Applications of Linear Algebra

by

Gordon C. Everstine

5 June 2010

Copyright © 1998–2010 by Gordon C. Everstine.
All rights reserved.

This book was typeset with LaTeX 2$_\varepsilon$ (MiKTeX).

Preface

These lecture notes are intended to supplement a one-semester graduate-level engineering course at The George Washington University in algebraic methods appropriate to the solution of engineering computational problems, including systems of linear equations, linear vector spaces, matrices, least squares problems, Fourier series, and eigenvalue problems. In general, the mix of topics and level of presentation are aimed at upper-level undergraduates and first-year graduate students in mechanical, aerospace, and civil engineering.

Gordon Everstine
Gaithersburg, Maryland
June 2010

Contents

List of Figures

1 Systems of Linear Equations

A system of n linear equations in n unknowns can be expressed in the form

$$\left.\begin{array}{rcrcrcrcrcl} A_{11}x_1 & + & A_{12}x_2 & + & A_{13}x_3 & + & \cdots & + & A_{1n}x_n & = & b_1 \\ A_{21}x_1 & + & A_{22}x_2 & + & A_{23}x_3 & + & \cdots & + & A_{2n}x_n & = & b_2 \\ & & & & & & \vdots & & & & \\ A_{n1}x_1 & + & A_{n2}x_2 & + & A_{n3}x_3 & + & \cdots & + & A_{nn}x_n & = & b_n. \end{array}\right\} \tag{1.1}$$

The problem is to find numbers $x_1, x_2, \cdots, x_n$, given the coefficients A_{ij} and the right-hand side $b_1, b_2, \cdots, b_n$.

In matrix notation,

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1.2}$$

where $\mathbf{A}$ is the $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}, \tag{1.3}$$

$\mathbf{b}$ is the right-hand side vector

$$\mathbf{b} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{Bmatrix}, \tag{1.4}$$

and $\mathbf{x}$ is the solution vector

$$\mathbf{x} = \begin{Bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{Bmatrix}. \tag{1.5}$$

1.1 Definitions

The *identity matrix* $\mathbf{I}$ is the square diagonal matrix with 1s on the diagonal:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{1.6}$$

In index notation, the identity matrix is the *Kronecker delta*:

$$I_{ij} = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{1.7}$$

The *inverse* $\mathbf{A}^{-1}$ of a square matrix $\mathbf{A}$ is the matrix such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{1.8}$$

The inverse is unique if it exists. For example, let $\mathbf{B}$ be the inverse, so that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. If $\mathbf{C}$ is another inverse, $\mathbf{C} = \mathbf{CI} = \mathbf{C}(\mathbf{AB}) = (\mathbf{CA})\mathbf{B} = \mathbf{IB} = \mathbf{B}$.

A square matrix $\mathbf{A}$ is said to be *orthogonal* if

$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}. \tag{1.9}$$

That is, an orthogonal matrix is one whose inverse is equal to the transpose.

An *upper triangular* matrix has only zeros below the diagonal, e.g.,

$$\mathbf{U} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ 0 & A_{22} & A_{23} & A_{24} & A_{25} \\ 0 & 0 & A_{33} & A_{34} & A_{35} \\ 0 & 0 & 0 & A_{44} & A_{45} \\ 0 & 0 & 0 & 0 & A_{55} \end{bmatrix}. \tag{1.10}$$

Similarly, a *lower triangular* matrix has only zeros above the diagonal, e.g.,

$$\mathbf{L} = \begin{bmatrix} A_{11} & 0 & 0 & 0 & 0 \\ A_{21} & A_{22} & 0 & 0 & 0 \\ A_{31} & A_{32} & A_{33} & 0 & 0 \\ A_{41} & A_{42} & A_{43} & A_{44} & 0 \\ A_{51} & A_{52} & A_{53} & A_{54} & A_{55} \end{bmatrix}. \tag{1.11}$$

A *tridiagonal* matrix has all its nonzeros on the main diagonal and the two adjacent diagonals, e.g.,

$$\mathbf{A} = \begin{bmatrix} \mathrm{x} & \mathrm{x} & 0 & 0 & 0 & 0 \\ \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 & 0 \\ 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 & 0 \\ 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} & 0 \\ 0 & 0 & 0 & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & \mathrm{x} & \mathrm{x} \end{bmatrix}. \tag{1.12}$$

Consider two three-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ given by

$$\mathbf{u} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + u_3\mathbf{e}_3, \quad \mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3, \tag{1.13}$$

where $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are the mutually orthogonal unit basis vectors in the Cartesian coordinate directions, and u_1, u_2, u_3, v_1, v_2, and v_3 are the vector components. The *dot product* (or *inner product*) of $\mathbf{u}$ and $\mathbf{v}$ is

$$\mathbf{u}\cdot\mathbf{v} = (u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + u_3\mathbf{e}_3)\cdot(v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3) = u_1v_1 + u_2v_2 + u_3v_3 = \sum_{i=1}^{3} u_iv_i, \tag{1.14}$$

where

$$\mathbf{e}_i\cdot\mathbf{e}_j = \delta_{ij}. \tag{1.15}$$

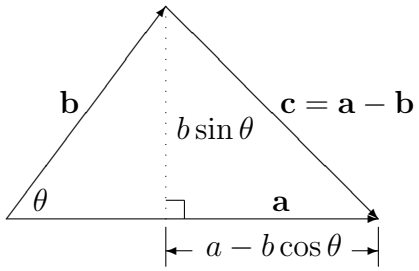

Figure 1: Two Vectors.

In n dimensions,

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{n} u_i v_i, \tag{1.16}$$

where n is the dimension of each vector (the number of components). There are several notations for this product, all of which are equivalent:

inner product notation	$(\mathbf{u}, \mathbf{v})$
vector notation	$\mathbf{u} \cdot \mathbf{v}$
index notation	$\sum_{i=1}^{n} u_i v_i$
matrix notation	$\mathbf{u}^T\mathbf{v}$ or $\mathbf{v}^T\mathbf{u}$

The scalar product can also be written in matrix notation as

$$\mathbf{u}^T\mathbf{v} = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{Bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{Bmatrix}. \tag{1.17}$$

For $\mathbf{u}$ a vector, the *length* of $\mathbf{u}$ is

$$u = |\mathbf{u}| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{\sum_{i=1}^{n} u_i u_i} = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2}. \tag{1.18}$$

A *unit vector* is a vector whose length is unity. Two vectors $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal* if $\mathbf{u} \cdot \mathbf{v} = 0$.

Consider the triangle defined by three vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ (Fig. 1), where

$$\mathbf{c} = \mathbf{a} - \mathbf{b}, \tag{1.19}$$

in which case

$$(a - b\cos\theta)^2 + (b\sin\theta)^2 = c^2, \tag{1.20}$$

where a, b, and c are the lengths of $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, respectively, amd θ is the angle between $\mathbf{a}$ and $\mathbf{b}$. Eq. 1.20 can be expanded to yield

$$a^2 - 2ab\cos\theta + b^2\cos^2\theta + b^2\sin^2\theta = c^2 \tag{1.21}$$

or

$$c^2 = a^2 + b^2 - 2ab\cos\theta, \tag{1.22}$$

which is the *law of cosines*. We can further expand the law of cosines in terms of components to obtain

$$(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 = a_1^2 + a_2^2 + a_3^2 + b_1^2 + b_2^2 + b_3^2 - 2ab\cos\theta \tag{1.23}$$

or

$$a_1b_1 + a_2b_2 + a_3b_3 = ab\cos\theta. \tag{1.24}$$

Thus, the dot product of two vectors can alternatively be expressed as

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta = ab\cos\theta, \tag{1.25}$$

where θ is the angle between the two vectors.

1.2 Nonsingular Systems of Equations

If the linear system $\mathbf{Ax} = \mathbf{b}$ has a unique solution $\mathbf{x}$ for every right-hand side $\mathbf{b}$, the system is said to be *nonsingular*. Usually, one refers to the matrix rather than the system of equations as being nonsingular, since this property is determined by the coefficients A_{ij} of the matrix and does not depend on the right-hand side $\mathbf{b}$. If $\mathbf{A}$ is *singular* (i.e., not nonsingular), then, for some right-hand sides, the system $\mathbf{Ax} = \mathbf{b}$ will be *inconsistent* and have no solutions, and, for other right-hand sides, the equations will be dependent and have an infinite number of solutions.

For example, the system

$$\left.\begin{array}{rcrcl} 2x_1 & + & 3x_2 & = & 5 \\ 4x_1 & + & 6x_2 & = & 1 \end{array}\right\} \tag{1.26}$$

is inconsistent and has no solution. The system

$$\left.\begin{array}{rcrcl} 2x_1 & + & 3x_2 & = & 5 \\ 4x_1 & + & 6x_2 & = & 10 \end{array}\right\} \tag{1.27}$$

is redundant (i.e., has dependent equations) and has an infinite number of solutions. The only possibilities for a square system of linear equations is that the system have no solution, one solution, or an infinite number of solutions.

1.3 Diagonal and Triangular Systems

Several types of systems of equations are particularly easy to solve, including diagonal and triangular systems. For the diagonal matrix

$$\mathbf{A} = \begin{bmatrix} A_{11} & & & & \\ & A_{22} & & & \\ & & A_{33} & & \\ & & & \ddots & \\ & & & & A_{nn} \end{bmatrix}, \tag{1.28}$$

the system $\mathbf{Ax} = \mathbf{b}$ has the solution

$$x_1 = \frac{b_1}{A_{11}}, \quad x_2 = \frac{b_2}{A_{22}}, \quad \cdots, \quad x_n = \frac{b_n}{A_{nn}}, \tag{1.29}$$

assuming that, for each i, $A_{ii} \neq 0$. If $A_{kk} = 0$ for some k, then x_k does not exist if $b_k \neq 0$, and x_k is arbitrary if $b_k = 0$.

Consider the upper triangular system ($A_{ij} = 0$ if $i > j$)

$$\left.\begin{array}{rcrcrcrcrcl} A_{11}x_1 & + & A_{12}x_2 & + & A_{13}x_3 & + & \cdots & + & A_{1n}x_n & = & b_1 \\ & & A_{22}x_2 & + & A_{23}x_3 & + & \cdots & + & A_{2n}x_n & = & b_2 \\ & & & & A_{33}x_3 & + & \cdots & + & A_{3n}x_n & = & b_3 \\ & & & & & & & & \vdots & & \\ & & & & & & A_{n-1,n-1}x_{n-1} & + & A_{n-1,n}x_n & = & b_{n-1} \\ & & & & & & & & A_{nn}x_n & = & b_n. \end{array}\right\} \tag{1.30}$$

This system is solved by *backsolving*:

$$x_n = b_n/A_{nn}, \tag{1.31}$$

$$x_{n-1} = (b_{n-1} - A_{n-1,n}x_n)/A_{n-1,n-1}, \tag{1.32}$$

and so on until all components of $\mathbf{x}$ are found.

The backsolving algorithm can be summarized as follows:

1. Input $n, \mathbf{A}, \mathbf{b}$
2. $x_n = b_n/A_{nn}$
3. For $k = n-1, n-2, \ldots, 1$: [k = row number]
   $x_k = b_k$
   For $i = k+1, k+2, \ldots, n$:
   $x_k = x_k - A_{ki}x_i$
   $x_k = x_k/A_{kk}$
4. Output $\mathbf{x}$

Since the only impediments to the successful completion of this algorithm are the divisions, it is apparent that an upper triangular system is nonsingular if, and only if, $A_{ii} \neq 0$ for all i. Alternatively, if any $A_{ii} = 0$, the system is singular.

For example, consider the upper triangular system

$$\left.\begin{array}{rcrcrcl} x_1 & - & 2x_2 & + & x_3 = & 1 \\ & & 2x_2 & + & 4x_3 = & 10 \\ & & & - & 2x_3 = & 8. \end{array}\right\} \tag{1.33}$$

From the backsolving algorithm, the solution is found as

$$\begin{aligned} x_3 &= -4, \\ x_2 &= [10 - 4(-4)]/2 = 13, \\ x_1 &= 1 + 2(13) - (-4) = 31. \end{aligned} \tag{1.34}$$

A variation on the back-solving algorithm listed above is obtained if we return the solution in the right-hand side array **b**. That is, array **x** could be eliminated by overwriting **b** and storing the solution in **b**. With this variation, the back-solving algorithm becomes

1. Input $n, \mathbf{A}, \mathbf{b}$
2. $b_n = b_n / A_{nn}$
3. For $k = n-1, n-2, \ldots, 1$: [k = row number]
   For $i = k+1, k+2, \ldots, n$:
   $b_k = b_k - A_{ki} b_i$
   $b_k = b_k / A_{kk}$
4. Output **b**

On output, the solution is returned in **b**, and the original **b** is destroyed.

A lower triangular system $\mathbf{Ax} = \mathbf{b}$ is one for which $A_{ij} = 0$ if $i < j$. This system can be solved directly by forward solving in a manner similar to backsolving for upper triangular systems.

1.4 Gaussian Elimination

The approach widely used to solve general systems of linear equations is Gaussian elimination, for which the procedure is to transform the general system into upper triangular form, and then backsolve. The transformation is based on the following *elementary operations*, which have no effect on the solution of the system:

1. An equation can be multiplied by a nonzero constant.
2. Two equations can be interchanged.
3. Two equations can be added together and either equation replaced by the sum.
4. An equation can be multiplied by a nonzero constant, added to another equation, and the result used to replace either of the first two equations. (This operation is a combination of the first and third operations.)

We illustrate Gaussian elimination with an example. Consider the system

$$\left.\begin{array}{rcrcrcrcl} -x_1 & + & 2x_2 & + & 3x_3 & + & x_4 & = & 1 \\ -3x_1 & + & 8x_2 & + & 8x_3 & + & x_4 & = & 2 \\ x_1 & + & 2x_2 & - & 6x_3 & + & 4x_4 & = & -1 \\ 2x_1 & - & 4x_2 & - & 5x_3 & - & x_4 & = & 0. \end{array}\right\} \tag{1.35}$$

Step 1: Eliminate all coefficients of x_1 except for that in the first equation:

$$\left.\begin{array}{rcrcrcrcl} -x_1 & + & 2x_2 & + & 3x_3 & + & x_4 & = & 1 \\ & & 2x_2 & - & x_3 & - & 2x_4 & = & -1 \\ & & 4x_2 & - & 3x_3 & + & 5x_4 & = & 0 \\ & & & & x_3 & + & x_4 & = & 2. \end{array}\right\} \tag{1.36}$$

Step 2: Eliminate all coefficients of x_2 except for those in the first two equations:

$$\left.\begin{array}{rcrcrcrcl} -x_1 & + & 2x_2 & + & 3x_3 & + & x_4 & = & 1 \\ & & 2x_2 & - & x_3 & - & 2x_4 & = & -1 \\ & & & - & x_3 & + & 9x_4 & = & 2 \\ & & & & x_3 & + & x_4 & = & 2. \end{array}\right\} \tag{1.37}$$

Step 3: Eliminate all coefficients of x_3 except for those in the first three equations:

$$\left.\begin{array}{rcrcrcrcl} -x_1 & + & 2x_2 & + & 3x_3 & + & x_4 & = & 1 \\ & & 2x_2 & - & x_3 & - & 2x_4 & = & -1 \\ & & & - & x_3 & + & 9x_4 & = & 2 \\ & & & & & & 10x_4 & = & 4. \end{array}\right\} \tag{1.38}$$

Ths system is now in upper triangular form, and we could complete the solution by back-solving the upper triangular system. Notice in the above procedure that we work with both sides of the equations simultaneously.

Notice also that we could save some effort in the above procedure by not writing the unknowns at each step, and instead use matrix notation. We define a matrix consisting of the left-hand side coefficient matrix augmented by the right-hand side, and apply to this matrix the same elementary operations used in the three steps above:

$$\left[\begin{array}{rrrr|r} -1 & 2 & 3 & 1 & 1 \\ -3 & 8 & 8 & 1 & 2 \\ 1 & 2 & -6 & 4 & -1 \\ 2 & -4 & -5 & -1 & 0 \end{array}\right] \longrightarrow \left[\begin{array}{rrrr|r} -1 & 2 & 3 & 1 & 1 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 4 & -3 & 5 & 0 \\ 0 & 0 & 1 & 1 & 2 \end{array}\right] \tag{1.39}$$

$$\longrightarrow \left[\begin{array}{rrrr|r} -1 & 2 & 3 & 1 & 1 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 0 & -1 & 9 & 2 \\ 0 & 0 & 1 & 1 & 2 \end{array}\right] \longrightarrow \left[\begin{array}{rrrr|r} -1 & 2 & 3 & 1 & 1 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 0 & -1 & 9 & 2 \\ 0 & 0 & 0 & 10 & 4 \end{array}\right]. \tag{1.40}$$

Note that this reduction can be performed "in-place" (i.e., by writing over the matrix and destroying both the original matrix and the right-hand side).

Multiple right-hand sides can be handled simultaneously. For example, if we want the solution of $\mathbf{A}\mathbf{x}_1 = \mathbf{b}_1$ and $\mathbf{A}\mathbf{x}_2 = \mathbf{b}_2$, we augment the coefficient matrix by both right-hand sides:

$$\left[\begin{array}{rrrr|rr} -1 & 2 & 3 & 1 & 1 & 2 \\ -3 & 8 & 8 & 1 & 2 & 3 \\ 1 & 2 & -6 & 4 & -1 & 3 \\ 2 & -4 & -5 & -1 & 0 & -1 \end{array}\right].$$

The Gaussian elimination algorithm (reduction to upper triangular form) can now be summarized as follows:

1. Input $n, \mathbf{A}, \mathbf{b}$
2. For $k = 1, 2, \ldots, n-1$: [k = pivot row number]
   For $i = k+1, k+2, \ldots, n$: [i = row number below k]

$m = -A_{ik}/A_{kk}$ [m = row multiplier]
For $j = k+1, k+2, \ldots, n$: [j = column number to right of k]
$A_{ij} = A_{ij} + mA_{kj}$
$b_i = b_i + mb_k$ [rhs]

3. Output $\mathbf{A}, \mathbf{b}$

The output of this procedure is an upper triangular system which can be solved by back-solving.

Note that each multiplier m could be saved in A_{ik} (which is in the lower triangle and not used in the above algorithm) so that additional right-hand sides could be solved later without having to reduce the original system again. Thus, the Gaussian elimination algorithm is modified as follows:

1. Input $n, \mathbf{A}, \mathbf{b}$

2. For $k = 1, 2, \ldots, n-1$: [k = pivot row number]
For $i = k+1, k+2, \ldots, n$: [i = row number below k]
$A_{ik} = -A_{ik}/A_{kk}$ [multiplier stored in A_{ik}]
For $j = k+1, k+2, \ldots, n$: [j = column number to right of k]
$A_{ij} = A_{ij} + A_{ik}A_{kj}$
$b_i = b_i + A_{ik}b_k$ [rhs]

3. Output $\mathbf{A}, \mathbf{b}$

Note further that, if the multipliers are saved in A_{ik}, the loops involving the coefficient matrix $\mathbf{A}$ can (and should) be separated from those involving the right-hand side $\mathbf{b}$:

1. Input $n, \mathbf{A}, \mathbf{b}$

2. Elimination
For $k = 1, 2, \ldots, n-1$: [k = pivot row number]
For $i = k+1, k+2, \ldots, n$: [i = row number below k]
$A_{ik} = -A_{ik}/A_{kk}$ [multiplier stored in A_{ik}]
For $j = k+1, k+2, \ldots, n$: [j = column number to right of k]
$A_{ij} = A_{ij} + A_{ik}A_{kj}$

3. RHS modification
For $k = 1, 2, \ldots, n-1$: [k = pivot row number]
For $i = k+1, k+2, \ldots, n$: [i = row number below k]
$b_i = b_i + A_{ik}b_k$ [rhs]

4. Output $\mathbf{A}, \mathbf{b}$

If the coefficient matrix is banded, the loops above can be shortened to avoid unnecessary operations on zeros. A band matrix, which occurs frequently in applications, is one for which the nonzeros are clustered about the main diagonal. We define the matrix semi-bandwidth w as the maximum "distance" of any nonzero off-diagonal term from the diagonal (including

the diagonal). Thus, for a band matrix, the upper limit n in the loops associated with pivot row k above can be replaced by

$$k_{\max} = \min(k + w - 1, n). \tag{1.41}$$

The modified algorithm thus becomes

1. Input $n, \mathbf{A}, \mathbf{b}, w$ $[w =$ matrix semi-bandwidth$]$

2. Elimination
   For $k = 1, 2, \ldots, n-1$: $[k =$ pivot row number$]$
   $k_{\max} = \min(k + w - 1, n)$ [replaces n in the loops]
   For $i = k+1, k+2, \ldots, k_{\max}$: $[i =$ row number below $k]$
   $A_{ik} = -A_{ik}/A_{kk}$ [multiplier stored in A_{ik}]
   For $j = k+1, k+2, \ldots, k_{\max}$: $[j =$ column number to right of $k]$
   $A_{ij} = A_{ij} + A_{ik}A_{kj}$

3. RHS modification
   For $k = 1, 2, \ldots, n-1$: $[k =$ pivot row number$]$
   $k_{\max} = \min(k + w - 1, n)$ [replaces n in the loop]
   For $i = k+1, k+2, \ldots, k_{\max}$: $[i =$ row number below $k]$
   $b_i = b_i + A_{ik}b_k$ [rhs]

4. Output $\mathbf{A}, \mathbf{b}$

A similar modification is made to the back-solve algorithm.

1.5 Operation Counts

Except for small systems of equations (small n), we can estimate the number of operations required to execute an algorithm by looking only at the operations in the innermost loop. We define an *operation* as the combination of a multiply and add, since they generally are executed together. Gaussian elimination involves a triply-nested loop. For fixed row number k, the inner loops on i and j are executed $n - k$ times each. Hence,

$$\begin{aligned} \text{Number of multiply-adds} &\approx \sum_{k=1}^{n-1}(n-k)^2 = \sum_{m=n-1}^{1} m^2 = \sum_{m=1}^{n-1} m^2 \\ &\approx \int_0^{n-1} m^2\, dm = (n-1)^3/3 \approx n^3/3, \end{aligned} \tag{1.42}$$

where the approximations all depend on n's being large. Thus, for a fully-populated matrix, Gaussian elimination requires about $n^3/3$ operations.

For a band matrix with a semi-bandwidth $w \ll n$, w^2 operations are performed for each of the n rows (to first order). Thus, for a general band matrix, Gaussian elimination requires about nw^2 operations.

For backsolving, we consider Step 3 of the backsolve algorithm. In Row k (counting from the bottom), there are approximately k multiply-add operations. Hence,

$$\text{Number of multiply-adds} \approx \sum_{k=1}^{n} k = n(n+1)/2 \approx n^2/2. \tag{1.43}$$

Thus, backsolving requires about $n^2/2$ operations for each right-hand side.

Notice that, for large n, the elimination step has many more operations than does the backsolve.

1.6 Partial Pivoting

For Gaussian elimination to work, the (current) diagonal entry A_{ii} must be nonzero at each intermediate step. For example, consider the same example used earlier, but with the equations rearranged:

$$\left[\begin{array}{rrrr|r} -1 & 2 & 3 & 1 & 1 \\ 2 & -4 & -5 & -1 & 0 \\ -3 & 8 & 8 & 1 & 2 \\ 1 & 2 & -6 & 4 & -1 \end{array}\right] \longrightarrow \left[\begin{array}{rrrr|r} -1 & 2 & 3 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 2 & -1 & -2 & -1 \\ 0 & 4 & -3 & 5 & 0 \end{array}\right]. \tag{1.44}$$

In this case, the next step would fail, since $A_{22} = 0$. The solution to this problem is to interchange two equations to avoid the zero divisor. If the system is nonsingular, this interchange is always possible.

However, complications arise in Gaussian elimination not only when $A_{ii} = 0$ at some step but also sometimes when A_{ii} is small. For example, consider the system

$$\left.\begin{array}{rcrcl} (1.000 \times 10^{-5})x_1 & + & 1.000x_2 & = & 1.000 \\ 1.000x_1 & + & 1.000x_2 & = & 2.000, \end{array}\right\} \tag{1.45}$$

the solution of which is $(x_1, x_2) = (1.00001000\ldots, 0.99998999\ldots)$. Assume that we perform the Gaussian elimination on a computer which has four digits of precision. After the one-step elimination, we obtain

$$\left.\begin{array}{rcrcl} (1.000 \times 10^{-5})x_1 & + & 1.000x_2 & = & 1.000 \\ & - & 1.000 \times 10^5 x_2 & = & -1.000 \times 10^5, \end{array}\right\} \tag{1.46}$$

whose solution is $x_2 = 1.000$, $x_1 = 0.000$.

If we now interchange the original equations, the system is

$$\left.\begin{array}{rcrcl} 1.000x_1 & + & 1.000x_2 & = & 2.000 \\ (1.000 \times 10^{-5})x_1 & + & 1.000x_2 & = & 1.000, \end{array}\right\} \tag{1.47}$$

which, after elimination, yields

$$\left.\begin{array}{rcrcl} 1.000x_1 & + & 1.000x_2 & = & 2.000 \\ & & 1.000x_2 & = & 1.000, \end{array}\right\} \tag{1.48}$$

whose solution is $x_2 = 1.000$, $x_1 = 1.000$, which is correct to the precision available.

We conclude from this example that it is better to interchange equations at Step i so that $|A_{ii}|$ is as large as possible. The interchanging of equations is called *partial pivoting*. Thus an alternative version of the Gaussian elimination algorithm would be that given previously except that, at each step (Step i), the ith equation is interchanged with one below so that $|A_{ii}|$ is as large as possible. In actual practice, the "interchange" can alternatively be performed by interchanging the equation subscripts rather than the equations themselves.

1.7 LU Decomposition

Define $\mathbf{M}_{ik}$ as the identity matrix with the addition of the number m_{ik} in Row i, Column k:

$$\mathbf{M}_{ik} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & m_{ik} & \ddots & \\ & & & & 1 \end{bmatrix}. \tag{1.49}$$

For example, for $n = 3$, $\mathbf{M}_{31}$ is

$$\mathbf{M}_{31} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix}. \tag{1.50}$$

We now consider the effect of multiplying a matrix $\mathbf{A}$ by $\mathbf{M}_{ik}$. For example, for $n = 3$,

$$\mathbf{M}_{31}\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ m_{31} & 0 & 1 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \tag{1.51}$$

$$= \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ m_{31}A_{11} + A_{31} & m_{31}A_{12} + A_{32} & m_{31}A_{13} + A_{33} \end{bmatrix}. \tag{1.52}$$

Thus, when $\mathbf{A}$ is multiplied by $\mathbf{M}_{31}$, we obtain a new matrix which has Row 3 replaced by the sum of Row 3 and $m_{31}\times$Row 1. In general, multiplication by $\mathbf{M}_{ik}$ has the effect of replacing Row i by Row $i + m_{ik}\times$Row k. Thus, each step of the Gaussian elimination process is equivalent to multiplying $\mathbf{A}$ by some $\mathbf{M}_{ik}$. For example, for $n = 4$ (a 4×4 matrix), the entire Gaussian elimination can be written as

$$\mathbf{M}_{43}\mathbf{M}_{42}\mathbf{M}_{32}\mathbf{M}_{41}\mathbf{M}_{31}\mathbf{M}_{21}\mathbf{A} = \mathbf{U}, \tag{1.53}$$

where

$$m_{ik} = -\frac{A_{ik}}{A_{kk}} \tag{1.54}$$

are the multipliers in Gaussian elimination, and $\mathbf{U}$ is the upper triangular matrix obtained in the Gaussian elimination. Note that the multipliers m_{ik} in Eq. 1.54 are computed using the current matrix entries obtained during Gaussian elimination.

If each $\mathbf{M}_{ik}^{-1}$ exists, then, for $n = 4$,

$$\mathbf{A} = (\mathbf{M}_{43}\mathbf{M}_{42}\mathbf{M}_{32}\mathbf{M}_{41}\mathbf{M}_{31}\mathbf{M}_{21})^{-1}\,\mathbf{U} = \mathbf{M}_{21}^{-1}\mathbf{M}_{31}^{-1}\mathbf{M}_{41}^{-1}\mathbf{M}_{32}^{-1}\mathbf{M}_{42}^{-1}\mathbf{M}_{43}^{-1}\mathbf{U}, \tag{1.55}$$

since the inverse of a matrix product is equal to the product of the matrix inverses in reverse order. We observe, by example, that $\mathbf{M}_{ik}^{-1}$ exists and is obtained from $\mathbf{M}_{ik}$ simply by changing the sign of m_{ik} (the ik element), since, for $n = 4$ and $\mathbf{M}_{42}$,

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & m_{42} & & 1 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & -m_{42} & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & m_{42} - m_{42} & 0 & 1 \end{bmatrix} = \mathbf{I}. \tag{1.56}$$

From Eq. 1.55, for $n = 3$, we obtain

$$\mathbf{A} = \mathbf{M}_{21}^{-1}\mathbf{M}_{31}^{-1}\mathbf{M}_{32}^{-1}\mathbf{U}, \tag{1.57}$$

where

$$\mathbf{M}_{21}^{-1}\mathbf{M}_{31}^{-1}\mathbf{M}_{32}^{-1} = \begin{bmatrix} 1 & & \\ -m_{21} & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ -m_{31} & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ & -m_{32} & 1 \end{bmatrix} \tag{1.58}$$

$$= \begin{bmatrix} 1 & & \\ -m_{21} & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ -m_{31} & -m_{32} & 1 \end{bmatrix} \tag{1.59}$$

$$= \begin{bmatrix} 1 & & \\ -m_{21} & 1 & \\ -m_{31} & -m_{32} & 1 \end{bmatrix}. \tag{1.60}$$

In general,

$$\mathbf{M}_{21}^{-1}\mathbf{M}_{31}^{-1}\cdots\mathbf{M}_{n1}^{-1}\cdots\mathbf{M}_{n,n-1}^{-1} = \begin{bmatrix} 1 & & & & \\ -m_{21} & 1 & & & \\ -m_{31} & -m_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -m_{n1} & -m_{n2} & \cdots & -m_{n,n-1} & 1 \end{bmatrix}, \tag{1.61}$$

and, for any nonsingular matrix $\mathbf{A}$,

$$\mathbf{A} = \mathbf{L}\mathbf{U}, \tag{1.62}$$

where $\mathbf{L}$ is the lower triangular matrix

$$\mathbf{L} = \begin{bmatrix} 1 & & & & \\ -m_{21} & 1 & & & \\ -m_{31} & -m_{32} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \\ -m_{n1} & -m_{n2} & \cdots & -m_{n,n-1} & 1 \end{bmatrix}, \tag{1.63}$$

and **U** is the upper triangular matrix obtained in Gaussian elimination. A lower triangular matrix with 1s on the diagonal is referred to as a *lower unit triangular matrix*.

