marching methods.

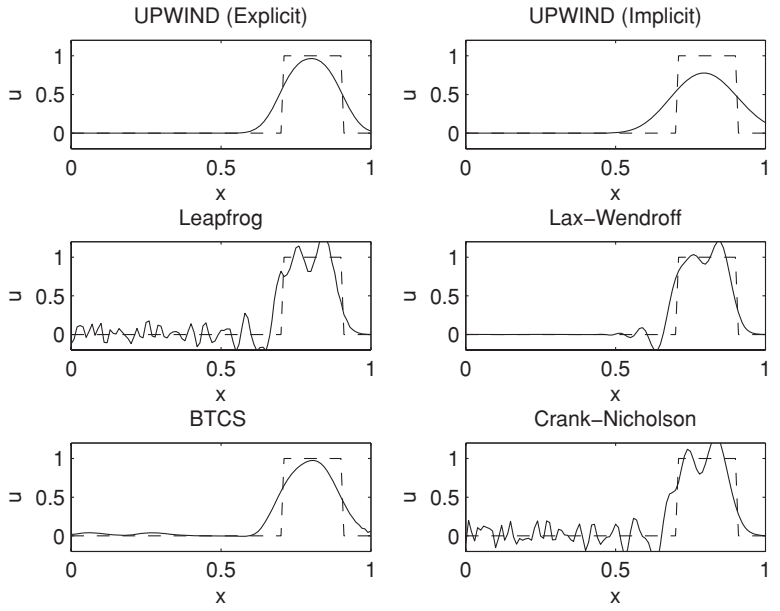


Figure M.6. Comparison with exact solutions for different schemes at $t = 1$.

M.4 Alternating Direction Implicit (ADI) Schemes

Let matrix $\mathbf{G}$ be a multidiagonal, banded matrix of width ω, that is, $\mathbf{G}_{ij} = 0$ for $|i - j| > \omega$. In general, the LU factorization of $\mathbf{G}$ will result in L and U matrices that are banded with the same width. Unfortunately, the matrices generated during finite-difference methods of two or three spatial-dimensional systems are likely to have very wide bands, even though the matrices are very sparse. For instance, matrix $\mathfrak{R}$ in Example 13.9 will have a band of width N. Yet in any row of $\mathfrak{R}$, there are only at most five-nonzero entries. This means that using a full LU factorization of sparse, multidiagonal matrices with large bandwidths may still end up with large amounts of storage and computations.

One group of schemes, known as the **Alternating Direction Implicit (ADI)** schemes, replaces a multidiagonal matrix by a product of two or more tri-diagonal matrices. More importantly, these schemes maintain the same levels of consistency

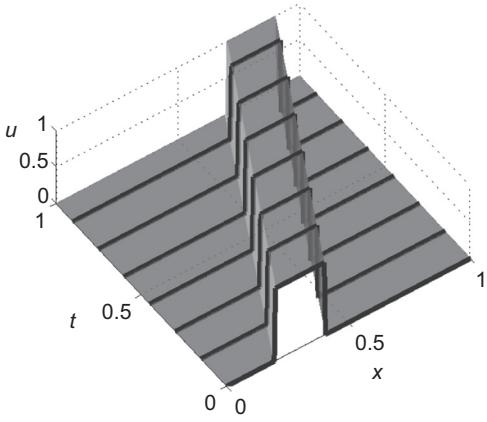


Figure M.7. Numerical solutions for discontinuous initial condition using the Lax-Wendroff with $|\eta| = 1$.

and convergence, as well as the same range of stability as the original schemes. Because the computations are now reduced to solving two or more sequences of tri-diagonal systems, via the Thomas algorithm, the improvements in computational efficiency, in terms of both storage and number of computations, become very significant compared with the direct LU factorizations.

