Applied Mathematics

Peter J. Olver
School of Mathematics
University of Minnesota
Minneapolis, MN 55455
olver@math.umn.edu
http://www.math.umn.edu/∼olver

Chehrzad Shakiban
Department of Mathematics
University of St. Thomas
St. Paul, MN 55105-1096
c9shakiban@stthomas.edu
http://webcampus3.stthomas.edu/c9shakiban

Table of Contents

The Airy Equation
The Legendre Equation
Regular Singular Points
Bessel's Equation

Chapter 1

Linear Algebra

The source of linear algebra is the solution of systems of linear algebraic equations. Linear algebra is the foundation upon which almost all applied mathematics rests. This is not to say that nonlinear equations are less important; rather, progress in the vastly more complicated nonlinear realm is impossible without a firm grasp of the fundamentals of linear systems. Furthermore, linear algebra underlies the numerical analysis of continuous systems, both linear and nonlinear, which are typically modeled by differential equations. Without a systematic development of the subject from the start, we will be ill equipped to handle the resulting large systems of linear equations involving many (e.g., thousands of) unknowns.

This first chapter is devoted to the systematic development of direct[†] algorithms for solving systems of linear algegbraic equations in a finite number of variables. Our primary focus will be the most important situation involving the same number of equations as unknowns, although in Section 1.8 we extend our techniques to completely general linear systems. While the former usually have a unique solution, more general systems more typically have either no solutions, or infinitely many, and so tend to be of less direct physical relevance. Nevertheless, the ability to confidently handle all types of linear systems is a basic prerequisite for the subject.

The basic solution algorithm is known as *Gaussian elimination*, in honor of one of the all-time mathematical greats — the nineteenth century German mathematician Carl Friedrich Gauss. As the father of linear algebra, his name will occur repeatedly throughout this text. Gaussian elimination is quite elementary, but remains one of *the* most important techniques in applied (as well as theoretical) mathematics. Section 1.7 discusses some practical issues and limitations in computer implementations of the Gaussian elimination method for large systems arising in applications.

The systematic development of the subject relies on the fundamental concepts of scalar, vector, and matrix, and we quickly review the basics of matrix arithmetic. Gaussian elimination can be reinterpreted as matrix factorization, the (permuted) LU decomposition, which provides additional insight into the solution algorithm. Matrix inverses and determinants are discussed in Sections 1.5 and 1.9, respectively. However, both play a relatively minor role in practical applied mathematics, and so will not assume their more traditional central role in this applications-oriented text.

[†] Indirect algorithms, which are based on iteration, will be the subject of Chapter 10.

1.1. Solution of Linear Systems.

Gaussian elimination is a simple, systematic approach to the solution of systems of linear equations. It is the workhorse of linear algebra, and as such of absolutely fundamental importance in applied mathematics. In this section, we review the method in the most important case in which there are the same number of equations as unknowns. The general situation will be deferred until Section 1.8.

To illustrate, consider an elementary system of three linear equations

$$\begin{aligned} x + 2y + z &= 2, \\ 2x + 6y + z &= 7, \\ x + y + 4z &= 3, \end{aligned} \tag{1.1}$$

in three unknowns x, y, z. Linearity refers to the fact that the unknowns only appear to the first power[†] in the equations. The basic solution method is to systematically employ the following fundamental operation:

Linear System Operation #1: Add a multiple of one equation to another equation.

Before continuing, you should convince yourself that this operation does not change the solutions to the system. As a result, our goal is to judiciously apply the operation and so be led to a much simpler linear system that is easy to solve, and, moreover has the same solutions as the original. Any linear system that is derived from the original system by successive application of such operations will be called an *equivalent system*. By the preceding remark, *equivalent linear systems have the same solutions*.

The systematic feature is that we successively eliminate the variables in our equations in order of appearance. We begin by eliminating the first variable, x, from the second equation. To this end, we subtract twice the first equation from the second, leading to the equivalent system

$$\begin{aligned} x + 2y + z &= 2, \\ 2y - z &= 3, \\ x + y + 4z &= 3. \end{aligned} \tag{1.2}$$

Next, we eliminate x from the third equation by subtracting the first equation from it:

$$\begin{aligned} x + 2y + z &= 2, \\ 2y - z &= 3, \\ -y + 3z &= 1. \end{aligned} \tag{1.3}$$

The equivalent system (1.3) is already simpler than the original system (1.1). Notice that the second and third equations do not involve x (by design) and so constitute a system of two linear equations for two unknowns. Moreover, once we have solved this subsystem for y and z, we can substitute the answer into the first equation, and we need only solve a single linear equation for x.

[†] Also, there are no product terms like xy or xyz. The "official" definition of linearity will be deferred until Chapter 7.

We continue on in this fashion, the next phase being the elimination of the second variable y from the third equation by adding $\frac{1}{2}$ the second equation to it. The result is

$$\begin{aligned} x + 2\,y + z &= 2, \\ 2\,y - z &= 3, \\ \tfrac{5}{2}\,z &= \tfrac{5}{2}, \end{aligned} \tag{1.4}$$

which is the simple system we are after. It is in what is called *triangular form*, which means that, while the first equation involves all three variables, the second equation only involves the second and third variables, and the last equation only involves the last variable.

Any triangular system can be straightforwardly solved by the method of *Back Substitution*. As the name suggests, we work backwards, solving the last equation first, which requires $z = 1$. We substitute this result back into the next to last equation, which becomes $2\,y - 1 = 3$, with solution $y = 2$. We finally substitute these two values for y and z into the first equation, which becomes $x + 5 = 2$, and so the solution to the triangular system (1.4) is

$$x = -3, \qquad y = 2, \qquad z = 1. \tag{1.5}$$

Moreover, since we only used our basic operation to pass from (1.1) to the triangular system (1.4), this is also the solution to the original system of linear equations. We note that the system (1.1) has a unique — meaning one and only one — solution, namely (1.5).

And that, barring a few complications that can crop up from time to time, is all that there is to the method of Gaussian elimination! It is very simple, but its importance cannot be overemphasized. Before discussing the relevant issues, it will help to reformulate our method in a more convenient matrix notation.

1.2. Matrices and Vectors.

A *matrix* is a rectangular array of numbers. Thus,

$$\begin{pmatrix} 1 & 0 & 3 \\ -2 & 4 & 1 \end{pmatrix}, \qquad \begin{pmatrix} \pi & 0 \\ e & \frac{1}{2} \\ -1 & .83 \\ \sqrt{5} & -\frac{4}{7} \end{pmatrix}, \qquad \begin{pmatrix} .2 & -1.6 & .32 \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 3 \\ -2 & 5 \end{pmatrix},$$

are all examples of matrices. We use the notation

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \tag{1.6}$$

for a general matrix of size $m \times n$ (read "m by n"), where m denotes the number of *rows* in A and n denotes the number of *columns*. Thus, the preceding examples of matrices have respective sizes 2×3, 4×2, 1×3, 2×1 and 2×2. A matrix is *square* if $m = n$, i.e., it has the same number of rows as columns. A *column vector* is a $m \times 1$ matrix, while a *row*

vector is a $1 \times n$ matrix. As we shall see, column vectors are by far the more important of the two, and the term "vector" without qualification will always mean "column vector". A 1×1 matrix, which has but a single entry, is both a row and column vector.

The number that lies in the i^{th} row and the j^{th} column of A is called the (i, j) *entry* of A, and is denoted by a_{ij}. The row index always appears first and the column index second[†]. Two matrices are equal, $A = B$, if and only if they have the same size, and *all* their entries are the same: $a_{ij} = b_{ij}$.

A general linear system of m equations in n unknowns will take the form

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 + \;\cdots\; + a_{1n}\,x_n &= b_1, \\ a_{21}\,x_1 + a_{22}\,x_2 + \;\cdots\; + a_{2n}\,x_n &= b_2, \\ \vdots \qquad\qquad \vdots \qquad\qquad & \vdots \\ a_{m1}\,x_1 + a_{m2}\,x_2 + \;\cdots\; + a_{mn}\,x_n &= b_m. \end{aligned} \tag{1.7}$$

As such, it has three basic constituents: the $m \times n$ coefficient matrix A, with entries a_{ij} as in (1.6), the column vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ containing the unknowns, and the column vector $\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$ containing right hand sides. For instance, in our previous example (1.1), the coefficient matrix is $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}$, the vector of unknowns is $\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$, while $\mathbf{b} = \begin{pmatrix} 2 \\ 7 \\ 3 \end{pmatrix}$ contains the right hand sides.

Remark: We will consistently use bold face lower case letters to denote vectors, and ordinary capital letters to denote general matrices.

Matrix Arithmetic

There are three basic operations in matrix arithmetic: *matrix addition*, *scalar multiplication*, and *matrix multiplication*. First we define *addition* of matrices. You are only allowed to add two matrices of the *same size*, and matrix addition is performed entry by

[†] In tensor analysis, [**2**], a sub- and super-script notation is adopted, with a^i_j denoting the (i, j) entry of the matrix A. This has certain advantages, but, to avoid possible confusion with powers, we shall stick with the simpler subscript notation throughout this text.

entry. Therefore, if A and B are $m \times n$ matrices, their sum $C = A + B$ is the $m \times n$ matrix whose entries are given by $c_{ij} = a_{ij} + b_{ij}$ for $i = 1, \dots, m, j = 1, \dots, n$. For example,

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

When defined, matrix addition is commutative, $A + B = B + A$, and associative, $A + (B + C) = (A + B) + C$, just like ordinary addition.

A *scalar* is a fancy name for an ordinary number — the term merely distinguishes it from a vector or a matrix. For the time being, we will restrict our attention to real scalars and matrices with real entries, but eventually complex scalars and complex matrices must be dealt with. We will often identify a scalar $c \in \mathbb{R}$ with the 1×1 matrix (c) in which it is the sole entry. *Scalar multiplication* takes a scalar c and an $m \times n$ matrix A and computes the $m \times n$ matrix $B = c\,A$ by multiplying each entry of A by c. Thus, $b_{ij} = c\,a_{ij}$ for $i = 1, \dots, m, j = 1, \dots, n$. For example,

$$3 \begin{pmatrix} 1 & 2 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ -3 & 0 \end{pmatrix}.$$

Basic properties of scalar multiplication are summarized at the end of this section.

Finally, we define *matrix multiplication*. First, the product between a row vector $\mathbf{a}$ and a column vector $\mathbf{x}$ having the *same* number of entries is the *scalar* defined by the following rule:

$$\mathbf{a}\,\mathbf{x} = (\, a_1 \;\, a_2 \;\, \dots \;\, a_n \,) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a_1 x_1 + a_2 x_2 + \, \cdots \, + a_n x_n = \sum_{k=1}^{n} a_k x_k. \tag{1.8}$$

More generally, if A is an $m \times n$ matrix and B is an $n \times p$ matrix, so that the number of columns in A equals the number of rows in B, then the matrix product $C = A\,B$ is defined as the $m \times p$ matrix whose (i, j) entry equals the vector product of the i^{th} row of A and the j^{th} column of B. Therefore,

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}. \tag{1.9}$$

Note that our restriction on the sizes of A and B guarantees that the relevant row and column vectors will have the same number of entries, and so their product is defined.

For example, the product of the coefficient matrix A and vector of unknowns $\mathbf{x}$ for our original system (1.1) is given by

$$A\,\mathbf{x} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y + z \\ 2x + 6y + z \\ x + y + 4z \end{pmatrix}.$$

The result is a column vector whose entries reproduce the left hand sides of the original linear system! As a result, we can rewrite the system

$$A\,\mathbf{x} = \mathbf{b} \tag{1.10}$$

as an equality between two column vectors. This result is general; a linear system (1.7) consisting of m equations in n unknowns can be written in the matrix form (1.10) where A is the $m \times n$ coefficient matrix (1.6), $\mathbf{x}$ is the $n \times 1$ column vectors of unknowns, and $\mathbf{b}$ is the $m \times 1$ column vector containing the right hand sides. This is the reason behind the non-evident definition of matrix multiplication. Component-wise multiplication of matrix entries turns out to be almost completely useless in applications.

Now, the bad news. Matrix multiplication is *not* commutative. For example, BA may not be defined even when AB is. Even if both are defined, they may be different sized matrices. For example the product of a row vector $\mathbf{r}$, a $1 \times n$ matrix, and a column vector $\mathbf{c}$, an $n \times 1$ matrix, is a 1×1 matrix or scalar $s = \mathbf{r}\,\mathbf{c}$, whereas the reversed product $C = \mathbf{c}\,\mathbf{r}$ is an $n \times n$ matrix. For example,

$$\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = 3, \qquad \text{whereas} \qquad \begin{pmatrix} 3 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 0 & 0 \end{pmatrix}.$$

In computing the latter product, don't forget that we multiply the *rows* of the first matrix by the *columns* of the second. Moreover, even if the matrix products AB and BA have the same size, which requires both A and B to be square matrices, we may still have $AB \neq BA$. For example,

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} -2 & 5 \\ -4 & 11 \end{pmatrix} \neq \begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

On the other hand, matrix multiplication is associative, so $A(BC) = (AB)C$ whenever A has size $m \times n$, B has size $n \times p$ and C has size $p \times q$; the result is a matrix of size $m \times q$. The proof of this fact is left to the reader. Consequently, the one significant difference between matrix algebra and ordinary algebra is that you need to be careful not to change the order of multiplicative factors without proper justification.

Since matrix multiplication multiplies rows times columns, one can compute the columns in a matrix product $C = AB$ by multiplying the matrix A by the individual columns of B. The k^{th} column of C is equal to the product of A with the k^{th} column of B. For example, the two columns of the matrix product

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} 3 & 4 \\ 0 & 2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ 8 & 6 \end{pmatrix}$$

are obtained by multiplying the first matrix with the individual columns of the second:

$$\begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \end{pmatrix}, \qquad \begin{pmatrix} 1 & -1 & 2 \\ 2 & 0 & -2 \end{pmatrix} \begin{pmatrix} 4 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 6 \end{pmatrix}.$$

In general, if we use $\mathbf{b}_j$ to denote the j^{th} column of B, then

$$A\,B = A\begin{pmatrix}\mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_p\end{pmatrix} = \begin{pmatrix}A\,\mathbf{b}_1 & A\,\mathbf{b}_2 & \dots & A\,\mathbf{b}_p\end{pmatrix}. \tag{1.11}$$

There are two important special matrices. The first is the *zero matrix* of size $m \times n$, denoted $\mathrm{O}_{m\times n}$ or just O if the size is clear from context. It forms the additive unit, so $A + \mathrm{O} = A = \mathrm{O} + A$ for any matrix A of the same size. The role of the multiplicative unit is played by the square *identity matrix*

$$\mathrm{I} = \mathrm{I}_n = \begin{pmatrix}1 & 0 & 0 & \dots & 0\\ 0 & 1 & 0 & \dots & 0\\ 0 & 0 & 1 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \dots & 1\end{pmatrix}$$

of size $n \times n$. The entries of I along the *main diagonal* (which runs from top left to bottom right) are equal to 1; the *off-diagonal* entries are all 0. As the reader can check, if A is any $m \times n$ matrix, then $\mathrm{I}_m\,A = A = A\,\mathrm{I}_n$. We will sometimes write the last equation as just $\mathrm{I}\,A = A = A\,\mathrm{I}$; even though the identity matrices can have different sizes, only one size is valid for each matrix product to be defined.

The identity matrix is a particular example of a diagonal matrix. In general, a matrix is *diagonal* if all its off-diagonal entries are zero: $a_{ij} = 0$ for all $i \neq j$. We will sometimes write $D = \mathrm{diag}\,(c_1, \dots, c_n)$ for the $n \times n$ diagonal matrix with diagonal entries $d_{ii} = c_i$. Thus, $\mathrm{diag}\,(1, 3, 0)$ refers to the diagonal matrix $\begin{pmatrix}1 & 0 & 0\\ 0 & 3 & 0\\ 0 & 0 & 0\end{pmatrix}$, while the $n \times n$ identity matrix can be written as $\mathrm{I}_n = \mathrm{diag}\,(1, 1, \dots, 1)$.

Let us conclude this section by summarizing the basic properties of matrix arithmetic. In the following table, A, B, C are matrices, c, d scalars, O is a zero matrix, and I is an identity matrix. The matrices are assumed to have the correct sizes so that the indicated operations are defined.