The decomposition $\mathbf{A} = \mathbf{LU}$ is referred to as the "LU decomposition" or "LU factorization" of **A**. Note that the LU factors are obtained directly from the Gaussian elimination algorithm, since **U** is the upper triangular result, and **L** consists of the negatives of the row multipliers. Thus, the LU decomposition is merely a matrix interpretation of Gaussian elimination.

It is also useful to show a common storage scheme for the LU decomposition. If **A** is factored in place (i.e., the storage locations for **A** are over-written during the Gaussian elimination process), both **L** and **U** can be stored in the same n^2 storage locations used for **A**. In particular,

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1n} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2n} \\ A_{31} & A_{32} & A_{33} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nn} \end{bmatrix} \text{ is replaced by } \begin{bmatrix} U_{11} & U_{12} & U_{13} & \cdots & U_{1n} \\ L_{21} & U_{22} & U_{23} & \cdots & U_{2n} \\ L_{31} & L_{32} & U_{33} & \cdots & U_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & L_{n3} & \cdots & U_{nn} \end{bmatrix},$$

where the L_{ij} entries are the negatives of the multipliers obtained during Gaussian elimination. Since **L** is a lower unit triangular matrix (with 1s on the diagonal), it is not necessary to store the 1s. Each L_{ij} entry can be stored as it is computed.

We have seen how a lower unit triangular matrix can be formed from the product of **M** matrices. It can also be shown that the inverse of a matrix product is the product of the individual inverses in reverse order. We have also seen that the inverse of an **M** matrix is another **M** matrix: the one with the off-diagonal term negated. Thus, one might infer that the inverse of a lower unit triangular matrix is equal to the same matrix with the terms below the diagonal negated. However, such an inference would be incorrect, since, although the inverse of a lower unit triangular matrix is a lower triangular matrix equal to the product of **M** matrices, the **M** matrices are in the wrong order. Thus, the inverse of a lower unit triangular matrix is not equal to the same matrix with the terms below the diagonal negated, as the following simple example illustrates. Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 1 & 8 & 1 \end{bmatrix}. \tag{1.64}$$

We note that

$$\mathbf{A}^{-1} \neq \begin{bmatrix} 1 & 0 & 0 \\ -5 & 1 & 0 \\ -1 & -8 & 1 \end{bmatrix}, \tag{1.65}$$

since

$$\begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 1 & 8 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -5 & 1 & 0 \\ -1 & -8 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -40 & 0 & 1 \end{bmatrix} \neq \mathbf{I}. \tag{1.66}$$

To see why we cannot compute $\mathbf{A}^{-1}$ by simply negating the terms of $\mathbf{A}$ below the diagonal, consider the three matrices

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 8 & 1 \end{bmatrix}. \tag{1.67}$$

Is $\mathbf{A} = \mathbf{BCD}$ or $\mathbf{A} = \mathbf{DCB}$ or both? Notice that

$$\mathbf{BCD} = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 8 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 1 & 8 & 1 \end{bmatrix} = \mathbf{A}, \tag{1.68}$$

and

$$\mathbf{DCB} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 8 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 5 & 1 & 0 \\ 41 & 8 & 1 \end{bmatrix} \neq \mathbf{A}. \tag{1.69}$$

Thus, we conclude that order matters when $\mathbf{M}$ matrices are multiplied, and

$$\mathbf{BCD} \neq \mathbf{DCB}. \tag{1.70}$$

That is, the nice behavior which occurs when multiplying $\mathbf{M}$ matrices occurs only if the matrices are multiplied in the right order: the same order as that which the Gaussian elimination multipliers are computed. To complete this discussion, we compute $\mathbf{A}^{-1}$ as

$$\mathbf{A}^{-1} = \mathbf{D}^{-1}\mathbf{C}^{-1}\mathbf{B}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -8 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -5 & 1 & 0 \\ 39 & -8 & 1 \end{bmatrix}. \tag{1.71}$$

1.8 Matrix Interpretation of Back Substitution

Consider the linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ has been factored into $\mathbf{A} = \mathbf{LU}$. Then

$$\mathbf{LUx} = \mathbf{b}. \tag{1.72}$$

If we now define $\mathbf{Ux} = \mathbf{y}$, Eq. 1.72 is equivalent to the pair of equations

$$\begin{cases} \mathbf{Ly} = \mathbf{b}, \\ \mathbf{Ux} = \mathbf{y}, \end{cases} \tag{1.73}$$

which can be solved in two steps, first to compute $\mathbf{y}$ and then to compute $\mathbf{x}$. The combination of Gaussian elimination and backsolve can therefore be interpreted as the sequence of two matrix solutions: Eq. 1.73a is the modification of the original right-hand side $\mathbf{b}$ by the multipliers to yield $\mathbf{y}$. Eq. 1.73b is the backsolve, since the backsolve operation involves the coefficient matrix $\mathbf{U}$ to yield the solution $\mathbf{x}$. The two steps in Eq. 1.73 are often referred to as forward substitution and back substitution, respectively. The combination is sometimes

called the forward-backward substitution (FBS). Since the forward and back substitutions each require $n^2/2$ operations for a fully populated matrix, FBS requires n^2 operations.

A good example of why equations are solved with LU factorization followed by FBS is provided by transient structural dynamics. Transient response analysis of structures is concerned with computing the forced response due to general time-dependent loads (e.g., blast or shock). Linear finite element modeling results in the matrix differential equation

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{B}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}(t), \tag{1.74}$$

where $\mathbf{u}(t)$ is the vector of unknown displacements, t is time, $\mathbf{M}$ is the system mass matrix, $\mathbf{B}$ is the viscous damping matrix, $\mathbf{K}$ is the system stiffness matrix, $\mathbf{F}$ is the time-dependent vector of applied forces, and dots denote differentiation with respect to the time t. $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{K}$ are constant (independent of time). The problem is to determine $\mathbf{u}(t)$, given the initial displacement $\mathbf{u}_0$ and initial velocity $\dot{\mathbf{u}}_0$.

There are several numerical approaches for solving this system of equations, one of which integrates the equations directly in the time domain. Another approach, the modal superposition approach, will be discussed as an application of eigenvalue problems in Section 6.10 (p. 88). An example of a direct integrator is the Newmark-beta finite difference method (in time), which results in the recursive relation

$$\begin{aligned}\left(\frac{1}{\Delta t^2}\mathbf{M} + \frac{1}{2\Delta t}\mathbf{B} + \frac{1}{3}\mathbf{K}\right)\mathbf{u}_{n+1} &= \frac{1}{3}\left(\mathbf{F}_{n+1} + \mathbf{F}_n + \mathbf{F}_{n-1}\right) \\ &+ \left(\frac{2}{\Delta t^2}\mathbf{M} - \frac{1}{3}\mathbf{K}\right)\mathbf{u}_n + \left(-\frac{1}{\Delta t^2}\mathbf{M} + \frac{1}{2\Delta t}\mathbf{B} - \frac{1}{3}\mathbf{K}\right)\mathbf{u}_{n-1},\end{aligned} \tag{1.75}$$

where Δt is the integration time step, and $\mathbf{u}_n$ is the solution at the nth step. If the current time solution is $\mathbf{u}_n$, this matrix system is a system of linear algebraic equations whose solution $\mathbf{u}_{n+1}$ is the displacement solution at the next time step. The right-hand side depends on the current solution $\mathbf{u}_n$ and the immediately preceding solution $\mathbf{u}_{n-1}$. Since the original differential equation is second-order, it makes sense that two consecutive time solutions are needed to move forward, since a minimum of three consecutive time values is needed to estimate a second derivative using finite differences. Newmark-beta uses the initial conditions in a special start-up procedure to obtain $\mathbf{u}_0$ and $\mathbf{u}_{-1}$ to get the process started.

There are various issues with integrators (including stability, time step selection, and the start-up procedure), but our interest here is in the computational effort. Notice that, in Eq. 1.75, the coefficient matrices in parentheses that involve the time-independent matrices $\mathbf{M}$, $\mathbf{B}$, and $\mathbf{K}$ do not change during the integration unless the time step Δt changes. Thus, Eq. 1.75 is an example of a linear equation system of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$ for which many solutions are needed (one for each right-hand side $\mathbf{b}$) for the same coefficient matrix $\mathbf{A}$, but the right-hand sides are not known in advance. Thus, the solution strategy is to factor the left-hand side coefficient matrix once into $\mathbf{LU}$, save the factors, form the new rhs at each step, and then complete the solution with FBS. The computational effort is therefore one matrix decomposition (LU) for each unique Δt followed by two matrix/vector muliplications (to form the rhs) and one FBS at each step. Since LU costs considerably more than multiplication or FBS, many time steps can be computed for the same effort as one LU. As a result, one should not change the time step size Δt unnecessarily.

1.9 LDU Decomposition

Recall that any nonsingular square matrix $\mathbf{A}$ can be factored into $\mathbf{A} = \mathbf{LU}$, where $\mathbf{L}$ is a lower unit triangular matrix, and $\mathbf{U}$ is an upper triangular matrix. The diagonal entries in $\mathbf{U}$ can be factored out into a separate diagonal matrix $\mathbf{D}$ to yield

$$\mathbf{A} = \mathbf{LDU}, \tag{1.76}$$

where $\mathbf{D}$ is a nonsingular diagonal matrix, and $\mathbf{U}$ is an upper unit triangular matrix (not the same $\mathbf{U}$ as in $\mathbf{LU}$).

To verify the correctness of this variation of the LU decomposition, let

$$\mathbf{LU} = \mathbf{LD}\hat{\mathbf{U}}. \tag{1.77}$$

For a 3×3 matrix,

$$\mathbf{U} = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix} = \begin{bmatrix} D_{11} & 0 & 0 \\ 0 & D_{22} & 0 \\ 0 & 0 & D_{33} \end{bmatrix} \begin{bmatrix} 1 & \hat{U}_{12} & \hat{U}_{13} \\ 0 & 1 & \hat{U}_{23} \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{D}\hat{\mathbf{U}}, \tag{1.78}$$

where $D_{ii} = U_{ii}$ (no sum on i), and $\hat{U}_{ij} = U_{ij}/D_{ii}$ (no sum on i). That is,

$$\begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix} = \begin{bmatrix} U_{11} & 0 & 0 \\ 0 & U_{22} & 0 \\ 0 & 0 & U_{33} \end{bmatrix} \begin{bmatrix} 1 & U_{12}/U_{11} & U_{13}/U_{11} \\ 0 & 1 & U_{23}/U_{22} \\ 0 & 0 & 1 \end{bmatrix}. \tag{1.79}$$

For example,

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 0 & -8 & -2 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.80}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & -8 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 1/2 \\ 0 & 1 & 1/4 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{LD}\hat{\mathbf{U}}. \tag{1.81}$$

Note that, if $\mathbf{A}$ is symmetric ($\mathbf{A}^T = \mathbf{A}$),

$$\mathbf{A} = \mathbf{LDU} = (\mathbf{LDU})^T = \mathbf{U}^T\mathbf{D}\mathbf{L}^T, \tag{1.82}$$

or

$$\mathbf{U} = \mathbf{L}^T. \tag{1.83}$$

Thus, nonsingular symmetric matrices can be factored into

$$\mathbf{A} = \mathbf{LDL}^T. \tag{1.84}$$

1.10 Determinants

Let $\mathbf{A}$ be an $n \times n$ square matrix. For $n = 2$, we define the determinant of $\mathbf{A}$ as

$$\det \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = A_{11}A_{22} - A_{12}A_{21}. \tag{1.85}$$

For example,

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 1 \cdot 4 - 2 \cdot 3 = -2, \tag{1.86}$$

where the number of multiply-add operations is 2. For $n = 3$, we can expand by cofactors to obtain, for example,

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = 1 \begin{vmatrix} 5 & 6 \\ 8 & 9 \end{vmatrix} - 2 \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 3 \begin{vmatrix} 4 & 5 \\ 7 & 8 \end{vmatrix} = 0, \tag{1.87}$$

where the number of operations is approximately $3 \cdot 2 = 6$. In general, if we expand a determinant using cofactors, the number of operations to expand an $n \times n$ determinant is n times the number of operations to expand an $(n-1) \times (n-1)$ determinant. Thus, using the cofactor approach, the evaluation of the determinant of an $n \times n$ matrix requires about $n!$ operations (to first order), which is prohibitively expensive for all but the smallest of matrices.

Several properties of determinants are of interest and stated here without proof:

1. A cofactor expansion can be performed along any matrix row or column.
2. $\det \mathbf{A}^T = \det \mathbf{A}$
3. $\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B})$
4. $\det \mathbf{A}^{-1} = (\det \mathbf{A})^{-1}$ (a special case of Property 3)
5. The determinant of a triangular matrix is the product of the diagonals.
6. Interchanging two rows or two columns of a matrix changes the sign of the determinant.
7. Multiplying any single row or column by a constant multiplies the determinant by that constant. Consequently, for an $n \times n$ matrix $\mathbf{A}$ and scalar c, $\det(c\mathbf{A}) = c^n \det(\mathbf{A})$.
8. A matrix with a zero row or column has a zero determinant.
9. A matrix with two identical rows or two identical columns has a zero determinant.
10. If a row (or column) of a matrix is a constant multiple of another row (or column), the determinant is zero.
11. Adding a constant multiple of a row (or column) to another row (or column) leaves the determinant unchanged.

To illustrate the first property using the matrix of Eq. 1.87, we could also write

$$\begin{vmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{vmatrix} = -2 \begin{vmatrix} 4 & 6 \\ 7 & 9 \end{vmatrix} + 5 \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} - 8 \begin{vmatrix} 1 & 3 \\ 4 & 6 \end{vmatrix} = 0. \tag{1.88}$$

This property is sometimes convenient for matrices with many zeros, e.g.,

$$\begin{vmatrix} 1 & 0 & 3 \\ 0 & 5 & 0 \\ 7 & 0 & 9 \end{vmatrix} = 5 \begin{vmatrix} 1 & 3 \\ 7 & 9 \end{vmatrix} = -60. \tag{1.89}$$

Now consider the LU factorization $\mathbf{A} = \mathbf{LU}$:

$$\det \mathbf{A} = (\det \mathbf{L})(\det \mathbf{U}) = (1 \cdot 1 \cdot 1 \cdots 1)(U_{11}U_{22} \cdots U_{nn}) = U_{11}U_{22} \cdots U_{nn}. \tag{1.90}$$

This result implies that a nonsingular matrix has a nonzero determinant, since, if Gaussian elimination results in a $\mathbf{U}$ matrix with only nonzeros on the diagonal, $\mathbf{U}$ is nonsingular, which implies that $\mathbf{A}$ is nonsingular, and $\det \mathbf{A} \neq 0$. On the other hand, if $\mathbf{U}$ has one or more zeros on the diagonal, $\mathbf{U}$ and $\mathbf{A}$ are both singular, and $\det \mathbf{A} = 0$. Thus, a matrix is nonsingular if, and only if, its determinant is nonzero.

Eq. 1.90 also shows that, given $\mathbf{U}$, the determinant can be computed with an additional n operations. Given $\mathbf{A}$, calculating $\det \mathbf{A}$ requires about $n^3/3$ operations, since computing the LU factorization by Gaussian elimination requires about $n^3/3$ operations, a very small number compared to the $n!$ required by a cofactor expansion, as seen in the table below:

n	$n^3/3$	$n!$
5	42	120
10	333	3.6×10^6
25	5208	1.6×10^{25}
60	7.2×10^4	8.3×10^{81}

Determinants are rarely of interest in engineering computations or numerical mathematics, but, if needed, can be economically computed using the LU factorization.

1.11 Multiple Right-Hand Sides and Matrix Inverses

Consider the matrix system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is an $n \times n$ matrix, and both $\mathbf{x}$ and $\mathbf{b}$ are $n \times m$ matrices. That is, both the right-hand side $\mathbf{b}$ and the solution $\mathbf{x}$ have m columns, each corresponding to an independent right-hand side. Each column of the solution $\mathbf{x}$ corresponds to the same column of the right-hand side $\mathbf{b}$, as can be seen if the system is written in block form as

$$\mathbf{A}[\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \mathbf{x}_3] = [\mathbf{b}_1 \;\; \mathbf{b}_2 \;\; \mathbf{b}_3], \tag{1.91}$$

or

$$[\mathbf{Ax}_1 \;\; \mathbf{Ax}_2 \;\; \mathbf{Ax}_3] = [\mathbf{b}_1 \;\; \mathbf{b}_2 \;\; \mathbf{b}_3]. \tag{1.92}$$

Thus, a convenient and economical approach to solve a system with multiple right-hand sides is to factor $\mathbf{A}$ into $\mathbf{A} = \mathbf{LU}$ (an $n^3/3$ operation), and then apply FBS (an n^2 operation) to

each right-hand side. Unless the number m of right-hand sides is large, additional right-hand sides add little to the overall computational cost of solving the system.

Given a square matrix $\mathbf{A}$, the matrix inverse $\mathbf{A}^{-1}$, if it exists, is the matrix satisfying

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{1.93}$$

The calculation of $\mathbf{A}^{-1}$ is equivalent to solving a set of equations with coefficient matrix $\mathbf{A}$ and n right-hand sides, each one column of the identity matrix. The level of effort required is one LU factorization and n FBSs; that is,

$$\text{Number of operations} \approx \frac{1}{3}n^3 + n(n^2) = \frac{4}{3}n^3. \tag{1.94}$$

Thus, computing an inverse is only four times more expensive than solving a system with one right-hand side. In fact, if one exploits the special form of the right-hand side (the identity matrix, which has only one nonzero in each column), it turns out that the inverse can be computed in n^3 operations.

1.12 Tridiagonal Systems

Tridiagonal systems arise in a variety of engineering applications, including the finite difference solution of differential equations (e.g., the Crank-Nicolson finite difference method for solving parabolic partial differential equations) and cubic spline interpolation.

Tridiagonal systems are particularly easy (and fast) to solve using Gaussian elimination. It is convenient to solve such systems using the following notation:

$$\left.\begin{array}{rcrcrcrcl} d_1x_1 & + & u_1x_2 & & & & & = & b_1 \\ l_2x_1 & + & d_2x_2 & + & u_2x_3 & & & = & b_2 \\ & & l_3x_2 & + & d_3x_3 & + & u_3x_4 & = & b_3 \\ & & & & & & & \vdots & \\ & & & & l_nx_{n-1} & + & d_nx_n & = & b_n, \end{array}\right\} \tag{1.95}$$

where d_i, u_i, and l_i are, respectively, the diagonal, upper, and lower matrix entries in Row i. All coefficients can now be stored in three one-dimensional arrays, `D(·)`, `U(·)`, and `L(·)`, instead of a full two-dimensional array `A(I,J)`. The solution algorithm (reduction to upper triangular form by Gaussian elimination followed by backsolving) can now be summarized as follows:

1. For $k = 1, 2, \ldots, n-1$: [k = pivot row]
   $m = -l_{k+1}/d_k$ [m = multiplier needed to annihilate term below]
   $d_{k+1} = d_{k+1} + mu_k$ [new diagonal entry in next row]
   $b_{k+1} = b_{k+1} + mb_k$ [new rhs in next row]

2. $x_n = b_n/d_n$ [start of backsolve]

3. For $k = n-1, n-2, \ldots, 1$: [backsolve loop]
   $x_k = (b_k - u_k x_{k+1})/d_k$

To first order, this algorithm requires $5n$ operations.

1.13 Iterative Methods

Gaussian elimination is an example of a direct method, for which the number of operations required to effect a solution is predictable and fixed. Some systems of equations have characteristics that can be exploited by iterative methods, where the number of operations is not predictable in advance. For example, systems with large diagonal entries will converge rapidly with iterative methods.

Consider the linear system $\mathbf{Ax} = \mathbf{b}$. In Jacobi's method, one starts by assuming a solution $\mathbf{x}^{(0)}$, say, and then computing a new approximation to the solution using the equations

$$\begin{cases} x_1^{(1)} = \left(b_1 - A_{12}x_2^{(0)} - A_{13}x_3^{(0)} - \cdots - A_{1n}x_n^{(0)}\right)/A_{11} \\ x_2^{(1)} = \left(b_2 - A_{21}x_1^{(0)} - A_{23}x_3^{(0)} - \cdots - A_{2n}x_n^{(0)}\right)/A_{22} \\ \quad\vdots \\ x_n^{(1)} = \left(b_n - A_{n1}x_1^{(0)} - A_{n2}x_2^{(0)} - \cdots - A_{n,n-1}x_{n-1}^{(0)}\right)/A_{nn}. \end{cases} \tag{1.96}$$

In general,

$$x_i^{(k)} = \left(b_i - A_{i1}x_1^{(k-1)} - A_{i2}x_2^{(k-1)} - \cdots - A_{i,i-1}x_{i-1}^{(k-1)} - A_{i,i+1}x_{i+1}^{(k-1)} - \cdots - A_{in}x_n^{(k-1)}\right)/A_{ii}, \quad i = 1, 2, \ldots, n. \tag{1.97}$$

or

$$x_i^{(k)} = \left(b_i - \sum_{j \neq i} A_{ij}x_j^{(k-1)}\right)/A_{ii}, \quad i = 1, 2, \ldots, n. \tag{1.98}$$

In these equations, superscripts refer to the iteration number. If the vector $\mathbf{x}$ does not change much from one iteration to the next, the process has converged, and we have a solution.

Jacobi's method is also called *iteration by total steps* and the *method of simultaneous displacements*, since the order in which the equations are processed is irrelevant, and the updates could in principle be done simultaneously. Thus, Jacobi's method is nicely suited to parallel computation.

Notice that, in this algorithm, two consecutive approximations to $\mathbf{x}$ are needed at one time so that convergence can be checked. That is, it is not possible to update the vector $\mathbf{x}$ in place.

The Jacobi algorithm can be summarized as follows:

1. Input $n, \mathbf{A}, \mathbf{b}, N_{\max}, \mathbf{x}$ [$\mathbf{x}$ = initial guess]

2. k=0

3. k=k+1 [k = iteration number]

4. For $i = 1, 2, \ldots, n$:
$$y_i = \left(b_i - A_{i1}x_1 - A_{i2}x_2 - \cdots - A_{i,i-1}x_{i-1} - A_{i,i+1}x_{i+1} - \cdots - A_{in}x_n\right)/A_{ii}$$

5. If converged (compare $\mathbf{y}$ with $\mathbf{x}$), output $\mathbf{y}$ and exit; otherwise ...

6. $\mathbf{x} = \mathbf{y}$ [save latest iterate in $\mathbf{x}$]

7. If $k < N_{\max}$, go to Step 3; otherwise ...

8. No convergence; error termination.

In the above algorithm, the old iterates at each step are called $\mathbf{x}$, and the new ones are called $\mathbf{y}$.

For example, consider the system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}. \tag{1.99}$$

This system can be written in the form

$$\begin{cases} x_1 = (1 - x_2)/2 \\ x_2 = (-x_1 - x_3)/4 \\ x_3 = -x_2/3. \end{cases} \tag{1.100}$$

As a starting vector, we use

$$\mathbf{x}^{(0)} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}. \tag{1.101}$$

We summarize the iterations in a table:

k	x_1	x_2	x_3
0	0	0	0
1	0.5000	0	0
2	0.5000	-0.1250	0
3	0.5625	-0.1250	0.0417
4	0.5625	-0.1510	0.0417
5	0.5755	-0.1510	0.0503
6	0.5755	-0.1564	0.0503
7	0.5782	-0.1564	0.0521
8	0.5782	-0.1576	0.0521
9	0.5788	-0.1576	0.0525
10	0.5788	-0.1578	0.0525
11	0.5789	-0.1578	0.0526
12	0.5789	-0.1579	0.0526

For the precision displayed, this iteration has converged.

Notice that, in Jacobi's method, when $x_3^{(k)}$ is computed, we already know $x_2^{(k)}$ and $x_1^{(k)}$. Why not use the new values rather than $x_2^{(k-1)}$ and $x_1^{(k-1)}$? The iterative algorithm which uses at each step the "improved" values if they are available is called the Gauss-Seidel method. In general, in Gauss-Seidel,

$$x_i^{(k)} = \left(b_i - A_{i1}x_1^{(k)} - A_{i2}x_2^{(k)} - \cdots - A_{i,i-1}x_{i-1}^{(k)} - A_{i,i+1}x_{i+1}^{(k-1)} - \cdots - A_{in}x_n^{(k-1)}\right)/A_{ii},$$
$$i = 1, 2, \ldots, n. \tag{1.102}$$

This formula is used instead of Eq. 1.97.

The Gauss-Seidel algorithm can be summarized as follows:

1. Input $n, \mathbf{A}, \mathbf{b}, N_{\max}, \mathbf{x}$ [$\mathbf{x}$ = initial guess]
2. k=0
3. k=k+1 [k = iteration number]
4. $\mathbf{y} = \mathbf{x}$ [save previous iterate in $\mathbf{y}$]
5. For $i = 1, 2, \ldots, n$:
   $x_i = (b_i - A_{i1}x_1 - A_{i2}x_2 - \cdots - A_{i,i-1}x_{i-1} - A_{i,i+1}x_{i+1} - \cdots - A_{in}x_n)/A_{ii}$
6. If converged (compare $\mathbf{x}$ with $\mathbf{y}$), output $\mathbf{x}$ and exit; otherwise ...
7. If $k < N_{\max}$, go to Step 3; otherwise ...
8. No convergence; error termination.

We return to the example used to illustrate Jacobi's method:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 3 \end{bmatrix}, \quad \mathbf{b} = \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}. \tag{1.103}$$

The system $\mathbf{Ax} = \mathbf{b}$ can again be written in the form

$$\begin{cases} x_1 = (1 - x_2)/2 \\ x_2 = (-x_1 - x_3)/4 \\ x_3 = -x_2/3, \end{cases} \tag{1.104}$$

where we again use the starting vector

$$\mathbf{x}^{(0)} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}. \tag{1.105}$$

We summarize the Gauss-Seidel iterations in a table:

k	x_1	x_2	x_3
0	0	0	0
1	0.5000	-0.1250	0.0417
2	0.5625	-0.1510	0.0503
3	0.5755	-0.1564	0.0521
4	0.5782	-0.1576	0.0525
5	0.5788	-0.1578	0.0526
6	0.5789	-0.1578	0.0526

For this example, Gauss-Seidel converges is about half as many iterations as Jacobi.

In Gauss-Seidel iteration, the updates cannot be done simultaneously as in the Jacobi method, since each component of the new iterate depends upon all previously computed components. Also, each iterate depends upon the *order* in which the equations are processed, so that the Gauss-Seidel method is sometimes called the *method of successive displacements* to indicate the dependence of the iterates on the ordering. If the ordering of the equations is changed, the components of the new iterate will also change. It also turns out that the *rate* of convergence depends on the ordering.

In general, if the Jacobi method converges, the Gauss-Seidel method will converge faster than the Jacobi method. However, this statement is not always true. Consider, for example, the system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix}. \tag{1.106}$$

The solution of this system is $(-3, 3, 1)$. For this system, Jacobi converges, and Gauss-Seidel diverges. On the other hand, consider the system $\mathbf{Ax} = \mathbf{b}$ with

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & 3 \\ 4 & 7 & 4 \\ 3 & 4 & 4 \end{bmatrix}, \quad \mathbf{b} = \begin{Bmatrix} 12 \\ 15 \\ 11 \end{Bmatrix}. \tag{1.107}$$

For this system, whose solution is $(1, 1, 1)$, Jacobi diverges, and Gauss-Seidel converges.

Generally, for iterative methods such as Jacobi or Gauss-Seidel, the stopping criterion is either that the absolute change from one iteration to the next is small or that the relative change from one iteration to the next is small. That is, either

$$\max_i \left| x_i^{(k)} - x_i^{(k-1)} \right| < \varepsilon \tag{1.108}$$

or

$$\max_i \left| x_i^{(k)} - x_i^{(k-1)} \right| < \varepsilon \max_i \left| x_i^{(k-1)} \right|, \tag{1.109}$$

where ε is a small number chosen by the analyst.

We now return to the question of convergence of iterative methods. Consider Jacobi's method of solution for the system $\mathbf{Ax} = \mathbf{b}$. Define $\mathbf{M}$ as the iteration matrix for Jacobi's method; i.e., $\mathbf{M}$ is defined as the matrix such that

$$\mathbf{x}^{(k)} = \mathbf{M}\mathbf{x}^{(k-1)} + \mathbf{d}, \tag{1.110}$$

where

$$\mathbf{M} = \begin{bmatrix} 0 & -\frac{A_{12}}{A_{11}} & -\frac{A_{13}}{A_{11}} & \cdots & -\frac{A_{1n}}{A_{11}} \\ -\frac{A_{21}}{A_{22}} & 0 & -\frac{A_{23}}{A_{22}} & \cdots & -\frac{A_{2n}}{A_{22}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{A_{n1}}{A_{nn}} & -\frac{A_{n2}}{A_{nn}} & \cdots & -\frac{A_{n,n-1}}{A_{nn}} & 0 \end{bmatrix}, \quad d_i = b_i/A_{ii}. \tag{1.111}$$

Define the error at the kth iteration as

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}, \tag{1.112}$$

where $\mathbf{x}$ is the exact solution (unknown). By definition (Eq. 1.110),

$$\mathbf{M}\mathbf{x} + \mathbf{d} = \mathbf{x}. \tag{1.113}$$

Thus,

$$\mathbf{M}\mathbf{e}^{(k)} = \mathbf{M}\mathbf{x}^{(k)} - \mathbf{M}\mathbf{x} = \left(\mathbf{x}^{(k+1)} - \mathbf{d}\right) - (\mathbf{x} - \mathbf{d}) = \mathbf{x}^{(k+1)} - \mathbf{x} = \mathbf{e}^{(k+1)}. \tag{1.114}$$

That is, if the iteration matrix $\mathbf{M}$ acts on the error $\mathbf{e}^{(k)}$, $\mathbf{e}^{(k+1)}$ results. Thus, to have the error decrease at each iteration, we conclude that $\mathbf{M}$ should be small in some sense. The last equation can be written in terms of components as

$$e_i^{(k+1)} = -\sum_{j \neq i} \frac{A_{ij}}{A_{ii}} e_j^{(k)} \tag{1.115}$$

or, with absolute values,

$$\left|e_i^{(k+1)}\right| = \left|\sum_{j \neq i} \frac{A_{ij}}{A_{ii}} e_j^{(k)}\right|. \tag{1.116}$$

Since, for any numbers c_i,

$$\left|\sum_i c_i\right| \leq \sum_i |c_i|, \tag{1.117}$$

it follows that

$$\left|e_i^{(k+1)}\right| \leq \sum_{j \neq i} \left|\frac{A_{ij}}{A_{ii}} e_j^{(k)}\right| = \sum_{j \neq i} \left|\frac{A_{ij}}{A_{ii}}\right| \left|e_j^{(k)}\right| \leq \left(\sum_{j \neq i} \left|\frac{A_{ij}}{A_{ii}}\right|\right) \left|e_{\max}^{(k)}\right| = \alpha_i \left|e_{\max}^{(k)}\right|. \tag{1.118}$$

where we define

$$\alpha_i = \sum_{j \neq i} \left|\frac{A_{ij}}{A_{ii}}\right| = \frac{1}{|A_{ii}|} \sum_{j \neq i} |A_{ij}|. \tag{1.119}$$

Thus, Eq. 1.118 implies

$$\left|e_{\max}^{(k+1)}\right| \leq \alpha_{\max} \left|e_{\max}^{(k)}\right|. \tag{1.120}$$

If $\alpha_{\max} < 1$, the error decreases with each iteration, where $\alpha_{\max} < 1$ implies

$$|A_{ii}| > \sum_{j \neq i} |A_{ij}|, \quad i = 1, 2, \ldots, n. \tag{1.121}$$

A matrix which satisfies this condition is said to be *strictly diagonally dominant*. Therefore, a sufficient condition for Jacobi's method to converge is that the coefficient matrix $\mathbf{A}$ be strictly diagonally dominant. Notice that the starting vector does not matter, since it did not enter into this derivation. It turns out that strict diagonal dominance is also a sufficient condition for the Gauss-Seidel method to converge.