The original ADI schemes were developed by Douglas, Peaceman, and Rachford to improve the Crank-Nicholson schemes for parabolic equations. For a simple illustration of the ADI approach, we take the linear second-order diffusion equation for 2D space, without any mixed partial derivatives, given by

$$\begin{aligned} \frac{\partial u}{\partial t} &= \mu_{xx}(t,x,y)\frac{\partial^2 u}{\partial x^2} + \mu_{yy}(t,x,y)\frac{\partial^2 u}{\partial y^2} + \beta_x(t,x,y)\frac{\partial u}{\partial x} \\ &\quad + \beta_y(t,x,y)\frac{\partial u}{\partial y} + \zeta(t,x,y)u + \eta(t,x,y) \end{aligned} \tag{M.26}$$

together with Dirichlet boundary conditions,

$$\begin{aligned} u(t,0,y) &= v_0(t,y) \quad ; \quad u(t,x,0) = w_0(t,y) \\ u(t,1,y) &= v_1(t,y) \quad ; \quad u(t,x,1) = w_1(t,y) \end{aligned}$$

Let u, η, ζ, μ_{xx}, μ_{yy}, β_x, and β_y be represented in matrix forms,

$$U = \begin{pmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & \ddots & \vdots \\ u_{K1} & \cdots & u_{KN} \end{pmatrix} ; \underline{\underline{\zeta}} = \begin{pmatrix} \zeta_{11} & \cdots & \zeta_{1N} \\ \vdots & \ddots & \vdots \\ \zeta_{K1} & \cdots & \zeta_{KN} \end{pmatrix} ; \text{ etc.}$$

where $u_{kn} = u(k\Delta x, n\Delta y)$, $\zeta_{kn} = \zeta(k\Delta x, n\Delta y)$, etc.

Following the results of (13.39), the semi-discrete approach yields

$$\frac{d}{dt}\,\mathbf{v} = \mathbf{F}(t)\,\mathbf{v} + \mathbf{B}(t) \tag{M.27}$$

where

$$\begin{aligned} \mathbf{v} &= \text{vec}(U) \\ \mathbf{F} &= \mathfrak{M}_x + \mathfrak{M}_y \\ \mathfrak{M}_x &= \left[\underline{\underline{\mu_{xx}}}\right]^{\text{dv}} I_N \otimes \mathcal{D}_{(2,x)} + \left[\underline{\underline{\beta_x}}\right]^{\text{dv}} I_N \otimes \mathcal{D}_{(1,x)} + \frac{1}{2}\underline{\underline{\zeta}}^{\text{dv}} \\ \mathfrak{M}_y &= \left[\underline{\underline{\mu_{yy}}}\right]^{\text{dv}} \mathcal{D}_{(2,y)} \otimes I_K + \left[\underline{\underline{\beta_y}}\right]^{\text{dv}} \mathcal{D}_{(1,y)} \otimes I_K + \frac{1}{2}\underline{\underline{\zeta}}^{\text{dv}} \\ \mathbf{B} &= \left[\underline{\underline{\mu_{xx}}}\right]^{\text{dv}} \text{vec}\left(\mathcal{B}_{(2,x)}\right) + \left[\underline{\underline{\mu_{yy}}}\right]^{\text{dv}} \text{vec}\left(\mathcal{B}^T_{(2,y)}\right) \\ &\quad + \left[\underline{\underline{\beta_x}}\right]^{\text{dv}} \text{vec}\left(\mathcal{B}_{(1,x)}\right) + \left[\underline{\underline{\beta_y}}\right]^{\text{dv}} \text{vec}\left(\mathcal{B}^T_{(1,y)}\right) - \text{vec}\left([\eta]\right) \end{aligned}$$

and the superscript "dv" is the notation for diagonal-vectorization operation.

Applying the Crank-Nicholson scheme, we have

$$\left(I - \frac{\Delta t}{2}\mathbf{F}^{(q+1)}\right)\mathbf{v}^{(q+1)} = \left(I + \frac{\Delta t}{2}\mathbf{F}^{(q)}\right)\mathbf{v}^{(q)} + \frac{\Delta t}{2}\left(\mathbf{B}^{(q+1)} + \mathbf{B}^{(q)}\right) \tag{M.28}$$

By subtracting the term, $\left(I - (\Delta t/2)\Delta t\mathbf{F}^{(q+1)}\right)\mathbf{v}^{(q)}$, from both sides of (M.28),

$$\left(I - \frac{\Delta t}{2}\mathbf{F}^{(q+1)}\right)\left(\mathbf{v}^{(q+1)} - \mathbf{v}^{(q)}\right) = \frac{\Delta t}{2}\left(\left(\mathbf{F}^{(q+1)} + \mathbf{F}^{(q)}\right)\mathbf{v}^{(q)} + \left(\mathbf{B}^{(q+1)} + \mathbf{B}^{(q)}\right)\right) \tag{M.29}$$