1.3. Gaussian Elimination — Regular Case.

With the basic matrix arithmetic operations in hand, let us now return to our primary task. The goal is to develop a systematic method for solving linear systems of equations. While we could continue to work directly with the equations, matrices provide a convenient alternative that begins by merely shortening the amount of writing, but ultimately leads to profound insight into the solution and its structure.

We begin by replacing the system (1.7) by its matrix constituents. It is convenient to ignore the vector of unknowns, and form the *augmented matrix*

$$M = \left(\begin{array}{c|c} A & \mathbf{b}\end{array}\right) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1\\ a_{21} & a_{22} & \dots & a_{2n} & b_2\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ a_{m1} & a_{m2} & \dots & a_{mn} & b_n\end{array}\right) \tag{1.12}$$

Basic Matrix Arithmetic

Commutativity — Matrix Addition	$A + B = B + A$
Associativity — Matrix Addition	$(A + B) + C = A + (B + C)$
Zero Matrix — Matrix Addition	$A + \mathrm{O} = A = \mathrm{O} + A$
Associativity — Scalar Multiplication	$c(dA) = (cd)A$
Additive Inverse	$A + (-A) = \mathrm{O}, \quad -A = (-1)A$
Unit — Scalar Multiplication	$1 \cdot A = A$
Zero — Scalar Multiplication	$0 \cdot A = \mathrm{O}$
Distributivity — Matrix Addition	$c\,(A + B) = (cA) + (cB)$
Distributivity — Scalar Addition	$(c + d)\,A = (cA) + (dA)$
Associativity — Matrix Multiplication	$(AB)C = A(BC)$
Identity Matrix	$A\,\mathrm{I} = A = \mathrm{I}\,A$
Zero Matrix — Matrix Multiplication	$A\,\mathrm{O} = \mathrm{O} = \mathrm{O}\,A$

which is an $m \times (n+1)$ matrix obtained by tacking the right hand side vector onto the original coefficient matrix. The extra vertical line is included just to remind us that the last column of this matrix is special. For example, the augmented matrix for the system (1.1), i.e.,

$$\begin{aligned} x + 2y + z &= 2, \\ 2x + 6y + z &= 7, \\ x + y + 4z &= 3, \end{aligned} \qquad \text{is} \qquad M = \left(\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 1 & 4 & 3 \end{array}\right). \tag{1.13}$$

Note that one can immediately recover the equations in the original linear system from the augmented matrix. Since operations on equations also affect their right hand sides, keeping track of everything is most easily done through the augmented matrix.

For the time being, we will concentrate our efforts on linear systems that have the same number, n, of equations as unknowns. The associated coefficient matrix A is square, of size $n \times n$. The corresponding augmented matrix $M = \left(A \mid \mathbf{b} \right)$ then has size $n \times (n+1)$.

The matrix operation that assumes the role of Linear System Operation #1 is:

Elementary Row Operation #1:
Add a scalar multiple of one row of the augmented matrix to another row.

For example, if we add -2 times the first row of the augmented matrix (1.13) to the second row, the result is the row vector

$$-2\,(1 \;\; 2 \;\; 1 \;\; 2) \,+\, (2 \;\; 6 \;\; 1 \;\; 7) = (0 \;\; 2 \;\; -1 \;\; 3).$$

The result can be recognized as the second row of the modified augmented matrix

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 1 & 1 & 4 & 3 \end{array}\right) \tag{1.14}$$

that corresponds to the first equivalent system (1.2). When elementary row operation #1 is performed, it is critical that the result replace the row being added to — *not* the row being multiplied by the scalar. Notice that the elimination of a variable in an equation — in this case, the first variable in the second equation — amounts to making its entry in the coefficient matrix equal to zero.

We shall call the $(1,1)$ entry of the coefficient matrix the *first pivot*. The precise definition of pivot will become clear as we continue; the one key requirement is that a pivot be *nonzero*. Eliminating the first variable x from the second and third equations amounts to making all the matrix entries in the column below the pivot equal to zero. We have already done this with the $(2,1)$ entry in (1.14). To make the $(3,1)$ entry equal to zero, we subtract the first row from the last row. The resulting augmented matrix is

$$\left(\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & -1 & 3 & 1 \end{array}\right),$$

which corresponds to the system (1.3). The *second pivot* is the $(2,2)$ entry of this matrix, which is 2, and is the coefficient of the second variable in the second equation. Again, the pivot must be nonzero. We use the elementary row operation of adding $\frac{1}{2}$ of the second row to the third row to make the entry below the second pivot equal to 0; the result is the augmented matrix

$$N = \left(\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -1 & 3 \\ 0 & 0 & \frac{5}{2} & \frac{5}{2} \end{array}\right).$$

that corresponds to the triangular system (1.4). We write the final augmented matrix as

$$N = \big(\, U \mid \mathbf{c} \,\big), \qquad \text{where} \qquad U = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 2 \\ 3 \\ \frac{5}{2} \end{pmatrix}.$$

The corresponding linear system has vector form

$$U\mathbf{x} = \mathbf{c}. \tag{1.15}$$

Its coefficient matrix U is *upper triangular*, which means that all its entries below the main diagonal are zero: $u_{ij} = 0$ whenever $i > j$. The three nonzero entries on its diagonal, $1, 2, \frac{5}{2}$, including the last one in the $(3,3)$ slot are the three pivots. Once the system has been reduced to triangular form (1.15), we can easily solve it, as discussed earlier, by back substitution.

Gaussian Elimination — Regular Case

```
start
    for j = 1 to n
        if m_jj = 0, stop; print "A is not regular"
        else for i = j + 1 to n
            set l_ij = m_ij / m_jj
            add −l_ij times row j of M to row i of M
        next i
    next j
end
```

The preceding algorithm for solving a linear system is known as *regular Gaussian elimination*. A square matrix A will be called *regular*[†] if the algorithm successfully reduces it to upper triangular form U with all non-zero pivots on the diagonal. In other words, for regular matrices, we identify each successive nonzero entry in a diagonal position as the current pivot. We then use the pivot row to make all the entries in the column below the pivot equal to zero through elementary row operations of Type #1. A system whose coefficient matrix is regular is solved by first reducing the augmented matrix to upper triangular form and then solving the resulting triangular system by back substitution.

Let us state this algorithm in the form of a program, written in a general "pseudocode" that can be easily translated into any specific language, e.g., C++, FORTRAN, JAVA, MAPLE, MATHEMATICA or MATLAB. We use a single letter $M = (m_{ij})$ to denote the current augmented matrix at each stage in the computation, and initialize $M = \big(\, A \mid \mathbf{b} \,\big)$. Note that the entries of M will change as the algorithm progresses. The final output of the program, assuming A is regular, is the augmented matrix $M = \big(\, U \mid \mathbf{c} \,\big)$, where U is the upper triangular matrix U whose diagonal entries u_{ii} are the pivots and $\mathbf{c}$ is the vector of right hand sides obtained after performing the elementary row operations.

Elementary Matrices

A key observation is that elementary row operations can, in fact, be realized by matrix multiplication.

Definition 1.1. The *elementary matrix* E associated with an elementary row operation for matrices with m rows is the matrix obtained by applying the row operation to the $m \times m$ identity matrix I_m.

[†] Strangely, there is no commonly accepted term for these kinds of matrices. Our proposed adjective "regular" will prove to be quite useful in the sequel.

For example, applying the elementary row operation that adds -2 times the first row to the second row to the 3×3 identity matrix $\mathrm{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ results in the corresponding elementary matrix $E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. We claim that, if A is *any* 3–rowed matrix, then multiplying $E_1 A$ has the same effect as the given elementary row operation. For example,

$$E_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 1 & 1 & 4 \end{pmatrix},$$

which you may recognize as the first elementary row operation we used to solve the illustrative example. Indeed, if we set

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \qquad E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}, \tag{1.16}$$

then multiplication by E_1 will subtract twice the first row from the second row, multiplication by E_2 will subtract the first row from the third row, and multiplication by E_3 will add $\frac{1}{2}$ the second row to the third row — precisely the row operations used to place our original system in triangular form. Therefore, performing them in the correct order (and using the associativity of matrix multiplication), we conclude that when

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}, \qquad \text{then} \qquad E_3 E_2 E_1 A = U = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix}. \tag{1.17}$$

The reader should check this by directly multiplying the indicated matrices.

In general, then, the *elementary matrix* E of size $m \times m$ will have all 1's on the diagonal, a nonzero entry c in position (i, j), for some $i \neq j$, and all other entries equal to zero. If A is any $m \times n$ matrix, then the matrix product $E A$ is equal to the matrix obtained from A by the elementary row operation adding c times row j to row i. (Note the reversal of order of i and j.)

The elementary row operation that undoes adding c times row j to row i is the inverse row operation that subtracts c (or, equivalently, adds $-c$) times row j from row i. The corresponding *inverse elementary matrix* again has 1's along the diagonal and $-c$ in the (i, j) slot. Let us denote the inverses of the particular elementary matrices (1.16) by L_i, so that, according to our general rule,

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \qquad L_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{pmatrix}. \tag{1.18}$$

Note that the product

$$L_i E_i = \mathrm{I} \tag{1.19}$$

is the 3×3 identity matrix, reflecting the fact that these are inverse operations. (A more thorough discussion of matrix inverses will be postponed until the following section.)

The product of the latter three elementary matrices is equal to

$$L = L_1 L_2 L_3 = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix}. \tag{1.20}$$

The matrix L is called a *special lower triangular* matrix, where "lower triangular" means that all the entries above the main diagonal are 0, while "special" indicates that all the entries on the diagonal are equal to 1. Observe that the entries of L below the diagonal are the same as the corresponding nonzero entries in the L_i. This is a general fact, that holds when the lower triangular elementary matrices are multiplied in the correct order. (For instance, the product $L_3 L_2 L_1$ is not so easily predicted.) More generally, the following elementary consequence of the laws of matrix multiplication will be used extensively.

Lemma 1.2. *If L and $\widehat{L}$ are lower triangular matrices of the same size, so is their product $L\widehat{L}$. If they are both special lower triangular, so is their product. Similarly, if $U, \widehat{U}$ are (special) upper triangular matrices, so is their product $U\widehat{U}$.*

The LU Factorization

We have almost arrived at our first important result. Consider the product of the matrices L and U in (1.17), (1.20). Using equation (1.19), along with the basic property of the identity matrix I and associativity of matrix multiplication, we conclude that

$$\begin{aligned} LU &= (L_1 L_2 L_3)(E_3 E_2 E_1 A) = L_1 L_2 (L_3 E_3) E_2 E_1 A = L_1 L_2 \,\mathrm{I}\, E_2 E_1 A \\ &= L_1 (L_2 E_2) E_1 A = L_1 \,\mathrm{I}\, E_1 A = L_1 E_1 A = \mathrm{I}\, A = A. \end{aligned}$$

In other words, we have *factorized* the coefficient matrix $A = LU$ into a product of a special lower triangular matrix L and an upper triangular matrix U with the nonzero pivots on its main diagonal. The same holds true for almost all square coefficient matrices.

Theorem 1.3. *A matrix A is regular if and only if it can be factorized*

$$A = LU, \tag{1.21}$$

where L is a special lower triangular matrix, having all 1's on the diagonal, and U is upper triangular with nonzero diagonal entries, which are its pivots. The nonzero off-diagonal entries l_{ij} for $i > j$ appearing in L prescribe the elementary row operations that bring A into upper triangular form; namely, one subtracts l_{ij} times row j from row i at the appropriate step of the Gaussian elimination process.

Example 1.4. Let us compute the LU factorization of the matrix $A = \begin{pmatrix} 2 & 1 & 1 \\ 4 & 5 & 2 \\ 2 & -2 & 0 \end{pmatrix}$. Applying the Gaussian elimination algorithm, we begin by subtracting twice the first row from the second row, and then subtract the first row from the third. The result is the

matrix $\begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & -3 & -1 \end{pmatrix}$. The next step adds the second row to the third row, leading to the upper triangular matrix $U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}$, with its diagonal entries $2, 3, -1$ indicating the pivots. The corresponding lower triangular matrix is $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix}$, whose entries below the diagonal are the *negatives* of the multiples we used during the elimination procedure. Namely, the $(2,1)$ entry of L indicates that we added -2 times the first row to the second row; the $(3,1)$ entry indicates that we added -1 times the first row to the third; and, finally, the $(3,2)$ entry indicates that we added the second row to the third row during the algorithm. The reader might wish to verify the factorization $A = LU$, or, explicitly,

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

Forward and Back Substitution

Once we know the LU factorization of a regular matrix A, we are able to solve any associated linear system $A\,\mathbf{x} = \mathbf{b}$ in two stages:

(1) First solve the lower triangular system

$$L\,\mathbf{c} = \mathbf{b} \tag{1.22}$$

for the vector $\mathbf{c}$ by *forward substitution*. This is the same as back substitution, except one solves the equations for the variables in the direct order — from first to last. Explicitly,

$$c_1 = b_1, \qquad c_i = b_i - \sum_{j=1}^{i} l_{ij}\, c_j, \qquad \text{for} \qquad i = 2, 3, \ \dots \ , n, \tag{1.23}$$

noting that the previously computed values of $c_1, \dots, c_{i-1}$ are used to determine c_i.

(2) Second, solve the resulting upper triangular system

$$U\,\mathbf{x} = \mathbf{c} \tag{1.24}$$

by *back substitution*. Explicitly, the values of the unknowns

$$x_n = \frac{c_n}{u_{nn}}\,, \qquad x_i = \frac{1}{u_{ii}} \left(c_i - \sum_{j=i+1}^{n} u_{ij}\, x_j \right), \qquad \text{for} \qquad i = n-1, \ \dots \ , 2, 1, \tag{1.25}$$

are successively computed, but now in reverse order.

Note that this algorithm does indeed solve the original system, since if

$$U\,\mathbf{x} = \mathbf{c} \qquad \text{and} \qquad L\,\mathbf{c} = \mathbf{b}, \qquad \text{then} \qquad A\,\mathbf{x} = LU\,\mathbf{x} = L\,\mathbf{c} = \mathbf{b}.$$

Once we have found the LU factorization of the coefficient matrix A, the Forward and Back Substitution processes quickly produce the solution, and are easy to program on a computer.

Example 1.5. With the LU decomposition

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

found in Example 1.4, we can readily solve any linear system with the given coefficient matrix by Forward and Back Substitution. For instance, to find the solution to

$$\begin{pmatrix} 2 & 1 & 1 \\ 4 & 5 & 2 \\ 2 & -2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix},$$

we first solve the lower triangular system

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \qquad \text{or, explicitly,} \qquad \begin{aligned} a &= 1, \\ 2\,a + b &= 2, \\ a - b + c &= 2. \end{aligned}$$

The first equation says $a = 1$; substituting into the second, we find $b = 0$; the final equation gives $c = 1$. We then solve the upper triangular system

$$\begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \qquad \text{which is} \qquad \begin{aligned} 2\,x + y + z &= 1, \\ 3\,y &= 0, \\ -\,z &= 1. \end{aligned}$$

In turn, we find $z = -1$, then $y = 0$, and then $x = 1$, which is the unique solution to the original system.

Of course, if we are not given the LU factorization in advance, we can just use direct Gaussian elimination on the augmented matrix. Forward and Back Substitution is useful if one has already computed the factorization by solving for a particular right hand side $\mathbf{b}$, but then later wants to know the solutions corresponding to alternative $\mathbf{b}$'s.

1.4. Pivoting and Permutations.

The method of Gaussian elimination presented so far applies only to regular matrices. But not every square matrix is regular; a simple class of examples are matrices whose upper left entry is zero, and so cannot serve as the first pivot. More generally, the regular elimination algorithm cannot proceed whenever a zero entry appears in the current pivot spot on the diagonal. Zero can never serve as a pivot, since we cannot use it to eliminate any nonzero entries in the column below it. What then to do? The answer requires revisiting the source of our algorithm.

Let us consider, as a specific example, the linear system

$$\begin{aligned} 3\,y + z &= 2, \\ 2\,x + 6\,y + z &= 7, \\ x + 4\,z &= 3. \end{aligned} \tag{1.26}$$

The augmented coefficient matrix is

$$\left(\begin{array}{ccc|c} 0 & 3 & 1 & 2 \\ 2 & 6 & 1 & 7 \\ 1 & 0 & 4 & 3 \end{array}\right).$$

In this case, the $(1,1)$ entry is 0, and is not a legitimate pivot. The problem, of course, is that the first variable x does not appear in the first equation, and so we cannot use it to eliminate x in the other two equations. But this "problem" is actually a bonus — we already have an equation with only two variables in it, and so we only need to eliminate x from one of the other two equations. To be systematic, we rewrite the system in a different order,

$$\begin{aligned} 2\,x + 6\,y + z &= 7, \\ 3\,y + z &= 2, \\ x + 4\,z &= 3, \end{aligned}$$

by interchanging the first two equations. In other words, we employ

Linear System Operation #2: Interchange two equations.