Note that strict diagonal dominance is a sufficient condition, but not necessary. As a practical matter, rapid convergence is desired, and rapid convergence can be assured only if the main diagonal coefficients are large compared to the off-diagonal coefficients. These sufficient conditions are weak in the sense that the iteration might converge even for systems not diagonally dominant, as seen in the examples of Eqs. 1.106 and 1.107.

2 Vector Spaces

A *vector space* is defined as a set of objects, called vectors, which can be combined using addition and multiplication under the following eight rules:

1. Addition is commutatative: $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$.
2. Addition is associative: $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$.
3. There exists a unique zero vector $\mathbf{0}$ such that $\mathbf{a} + \mathbf{0} = \mathbf{a}$.
4. There exists a unique negative vector such that $\mathbf{a} + (-\mathbf{a}) = \mathbf{0}$.
5. $1\,\mathbf{a} = \mathbf{a}$.
6. Multiplication is associative: $\alpha(\beta\mathbf{a}) = (\alpha\beta)\mathbf{a}$ for scalars α, β.
7. Multiplication is distributive with respect to vector addition: $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$.
8. Multiplication is distributive with respect to scalar addition: $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$.

Examples of vector spaces include the following:

1. The space of n-dimensional vectors: $\mathbf{a} = (a_1, a_2, \ldots, a_n)$.
2. The space of $m \times n$ matrices.
3. The space of functions $f(x)$.

Any set of objects that obeys the above eight rules is a vector space.

A *subspace* is a subset of a vector space which is closed under addition and scalar multiplication. The phrase "closed under addition and scalar multiplication" means that (1) if two vectors $\mathbf{a}$ and $\mathbf{b}$ are in the subspace, their sum $\mathbf{a}+\mathbf{b}$ also lies in the subspace, and (2) if a vector $\mathbf{a}$ is in the subspace, the scalar multiple $\alpha\mathbf{a}$ lies in the subspace. For example, a two-dimensional plane in 3-D is a subspace. However, polynomials of degree 2 are not a subspace, since, for example, we can add two polynomials of degree 2,

$$\left(-x^2+x+1\right)+\left(x^2+x+1\right)=2x+2, \tag{2.1}$$

and obtain a result which is a polynomial of degree 1. On the other hand, polynomials of degree ≤ 2 are a subspace.

The *column space* of a matrix $\mathbf{A}$ consists of all combinations of the columns of $\mathbf{A}$. For example, consider the 3×2 system

$$\begin{bmatrix} 1 & 0 \\ 5 & 4 \\ 2 & 4 \end{bmatrix} \begin{Bmatrix} u \\ v \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix}. \tag{2.2}$$

Since this system has three equations and two unknowns, it may not have a solution. An alternative way to write this system is

$$u\begin{Bmatrix} 1 \\ 5 \\ 2 \end{Bmatrix} + v\begin{Bmatrix} 0 \\ 4 \\ 4 \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix}, \tag{2.3}$$

where u and v are scalar multipliers of the two vectors. This system has a solution if, and only if, the right-hand side $\mathbf{b}$ can be written as a linear combination of the two columns. Geometrically, the system has a solution if, and only if, $\mathbf{b}$ lies in the plane formed by the two vectors. That is, the system has a solution if, and only if, $\mathbf{b}$ lies in the column space of $\mathbf{A}$.

We note that the column space of an $m\times n$ matrix $\mathbf{A}$ is a subspace of m-dimensional space $\mathbb{R}^m$. For example, in the 3×2 matrix above, the column space is the plane in 3-D formed by the two vectors $(1,5,2)$ and $(0,4,4)$.

The *null space* of a matrix $\mathbf{A}$ consists of all vectors $\mathbf{x}$ such that $\mathbf{Ax}=\mathbf{0}$.

2.1 Rectangular Systems of Equations

For square systems of equations $\mathbf{Ax}=\mathbf{b}$, there is one solution (if $\mathbf{A}^{-1}$ exists) or zero or infinite solutions (if $\mathbf{A}^{-1}$ does not exist). The situation with rectangular systems is a little different.

For example, consider the 3×4 system

$$\begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix}. \tag{2.4}$$

We can simplify this system using elementary row operations to obtain

$$\left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & b_1 \\ 2 & 6 & 9 & 5 & b_2 \\ -1 & -3 & 3 & 0 & b_3 \end{array}\right] \longrightarrow \left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & b_1 \\ 0 & 0 & 3 & 1 & b_2 - 2b_1 \\ 0 & 0 & 6 & 2 & b_1 + b_3 \end{array}\right] \tag{2.5}$$

$$\longrightarrow \left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & b_1 \\ 0 & 0 & 3 & 1 & b_2 - 2b_1 \\ 0 & 0 & 0 & 0 & b_3 - 2b_2 + 5b_1 \end{array}\right], \tag{2.6}$$

which implies that the solution exists if, and only if,

$$b_3 - 2b_2 + 5b_1 = 0. \tag{2.7}$$

Thus, for this system, even though there are more unknowns than equations, there may be no solution.

Now assume that a solution of Eq. 2.4 exists. For example, let

$$\mathbf{b} = \left\{\begin{array}{c} 1 \\ 5 \\ 5 \end{array}\right\}, \tag{2.8}$$

in which case the row-reduced system is

$$\left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & 1 \\ 0 & 0 & 3 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right], \tag{2.9}$$

which implies, from the second equation,

$$3x_3 + x_4 = 3. \tag{2.10}$$

One of the two unknowns (x_4, say) can be picked arbitrarily, in which case

$$x_3 = 1 - \frac{1}{3}x_4. \tag{2.11}$$

From the first equation in Eq. 2.9, we obtain

$$x_1 = 1 - 3x_2 - 3x_3 - 2x_4 = 1 - 3x_2 - 3(1 - x_4/3) - 2x_4 = -2 - 3x_2 - x_4, \tag{2.12}$$

which implies that the solution can be written in the form

$$\mathbf{x} = \left\{\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\right\} = \left\{\begin{array}{c} -2 \\ 0 \\ 1 \\ 0 \end{array}\right\} + x_2 \left\{\begin{array}{c} -3 \\ 1 \\ 0 \\ 0 \end{array}\right\} + x_4 \left\{\begin{array}{c} -1 \\ 0 \\ -1/3 \\ 1 \end{array}\right\}. \tag{2.13}$$

This system thus has a doubly infinite set of solutions, i.e., there are two free parameters.

Now consider the corresponding homogeneous system (with zero right-hand side):

$$\mathbf{A}\mathbf{x}_\mathrm{h} = \mathbf{0}, \tag{2.14}$$

where, from Eq. 2.9, the row-reduced system is

$$\left[\begin{array}{cccc|c} 1 & 3 & 3 & 2 & 0 \\ 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right], \tag{2.15}$$

which implies

$$3x_3 = -x_4 \tag{2.16}$$

and

$$x_1 = -3x_2 - 3x_3 - 2x_4 = -3x_2 - x_4. \tag{2.17}$$

Thus, the solution of the homogeneous system can be written

$$\mathbf{x}_\mathrm{h} = x_2 \left\{\begin{array}{r} -3 \\ 1 \\ 0 \\ 0 \end{array}\right\} + x_4 \left\{\begin{array}{r} -1 \\ 0 \\ -1/3 \\ 1 \end{array}\right\}. \tag{2.18}$$

Therefore, from Eqs. 2.13 and 2.18, the solution of the nonhomogeneous system $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be written as the sum of a particular solution $\mathbf{x}_\mathrm{p}$ and the solution of the corresponding homogeneous problem $\mathbf{A}\mathbf{x}_\mathrm{h} = \mathbf{0}$:

$$\mathbf{x} = \mathbf{x}_\mathrm{p} + \mathbf{x}_\mathrm{h}. \tag{2.19}$$

Alternatively, for the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, given a solution $\mathbf{x}$, any solution of the corresponding homogeneous system can be added to get another solution, since

$$\mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}_\mathrm{p} + \mathbf{x}_\mathrm{h}) = \mathbf{A}\mathbf{x}_\mathrm{p} + \mathbf{A}\mathbf{x}_\mathrm{h} = \mathbf{b} + \mathbf{0} = \mathbf{b}. \tag{2.20}$$

The final row-reduced matrix is said to be in *echelon* form if it looks like this:

$$\left[\begin{array}{cccccc} \textcircled{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & \textcircled{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & \textcircled{x} & \mathrm{x} & \mathrm{x} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

where nonzeros are denoted with the letter "x". The echelon form of a matrix has the following characteristics:

1. The nonzero rows appear first.
2. The pivots (circled above) are the first nonzeros in each row.
3. Only zeros appear below each pivot.
4. Each pivot appears to the right of the pivot in the row above.

The number of nonzero rows in the row-reduced (echelon) matrix is referred to as the *rank* of the system. Thus, matrix rank is the number of independent rows in a matrix.

2.2 Linear Independence

The vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are *linearly independent* if the linear combination

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \tag{2.21}$$

implies $c_1 = c_2 = \cdots = c_k = 0$. We note that, if the coefficients c_i are not all zero, then one vector can be written as a linear combination of the others.

For the special case of two vectors, two vectors (of any dimension) are linearly dependent if one is a multiple of the other, i.e., $\mathbf{v}_1 = c\mathbf{v}_2$. Geometrically, if two vectors are dependent, they are parallel. Three vectors (of any dimension) are linearly dependent if they are coplanar.

For example, let us determine if the three vectors

$$\mathbf{v}_1 = \begin{Bmatrix} 3 \\ 3 \\ 3 \end{Bmatrix}, \quad \mathbf{v}_2 = \begin{Bmatrix} 4 \\ 5 \\ 4 \end{Bmatrix}, \quad \mathbf{v}_3 = \begin{Bmatrix} 2 \\ 7 \\ 4 \end{Bmatrix} \tag{2.22}$$

are independent. Note that Eq. 2.21 implies

$$\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \end{bmatrix} \begin{Bmatrix} c_1 \\ c_2 \\ c_3 \end{Bmatrix} = \mathbf{0} \tag{2.23}$$

or

$$\begin{bmatrix} 3 & 4 & 2 \\ 3 & 5 & 7 \\ 3 & 4 & 4 \end{bmatrix} \begin{Bmatrix} c_1 \\ c_2 \\ c_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}, \tag{2.24}$$

so that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are independent if $c_1 = c_2 = c_3 = 0$ is the only solution of this system. Thus, to determine the independence of vectors, we can arrange the vectors in the columns of a matrix, and determine the null space of the matrix:

$$\left[\begin{array}{ccc|c} 3 & 4 & 2 & 0 \\ 3 & 5 & 7 & 0 \\ 3 & 4 & 4 & 0 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 3 & 4 & 2 & 0 \\ 0 & 1 & 5 & 0 \\ 0 & 0 & 2 & 0 \end{array}\right]. \tag{2.25}$$

This system implies $c_1 = c_2 = c_3 = 0$, since an upper triangular system with a nonzero diagonal is nonsingular. Thus, the three given vectors are independent.

The previous example also shows that the columns of a matrix are linearly independent if, and only if, the null space of the matrix is the zero vector.

Let r denote the rank of a matrix (the number of independent rows). We observe that the number of independent columns is also r, since each column of an echelon matrix with a pivot has a nonzero component in a new location. For example, consider the echelon matrix

$$\begin{bmatrix} \textcircled{1} & 3 & 3 & 2 \\ 0 & 0 & \textcircled{3} & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where the pivots are circled. Column 1 has one nonzero component. Column 2 is not independent, since it is proportional to Column 1. Column 3 has two nonzero components,

Figure 2: Some Vectors That Span the xy-Plane.

so Column 3 cannot be dependent on the first two columns. Column 4 also has two nonzero components, but the column can be written uniquely as a linear combination of Columns 1 and 3 by picking first the Column 3 multiplier (1/3) followed by the Column 1 multiplier. Thus we see that, for a matrix, the row rank is equal to the column rank.

2.3 Spanning a Subspace

If every vector $\mathbf{v}$ in a vector space can be expressed as a linear combination

$$\mathbf{v} = c_1\mathbf{w}_1 + c_2\mathbf{w}_2 + \cdots + c_k\mathbf{w}_k \tag{2.26}$$

for some coefficients c_i, then the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k$ are said to *span* the space. For example, in 2-D, the vectors (1,0) and (0,1) span the xy-plane (Fig. 2). The two vectors (1,0) and (1,1) also span the xy-plane. Since the number of spanning vectors need not be minimal, the three vectors (1,0), (0,1), and (1,1) also span the xy-plane, although only two of the three vectors are needed.

The *column space* of a matrix is the space spanned by the columns of the matrix. This definition is equivalent to that given earlier (all possible combinations of the columns).

A *basis* for a vector space is a set of vectors which is linearly independent and spans the space. The requirement for linear independence means that there can be no extra vectors. For example, the two vectors (1,0) and (0,1) in Fig. 2 form a basis for $\mathbb{R}^2$. The two vectors (1,0) and (1,1) also form a basis for $\mathbb{R}^2$. However, the three vectors (1,0), (0,1), and (1,1) do not form a basis for $\mathbb{R}^2$, since one of the three vectors is linearly dependent on the other two. Thus, we see that all bases for a vector space V contain the same number of vectors; that number is referred to as the *dimension* of V.

2.4 Row Space and Null Space

The *row space* of a matrix $\mathbf{A}$ consists of all possible combinations of the rows of $\mathbf{A}$. Consider again the 3×4 example used in the discussion of rectangular systems:

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 2 & 6 & 9 & 5 \\ -1 & -3 & 3 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 6 & 2 \end{bmatrix} \longrightarrow \begin{bmatrix} \textcircled{1} & 3 & 3 & 2 \\ 0 & 0 & \textcircled{3} & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \mathbf{U}, \tag{2.27}$$

where the pivots are circled. $\mathbf{A}$ has two independent rows, i.e., $\text{rank}(\mathbf{A}) = 2$. Also, since the rows of the final matrix $\mathbf{U}$ are simply linear combinations of the rows of $\mathbf{A}$, the row spaces of both $\mathbf{A}$ and $\mathbf{U}$ are the same. Again we see that the number of pivots is equal to the rank r,

and that the number of independent columns is also r. That is, the number of independent columns is equal to the number of independent rows. Hence, for a matrix,

$$\text{Row rank} = \text{Column rank}. \tag{2.28}$$

We recall that the null space of a matrix $\mathbf{A}$ consists of all vectors $\mathbf{x}$ such that $\mathbf{Ax} = \mathbf{0}$. With elementary row operations, $\mathbf{A}$ can be transformed into echelon form, i.e.,

$$\mathbf{Ax} = \mathbf{0} \longrightarrow \mathbf{Ux} = \mathbf{0}. \tag{2.29}$$

For the example of the preceding section,

$$\mathbf{U} = \begin{bmatrix} 1 & 3 & 3 & 2 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2.30}$$

which implies

$$x_1 + 3x_2 + 3x_3 + 2x_4 = 0 \tag{2.31}$$

and

$$3x_3 + x_4 = 0. \tag{2.32}$$

With four unknowns and two independent equations, we could, for example, choose x_4 arbitrarily in Eq. 2.32 and x_2 arbitrarily in Eq. 2.31. In general, for a rectangular system with m equations and n unknowns (i.e., $\mathbf{A}$ is an $m \times n$ matrix), the number of free variables (the dimension of the null space) is $n - r$, where r is the matrix rank.

We thus summarize the dimensions of the various spaces in the following table. For an $m \times n$ matrix $\mathbf{A}$ (m rows, n columns):

Space	Dimension	Subspace of
Row	r	$\mathbb{R}^n$
Column	r	$\mathbb{R}^m$
Null	$n - r$	$\mathbb{R}^n$

Note that

$$\dim(\text{row space}) + \dim(\text{null space}) = n = \text{number of columns}. \tag{2.33}$$

This property will be useful in the discussions of pseudoinverses and least squares.

2.5 Pseudoinverses

Rectangular $m \times n$ matrices cannot be inverted in the usual sense. However, one can compute a *pseudoinverse* by considering either $\mathbf{AA}^T$ or $\mathbf{A}^T\mathbf{A}$, whichever (it turns out) is smaller. We consider two cases: $m < n$ and $m > n$. Notice that, for any matrix $\mathbf{A}$, both $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$ are square and symmetric.

For $m < n$, consider $\mathbf{AA}^T$, which is a square $m \times m$ matrix. If $\mathbf{AA}^T$ is invertible,

$$\left(\mathbf{AA}^T\right)\left(\mathbf{AA}^T\right)^{-1} = \mathbf{I} \tag{2.34}$$

or

$$\mathbf{A}\left[\mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}\right] = \mathbf{I}, \tag{2.35}$$

where the matrix $\mathbf{C} = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}$ in brackets is referred to as the pseudo right inverse. The reason for choosing the order of the two inverses given in Eq. 2.34 is that this order allows the matrix $\mathbf{A}$ to be isolated in the next equation.

For the second case, $m > n$, consider $\mathbf{A}^T\mathbf{A}$, which is an $n \times n$ matrix. If $\mathbf{A}^T\mathbf{A}$ is invertible,

$$\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{A}\right) = \mathbf{I} \tag{2.36}$$

or

$$\left[\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\right]\mathbf{A} = \mathbf{I}, \tag{2.37}$$

where the matrix $\mathbf{B} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T$ in brackets is referred to as the pseudo left inverse.

For example, consider the 2×3 matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix}, \tag{2.38}$$

in which case

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix}\begin{bmatrix} 4 & 0 \\ 0 & 5 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 25 \end{bmatrix}, \tag{2.39}$$

and the pseudo right inverse is

$$\mathbf{C} = \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1} = \begin{bmatrix} 4 & 0 \\ 0 & 5 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1/16 & 0 \\ 0 & 1/25 \end{bmatrix} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/5 \\ 0 & 0 \end{bmatrix}. \tag{2.40}$$

We check the correctness of this result by noting that

$$\mathbf{A}\mathbf{C} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \end{bmatrix}\begin{bmatrix} 1/4 & 0 \\ 0 & 1/5 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}. \tag{2.41}$$

One of the most important properties associated with pseudoinverses is that, for any matrix $\mathbf{A}$,

$$\text{rank}\left(\mathbf{A}^T\mathbf{A}\right) = \text{rank}\left(\mathbf{A}\mathbf{A}^T\right) = \text{rank}(\mathbf{A}). \tag{2.42}$$

To show that

$$\text{rank}\left(\mathbf{A}^T\mathbf{A}\right) = \text{rank}(\mathbf{A}), \tag{2.43}$$

we first recall that, for any matrix $\mathbf{A}$,

$$\dim(\text{row space}) + \dim(\text{null space}) = \text{number of columns} \tag{2.44}$$

or

$$r + \dim(\text{null space}) = n. \tag{2.45}$$

Since $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}$ have the same number of columns (n), if suffices to show that $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}$ have the same null space. That is, we want to prove that

$$\mathbf{A}\mathbf{x} = \mathbf{0} \iff \mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}. \tag{2.46}$$

The forward direction ($\Longrightarrow$) is clear by inspection. For the other direction ($\Longleftarrow$), consider

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}. \tag{2.47}$$

If we take the dot product of both sides with $\mathbf{x}$,

$$0 = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{x}) = |\mathbf{A}\mathbf{x}|^2, \tag{2.48}$$

which implies (since a vector of zero length must be the zero vector)

$$\mathbf{A}\mathbf{x} = \mathbf{0}. \tag{2.49}$$

Thus, since $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}$ have both the same null space and the same number of columns (n), the dimension r of the row space of the two matrices is the same, and the first part of Eq. 2.42 is proved. Also, with $\mathbf{A}^T = \mathbf{B}$,

$$\text{rank}\left(\mathbf{A}\mathbf{A}^T\right) = \text{rank}\left(\mathbf{B}^T\mathbf{B}\right) = \text{rank}(\mathbf{B}) = \text{rank}(\mathbf{B}^T) = \text{rank}(\mathbf{A}), \tag{2.50}$$

and the second part of Eq. 2.42 is proved. The implication of Eq. 2.42 is that, if rank($\mathbf{A}$) = n (i.e., $\mathbf{A}$'s columns are independent), then $\mathbf{A}^T\mathbf{A}$ is invertible.

It will also be shown in §4 that, for the overdetermined system $\mathbf{A}\mathbf{x} = \mathbf{b}$ (with $m > n$), the pseudoinverse solution using $\mathbf{A}^T\mathbf{A}$ is equivalent to the least squares solution.

2.6 Linear Transformations

If $\mathbf{A}$ is an $m \times n$ matrix, and $\mathbf{x}$ is an n-vector, the matrix product $\mathbf{A}\mathbf{x}$ can be thought of as a transformation of $\mathbf{x}$ into another vector $\mathbf{x}'$:

$$\mathbf{A}\mathbf{x} = \mathbf{x}', \tag{2.51}$$

where $\mathbf{x}'$ is an $m \times 1$ vector. We consider some examples:

1. Stretching

   Let

   $$\mathbf{A} = \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} = c\mathbf{I}, \tag{2.52}$$

   in which case

   $$\mathbf{A}\mathbf{x} = \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} cx_1 \\ cx_2 \end{Bmatrix} = c \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}. \tag{2.53}$$

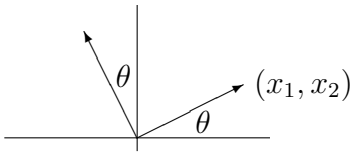

Figure 3: 90° Rotation.

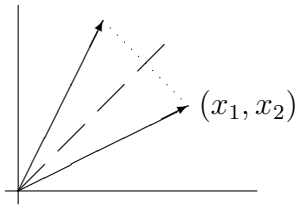

Figure 4: Reflection in 45° Line.

2. 90° Rotation (Fig. 3)

   Let

$$\mathbf{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \tag{2.54}$$

   in which case

$$\mathbf{Ax} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} -x_2 \\ x_1 \end{Bmatrix}. \tag{2.55}$$

3. Reflection in 45° Line (Fig. 4)

   Let

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{2.56}$$

   in which case

$$\mathbf{Ax} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} x_2 \\ x_1 \end{Bmatrix}. \tag{2.57}$$

4. Projection Onto Horizontal Axis (Fig. 5)

   Let

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.58}$$

   in which case

$$\mathbf{Ax} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} x_1 \\ 0 \end{Bmatrix}. \tag{2.59}$$

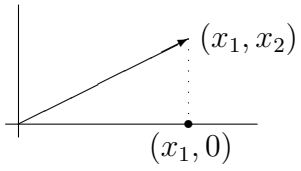

Figure 5: Projection Onto Horizontal Axis.

Note that transformations which can be represented by matrices are linear transformations, since

$$\mathbf{A}(c\mathbf{x} + d\mathbf{y}) = c(\mathbf{A}\mathbf{x}) + d(\mathbf{A}\mathbf{y}), \tag{2.60}$$

where $\mathbf{x}$ and $\mathbf{y}$ are vectors, and c and d are scalars. This property implies that knowing how $\mathbf{A}$ transforms each vector in a basis determines how $\mathbf{A}$ transforms every vector in the entire space, since the vectors in the space are linear combinations of the basis vectors:

$$\mathbf{A}(c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n) = c_1(\mathbf{A}\mathbf{x}_1) + c_2(\mathbf{A}\mathbf{x}_2) + \cdots + c_n(\mathbf{A}\mathbf{x}_n), \tag{2.61}$$

where $\mathbf{x}_i$ is the ith basis vector, and c_i is the ith scalar multiplier.

We now continue with the examples of linear transformations:

5. Differentiation of Polynomials

   Consider the polynomial of degree 4

   $$p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4. \tag{2.62}$$

   The basis vectors are

   $$p_1 = 1, \quad p_2 = t, \quad p_3 = t^2, \quad p_4 = t^3, \quad p_5 = t^4, \tag{2.63}$$

   which implies that, in terms of the basis vectors,

   $$p(t) = a_0 p_1 + a_1 p_2 + a_2 p_3 + a_3 p_4 + a_4 p_5. \tag{2.64}$$

   The coefficients a_i become the components of the vector p. If we differentiate the polynomial, we obtain

   $$p'(t) = a_1 + 2a_2 t + 3a_3 t^2 + 4a_4 t^3 = a_1 p_1 + 2a_2 p_2 + 3a_3 p_3 + 4a_4 p_4. \tag{2.65}$$

   Differentiation of polynomials of degree 4 can thus be written in matrix form as

   $$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{Bmatrix} = \begin{Bmatrix} a_1 \\ 2a_2 \\ 3a_3 \\ 4a_4 \end{Bmatrix}, \tag{2.66}$$

   so that the matrix which represents differentiation of polynomials of degree 4 is

   $$\mathbf{A}_{\mathrm{D}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}, \tag{2.67}$$

   where the subscript "D" denotes differentiation. Note that this matrix has been written as a rectangular 4×5 matrix, since differentiation reduces the degree of the polynomial by unity.

6. Integration of Polynomials

Let

$$p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3, \tag{2.68}$$

which can be integrated to obtain

$$\int_0^t p(\tau)\, d\tau = a_0 t + \frac{a_1}{2} t^2 + \frac{a_2}{3} t^3 + \frac{a_3}{4} t^4 = a_0 p_2 + \frac{a_1}{2} p_3 + \frac{a_2}{3} p_4 + \frac{a_3}{4} p_5. \tag{2.69}$$

Thus, integration of polynomials of degree 3 can be written in matrix form as

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ a_0 \\ \frac{1}{2} a_1 \\ \frac{1}{3} a_2 \\ \frac{1}{4} a_3 \end{Bmatrix}, \tag{2.70}$$

so that the matrix which represents integration of polynomials of degree 3 is

$$\mathbf{A}_{\mathrm{I}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix}, \tag{2.71}$$

where the subscript "I" denotes integration. We note that integration followed by differentiation is an inverse operation which restores the vector to its original:

$$\mathbf{A}_{\mathrm{D}} \mathbf{A}_{\mathrm{I}} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbf{I}. \tag{2.72}$$

Thus, $\mathbf{A}_{\mathrm{D}}$ is the pseudo left inverse of $\mathbf{A}_{\mathrm{I}}$. On the other hand, performing the operations in reverse order (differentiation followed by integation) does not restore the vector to its original:

$$\mathbf{A}_{\mathrm{I}} \mathbf{A}_{\mathrm{D}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \neq \mathbf{I}. \tag{2.73}$$

Integration after differentiation does not restore the original constant term.

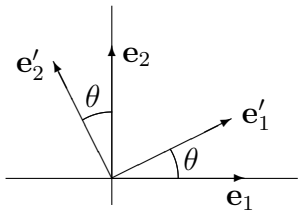

Figure 6: Rotation by Angle θ.

7. Rotation by Angle θ (Fig. 6)

Consider the 2-D transformation

$$\mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2.74}$$

If $\mathbf{v}$ is a typical vector in 2-D, $\mathbf{v}$ can be written

$$\mathbf{v} = \begin{Bmatrix} v_1 \\ v_2 \end{Bmatrix} = v_1 \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} + v_2 \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2, \tag{2.75}$$

where $\mathbf{e}_1$ and $\mathbf{e}_2$ are the basis vectors. We now consider the effect of $\mathbf{R}$ on the two basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$:

$$\mathbf{R}\mathbf{e}_1 = \begin{Bmatrix} \cos\theta \\ \sin\theta \end{Bmatrix} = \mathbf{e}_1', \quad \mathbf{R}\mathbf{e}_2 = \begin{Bmatrix} -\sin\theta \\ \cos\theta \end{Bmatrix} = \mathbf{e}_2', \tag{2.76}$$

as shown in Fig. 6. Since $\mathbf{e}_1$ and $\mathbf{e}_2$ are both rotated by θ, so is $\mathbf{v}$. The transformation $\mathbf{R}$ has the following properties:

(a) $\mathbf{R}_\theta$ and $\mathbf{R}_{-\theta}$ are inverses (rotation by θ and $-\theta$):

$$\mathbf{R}_{-\theta}\mathbf{R}_\theta = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}, \tag{2.77}$$

where we used $\sin(-\theta) = -\sin\theta$ and $\cos(-\theta) = \cos\theta$ to compute $\mathbf{R}_{-\theta}$. Thus, $\mathbf{R}$ is an orthogonal matrix ($\mathbf{R}\mathbf{R}^T = \mathbf{I}$). Note also that $\mathbf{R}_{-\theta} = \mathbf{R}_\theta^{-1} = \mathbf{R}^T$.

(b) $\mathbf{R}_{2\theta} = \mathbf{R}_\theta\mathbf{R}_\theta$ (rotation through the double angle):

$$\mathbf{R}_\theta\mathbf{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{2.78}$$

$$= \begin{bmatrix} \cos^2\theta - \sin^2\theta & -2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \cos^2\theta - \sin^2\theta \end{bmatrix} \tag{2.79}$$

$$= \begin{bmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix} = \mathbf{R}_{2\theta} \tag{2.80}$$

The above equations assume the double-angle trigonometric identities as known, but an alternative view of these equations would be to assume that $\mathbf{R}_{2\theta} = \mathbf{R}_\theta\mathbf{R}_\theta$ must be true, and view these equations as a derivation of the double-angle identities.