Let

$$\Delta_t\mathbf{v}^{(q)} = \mathbf{v}^{(q+1)} - \mathbf{v}^{(q)}$$

then with $\mathbf{F}^{(q+1)} = \mathfrak{M}_x^{(q+1)} + \mathfrak{M}_y^{(q+1)}$ (see (M.28)),

$$\begin{aligned}\left(I - \frac{\Delta t}{2}\mathbf{F}^{(q+1)}\right)\left(\Delta_t\mathbf{v}^{(q)}\right) &= \left(I - \frac{\Delta t}{2}\mathfrak{M}_x^{(q+1)} - \frac{\Delta t}{2}\mathfrak{M}_y^{(q+1)}\right)\left(\Delta_t\mathbf{v}^{(q)}\right)\\ &= \left(I - \frac{\Delta t}{2}\mathfrak{M}_x^{(q+1)}\right)\left(I - \frac{\Delta t}{2}\mathfrak{M}_y^{(q+1)}\right)\left(\Delta_t\mathbf{v}^{(q)}\right)\\ &\quad - \frac{\Delta t^2}{4}\mathfrak{M}_x^{(q+1)}\mathfrak{M}_y^{(q+1)}\left(\Delta_t\mathbf{v}^{(q)}\right)\\ &= \mathbf{G}_x^{(q+1)}\mathbf{G}_y^{(q+1)}\Delta_t\mathbf{v}^{(q)} - \mathcal{O}\left(\Delta t^4\right)\end{aligned}$$

where

$$\mathbf{G}_x^{(q)} = I - \frac{\Delta t}{2}\mathfrak{M}_x^{(q)} \qquad ; \qquad \mathbf{G}_y^{(q)} = I - \frac{\Delta t}{2}\mathfrak{M}_y^{(q)} \tag{M.30}$$

The last term is $\mathcal{O}\left(\Delta t^4\right)$ because of the fact that the Crank-Nicholson scheme guarantees that $\Delta_t\mathbf{v}^{(q)} = \mathbf{v}^{(q+1)} - \mathbf{v}^{(q)} = \mathcal{O}\left(\Delta t^2\right)$. By neglecting terms of order $\mathcal{O}\left(\Delta t^4\right)$, (M.29) can then be replaced by

$$\mathbf{G}_x^{(q+1)}\mathbf{G}_y^{(q+1)}\left(\Delta_t[u]^{(q)}\right) = \frac{\Delta t}{2}\left(\left(\mathbf{F}^{(q+1)} + \mathbf{F}^{(q)}\right)[u]^q + \left(\mathbf{B}^{(q+1)} + \mathbf{B}^{(q)}\right)\right) \tag{M.31}$$

However, $\mathbf{G}_x$ and $\mathbf{G}_y$ are block tri-diagonal matrices whose nonzero submatrices are diagonal in which the main blocks in the diagonal are also tri-diagonal, thus allowing easy implementation of the Thomas and block-Thomas algorithms. Equation (M.31) is known as the **delta-form** of the ADI scheme.[2] The values of $U^{(q+1)}$ are them obtained from

$$U^{(q+1)} = \left(\Delta_t U^{(q)}\right) + u^{(q)} \tag{M.32}$$

It can be shown by direct application of the von Neumann analysis that the ADI scheme given in (M.31) will not change the stability conditions; that is, if the Crank-Nicholson scheme is unconditionally stable, then the corresponding ADI schemes will also be unconditionally stable. Furthermore, because the only change from the original Crank-Nicholson scheme was the removal of terms that are fourth order in Δt, the ADI scheme is also consistent. The application of the Lax equivalence theorem then implies that the ADI schemes will be convergent. The extension of the ADI approach to 3D space is straightforward and is given as an exercise.

[2] The scheme is named **Alternating Direction Implicit** (ADI) based on the fact that the factors $\mathbf{G}_x^{(q)}$ and $\mathbf{G}_y^{(q)}$ deal separately along the x and y directions, respectively. Also, the term **Implicit** (the "I" in ADI) is a reminder that ADI schemes are developed to improve the computation of implicit schemes such as the backward-Euler or Crank-Nicholson, where matrix inversions or LU factorizations are required.