Clearly this operation does not change the solution, and so produces an equivalent system. In our case, the resulting augmented coefficient matrix is

$$\left(\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 1 & 0 & 4 & 3 \end{array}\right),$$

and is obtained from the original by performing the second type of row operation:

Elementary Row Operation #2: Interchange two rows of the matrix.

The new nonzero upper left entry, 2, can now serve as the first pivot, and we may continue to apply elementary row operations of Type #1 to reduce our matrix to upper triangular form. For this particular example, we eliminate the remaining nonzero entry in the first column by subtracting $\frac{1}{2}$ the first row from the last:

$$\left(\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 0 & -3 & \frac{7}{2} & -\frac{1}{2} \end{array}\right).$$

The $(2,2)$ entry serves as the next pivot. To eliminate the nonzero entry below it, we add the second to the third row:

$$\left(\begin{array}{ccc|c} 2 & 6 & 1 & 7 \\ 0 & 3 & 1 & 2 \\ 0 & 0 & \frac{9}{2} & \frac{3}{2} \end{array}\right).$$

Gaussian Elimination — Nonsingular Case

```
start
    for j = 1 to n
        if m_kj = 0 for all k ≥ j, stop; print "A is singular"
        if m_jj = 0 but m_kj ≠ 0 for some k > j, switch rows k and j
        for i = j + 1 to n
            set l_ij = m_ij / m_jj
            add −l_ij times row j to row i of M
        next i
    next j
end
```

We have now placed the system in upper triangular form, with the three pivots, $2, 3, \frac{9}{2}$ along the diagonal. Back substitution produces the solution $x = \frac{5}{3}$, $y = \frac{5}{9}$, $z = \frac{1}{3}$.

The row interchange that is required when a zero shows up on the diagonal in pivot position is known as *pivoting*. Later, in Section 1.7, we shall discuss practical reasons for pivoting even when a diagonal entry is nonzero. The coefficient matrices for which the Gaussian elimination algorithm with pivoting produces the solution are of fundamental importance.

Definition 1.6. A square matrix is called *nonsingular* if it can be reduced to upper triangular form with all non-zero elements on the diagonal by elementary row operations of Types 1 and 2. Conversely, a square matrix that cannot be reduced to upper triangular form because at some stage in the elimination procedure the diagonal entry and all the entries below it are zero is called *singular*.

Every regular matrix is nonsingular, but, as we just saw, nonsingular matrices are more general. Uniqueness of solutions is the key defining characteristic of nonsingularity.

Theorem 1.7. *A linear system $A\mathbf{x} = \mathbf{b}$ has a unique solution for* every *choice of right hand side $\mathbf{b}$ if and only if its coefficient matrix A is square and nonsingular.*

We are able to prove the "if" part of this theorem, since nonsingularity implies reduction to an equivalent upper triangular form that has the same solutions as the original system. The unique solution to the system is found by back substitution. The "only if" part will be proved in Section 1.8.

The revised version of the Gaussian Elimination algorithm, valid for all nonsingular coefficient matrices, is implemented by the accompanying program. The starting point is the augmented matrix $M = \big(\, A \mid \mathbf{b} \,\big)$ representing the linear system $A\mathbf{x} = \mathbf{b}$. After successful termination of the program, the result is an augmented matrix in upper triangular form $M = \big(\, U \mid \mathbf{c} \,\big)$ representing the equivalent linear system $U\mathbf{x} = \mathbf{c}$. One then uses Back Substitution to determine the solution $\mathbf{x}$ to the linear system.

Permutation Matrices

As with the first type of elementary row operation, row interchanges can be accomplished by multiplication by a second type of elementary matrix. Again, the elementary matrix is found by applying the row operation in question to the identity matrix of the appropriate size. For instance, interchanging rows 1 and 2 of the 3×3 identity matrix produces the elementary interchange matrix

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

As the reader can check, the effect of multiplying a 3 rowed matrix A on the left by P, producing PA, is the same as interchanging the first two rows of A. For instance,

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 4 & 5 & 6 \\ 1 & 2 & 3 \\ 7 & 8 & 9 \end{pmatrix}.$$

Multiple row interchanges are accomplished by combining such elementary interchange matrices. Each such combination of row interchanges corresponds to a unique permutation matrix.

Definition 1.8. A *permutation matrix* is a matrix obtained from the identity matrix by any combination of row interchanges.

In particular, applying a row interchange to a permutation matrix produces another permutation matrix. The following result is easily established.

Lemma 1.9. *A matrix P is a permutation matrix if and only if each row of P contains all 0 entries except for a single 1, and, in addition, each column of P also contains all 0 entries except for a single 1.*

In general, if a permutation matrix P has a 1 in position (i, j), then the effect of multiplication by P is to move the j^{th} row of A into the i^{th} row of the product PA.

Example 1.10. There are six different 3×3 permutation matrices, namely

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{1.27}$$

These have the following effects: if A is a matrix with row vectors $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$, then multiplication on the left by each of the six permutation matrices produces

$$\begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_2 \\ \mathbf{r}_1 \\ \mathbf{r}_3 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_3 \\ \mathbf{r}_2 \\ \mathbf{r}_1 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_3 \\ \mathbf{r}_2 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_2 \\ \mathbf{r}_3 \\ \mathbf{r}_1 \end{pmatrix}, \quad \begin{pmatrix} \mathbf{r}_3 \\ \mathbf{r}_1 \\ \mathbf{r}_2 \end{pmatrix},$$

respectively. Thus, the first permutation matrix, which is the identity, does nothing. The second, third and fourth represent row interchanges. The last two are non-elementary permutations; each can be realized as a pair of row interchanges.

An elementary combinnatorial argument proves that there are a total of

$$n! = n\,(n-1)\,(n-2)\;\cdots\;3\cdot 2\cdot 1 \tag{1.28}$$

different permutation matrices of size $n \times n$. Moreover, the product $P = P_1 P_2$ of any two permutation matrices is also a permutation matrix. An important point is that multiplication of permutation matrices is *noncommutative* — the order in which one permutes makes a difference. Switching the first and second rows, and then switching the second and third rows *does not* have the same effect as first switching the second and third rows and then switching the first and second rows!

The Permuted $L\,U$ Factorization

As we now know, any nonsingular matrix A can be reduced to upper triangular form by elementary row operations of types #1 and #2. The row interchanges merely reorder the equations. If one performs all of the required row interchanges in advance, then the elimination algorithm can proceed without requiring any further pivoting. Thus, the matrix obtained by permuting the rows of A in the prescribed manner is regular. In other words, if A is a nonsingular matrix, then there is a permutation matrix P such that the product $P\,A$ is regular, and hence admits an $L\,U$ factorization. As a result, we deduce the general *permuted $L\,U$ factorization*

$$P\,A = L\,U, \tag{1.29}$$

where P is a permutation matrix, L is special lower triangular, and U is upper triangular with the pivots on the diagonal. For instance, in the preceding example, we permuted the first and second rows, and hence equation (1.29) has the explicit form

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & \frac{9}{2} \end{pmatrix}. \tag{1.30}$$

As a result of these considerations, we have established the following generalization of Theorem 1.3.

Theorem 1.11. *Let A be an $n \times n$ matrix. Then the following conditions are equivalent:*

(*i*) *A is nonsingular.*
(*ii*) *A has n nonzero pivots.*
(*iii*) *A admits a permuted $L\,U$ factorization: $P\,A = L\,U$,.*

One should be aware of a couple of practical complications. First, to implement the permutation P of the rows that makes A regular, one needs to be clairvoyant: it is not always clear in advance when and where a required row interchange will crop up. Second, any row interchange performed during the course of the Gaussian Elimination algorithm will affect the lower triangular matrix L, and precomputed entries must be permuted accordingly; an example appears in Exercise ∎.

Once the permuted $L\,U$ factorization is established, the solution to the original system $A\,\mathbf{x} = \mathbf{b}$ is obtained by using the same Forward and Back Substitution algorithm presented

above. Explicitly, we first multiply the system $A\,\mathbf{x} = \mathbf{b}$ by the permutation matrix, leading to

$$P A\,\mathbf{x} = P\,\mathbf{b} \equiv \widehat{\mathbf{b}}, \tag{1.31}$$

whose right hand side $\widehat{\mathbf{b}}$ has been obtained by permuting the entries of $\mathbf{b}$ in the same fashion as the rows of A. We then solve the two systems

$$L\,\mathbf{c} = \widehat{\mathbf{b}}, \qquad \text{and} \qquad U\,\mathbf{x} = \mathbf{c}, \tag{1.32}$$

by, respectively, Forward and Back Substitution as before.

Example 1.12. Suppose we wish to solve

$$\begin{pmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}.$$

In view of the $P A = L U$ factorization established in (1.30), we need only solve the two auxiliary systems (1.32) by Forward and Back Substitution, respectively. The lower triangular system is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix},$$

with solution $a = -2$, $b = 1$, $c = 2$. The resulting upper triangular system is

$$\begin{pmatrix} 2 & 6 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & \frac{9}{2} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

The solution, which is also the solution to the original system, is obtained by back substitution, with $z = \frac{4}{9}$, $y = \frac{5}{18}$, $x = -\frac{37}{18}$.

1.5. Matrix Inverses.

The inverse of a matrix is analogous to the reciprocal $a^{-1} = 1/a$ of a scalar, which is the 1×1 case. We already introduced the inverses of matrices corresponding to elementary row operations. In this section, we will analyze inverses of general square matrices. We begin with the formal definition.

Definition 1.13. Let A be a square matrix of size $n \times n$. An $n \times n$ matrix X is called the *inverse* of A if it satisfies

$$X A = \mathrm{I} = A X, \tag{1.33}$$

where $\mathrm{I} = \mathrm{I}_n$ is the $n \times n$ identity matrix. The inverse is commonly denoted by $X = A^{-1}$.

Remark: Noncommutativity of matrix multiplication requires that we impose both conditions in (1.33) in order to properly define an inverse to the matrix A. The first condition $X A = \mathrm{I}$ says that X is a *left inverse*, while the second $A X = \mathrm{I}$ requires that X also be a *right inverse*, in order that it fully qualify as a *bona fide* inverse of A.

Example 1.14. Since

$$\begin{pmatrix} 1 & 2 & -1 \\ -3 & 1 & 2 \\ -2 & 2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 4 & -5 \\ 1 & 1 & -1 \\ 4 & 6 & -7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 4 & -5 \\ 1 & 1 & -1 \\ 4 & 6 & -7 \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ -3 & 1 & 2 \\ -2 & 2 & 1 \end{pmatrix},$$

we conclude that when $A = \begin{pmatrix} 1 & 2 & -1 \\ -3 & 1 & 2 \\ -2 & 2 & 1 \end{pmatrix}$ then $A^{-1} = \begin{pmatrix} 3 & 4 & -5 \\ 1 & 1 & -1 \\ 4 & 6 & -7 \end{pmatrix}$. Note that the entries of A^{-1} do not follow any easily discernable pattern in terms of the entries of A.

Not every square matrix has an inverse. Indeed, not every scalar has an inverse — the one counterexample being $a = 0$. There is *no* general concept of inverse for rectangular matrices.

Example 1.15. Let us compute the inverse $X = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$ of a general 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The right inverse condition

$$A\,X = \begin{pmatrix} a\,x + b\,z & a\,y + b\,w \\ c\,x + d\,z & c\,y + d\,w \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{I}$$

holds if and only if x, y, z, w satisfy the linear system

$$\begin{aligned} a\,x + b\,z &= 1, & a\,y + b\,w &= 0, \\ c\,x + d\,z &= 0, & c\,y + d\,w &= 0. \end{aligned}$$

Solving by Gaussian elimination (or directly), we find

$$x = \frac{d}{a\,d - b\,c}\,, \qquad y = -\,\frac{b}{a\,d - b\,c}\,, \qquad z = -\,\frac{c}{a\,d - b\,c}\,, \qquad w = \frac{a}{a\,d - b\,c}\,,$$

provided the common denominator $a\,d - b\,c \neq 0$ does not vanish. Therefore, the matrix

$$X = \frac{1}{a\,d - b\,c} \begin{pmatrix} d & -\,b \\ -\,c & a \end{pmatrix}$$

forms a right inverse to A. However, a short computation shows that it also defines a left inverse:

$$X\,A = \begin{pmatrix} x\,a + y\,c & x\,b + y\,d \\ z\,a + w\,c & z\,b + w\,d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathrm{I}\,,$$

and hence $X = A^{-1}$ is the inverse to A.

The denominator appearing in the preceding formulae has a special name; it is called the *determinant* of the 2×2 matrix A, and denoted

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\,d - b\,c. \tag{1.34}$$

Thus, the determinant of a 2×2 matrix is the product of the diagonal entries minus the product of the off-diagonal entries. (Determinants of larger square matrices will be discussed in Section 1.9.) Thus, the 2×2 matrix A is invertible, with

$$A^{-1} = \frac{1}{a\,d - b\,c} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \tag{1.35}$$

if and only if $\det A \neq 0$. For example, if $A = \begin{pmatrix} 1 & 3 \\ -2 & -4 \end{pmatrix}$, then $\det A = 2 \neq 0$. We conclude that A has an inverse, which, by (1.35), is $A^{-1} = \frac{1}{2} \begin{pmatrix} -4 & -3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} -2 & -\frac{3}{2} \\ 1 & \frac{1}{2} \end{pmatrix}$.

The following key result will be established later in this chapter.

Theorem 1.16. *A square matrix A has an inverse if and only if it is nonsingular.*

Consequently, an $n \times n$ matrix will have an inverse if and only if it can be reduced to upper triangular form with n nonzero pivots on the diagonal by a combination of elementary row operations. Indeed, "invertible" is often used as a synoinym for "nonsingular". All other matrices are singular and do not have an inverse as defined above. Before attempting to prove this fundamental result, we need to first become familiar with some elementary properties of matrix inverses.

Lemma 1.17. *The inverse of a square matrix, if it exists, is unique.*

Proof: If X and Y both satisfy (1.33), so $X\,A = \mathrm{I} = A\,X$ and $Y\,A = \mathrm{I} = A\,Y$, then, by associativity, $X = X\,\mathrm{I} = X(A\,Y) = (X\,A)\,Y = \mathrm{I}\,Y = Y$, and hence $X = Y$. *Q.E.D.*

Inverting a matrix twice gets us back to where we started.

Lemma 1.18. *If A is invertible, then A^{-1} is also invertible and $(A^{-1})^{-1} = A$.*

Proof: The matrix inverse equations $A^{-1}\,A = \mathrm{I} = A\,A^{-1}$, are sufficient to prove that A is the inverse of A^{-1}. *Q.E.D.*

Example 1.19. We already learned how to find the inverse of an elementary matrix of type #1; we just negate the one nonzero off-diagonal entry. For example, if

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}, \qquad \text{then} \qquad E^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

This reflects the fact that the inverse of the elementary row operation that adds twice the first row to the third row is the operation of subtracting twice the first row from the third row.

Example 1.20. Let $P = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ denote the elementary matrix that has the effect of interchanging rows 1 and 2 of a matrix. Then $P^2 = \mathrm{I}$, since doing the same

operation twice in a row has no net effect. This implies that $P^{-1} = P$ is its own inverse. Indeed, the same result holds for all elementary permutation matrices that correspond to row operations of type #2. However, it is not true for more general permutation matrices.

Lemma 1.21. *If A and B are invertible matrices of the same size, then their product, AB, is invertible, and*

$$(AB)^{-1} = B^{-1}A^{-1}. \tag{1.36}$$

Note particularly the reversal in order of the factors.

Proof: Let $X = B^{-1}A^{-1}$. Then, by associativity,

$$X(AB) = B^{-1}A^{-1}AB = B^{-1}B = \mathrm{I}, \qquad (AB)X = ABB^{-1}A^{-1} = AA^{-1} = \mathrm{I}.$$

Thus X is both a left and a right inverse for the product matrix AB and the result follows. *Q.E.D.*

Example 1.22. One verifies, directly, that the inverse of $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ is $A^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$, while the inverse of $B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ is $B^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Therefore, the inverse of thier product $C = AB = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ -1 & 0 \end{pmatrix}$ is given by $C^{-1} = B^{-1}A^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & -2 \end{pmatrix}$.