(c) $\mathbf{R}_{\theta+\phi} = \mathbf{R}_\phi \mathbf{R}_\theta$ (rotation through θ, then ϕ):

$$\mathbf{R}_\phi \mathbf{R}_\theta = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{2.81}$$

$$= \begin{bmatrix} \cos\phi\cos\theta - \sin\phi\sin\theta & -\cos\phi\sin\theta - \sin\phi\cos\theta \\ \sin\phi\cos\theta + \cos\phi\sin\theta & -\sin\phi\sin\theta + \cos\phi\cos\theta \end{bmatrix} \tag{2.82}$$

$$= \begin{bmatrix} \cos(\phi+\theta) & -\sin(\phi+\theta) \\ \sin(\phi+\theta) & \cos(\phi+\theta) \end{bmatrix} = \mathbf{R}_{\phi+\theta} \tag{2.83}$$

We note that, in 2-D rotations (but not in 3-D), the order of rotations does not matter, i.e.,

$$\mathbf{R}_\phi \mathbf{R}_\theta = \mathbf{R}_\theta \mathbf{R}_\phi. \tag{2.84}$$

Properties (a) and (b) are special cases of Property (c).

2.7 Orthogonal Subspaces

For two vectors $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, we recall that the inner product (or dot product) of $\mathbf{u}$ and $\mathbf{v}$ is defined as

$$(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^{n} u_i v_i. \tag{2.85}$$

The length of $\mathbf{u}$ is

$$|\mathbf{u}| = \sqrt{u_1^2 + u_2^2 + \cdots + u_n^2} = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{\sum_{i=1}^{n} u_i u_i} \tag{2.86}$$

The vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if $\mathbf{u} \cdot \mathbf{v} = 0$.

We note first that a set of mutually orthogonal vectors is linearly independent. To prove this assertion, let vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ be mutually orthogonal. To prove independence, we must show that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \tag{2.87}$$

implies $c_i = 0$ for all i. If we take the dot product of both sides of Eq. 2.87 with $\mathbf{v}_1$, we obtain

$$\mathbf{v}_1 \cdot (c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k) = 0, \tag{2.88}$$

where orthogonality implies $\mathbf{v}_1 \cdot \mathbf{v}_i = 0$ for $i \neq 1$. Thus, from Eq. 2.88,

$$c_1\mathbf{v}_1 \cdot \mathbf{v}_1 = c_1|\mathbf{v}_1|^2 = 0. \tag{2.89}$$

Since $|\mathbf{v}_1| \neq 0$, we conclude that $c_1 = 0$. Similarly, $c_2 = c_3 = \cdots = c_k = 0$, and the original assertion is proved.

We define two subspaces as orthogonal if every vector in one subspace is orthogonal to every vector in the other subspace. For example, the xy-plane is a subspace of $\mathbb{R}^3$, and the orthogonal subspace is the z-axis.

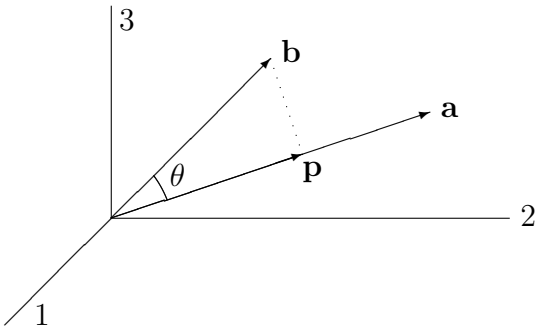

Figure 7: Projection Onto Line.

We now show that the row space of a matrix is orthogonal to the null space. To prove this assertion, consider a matrix $\mathbf{A}$ for which $\mathbf{x}$ is in the nullspace, and $\mathbf{v}$ is in the row space. Thus,

$$\mathbf{A}\mathbf{x} = \mathbf{0} \tag{2.90}$$

and

$$\mathbf{v} = \mathbf{A}^T\mathbf{w} \tag{2.91}$$

for some $\mathbf{w}$. That is, $\mathbf{v}$ is a combination of the columns of $\mathbf{A}^T$ or rows of $\mathbf{A}$. Then,

$$\mathbf{v}^T\mathbf{x} = \mathbf{w}^T\mathbf{A}\mathbf{x} = 0, \tag{2.92}$$

thus proving the assertion.

2.8 Projections Onto Lines

Consider two n-dimensional vectors $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$. We wish to project $\mathbf{b}$ onto $\mathbf{a}$, as shown in Fig. 7 in 3-D. Let $\mathbf{p}$ be the vector obtained by projecting $\mathbf{b}$ onto $\mathbf{a}$. The scalar projection of $\mathbf{b}$ onto $\mathbf{a}$ is $|\mathbf{b}|\cos\theta$, where θ is the angle between $\mathbf{b}$ and $\mathbf{a}$. Thus,

$$\mathbf{p} = (|\mathbf{b}|\cos\theta)\frac{\mathbf{a}}{|\mathbf{a}|}, \tag{2.93}$$

where the fraction in this expression is the unit vector in the direction of $\mathbf{a}$. Since

$$\mathbf{a}\cdot\mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta, \tag{2.94}$$

Eq. 2.93 becomes

$$\mathbf{p} = |\mathbf{b}|\,\frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}\,\frac{\mathbf{a}}{|\mathbf{a}|} = \frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}}\,\mathbf{a}. \tag{2.95}$$

This result can be obtained by inspection by noting that the scalar projection of $\mathbf{b}$ onto $\mathbf{a}$ is the dot product of $\mathbf{b}$ with a unit vector in the direction of $\mathbf{a}$. The vector projection would then be this scalar projection times a unit vector in the direction of $\mathbf{a}$:

$$\mathbf{p} = \left(\mathbf{b}\cdot\frac{\mathbf{a}}{|\mathbf{a}|}\right)\frac{\mathbf{a}}{|\mathbf{a}|}. \tag{2.96}$$

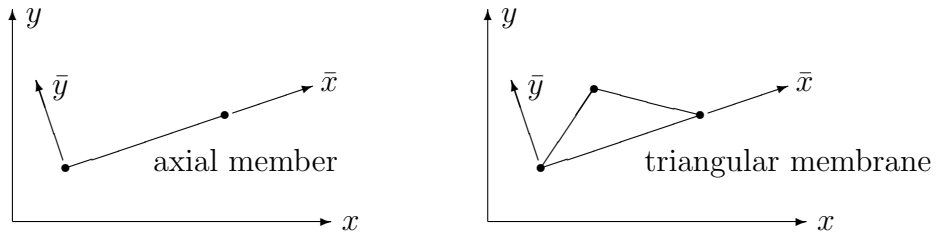

Figure 8: Element Coordinate Systems in the Finite Element Method.

We can also determine the projection matrix for this projection. The projection matrix is the matrix $\mathbf{P}$ that transforms $\mathbf{b}$ into $\mathbf{p}$, i.e., $\mathbf{P}$ is the matrix such that

$$\mathbf{p} = \mathbf{P}\mathbf{b}. \tag{2.97}$$

Here it is convenient to switch to index notation with the summation convention (in which repeated indices are summed). From Eq. 2.95,

$$p_i = \frac{a_j b_j}{\mathbf{a}\cdot\mathbf{a}}\, a_i = \frac{a_i a_j}{\mathbf{a}\cdot\mathbf{a}}\, b_j = P_{ij} b_j. \tag{2.98}$$

Thus, the projection matrix is

$$P_{ij} = \frac{a_i a_j}{\mathbf{a}\cdot\mathbf{a}} \tag{2.99}$$

or, in vector and matrix notations,

$$\mathbf{P} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}\cdot\mathbf{a}} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}. \tag{2.100}$$

The projection matrix $\mathbf{P}$ has the following properties:

1. $\mathbf{P}$ is symmetric, i.e., $P_{ij} = P_{ji}$.

2. $\mathbf{P}^2 = \mathbf{P}$. That is, once a vector has been projected, there is no place else to go with further projection. This property can be proved using index notation:

$$\left(\mathbf{P}^2\right)_{ij} = (\mathbf{P}\mathbf{P})_{ij} = P_{ik}P_{kj} = \frac{a_i a_k a_k a_j}{(\mathbf{a}\cdot\mathbf{a})(\mathbf{a}\cdot\mathbf{a})} = \frac{a_i a_j (\mathbf{a}\cdot\mathbf{a})}{(\mathbf{a}\cdot\mathbf{a})(\mathbf{a}\cdot\mathbf{a})} = \frac{a_i a_j}{\mathbf{a}\cdot\mathbf{a}} = P_{ij}. \tag{2.101}$$

3. $\mathbf{P}$ is singular. That is, given $\mathbf{p}$, it is not possible to reconstruct the original vector $\mathbf{b}$, since $\mathbf{b}$ is not unique.

3 Change of Basis

On many occasions in engineering applications, the need arises to transform vectors and matrices from one coordinate system to another. For example, in the finite element method, it is frequently more convenient to derive element matrices in a local element coordinate system and then transform those matrices to a global system (Fig. 8). Transformations

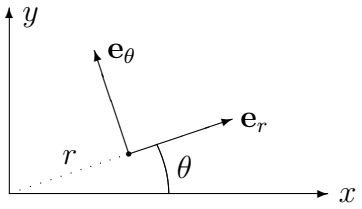

Figure 9: Basis Vectors in Polar Coordinate System.

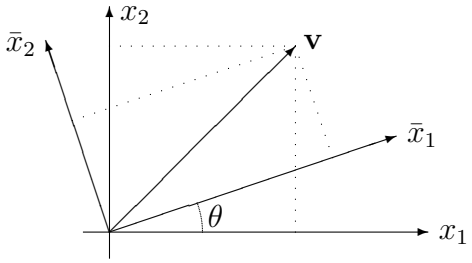

Figure 10: Change of Basis.

are also needed to transform from other orthogonal coordinate systems (e.g., cylindrical or spherical) to Cartesian coordinates (Fig. 9).

Let the vector $\mathbf{v}$ be given by

$$\mathbf{v} = v_1\mathbf{e}_1 + v_2\mathbf{e}_2 + v_3\mathbf{e}_3 = \sum_{i=1}^{3} v_i\mathbf{e}_i, \tag{3.1}$$

where $\mathbf{e}_i$ are the basis vectors, and v_i are the components of $\mathbf{v}$. Using the *summation convention*, we can omit the summation sign and write

$$\mathbf{v} = v_i\mathbf{e}_i, \tag{3.2}$$

where, if a subscript appears exactly twice, a summation is implied over the range.

An *orthonormal basis* is defined as a basis whose basis vectors are mutually orthogonal unit vectors (i.e., vectors of unit length). If $\mathbf{e}_i$ is an orthonormal basis,

$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \tag{3.3}$$

where δ_{ij} is the Kronecker delta.

Since bases are not unique, we can express $\mathbf{v}$ in two different orthonormal bases:

$$\mathbf{v} = \sum_{i=1}^{3} v_i\mathbf{e}_i = \sum_{i=1}^{3} \bar{v}_i\bar{\mathbf{e}}_i, \tag{3.4}$$

where v_i are the components of $\mathbf{v}$ in the unbarred coordinate system, and $\bar{v}_i$ are the components in the barred system (Fig. 10). If we take the dot product of both sides of Eq. 3.4

with $\mathbf{e}_j$, we obtain

$$\sum_{i=1}^{3} v_i \, \mathbf{e}_i \cdot \mathbf{e}_j = \sum_{i=1}^{3} \bar{v}_i \, \bar{\mathbf{e}}_i \cdot \mathbf{e}_j, \tag{3.5}$$

where $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$, and we define the 3×3 matrix $\mathbf{R}$ as

$$R_{ij} = \bar{\mathbf{e}}_i \cdot \mathbf{e}_j. \tag{3.6}$$

Thus, from Eq. 3.5,

$$v_j = \sum_{i=1}^{3} R_{ij} \bar{v}_i = \sum_{i=1}^{3} R_{ji}^T \bar{v}_i. \tag{3.7}$$

Since the matrix product

$$\mathbf{C} = \mathbf{AB} \tag{3.8}$$

can be written using subscript notation as

$$C_{ij} = \sum_{k=1}^{3} A_{ik} B_{kj}, \tag{3.9}$$

Eq. 3.7 is equivalent to the matrix product

$$\mathbf{v} = \mathbf{R}^T \bar{\mathbf{v}}. \tag{3.10}$$

Similarly, if we take the dot product of Eq. 3.4 with $\bar{\mathbf{e}}_j$, we obtain

$$\sum_{i=1}^{3} v_i \, \mathbf{e}_i \cdot \bar{\mathbf{e}}_j = \sum_{i=1}^{3} \bar{v}_i \, \bar{\mathbf{e}}_i \cdot \bar{\mathbf{e}}_j, \tag{3.11}$$

where $\bar{\mathbf{e}}_i \cdot \bar{\mathbf{e}}_j = \delta_{ij}$, and $\mathbf{e}_i \cdot \bar{\mathbf{e}}_j = R_{ji}$. Thus,

$$\bar{v}_j = \sum_{i=1}^{3} R_{ji} v_i \quad \text{or} \quad \bar{\mathbf{v}} = \mathbf{R}\mathbf{v} \quad \text{or} \quad \mathbf{v} = \mathbf{R}^{-1}\bar{\mathbf{v}}. \tag{3.12}$$

A comparison of Eqs. 3.10 and 3.12 yields

$$\mathbf{R}^{-1} = \mathbf{R}^T \quad \text{or} \quad \mathbf{R}\mathbf{R}^T = \mathbf{I} \quad \text{or} \quad \sum_{k=1}^{3} R_{ik} R_{jk} = \delta_{ij}, \tag{3.13}$$

where $\mathbf{I}$ is the identity matrix ($I_{ij} = \delta_{ij}$):

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.14}$$

This type of transformation is called an *orthogonal coordinate transformation* (OCT). A matrix $\mathbf{R}$ satisfying Eq. 3.13 is said to be an *orthogonal* matrix. That is, an orthogonal

matrix is one whose inverse is equal to the transpose. $\mathbf{R}$ is sometimes called a *rotation matrix*.

For example, for the coordinate rotation shown in Fig. 10, in 3-D,

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3.15}$$

In 2-D,

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{3.16}$$

and

$$\begin{cases} v_x = \bar{v}_x \cos\theta - \bar{v}_y \sin\theta \\ v_y = \bar{v}_x \sin\theta + \bar{v}_y \cos\theta. \end{cases} \tag{3.17}$$

We recall that the determinant of a matrix product is equal to the product of the determinants. Also, the determinant of the transpose of a matrix is equal to the determinant of the matrix itself. Thus, from Eq. 3.13,

$$\det(\mathbf{R}\mathbf{R}^T) = (\det\mathbf{R})(\det\mathbf{R}^T) = (\det\mathbf{R})^2 = \det\mathbf{I} = 1, \tag{3.18}$$

and we conclude that, for an orthogonal matrix $\mathbf{R}$,

$$\det\mathbf{R} = \pm 1. \tag{3.19}$$

The plus sign occurs for rotations, and the minus sign occurs for combinations of rotations and reflections. For example, the orthogonal matrix

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \tag{3.20}$$

indicates a reflection in the z direction (i.e., the sign of the z component is changed).

Another property of orthogonal matrices that can be deduced directly from the definition, Eq. 3.13, is that the rows and columns of an orthogonal matrix must be unit vectors and mutually orthogonal. That is, the rows and columns form an orthonormal set.

We note that the length of a vector is unchanged under an orthogonal coordinate transformation, since the square of the length is given by

$$\bar{v}_i \bar{v}_i = R_{ij} v_j R_{ik} v_k = \delta_{jk} v_j v_k = v_j v_j, \tag{3.21}$$

where the summation convention was used. That is, the square of the length of a vector is the same in both coordinate systems.

To summarize, under an orthogonal coordinate transformation, vectors transform according to the rule

$$\bar{\mathbf{v}} = \mathbf{R}\mathbf{v} \quad \text{or} \quad \bar{v}_i = \sum_{j=1}^{3} R_{ij} v_j, \tag{3.22}$$

where

$$R_{ij} = \bar{\mathbf{e}}_i \cdot \mathbf{e}_j, \tag{3.23}$$

and

$$\mathbf{R}\mathbf{R}^T = \mathbf{R}^T\mathbf{R} = \mathbf{I}. \tag{3.24}$$

3.1 Tensors

A vector which transforms under an orthogonal coordinate transformation according to the rule $\bar{\mathbf{v}} = \mathbf{R}\mathbf{v}$ is defined as a tensor of rank 1. A tensor of rank 0 is a scalar (a quantity which is unchanged by an orthogonal coordinate transformation). For example, temperature and pressure are scalars, since $\bar{T} = T$ and $\bar{p} = p$.

We now introduce tensors of rank 2. Consider a matrix $\mathbf{M} = (M_{ij})$ which relates two vectors $\mathbf{u}$ and $\mathbf{v}$ by

$$\mathbf{v} = \mathbf{M}\mathbf{u} \quad \text{or} \quad v_i = \sum_{j=1}^{3} M_{ij} u_j \tag{3.25}$$

(i.e., the result of multiplying a matrix and a vector is a vector). Also, in a rotated coordinate system,

$$\bar{\mathbf{v}} = \bar{\mathbf{M}}\bar{\mathbf{u}}. \tag{3.26}$$

Since both $\mathbf{u}$ and $\mathbf{v}$ are vectors (tensors of rank 1), Eq. 3.25 implies

$$\mathbf{R}^T\bar{\mathbf{v}} = \mathbf{M}\mathbf{R}^T\bar{\mathbf{u}} \quad \text{or} \quad \bar{\mathbf{v}} = \mathbf{R}\mathbf{M}\mathbf{R}^T\bar{\mathbf{u}}. \tag{3.27}$$

By comparing Eqs. 3.26 and 3.27, we obtain

$$\bar{\mathbf{M}} = \mathbf{R}\mathbf{M}\mathbf{R}^T \tag{3.28}$$

or, in index notation,

$$\bar{M}_{ij} = \sum_{k=1}^{3}\sum_{l=1}^{3} R_{ik} R_{jl} M_{kl}, \tag{3.29}$$

which is the transformation rule for a tensor of rank 2. In general, a tensor of rank n, which has n indices, transforms under an orthogonal coordinate transformation according to the rule

$$\bar{A}_{ij\cdots k} = \sum_{p=1}^{3}\sum_{q=1}^{3}\cdots\sum_{r=1}^{3} R_{ip} R_{jq} \cdots R_{kr} A_{pq\cdots r}. \tag{3.30}$$

3.2 Examples of Tensors

1. Stress and strain in elasticity

The stress tensor $\boldsymbol{\sigma}$ is

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}, \tag{3.31}$$

where σ_{11}, σ_{22}, σ_{33} are the direct (normal) stresses, and σ_{12}, σ_{13}, and σ_{23} are the shear stresses. The corresponding strain tensor is

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} \\ \varepsilon_{31} & \varepsilon_{32} & \varepsilon_{33} \end{bmatrix}, \tag{3.32}$$

where, in terms of displacements,

$$\varepsilon_{ij} = \frac{1}{2}(u_{i,j} + u_{j,i}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{3.33}$$

The shear strains in this tensor are equal to half the corresponding engineering shear strains. Both $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$ transform like second rank tensors.

2. Generalized Hooke's law

According to Hooke's law in elasticity, the extension in a bar subjected to an axial force is proportional to the force, or stress is proportional to strain. In 1-D,

$$\sigma = E\varepsilon, \tag{3.34}$$

where E is Young's modulus, an experimentally determined material property.

In general three-dimensional elasticity, there are nine components of stress σ_{ij} and nine components of strain ε_{ij}. According to generalized Hooke's law, each stress component can be written as a linear combination of the nine strain components:

$$\sigma_{ij} = c_{ijkl}\varepsilon_{kl}, \tag{3.35}$$

where the 81 material constants c_{ijkl} are experimentally determined, and the summation convention is being used.

We now prove that c_{ijkl} is a tensor of rank 4. We can write Eq. 3.35 in terms of stress and strain in a second orthogonal coordinate system as

$$R_{ki}R_{lj}\bar{\sigma}_{kl} = c_{ijkl}R_{mk}R_{nl}\bar{\varepsilon}_{mn}. \tag{3.36}$$

If we multiply both sides of this equation by $R_{pj}R_{oi}$, and sum repeated indices, we obtain

$$R_{pj}R_{oi}R_{ki}R_{lj}\bar{\sigma}_{kl} = R_{oi}R_{pj}R_{mk}R_{nl}c_{ijkl}\bar{\varepsilon}_{mn}, \tag{3.37}$$

or, because $\mathbf{R}$ is an orthogonal matrix,

$$\delta_{ok}\delta_{pl}\bar{\sigma}_{kl} = \bar{\sigma}_{op} = R_{oi}R_{pj}R_{mk}R_{nl}c_{ijkl}\bar{\varepsilon}_{mn}. \tag{3.38}$$

Since, in the second coordinate system,

$$\bar{\sigma}_{op} = \bar{c}_{opmn}\bar{\varepsilon}_{mn}, \tag{3.39}$$

we conclude that

$$\bar{c}_{opmn} = R_{oi}R_{pj}R_{mk}R_{nl}c_{ijkl}, \tag{3.40}$$

which proves that c_{ijkl} is a tensor of rank 4.

3. Stiffness matrix in finite element analysis

In the finite element method of analysis for structures, the forces $\mathbf{F}$ acting on an object in static equilibrium are a linear combination of the displacements $\mathbf{u}$ (or *vice versa*):

$$\mathbf{Ku} = \mathbf{F}, \tag{3.41}$$

where $\mathbf{K}$ is referred to as the stiffness matrix (with dimensions of force/displacement). In this equation, $\mathbf{u}$ and $\mathbf{F}$ contain several subvectors, since $\mathbf{u}$ and $\mathbf{F}$ are the displacement and force vectors for all grid points, i.e.,

$$\mathbf{u} = \begin{Bmatrix} \mathbf{u}_a \\ \mathbf{u}_b \\ \mathbf{u}_c \\ \vdots \end{Bmatrix}, \quad \mathbf{F} = \begin{Bmatrix} \mathbf{F}_a \\ \mathbf{F}_b \\ \mathbf{F}_c \\ \vdots \end{Bmatrix} \tag{3.42}$$

for grid points $a, b, c, \ldots$, where

$$\bar{\mathbf{u}}_a = \mathbf{R}_a \mathbf{u}_a, \quad \bar{\mathbf{u}}_b = \mathbf{R}_b \mathbf{u}_b, \quad \cdots . \tag{3.43}$$

Thus,

$$\bar{\mathbf{u}} = \begin{Bmatrix} \bar{\mathbf{u}}_a \\ \bar{\mathbf{u}}_b \\ \bar{\mathbf{u}}_c \\ \vdots \end{Bmatrix} = \begin{bmatrix} \mathbf{R}_a & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{R}_b & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_c & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{Bmatrix} \mathbf{u}_a \\ \mathbf{u}_b \\ \mathbf{u}_c \\ \vdots \end{Bmatrix} = \boldsymbol{\Gamma} \mathbf{u}, \tag{3.44}$$

where $\boldsymbol{\Gamma}$ is an orthogonal block-diagonal matrix consisting of rotation matrices $\mathbf{R}$:

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mathbf{R}_a & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{R}_b & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_c & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.45}$$

and

$$\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \mathbf{I}. \tag{3.46}$$

Similarly,

$$\bar{\mathbf{F}} = \boldsymbol{\Gamma} \mathbf{F}. \tag{3.47}$$

Thus, if

$$\bar{\mathbf{K}} \bar{\mathbf{u}} = \bar{\mathbf{F}}, \tag{3.48}$$

$$\bar{\mathbf{K}} \boldsymbol{\Gamma} \mathbf{u} = \boldsymbol{\Gamma} \mathbf{F} \tag{3.49}$$

or

$$\left(\boldsymbol{\Gamma}^T \bar{\mathbf{K}} \boldsymbol{\Gamma} \right) \mathbf{u} = \mathbf{F}. \tag{3.50}$$

That is, the stiffness matrix transforms like other tensors of rank 2:

$$\mathbf{K} = \boldsymbol{\Gamma}^T \bar{\mathbf{K}} \boldsymbol{\Gamma}. \tag{3.51}$$

We illustrate the transformation of a finite element stiffness matrix by transforming the stiffness matrix for the pin-jointed rod element from a local element coordinate system to a global Cartesian system. Consider the rod shown in Fig. 11. For this element, the 4×4 2-D

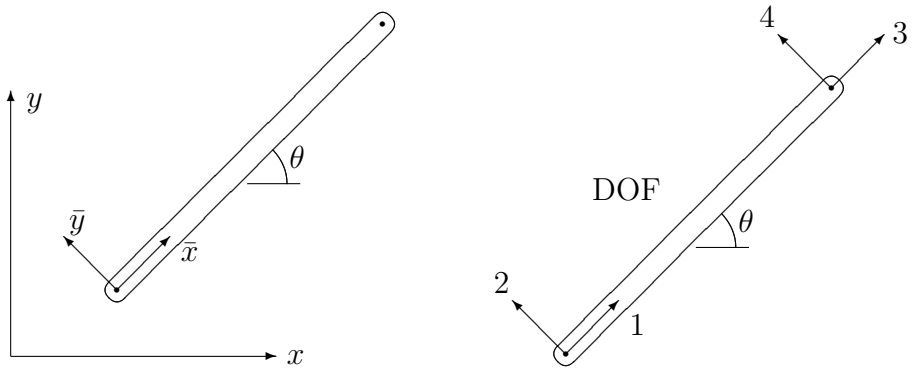

Figure 11: Element Coordinate System for Pin-Jointed Rod.

stiffness matrix in the element coordinate system is

$$\bar{\mathbf{K}} = \frac{AE}{L}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.52}$$

where A is the cross-sectional area of the rod, E is Young's modulus of the rod material, and L is the rod length. In the global coordinate system,

$$\mathbf{K} = \mathbf{\Gamma}^T \bar{\mathbf{K}} \mathbf{\Gamma}, \tag{3.53}$$

where the 4×4 transformation matrix is

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{R} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}, \tag{3.54}$$

and the rotation matrix is

$$\mathbf{R} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{3.55}$$

Thus,

$$\begin{aligned} \mathbf{K} &= \frac{AE}{L}\begin{bmatrix} c & -s & 0 & 0 \\ s & c & 0 & 0 \\ 0 & 0 & c & -s \\ 0 & 0 & s & c \end{bmatrix}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} c & s & 0 & 0 \\ -s & c & 0 & 0 \\ 0 & 0 & c & s \\ 0 & 0 & -s & c \end{bmatrix} \\ &= \frac{AE}{L}\begin{bmatrix} c^2 & cs & -c^2 & -cs \\ cs & s^2 & -cs & -s^2 \\ -c^2 & -cs & c^2 & cs \\ -cs & -s^2 & cs & s^2 \end{bmatrix}, \end{aligned} \tag{3.56}$$

where $c = \cos\theta$ and $s = \sin\theta$.

3.3 Isotropic Tensors

An *isotropic tensor* is a tensor which is independent of coordinate system (i.e., invariant under an orthogonal coordinate transformation). The Kronecker delta δ_{ij} is a second rank tensor and isotropic, since $\bar{\delta}_{ij} = \delta_{ij}$, and

$$\bar{\mathbf{I}} = \mathbf{R}\mathbf{I}\mathbf{R}^T = \mathbf{R}\mathbf{R}^T = \mathbf{I}. \tag{3.57}$$

That is, the identity matrix $\mathbf{I}$ is invariant under an orthogonal coordinate transformation.

It can be shown in tensor analysis that δ_{ij} is the only isotropic tensor of rank 2 and, moreover, δ_{ij} is the characteristic tensor for all isotropic tensors:

Rank	Isotropic Tensors
1	none
2	$c\delta_{ij}$
3	none
4	$a\delta_{ij}\delta_{kl} + b\delta_{ik}\delta_{jl} + c\delta_{il}\delta_{jk}$
odd	none

That is, all isotropic tensors of rank 4 must be of the form shown above, which has three constants. For example, in generalized Hooke's law (Eq. 3.35), the material property tensor c_{ijkl} has $3^4 = 81$ constants (assuming no symmetry). For an isotropic material, c_{ijkl} must be an isotropic tensor of rank 4, thus implying at most three independent elastic material constants (on the basis of tensor analysis alone). The actual number of independent material constants for an isotropic material turns out to be two rather than three, a result which depends on the existence of a strain energy function, which implies the additional symmetry $c_{ijkl} = c_{klij}$.

4 Least Squares Problems

Consider the rectangular system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{4.1}$$

where $\mathbf{A}$ is an $m \times n$ matrix with $m > n$. That is, we are interested in the case where there are more equations than unknowns. Assume that the system is inconsistent, so that Eq. 4.1 has no solution. Our interest here is to find a vector $\mathbf{x}$ which "best" fits the data in some sense.

We state the problem as follows: If $\mathbf{A}$ is a $m \times n$ matrix, with $m > n$, and $\mathbf{A}\mathbf{x} = \mathbf{b}$, find $\mathbf{x}$ such that the scalar error measure $E = |\mathbf{A}\mathbf{x} - \mathbf{b}|$ is minimized, where E is the length of the residual $\mathbf{A}\mathbf{x} - \mathbf{b}$ (error).

To solve this problem, we consider the square of the error:

$$\begin{aligned} E^2 &= (\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= (\mathbf{x}^T\mathbf{A}^T - \mathbf{b}^T)(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} - \mathbf{b}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{b}, \end{aligned} \tag{4.2}$$

or, in index notation using the summation convention,

$$\begin{aligned} E^2 &= x_i(A^T)_{ij}A_{jk}x_k - x_i(A^T)_{ij}b_j - b_iA_{ij}x_j + b_ib_i \\ &= x_iA_{ji}A_{jk}x_k - x_iA_{ji}b_j - b_iA_{ij}x_j + b_ib_i. \end{aligned} \tag{4.3}$$

We wish to find the vector $\mathbf{x}$ which minimizes E^2. Thus, we set

$$\frac{\partial(E^2)}{\partial x_l} = 0 \quad \text{for each } l, \tag{4.4}$$

where we note that

$$\frac{\partial x_i}{\partial x_l} = \delta_{il}, \tag{4.5}$$

where δ_{ij} is the Kronecker delta defined as

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{4.6}$$

We then differentiate Eq. 4.3 to obtain

$$0 = \frac{\partial(E^2)}{\partial x_l} = \delta_{il}A_{ji}A_{jk}x_k + x_iA_{ji}A_{jk}\delta_{kl} - \delta_{il}A_{ji}b_j - b_iA_{ij}\delta_{jl} + 0 \tag{4.7}$$

$$= A_{jl}A_{jk}x_k + x_iA_{ji}A_{jl} - A_{jl}b_j - b_iA_{il}, \tag{4.8}$$

where the dummy subscripts k in the first term and j in the third term can both be changed to i. Thus,

$$0 = \frac{\partial(E^2)}{\partial x_l} = A_{jl}A_{ji}x_i + A_{ji}A_{jl}x_i - A_{il}b_i - A_{il}b_i = (2\mathbf{A}^T\mathbf{A}\mathbf{x} - 2\mathbf{A}^T\mathbf{b})_l \tag{4.9}$$

or

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}. \tag{4.10}$$

The problem solved is referred to as the *least squares problem*, since we have minimized E, the square root of the sum of the squares of the components of the vector $\mathbf{A}\mathbf{x} - \mathbf{b}$. The equations, Eq. 4.10, which solve the least squares problem are referred to as the *normal equations*.