An important issue with ADI schemes is that for accurate time-marching profiles, a small time step is still needed. Recall that the removal of the $\mathcal{O}\left(\Delta t^4\right)$ terms will introduce errors to the original schemes. This additional error is negligible as long as Δt is chosen small enough. However, time-marching approaches are sometimes used primarily to find steady-state solution. In those cases, accuracy only matters at large time values. Because of stability properties, the errors should then have asymptotically settled out toward zero. The ADI schemes are very often used to obtain steady-state solutions because they handle the complexity and size requirements of 2D and 3D systems efficiently.[3]

[3] Other approaches to steady-state solutions include relaxation methods for solving large sparse linear equations such as Jacobi, Gauss-Seidel, SOR. Currently, various Krylov subspace approaches such as conjugate gradient and GMRES (see Sections 2.7 and 2.8) are used for very large sparse problems.

APPENDIX N

Additional Details and Fortification for Chapter 14

N.1 Convex Hull Algorithm

In this section, we describe an algorithm to find a polygonal convex hull of a set of 3D points. The algorithm is a simplified variant of the QuickHull algorithm.[1] Furthermore, we restrict the algorithm only to points in three dimensions where all the points are boundary points of the convex hull. This case applies to points that all come from a paraboloid surface.

We begin by introducing some terms and operators to be used in the algorithm.

1. **Outside sets and visible sets**. For a given facet F, let **Hyp**(F) be the hyperplane that includes F. Then a point p is **outside** of F if it is located on the side of **Hyp**(F) along with the outward unit normal vector (see Figure N.1). Also, the **outside set of** F, denoted by **Out**$(F) = \{p_1, \ldots, p_\ell\}$, is the set of all points that are outside of F.

   Switching perspectives, for a given point p, a facet F is **visible** to p if p is outside of F. The **visible set of** p, denoted by **Vis**$(p) = \{F_1, \ldots, F_q\}$, is the set of all facets that are visible to p.

2. **Ridge sets**. Let $\Phi = \{F_1, \ldots, F_m\}$ be a set of facets that collectively forms a simply connected region D. Then each boundary edge of D, denoted by $\mathcal{R}_i$, is called a **ridge of** Φ, and the collection $\mathcal{R}(\Phi) = \{\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_m\}$ is referred to as the **ridge set of facets in** Φ.

   For example, from the group of facets shown in Figure N.2, let

$$\Phi = \{F_7, F_8, F_9, F_{10}, F_{13}F_{14}F_{15}\}$$

[1] Barber, C. B., Dobkin, D. B., and Huhdanpaa, H. *The QuickHull Algorithm for Convex Hulls*. ACM Trans. on Math. Software, 1995.

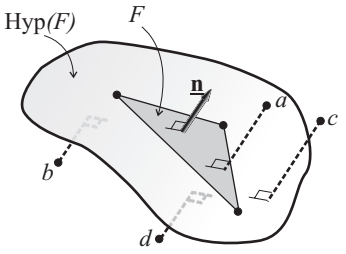


Figure N.1. Points a and c are outside points of facet F, whereas b and d are not. **Hyp**(F) is the hyperplane containing F and **n** is the outward unit normal vector.

then

$$\mathcal{R}(\Phi) = \begin{pmatrix} (p_a, p_b) \\ (p_b, p_c) \\ (p_c, p_d) \\ (p_d, p_e) \\ (p_e, p_f) \\ (p_f, p_g) \\ (p_g, p_a) \end{pmatrix}$$

Note that it will be beneficial for our purposes to specify the sequence of each edge such that they follow a counter-clockwise traversal, for example (p_a, p_b) instead of (p_b, p_a), and so forth.

3. **Extend operation.** Let p be an outside point to a set of connected facets, Φ. Then the operation **Extend**(p, Φ) will take p and attach it to each ridge in $\mathcal{R}(\Phi)$ to form m new facets, where m is the number of ridges of Φ, that is,

$$\begin{pmatrix} F_{M+1} \\ \vdots \\ F_{M+m} \end{pmatrix} = \textbf{Extend}(p, \Phi) \tag{N.1}$$

M is the number of facets before the operation, and

$$F_{M+i} = (p, p_{i,a}, p_{i,b})$$

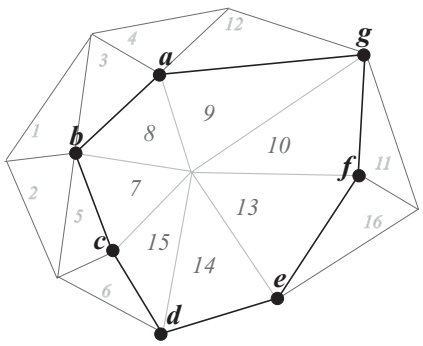


Figure N.2. The set of facets $\Phi = \{F_7, F_8, F_9, F_{10}, F_{13}$ $F_{14}F_{15}\}$ forms a simply connected region whose edges form the ridge set of Φ.