We can straightforwardly generalize the preceding result. The inverse of a multiple product of invertible matrices is the product of their inverses, *in the reverse order*:

$$(A_1A_2\cdots A_{m-1}A_m)^{-1} = A_m^{-1}A_{m-1}^{-1}\cdots A_2^{-1}A_1^{-1}. \tag{1.37}$$

Warning: In general, $(A + B)^{-1} \neq A^{-1} + B^{-1}$. This equation is not even true for scalars (1×1 matrices)!

Gauss–Jordan Elimination

The basic algorithm used to compute the inverse of a square matrix is known as *Gauss–Jordan Elimination*, in honor of Gauss and Wilhelm Jordan, a nineteenth century German engineer. A key fact is that we only need to solve the right inverse equation

$$AX = \mathrm{I} \tag{1.38}$$

in order to compute $X = A^{-1}$. The other equation in (1.33), namely $XA = \mathrm{I}$, will then follow as an automatic consequence. In other words, for square matrices, a right inverse is automatically a left inverse, and conversely! A proof will appear below.

The reader may well ask, then, why use both left and right inverse conditions in the original definition? There are several good reasons. First of all, a rectangular matrix may satisfy one of the two conditions — having either a left inverse or a right inverse — but can never satisfy both. Moreover, even when we restrict our attention to square

matrices, starting with only one of the conditions makes the logical development of the subject considerably more difficult, and not really worth the extra effort. Once we have established the basic properties of the inverse of a square matrix, we can then safely discard the superfluous left inverse condition. Finally, when we generalize the notion of an inverse to a linear operator in Chapter 7, then, unlike square matrices, we *cannot* dispense with either of the conditions.

Let us write out the individual columns of the right inverse equation (1.38). The i^{th} column of the $n \times n$ identity matrix I is the vector $\mathbf{e}_i$ that has a single 1 in the i^{th} slot and 0's elsewhere, so

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \dots \qquad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{1.39}$$

According to (1.11), the i^{th} column of the matrix product AX is equal to $A\,\mathbf{x}_i$, where $\mathbf{x}_i$ denotes the i^{th} column of $X = (\,\mathbf{x}_1\ \mathbf{x}_2\ \dots\ \mathbf{x}_n\,)$. Therefore, the single matrix equation (1.38) is equivalent to n linear systems

$$A\,\mathbf{x}_1 = \mathbf{e}_1, \qquad A\,\mathbf{x}_2 = \mathbf{e}_2, \qquad \dots \qquad A\,\mathbf{x}_n = \mathbf{e}_n, \tag{1.40}$$

all having the same coefficient matrix. As such, to solve them we are led to form the n augmented matrices $M_1 = \big(\,A \mid \mathbf{e}_1\,\big)$, $\dots$, $M_n = \big(\,A \mid \mathbf{e}_n\,\big)$, and then perform our Gaussian elimination algorithm on each one. But this would be a waste of effort. Since the coefficient matrix is the same, we will end up performing *identical* row operations on each augmented matrix. Consequently, it will be more efficient to combine them into one large augmented matrix $M = \big(\,A \mid \mathbf{e}_1\ \dots\ \mathbf{e}_n\,\big) = \big(\,A \mid \mathrm{I}\,\big)$, of size $n \times (2n)$, in which the right hand sides $\mathbf{e}_1, \dots, \mathbf{e}_n$ of our systems are placed into n different columns, which we then recognize as reassembling the columns of an $n \times n$ identity matrix. We may then apply our elementary row operations to reduce, if possible, the large augmented matrix so that its first n columns are in upper triangular form.

Example 1.23. For example, to find the inverse of the matrix $A = \begin{pmatrix} 0 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}$, we form the large augmented matrix

$$\left(\begin{array}{ccc|ccc} 0 & 2 & 1 & 1 & 0 & 0 \\ 2 & 6 & 1 & 0 & 1 & 0 \\ 1 & 1 & 4 & 0 & 0 & 1 \end{array}\right).$$

Applying the same sequence of elementary row operations as in Section 1.4, we first interchange the rows

$$\left(\begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 1 & 1 & 4 & 0 & 0 & 1 \end{array}\right),$$

and then eliminate the nonzero entries below the first pivot,

$$\left(\begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & -2 & \frac{7}{2} & 0 & -\frac{1}{2} & 1 \end{array}\right).$$

Next we eliminate the entry below the second pivot:

$$\left(\begin{array}{ccc|ccc} 2 & 6 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 \\ 0 & 0 & \frac{9}{2} & 1 & -\frac{1}{2} & 1 \end{array}\right).$$

At this stage, we have reduced our augmented matrix to the upper triangular form $\big(\, U \mid C \,\big)$, which is equivalent to reducing the original n linear systems $A\,\mathbf{x}_i = \mathbf{e}_i$ to n upper triangular systems $U\,\mathbf{x}_i = \mathbf{c}_i$. We could therefore perform n back substitutions to produce the solutions $\mathbf{x}_i$, which would form the individual columns of the inverse matrix $X = (\mathbf{x}_1 \,\ldots\, \mathbf{x}_n)$.

In the standard Gauss–Jordan scheme, one instead continues to employ the usual sequence of elementary row operations to fully reduce the augmented matrix to the form $\big(\, \mathrm{I} \mid X \,\big)$ in which the left hand $n \times n$ matrix has become the identity, while the right hand matrix is the desired solution $X = A^{-1}$. Indeed, $\big(\, \mathrm{I} \mid X \,\big)$ represents the n trivial, but equivalent, linear systems $\mathrm{I}\,\mathbf{x}_i = \mathbf{x}_i$ with identity coefficient matrix.

Now, the identity matrix has 0's below the diagonal, just like U. It also has 1's along the diagonal, whereas U has the pivots (which are all nonzero) along the diagonal. Thus, the next phase in the procedure is to make all the diagonal entries of U equal to 1. To do this, we need to introduce the last, and least, of our linear systems operations.

Linear System Operation #3: Multiply an equation by a nonzero constant.

This operation does not change the solution, and so yields an equivalent linear system. The corresponding elementary row operation is:

Elementary Row Operation #3: Multiply a row of the matrix by a nonzero scalar.

Dividing the rows of the upper triangular augmented matrix $\big(\, U \mid C \,\big)$ by the diagonal pivots of U will produce a matrix of the form $\big(\, V \mid K \,\big)$ where V is *special upper triangular*, meaning it has all 1's along the diagonal. In the particular example, the result of these three elementary row operations of Type #3 is

$$\left(\begin{array}{ccc|ccc} 1 & 3 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array}\right),$$

where we multiplied the first and second rows by $\frac{1}{2}$ and the third row by $\frac{2}{9}$.

We are now over half way towards our goal of an identity matrix on the left. We need only make the entries above the diagonal equal to zero. This can be done by elementary row operations of Type #1, but now we work backwards as in back substitution. First,

eliminate the nonzero entries in the third column lying above the $(3,3)$ entry; this is done by subtracting one half the third row from the second and also from the first:

$$\left(\begin{array}{ccc|ccc} 1 & 3 & 0 & -\frac{1}{9} & \frac{5}{9} & -\frac{1}{9} \\ 0 & 1 & 0 & \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ 0 & 0 & 1 & \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array}\right).$$

Finally, subtract $\frac{1}{3}$ the second from the first to eliminate the remaining nonzero off-diagonal entry:

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & -\frac{23}{18} & \frac{7}{18} & \frac{2}{9} \\ 0 & 1 & 0 & \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ 0 & 0 & 1 & \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{array}\right).$$

The final right hand matrix is our desired inverse:

$$A^{-1} = \begin{pmatrix} -\frac{23}{18} & \frac{7}{18} & \frac{2}{9} \\ \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{pmatrix},$$

thereby completing the Gauss–Jordan procedure. The reader may wish to verify that the final result does satisfy both inverse conditions $A\,A^{-1} = \mathrm{I} = A^{-1}A$.

We are now able to complete the proofs of the basic results on inverse matrices. First, we need to determine the elementary matrix corresponding to an elementary row operation of type #3. Again, this is obtained by performing the indicated elementary row operation on the identity matrix. Thus, the elementary matrix that multiplies row i by the nonzero scalar $c \neq 0$ is the diagonal matrix having c in the i^{th} diagonal position, and 1's elsewhere along the diagonal. The inverse elementary matrix is the diagonal matrix with $1/c$ in the i^{th} diagonal position and 1's elsewhere on the main diagonal; it corresponds to the inverse operation that divides row i by c. For example, the elementary matrix that multiplies the second row of a $3 \times n$ matrix by the scalar 5 is

$$E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \text{and has inverse} \qquad E^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The Gauss–Jordan method tells us how to reduce any nonsingular square matrix A to the identity matrix by a sequence of elementary row operations. Let $E_1, E_2, \dots, E_N$ be the corresponding elementary matrices. Therefore,

$$E_N\, E_{N-1}\, \cdots\, E_2\, E_1\, A = \mathrm{I}\,. \tag{1.41}$$

We claim that the matrix product

$$X = E_N\, E_{N-1}\, \cdots\, E_2\, E_1 \tag{1.42}$$

is the inverse of A. Indeed, formula (1.41) says that $X\,A = \mathrm{I}\,$, and so X is a left inverse. Furthermore, each elementary matrix has an inverse, and so by (1.37), X itself is invertible, with

$$X^{-1} = E_1^{-1}\, E_2^{-1}\, \cdots\, E_{N-1}^{-1}\, E_N^{-1}. \tag{1.43}$$

Therefore, multiplying the already established formula $X\,A = \mathrm{I}$ on the left by X^{-1}, we find $A = X^{-1}$, and so, by Lemma 1.18, $X = A^{-1}$ as claimed. This completes the proof of Theorem 1.16. Finally, equating $A = X^{-1}$ to (1.43), and using the fact that the inverse of an elementary matrix is also an elementary matrix, we have established:

Proposition 1.24. *Any nonsingular matrix A can be written as the product of elementary matrices.*

For example, the 2×2 matrix $A = \begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}$ is converted into the identity matrix by row operations corresponding to the matrices $E_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, corresponding to a row interchange, $E_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, scaling the second row by -1, and $E_3 = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix}$ that subtracts 3 times the second row from the first. Therefore,

$$A^{-1} = E_3\,E_2\,E_1 = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -1 & 0 \end{pmatrix},$$

while

$$A = E_1^{-1}\,E_2^{-1}\,E_3^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 3 \end{pmatrix}.$$

As an application, let us prove that the inverse of a nonsingular triangular matrix is also triangular. Specifically:

Lemma 1.25. *If L is a lower triangular matrix with all nonzero entries on the main diagonal, then L is nonsingular and its inverse L^{-1} is also lower triangular. In particular, if L is special lower triangular, so is L^{-1}. A similar result holds for upper triangular matrices.*

Proof: It suffices to note that if L has all nonzero diagonal entries, one can reduce L to the identity by elementary row operations of Types #1 and #3, whose associated elementary matrices are all lower triangular. Lemma 1.2 implies that the product (1.42) is then also lower triangular. If L is special, then all the pivots are equal to 1 and so no elementary row operations of Type #3 are required, so the inverse is a product of special lower triangular matrices, and hence is special lower triangular. *Q.E.D.*

Solving Linear Systems with the Inverse

An important motivation for the matrix inverse is that it enables one to effect an immediate solution to a nonsingular linear system.

Theorem 1.26. *If A is invertible, then the unique solution to the linear system $A\,\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = A^{-1}\,\mathbf{b}$.*

Proof: We merely multiply the system by A^{-1}, which yields $\mathbf{x} = A^{-1}A\,\mathbf{x} = A^{-1}\mathbf{b}$, as claimed. *Q.E.D.*

Thus, with the inverse in hand, a "more direct" way to solve our example (1.26) is to multiply the right hand side by the inverse matrix:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\frac{23}{18} & \frac{7}{18} & \frac{2}{9} \\ \frac{7}{18} & \frac{1}{18} & -\frac{1}{9} \\ \frac{2}{9} & -\frac{1}{9} & \frac{2}{9} \end{pmatrix} \begin{pmatrix} 2 \\ 7 \\ 3 \end{pmatrix} = \begin{pmatrix} \frac{5}{6} \\ \frac{5}{6} \\ \frac{1}{3} \end{pmatrix},$$

reproducing our earlier solution.

However, while æsthetically appealing, the solution method based on the inverse matrix is hopelessly inefficient as compared to forward and back substitution based on a (permuted) LU factorization, and *should not be used*. A complete justification of this dictum will be provided in Section 1.7. In contrast to what you might have learned in an introductory linear algebra course, you should never use the matrix inverse for practical computations! This is *not* to say that the inverse is completely without merit. Far from it! The inverse continues to play a fundamental role in the theoretical side of linear algebra, as well as providing important insight into the algorithms that are used in practice. But the basic message of practical, applied linear algebra is that LU decomposition and Gaussian Elimination are fundamental; inverses are only used for theoretical purposes, and are to be avoided in all but the most elementary practical computations.

Remark: The reader may have learned a version of the Gauss–Jordan algorithm for solving a single linear system that replaces the back substitution step by a further application of all three types of elementary row operations in order to reduce the coefficient matrix to the identity. In other words, to solve $A\,\mathbf{x} = \mathbf{b}$, we start with the augmented matrix $M = \big(\, A \mid \mathbf{b} \,\big)$ and use all three types of elementary row operations to produce (assuming nonsingularity) the fully reduced form $\big(\, \mathrm{I} \mid \mathbf{x} \,\big)$, representing the trivial, equivalent system $\mathrm{I}\,\mathbf{x} = \mathbf{x}$, with the solution $\mathbf{x}$ to the original system in its final column. However, as we shall see, back substitution is much more efficient, and is the method of choice in all practical situations.

The LDV Factorization

The Gauss–Jordan construction leads to a slightly more detailed version of the LU factorization, which is useful in certain situations. Let D denote the diagonal matrix having the same diagonal entries as U; in other words, D has the pivots on its diagonal and zeros everywhere else. Let V be the special upper triangular matrix obtained from U by dividing each row by its pivot, so that V has all 1's on the diagonal. We already encountered V during the course of the Gauss–Jordan method. It is easily seen that $U = DV$, which implies the following result.

Theorem 1.27. *A matrix A is regular if and only if it admits a factorization*

$$A = LDV, \tag{1.44}$$

where L is special lower triangular matrix, D is a diagonal matrix having the nonzero pivots on the diagonal, and V is special upper triangular.

For the matrix appearing in Example 1.5, we have $U = DV$, where

$$U = \begin{pmatrix} 2 & 1 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad D = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad V = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

producing the $A = LDV$ factorization

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

Proposition 1.28. *If $A = LU$ is regular, then the factors L and U are each uniquely determined. The same holds for its $A = LDV$ factorization.*

Proof: Suppose $LU = \widetilde{L}\,\widetilde{U}$. Since the diagonal entries of all four matrices are nonzero, Lemma 1.25 implies that they are invertible. Therefore,

$$\widetilde{L}^{-1}L = \widetilde{L}^{-1}LU\,U^{-1} = \widetilde{L}^{-1}\widetilde{L}\,\widetilde{U}\,U^{-1} = \widetilde{U}\,U^{-1}. \tag{1.45}$$

The left hand side of the matrix equation (1.45) is the product of two special lower triangular matrices, and so, according to Lemma 1.2, is itself special lower triangular — with 1's on the diagonal. The right hand side is the product of two upper triangular matrices, and hence is itself upper triangular. Comparing the individual entries, the only way such a special lower triangular matrix could equal an upper triangular matrix is if they both equal the diagonal identity matrix. Therefore, $\widetilde{L}^{-1}L = \mathrm{I} = \widetilde{U}\,U^{-1}$, which implies that $\widetilde{L} = L$ and $\widetilde{U} = U$, and proves the result. The LDV version is an immediate consequence. *Q.E.D.*

As you may have guessed, the more general cases requiring one or more row interchanges lead to a permuted LDV factorization in the following form.

Theorem 1.29. *A matrix A is nonsingular if and only if there is a permutation matrix P such that*

$$PA = LDV, \tag{1.46}$$

where L, D, V are as before.

Uniqueness does not hold for the more general permuted factorizations (1.29), (1.46) since there may be various permutation matrices that place a matrix A in regular form PA; see Exercise ■ for an explicit example. Moreover, unlike the regular case, the pivots, i.e., the diagonal entries of U, are no longer uniquely defined, but depend on the particular combination of row interchanges employed during the course of the computation.