Since $\mathbf{A}$ is an $m \times n$ matrix, the coefficient matrix $\mathbf{A}^T\mathbf{A}$ in Eq. 4.10 is a square $n \times n$ matrix. Moreover, $\mathbf{A}^T\mathbf{A}$ is also symmetric, since

$$(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T(\mathbf{A}^T)^T = \mathbf{A}^T\mathbf{A}. \tag{4.11}$$

If $\mathbf{A}^T\mathbf{A}$ is invertible, we could "solve" the normal equations to obtain

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}. \tag{4.12}$$

However, $(\mathbf{A}^T\mathbf{A})^{-1}$ may not exist, since the original system, Eq. 4.1, may not have been over-determined (i.e., rank $r < n$).

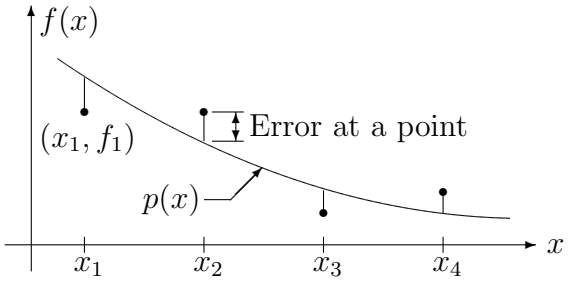

Figure 12: Example of Least Squares Fit of Data.

We note that Eq. 4.10 could have been obtained simply by multiplying both sides of Eq. 4.1 by $\mathbf{A}^T$, but we would not have known that the result is a least-squares solution. We also recall that the matrix $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ in Eq. 4.12 is the *pseudo left inverse* of $\mathbf{A}$. Thus, the pseudo left inverse is equivalent to a least squares solution.

We also note that, if $\mathbf{A}$ is square and invertible, then

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} = \mathbf{A}^{-1}(\mathbf{A}^T)^{-1}\mathbf{A}^T\mathbf{b} = \mathbf{A}^{-1}\mathbf{b}, \tag{4.13}$$

as expected.

One of the most important properties associated with the normal equations is that, for any matrix $\mathbf{A}$,

$$\text{rank}(\mathbf{A}^T\mathbf{A}) = \text{rank}(\mathbf{A}), \tag{4.14}$$

as proved in §2.5 following Eq. 2.42. We recall that the implication of this result is that, if $\text{rank}(\mathbf{A}) = n$ (i.e., $\mathbf{A}$'s columns are independent), then $\mathbf{A}^T\mathbf{A}$ is invertible.

4.1 Least Squares Fitting of Data

Sometimes one has a large number of discrete values of some function and wants to determine a fairly simple approximation to the data, but make use of all the data. For example, as shown in Fig. 12, suppose we know the points $(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)$, and we want to approximate f with the low-order polynomial

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n, \tag{4.15}$$

where the number m of points is large, and the order n of the polynomial is small.

The goal of the fit is for p to agree as closely as possible with the data. For example, we could require that the algebraic sum of the errors would be minimized by p, but positive and negative errors would cancel each other, thus distorting the evaluation of the fit. We could alternatively require that the sum of the *absolute* errors be minimum, but it turns out that that choice results in a nasty mathematical problem. Hence we pick $p(x)$ such that the sum of the *squares* of the errors,

$$R = [f_1 - p(x_1)]^2 + [f_2 - p(x_2)]^2 + \cdots + [f_m - p(x_m)]^2, \tag{4.16}$$

is minimum. We note that the quantities in brackets are the errors at each of the discrete points where the function is known. R is referred to as the *residual*.

For example, consider the linear approximation

$$p(x) = a_0 + a_1 x, \tag{4.17}$$

where the measure of error is given by

$$R(a_0, a_1) = [f_1 - (a_0 + a_1 x_1)]^2 + [f_2 - (a_0 + a_1 x_2)]^2 + \cdots + [f_m - (a_0 + a_1 x_m)]^2. \tag{4.18}$$

For R a minimum,

$$\frac{\partial R}{\partial a_0} = 0 \quad \text{and} \quad \frac{\partial R}{\partial a_1} = 0, \tag{4.19}$$

which implies, from Eq. 4.18,

$$\begin{aligned} 2[f_1 - (a_0 + a_1 x_1)](-1) + 2[f_2 - (a_0 + a_1 x_2)](-1) + \cdots + 2[f_m - (a_0 + a_1 x_m)](-1) = 0 \\ 2[f_1 - (a_0 + a_1 x_1)](-x_1) + 2[f_2 - (a_0 + a_1 x_2)](-x_2) + \cdots + 2[f_m - (a_0 + a_1 x_m)](-x_m) = 0 \end{aligned} \tag{4.20}$$

or

$$\begin{array}{rcrcl} m a_0 & + & (x_1 + x_2 + \cdots + x_m) a_1 & = & f_1 + f_2 + \cdots + f_m \\ (x_1 + x_2 + \cdots + x_m) a_0 & + & (x_1^2 + x_2^2 + \cdots + x_m^2) a_1 & = & f_1 x_1 + f_2 x_2 + \cdots + f_m x_m. \end{array} \tag{4.21}$$

This is a symmetric system of two linear algebraic equations in two unknowns (a_0, a_1) which can be solved for a_0 and a_1 to yield $p(x)$.

An alternative approach would be to attempt to find (a_0, a_1) such that the linear function goes through all m points:

$$\begin{aligned} a_0 + a_1 x_1 &= f_1 \\ a_0 + a_1 x_2 &= f_2 \\ a_0 + a_1 x_3 &= f_3 \\ &\vdots \\ a_0 + a_1 x_m &= f_m, \end{aligned} \tag{4.22}$$

which is an overdetermined system if $m > 2$. In matrix notation, this system can be written

$$\mathbf{A}\mathbf{a} = \mathbf{f}, \tag{4.23}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix}, \quad \mathbf{f} = \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_m \end{Bmatrix}, \quad \mathbf{a} = \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix}. \tag{4.24}$$

This system was solved previously by multiplying by A^T:

$$\mathbf{A}^T \mathbf{A}\mathbf{a} = \mathbf{A}^T \mathbf{f}, \tag{4.25}$$

where

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_m \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} = \begin{bmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{bmatrix}, \tag{4.26}$$

and

$$\mathbf{A}^T\mathbf{f} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_m \end{bmatrix} \begin{Bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_m \end{Bmatrix} = \begin{Bmatrix} \sum_{i=1}^{m} f_i \\ \sum_{i=1}^{m} x_i f_i \end{Bmatrix}. \tag{4.27}$$

Thus, Eq. 4.25 becomes

$$\begin{bmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{Bmatrix} \sum_{i=1}^{m} f_i \\ \sum_{i=1}^{m} x_i f_i \end{Bmatrix}, \tag{4.28}$$

which is identical to Eq. 4.21. That is, both approaches give the same result. The *geometric* interpretation of minimizing the sum of squares of errors is equivalent to the *algebraic* interpretation of finding $\mathbf{x}$ such that $|\mathbf{Aa} - \mathbf{f}|$ is minimized. Both approaches yield Eq. 4.28. Hence, the two error measures are also the same; i.e., $E^2 = R$.

4.2 The General Linear Least Squares Problem

Given m points $(x_1, f_1), (x_2, f_2), \ldots, (x_m, f_m)$, we could, in general, perform a least squares fit using a linear combination of other functions $\phi_1(x), \phi_2(x), \ldots, \phi_n(x)$. For example, we could attempt a fit using the set of functions $\{e^x, 1/x, \sin x\}$. The functions used in the linear combination are referred to as *basis functions* (or *basis vectors*). Thus, the more general least squares problem is to find the coefficients a_i in

$$p(x) = a_1\phi_1(x) + a_2\phi_2(x) + \cdots + a_n\phi_n(x) \tag{4.29}$$

so that the mean square error, Eq. 4.16, between the function $p(x)$ and the original points is minimized.

If we use the algebraic approach to set up the equations, we attempt to force $p(x)$ to go through all m points:

$$\begin{aligned} a_1\phi_1(x_1) + a_2\phi_2(x_1) + \cdots + a_n\phi_n(x_1) &= f_1 \\ a_1\phi_1(x_2) + a_2\phi_2(x_2) + \cdots + a_n\phi_n(x_2) &= f_2 \\ &\vdots \\ a_1\phi_1(x_m) + a_2\phi_2(x_m) + \cdots + a_n\phi_n(x_m) &= f_m, \end{aligned} \tag{4.30}$$

which is an over-determined system if $m > n$. Thus,

$$\mathbf{A}\mathbf{a} = \mathbf{f} \quad \text{or} \quad A_{ij}a_j = f_i, \tag{4.31}$$

where the summation convention is used, and the coefficient matrix is

$$A_{ij} = \phi_j(x_i). \tag{4.32}$$

The least squares solution of this problem is therefore found by multiplying by $\mathbf{A}^T$:

$$(\mathbf{A}^T\mathbf{A})\mathbf{a} = \mathbf{A}^T\mathbf{f}, \tag{4.33}$$

where (using the summation convention)

$$(\mathbf{A}^T\mathbf{A})_{ij} = (A^T)_{ik}A_{kj} = A_{ki}A_{kj} = \phi_i(x_k)\phi_j(x_k) \tag{4.34}$$

and

$$(\mathbf{A}^T\mathbf{f})_i = (A^T)_{ij}f_j = A_{ji}f_j = \phi_i(x_j)f_j. \tag{4.35}$$

Thus, the normal equations for the general linear least squares problem are (in index notation)

$$[\phi_i(x_k)\phi_j(x_k)]a_j = \phi_i(x_j)f_j. \tag{4.36}$$

The order of this system is n, the number of basis functions. The straight line fit considered earlier is a special case of this formula.

4.3 Gram-Schmidt Orthogonalization

One of the potential problems with the least squares fits described above is that the set of basis functions is not an orthogonal set (even though the functions may be independent), and the resulting system of equations (the normal equations) may be hard to solve (i.e, the system may be poorly conditioned). This situation can be improved by making the set of basis functions an orthogonal set.

The Gram-Schmidt procedure is a method for finding an orthonormal basis for a set of independent vectors. Note that, given a set of vectors, we cannot simply use as our orthonormal basis the set $\{\mathbf{e}^{(1)}, \mathbf{e}^{(2)}, \mathbf{e}^{(3)}, \ldots\}$, since the given vectors may define a subspace of a larger dimensional space (e.g., a skewed plane in $\mathbb{R}^3$).

Consider the following problem: Given a set of independent vectors $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \mathbf{a}^{(3)}, \ldots$, find a set of orthonormal vectors $\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \mathbf{q}^{(3)}, \ldots$ that span the same space. (A set of vectors is *orthonormal* if the vectors in the set are both mutually orthogonal and normalized to unit length.)

We start the Gram-Schmidt procedure by choosing the first unit basis vector to be parallel to the first vector in the set:

$$\mathbf{q}^{(1)} = \frac{\mathbf{a}^{(1)}}{|\mathbf{a}^{(1)}|}, \tag{4.37}$$

as shown in Fig. 13. We define the second unit basis vector so that it is coplanar with $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$:

$$\mathbf{q}^{(2)} = \frac{\mathbf{a}^{(2)} - (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)}}{|\mathbf{a}^{(2)} - (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)}|}. \tag{4.38}$$

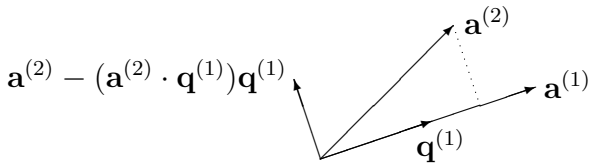

Figure 13: The First Two Vectors in Gram-Schmidt.

Since the third unit basis vector must be orthogonal to the first two, we create $\mathbf{q}^{(3)}$ by subtracting from $\mathbf{a}^{(3)}$ any part of $\mathbf{a}^{(3)}$ which is parallel to either $\mathbf{q}^{(1)}$ or $\mathbf{q}^{(2)}$. Hence,

$$\mathbf{q}^{(3)} = \frac{\mathbf{a}^{(3)} - (\mathbf{a}^{(3)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} - (\mathbf{a}^{(3)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)}}{|\mathbf{a}^{(3)} - (\mathbf{a}^{(3)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} - (\mathbf{a}^{(3)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)}|}, \tag{4.39}$$

$$\mathbf{q}^{(4)} = \frac{\mathbf{a}^{(4)} - (\mathbf{a}^{(4)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} - (\mathbf{a}^{(4)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)} - (\mathbf{a}^{(4)} \cdot \mathbf{q}^{(3)})\mathbf{q}^{(3)}}{|\mathbf{a}^{(4)} - (\mathbf{a}^{(4)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} - (\mathbf{a}^{(4)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)} - (\mathbf{a}^{(4)} \cdot \mathbf{q}^{(3)})\mathbf{q}^{(3)}|}, \tag{4.40}$$

and so on.

4.4 QR Factorization

Consider an $m \times n$ matrix $\mathbf{A}$ whose columns $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \mathbf{a}^{(3)}, \ldots, \mathbf{a}^{(n)}$ are independent. If we have an orthonormal basis $\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \mathbf{q}^{(3)}, \ldots, \mathbf{q}^{(n)}$ for the columns, then each column can be expanded in terms of the orthonormal basis:

$$\left.\begin{aligned} \mathbf{a}^{(1)} &= (\mathbf{a}^{(1)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} + (\mathbf{a}^{(1)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)} + \cdots + (\mathbf{a}^{(1)} \cdot \mathbf{q}^{(n)})\mathbf{q}^{(n)} \\ \mathbf{a}^{(2)} &= (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} + (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)} + \cdots + (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(n)})\mathbf{q}^{(n)} \\ &\vdots \\ \mathbf{a}^{(n)} &= (\mathbf{a}^{(n)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)} + (\mathbf{a}^{(n)} \cdot \mathbf{q}^{(2)})\mathbf{q}^{(2)} + \cdots + (\mathbf{a}^{(n)} \cdot \mathbf{q}^{(n)})\mathbf{q}^{(n)} \end{aligned}\right\}, \tag{4.41}$$

or, in terms of matrices,

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}^{(1)} & \mathbf{a}^{(2)} & \cdots & \mathbf{a}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{q}^{(1)} & \mathbf{q}^{(2)} & \cdots & \mathbf{q}^{(n)} \end{bmatrix} \begin{bmatrix} \mathbf{a}^{(1)} \cdot \mathbf{q}^{(1)} & \mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)} & \cdots \\ \mathbf{a}^{(1)} \cdot \mathbf{q}^{(2)} & \mathbf{a}^{(2)} \cdot \mathbf{q}^{(2)} & \cdots \\ \vdots & \vdots & \vdots \\ \mathbf{a}^{(1)} \cdot \mathbf{q}^{(n)} & \mathbf{a}^{(2)} \cdot \mathbf{q}^{(n)} & \cdots \end{bmatrix}. \tag{4.42}$$

However, if the $\mathbf{q}$ vectors were obtained from a Gram-Schmidt process, $\mathbf{q}^{(1)}$ is parallel to $\mathbf{a}^{(1)}$, and

$$\mathbf{a}^{(1)} \cdot \mathbf{q}^{(2)} = \mathbf{a}^{(1)} \cdot \mathbf{q}^{(3)} = \mathbf{a}^{(1)} \cdot \mathbf{q}^{(4)} = \cdots = 0. \tag{4.43}$$

Also, $\mathbf{q}^{(2)}$ is in the plane of $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$, in which case

$$\mathbf{q}^{(3)} \cdot \mathbf{a}^{(2)} = \mathbf{q}^{(4)} \cdot \mathbf{a}^{(2)} = \mathbf{q}^{(5)} \cdot \mathbf{a}^{(2)} = \cdots = 0. \tag{4.44}$$

Hence, Eq. 4.42 becomes

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}^{(1)} & \mathbf{a}^{(2)} & \cdots & \mathbf{a}^{(n)} \end{bmatrix} = \begin{bmatrix} \mathbf{q}^{(1)} & \mathbf{q}^{(2)} & \cdots & \mathbf{q}^{(n)} \end{bmatrix} \begin{bmatrix} \mathbf{a}^{(1)} \cdot \mathbf{q}^{(1)} & \mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)} & \mathbf{a}^{(3)} \cdot \mathbf{q}^{(1)} & \cdots \\ 0 & \mathbf{a}^{(2)} \cdot \mathbf{q}^{(2)} & \mathbf{a}^{(3)} \cdot \mathbf{q}^{(2)} & \cdots \\ 0 & 0 & \mathbf{a}^{(3)} \cdot \mathbf{q}^{(3)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4.45}$$

or

$$\mathbf{A} = \mathbf{QR}, \tag{4.46}$$

where $\mathbf{Q}$ is the matrix formed with the orthonormal columns $\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \mathbf{q}^{(3)}, \ldots, \mathbf{q}^{(n)}$ of the Gram-Schmidt process, and $\mathbf{R}$ is an upper triangular matrix. We thus conclude that every $m \times n$ matrix with independent columns can be factored into $\mathbf{A} = \mathbf{QR}$.

The benefit of the QR factorization is that it simplifies the least squares solution. We recall that the rectangular system

$$\mathbf{Ax} = \mathbf{b} \tag{4.47}$$

with $m > n$ was solved using least squares by solving the normal equations

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}. \tag{4.48}$$

If $\mathbf{A}$ is factored into

$$\mathbf{A} = \mathbf{QR} \tag{4.49}$$

(where $\mathbf{A}$ and $\mathbf{Q}$ are $m \times n$ matrices, and $\mathbf{R}$ is an $n \times n$ matrix), Eq. 4.48 becomes

$$\mathbf{R}^T\mathbf{Q}^T\mathbf{QRx} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}, \tag{4.50}$$

where, in block form,

$$\mathbf{Q}^T\mathbf{Q} = \begin{Bmatrix} \mathbf{q}^{(1)T} \\ \mathbf{q}^{(2)T} \\ \vdots \\ \mathbf{q}^{(n)T} \end{Bmatrix} \begin{bmatrix} \mathbf{q}^{(1)} & \mathbf{q}^{(2)} & \cdots & \mathbf{q}^{(n)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}. \tag{4.51}$$

Thus,

$$\mathbf{R}^T\mathbf{Rx} = \mathbf{R}^T\mathbf{Q}^T\mathbf{b}, \tag{4.52}$$

and, since $\mathbf{R}$ is upper triangular and invertible, the normal equations become

$$\mathbf{Rx} = \mathbf{Q}^T\mathbf{b}. \tag{4.53}$$

This system can be solved very quickly, since $\mathbf{R}$ is an upper triangular matrix. However, the major computational expense of the least squares problem is factoring $\mathbf{A}$ into $\mathbf{QR}$, i.e., performing the Gram-Schmidt process to find $\mathbf{Q}$ and $\mathbf{R}$.

5 Fourier Series

Consider two functions $f(x)$ and $g(x)$ defined in the interval $[a, b]$ (i.e., $a \leq x \leq b$). We define the *inner product* (or *scalar product*) of f and g as

$$(f, g) = \int_a^b f(x)g(x)\,dx. \tag{5.1}$$

Note the analogy between this definition for functions and the dot product for vectors.

We define the *norm* N_f of a function f as

$$N_f = (f,f)^{\frac{1}{2}} = \sqrt{\int_a^b [f(x)]^2\,dx}. \tag{5.2}$$

Note that a new function $\bar{f}(x)$ defined as

$$\bar{f}(x) = \frac{f(x)}{N_f} \tag{5.3}$$

has unit norm, i.e., $(\bar{f},\bar{f}) = 1$. The function $\bar{f}$ is said to be *normalized* over $[a,b]$. Thus, this definition of the norm of a function is analogous to the length of a vector.

We define two functions f and g as *orthogonal* over $[a,b]$ if $(f,g) = 0$. Orthogonality of functions is analogous to the orthogonality of vectors. We define the set of functions $\phi_i(x)$, $i = 1,2,\ldots,$ as an *orthonormal* set over $[a,b]$ if $(\phi_i,\phi_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta defined as

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{5.4}$$

For example, the set of functions

$$\{1,\ \cos x,\ \sin x,\ \cos 2x,\ \sin 2x,\ \cos 3x,\ \sin 3x,\ \ldots\}$$

is an orthogonal set over $[-\pi,\pi]$, since, for integers m and n,

$$\begin{gathered} (\cos nx, \sin mx) = 0, \\ (\sin nx, \sin mx) = (\cos nx, \cos mx) = 0, \quad m \neq n. \end{gathered} \tag{5.5}$$

To compute the norms of these functions, we note that

$$\begin{gathered} \int_{-\pi}^{\pi} (1)^2\,dx = 2\pi, \\ \int_{-\pi}^{\pi} \sin^2 nx\,dx = \int_{-\pi}^{\pi} \cos^2 nx\,dx = \pi \end{gathered} \tag{5.6}$$

for all integers n. Thus, the set of functions

$$\left\{\frac{1}{\sqrt{2\pi}},\ \frac{\cos x}{\sqrt{\pi}},\ \frac{\sin x}{\sqrt{\pi}},\ \frac{\cos 2x}{\sqrt{\pi}},\ \frac{\sin 2x}{\sqrt{\pi}},\ \ldots\right\}$$

is an orthonormal set over $[-\pi,\pi]$.

The Fourier series of a function defined over $[-L,L]$ is an expansion of the function into sines and cosines:

$$f(x) = A_0 + \sum_{n=1}^{\infty} A_n \cos\frac{n\pi x}{L} + \sum_{n=1}^{\infty} B_n \sin\frac{n\pi x}{L}. \tag{5.7}$$

This equation is an expansion in functions orthogonal over $[-L,L]$, analogous to the expansion of a vector in orthogonal basis vectors. The *basis functions* here are

$$\left\{1,\ \cos\frac{\pi x}{L},\ \sin\frac{\pi x}{L},\ \cos\frac{2\pi x}{L},\ \sin\frac{2\pi x}{L},\ \cos\frac{3\pi x}{L},\ \sin\frac{3\pi x}{L},\ \ldots\right\}.$$

To determine the coefficients A_n, B_n of the Fourier series, we take the inner product of both sides of Eq. 5.7 with each basis function; that is, we multiply both sides of that equation by a basis function, and integrate. For example,

$$\int_{-L}^{L} f(x)\cos\frac{m\pi x}{L}\,dx = A_0\int_{-L}^{L}\cos\frac{m\pi x}{L}\,dx + \sum_{n=1}^{\infty}A_n\int_{-L}^{L}\cos\frac{n\pi x}{L}\cos\frac{m\pi x}{L}\,dx + \sum_{n=1}^{\infty}B_n\int_{-L}^{L}\sin\frac{n\pi x}{L}\cos\frac{m\pi x}{L}\,dx = \begin{cases} A_0(2L), & m=0,\\ A_m(L), & m\neq 0, \end{cases} \tag{5.8}$$

since the integrals in Eq. 5.8 are

$$\int_{-L}^{L}\cos\frac{m\pi x}{L}\,dx = \begin{cases} 2L, & m=0,\\ 0, & m\neq 0, \end{cases} \tag{5.9}$$

$$\int_{-L}^{L}\cos\frac{n\pi x}{L}\cos\frac{m\pi x}{L}\,dx = \int_{-L}^{L}\sin\frac{n\pi x}{L}\sin\frac{m\pi x}{L}\,dx = \begin{cases} L, & m=n,\\ 0, & m\neq n, \end{cases} \tag{5.10}$$

$$\int_{-L}^{L}\sin\frac{n\pi x}{L}\cos\frac{m\pi x}{L}\,dx = 0 \tag{5.11}$$

for integers m and n. Similarly, we can evaluate B_n. Thus, the coefficients A_n, B_n in the Fourier series, Eq. 5.7, are

$$A_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\,dx, \tag{5.12}$$

$$A_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L}\,dx, \quad n>0, \tag{5.13}$$

$$B_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}\,dx. \tag{5.14}$$

Note that A_0 is the average value of $f(x)$ over the domain $[-L, L]$.

The evaluation of the coefficients using integrals could have alternatively been expressed in terms of inner products, in which case, from Eq. 5.7 (with $L=\pi$),

$$(f, \cos mx) = A_m(\cos mx, \cos mx) \tag{5.15}$$

or

$$A_n = \frac{(f, \cos nx)}{(\cos nx, \cos nx)}. \tag{5.16}$$

It is interesting to note the similarity between this last formula and the formula of the vector projection $\mathbf{p}$ of a vector $\mathbf{b}$ onto another vector $\mathbf{a}$ (Fig. 14):

$$\mathbf{p} = \left(\mathbf{b}\cdot\frac{\mathbf{a}}{|\mathbf{a}|}\right)\frac{\mathbf{a}}{|\mathbf{a}|} = \frac{\mathbf{b}\cdot\mathbf{a}}{\mathbf{a}\cdot\mathbf{a}}\,\mathbf{a}. \tag{5.17}$$

Thus, the nth Fourier coefficient can be interpreted as being the "projection" of the function f on the nth basis function $(\cos nx)$.

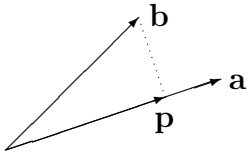

Figure 14: The Projection of One Vector onto Another.

5.1 Example

We evaluate the Fourier series for the function

$$f(x) = \begin{cases} 0, & -5 < x < 0, \\ 3, & 0 < x < 5. \end{cases} \tag{5.18}$$

From Eqs. 5.12–5.14 (with $L = 5$), the Fourier coefficients are

$$A_0 = \frac{1}{10}\int_{-5}^{5} f(x)\,dx = \frac{1}{10}\int_{0}^{5} 3\,dx = 3/2, \tag{5.19}$$

$$A_n = \frac{1}{5}\int_{-5}^{5} f(x)\cos\frac{n\pi x}{5}\,dx = \frac{1}{5}\int_{0}^{5} 3\cos\frac{n\pi x}{5}\,dx = 0, \quad n > 0, \tag{5.20}$$

$$B_n = \frac{1}{5}\int_{-5}^{5} f(x)\sin\frac{n\pi x}{5}\,dx = \frac{1}{5}\int_{0}^{5} 3\sin\frac{n\pi x}{5}\,dx = \begin{cases} 6/(n\pi), & n \text{ odd}, \\ 0, & n \text{ even}. \end{cases} \tag{5.21}$$

Thus the corresponding Fourier series is

$$f(x) = \frac{3}{2} + \frac{6}{\pi}\sum_{n\ \text{odd}} \frac{1}{n}\sin\frac{n\pi x}{5} = \frac{3}{2} + \frac{6}{\pi}\left(\sin\frac{\pi x}{5} + \frac{1}{3}\sin\frac{3\pi x}{5} + \frac{1}{5}\sin\frac{5\pi x}{5} + \cdots\right). \tag{5.22}$$

The convergence of this series can be seen by evaluating the partial sums

$$f(x) \approx f_N(x) = A_0 + \sum_{n=1}^{N} A_n \cos\frac{n\pi x}{L} + \sum_{n=1}^{N} B_n \sin\frac{n\pi x}{L} \tag{5.23}$$

for different values of the upper limit N. Four such representations of $f(x)$ are shown in Fig. 15 for $N = 5$, 10, 20, and 40. Since half the sine terms and all the cosine terms (except the constant term) are zero, the four partial sums have 4, 6, 11, and 21 nonzero terms, respectively.

Notice also in Fig. 15 that, as the number of terms included in the Fourier series increases, the overshoot which occurs at a discontinuity does not diminish. This property of Fourier series is referred to as the Gibbs phenomenom. Notice also that, at a discontinuity of the function, the series converges to the average value of the function at the discontinuity.

5.2 Generalized Fourier Series

The expansion of a function in a series of sines and cosines is a special case of an expansion in other orthogonal functions. In the generalized Fourier series representation of the function

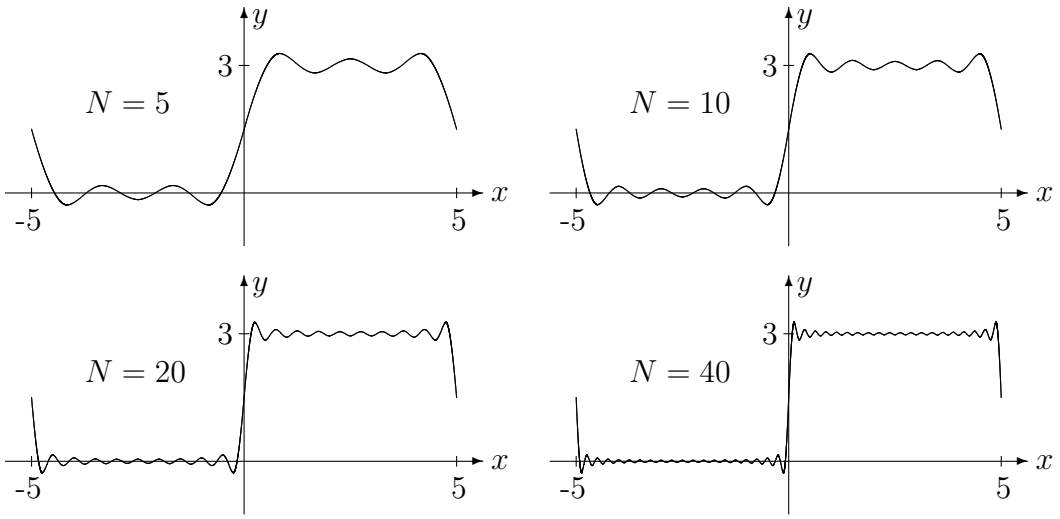

Figure 15: Convergence of Series in Example.

$f(x)$,

$$f(x) \approx \sum_{i=1}^{N} c_i \phi_i(x), \tag{5.24}$$

we wish to evaluate the coefficients c_i so as to minimize the mean square error

$$M_N = \int_a^b \left[f(x) - \sum_{i=1}^{N} c_i \phi_i(x) \right]^2 dx, \tag{5.25}$$

where $(\phi_i, \phi_j) = \delta_{ij}$ (Kronecker delta). Thus, the series representation in Eq. 5.24 is an expansion in orthonormal functions.