For example, using the same set Φ shown in Figure N.2, suppose p_h is an outside point to the facets in Φ, then we have

$$\textbf{Extend}(p, \Phi) = \begin{pmatrix} F_{17} = (p_h, p_a, p_b) \\ F_{18} = (p_h, p_b, p_c) \\ F_{19} = (p_h, p_c, p_d) \\ F_{20} = (p_h, p_d, p_e) \\ F_{21} = (p_h, p_e, p_f) \\ F_{22} = (p_h, p_f, p_g) \\ F_{23} = (p_h, p_g, p_a) \end{pmatrix}$$

Note that each new facet generated will also have a sequence that goes counterclockwise.

Simplified-QuickHull Algorithm:

Let $P = \{p_1, \ldots, p_N\}$ be the set of available points.

1. **Initialization.**

   (a) Create a tetrahedron as the initial convex hull (e.g., using the points in P corresponding to the three largest z-components and connect them to the point with the smallest z-component):

   $$F = \{F_1, F_2, F_3, F_4\}$$

   (b) Remove, from P, the points that were assigned to F.
   (c) Obtain the collection of current visible sets:

   $$\mathbf{V} = \{\textbf{Vis}(p_i), p_i \in P\}$$

2. **Expand the convex hull using unassigned point p_i.**

   (a) Obtain the ridge set of the visible set of p_i:

   $$\mathcal{R} = \mathcal{R}(\textbf{Vis}(p_i))$$

   (b) Update the facets of the hull:
      i. Generate new facets: $F_{\text{add}} = \textbf{EXTEND}(p_i, \mathcal{R})$.
      ii. Combine with F: $F \leftarrow F \bigcup F_{\text{add}}$.
      iii. Remove $\textbf{Vis}(p_i)$ from F: $F \leftarrow F - \textbf{Vis}(p_i)$.
   (c) Update the collection of visibility sets:
      i. Remove, from each set in $\mathbf{V}$, any reference to the facets in $\textbf{Vis}(p_i)$ (thus also removing $\textbf{Vis}(p_i)$ from $\mathbf{V}$).
      ii. Add facet $F_k \in F_{\text{add}}$ to $\textbf{Vis}(p_j)$ if point p_j is outside of facet of F_k.
   (d) Remove p_i from the set of available points.

This version is a simplification of the QuickHull algorithm. We have assumed that all the points are boundary points; that is, each point will end up as vertices of

the triangular patches forming the convex hull. Because of this, the algorithm steps through each unassigned point and modifies the visibility sets of these points as the convex hull grows in size.[2]

N.2 Stabilization via Streamline-Upwind Petrov-Galerkin (SUPG)

The finite element method discussed in Sections 14.2 through 14.5 used a specific choice for the weights δu, which was defined using the same shape functions for u. As mentioned before, this is known as the Galerkin method. Unfortunately, we expect that as the norm of $\mathbf{M}$ decreases relative to the norm of $\mathbf{b}$, we approach what is known as the **convection-dominated case**, and we expect the Galerkin method to start becoming inaccurate, because the Galerkin method is optimal only for the other extreme case in which $\mathbf{b} = 0$.

For the convection-dominated case, an alternative method known as the **Streamline-Upwind Petrov-Galerkin (SUPG) method** can improve the accuracy. It uses a different set of weights given by

$$\delta u = \widehat{\delta u} + \tau \mathbf{b} \cdot \nabla \left(\widehat{\delta u} \right) \tag{N.2}$$

where τ is known as the **stabilization parameter** that depends on the ratio of $\|\mathbf{b}\|$ over $\|\mathbf{M}\|$ and a characteristic length ℓ of the finite element. The label "streamline-upwind" indicates the presence of $\mathbf{b}$, which is a vector usually known as the advection coefficient or velocity.