1.6. Transposes and Symmetric Matrices.

Another basic operation on a matrix is to interchange its rows and columns. If A is an $m \times n$ matrix, then its *transpose*, denoted A^T, is the $n \times m$ matrix whose (i, j) entry equals the (j, i) entry of A; thus

$$B = A^T \qquad \text{means that} \qquad b_{ij} = a_{ji}.$$

For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \qquad \text{then} \qquad A^T = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}.$$

Note that the rows of A are the columns of A^T and vice versa. In particular, the transpose of a row vector is a column vector, while the transpose of a column vector is a row vector.

$$\text{For example, if} \qquad \mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \qquad \text{then} \qquad \mathbf{v}^T = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}.$$

The transpose of a scalar, considered as a 1×1 matrix, is itself: $c^T = c$ for $c \in \mathbb{R}$.

Remark: Most vectors appearing in applied mathematics are column vectors. To conserve vertical space in this text, we will often use the transpose notation, e.g., $\mathbf{v} = \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}^T$, as a compact way of writing column vectors.

In the square case, transpose can be viewed as "reflecting" the matrix entries across the main diagonal. For example,

$$\begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & 5 \\ -2 & -4 & 8 \end{pmatrix}^T = \begin{pmatrix} 1 & 3 & -2 \\ 2 & 0 & -4 \\ -1 & 5 & 8 \end{pmatrix}.$$

In particular, the transpose of a lower triangular matrix is upper triangular and vice-versa.

Performing the transpose twice gets you back to where you started:

$$(A^T)^T = A. \tag{1.47}$$

Unlike the inverse, the transpose *is* compatible with matrix addition and scalar multiplication:

$$(A + B)^T = A^T + B^T, \qquad (cA)^T = cA^T. \tag{1.48}$$

The transpose is also compatible with matrix multiplication, but with a twist. Like the inverse, the transpose *reverses* the order of multiplication:

$$(AB)^T = B^T A^T. \tag{1.49}$$

The proof of (1.49) is a straightforward consequence of the basic laws of matrix multiplication. An important special case is the product between a row vector $\mathbf{v}^T$ and a column vector $\mathbf{w}$. In this case,

$$\mathbf{v}^T \mathbf{w} = (\mathbf{v}^T \mathbf{w})^T = \mathbf{w}^T \mathbf{v}, \tag{1.50}$$

because the product is a scalar and so equals its own transpose.

Lemma 1.30. *The operations of transpose and inverse commute. In other words, if A is invertible, so is A^T, and its inverse is*

$$A^{-T} \equiv (A^T)^{-1} = (A^{-1})^T. \tag{1.51}$$

Proof: Let $Y = (A^{-1})^T$. Then, according to (1.49),

$$Y\,A^T = (A^{-1})^T A^T = (A\,A^{-1})^T = \mathrm{I}^{\,T} = \mathrm{I}\,.$$

The proof that $A^T Y = \mathrm{I}$ is similar, and so we conclude that $Y = (A^T)^{-1}$. *Q.E.D.*

Factorization of Symmetric Matrices

The most important class of square matrices are those that are unchanged by the transpose operation.

Definition 1.31. A square matrix is *symmetric* if it equals its own transpose: $A = A^T$.

Thus, A is symmetric if and only if its entries satisfy $a_{ji} = a_{ij}$ for all i, j. In other words, entries lying in "mirror image" positions relative to the main diagonal must be equal. For example, the most general symmetric 3×3 matrix has the form

$$A = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}.$$

Note that any diagonal matrix, including the identity, is symmetric. A lower or upper triangular matrix is symmetric if and only if it is, in fact, a diagonal matrix.

The $L\,D\,V$ factorization of a nonsingular matrix takes a particularly simple form if the matrix also happens to be symmetric. This result will form the foundation of some significant later developments.

Theorem 1.32. *A symmetric matrix A is regular if and only if it can be factored as*

$$A = L\,D\,L^T, \tag{1.52}$$

where L is a special lower triangular matrix and D is a diagonal matrix with nonzero diagonal entries.

Proof: We already know, according to Theorem 1.27, that we can factorize

$$A = L\,D\,V. \tag{1.53}$$

We take the transpose of both sides of this equation and use the fact that the tranpose of a matrix product is the product of the transposes in the reverse order, whence

$$A^T = (L\,D\,V)^T = V^T D^T L^T = V^T D\,L^T, \tag{1.54}$$

where we used the fact that a diagonal matrix is automatically symmetric, $D^T = D$. Note that V^T is special lower triangular, and L^T is special upper triangular. Therefore (1.54) gives the $L\,D\,V$ factorization of A^T.

In particular, if $A = A^T$, then we can invoke the uniqueness of the $L\,D\,V$ factorization, cf. Proposition 1.28, to conclude that $L = V^T$, and $V = L^T$, (which are two versions of the same equation). Replacing V by L^T in (1.53) proves the factorization (1.52). *Q.E.D.*

Example 1.33. Let us find the $L\,D\,L^T$ factorization of the particular symmetric matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix}$. This is done by performing the usual Gaussian elimination algorithm. Subtracting twice the first row from the second and also the first row from the third produces the matrix $\begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$. We then add one half of the second row of the latter matrix to its third row, resulting in the upper triangular form

$$U = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix} = D\,V,$$

which we further factorize by dividing each row of U by its pivot. On the other hand, the special lower triangular matrix associated with the row operations is $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix}$, which, as guaranteed by Theorem 1.32, is the transpose of $V = L^T$. Therefore, the desired $A = LU = L\,D\,L^T$ factorizations of this particular symmetric matrix are

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & -\frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \frac{5}{2} \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}.$$

Example 1.34. Let us look at a general 2×2 symmetric matrix

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}. \tag{1.55}$$

Regularity requires that the first pivot be $a \neq 0$. A single row operation will place A in upper triangular form $U = \begin{pmatrix} a & c \\ 0 & \dfrac{a\,c - b^2}{a} \end{pmatrix}$. The associated lower triangular matrix is $L = \begin{pmatrix} 1 & 0 \\ \dfrac{b}{a} & 1 \end{pmatrix}$. Thus, $A = LU$. Finally, $D = \begin{pmatrix} a & 0 \\ 0 & \dfrac{a\,c - b^2}{a} \end{pmatrix}$ is just the diagonal part of U, and we find $U = D\,L^T$, so that the $L\,D\,L^T$ factorization is explicitly given by

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \dfrac{b}{a} & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & \dfrac{a\,c - b^2}{a} \end{pmatrix} \begin{pmatrix} 1 & \dfrac{b}{a} \\ 0 & 1 \end{pmatrix}. \tag{1.56}$$

Remark: If $A = L\,D\,L^T$, then A is necessarily symmetric. Indeed,

$$A^T = (L\,D\,L^T)^T = (L^T)^T D^T L^T = L\,D\,L^T = A.$$

However, not every symmetric matrix has an $L\,D\,L^T$ factorization. A simple example is the irregular but invertible 2×2 matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

1.7. Practical Linear Algebra.

For pedagogical reasons, the examples and exercise that have been used to illustrate the algorithms are all based on rather small (2×2 or 3×3) matrices. In such cases, or even for matrices of moderate size, the differences between the various approaches to solving linear systems (Gauss, Gauss–Jordan, matrix inverse, etc.) are relatively unimportant, particularly if one has a decent computer or even hand calculator to perform the tedious parts. However, real-world applied mathematics deals with much larger linear systems, and the design of efficient algorithms is critical. For example, numerical solutions of ordinary differential equations will typically lead to matrices with hundreds or thousands of entries, while numerical solution of partial differential equations arising in fluid and solid mechanics, weather prediction, image and video processing, chemical reactions, quantum mechanics, molecular dynamics, and many other areas will often lead to matrices with millions of entries. It is not hard for such systems to tax even the most sophisticated supercomputer. Thus, it is essential that we look into the computational details of competing algorithms in order to compare their efficiency, and thereby gain some experience with the issues underlying the design of high performance numerical algorithms.

The most basic question is: how many arithmetic operations are required for each of our algorithms? We shall keep track of additions and multiplications separately, since the latter typically take slightly longer to perform in a computer processor. However, we shall not distinguish between addition and subtraction, nor between multiplication and division, as these typically rely on the same floating point algorithm. We shall also assume that the matrices and vectors are *generic*, with few, if any, zero entries. Modifications of the basic algorithms for *sparse matrices*, meaning those that have lots of zero entries, are an important topic of research, since these include many of the large matrices that appear in applications to differential equations. We refer the interested reader to more advanced numerical linear algebra texts, e.g., [**121**, **119**], for further developments.

First, for ordinary multiplication of an $n \times n$ matrix A and a vector $\mathbf{b}$, each entry of the product $A\,\mathbf{b}$ requires n multiplications of the form $a_{ij}\,b_j$ and $n-1$ additions to sum the resulting products. Since there are n entries, this means a total of n^2 multiplications and $n(n-1) = n^2 - n$ additions. Thus, for a matrix of size $n = 100$, one needs about $10,000$ distinct multiplications and a similar (but slightly fewer) number of additions. If $n = 1,000,000 = 10^6$ then $n^2 = 10^{12}$, which is phenomenally large, and the total time required to perform the computation becomes a significant issue[†].

Let us look at the regular Gaussian Elimination algorithm, referring back to our program. First, we count how many arithmetic operations are based on the j^{th} pivot m_{jj}. For each of the $n-j$ rows lying below it, we must perform one division to compute the factor $l_{ij} = m_{ij}/m_{jj}$ used in the elementary row operation. The entries in the column below the pivot will be made equal to zero automatically, and so we need only compute the updated entries lying below and to the right of the pivot. There are $(n-j)^2$ such entries in the coefficient matrix and an additional $n-j$ entries in the last column of the augmented matrix. Let us concentrate on the former for the moment. For each of these, we replace

[†] See [**31**] for more sophisticated computational methods to speed up matrix multiplication.

m_{ik} by $m_{ik} - l_{ij}\, m_{jk}$, and so must perform one multiplication and one addition. Therefore, for the j^{th} pivot there are a total of $(n-j)(n-j+1)$ multiplications — including the initial $n-j$ divisions needed to produce the l_{ij} — and $(n-j)^2$ additions needed to update the coefficient matrix. Therefore, to reduce a regular $n \times n$ matrix to upper triangular form requires a total[‡] of

$$\sum_{j=1}^{n} (n-j)(n-j+1) = \frac{n^3 - n}{3} \qquad \text{multiplications, and} \tag{1.57}$$

$$\sum_{j=1}^{n} (n-j)^2 = \frac{2\,n^3 - 3\,n^2 + n}{6} \qquad \text{additions.} \tag{1.58}$$

Thus, when n is large, both require approximately $\frac{1}{3}\, n^3$ operations.

We should also be keeping track of the number of operations on the right hand side of the system. No pivots appear there, and so there are

$$\sum_{j=1}^{n} (n-j) = \frac{n^2 - n}{2} \tag{1.59}$$

multiplications and the same number of additions required to produce the right hand side in the resulting triangular system $U\mathbf{x} = \mathbf{c}$. For large n, this number is considerably smaller than the coefficient matrix totals (1.57), (1.58).

The next phase of the algorithm can be similarly analyzed. To find the value of

$$x_j = \frac{1}{u_{jj}} \left(c_j - \sum_{i=j+1}^{n} u_{ji}\, x_i \right)$$

once we have computed $x_{j+1}, \dots, x_n$, requires $n-j+1$ multiplications/divisions and $n-j$ additions. Therefore, the Back Substitution phase of the algorithm requires

$$\sum_{j=1}^{n} (n-j+1) = \frac{n^2 + n}{2} \ \text{ multiplications, and } \ \sum_{j=1}^{n} (n-j) = \frac{n^2 - n}{2} \ \text{ additions.} \tag{1.60}$$

For n large, both of these are approximately equal to $\frac{1}{2}\, n^2$. Comparing these results, we conclude that the bulk of the computational effort goes into the reduction of the coefficient matrix to upper triangular form.

Forward substitution, to solve $L\mathbf{c} = \mathbf{b}$, has the same operations count, except that since the diagonal entries of L are all equal to 1, no divisions are required, and so we use a total of $\frac{1}{2}(n^2 - n)$ multiplications and the same number of additions. Thus, once we have computed the LU decomposition of the matrix A, the Forward and Back Substitution process requires about n^2 arithmetic operations of the two types, which is the *same* as the

[‡] In Exercise ■ the reader is asked to prove these summation formulae by induction.

number of operations needed to perform the matrix multiplication $A^{-1}\mathbf{b}$. Thus, even if we know the inverse of the coefficient matrix A, it is still just as efficient to use Forward and Back Substitution to compute the solution!

As noted above, the computation of L and U requires about $\frac{1}{3}n^3$ arithmetic operations of each type. On the other hand, to complete the full-blown Gauss–Jordan elimination scheme, we must perform all the elementary row operations on the large augmented matrix, which has size $n \times 2n$. Therefore, during the reduction to upper triangular form, there are an additional $\frac{1}{2}n^3$ operations of each type required. Moreover, we then need to perform an additional $\frac{1}{3}n^3$ operations to reduce U to the identity matrix, and a corresponding $\frac{1}{2}n^3$ operations on the right hand matrix, too. (All these are approximate totals, based on the leading term in the actual count.) Therefore, Gauss–Jordan requires a grand total of $\frac{5}{3}n^3$ operations to complete, just to find A^{-1}; multiplying the right hand side to obtain the solution $\mathbf{x} = A^{-1}\mathbf{b}$ involves another n^2 operations. Thus, the Gauss–Jordan method requires approximately *five* times as many arithmetic operations, and so would take five times as long to complete, as compared to the more elementary Gaussian Elimination and Back Substitution algorithm. These observations serve to justify our earlier contention that matrix inversion is inefficient, and should never be used to solve linear systems in practice.

Tridiagonal Matrices

Of course, in special cases, the arithmetic operation count might be considerably reduced, particularly if A is a sparse matrix with many zero entries. A number of specialized techniques have been designed to handle such sparse linear systems. A particularly important class are the *tridiagonal matrices*

$$A = \begin{pmatrix} q_1 & r_1 & & & & \\ p_1 & q_2 & r_2 & & & \\ & p_2 & q_3 & r_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & p_{n-2} & q_{n-1} & r_{n-1} \\ & & & & p_{n-1} & q_n \end{pmatrix} \tag{1.61}$$

with all entries zero except for those on the main diagonal, $a_{i,i} = q_i$, on the *subdiagonal*, meaning the $n-1$ entries $a_{i+1,i} = p_i$ immediately below the main diagonal, and the *superdiagonal*, meaning the entries $a_{i,i+1} = r_i$ immediately above the main diagonal. (Zero entries are left blank.) Such matrices arise in the numerical solution of ordinary differential equations and the spline fitting of curves for interpolation and computer graphics. If

$A = LU$ is regular, it turns out that the factors are lower and upper *bidiagonal matrices*,

$$L = \begin{pmatrix} 1 & & & & & \\ l_1 & 1 & & & & \\ & l_2 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & l_{n-2} & 1 & \\ & & & & l_{n-1} & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} d_1 & u_1 & & & & \\ & d_2 & u_2 & & & \\ & & d_3 & u_3 & & \\ & & & \ddots & \ddots & \\ & & & & d_{n-1} & u_{n-1} \\ & & & & & d_n \end{pmatrix}. \tag{1.62}$$

Multiplying out LU, and equating the result to A leads to the equations

$$\begin{aligned} & d_1 = q_1, && u_1 = r_1, && l_1 d_1 = p_1, \\ & l_1 u_1 + d_2 = q_2, && u_2 = r_2, && l_2 d_2 = p_2, \\ & \quad \vdots && \quad \vdots && \quad \vdots \\ & l_{j-1} u_{j-1} + d_j = q_j, && u_j = r_j, && l_j d_j = p_j, \\ & \quad \vdots && \quad \vdots && \quad \vdots \\ & l_{n-2} u_{n-2} + d_{n-1} = q_{n-1}, && u_{n-1} = r_{n-1}, && l_{n-1} d_{n-1} = p_{n-1}, \\ & l_{n-1} u_{n-1} + d_n = q_n. \end{aligned} \tag{1.63}$$

These elementary algebraic equations can be successively solved for the entries of L and U in the order $d_1, u_1, l_1, d_2, u_2, l_2, d_3, u_3 \dots$. The original matrix A is regular provided none of the diagonal entries $d_1, d_2, \dots$ are zero, which allows the recursive procedure to proceed.