To minimize M_N, we set, for each j,

$$0 = \frac{\partial M_N}{\partial c_j} = \int_a^b 2 \left[f(x) - \sum_{i=1}^{N} c_i \phi_i(x) \right] [-\phi_j(x)] \, dx, \tag{5.26}$$

or

$$\int_a^b f(x)\phi_j(x)\, dx = \sum_{i=1}^{N} c_i \int_a^b \phi_i(x)\phi_j(x)\, dx = \sum_{i=1}^{N} c_i \delta_{ij} = c_j. \tag{5.27}$$

That is,

$$c_i = (f, \phi_i). \tag{5.28}$$

Alternatively, had we started with the series, Eq. 5.24, we would have obtained the same result by taking the inner product of each side with ϕ_j:

$$(f, \phi_j) = \sum_{i=1}^{N} c_i (\phi_i, \phi_j) = \sum_{i=1}^{N} c_i \delta_{ij} = c_j. \tag{5.29}$$

The expansion coefficients c_i are called the *Fourier coefficients*, and the series is called the *generalized Fourier series*. The generalized Fourier series thus has the property that the

series representation of a function minimizes the mean square error between the series and the function.

The formula for the generalized Fourier coefficient c_i in Eq. 5.28 can be interpreted as the projection of the function $f(x)$ on the ith basis function ϕ_i. This interpretation is analogous to the expansion of a vector in terms of the unit basis vectors in the coordinate directions. For example, in three-dimensions, the vector $\mathbf{v}$ is, in Cartesian coordinates,

$$\mathbf{v} = v_x \mathbf{e}_x + v_y \mathbf{e}_y + v_z \mathbf{e}_z, \tag{5.30}$$

where the components v_x, v_y, and v_z are the projections of $\mathbf{v}$ on the coordinate basis vectors $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$. That is,

$$v_x = \mathbf{v} \cdot \mathbf{e}_x, \;\; v_y = \mathbf{v} \cdot \mathbf{e}_y, \;\; v_z = \mathbf{v} \cdot \mathbf{e}_z. \tag{5.31}$$

These components are analogous to the Fourier coefficients c_i.

From Eq. 5.25, we can deduce some additional properties of the Fourier series. From that equation, the mean square error after summing N terms of the series is

$$\begin{aligned} M_N &= (f,f) - 2\sum_{i=1}^{N} c_i (f,\phi_i) + \int_a^b \left(\sum_{i=1}^{N} c_i \phi_i \right) \left(\sum_{j=1}^{N} c_j \phi_j \right) dx \\ &= (f,f) - 2\sum_{i=1}^{N} c_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j (\phi_i, \phi_j) \\ &= (f,f) - 2\sum_{i=1}^{N} c_i^2 + \sum_{i=1}^{N} c_i^2 \\ &= (f,f) - \sum_{i=1}^{N} c_i^2. \end{aligned} \tag{5.32}$$

Since, by definition, $M_N \geq 0$, this last equation implies

$$\sum_{i=1}^{N} c_i^2 \leq (f,f), \tag{5.33}$$

which is referred to as Bessel's Inequality. Since the right-hand side of this inequality (the square of the norm) is fixed for a given function, and each term of the left-hand side is non-negative, a necessary condition for a Fourier series to converge is that $c_i \to 0$ as $i \to \infty$.

From Eq. 5.32, the error after summing N terms of the series is

$$M_N = (f,f) - \sum_{i=1}^{N} c_i^2, \tag{5.34}$$

so that the error after $N+1$ terms is

$$M_{N+1} = (f,f) - \sum_{i=1}^{N+1} c_i^2 = (f,f) - \sum_{i=1}^{N} c_i^2 - c_{N+1}^2 = M_N - c_{N+1}^2. \tag{5.35}$$

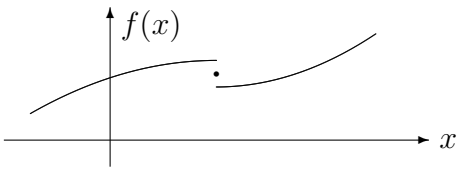

Figure 16: A Function with a Jump Discontinuity.

That is, adding a new term in the series decreases the error (if that new term is nonzero). Thus, in general, higher order approximations are better than lower order approximations.

The set of basis functions $\phi_i(x)$ is defined as a *complete* set if

$$\lim_{N\to\infty} M_N = \lim_{N\to\infty} \int_a^b \left[f(x) - \sum_{i=1}^{N} c_i \phi_i(x) \right]^2 dx = 0. \tag{5.36}$$

If this equation holds, the Fourier series is said to *converge in the mean*. Having a complete set means that the set of basis functions $\phi_i(x)$ is not missing any members. Thus, we could alternatively define the set $\phi_i(x)$ as complete if there is no function with positive norm which is orthogonal to each basis function $\phi_i(x)$ in the set. It turns out that the set of basis functions in Eq. 5.7 forms a complete set in $[-L, L]$.

Note that convergence in the mean does not imply convergence pointwise. That is, it may not be true that

$$\lim_{N\to\infty} \left| f(x) - \sum_{i=1}^{N} c_i \phi_i(x) \right| = 0 \tag{5.37}$$

for all x. For example, the series representation of a discontinuous function $f(x)$ will not have pointwise convergence at the discontinuity (Fig. 16).

The vector analog to completeness for sets of functions can be illustrated by observing that, to express an arbitrary three-dimensional vector in terms of a sum of coordinate basis vectors, three basis vectors are needed. A basis consisting of only two unit vectors is incomplete, since it in incapable of representing arbitrary vectors in three dimensions.

5.3 Fourier Expansions Using a Polynomial Basis

Recall, in the generalized Fourier series, we represent the function $f(x)$ with the expansion

$$f(x) = \sum_{i=1}^{N} c_i \phi_i(x), \tag{5.38}$$

although the equality is approximate if N is finite. Let the basis functions be

$$\{\phi_i\} = \{1, x, x^2, x^3, \ldots\}, \tag{5.39}$$

which is a set of independent, nonorthogonal functions. If we take the inner product of both sides of Eq. 5.38 with ϕ_j, we obtain

$$(f, \phi_j) = \sum_{i=1}^{N} c_i (\phi_i, \phi_j), \tag{5.40}$$

where, in general, since the basis functions are nonorthogonal,

$$(\phi_i, \phi_j) \neq 0 \quad \text{for } i \neq j. \tag{5.41}$$

Thus, Eq. 5.40 is a set of *coupled* equations which can be solved to yield c_i:

$$\begin{array}{l} (f, \phi_1) = (\phi_1, \phi_1)c_1 + (\phi_1, \phi_2)c_2 + (\phi_1, \phi_3)c_3 + \cdots \\ (f, \phi_2) = (\phi_2, \phi_1)c_1 + (\phi_2, \phi_2)c_2 + (\phi_2, \phi_3)c_3 + \cdots \\ (f, \phi_3) = (\phi_3, \phi_1)c_1 + (\phi_3, \phi_2)c_2 + (\phi_3, \phi_3)c_3 + \cdots \\ \vdots \end{array} \tag{5.42}$$

or

$$\mathbf{Ac} = \mathbf{b}, \tag{5.43}$$

where

$$A_{ij} = (\phi_i, \phi_j), \quad \mathbf{c} = \begin{Bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_N \end{Bmatrix}, \quad \mathbf{b} = \begin{Bmatrix} (f, \phi_1) \\ (f, \phi_2) \\ (f, \phi_3) \\ \vdots \\ (f, \phi_N) \end{Bmatrix}, \tag{5.44}$$

and $\phi_i(x) = x^{i-1}$.

If we assume the domain $0 \leq x \leq 1$, we obtain

$$A_{ij} = \int_0^1 x^{i-1} x^{j-1}\, dx = \int_0^1 x^{i+j-2}\, dx = \left. \frac{x^{i+j-1}}{i+j-1} \right|_0^1 = \frac{1}{i+j-1} \tag{5.45}$$

or

$$\mathbf{A} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{5.46}$$

which is referred to as the Hilbert matrix.

There are several problems with this approach:

1. Since the set $\{\phi_i\}$ is not orthogonal, we must solve a set of *coupled* equations to obtain the coefficients c_i.

2. The coefficients c_i depend on the number N of terms in the expansion. That is, to add more terms, all c_i change, not just the new coefficients.

3. The resulting coefficient matrix, the Hilbert matrix, is terribly conditioned even for very small matrices.

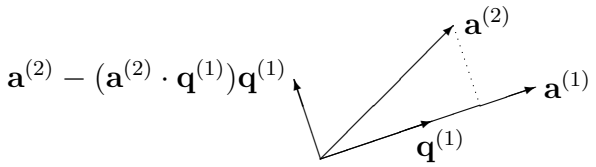

Figure 17: The First Two Vectors in Gram-Schmidt.

These problems can be eliminated if we switch to a set of *orthogonal* polynomials, which can be found using the Gram-Schmidt process. Let $\phi_i(x)$ denote the nonorthogonal set of basis functions

$$\phi_0(x) = 1, \quad \phi_1(x) = x, \quad \phi_2(x) = x^2, \quad \ldots, \tag{5.47}$$

where we have chosen to count the functions starting from zero. We let $P_i(x)$ denote the set of orthogonal basis functions to be determined. These functions will not be required to have unit norm.

To facilitate the application of Gram-Schmidt to functions, we recall the Gram-Schmidt algorithm for vectors. Given a set of independent, nonorthogonal vectors $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \mathbf{a}^{(3)}, \ldots$, this procedure computes a set of orthonormal vectors $\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \mathbf{q}^{(3)}, \ldots$ that span the same space. For example, the first two orthonormal vectors are

$$\mathbf{q}^{(1)} = \frac{\mathbf{a}^{(1)}}{|\mathbf{a}^{(1)}|}, \quad \mathbf{q}^{(2)} = \frac{\mathbf{a}^{(2)} - (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)}}{|\mathbf{a}^{(2)} - (\mathbf{a}^{(2)} \cdot \mathbf{q}^{(1)})\mathbf{q}^{(1)}|}, \tag{5.48}$$

as illustrated in Fig. 17.

For convenience, we choose the domain $-1 \le x \le 1$, in which case

$$(\phi_i, \phi_j) = 0, \quad i + j = \text{odd}, \tag{5.49}$$

since the integral of an odd power of x over symmetrical limits is zero (i.e., the integrand is an odd function of x).

Thus, by analogy with Gram-Schmidt for vectors, we obtain

$$P_0(x) = \phi_0(x) = 1, \tag{5.50}$$

$$P_1(x) = \phi_1(x) - (\phi_1, P_0)\frac{P_0(x)}{(P_0, P_0)} = x - (x, 1)\frac{1}{(1, 1)} = x - 0 = x, \tag{5.51}$$

$$P_2(x) = \phi_2(x) - (\phi_2, P_0)\frac{P_0(x)}{(P_0, P_0)} - (\phi_2, P_1)\frac{P_1(x)}{(P_1, P_1)} = x^2 - (x^2, 1)\frac{1}{(1, 1)} - (x^2, x)\frac{x}{(x, x)}, \tag{5.52}$$

where

$$(x^2, 1) = \int_{-1}^{1} x^2\, dx = \frac{2}{3}, \quad (1, 1) = \int_{-1}^{1} dx = 2, \quad (x^2, x) = 0. \tag{5.53}$$

Thus,

$$P_2(x) = x^2 - \frac{1}{3}, \tag{5.54}$$

and so on. These polynomials, called *Legendre polynomials*, are orthogonal over [-1,1]. Because of their orthogonality, the Legendre polynomials are a better choice for polynomial expansions than the polynomials of the Taylor series, Eq. 5.47.

With the basis functions ϕ_i given by the Legendre polynomials, Eq. 5.40 becomes

$$(f, \phi_j) = \sum_{i=1}^{N} c_i(\phi_i, \phi_j) = c_j(\phi_j, \phi_j) \quad \text{(no sum on } j\text{)}, \tag{5.55}$$

since each equation in Eq. 5.40 has only one nonzero term on the right-hand side (the term $i = j$).

5.4 Similarity of Fourier Series With Least Squares Fitting

In the least squares problem, given a finite number m of points (x_i, f_i) and the n-term fitting function

$$p(x) = a_i\phi_i(x), \tag{5.56}$$

the normal equations are, from Eq. 4.36,

$$[\phi_i(x_k)\phi_j(x_k)]a_j = f_j\phi_i(x_j), \tag{5.57}$$

where the order of the resulting system is n, the number of basis functions. The summation convention is used, so that a summation is implied over repeated indices.

In the Fourier series problem, given a function $f(x)$ and the generalized Fourier series

$$f(x) = c_i\phi_i(x), \tag{5.58}$$

the coefficients are determined from

$$(\phi_i, \phi_j)c_j = (f, \phi_i), \tag{5.59}$$

where the summation convention is used again.

Note the similarities between Eqs. 5.57 and 5.59. These two equations are essentially the same, except that Eq. 5.57 is the discrete form of Eq. 5.59. Also, the basis functions in Eq. 5.59 are orthogonal, so that Eq. 5.59 is a diagonal system.

6 Eigenvalue Problems

If, for a square matrix $\mathbf{A}$ of order n, there exists a vector $\mathbf{x} \neq \mathbf{0}$ and a number λ such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{6.1}$$

λ is called an *eigenvalue* of $\mathbf{A}$, and $\mathbf{x}$ is the corresponding *eigenvector*. Note that Eq. 6.1 can alternatively be written in the form

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{I}\mathbf{x}, \tag{6.2}$$

where $\mathbf{I}$ is the identity matrix, or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \tag{6.3}$$

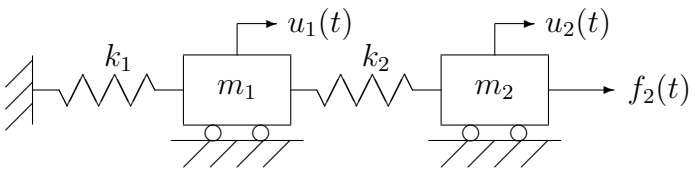

Figure 18: 2-DOF Mass-Spring System.

which is a system of n linear algebraic equations. If the coefficient matrix $\mathbf{A} - \lambda\mathbf{I}$ were nonsingular, the unique solution of Eq. 6.3 would be $\mathbf{x} = \mathbf{0}$, which is not of interest. Thus, to obtain a *nonzero* solution of the eigenvalue problem, $\mathbf{A} - \lambda\mathbf{I}$ must be singular, which implies that

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} - \lambda & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{6.4}$$

We note that this determinant, when expanded, yields a polynomial of degree n referred to as the *characteristic polynomial* associated with the matrix. The equation obtained by equating this polynomial with zero is referred to as the *characteristic equation* for the matrix. The eigenvalues are thus the roots of the characteristic polynomial. Since a polynomial of degree n has n roots, $\mathbf{A}$ has n eigenvalues (not necessarily real).

Eigenvalue problems arise in many physical applications, including free vibrations of mechanical systems, buckling of structures, and the calculation of principal axes of stress, strain, and inertia.

6.1 Example 1: Mechanical Vibrations

Consider the undamped two-degree-of-freedom mass-spring system shown in Fig. 18. We let u_1 and u_2 denote the displacements from the equilibrium of the two masses m_1 and m_2. The stiffnesses of the two springs are k_1 and k_2.

For the purposes of computation, let $k_1 = k_2 = 1$ and $m_1 = m_2 = 1$. From Newton's second law of motion ($F = ma$), the equations of motion of this system are

$$\left.\begin{aligned} \ddot{u}_1 + 2u_1 - u_2 &= 0 \\ \ddot{u}_2 - u_1 + u_2 &= f_2 \end{aligned}\right\}, \tag{6.5}$$

where dots denote differentiation with respect to the time t. In matrix notation, this system can be written

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{Bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \end{Bmatrix} + \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ f_2 \end{Bmatrix} \tag{6.6}$$

or

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}, \tag{6.7}$$

where

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{F} = \begin{Bmatrix} 0 \\ f_2 \end{Bmatrix}. \tag{6.8}$$

These three matrices are referred to as the system's mass, stiffness, and force matrices, respectively.

With zero applied force (the free undamped vibration problem),

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{0}. \tag{6.9}$$

We look for nonzero solutions of this equation in the form

$$\mathbf{u} = \mathbf{u}_0 \cos \omega t. \tag{6.10}$$

That is, we look for solutions of Eq. 6.9 which are sinusoidal in time and where both DOF oscillate with the same circular frequency ω. The vector $\mathbf{u}_0$ is the amplitude vector for the solution. The substitution of Eq. 6.10 into Eq. 6.9 yields

$$-\mathbf{M}\omega^2\mathbf{u}_0 \cos \omega t + \mathbf{K}\mathbf{u}_0 \cos \omega t = \mathbf{0}. \tag{6.11}$$

Since this equation must hold for all time t,

$$(-\omega^2\mathbf{M} + \mathbf{K})\mathbf{u}_0 = \mathbf{0} \tag{6.12}$$

or

$$\mathbf{K}\mathbf{u}_0 = \omega^2\mathbf{M}\mathbf{u}_0 \tag{6.13}$$

or

$$\mathbf{M}^{-1}\mathbf{K}\mathbf{u}_0 = \omega^2\mathbf{u}_0. \tag{6.14}$$

This is an eigenvalue problem in standard form if we define $\mathbf{A} = \mathbf{M}^{-1}\mathbf{K}$ and $\omega^2 = \lambda$:

$$\mathbf{A}\mathbf{u}_0 = \lambda\mathbf{u}_0 \tag{6.15}$$

or

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u}_0 = \mathbf{0}. \tag{6.16}$$

For this problem,

$$\mathbf{A} = \mathbf{M}^{-1}\mathbf{K} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}. \tag{6.17}$$

Since we want the eigenvalues and eigenvectors of $\mathbf{A}$, it follows from Eq. 6.16 that

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 2-\lambda & -1 \\ -1 & 1-\lambda \end{vmatrix} = (2-\lambda)(1-\lambda) - 1 = \lambda^2 - 3\lambda + 1 = 0. \tag{6.18}$$

Thus,

$$\lambda = \frac{3 \pm \sqrt{5}}{2} \approx 0.382 \text{ and } 2.618\,, \tag{6.19}$$

and the circular frequencies ω (the square root of λ) are

$$\omega_1 = \sqrt{\lambda_1} = 0.618, \quad \omega_2 = \sqrt{\lambda_2} = 1.618. \tag{6.20}$$

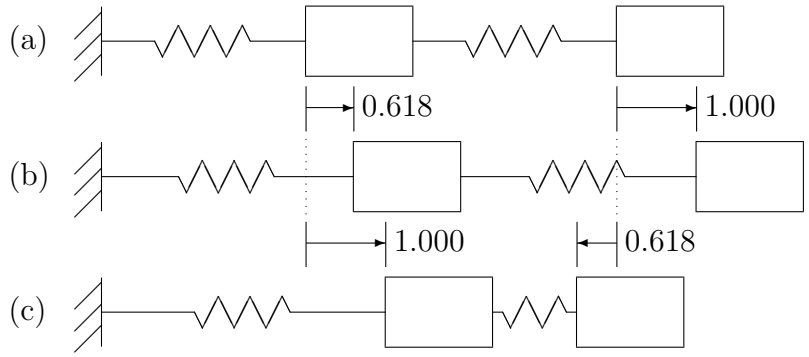

Figure 19: Mode Shapes for 2-DOF Mass-Spring System. (a) Undeformed shape, (b) Mode 1 (masses in-phase), (c) Mode 2 (masses out-of-phase).

To obtain the eigenvectors, we solve Eq. 6.16 for each eigenvalue found. For the first eigenvalue, $\lambda = 0.382$, we obtain

$$\begin{bmatrix} 1.618 & -1 \\ -1 & 0.618 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}, \tag{6.21}$$

which is a redundant system of equations satisfied by

$$\mathbf{u}^{(1)} = \begin{Bmatrix} 0.618 \\ 1 \end{Bmatrix}. \tag{6.22}$$

The superscript indicates that this is the eigenvector associated with the first eigenvalue. Note that, since the eigenvalue problem is a homogeneous problem, any nonzero multiple of $\mathbf{u}^{(1)}$ is also an eigenvector.

For the second eigenvalue, $\lambda = 2.618$, we obtain

$$\begin{bmatrix} -0.618 & -1 \\ -1 & -1.618 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}, \tag{6.23}$$

a solution of which is

$$\mathbf{u}^{(2)} = \begin{Bmatrix} 1 \\ -0.618 \end{Bmatrix}. \tag{6.24}$$

Physically, ω_i is a *natural frequency* of the system (in radians per second), and $\mathbf{u}^{(i)}$ is the corresponding *mode shape*. An n-DOF system has n natural frequencies and mode shapes. Note that natural frequencies are not commonly expressed in radians per second, but in cycles per second or Hertz (Hz), where $\omega = 2\pi f$. The lowest natural frequency for a system is referred to as the *fundamental frequency* for the system.

The mode shapes for this example are sketched in Fig. 19. For Mode 1, the two masses are vibrating in-phase, and m_2 has the larger displacement, as expected. For Mode 2, the two masses are vibrating out-of-phase.

The determinant approach used above is fine for small matrices but not well-suited to numerical computation involving large matrices.

6.2 Properties of the Eigenvalue Problem

There are several useful properties associated with the eigenvalue problem:

Property 1. Eigenvectors are unique only up to an arbitrary multiplicative constant. From the eigenvalue problem, Eq. 6.1, if $\mathbf{x}$ is a solution, $\alpha\mathbf{x}$ is also a solution for any nonzero scalar α. We note that Eq. 6.1 is a homogeneous equation.

Property 2. The eigenvalues of a triangular matrix are the diagonal entries. For example, let

$$\mathbf{A} = \begin{bmatrix} 1 & 4 & 5 \\ 0 & 4 & 7 \\ 0 & 0 & 9 \end{bmatrix}, \tag{6.25}$$

for which the characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 4 & 5 \\ 0 & 4-\lambda & 7 \\ 0 & 0 & 9-\lambda \end{vmatrix} = (1-\lambda)(4-\lambda)(9-\lambda) = 0, \tag{6.26}$$

which implies $\lambda = 1, 4, 9$. Although the eigenvalues of a triangular matrix are obvious, it turns out that the eigenvectors are not obvious.

Property 3. The eigenvalues of a diagonal matrix are the diagonal entries, and the eigenvectors are

$$\begin{Bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}, \quad \begin{Bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{Bmatrix}, \quad \begin{Bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{Bmatrix}, \quad \text{etc.} \tag{6.27}$$

For example, let

$$\mathbf{A} = \begin{bmatrix} 11 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}. \tag{6.28}$$

For $\lambda_1 = 11$,

$$\mathbf{x}^{(1)} = \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}, \tag{6.29}$$

since

$$\mathbf{A}\mathbf{x}^{(1)} = \begin{bmatrix} 11 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix} \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix} = \begin{Bmatrix} 11 \\ 0 \\ 0 \end{Bmatrix} = 11 \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix} = \lambda_1 \mathbf{x}^{(1)}. \tag{6.30}$$

Similarly,

$$\mathbf{A}\mathbf{x}^{(2)} = 4 \begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix}, \quad \mathbf{A}\mathbf{x}^{(3)} = 9 \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix}. \tag{6.31}$$

Property 4. The sum of the eigenvalues of $\mathbf{A}$ is equal to the trace of $\mathbf{A}$; i.e., for a matrix of order n,

$$\sum_{i=1}^{n} \lambda_i = \operatorname{tr} \mathbf{A}. \tag{6.32}$$

(The *trace* of a matrix is defined as the sum of the main diagonal terms.) To prove this property, we first note, from Eq. 6.4, that

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) &= \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} - \lambda & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} - \lambda \end{vmatrix} \\ &= (-\lambda)^n + (a_{11} + a_{22} + \cdots + a_{nn})(-\lambda)^{n-1} + \cdots . \end{aligned} \tag{6.33}$$

No other products in the determinant contribute to the $(-\lambda)^{n-1}$ term. For example, the cofactor of a_{12} deletes two matrix terms involving λ, so that the largest power of λ from that cofactor is $n-2$. On the other hand, since $\det(\mathbf{A} - \lambda \mathbf{I})$ is a polynomial of degree n, we can write it in the factored form

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}) &= (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda) \\ &= (-\lambda)^n + (\lambda_1 + \lambda_2 + \cdots + \lambda_n)(-\lambda)^{n-1} + \cdots + \lambda_1 \lambda_2 \lambda_3 \cdots \lambda_n. \end{aligned} \tag{6.34}$$

Since the polynomials in Eqs. 6.33 and 6.34 must be identically equal for all λ, we equate the coefficients of the $(-\lambda)^{n-1}$ terms to obtain

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = a_{11} + a_{22} + \cdots + a_{nn}, \tag{6.35}$$

which is the desired result.

Property 5. The product of the eigenvalues equals the determinant of $\mathbf{A}$, i.e.,

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det \mathbf{A}. \tag{6.36}$$

This property results immediately from Eq. 6.34 by setting $\lambda = 0$.

Property 6. If the eigenvectors of $\mathbf{A}$ are linearly independent, and a matrix $\mathbf{S}$ is formed whose columns are the eigenvectors, then the matrix product

$$\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_n \end{bmatrix} \tag{6.37}$$

is diagonal. This property is proved in block form as follows:

$$\begin{aligned}
\mathbf{AS} &= \mathbf{A}\left[\begin{array}{cccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(n)} \end{array}\right] \\
&= \left[\begin{array}{cccc} \mathbf{A}\mathbf{x}^{(1)} & \mathbf{A}\mathbf{x}^{(2)} & \cdots & \mathbf{A}\mathbf{x}^{(n)} \end{array}\right] \\
&= \left[\begin{array}{cccc} \lambda_1\mathbf{x}^{(1)} & \lambda_2\mathbf{x}^{(2)} & \cdots & \lambda_n\mathbf{x}^{(n)} \end{array}\right] \\
&= \left[\begin{array}{cccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(n)} \end{array}\right] \left[\begin{array}{ccccc} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_n \end{array}\right] = \mathbf{S}\boldsymbol{\Lambda}.
\end{aligned} \tag{6.38}$$

Since the columns of $\mathbf{S}$ are independent, $\mathbf{S}$ is invertible, and

$$\mathbf{S}^{-1}\mathbf{AS} = \boldsymbol{\Lambda} \quad \text{or} \quad \mathbf{A} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^{-1}. \tag{6.39}$$

Property 7. Distinct eigenvalues yield independent eigenvectors. To prove this property, we consider two eigensolutions:

$$\mathbf{A}\mathbf{x}^{(1)} = \lambda_1\mathbf{x}^{(1)} \tag{6.40}$$

$$\mathbf{A}\mathbf{x}^{(2)} = \lambda_2\mathbf{x}^{(2)}. \tag{6.41}$$

To show that $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are independent, we must show that

$$c_1\mathbf{x}^{(1)} + c_2\mathbf{x}^{(2)} = \mathbf{0} \tag{6.42}$$

implies $c_1 = c_2 = 0$. We first multiply Eq. 6.42 by $\mathbf{A}$ to obtain

$$c_1\mathbf{A}\mathbf{x}^{(1)} + c_2\mathbf{A}\mathbf{x}^{(2)} = c_1\lambda_1\mathbf{x}^{(1)} + c_2\lambda_2\mathbf{x}^{(2)} = \mathbf{0}. \tag{6.43}$$

We then multiply Eq. 6.42 by λ_2 to obtain

$$c_1\lambda_2\mathbf{x}^{(1)} + c_2\lambda_2\mathbf{x}^{(2)} = \mathbf{0}. \tag{6.44}$$

If we subtract these last two equations, we obtain

$$c_1(\lambda_1 - \lambda_2)\mathbf{x}^{(1)} = \mathbf{0}, \tag{6.45}$$

which implies $c_1 = 0$, since the eigenvalues are distinct ($\lambda_1 \neq \lambda_2$). Similarly, $c_2 = 0$, and the eigenvectors are independent. From the last two properties, we conclude that a matrix with distinct eigenvalues can always be diagonalized. It turns out that a matrix with repeated eigenvalues cannot always be diagonalized.

Property 8. The eigenvalues and eigenvectors of a real, symmetric matrix are real. To prove this property, we consider the eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{6.46}$$

where $\mathbf{A}$ is a real, symmetric matrix. If we take the complex conjugate of both sides of this equation, we obtain

$$\mathbf{A}\mathbf{x}^* = \lambda^*\mathbf{x}^*, \tag{6.47}$$

where the asterisk (*) denotes the complex conjugate. The goal is to show that $\lambda = \lambda^*$. We multiply Eq. 6.46 by $\mathbf{x}^{*T}$ (the conjugate transpose) and Eq. 6.47 by $\mathbf{x}^T$ to obtain

$$\mathbf{x}^{*T}\mathbf{A}\mathbf{x} = \lambda\mathbf{x}^{*T}\mathbf{x}, \tag{6.48}$$

$$\mathbf{x}^T\mathbf{A}\mathbf{x}^* = \lambda^*\mathbf{x}^T\mathbf{x}^*, \tag{6.49}$$

where the two left-hand sides are equal, since they are the transposes of each other, and both are scalars (1×1 matrices). Thus,

$$\lambda\mathbf{x}^{*T}\mathbf{x} = \lambda^*\mathbf{x}^T\mathbf{x}^*, \tag{6.50}$$

where $\mathbf{x}^{*T}\mathbf{x} = \mathbf{x}^T\mathbf{x}^* \neq 0$ implies $\lambda = \lambda^*$ (i.e., λ is real). The eigenvectors are also real, since they are obtained by solving the equation

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \tag{6.51}$$

for $\mathbf{x}$, given λ.

In engineering applications, the form of the eigenvalue problem that is of greater interest than Eq. 6.46 is the *generalized eigenvalue problem*

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}, \tag{6.52}$$

where both $\mathbf{A}$ and $\mathbf{B}$ are real and symmetric. Having $\mathbf{A}$ and $\mathbf{B}$ real and symmetric is not sufficient, as is, to show that the generalized eigenvalue problem has only real eigenvalues, as can be seen from the example

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{6.53}$$

for which Eq. 6.52 has imaginary eigenvalues $\pm i$, where $i = \sqrt{-1}$. However, it turns out that, if $\mathbf{B}$ is not only real and symmetric but also positive definite (a term to be defined later in §6.8, p. 82), the generalized eigenvalue problem has real eigenvalues [10]. Since, in engineering applications, $\mathbf{B}$ is usually positive definite, the generalized eigenvalue problems that arise in engineering have real eigenvalues.

Matrix formulations of engineering problems generally yield symmetric coefficient matrices, so this property is of great practical significance. As a result, various physical quantities of interest, which must be real, are guaranteed to be real, including the natural frequencies of vibration and mode shapes (the eigenvectors) of undamped mechanical systems, buckling loads in elastic stability problems, and principal stresses.