With our choice of using triangular linear elements, we can again use the same approach of applying the same shape functions used for u, that is, with

$$\widehat{\delta u} \approx \psi_1 \widehat{\delta u}_1 + \psi_2 \widehat{\delta u}_2 + \psi_3 \widehat{\delta u}_3 \tag{N.3}$$

Doing so, the modifications will simply end up with the addition of one term each for K_n and Γ_n as defined in (14.43) and (14.44), respectively; that is, we now instead use

$$K_n = \left\{ \left(\mathbf{T}^T \mathbf{M}_{(\mathbf{p}^*)} \mathbf{T} - \zeta \mathbf{b}^T_{(\mathbf{p}^*)} \mathbf{T} - g_{(\mathbf{p}^*)} \zeta \zeta^T - \Lambda - \tau T^T \mathbf{b}_{(\mathbf{p}^*)} \mathbf{b}^T_{(\mathbf{p}^*)} T \right) \frac{D}{2} \right\}_n \tag{N.4}$$

$$\Gamma_n = \left\{ \left(h_{(\mathbf{p}^*)} \zeta + Q + Q^{(rbc)} + \tau T^T \mathbf{b}_{(\mathbf{p}^*)} h_{(\mathbf{p}^*)} \right) \frac{D}{2} \right\}_n \tag{N.5}$$

When $\tau = 0$, we get back the Galerkin method.

The last detail is the evaluation of the stabilization parameter τ. Although several studies have found an optimal value for τ in the one-dimensional case, the formulation for the optimal values for 2D and 3D cases remain to be largely heuristic. For simplicity, we can choose the rule we refer to as the **Shakib formula**,

$$\tau = \left[\left(\frac{2b}{\ell} \right)^2 + 9 \left(\frac{4\mu}{\ell^2} \right)^2 + \sigma^2 \right]^{-1/2} \tag{N.6}$$

[2] In the original QuickHull algorithm of Barber and co-workers, the procedure steps through each facet that have non-empty outside sets and then builds the visible set of the farthest outside point. This will involve checking whether the chosen point is outside of the adjacent facets. In case there are points that eventually reside inside the convex hull, the original version will likely be more efficient. Nonetheless, we opted to describe the revised approach because of its relative simplicity.

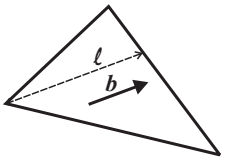


Figure N.3. The characteristic length ℓ based on the direction of $\mathbf{b}$.

where ℓ is the characteristic length of the triangle, $b = \left\|\mathbf{b}_{(\mathbf{p}^*)}\right\|$, $\mu = \left\|M_{(\mathbf{p}^*)}\right\|$, and $\sigma = g_{(\mathbf{p}^*)}$. The length ℓ is the distance of the segment from one vertex of the triangle to the opposite edge in the direction of $\mathbf{b}_{(\mathbf{p}^*)}$ as shown in Figure N.3. (Note that only one of the vertices can satisfy this condition.) The length ℓ can be found as follows: Let $\mathbf{v} = \mathbf{b}/\left\|\mathbf{b}\right\|$. Find node i such that solving

$$\begin{pmatrix} s \\ \lambda \end{pmatrix} = \left(\mathbf{v} \;\middle|\; (\mathbf{p}_k - \mathbf{p}_i) \right)^{-1} \left(\mathbf{p}_k - \mathbf{p}_j \right) \tag{N.7}$$

will yield $0 \leq \lambda \leq 1$. Then $\ell = |s|$ is the length of the segment from node i to the edge containing nodes j and k.

EXAMPLE N.1. To test the SUPG method, consider the differential equation

$$[\nabla \cdot (\mathbf{M}(x, y) \cdot \nabla u)] \; + \; [\mathbf{b}(x, y) \cdot \nabla u] \; + \; g(x, y)u \; + \; h(x, y) \; = \; 0$$

with

$$\mathbf{M} = \begin{pmatrix} 0.001 & 0 \\ 0 & 0.001 \end{pmatrix} ; \; \mathbf{b} = \begin{pmatrix} -2 \\ 3 \end{pmatrix} ; \; g = 0$$

and

$$h = -1.5\,(3x - 2y) - \left(0.32\left(x^2 + y^2\right) - 80\,(1.5y - x + 0.001)\,e^{-4\left(x^2+y^2\right)}\right)$$