Once the LU factors are in place, we can apply Forward and Back Substitution to solve the tridiagonal linear system $A\mathbf{x} = \mathbf{b}$. We first solve $L\mathbf{c} = \mathbf{b}$ by Forward Substitution, which leads to the recursive equations

$$c_1 = b_1, \qquad c_2 = b_2 - l_1 c_1, \qquad \dots \qquad c_n = b_n - l_{n-1} c_{n-1}. \tag{1.64}$$

We then solve $U\mathbf{x} = \mathbf{c}$ by Back Substitution, again recursively:

$$x_n = \frac{c_n}{d_n}, \qquad x_{n-1} = \frac{c_{n-1} - u_{n-1} x_n}{d_{n-1}}, \qquad \dots \qquad x_1 = \frac{c_1 - u_1 x_2}{d_1}. \tag{1.65}$$

As you can check, there are a total of $5n - 4$ multiplications/divisions and $3n - 3$ additions/subtractions required to solve a general tridiagonal system of n linear equations — a striking improvement over the general case.

Example 1.35. Consider the $n \times n$ tridiagonal matrix

$$A = \begin{pmatrix} 4 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & 1 & 4 & 1 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & 1 & 4 & 1 \\ & & & & & 1 & 4 \end{pmatrix}$$

in which the diagonal entries are all $q_i = 4$, while the entries immediately above and below the main diagonal are all $p_i = r_i = 1$. According to (1.63), the tridiagonal factorization (1.62) has $u_1 = u_2 = \ldots = u_{n-1} = 1$, while

$$d_1 = 4, \qquad l_j = 1/d_j, \qquad d_{j+1} = 4 - l_j, \qquad j = 1, 2, \ldots, n-1.$$

The computed values are

j	1	2	3	4	5	6	7
d_j	4	3.75	3.733333	3.732143	3.732057	3.732051	3.732051
l_j	.25	.266666	.267857	.267942	.267948	.267949	.267949

These converge rapidly to

$$d_j \longrightarrow 2 + \sqrt{3} = 3.732051\ldots, \qquad l_j \longrightarrow 2 - \sqrt{3} = .2679492\ldots,$$

which makes the factorization for large n almost trivial. The numbers $2 \pm \sqrt{3}$ are the roots of the quadratic equation $x^2 - 4x + 1 = 0$; an explanation of this observation will be revealed in Chapter 19.

Pivoting Strategies

Let us now consider the practical side of pivoting. As we know, in the irregular situations when a zero shows up in a diagonal pivot position, a row interchange is required to proceed with the elimination algorithm. But even when a nonzero element appear in the current pivot position, there may be good numerical reasons for exchanging rows in order to install a more desirable element in the pivot position. Here is a simple example:

$$.01\,x + 1.6\,y = 32.1, \qquad x + .6\,y = 22. \tag{1.66}$$

The exact solution to the system is $x = 10$, $y = 20$. Suppose we are working with a very primitive calculator that only retains 3 digits of accuracy. (Of course, this is not a very realistic situation, but the example could be suitably modified to produce similar difficulties no matter how many digits of accuracy our computer retains.) The augmented matrix is

$$\left(\begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 1 & .6 & 22 \end{array}\right).$$

Choosing the $(1,1)$ entry as our pivot, and subtracting 100 times the first row from the second produces the upper triangular form

$$\left(\begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 0 & -159.4 & -3188 \end{array}\right).$$

Since our calculator has only three–place accuracy, it will round the entries in the second row, producing the augmented coefficient matrix

$$\left(\begin{array}{cc|c} .01 & 1.6 & 32.1 \\ 0 & -159.0 & -3190 \end{array}\right).$$

Gaussian Elimination With Partial Pivoting

```
start
    for i = 1 to n
        set σ(i) = i
    next i
    for j = 1 to n
        if m_{σ(i),j} = 0 for all i ≥ j, stop; print "A is singular"
        choose i > j such that m_{σ(i),j} is maximal
        interchange σ(i) ⟷ σ(j)
        for i = j + 1 to n
            set l_{σ(i)j} = m_{σ(i)j} / m_{σ(j)j}
            for k = j + 1 to n + 1
                set m_{σ(i)k} = m_{σ(i)k} − l_{σ(i)j} m_{σ(j)k}
            next k
        next i
    next j
end
```

The solution by back substitution gives $y = \dfrac{3190}{159} = 20.0628\ldots \simeq 20.1$, and then $x = 100\,(32.1 - 1.6\,y) = 100\,(32.1 - 32.16) \simeq 100\,(32.1 - 32.2) = -10$. The relatively small error in y has produced a very large error in x — not even its sign is correct!

The problem is that the first pivot, .01, is much smaller than the other element, 1, that appears in the column below it. Interchanging the two rows before performing the row operation would resolve the difficulty — even with such an inaccurate calculator! After the interchange, we have

$$\left(\begin{array}{cc|c} 1 & .6 & 22 \\ .01 & 1.6 & 32.1 \end{array}\right),$$

which results in the rounded-off upper triangular form

$$\left(\begin{array}{cc|c} 1 & .6 & 22 \\ 0 & 1.594 & 31.88 \end{array}\right) \simeq \left(\begin{array}{cc|c} 1 & .6 & 22 \\ 0 & 1.59 & 31.9 \end{array}\right).$$

The solution by back substitution now gives a respectable answer:

$$y = 31.9/1.59 = 20.0628\ldots \simeq 20.1, \qquad x = 22 - .6\,y = 22 - 12.06 \simeq 22 - 12.1 = 9.9.$$

The general strategy, known as *Partial Pivoting*, says that at each stage, we should use the largest legitimate (i.e., lying on or below the diagonal) element as the pivot, even

if the diagonal element is nonzero. In a computer implementation of pivoting, there is no need to waste processor time physically exchanging the row entries in memory. Rather, one introduces a separate array of pointers that serve to indicate which original row is currently in which permuted position. More specifically, one initializes n row pointers $\sigma(1) = 1, \dots, \sigma(n) = n$. Interchanging row i and row j of the coefficient or augmented matrix is then accomplished by merely interchanging $\sigma(i)$ and $\sigma(j)$. Thus, to access a matrix element that is currently in row i of the augmented matrix, one merely retrieves the element that is in row $\sigma(i)$ in the computer's memory. An explicit implementation of this strategy is provided below. A program for partial pivoting that includes row pointers appears above.

Partial pivoting will solve most problems, although there can still be difficulties. For instance, it will not handle the system

$$10\,x + 1600\,y = 3210, \qquad x + .6\,y = 22,$$

obtained by multiplying the first equation in (1.66) by 1000. The tip-off is that, while the entries in the column containing the pivot are smaller, those in its row are much larger. The solution to this difficulty is Full Pivoting, in which one also performs column interchanges — preferably with a column pointer — to move the largest legitimate element into the pivot position. In practice, a column interchange is just a reordering of the variables in the system, which, as long as one keeps proper track of the order, also doesn't change the solutions.

Finally, there are some matrices that are hard to handle even with pivoting tricks. Such *ill-conditioned* matrices are typically characterized by being "almost" singular[†]. A famous example of an ill-conditioned matrix is the $n \times n$ *Hilbert matrix*

$$H_n = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \cdots & \frac{1}{n+1} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \cdots & \frac{1}{n+2} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \cdots & \frac{1}{n+3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \frac{1}{n+2} & \frac{1}{n+3} & \cdots & \frac{1}{2n-1} \end{pmatrix}. \tag{1.67}$$

In Proposition 3.36 we will prove that H_n is nonsingular for all n. However, the solution of a linear system whose coefficient matrix is a Hilbert matrix H_n, even for moderately sized n, is a very challenging problem, even if one uses high precision computer arithmetic[‡].

[†] This can be quantified by saying that their determinant is very small, but non-zero; see also Sections 8.5 and 10.3.

[‡] In computer algebra systems such as MAPLE or MATHEMATICA, one can use exact rational arithmetic to perform the computations. Then the important issues are time and computational efficiency.

This is because the larger n is, the closer H_n is, in a sense, to being singular.

The reader is urged to try the following computer experiment. Fix a moderately large value of n, say 20. Choose a column vector $\mathbf{x}$ with n entries chosen at random. Compute $\mathbf{b} = H_n \mathbf{x}$ directly. Then try to solve the system $H_n \mathbf{x} = \mathbf{b}$ by Gaussian Elimination. If it works for $n = 20$, try $n = 50$ or 100. This will give you a good indicator of the degree of precision used by your computer program, and the accuracy of the numerical solution algorithm.

1.8. General Linear Systems.

So far, we have only treated linear systems involving the same number of equations as unknowns, and then only those with nonsingular coefficient matrices. These are precisely the systems that always have a unique solution. We now turn to the problem of solving a general linear system of m equations in n unknowns. The cases not covered as yet are rectangular systems, with $m \neq n$, as well as square systems with singular coefficient matrices. The basic idea underlying the Gaussian Elimination Algorithm for nonsingular systems can be straightforwardly adapted to these cases, too. One systematically utilizes the same elementary row operations so as to manipulate the coefficient matrix into a particular reduced form generalizing the upper triangular form we aimed for in the earlier square, nonsingular cases.

Definition 1.36. An $m \times n$ matrix is said to be in *row echelon form* if it has the following "staircase" structure:

$$U = \begin{pmatrix} \circledast & * & \dots & * & * & \dots & * & * & \dots & \dots & * & * & * & \dots & * \\ 0 & 0 & \dots & 0 & \circledast & \dots & * & * & \dots & \dots & * & * & * & \dots & * \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & \circledast & \dots & \dots & * & * & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & & \ddots & & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & \circledast & * & \dots & * \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & \dots & 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

The entries indicated by $\circledast$ are the pivots, and must be nonzero. The first r rows of U each contain one pivot, but not all columns need to contain a pivot. The entries below the "staircase", indicated by the solid line, are all zero, while the non-pivot entries above the staircase, indicated by stars, can be anything. The last $m-r$ rows are identically zero, and do not contain any pivots. Here is an explicit example of a matrix in row echelon form:

$$\begin{pmatrix} 3 & 1 & 0 & 2 & 5 & -1 \\ 0 & -1 & -2 & 1 & 8 & 0 \\ 0 & 0 & 0 & 0 & 2 & -4 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The three pivots, which are the first three nonzero entries in the nonsero rows, are, respectively, $3, -1, 2$. There may, in exceptional situations, be one or more initial all zero columns.

Proposition 1.37. *Any matrix can be reduced to row echelon form by a sequence of elementary row operations of Types #1 and #2.*

In matrix language, Proposition 1.37 implies that if A is any $m \times n$ matrix, then there exists an $m \times m$ permutation matrix P and an $m \times m$ special lower triangular matrix L such that

$$PA = LU, \tag{1.68}$$

where U is in row echelon form. The factorization is not unique.

A constructive proof of this result is based on the general Gaussian elimination algorithm, which proceeds as follows. Starting at the top left of the matrix, one searches for the first column which is not identically zero. Any of the nonzero entries in that column may serve as the pivot. Partial pivoting indicates that it is probably best to choose the largest one, although this is not essential for the algorithm to proceed. One places the chosen pivot in the first row of the matrix via a row interchange, if necessary. The entries below the pivot are made equal to zero by the appropriate elementary row operations of Type #1. One then proceeds iteratively, performing the same reduction algorithm on the submatrix consisting of all entries strictly to the right and below the pivot. The algorithm terminates when either there is a pivot in the last row, or all of the rows lying below the last pivot are identically zero, and so no more pivots can be found.

Example 1.38. Let us illustrate the general Gaussian Elimination algorithm with a particular example. Consider the linear system

$$\begin{aligned} x + 3y + 2z - u &= a, \\ 2x + 6y + z + 4u + 3v &= b, \\ -x - 3y - 3z + 3u + v &= c, \\ 3x + 9y + 8z - 7u + 2v &= d, \end{aligned} \tag{1.69}$$

of 4 equations in 5 unknowns, where a, b, c, d are specified numbers[†]. The coefficient matrix is

$$A = \begin{pmatrix} 1 & 3 & 2 & -1 & 0 \\ 2 & 6 & 1 & 4 & 3 \\ -1 & -3 & -3 & 3 & 1 \\ 3 & 9 & 8 & -7 & 2 \end{pmatrix}. \tag{1.70}$$

To solve the system, we introduce the augmented matrix

$$\left(\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 2 & 6 & 1 & 4 & 3 & b \\ -1 & -3 & -3 & 3 & 1 & c \\ 3 & 9 & 8 & -7 & 2 & d \end{array}\right)$$

[†] It will be convenient to work with the right hand side in general form, although the reader may prefer, at least initially, to assign specific values to a, b, c, d.

obtained by appending the right hand side of the system. The upper left entry is nonzero, and so can serve as the first pivot; we eliminate the entries below it by elementary row operations, resulting in

$$\left(\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b-2a \\ 0 & 0 & -1 & 2 & 1 & c+a \\ 0 & 0 & 2 & -4 & 2 & d-3a \end{array}\right).$$

Now, the second column contains no suitable nonzero entry to serve as the second pivot. (The top entry already lies in a row with a pivot in it, and so cannot be used.) Therefore, we move on to the third column, choosing the $(2,3)$ entry, -3, as our second pivot. Again, we eliminate the entries below it, leading to

$$\left(\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b-2a \\ 0 & 0 & 0 & 0 & 0 & c-\frac{1}{3}b+\frac{5}{3}a \\ 0 & 0 & 0 & 0 & 4 & d+\frac{2}{3}b-\frac{13}{3}a \end{array}\right).$$

The final pivot is in the last column, and we interchange the last two rows in order to place the coefficient matrix in row echelon form:

$$\left(\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b-2a \\ 0 & 0 & 0 & 0 & 4 & d+\frac{2}{3}b-\frac{13}{3}a \\ 0 & 0 & 0 & 0 & 0 & c-\frac{1}{3}b+\frac{5}{3}a \end{array}\right). \tag{1.71}$$

There are three pivots, $-1, -3, 4$, sitting in positions $(1,1)$, $(2,3)$ and $(3,5)$. Note the staircase form, with the pivots on the steps and everything below the staircase being zero. Recalling the row operations used to construct the solution (and keeping in mind that the row interchange that appears at the end also affects the entries of L), we find the factorization (1.68) has the explicit form

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 3 & 2 & -1 & 0 \\ 2 & 6 & 1 & 4 & 3 \\ -1 & -3 & -3 & 3 & 1 \\ 3 & 9 & 8 & -7 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -\frac{2}{3} & 1 & 0 \\ -1 & \frac{1}{3} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 & 2 & -1 & 0 \\ 0 & 0 & -3 & 6 & 3 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

We shall return to find the solution to our system after a brief theoretical interlude.

Warning: In the augmented matrix, pivots can *never* appear in the last column, representing the right hand side of the system. Thus, even if $c - \frac{1}{3}b + \frac{5}{3}a \neq 0$, that entry does not qualify as a pivot.

We now introduce the most important numerical quantity associated with a matrix.

Definition 1.39. The *rank* of a matrix A is the number of pivots.

For instance, the rank of the matrix (1.70) equals 3, since its reduced row echelon form, i.e., the first five columns of (1.71), has three pivots. Since there is at most one pivot per row and one pivot per column, the rank of an $m \times n$ matrix is bounded by both m and n, and so $0 \leq r \leq \min\{m, n\}$. The only matrix of rank 0 is the zero matrix, which has no pivots.

Proposition 1.40. *A square matrix of size $n \times n$ is nonsingular if and only if its rank is equal to n.*

Indeed, the only way an $n \times n$ matrix can end up having n pivots is if its reduced row echelon form is upper triangular with nonzero diagonal entries. But a matrix that reduces to such triangular form is, by definition, nonsingular.

Interestingly, the rank of a matrix *does not depend* on which elementary row operations are performed along the way to row echelon form. Indeed, performing a different sequence of row operations — say using partial pivoting versus no pivoting — can produce a completely different row echelon form. The remarkable fact, though, is that all such row echelon forms end up having exactly the same number of pivots, and this number is the rank of the matrix. A formal proof of this fact will appear in Chapter 2.