Property 9. The eigenvectors of a real, symmetric matrix with distinct eigenvalues are mutually orthogonal. To prove this property, we consider from the outset the generalized eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}, \tag{6.54}$$

where $\mathbf{A}$ and $\mathbf{B}$ are both real, symmetric matrices. For two eigenvalues λ_1 and λ_2,

$$\mathbf{A}\mathbf{x}^{(1)} = \lambda_1 \mathbf{B}\mathbf{x}^{(1)}, \tag{6.55}$$

$$\mathbf{A}\mathbf{x}^{(2)} = \lambda_2 \mathbf{B}\mathbf{x}^{(2)}, \tag{6.56}$$

from which it follows that the scalar

$$\lambda_1 \mathbf{x}^{(2)T}\mathbf{B}\mathbf{x}^{(1)} = \mathbf{x}^{(2)T}\mathbf{A}\mathbf{x}^{(1)} = \mathbf{x}^{(1)T}\mathbf{A}\mathbf{x}^{(2)} = \lambda_2 \mathbf{x}^{(1)T}\mathbf{B}\mathbf{x}^{(2)} = \lambda_2 \mathbf{x}^{(2)T}\mathbf{B}\mathbf{x}^{(1)}. \tag{6.57}$$

Thus,

$$(\lambda_1 - \lambda_2)\mathbf{x}^{(2)T}\mathbf{B}\mathbf{x}^{(1)} = 0, \tag{6.58}$$

and, if the eigenvalues are distinct ($\lambda_1 \neq \lambda_2$),

$$\mathbf{x}^{(2)T}\mathbf{B}\mathbf{x}^{(1)} = 0. \tag{6.59}$$

Also, since $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$,

$$\mathbf{x}^{(2)T}\mathbf{A}\mathbf{x}^{(1)} = 0. \tag{6.60}$$

The eigenvectors are said to be orthogonal with respect to the matrices $\mathbf{A}$ and $\mathbf{B}$.

In mechanical vibrations, $\mathbf{A}$ represents the stiffness matrix, and $\mathbf{B}$ represents the mass matrix. For $\mathbf{x}$ an eigenvector, Eq. 6.54 implies the scalar equation

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda \mathbf{x}^T\mathbf{B}\mathbf{x} \tag{6.61}$$

or

$$\lambda = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{B}\mathbf{x}}, \tag{6.62}$$

which is referred to as the *Rayleigh quotient*. In mechanical vibrations, the numerator and denominator are the generalized stiffness and mass, respectively, for a given vibration mode, and $\lambda = \omega^2$ is the square of the circular frequency.

The orthogonality relations proved here also imply that

$$\mathbf{S}^T\mathbf{A}\mathbf{S} = \mathbf{\Lambda}\mathbf{S}^T\mathbf{B}\mathbf{S}, \tag{6.63}$$

where $\mathbf{S}$ is the matrix of eigenvectors (in the columns), $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues defined in Eq. 6.37, and $\mathbf{S}^T\mathbf{A}\mathbf{S}$ and $\mathbf{S}^T\mathbf{B}\mathbf{S}$ are both diagonal matrices. To prove Eq. 6.63, we note that, with the matrix $\mathbf{S}$ of eigenvectors defined as

$$\mathbf{S} = \begin{bmatrix} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots \end{bmatrix}, \tag{6.64}$$

it follows that

$$\mathbf{S}^T\mathbf{A}\mathbf{S} = \begin{Bmatrix} \mathbf{x}^{(1)T} \\ \mathbf{x}^{(2)T} \\ \vdots \end{Bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots \end{bmatrix} = \begin{bmatrix} \mathbf{x}^{(1)T}\mathbf{A}\mathbf{x}^{(1)} & \mathbf{x}^{(1)T}\mathbf{A}\mathbf{x}^{(2)} & \cdots \\ \mathbf{x}^{(2)T}\mathbf{A}\mathbf{x}^{(1)} & \mathbf{x}^{(2)T}\mathbf{A}\mathbf{x}^{(2)} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \tag{6.65}$$

where the off-diagonal terms are zero by orthogonality. A similar expression applies for $\mathbf{B}$.

If $\mathbf{Ax} = \lambda\mathbf{x}$ (i.e., $\mathbf{B} = \mathbf{I}$), and the eigenvectors are normalized to unit length, then the matrix $\mathbf{S}$ is orthogonal, and

$$\mathbf{S}^T\mathbf{AS} = \mathbf{\Lambda}\mathbf{S}^T\mathbf{IS} = \mathbf{\Lambda} \tag{6.66}$$

or

$$\mathbf{A} = \mathbf{S\Lambda S}^T. \tag{6.67}$$

This last result, referred to as the *spectral theorem*, means that a real, symmetric matrix can be factored into $\mathbf{A} = \mathbf{S\Lambda S}^T$, where $\mathbf{S}$ has orthonormal eigenvectors in the columns, and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues.

Conversely, if the eigenvalues of a matrix $\mathbf{A}$ are real, and the eigenvectors are mutually orthogonal, $\mathbf{A}$ is necessarily symmetric. Since orthogonal eigenvectors are independent, Eq. 6.39 implies that $\mathbf{A}$ can be written in the form

$$\mathbf{A} = \mathbf{S\Lambda S}^{-1}, \tag{6.68}$$

where $\mathbf{S}$ is the matrix whose columns are the eigenvectors. If the orthogonal eigenvectors are normalized to unit length, $\mathbf{S}$ is also orthogonal (i.e., $\mathbf{S}^{-1} = \mathbf{S}^T$), and

$$\mathbf{A} = \mathbf{S\Lambda S}^T, \tag{6.69}$$

which is a symmetric matrix.

6.3 Example 2: Principal Axes of Stress

The matrix of stress components in 3-D elasticity is a tensor of rank 2:

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}, \tag{6.70}$$

where the diagonal components of the matrix are the normal (or direct) stresses, and the off-diagonal components are the shear stresses. This matrix is symmetric.

Recall that an orthogonal transformation of coordinates transforms $\boldsymbol{\sigma}$ according to the rule

$$\boldsymbol{\sigma}' = \mathbf{R}\boldsymbol{\sigma}\mathbf{R}^T, \tag{6.71}$$

where

$$R_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j, \tag{6.72}$$

$\mathbf{e}'_i$ and $\mathbf{e}_j$ are unit vectors in the coordinate directions, and $\mathbf{R}$ is an orthogonal matrix ($\mathbf{RR}^T = \mathbf{I}$).

A key problem is to determine whether there is an orthogonal transformation of coordinates (i.e., a rotation of axes) $\mathbf{x}' = \mathbf{Rx}$ such that the stress tensor $\boldsymbol{\sigma}$ is diagonalized:

$$\boldsymbol{\sigma}' = \begin{bmatrix} \sigma'_{11} & 0 & 0 \\ 0 & \sigma'_{22} & 0 \\ 0 & 0 & \sigma'_{33} \end{bmatrix} = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix}, \tag{6.73}$$

where the diagonal stresses in this new coordinate system are denoted s_1, s_2, s_3.

From Eq. 6.71,

$$\boldsymbol{\sigma}\mathbf{R}^T = \mathbf{R}^T\boldsymbol{\sigma}'. \tag{6.74}$$

We now let $\mathbf{v}_i$ denote the ith column of $\mathbf{R}^T$; i.e.,

$$\mathbf{R}^T = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3], \tag{6.75}$$

in which case Eq. 6.74 can be written in matrix form as

$$\boldsymbol{\sigma}[\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3] = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]\begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix} = [s_1\mathbf{v}_1\ \ s_2\mathbf{v}_2\ \ s_3\mathbf{v}_3]. \tag{6.76}$$

(The various matrices in this equation are conformable from a "block" point of view, since the left-hand side is, for example, the product of a 1×1 matrix with a 1×3 matrix.) Each column of this equation is

$$\boldsymbol{\sigma}\mathbf{v}_i = s_i\mathbf{v}_i \ \ \text{(no sum on } i\text{)}. \tag{6.77}$$

Thus, the original desire to find a coordinate rotation which would transform the stress tensor to diagonal form reduces to seeking vectors $\mathbf{v}$ such that

$$\boldsymbol{\sigma}\mathbf{v} = s\mathbf{v}. \tag{6.78}$$

Eq. 6.78 is an eigenvalue problem with s the eigenvalue, and $\mathbf{v}$ the corresponding eigenvector. The goal in solving Eq. 6.78 (and eigenvalue problems in general) is to find nonzero vectors $\mathbf{v}$ which satisfy Eq. 6.78 for some scalar s. Geometrically, the goal in solving Eq. 6.78 is to find nonzero vectors $\mathbf{v}$ which, when multiplied by the matrix $\boldsymbol{\sigma}$, result in new vectors which are parallel to $\mathbf{v}$.

Eq. 6.78 is equivalent to the matrix system

$$(\boldsymbol{\sigma} - s\mathbf{I})\mathbf{v} = \mathbf{0}. \tag{6.79}$$

For nontrivial solutions ($\mathbf{v} \neq \mathbf{0}$), the matrix $\boldsymbol{\sigma} - s\mathbf{I}$ must be singular, implying that

$$\det(\boldsymbol{\sigma} - s\mathbf{I}) = \begin{vmatrix} \sigma_{11} - s & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} - s & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} - s \end{vmatrix} = 0, \tag{6.80}$$

which is referred to as the *characteristic equation* associated with the eigenvalue problem. The characteristic equation is a cubic polynomial in s, since, when expanded, yields

$$\begin{aligned}(\sigma_{11} - s)[(\sigma_{22} - s)(\sigma_{33} - s) - \sigma_{23}\sigma_{32}] - \sigma_{12}[\sigma_{21}(\sigma_{33} - s) - \sigma_{23}\sigma_{31}] \\ + \sigma_{13}[\sigma_{21}\sigma_{32} - \sigma_{31}(\sigma_{22} - s)] = 0\end{aligned} \tag{6.81}$$

or

$$-s^3 + \theta_1 s^2 - \theta_2 s + \theta_3 = 0, \tag{6.82}$$

where

$$\begin{cases} \theta_1 = \sigma_{11} + \sigma_{22} + \sigma_{33} = \sigma_{ii} = \operatorname{tr}\boldsymbol{\sigma} \\ \theta_2 = \sigma_{22}\sigma_{33} + \sigma_{33}\sigma_{11} + \sigma_{11}\sigma_{22} - \sigma_{31}^2 - \sigma_{12}^2 - \sigma_{23}^2 = \frac{1}{2}(\sigma_{ii}\sigma_{jj} - \sigma_{ij}\sigma_{ji}) \\ \theta_3 = \det\boldsymbol{\sigma}. \end{cases} \tag{6.83}$$

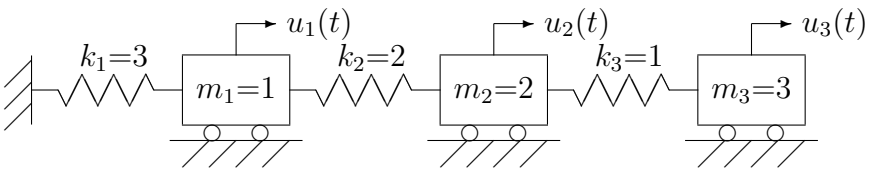

Figure 20: 3-DOF Mass-Spring System.

The stresses s_1, s_2, s_3, which are the three solutions of the characteristic equation, are referred to as the *principal stresses*. The resulting stress tensor is

$$\boldsymbol{\sigma} = \begin{bmatrix} s_1 & 0 & 0 \\ 0 & s_2 & 0 \\ 0 & 0 & s_3 \end{bmatrix}, \tag{6.84}$$

and the coordinate axes of the coordinate system in which the stress tensor is diagonal is referred to as the *principal axes of stress* or the *principal coordinates*.

The principal axes of stress are the eigenvectors of the stress tensor, since the (normalized) eigenvectors are columns of $\mathbf{R}^T$ (rows of $\mathbf{R}$), where

$$R_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j. \tag{6.85}$$

Thus, Row i of $\mathbf{R}$ consists of the direction cosines of the ith principal axis.

Since the principal stresses are independent of the original coordinate system, the coefficients of the characteristic polynomial, Eq. 6.82, must be *invariant* with respect to a coordinate rotation. Thus, θ_1, θ_2, and θ_3 are referred to as the *invariants* of the stress tensor. That is, the three invariants have the same values in all coordinate systems.

In principal coordinates, the stress invariants are

$$\begin{cases} \theta_1 = s_1 + s_2 + s_3 \\ \theta_2 = s_2 s_3 + s_3 s_1 + s_1 s_2 \\ \theta_3 = s_1 s_2 s_3 \end{cases} \tag{6.86}$$

We note that the above theory for eigenvalues, principal axes, and invariants is applicable to all tensors of rank 2, since we did not use the fact that we were dealing specifically with stress. For example, strain is also a tensor of rank 2. A geometrical example of a tensor of rank 2 is the inertia matrix I_{ij} whose matrix elements are moments of inertia. The determination of principal axes of inertia in two dimensions in an eigenvalue problem.

6.4 Computing Eigenvalues by Power Iteration

Determinants are not a practical way to compute eigenvalues and eigenvectors except for very small systems ($n = 2$ or 3). Here we show a classical approach to solving the eigenproblem for large matrices. We will illustrate the procedure with a mechanical vibrations example. Consider the 3-DOF system shown in Fig. 20. For this system, the mass and stiffness matrices are

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \mathbf{K} = \begin{bmatrix} 5 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 1 \end{bmatrix}. \tag{6.87}$$

The eigenvalue problem is, from Eq. 6.13,

$$\mathbf{K}\mathbf{u} = \lambda\mathbf{M}\mathbf{u} \tag{6.88}$$

or

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \tag{6.89}$$

where $\mathbf{u}$ is the vector of displacement amplitudes, and

$$\mathbf{A} = \mathbf{M}^{-1}\mathbf{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 5 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 0 \\ -1 & \frac{3}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}. \tag{6.90}$$

Notice that the product of two symmetric matrices need not be symmetric, even if the first matrix is diagonal.

By definition, a solution (eigenvector) $\mathbf{u}$ of Eq. 6.89 is a vector such that $\mathbf{A}\mathbf{u}$ is a scalar multiple of $\mathbf{u}$, where the scalar multiple is the corresponding eigenvalue λ. Geometrically, $\mathbf{u}$ is an eigenvector if $\mathbf{A}\mathbf{u}$ is parallel to $\mathbf{u}$. That is, an eigenvector of a matrix $\mathbf{A}$ is a vector which is transformed by $\mathbf{A}$ into a new vector with the same orientation but (possibly) different length. Thus, our approach to computing a solution of Eq. 6.89 will be to attempt to guess an eigenvector, check if that vector is transformed by $\mathbf{A}$ into a parallel vector, and if, as expected, our guess is wrong, we try to improve upon the guess iteratively until convergence is achieved.

For the matrix $\mathbf{A}$ given in Eq. 6.90, let our initial guess for an eigenvector be

$$\mathbf{u}^{(0)} = \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}, \tag{6.91}$$

in which case

$$\mathbf{A}\mathbf{u}^{(0)} = \begin{bmatrix} 5 & -2 & 0 \\ -1 & \frac{3}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix} = \begin{Bmatrix} 5 \\ -1 \\ 0 \end{Bmatrix} = 5 \begin{Bmatrix} 1.0 \\ -0.2 \\ 0.0 \end{Bmatrix} = \lambda_1 \mathbf{u}^{(1)}, \tag{6.92}$$

where, to allow easy comparison of $\mathbf{u}^{(1)}$ with $\mathbf{u}^{(0)}$, we factored the largest component from the right-hand side vector. Since $\mathbf{u}^{(0)}$ and $\mathbf{u}^{(1)}$ are not parallel, $\mathbf{u}^{(0)}$ is not an eigenvector. To determine if $\mathbf{u}^{(1)}$ is an eigenvector, we repeat this calculation using $\mathbf{u}^{(1)}$:

$$\mathbf{A}\mathbf{u}^{(1)} = \begin{bmatrix} 5 & -2 & 0 \\ -1 & \frac{3}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{Bmatrix} 1.0 \\ -0.2 \\ 0.0 \end{Bmatrix} = \begin{Bmatrix} 5.40 \\ -1.30 \\ 0.067 \end{Bmatrix} = 5.40 \begin{Bmatrix} 1.00 \\ -0.24 \\ 0.01 \end{Bmatrix} = \lambda_2 \mathbf{u}^{(2)}. \tag{6.93}$$

Since $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$ are not parallel, we conclude that $\mathbf{u}^{(1)}$ is not an eigenvector, and we iterate again. It is useful to summarize all the iterations:

$$\left\{\begin{array}{c}1\\0\\0\end{array}\right\},\ \left\{\begin{array}{r}1.0\\-0.2\\0.0\end{array}\right\},\ \left\{\begin{array}{r}1.00\\-0.24\\0.01\end{array}\right\},\ \left\{\begin{array}{r}1.000\\-0.249\\0.015\end{array}\right\},\ \left\{\begin{array}{r}1.000\\-0.252\\0.016\end{array}\right\},\ \cdots,\ \left\{\begin{array}{r}1.0000\\-0.2518\\0.0162\end{array}\right\}$$
$$5.0 \qquad 5.40 \qquad 5.48 \qquad 5.49 \qquad \cdots \qquad 5.5036$$

Under each iterate for the eigenvector is shown the corresponding iterate for the eigenvalue. Thus, we conclude that one eigenvalue of $\mathbf{A}$ is 5.5036, and the corresponding eigenvector is

$$\left\{\begin{array}{r}1.0000\\-0.2518\\0.0162\end{array}\right\}.$$

This iterative procedure is called the *power method* or *power iteration*.

We now wish to determine which of the three eigenvalues has been found. A symmetric system of order n has n eigenvectors which are orthogonal. Hence, any vector of dimension n can be expanded in a linear combination of such vectors. Let the n eigenvectors of $\mathbf{A}$ be $\boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}, \cdots, \boldsymbol{\phi}^{(n)}$, where $|\lambda_1| < |\lambda_2| < \cdots < |\lambda_n|$. Then

$$\mathbf{u}^{(0)} = c_1\boldsymbol{\phi}^{(1)} + c_2\boldsymbol{\phi}^{(2)} + \cdots + c_n\boldsymbol{\phi}^{(n)} \tag{6.94}$$

and

$$\begin{aligned}\mathbf{u}^{(1)} = \mathbf{A}\mathbf{u}^{(0)} &= c_1\mathbf{A}\boldsymbol{\phi}^{(1)} + c_2\mathbf{A}\boldsymbol{\phi}^{(2)} + \cdots + c_n\mathbf{A}\boldsymbol{\phi}^{(n)}\\ &= c_1\lambda_1\boldsymbol{\phi}^{(1)} + c_2\lambda_2\boldsymbol{\phi}^{(2)} + \cdots + c_n\lambda_n\boldsymbol{\phi}^{(n)}.\end{aligned} \tag{6.95}$$

Similarly,

$$\mathbf{u}^{(2)} = c_1\lambda_1^2\boldsymbol{\phi}^{(1)} + c_2\lambda_2^2\boldsymbol{\phi}^{(2)} + \cdots + c_n\lambda_n^2\boldsymbol{\phi}^{(n)}, \tag{6.96}$$

and, in general, after r iterations,

$$\begin{aligned}\mathbf{u}^{(r)} &= c_1\lambda_1^r\boldsymbol{\phi}^{(1)} + c_2\lambda_2^r\boldsymbol{\phi}^{(2)} + \cdots + c_n\lambda_n^r\boldsymbol{\phi}^{(n)}\\ &= c_n\lambda_n^r\left[\frac{c_1}{c_n}\left(\frac{\lambda_1}{\lambda_n}\right)^r\boldsymbol{\phi}^{(1)} + \frac{c_2}{c_n}\left(\frac{\lambda_2}{\lambda_n}\right)^r\boldsymbol{\phi}^{(2)} + \cdots + \boldsymbol{\phi}^{(n)}\right].\end{aligned} \tag{6.97}$$

Since

$$\left|\frac{\lambda_i}{\lambda_n}\right| < 1 \tag{6.98}$$

for all $i < n$, if follows that

$$\mathbf{u}^{(r)} \to c_n\lambda_n^r\boldsymbol{\phi}^{(n)} \quad \text{as} \quad r \to \infty. \tag{6.99}$$

This last equation shows that power iteration converges to the nth eigenvector (the one corresponding to the largest eigenvalue). However, the largest eigenvalue is rarely of interest in engineering applications. For example, in mechanical vibration problems, it is the low modes that are of interest. In fact, discrete element models such as finite element models are necessarily low frequency models, and the higher modes are not physically meaningful if the original problem was a continuum. In elastic stability problems, the buckling load factor turns out to be an eigenvalue, and it is the lowest buckling load which is normally of interest.

6.5 Inverse Iteration

We recall that iteration with $\mathbf{A}$ in

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \tag{6.100}$$

yields the largest eigenvalue and corresponding eigenvector of $\mathbf{A}$. We could alternatively write the eigenvalue problem as

$$\mathbf{A}^{-1}\mathbf{u} = \frac{1}{\lambda}\mathbf{u}, \tag{6.101}$$

so that iteration with $\mathbf{A}^{-1}$ would converge to the largest value of $1/\lambda$ and hence the smallest λ. This procedure is called *inverse iteration*. In practice, $\mathbf{A}^{-1}$ is not found, since computing $\mathbf{A}^{-1}$ is both expensive and unnecessary. Instead, one solves the system of equations with $\mathbf{A}$ as coefficient matrix:

$$\mathbf{u}^{(k)} = \frac{1}{\lambda}\mathbf{A}\mathbf{u}^{(k+1)}, \tag{6.102}$$

where k is the iteration number. Since, for each iteration, the same coefficient matrix $\mathbf{A}$ is used, one can factor $\mathbf{A} = \mathbf{L}\mathbf{U}$, save the factors, and repeat the forward-backward substitution (FBS) for each iteration.

For the mechanical vibration problem, where $\mathbf{A} = \mathbf{M}^{-1}\mathbf{K}$, Eq. 6.102 becomes

$$\mathbf{M}\mathbf{u}^{(k)} = \frac{1}{\lambda}\mathbf{K}\mathbf{u}^{(k+1)}. \tag{6.103}$$

$\mathbf{K}$ is factored once, and the LU factors saved. Each iteration requires an additional matrix multiplication and FBS.

Consider the generalized eigenvalue problem

$$\mathbf{K}\mathbf{u} = \lambda\mathbf{M}\mathbf{u}, \tag{6.104}$$

where $\mathbf{K}$ and $\mathbf{M}$ are both real, symmetric matrices. We can shift the origin of the eigenvalue scale with the transformation

$$\lambda = \lambda_0 + \mu, \tag{6.105}$$

where λ_0 is a specified shift. Eq. 6.104 then becomes

$$\mathbf{K}\mathbf{u} = (\lambda_0 + \mu)\mathbf{M}\mathbf{u}, \tag{6.106}$$

or

$$(\mathbf{K} - \lambda_0\mathbf{M})\mathbf{u} = \mu\mathbf{M}\mathbf{u}, \tag{6.107}$$

which is an eigenvalue problem with μ as the eigenvalue. This equation can be arranged for inverse iteration to obtain

$$\mathbf{M}\mathbf{u}^{(k)} = \frac{1}{\mu}(\mathbf{K} - \lambda_0\mathbf{M})\mathbf{u}^{(k+1)}, \tag{6.108}$$

where k is the iteration number, so that inverse iteration would converge to the smallest eigenvalue μ. Thus, by picking λ_0, we can find the eigenvalue closest to the shift point λ_0. This procedure is referred to as inverse iteration with a shift. The matrix which must be factored is $\mathbf{K} - \lambda_0\mathbf{M}$. A practical application of inverse iteration with a shift would be the mechanical vibration problem in which one seeks the vibration modes closest to a particular frequency.

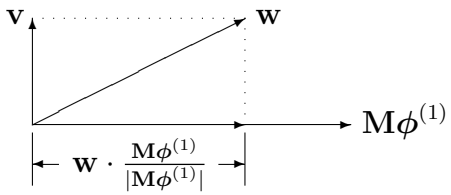

Figure 21: Geometrical Interpretation of Sweeping.

6.6 Iteration for Other Eigenvalues

We saw from the preceding sections that power iteration converges to the largest eigenvalue, and inverse iteration converges to the smallest eigenvalue. It is often of interest to find other eigenvalues as well. Here we describe the modification to the power iteration algorithm to allow convergence to the second largest eigenvalue.

Consider again the eigenvalue problem

$$\mathbf{A}\mathbf{u} = \lambda\mathbf{u}, \tag{6.109}$$

where $\mathbf{A} = \mathbf{M}^{-1}\mathbf{K}$, and assume we have already found the largest eigenvalue and its corresponding eigenvector $\boldsymbol{\phi}^{(1)}$. Let $\boldsymbol{\phi}^{(2)}$ denote the eigenvector corresponding to the second largest eigenvalue. Thus,

$$\mathbf{A}\boldsymbol{\phi}^{(2)} = \lambda\boldsymbol{\phi}^{(2)}, \tag{6.110}$$

where, by orthogonality,

$$\boldsymbol{\phi}^{(2)T}\mathbf{M}\boldsymbol{\phi}^{(1)} = 0 \quad \text{or} \quad \boldsymbol{\phi}^{(2)} \cdot \mathbf{M}\boldsymbol{\phi}^{(1)} = 0. \tag{6.111}$$

(The orthogonality relation has been written in both matrix and vector notations.)

To find $\boldsymbol{\phi}^{(2)}$, given $\boldsymbol{\phi}^{(1)}$, we will perform another power iteration but with the added requirement that, at each iteration, the vector used in the iteration is orthogonal to $\boldsymbol{\phi}^{(1)}$. For example, consider the vector $\mathbf{w}$ as a trial second eigenvector. Fig. 21 shows the plane of the two vectors $\mathbf{w}$ and $\mathbf{M}\boldsymbol{\phi}^{(1)}$. Since a unit vector in the direction of $\mathbf{M}\boldsymbol{\phi}^{(1)}$ is

$$\frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|},$$

the length of the projection of $\mathbf{w}$ onto $\mathbf{M}\boldsymbol{\phi}^{(1)}$ is

$$\mathbf{w} \cdot \frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|},$$

and the vector projection of $\mathbf{w}$ onto $\mathbf{M}\boldsymbol{\phi}^{(1)}$ is

$$\left(\mathbf{w} \cdot \frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|}\right) \frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|}.$$

Thus, a new trial vector $\mathbf{v}$ which is orthogonal to $\mathbf{M}\boldsymbol{\phi}^{(1)}$ is

$$\hat{\mathbf{v}} = \mathbf{w} - \left(\mathbf{w} \cdot \frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|}\right) \frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|}. \tag{6.112}$$

This procedure of subtracting from each trial vector (at each iteration) any components parallel to $\mathbf{M}\boldsymbol{\phi}^{(1)}$ is sometimes referred to as a *sweeping procedure*. This procedure is also equivalent to the Gram-Schmidt orthogonalization procedure discussed previously, except that here we do not require that $\mathbf{v}$ be a unit vector.

Note that, even though $\mathbf{A}$ is used in the iteration, we must use either $\mathbf{M}$ or $\mathbf{K}$ in the orthogonality relation rather than $\mathbf{A}$ or the identity $\mathbf{I}$, since, in general, $\mathbf{A}$ is not symmetric (e.g., Eq. 6.90). With a nonsymmetric $\mathbf{A}$, there is no orthogonality of the eigenvectors with respect to $\mathbf{A}$ or $\mathbf{I}$. If $\mathbf{A}$ happens to be symmetric (a special case of the symmetric generalized eigenvalue problem), then the eigenvectors would be orthogonal with respect to $\mathbf{A}$ and $\mathbf{I}$. Thus, even though we may iterate using $\mathbf{A}$, the orthogonality requires $\mathbf{M}$ or $\mathbf{K}$ weighting. $\mathbf{M}$ is chosen, since it usually contains more zeros than $\mathbf{K}$ and, for the example to follow, is diagonal.