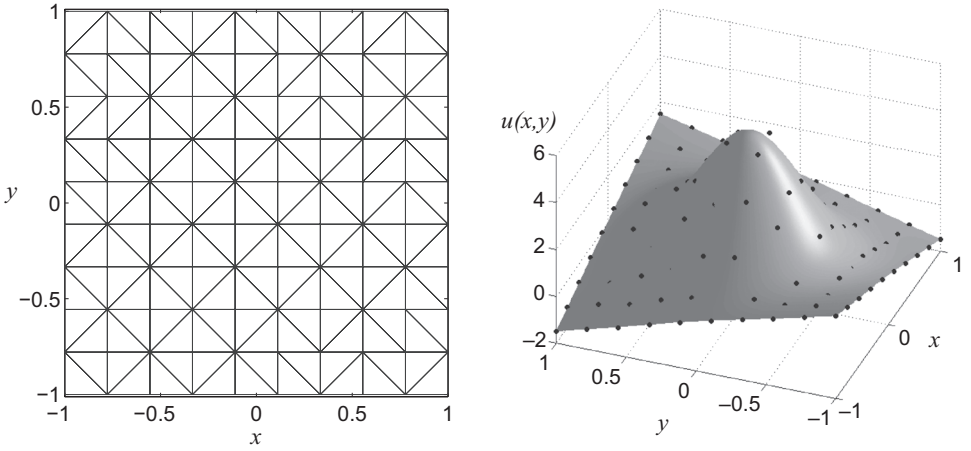


Figure N.4. The triangulation mesh is shown in the left plot, whereas the SUPG solution (dots) is shown together with exact solution (surface) in the right plot.

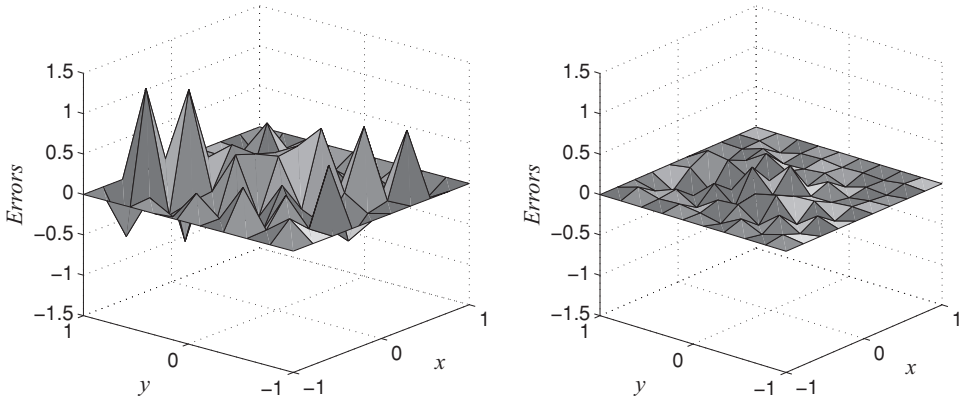


Figure N.5. The errors obtained using the Galerkin method are shown in the left plot, whereas the errors obtained using the SUPG method are shown in the right plot.

Let the domain to be a square of width 2, centered at the origin. Also, let all the boundary conditions be Dirichlet, with

$$u = 1.5xy + 5e^{-4\left(x^2+y^2\right)} \qquad \text{for} \qquad \begin{cases} x = -1 & , \quad -1 \le y \le 1 \\ x = 1 & , \quad -1 \le y \le 1 \\ -1 \le x \le 1 & , \quad y = -1 \\ -1 \le x \le 1 & , \quad y = 1 \end{cases}$$

The exact solution of this problem is known (which was in fact used to set h and the boundary conditions) and given by

$$u = 1.5xy + 5e^{-4\left(x^2+y^2\right)}$$

After applying the SUPG methods based on a Delaunay mesh shown in the left plot of Figure N.4, we obtain the solution shown in the right plot of Figure N.4. The improvements of the SUPG method over the Galerkin method are shown Figure N.5. The errors for the Galerkin and the SUPG are ± 1.2 and ± 0.3, respectively.

Of course, as the mesh sizes are decreased, the accuracy will also increase. Furthermore, note that from (N.6), the stabilization parameter τ for each element will approach 0 as $\ell \to 0$, reducing the SUPG method to a simple Galerkin method.

Remarks: The results for this example were generated by the MATLAB function `fem_sq_test2.m`, which uses the function `linear_2d_supg.m` – a general SUPG finite element solver for the linear second-order partial differential equation. Both of these files are available on the book's webpage.