Once the coefficient matrix has been reduced to row echelon form, the solution proceeds as follows. The first step is to see if there are any incompatibilities. Suppose one of the rows in the row echelon form of the coefficient matrix is identically zero, but the corresponding entry in the last column of the augmented matrix is nonzero. What linear equation would this represent? Well, the coefficients of all the variables are zero, and so the equation is of the form $0 = c$, where c, the number on the right hand side of the equation, is the entry in the last column. If $c \neq 0$, then the equation cannot be satisfied. Consequently, the entire system has *no solutions*, and is an *incompatible* linear system. On the other hand, if $c = 0$, then the equation is merely $0 = 0$, and so is trivially satisfied. For example, the last row in the echelon form (1.71) is all zero, and hence the last entry in the final column must also vanish in order that the system be compatible. Therefore, the linear system (1.69) will have a solution if and only if the right hand sides a, b, c, d satisfy the linear constraint

$$\tfrac{5}{3}\, a - \tfrac{1}{3}\, b + c = 0. \tag{1.72}$$

In general, if the system is incompatible, there is nothing else to do. Otherwise, every zero row in the echelon form of the augmented matrix also has a zero entry in the last column, and the system is *compatible*, so one or more solutions exist. To find the solution(s), we work backwards, starting with the last row that contains a pivot. The variables in the system naturally split into two classes.

Definition 1.41. In a linear system $U\mathbf{x} = \mathbf{c}$ in row echelon form, the variables corresponding to columns containing a pivot are called *basic variables*, while the variables corresponding to the columns without a pivot are called *free variables*.

The solution to the system proceeds by a version of the Back Substitution procedure. The nonzero equations are solved, in reverse order, for the basic variable corresponding to its pivot. Each result is substituted into the preceding equations before they in turn are

solved. The remaining free variables, if any, are allowed to take on any values whatsoever, and the solution then specifies all the basic variables in terms of the free variables, which serve to parametrize the general solution.

Example 1.42. Let us illustrate this construction with our particular example. Assuming the compatibility condition (1.72), the reduced augmented matrix (1.71) is

$$\left(\begin{array}{ccccc|c} 1 & 3 & 2 & -1 & 0 & a \\ 0 & 0 & -3 & 6 & 3 & b - 2\,a \\ 0 & 0 & 0 & 0 & 4 & d + \frac{2}{3}\,b - \frac{13}{3}\,a \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

The pivots are found in columns $1, 3, 5$, and so the corresponding variables, x, z, v, are basic; the other variables, y, u, are free. We will solve the reduced system for the basic variables in terms of the free variables.

As a specific example, the values $a = 0$, $b = 3$, $c = 1$, $d = 1$, satisfy the compatibility constraint (1.72). The resulting augmented echelon matrix (1.71) corresponds to the system

$$\begin{aligned} x + 3\,y + 2\,z - u &= 0, \\ -3\,z + 6\,u + 3\,v &= 3, \\ 4\,v &= 3, \\ 0 &= 0. \end{aligned}$$

We now solve the equations, in reverse order, for the basic variables, and then substitute the resulting values in the preceding equations. The result is the general solution

$$v = \tfrac{3}{4}, \qquad z = -1 + 2u + v = -\tfrac{1}{4} + 2\,u, \qquad x = -3\,y - 2\,z + u = \tfrac{1}{2} - 3\,y - 3\,u.$$

The free variables y, u are completely arbitrary; any value they assume will produce a solution to the original system. For instance, if $y = -2, u = 1 - \pi$, then $x = 3\pi + \frac{7}{2}$, $z = \frac{7}{4} - 2\,\pi$, $v = \frac{3}{4}$. But keep in mind that this is merely one of an infinite number of different solutions.

In general, if the $m \times n$ coefficient matrix of a system of m linear equations in n unknowns has rank r, there are $m - r$ all zero rows in the row echelon form, and these $m - r$ equations must have zero right hand side in order that the system be compatible and have a solution. Moreover, there are a total of r basic variables and $n - r$ free variables, and so the general solution depends upon $n - r$ parameters.

Summarizing the preceding discussion, we have learned that there are only three possible outcomes for the solution to a general linear system.

Theorem 1.43. *A system $A\,\mathbf{x} = \mathbf{b}$ of m linear equations in n unknowns has either (i) exactly one solution, (ii) no solutions, or (iii) infinitely many solutions.*

Case (*ii*) occurs if the system is incompatible, producing a zero row in the echelon form that has a nonzero right hand side. Case (*iii*) occurs if the system is compatible and there are one or more free variables. This happens when the system is compatible and the rank of the coefficient matrix is strictly less than the number of columns: $r < n$. Case

(i) occurs for nonsingular square coefficient matrices, and, more generally, for compatible systems for which $r = n$, implying there are no free varaibles. Since $r \leq m$, this case can only arise if the coefficient matrix has at least as many rows as columns, i.e., the linear system has at least as many equations as unknowns.

A linear system can *never* have a finite number — other than 0 or 1 — of solutions. Thus, any linear system that has more than one solution automatically has infinitely many. This result does *not* apply to nonlinear systems. As you know, a real quadratic equation $a\,x^2 + b\,x + c = 0$ can have either 2, 1, or 0 real solutions.

Example 1.44. Consider the linear system

$$y + 4\,z = a, \qquad 3\,x - y + 2\,z = b, \qquad x + y + 6\,z = c,$$

consisting of three equations in three unknowns. The augmented coefficient matrix is

$$\left(\begin{array}{ccc|c} 0 & 1 & 4 & a \\ 3 & -1 & 2 & b \\ 1 & 1 & 6 & c \end{array}\right).$$

Interchanging the first two rows, and then eliminating the elements below the first pivot leads to

$$\left(\begin{array}{ccc|c} 3 & -1 & 2 & b \\ 0 & 1 & 4 & a \\ 0 & \frac{4}{3} & \frac{16}{3} & c - \frac{1}{3}\,b \end{array}\right).$$

The second pivot is in the $(2,2)$ position, but after eliminating the entry below it, we find the row echelon form to be

$$\left(\begin{array}{ccc|c} 3 & -1 & 2 & b \\ 0 & 1 & 4 & a \\ 0 & 0 & 0 & c - \frac{1}{3}\,b - \frac{4}{3}\,a \end{array}\right).$$

Since we have a row of all zeros, the original coefficient matrix is singular, and its rank is only 2.

The compatibility condition for the system follows from this last row in the reduced echelon form, and so requires

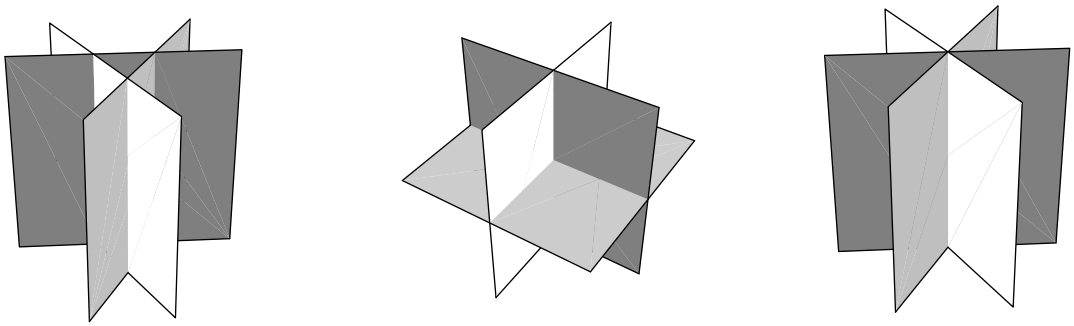
$$\tfrac{4}{3}\,a + \tfrac{1}{3}\,b - c = 0.$$

If this is not satisfied, the system has no solutions; otherwise it has infinitely many. The free variable is z, since there is no pivot in the third column. The general solution is

$$y = a - 4\,z, \qquad x = \tfrac{1}{3}\,b + \tfrac{1}{3}\,y - \tfrac{2}{3}\,z = \tfrac{1}{3}\,a + \tfrac{1}{3}\,b - 2z,$$

where z is arbitrary.

Geometrically, Theorem 1.43 is indicating something about the possible configurations of linear subsets (lines, planes, etc.) of an n-dimensional space. For example, a single linear equation $a\,x + b\,y + c\,z = d$ defines a plane P in three-dimensional space. The solutions to a system of three linear equations in three unknowns is the *intersection* $P_1 \cap P_2 \cap P_3$ of three planes. Generically, three planes intersect in a single common point; this is case (i)



No Solution Unique Solution Infinite # Solutions

Figure 1.1. Intersecting Planes.

of the theorem, and occurs if and only if the coefficient matrix is nonsingular. The case of infinitely many solutions occurs when the three planes intersect on a common line, or, even more degenerately, when they all coincide. On the other hand, parallel planes, or planes intersecting in parallel lines, have no common point of intersection, and this corresponds to the third case of a system with no solutions. Again, no other possibilities occur; clearly one cannot have three planes having exactly 2 points in their common intersection — it is either 0, 1 or ∞. Some possible geometric configurations are illustrated in Figure 1.1.

Homogeneous Systems

A linear system with all 0's on the right hand side is called a *homogeneous system*. In matrix notation, a homogeneous system takes the form

$$A\,\mathbf{x} = \mathbf{0}. \tag{1.73}$$

Homogeneous systems are always compatible, since $\mathbf{x} = \mathbf{0}$ is a solution, known as the *trivial solution*. If the homogeneous system has a nontrivial solution $\mathbf{x} \neq \mathbf{0}$, then Theorem 1.43 assures that it must have infinitely many solutions. This will occur if and only if the reduced system has one or more free variables. Thus, we find:

Theorem 1.45. *A homogeneous linear system $A\,\mathbf{x} = \mathbf{0}$ of m equations in n unknowns has a nontrivial solution $\mathbf{x} \neq 0$ if and only if the rank of A is $r < n$. If $m < n$, the system always has a nontrivial solution. If $m = n$, the system has a nontrivial solution if and only if A is singular.*

Example 1.46. Consider the homogeneous linear system

$$2x_1 + x_2 + 5x_4 = 0, \qquad 4x_1 + 2x_2 - x_3 + 8x_4 = 0, \qquad -2x_1 - x_2 + 3x_3 - 4x_4 = 0,$$

with coefficient matrix

$$A = \begin{pmatrix} 2 & 1 & 0 & 5 \\ 4 & 2 & -1 & 8 \\ -2 & -1 & 3 & -4 \end{pmatrix}.$$

Since the system is homogeneous and has fewer equations than unknowns, Theorem 1.45 assures us that it has infinitely many solutions, including the trivial solution $x_1 = x_2 =$

$x_3 = x_4 = 0$. When solving a homogeneous system, the final column of the augmented matrix consists of all zeros. As it will never be altered by row operations, it is a waste of effort to carry it along during the process. We therefore perform the Gaussian Elimination algorithm directly on the coefficient matrix A. Working with the $(1,1)$ entry as the first pivot, we first obtain

$$\begin{pmatrix} 2 & 1 & 0 & 5 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 3 & 1 \end{pmatrix}.$$

The $(2,3)$ entry is the second pivot, and we apply one final row operation to place the matrix in row echelon form

$$\begin{pmatrix} 2 & 1 & 0 & 5 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & 0 & -5 \end{pmatrix}.$$

This corresponds to the reduced homogeneous system

$$2\,x_1 + x_2 + 5\,x_4 = 0, \qquad -x_3 - 2\,x_4 = 0, \qquad -5\,x_4 = 0.$$

Since there are three pivots in the final row echelon form, the rank of the matrix A is 3. There is one free variable, namely x_2. Using Back Substitution, we easily obtain the general solution

$$x_1 = -\tfrac{1}{2}\,t, \qquad x_2 = t, \qquad x_3 = x_4 = 0,$$

which depends upon a single free parameter $t = x_2$.

Example 1.47. Consider the homogeneous linear system

$$2\,x - y + 3\,z = 0, \qquad -4\,x + 2\,y - 6\,z = 0, \qquad 2\,x - y + z = 0, \qquad 6\,x - 3\,y + 3\,z = 0,$$

with coefficient matrix $A = \begin{pmatrix} 2 & -1 & 3 \\ -4 & 2 & -6 \\ 2 & -1 & 1 \\ 6 & -3 & 3 \end{pmatrix}$. The system admits the trivial solution $x = y = z = 0$, but in this case we need to complete the elimination algorithm before we can state whether or not there are other solutions. After the first stage, the coefficient matrix has the form $\begin{pmatrix} 2 & -1 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & -6 \end{pmatrix}$. To continue, we need to interchange the second and third rows to place a nonzero entry in the final pivot position; after that the reduction to row echelon form is immediate: $\begin{pmatrix} 2 & -1 & 3 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & -6 \end{pmatrix} \longmapsto \begin{pmatrix} 2 & -1 & 3 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. Thus, the system reduces to the equations

$$2\,x - y + 3\,z = 0, \qquad -2\,z = 0, \qquad 0 = 0, \qquad 0 = 0,$$

where the third and fourth equations are trivially compatible, as they must be in the homogeneous case. The rank is equal to two, which is less than the number of columns, and so, even though the system has more equations than unknowns, it has infinitely many solutions. These can be written in terms of the free variable y, and so the general solution is $x = \frac{1}{2}\, y$, $z = 0$, where y is arbitrary.

1.9. Determinants.

You may be surprised that, so far, we have left undeveloped a topic that often assumes a central role in basic linear algebra: determinants. As with matrix inverses, while determinants can be useful in low dimensions and for theoretical purposes, they are mostly irrelevant when it comes to large scale applications and practical computations. Indeed, the best way to compute a determinant is (surprise) Gaussian Elimination! However, you should be familiar with the basics of determinants, and so for completeness, we shall provide a very brief introduction.

The determinant of a square matrix A, written $\det A$, is a number that immediately tells whether the matrix is singular or not. (Rectangular matrices do not have determinants.) We already encountered, (1.34), the determinant of a 2×2 matrix, which is equal to the product of the diagonal entries minus the product of the off-diagonal entries: $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\,d - b\,c$. The determinant is nonzero if and only if the matrix has an inverse. Our goal is to generalize this construction to general square matrices.

There are many different ways to define determinants. The difficulty is that the actual formula is very unwieldy — see (1.81) below — and not well motivated. We prefer an axiomatic approach that explains how our elementary row operations affect the determinant. In this manner, one can compute the determinant by Gaussian elimination, which is, in fact, the fastest and most practical computational method in all but the simplest situations. In effect, this remark obviates the need to ever compute a determinant.

Theorem 1.48. *The* determinant *of a square matrix A is the uniquely defined scalar quantity* $\det A$ *that satisfies the following axioms:*

(1) *Adding a multiple of one row to another does not change the determinant.*

(2) *Interchanging two rows changes the sign of the determinant.*

(3) *Multiplying a row by any scalar (including zero) multiplies the determinant by the same scalar.*

(4) *Finally, the determinant function is fixed by setting*

$$\det \mathrm{I} = 1. \tag{1.74}$$

Checking that all four of these axioms hold in the 2×2 case (1.34) is left as an elementary exercise for the reader. A particular consequence of axiom 3 is that when we multiply a row of any matrix A by the zero scalar, the resulting matrix, which has a row of all zeros, necessarily has zero determinant.

Lemma 1.49. *Any matrix with one or more all zero rows has zero determinant.*

Since the determinantal axioms tell how determinants behave under all three of our elementary row operations, we can use Gaussian elimination to compute a general determinant, recovering $\det A$ from its permuted LU factorization.

Theorem 1.50. *If A is a regular matrix, with $A = LU$ factorization as in* (1.21), *then*

$$\det A = \det U = \prod_{i=1}^{n} u_{ii} \tag{1.75}$$

equals the product of the pivots. More generally, if A is nonsingular, and requires k row interchanges to arrive at its permuted LU factorization $PA = LU$, then

$$\det A = \det P \, \det U = (-1)^k \prod_{i=1}^{n} u_{ii}. \tag{1.76}$$

Finally, A is singular if and only if $\det A = 0$.

Proof: In the regular case, one only needs elementary row operations of type #1 to reduce A to upper triangular form U, and axiom 1 says these do not change the determinant. Therefore $\det A = \det U$. Proceeding with the full Gauss–Jordan scheme, the next phase is to divide each row in U by its pivot, leading to the special upper triangular matrix V with all 1's on the diagonal. Axiom 3 implies

$$\det A = \det U = \left(\prod_{i=1}^{n} u_{ii} \right) \det V. \tag{1.77}$$

Finally, we can reduce V to the identity by further row operations of Type #1, and so by (1.74),

$$\det V = \det \mathrm{I} \, = 1. \tag{1.78}$$

Combining equations (1.77), (1.78) proves the theorem for the regular case. The nonsingular case follows without difficulty — each row interchange changes the sign of the determinant, and so $\det A$ equals $\det U$ if there have been an even number of interchanges, but equals $-\det U$ if there have been an odd number.