To illustrate the sweeping procedure, we continue the mechanical vibrations example started earlier. Recall that, from power iteration,

$$\mathbf{A} = \mathbf{M}^{-1}\mathbf{K} = \begin{bmatrix} 5 & -2 & 0 \\ -1 & \frac{3}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}, \quad \mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \boldsymbol{\phi}^{(1)} = \begin{Bmatrix} 1.0000 \\ -0.2518 \\ 0.0162 \end{Bmatrix}, \tag{6.113}$$

where

$$\mathbf{M}\boldsymbol{\phi}^{(1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{Bmatrix} 1.0000 \\ -0.2518 \\ 0.0162 \end{Bmatrix} = \begin{Bmatrix} 1.0000 \\ -0.5036 \\ 0.0486 \end{Bmatrix}, \tag{6.114}$$

a vector whose length is $|\mathbf{M}\boldsymbol{\phi}^{(1)}| = 1.1207$. Thus, the unit vector in the direction of $\mathbf{M}\boldsymbol{\phi}^{(1)}$ is

$$\frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{|\mathbf{M}\boldsymbol{\phi}^{(1)}|} = \begin{Bmatrix} 0.8923 \\ -0.4494 \\ 0.0434 \end{Bmatrix}. \tag{6.115}$$

The iterations are most conveniently displayed with a spreadsheet:

i	$\mathbf{w}^{(i)}$	$\alpha = \frac{\mathbf{w}^{(i)} \cdot \mathbf{M}\boldsymbol{\phi}^{(1)}}{\lvert\mathbf{M}\boldsymbol{\phi}^{(1)}\rvert}$	$\mathbf{w}^{(i)} - \alpha \frac{\mathbf{M}\boldsymbol{\phi}^{(1)}}{\lvert\mathbf{M}\boldsymbol{\phi}^{(1)}\rvert}$	λ_i	$\mathbf{v}^{(i)}$	$\mathbf{A}\mathbf{v}^{(i)} = \mathbf{w}^{(i+1)}$
0	$\begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}$	0.8923	$\begin{Bmatrix} 0.2038 \\ 0.4010 \\ -0.0387 \end{Bmatrix}$	0.4010	$\begin{Bmatrix} 0.5082 \\ 1.0000 \\ -0.0965 \end{Bmatrix}$	$\begin{Bmatrix} 0.5410 \\ 1.0401 \\ -0.3655 \end{Bmatrix}$
1	$\begin{Bmatrix} 0.5410 \\ 1.0401 \\ -0.3655 \end{Bmatrix}$	−0.0005	$\begin{Bmatrix} 0.5414 \\ 1.0399 \\ -0.3655 \end{Bmatrix}$	1.0399	$\begin{Bmatrix} 0.5206 \\ 1.0000 \\ -0.3515 \end{Bmatrix}$	$\begin{Bmatrix} 0.6030 \\ 1.1552 \\ -0.4505 \end{Bmatrix}$
2	$\begin{Bmatrix} 0.6030 \\ 1.1552 \\ -0.4505 \end{Bmatrix}$	−0.0006	$\begin{Bmatrix} 0.6035 \\ 1.1549 \\ -0.4505 \end{Bmatrix}$	1.1549	$\begin{Bmatrix} 0.5226 \\ 1.0000 \\ -0.3901 \end{Bmatrix}$	$\begin{Bmatrix} 0.6130 \\ 1.1725 \\ -0.4634 \end{Bmatrix}$
3	$\begin{Bmatrix} 0.6130 \\ 1.1725 \\ -0.4634 \end{Bmatrix}$	−0.0001	$\begin{Bmatrix} 0.6131 \\ 1.1725 \\ -0.4634 \end{Bmatrix}$	1.1725	$\begin{Bmatrix} 0.5229 \\ 1.0000 \\ -0.3952 \end{Bmatrix}$	$\begin{Bmatrix} 0.6145 \\ 1.1747 \\ -0.4651 \end{Bmatrix}$
4	$\begin{Bmatrix} 0.6145 \\ 1.1747 \\ -0.4651 \end{Bmatrix}$	0.0002	$\begin{Bmatrix} 0.6143 \\ 1.1748 \\ -0.4651 \end{Bmatrix}$	1.1748	$\begin{Bmatrix} 0.5229 \\ 1.0000 \\ -0.3959 \end{Bmatrix}$	$\begin{Bmatrix} 0.6145 \\ 1.1751 \\ -0.4653 \end{Bmatrix}$
5	$\begin{Bmatrix} 0.6145 \\ 1.1751 \\ -0.4653 \end{Bmatrix}$	0.0000	$\begin{Bmatrix} 0.6145 \\ 1.1751 \\ -0.4653 \end{Bmatrix}$	1.1751	$\begin{Bmatrix} 0.5229 \\ 1.0000 \\ -0.3960 \end{Bmatrix}$	$\begin{Bmatrix} 0.6145 \\ 1.1751 \\ -0.4653 \end{Bmatrix}$
6	$\begin{Bmatrix} 0.6145 \\ 1.1751 \\ -0.4653 \end{Bmatrix}$	0.0000	$\begin{Bmatrix} 0.6145 \\ 1.1751 \\ -0.4653 \end{Bmatrix}$	1.1751	$\begin{Bmatrix} 0.5229 \\ 1.0000 \\ -0.3960 \end{Bmatrix}$	

Thus, the second largest eigenvalue of $\mathbf{A}$ is 1.1751, and the corresponding eigenvector is

$$\begin{Bmatrix} 0.5229 \\ 1.0000 \\ -0.3960 \end{Bmatrix}.$$

6.7 Similarity Transformations

Two matrices $\mathbf{A}$ and $\mathbf{B}$ are defined as *similar* if $\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$ for some invertible matrix $\mathbf{M}$. The transformation from $\mathbf{A}$ to $\mathbf{B}$ (or *vice versa*) is called a *similarity transformation.*

The main property of interest is that similar matrices have the same eigenvalues. To prove this property, consider the eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{6.116}$$

and let

$$\mathbf{B} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M} \quad \text{or} \quad \mathbf{A} = \mathbf{M}\mathbf{B}\mathbf{M}^{-1}. \tag{6.117}$$

The substitution of Eq. 6.117 into Eq. 6.116 yields

$$\mathbf{M}\mathbf{B}\mathbf{M}^{-1}\mathbf{x} = \lambda\mathbf{x} \tag{6.118}$$

or

$$\mathbf{B}(\mathbf{M}^{-1}\mathbf{x}) = \lambda(\mathbf{M}^{-1}\mathbf{x}), \tag{6.119}$$

which completes the proof. This equation also shows that, if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, $\mathbf{M}^{-1}\mathbf{x}$ is an eigenvector of $\mathbf{B}$, and λ is the eigenvalue.

A similarity transformation can be interpreted as a change of basis. Consider, for example, mechanical vibrations, where the eigenvalue problem computes the natural frequencies and mode shapes of vibration. If the eigenvalues (natural frequencies) do not change under the transformation, the system represented by $\mathbf{A}$ has not changed. However, the eigenvectors are transformed from $\mathbf{x}$ to $\mathbf{M}^{-1}\mathbf{x}$, a transformation which can be thought of as simply a change of coordinates.

We recall from Property 6 of the eigenvalue problem (page 69) that, if a matrix $\mathbf{M}$ is formed with the eigenvectors of $\mathbf{A}$ in the columns, and the eigenvectors are independent,

$$\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}, \tag{6.120}$$

where $\mathbf{\Lambda}$ is a diagonal matrix populated with the eigenvalues of $\mathbf{A}$. Thus, the way to "simplify" a matrix (by diagonalizing it) is to find its eigenvectors. The eigenvalue problem is thus the basis for finding principal stresses and principal strains in elasticity and principal axes of inertia.

6.8 Positive Definite Matrices

For a square matrix $\mathbf{M}$, the scalar $Q = \mathbf{x}^T\mathbf{M}\mathbf{x} = \mathbf{x}\cdot\mathbf{M}\mathbf{x}$ is called a *quadratic form*. In index notation, the quadratic form is

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}x_i x_j \tag{6.121}$$

$$= M_{11}x_1^2 + M_{22}x_2^2 + \cdots + (M_{12}+M_{21})x_1x_2 + (M_{13}+M_{31})x_1x_3 + \cdots . \tag{6.122}$$

A matrix $\mathbf{S}$ is said to be *symmetric* if $\mathbf{S} = \mathbf{S}^T$ (i.e., $S_{ij} = S_{ji}$). A matrix $\mathbf{A}$ is said to be *antisymmetric* (or *skew symmetric*) if $\mathbf{A} = -\mathbf{A}^T$ (i.e., $A_{ij} = -A_{ji}$). For example, an antisymmetric matrix of order 3 is necessarily of the form

$$\mathbf{A} = \begin{bmatrix} 0 & A_{12} & A_{13} \\ -A_{12} & 0 & A_{23} \\ -A_{13} & -A_{23} & 0 \end{bmatrix}. \tag{6.123}$$

Any square matrix $\mathbf{M}$ can be written as the unique sum of a symmetric and an antisymmetric matrix $\mathbf{M} = \mathbf{S} + \mathbf{A}$, where

$$\mathbf{S} = \mathbf{S}^T = \frac{1}{2}(\mathbf{M} + \mathbf{M}^T), \quad \mathbf{A} = -\mathbf{A}^T = \frac{1}{2}(\mathbf{M} - \mathbf{M}^T). \tag{6.124}$$

For example, for a 3×3 matrix $\mathbf{M}$,

$$\mathbf{M} = \begin{bmatrix} 3 & 5 & 7 \\ 1 & 2 & 8 \\ 9 & 6 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 3 & 8 \\ 3 & 2 & 7 \\ 8 & 7 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 2 & -1 \\ -2 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} = \mathbf{S} + \mathbf{A}. \tag{6.125}$$

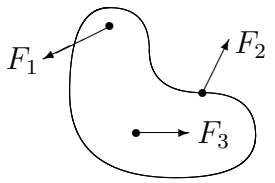

Figure 22: Elastic System Acted Upon by Forces.

The antisymmetric part of a matrix does not contribute to the quadratic form, i.e., if $\mathbf{A}$ is antisymmetric, $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ for all vectors $\mathbf{x}$. To prove this property, we note that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = (\mathbf{x}^T\mathbf{A}\mathbf{x})^T = \mathbf{x}^T\mathbf{A}^T\mathbf{x} = -\mathbf{x}^T\mathbf{A}\mathbf{x} = 0, \tag{6.126}$$

since the scalar is equal to its own negative, thus completing the proof. Thus, for any square matrix $\mathbf{M}$,

$$\mathbf{x}^T\mathbf{M}\mathbf{x} = \mathbf{x}^T\mathbf{S}\mathbf{x}, \tag{6.127}$$

where $\mathbf{S}$ is the symmetric part of $\mathbf{M}$.

A square matrix $\mathbf{M}$ is defined as *positive definite* if $\mathbf{x}^T\mathbf{M}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. $\mathbf{M}$ is defined as *positive semi-definite* if $\mathbf{x}^T\mathbf{M}\mathbf{x} \geq 0$ for all $\mathbf{x} \neq \mathbf{0}$.

Quadratic forms are of interest in engineering applications, since, physically, a quadratic form often corresponds to energy. For example, consider an elastic system (Fig. 22) acted upon by forces in static equilibrium. Let $\mathbf{u}$ denote the vector of displacements, and let $\mathbf{F}$ denote the vector of forces. If the forces and displacements are linearly related by generalized Hooke's law,

$$\mathbf{F} = \mathbf{K}\mathbf{u}, \tag{6.128}$$

where $\mathbf{K}$ is the system stiffness matrix.

The work W done by the forces $\mathbf{F}$ acting through the displacements $\mathbf{u}$ is then

$$W = \frac{1}{2}\mathbf{u}\cdot\mathbf{F} = \frac{1}{2}\mathbf{u}\cdot\mathbf{K}\mathbf{u} = \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u}, \tag{6.129}$$

which is a quadratic form. Since the work required to deform a mechanical system in equilibrium (and with sufficient constraints to prevent rigid body motion) is positive for any set of nonzero displacements,

$$W = \frac{1}{2}\mathbf{u}\cdot\mathbf{K}\mathbf{u} > 0. \tag{6.130}$$

That is, the stiffness matrix $\mathbf{K}$ for a mechanical system in equilibrium and constrained to prevent rigid body motion must be positive definite.

The main property of interest for positive definite matrices is that a matrix $\mathbf{M}$ is positive definite if, and only if, all the eigenvalues of $\mathbf{M}$ are positive. To prove this statement, we must prove two properties: (1) positive definiteness implies positive eigenvalues, and (2) positive eigenvalues imply positive definiteness.

We first prove that positive definiteness implies positive eigenvalues. For the ith eigenvalue λ_i,

$$\mathbf{M}\mathbf{x}^{(i)} = \lambda_i\mathbf{x}^{(i)}, \tag{6.131}$$

where $\mathbf{x}^{(i)}$ is the corresponding eigenvector. If we form the dot product of both sides of this equation with $\mathbf{x}^{(i)}$, we obtain

$$0 < \mathbf{x}^{(i)} \cdot \mathbf{M}\mathbf{x}^{(i)} = \lambda_i \mathbf{x}^{(i)} \cdot \mathbf{x}^{(i)} = \lambda_i \left|\mathbf{x}^{(i)}\right|^2, \tag{6.132}$$

which implies $\lambda_i > 0$, thus completing the first part of the proof.

We now prove that positive eigenvalues imply positive definiteness. Since, for a quadratic form, only the symmetric part of $\mathbf{M}$ matters, we can assume that $\mathbf{M}$ is symmetric. From Property 9 of the eigenvalue problem (page 71), the eigenvectors of $\mathbf{M}$ are mutually orthogonal, which implies that the eigenvectors are independent. We can therefore expand any vector $\mathbf{x}$ using an eigenvector basis:

$$\mathbf{x} = \sum_{i=1}^{n} c_i \mathbf{x}^{(i)}, \tag{6.133}$$

where $\mathbf{x}^{(i)}$ is the ith eigenvector. Then,

$$\mathbf{M}\mathbf{x} = \sum_{i=1}^{n} c_i \mathbf{M}\mathbf{x}^{(i)} = \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}^{(i)}, \tag{6.134}$$

which implies that

$$\mathbf{x} \cdot \mathbf{M}\mathbf{x} = \sum_{j=1}^{n} c_j \mathbf{x}^{(j)} \cdot \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}^{(i)} = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \lambda_i \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}. \tag{6.135}$$

However, since the eigenvectors are mutually orthogonal,

$$\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} = 0 \quad \text{for} \quad i \neq j. \tag{6.136}$$

Therefore,

$$\mathbf{x} \cdot \mathbf{M}\mathbf{x} = \sum_{i=1}^{n} c_i^2 \lambda_i \left|\mathbf{x}^{(i)}\right|^2, \tag{6.137}$$

which is positive if all $\lambda_i > 0$, thus completing the second part of the proof. This important property establishes the connection between positive definiteness and eigenvalues.

Two other consequences of positive-definiteness are

1. A positive definite matrix is non-singular, since $\mathbf{A}\mathbf{x} = \mathbf{0}$ implies $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ implies $\mathbf{x} = \mathbf{0}$ implies $\mathbf{A}^{-1}$ exists.

2. Gaussian elimination can be performed without pivoting.

6.9 Application to Differential Equations

Consider the ordinary differential equation

$$y' = ay, \tag{6.138}$$

where $y(t)$ is an unknown function to be determined, $y'(t)$ is its derivative, and a is a constant. All solutions of this equation are of the form

$$y(t) = y(0)e^{at}, \tag{6.139}$$

where $y(0)$ is the initial condition.

The generalization of Eq. 6.138 to systems of ordinary differential equations is

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \\ y_2' &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \\ &\vdots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \cdots + a_{nn}y_n, \end{aligned} \tag{6.140}$$

where $y_1(t)$, $y_2(t)$, ..., $y_n(t)$ are the unknown functions to be determined, and the a_{ij} are constants. In matrix notation, this system can be written

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \tag{6.141}$$

where

$$\mathbf{y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}. \tag{6.142}$$

Our principal interest in this section is in solving systems of coupled differential equations like Eq. 6.141. Note that, if the matrix $\mathbf{A}$ is a diagonal matrix, Eq. 6.141 is easy to solve, since each equation in the system involves only one unknown function (i.e., the system is uncoupled). Thus, to solve the system $\mathbf{y}' = \mathbf{A}\mathbf{y}$ for which $\mathbf{A}$ is not diagonal, we will attempt a change of variable

$$\mathbf{y} = \mathbf{S}\mathbf{u}, \tag{6.143}$$

where $\mathbf{S}$ is a square matrix of constants picked to yield a new system with a diagonal coefficient matrix. The substitution of Eq. 6.143 into Eq. 6.141 yields

$$\mathbf{S}\mathbf{u}' = \mathbf{A}\mathbf{S}\mathbf{u} \tag{6.144}$$

or

$$\mathbf{u}' = \left(\mathbf{S}^{-1}\mathbf{A}\mathbf{S}\right)\mathbf{u} = \mathbf{D}\mathbf{u}, \tag{6.145}$$

where

$$\mathbf{D} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \tag{6.146}$$

and we have assumed that $\mathbf{S}$ is invertible. The choice for $\mathbf{S}$ is now clear: $\mathbf{S}$ is the matrix which diagonalizes $\mathbf{A}$. From Property 6 of the eigenvalue problem (page 69), we see that $\mathbf{S}$ is the matrix whose columns are the linearly independent eigenvectors of $\mathbf{A}$, and $\mathbf{D}$ is the diagonal matrix of eigenvalues:

$$\mathbf{D} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_n \end{bmatrix} = \mathbf{\Lambda}, \tag{6.147}$$

where λ_i is the ith eigenvalue of $\mathbf{A}$.

Thus, we can use the following procedure for solving the coupled system of differential equations $\mathbf{y}' = \mathbf{A}\mathbf{y}$:

1. Find the eigenvalues and eigenvectors of $\mathbf{A}$.
2. Solve the uncoupled (diagonal) system $\mathbf{u}' = \mathbf{\Lambda}\mathbf{u}$, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{A}$.
3. Determine $\mathbf{y}$ from the equation $\mathbf{y} = \mathbf{S}\mathbf{u}$, where $\mathbf{S}$ is the matrix whose columns are the eigenvectors of $\mathbf{A}$.

We illustrate this procedure by solving the system

$$\begin{aligned} y_1' &= y_1 + y_2 \\ y_2' &= 4y_1 - 2y_2 \end{aligned} \tag{6.148}$$

with the initial conditions $y_1(0) = 1$, $y_2(0) = 6$. The coefficient matrix for this system is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 4 & -2 \end{bmatrix}. \tag{6.149}$$

Since the characteristic polynomial is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 1 \\ 4 & -2-\lambda \end{vmatrix} = \lambda^2 + \lambda - 6 = (\lambda+3)(\lambda-2), \tag{6.150}$$

the eigenvalues of $\mathbf{A}$ are $\lambda = 2$ and $\lambda = -3$. The corresponding eigenvectors are $(1, 1)$ and $(1, -4)$, respectively. Thus, a diagonalizing matrix for $\mathbf{A}$ is

$$\mathbf{S} = \begin{bmatrix} 1 & 1 \\ 1 & -4 \end{bmatrix}, \tag{6.151}$$

and

$$\mathbf{\Lambda} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \frac{1}{5}\begin{bmatrix} 4 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 4 & -2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -4 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix}. \tag{6.152}$$

Therefore, the substitution $\mathbf{y} = \mathbf{S}\mathbf{u}$ yields the new diagonal system

$$\mathbf{u}' = \mathbf{\Lambda}\mathbf{u} = \begin{bmatrix} 2 & 0 \\ 0 & -3 \end{bmatrix}\mathbf{u} \tag{6.153}$$

or

$$u_1' = 2u_1, \quad u_2' = -3u_2, \tag{6.154}$$

whose general solutions are

$$u_1 = c_1 e^{2t}, \quad u_2 = c_2 e^{-3t}. \tag{6.155}$$

Thus, $\mathbf{y}$, the original dependent variable of interest, is given by

$$\mathbf{y} = \mathbf{S}\mathbf{u} = \begin{bmatrix} 1 & 1 \\ 1 & -4 \end{bmatrix}\begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = \begin{Bmatrix} c_1 e^{2t} + c_2 e^{-3t} \\ c_1 e^{2t} - 4c_2 e^{-3t} \end{Bmatrix} \tag{6.156}$$

or

$$\begin{cases} y_1 = c_1 e^{2t} + c_2 e^{-3t} \\ y_2 = c_1 e^{2t} - 4c_2 e^{-3t}. \end{cases} \tag{6.157}$$

The application of the initial conditions yields

$$\begin{aligned} c_1 + c_2 &= 1 \\ c_1 - 4c_2 &= 6 \end{aligned} \tag{6.158}$$

or $c_1 = 2$, $c_2 = -1$. Thus, the solution of the original system is

$$\begin{cases} y_1 = 2e^{2t} - e^{-3t} \\ y_2 = 2e^{2t} + 4e^{-3t}. \end{cases} \tag{6.159}$$

Notice from Eq. 6.156 that the solution can be written in the form

$$\mathbf{y} = [\mathbf{v}_1 \;\; \mathbf{v}_2] \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} = u_1 \mathbf{v}_1 + u_2 \mathbf{v}_2 = c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2, \tag{6.160}$$

where the eigenvectors of $\mathbf{A}$ are the columns of $\mathbf{S}$:

$$\mathbf{v}_1 = \begin{Bmatrix} 1 \\ 1 \end{Bmatrix}, \quad \mathbf{v}_2 = \begin{Bmatrix} 1 \\ -4 \end{Bmatrix}. \tag{6.161}$$

Thus, in general, if the coefficient matrix $\mathbf{A}$ of the system $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is diagonalizable, the general solution can be written in the form

$$\mathbf{y} = c_1 e^{\lambda_1 t} \mathbf{v}_1 + c_2 e^{\lambda_2 t} \mathbf{v}_2 + \cdots + c_n e^{\lambda_n t} \mathbf{v}_n, \tag{6.162}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$, and $\mathbf{v}_i$ is the eigenvector corresponding to λ_i.

A comparison of the matrix differential equation, Eq. 6.141, with the scalar equation, Eq. 6.138, indicates that the solution of Eq. 6.141 can be written in the form

$$\mathbf{y}(t) = e^{\mathbf{A}t} \mathbf{y}(0), \tag{6.163}$$

where the exponential of the matrix $\mathbf{A}t$ is formally defined [8] by the convergent power series

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \frac{(\mathbf{A}t)^2}{2!} + \frac{(\mathbf{A}t)^3}{3!} + \cdots . \tag{6.164}$$

From Eq. 6.147,

$$\mathbf{A} = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}, \tag{6.165}$$

where $\mathbf{S}$ is a matrix whose columns are the eigenvectors of $\mathbf{A}$. Notice that, for integers k,

$$\mathbf{A}^k = (\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1})^k = (\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1})(\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}) \cdots (\mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}) = \mathbf{S}\mathbf{\Lambda}^k\mathbf{S}^{-1}, \tag{6.166}$$

since $\mathbf{S}^{-1}$ cancels $\mathbf{S}$. Thus,

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{S}\mathbf{\Lambda}\mathbf{S}^{-1}t + \frac{\mathbf{S}\mathbf{\Lambda}^2\mathbf{S}^{-1}t^2}{2!} + \frac{\mathbf{S}\mathbf{\Lambda}^3\mathbf{S}^{-1}t^3}{3!} + \cdots \tag{6.167}$$

$$= \mathbf{S}\left[\mathbf{I} + \mathbf{\Lambda}t + \frac{(\mathbf{\Lambda}t)^2}{2!} + \frac{(\mathbf{\Lambda}t)^3}{3!} + \cdots\right]\mathbf{S}^{-1} = \mathbf{S}e^{\mathbf{\Lambda}t}\mathbf{S}^{-1}. \tag{6.168}$$

Since $\mathbf{\Lambda}t$ is a diagonal matrix, the powers $(\mathbf{\Lambda}t)^k$ are also diagonal, and $e^{\mathbf{\Lambda}t}$ is therefore the diagonal matrix

$$e^{\mathbf{\Lambda}t} = \begin{bmatrix} e^{\lambda_1 t} & & & \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ & & & e^{\lambda_n t} \end{bmatrix}. \tag{6.169}$$

Thus, the solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ can be written in the form

$$\mathbf{y}(t) = \mathbf{S}e^{\mathbf{\Lambda}t}\mathbf{S}^{-1}\mathbf{y}(0), \tag{6.170}$$

where $\mathbf{S}$ is a matrix whose columns are the eigenvectors of $\mathbf{A}$, $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues, and the exponential is given by Eq. 6.169.

For the example of Eq. 6.148, Eq. 6.170 yields

$$\begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} = \frac{1}{5}\begin{bmatrix} 1 & 1 \\ 1 & -4 \end{bmatrix}\begin{bmatrix} e^{2t} & 0 \\ 0 & e^{-3t} \end{bmatrix}\begin{bmatrix} 4 & 1 \\ 1 & -1 \end{bmatrix}\begin{Bmatrix} 1 \\ 6 \end{Bmatrix} = \begin{Bmatrix} 2e^{2t} - e^{-3t} \\ 2e^{2t} + 4e^{-3t} \end{Bmatrix}, \tag{6.171}$$

in agreement with Eq. 6.159.

Note that this discussion of first-order systems includes as a special case the nth-order equation with constant coefficients, since a single nth-order ODE is equivalent to a system of n first-order ODEs. For example, the n-th order system

$$a_n y^{(n)} + a_{n-1}y^{(n-1)} + a_{n-2}y^{(n-2)} + \cdots + a_1 y' + a_0 y = 0, \quad a_n \neq 0, \tag{6.172}$$

can be replaced by a first-order system if we rename the derivatives as $y^{(i)} = y_{i+1}(t)$, in which case

$$\begin{cases} y_1' = y_2 \\ y_2' = y_3 \\ \quad \vdots \\ y_{n-1}' = y_n \\ y_n' = (-a_0 y_1 - a_1 y_2 - \cdots - a_{n-2}y_{n-1} - a_{n-1}y_n)/a_n. \end{cases} \tag{6.173}$$

6.10 Application to Structural Dynamics

We saw in the discussion of mechanical vibrations (§6.1, p. 65) that the equation of motion for a linear, multi-DOF, undamped mechanical system is

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}(t), \tag{6.174}$$

where $\mathbf{M}$ is the system mass matrix, $\mathbf{K}$ is the system stiffness matrix, $\mathbf{F}$ is the time-dependent applied force vector, $\mathbf{u}(t)$ is the unknown displacement vector, and dots denote differentiation with respect to the time t. If the system is damped, a viscous damping matrix $\mathbf{B}$ is introduced, and the equations of motion become

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{B}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{F}(t). \tag{6.175}$$

All three system matrices are real and symmetric. Equations like this arise whether the system is a discrete element system such as in Fig. 20 or a continuum modeled with finite elements. The general problem is to integrate these equations in time to determine $\mathbf{u}(t)$, given the initial displacement $\mathbf{u}_0$ and initial velocity $\dot{\mathbf{u}}_0$.

There are several approaches used to solve these equations, one of which is the Newmark method (Eq. 1.75) discussed on p. 15. Another approach expands the solution in terms of vibration modes (the modal superposition approach). We recall that, for the free, undamped vibration problem

$$\mathbf{K}\mathbf{u} = \lambda\mathbf{M}\mathbf{u}, \tag{6.176}$$

the eigenvectors are orthogonal with respect to both the mass $\mathbf{M}$ and stiffness $\mathbf{K}$ (Property 9, p. 71). Thus, the eigenvectors could be used to form a basis for the solution space, and the unknown displacement vector in Eq. 6.175 could be written as a linear combination of the modes:

$$\mathbf{u}(t) = \xi_1(t)\mathbf{x}^{(1)} + \xi_2(t)\mathbf{x}^{(2)} + \cdots + \xi_m(t)\mathbf{x}^{(m)}. \tag{6.177}$$

The number of terms in this expansion would be n (the number of physical DOF) if all eigenvectors were used in the expansion, but, in practice, a relatively small number m of modes is usually sufficient for acceptable accuracy. The modes selected are generally those having lowest frequency (eigenvalue). In this equation, the multipliers $\xi_i(t)$ are referred to as the modal coordinates, which can be collected into the array

$$\boldsymbol{\xi}(t) = \left\{ \begin{array}{c} \xi_1(t) \\ \xi_2(t) \\ \vdots \\ \xi_m(t) \end{array} \right\}. \tag{6.178}$$

Also, like in Property 6 (p. 69), we form the matrix $\mathbf{S}$ whose columns are the eigenvectors:

$$\mathbf{S} = \left[\begin{array}{cccc} \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \cdots & \mathbf{x}^{(m)} \end{array} \right], \tag{6.179}$$

in which case Eq. 6.177 can be written in matrix form as

$$\mathbf{u}(t) = \mathbf{S}\boldsymbol{\xi}(t). \tag{6.180}$$

This equation can be interpreted as a transformation from physical coordinates $\mathbf{u}$ to modal coordinates $\boldsymbol{\xi}$. If this equation is substituted into the equation of motion, Eq. 6.175, and multiplied by $\mathbf{S}^T$, we obtain

$$\mathbf{S}^T\mathbf{M}\mathbf{S}\ddot{\boldsymbol{\xi}} + \mathbf{S}^T\mathbf{B}\mathbf{S}\dot{\boldsymbol{\xi}} + \mathbf{S}^T\mathbf{K}\mathbf{S}\boldsymbol{\xi} = \mathbf{S}^T\mathbf{F}(t) = \mathbf{P}(t), \tag{6.181}$$

where $\widehat{\mathbf{M}} = \mathbf{S}^T\mathbf{M}\mathbf{S}$ and $\widehat{\mathbf{K}} = \mathbf{S}^T\mathbf{K}\mathbf{S}$ are, by orthogonality, diagonal matrices referred to as the modal mass and stiffness matrices, respectively. The diagonal entries in these two matrices are the generalized masses and stiffnesses for the individual vibration modes. The modal damping matrix $\widehat{\mathbf{B}} = \mathbf{S}^T\mathbf{B}\mathbf{S}$ is, in general, not a diagonal matrix, since the eigenvalue problem did not involve $\mathbf{B}$. Thus, in modal coordinates, the equations of motion are

$$\widehat{\mathbf{M}}\ddot{\boldsymbol{\xi}} + \widehat{\mathbf{B}}\dot{\boldsymbol{\xi}} + \widehat{\mathbf{K}}\boldsymbol{\xi} = \mathbf{P}(t). \tag{6.182}$$

This last equation is similar to Eq. 6.175 except that this equation has m DOF (the number of eigenvectors in the modal expansion), which generally is much smaller than the number n of physical DOF in Eq. 6.175. If $\widehat{\mathbf{B}}$ is non-diagonal, both equations require a similar time integration. The main benefit of the modal formulation is that, in structural dynamics, it is common to approximate damping with a mode-by-mode frequency-dependent viscous damping coefficient so that $\widehat{\mathbf{B}}$ is diagonal, in which case the modal equations (Eq. 6.182) uncouple into m scalar equations of the form

$$m_i\ddot{\xi}_i + b_i\dot{\xi}_i + k_i\xi_i = P_i(t), \tag{6.183}$$

where m_i and k_i are the generalized mass and stiffness, respectively, for mode i. This is the equation of motion for the ith modal coordinate. Since it is a scalar equation with an arbitrary right-hand side, it has a known analytic solution in terms of an integral. Once all the scalar equations have been solved, the solution $\mathbf{u}(t)$ in terms of physical coordinates is obtained from Eq. 6.177. Thus we see that a multi-DOF dynamical system can be viewed as a collection of single-DOF mass-spring-damper systems associated with the modes.

Most of the computational expense in the modal expansion method arises from solving the eigenvalue problem. Thus, the modal method is generally advantageous in structural dynamics if a small number of modes gives sufficient accuracy, if many time steps are required in the analysis, or if the structure has no localized damping treatments.

Bibliography

[1] J.W. Brown and R.V. Churchill. *Fourier Series and Boundary Value Problems*. McGraw-Hill, Inc., New York, seventh edition, 2006.

[2] R.W. Clough and J. Penzien. *Dynamics of Structures*. McGraw-Hill, Inc., New York, second edition, 1993.

[3] K.H. Huebner, D.L. Dewhirst, D.E. Smith, and T.G. Byrom. *The Finite Element Method for Engineers*. John Wiley and Sons, Inc., New York, fourth edition, 2001.

[4] D.C. Kay. *Schaum's Outline of Theory and Problems of Tensor Calculus*. Schaum's Outline Series. McGraw-Hill, Inc., New York, 1988.

[5] A.J. Laub. *Matrix Analysis for Scientists and Engineers*. Society for Industrial and Applied Mathematics, Philadelphia, 2005.

[6] S. Lipschutz. *3000 Solved Problems in Linear Algebra*. Schaum's Outline Series. McGraw-Hill, Inc., New York, 1988.

[7] S. Lipschutz. *Schaum's Outline of Theory and Problems of Linear Algebra*. Schaum's Outline Series. McGraw-Hill, Inc., New York, second edition, 1991.

[8] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

[9] M.R. Spiegel. *Schaum's Outline of Fourier Analysis With Applications to Boundary Value Problems*. Schaum's Outline Series. McGraw-Hill, Inc., New York, 1974.

[10] G. Strang. *Linear Algebra and Its Applications*. Thomson Brooks/Cole, Belmont, CA, fourth edition, 2006.

[11] J.S. Vandergraft. *Introduction to Numerical Calculations*. Academic Press, Inc., New York, second edition, 1983.

Index