Finally, if A is singular, then we can reduce it to a matrix with at least one row of zeros by elementary row operations of types #1 and #2. Lemma 1.49 implies that the resulting matrix has zero determinant, and so $\det A = 0$, also. *Q.E.D.*

Corollary 1.51. *The determinant of a diagonal matrix is the product of the diagonal entries. The same result holds for both lower triangular and upper triangular matrices.*

Example 1.52. Let us compute the determinant of the 4×4 matrix

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

We perform our usual Gaussian Elimination algorithm, successively leading to the matrices

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

where we used a single row interchange to obtain the final upper triangular form. Owing to the row interchange, the determinant of the original matrix is -1 times the product of the pivots:

$$\det A = -1 \cdot 1 \cdot 1 \cdot (-2) \cdot 3 = 6.$$

In particular, this tells us that A is nonsingular. But, of course, this was already implied by the elimination, since the matrix reduced to upper triangular form with 4 pivots.

Let us now present some of the basic properties of determinants.

Lemma 1.53. *The determinant of the product of two square matrices of the same size is the product of the determinants:*

$$\det(AB) = \det A \ \det B. \tag{1.79}$$

Proof: The product formula holds if A is an elementary matrix; this is a consequence of the determinantal axioms, combined with Corollary 1.51. By induction, if $A = E_1 E_2 \cdots E_N$ is a product of elementary matrices, then (1.79) also holds. Therefore, the result holds whenever A is nonsingular. On the other hand, if A is singular, then according to Exercise ■, $A = E_1 E_2 \cdots E_N Z$, where the E_i are elementary matrices, and Z, the row echelon form, is a matrix with a row of zeros. But then $ZB = W$ also has a row of zeros, and so $AB = E_1 E_2 \cdots E_N W$ is also singular. Thus, both sides of (1.79) are zero in this case. *Q.E.D.*

It is a remarkable fact that, even though matrix multiplication is not commutative, and so $AB \neq BA$ in general, it is nevertheless always true that both products have the same determinant: $\det(AB) = \det(BA)$. Indeed, both are equal to the product $\det A \ \det B$ of the individual determinants because ordinary (scalar) multiplication is commutative.

Lemma 1.54. *Transposing a matrix does not change its determinant:*

$$\det A^T = \det A. \tag{1.80}$$

Proof: By inspection, this formula holds if A is an elementary matrix. If $A = E_1 E_2 \cdots E_N$ is a product of elementary matrices, then using (1.49), (1.79) and induction

$$\begin{aligned} \det A^T &= \det(E_1 E_2 \cdots E_N)^T = \det(E_N^T E_{N-1}^T \cdots E_1^T) = \det E_N^T \ \det E_{N-1}^T \ \cdots \ \det E_1^T \\ &= \det E_N \ \det E_{N-1} \ \cdots \ \det E_1 = \det E_1 \ \det E_2 \ \cdots \ \det E_N \\ &= \det(E_1 E_2 \ \cdots \ E_N) = \det A. \end{aligned}$$

The middle equality follows from the commutativity of ordinary multiplication. This proves the nonsingular case; the singular case follows from Lemma 1.30, which implies that A^T is singular if and only if A is. *Q.E.D.*

Remark: Lemma 1.54 has the interesting consequence that one can equally well use "elementary column operations" to compute determinants. We will not develop this approach in any detail here, since it does not help us to solve linear equations.

Finally, we state the general formula for a determinant; a proof can be found in [**135**].

Theorem 1.55. *If A is an $n \times n$ matrix with entries a_{ij}, then*

$$\det A = \sum_{\pi} \pm a_{1,\pi(1)}\, a_{2,\pi(2)} \;\cdots\; a_{n,\pi(n)}. \tag{1.81}$$

The sum in (1.81) is over all possible permutations π of the columns of A. The summands consist of all possible ways of choosing n entries of A with one entry in each column and 1 entry in each row of A. The sign in front of the indicated term depends on the permutation π; it is $+$ if π is an even permutation, meaning that its matrix can be reduced to the identity by an even number of row interchanges, and $-$ is π is odd. For example, the six terms in the well-known formula

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{array}{l} a_{11}\, a_{22}\, a_{33} + a_{12}\, a_{23}\, a_{31} + a_{13}\, a_{21} a_{32} - \\ \qquad - a_{11}\, a_{23}\, a_{32} - a_{12}\, a_{21}\, a_{33} - a_{13}\, a_{22}\, a_{31} \end{array} \tag{1.82}$$

for a 3×3 determinant correspond to the six possible 3×3 permutation matrices (1.27).

The proof that (1.81) obeys the basic determinantal axioms is straightforward, but, will not be done here. The reader might wish to try the 3×3 case to be convinced that it works. This explicit formula proves that the determinant function is well-defined, and formally completes the proof of Theorem 1.48.

Unfortunately, the explicit determinant formula (1.81) contains $n!$ terms, and so, as soon as n is even moderately large, is completely impractical for computation. The most efficient way is still our mainstay — Gaussian Elimination coupled the fact that the determinant is $\pm$ the product of the pivots!

Determinants have many fascinating and theoretically important properties. However, in our applications, these will not be required, and so we conclude this very brief introduction to the subject.

Chapter 2

Vector Spaces

Vector spaces and their ancillary structures provide the common language of linear algebra, and, as such are an essential prerequisite for understanding contemporary applied mathematics. The key concepts of vector space, subspace, linear independence, span, and basis will appear, not only in linear systems of equations and the geometry of n-dimensional Euclidean space, but also in the analysis of linear ordinary differential equations, linear partial differential equations, linear boundary value problems, all of Fourier analysis, numerical approximations like the finite element method, and many, many other fields. Therefore, in order to develop the wide variety of analytical methods and applications covered in this text, we need to acquire a firm working knowledge of basic vector space analysis.

One of the great triumphs of modern mathematics was the recognition that many seemingly distinct constructions are, in fact, different manifestations of the same general mathematical structure. The abstract notion of a vector space serves to unify spaces of ordinary vectors, spaces of functions, such as polynomials, exponentials, trigonometric functions, as well as spaces of matrices, linear operators, etc., all in a common conceptual framework. Moreover, proofs that might look rather complicated in any particular context often turn out to be relatively transparent when recast in the abstract vector space framework. The price that one pays for the increased level of abstraction is that, while the underlying mathematics is not all that difficult, the student typically takes a long time to assimilate the material. In our opinion, the best way to approach the subject is to think in terms of concrete examples. First, make sure you understand what the concept or theorem says in the case of ordinary Euclidean space. Once this is grasped, the next important case to consider is an elementary function space, e.g., the space of continuous scalar functions. With these two examples firmly in hand, the leap to the general abstract version should not be too painful. Patience is essential; ultimately the only way to truly understand an abstract concept like a vector space is by working with it! And always keep in mind that the effort expended here will be amply rewarded later on.

Following an introduction to vector spaces and subspaces, we introduce the notions of span and linear independence of a collection of vector space elements. These are combined into the all-important concept of a basis of a vector space, leading to a linear algebraic characterization of its dimension. We will then study the four fundamental subspaces associated with a matrix — range, kernel, corange and cokernel — and explain how they help us understand the solution to linear algebraic systems. Of particular note is the all-pervasive linear superposition principle that enables one to construct more general solutions to linear systems by combining known solutions. Superposition is the hallmark

of linearity, and will apply not only to linear algebraic equations, but also linear ordinary differential equations, linear partial differential equations, linear boundary value problems, and so on. Some interesting applications in graph theory, to be used in our later study of electrical circuits, will form the final topic of this chapter.

2.1. Vector Spaces.

A vector space is the abstract formulation of the most basic underlying properties of n-dimensional[†] Euclidean space $\mathbb{R}^n$, which is defined as the set of all real (column) vectors with n entries. The basic laws of vector addition and scalar multiplication in $\mathbb{R}^n$ serve as the motivation for the general, abstract definition of a vector space. In the beginning, we will refer to the elements of a vector space as "vectors", even though, as we shall see, they might also be functions or matrices or even more general objects. Unless dealing with certain specific examples such as a space of functions, we will use bold face, lower case Latin letters to denote the elements of our vector space. We begin with the general definition.

Definition 2.1. A *vector space* is a set V equipped with two operations:

(i) *Addition*: adding any pair of vectors $\mathbf{v}, \mathbf{w} \in V$ produces another vector $\mathbf{v} + \mathbf{w} \in V$;

(ii) *Scalar Multiplication*: multiplying a vector $\mathbf{v} \in V$ by a scalar $c \in \mathbb{R}$ produces a vector $c\mathbf{v} \in V$.

which are required to satisfy the following axioms for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and all scalars $c, d \in \mathbb{R}$:

(a) *Commutativity of Addition*: $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$.

(b) *Associativity of Addition*: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.

(c) *Additive Identity*: There is a zero element $\mathbf{0} \in V$ satisfying $\mathbf{v} + \mathbf{0} = \mathbf{v} = \mathbf{0} + \mathbf{v}$.

(d) *Additive Inverse*: For each $\mathbf{v} \in V$ there is an element $-\mathbf{v} \in V$ such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0} = (-\mathbf{v}) + \mathbf{v}$.

(e) *Distributivity*: $(c + d)\mathbf{v} = (c\mathbf{v}) + (d\mathbf{v})$, and $c(\mathbf{v} + \mathbf{w}) = (c\mathbf{v}) + (c\mathbf{w})$.

(f) *Associativity of Scalar Multiplication*: $c(d\mathbf{v}) = (cd)\mathbf{v}$.

(g) *Unit for Scalar Multiplication*: the scalar $1 \in \mathbb{R}$ satisfies $1\mathbf{v} = \mathbf{v}$.

Note: We will use bold face $\mathbf{0}$ to denote the zero element of our vector space, while ordinary 0 denotes the real number zero. The following identities are elementary consequences of the vector space axioms:

(h) $0\mathbf{v} = \mathbf{0}$. (i) $(-1)\mathbf{v} = -\mathbf{v}$. ($j$) $c\mathbf{0} = \mathbf{0}$. (k) If $c\mathbf{v} = \mathbf{0}$, then either $c = 0$ or $\mathbf{v} = \mathbf{0}$.

Let us, as an example, prove (h). Let $\mathbf{z} = 0\mathbf{v}$. Then, by the distributive property,

$$\mathbf{z} + \mathbf{z} = 0\mathbf{v} + 0\mathbf{v} = (0 + 0)\mathbf{v} = 0\mathbf{v} = \mathbf{z}.$$

Adding $-\mathbf{z}$ to both sides of this equation, and making use of axioms (b), (d), and then (c), implies that $\mathbf{z} = \mathbf{0}$, which completes the proof. Verification of the other three properties is left as an exercise for the reader.

[†] The precise definition of dimension will appear later, in Theorem 2.28,

Remark: For most of this chapter we will deal with real vector spaces, in which the scalars are the real numbers $\mathbb{R}$. Complex vector spaces, where complex scalars are allowed, will be introduced in Section 3.6. Vector spaces over other fields are studied in abstract algebra, [**77**].

Example 2.2. As noted above, the prototypical example of a real vector space is the space $\mathbb{R}^n$ consisting of column vectors or n-tuples of real numbers $\mathbf{v} = (\, v_1, v_2, \dots, v_n \,)^T$. Vector addition and scalar multiplication are defined in the usual manner:

$$\mathbf{v} + \mathbf{w} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_n + w_n \end{pmatrix}, \qquad c\mathbf{v} = \begin{pmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_n \end{pmatrix}, \qquad \text{whenever} \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}.$$

The zero vector is $\mathbf{0} = (\, 0, \dots, 0 \,)^T$. The fact that vectors in $\mathbb{R}^n$ satisfy all of the vector space axioms is an immediate consequence of the laws of vector addition and scalar multiplication. Details are left to the reader.

Example 2.3. Let $\mathcal{M}_{m\times n}$ denote the space of all real matrices of size $m \times n$. Then $\mathcal{M}_{m\times n}$ forms a vector space under the laws of matrix addition and scalar multiplication. The zero element is the zero matrix O. Again, the vector space axioms are immediate consequences of the basic laws of matrix arithmetic. (For the purposes of this example, we ignore additional matrix properties, like matrix multiplication.) The preceding example of the vector space $\mathbb{R}^n = \mathcal{M}_{1\times n}$ is a particular case when the matrices have only one column.

Example 2.4. Consider the space

$$\mathcal{P}^{(n)} = \left\{\; p(x) = a_n x^n + a_{n-1} x^{n-1} + \;\cdots\; + a_1 x + a_0 \;\right\} \tag{2.1}$$

consisting of all polynomials of degree $\leq n$. Addition of polynomials is defined in the usual manner; for example,

$$(x^2 - 3x) + (2x^2 - 5x + 4) = 3x^2 - 8x + 4.$$

Note that the sum $p(x) + q(x)$ of two polynomials of degree $\leq n$ also has degree $\leq n$. (However, it is not true that the sum of two polynomials of degree $= n$ also has degree n; for example $(x^2+1)+(-x^2+x) = x+1$ has degree 1 even though the two summands have degree 2. This means that the set of polynomials of degree $= n$ is *not* a vector space.) The zero element of $\mathcal{P}^{(n)}$ is the zero polynomial. We can multiply polynomials by scalars — real constants — in the usual fashion; for example if $p(x) = x^2 - 2x$, then $3p(x) = 3x^2 - 6x$. The proof that $\mathcal{P}^{(n)}$ satisfies the vector space axioms is an easy consequence of the basic laws of polynomial algebra.

Remark: We are ignoring the fact that one can also multiply polynomials; this is *not* a vector space operation. Also, any scalar can be viewed as a constant polynomial, but one should really regard these as two completely different objects — one is a *number*, while the other is a *constant function*. To add to the confusion, one typically uses the same notation for these two objects; for instance, 1 could either mean the real number 1 or the constant function taking the value 1 everywhere. The reader needs to exercise due care when interpreting each occurrence.

For much of analysis, including differential equations, Fourier theory, numerical methods, and so on, the most important vector spaces consist of sets of functions with certain specified properties. The simplest such example is the following.

Example 2.5. Let $I \subset \mathbb{R}$ be an interval. Consider the *function space* $\mathcal{F} = \mathcal{F}(I)$ that consists of all real-valued functions $f(x)$ defined for all $x \in I$, which we also write as $f\colon I \to \mathbb{R}$. The claim is that the function space $\mathcal{F}$ has the structure of a vector space. Addition of functions in $\mathcal{F}$ is defined in the usual manner: $(f + g)(x) = f(x) + g(x)$. Multiplication by scalars $c \in \mathbb{R}$ is the same as multiplication by constants, $(c\,f)(x) = c\,f(x)$. The zero element is the constant function that is identically 0 for all $x \in I$. The proof of the vector space axioms is straightforward, just as in the case of polynomials. As in the preceding remark, we are ignoring all additional operations — multiplication, division, inversion, composition, etc. — that can be done with functions; these are irrelevant as far as the vector space structure of $\mathcal{F}$ goes.

Remark: An interval can be (a) *closed*, meaning that it includes its endpoints: $I = [a, b]$, (b) *open*, which does not include either endpoint: $I = (a, b)$, or (c) *half open*, which includes one but not the other endpoint, so $I = [a, b)$ or $(a, b]$. An open endpoint is allowed to be infinite; in particular, $(-\infty, \infty) = \mathbb{R}$ is another way of writing the real line.

Example 2.6. The preceding examples are all, in fact, special cases of an even more general construction. A clue is to note that the last example of a function space does not make any use of the fact that the domain of definition of our functions is a real interval. Indeed, the construction produces a function space $\mathcal{F}(I)$ corresponding to any subset $I \subset \mathbb{R}$.

Even more generally, let S be *any* set. Let $\mathcal{F} = \mathcal{F}(S)$ denote the space of all real-valued functions $f\colon S \to \mathbb{R}$. Then we claim that V is a vector space under the operations of function addition and scalar multiplication. More precisely, given functions f and g, we define their sum to be the function $h = f + g$ such that $h(x) = f(x) + g(x)$ for all $x \in S$. Similarly, given a function f and a real scalar $c \in \mathbb{R}$, we define the scalar multiple $k = c\,f$ to be the function such that $k(x) = c\,f(x)$ for all $x \in S$. The verification of the vector space axioms proceeds straightforwardly, and the reader should be able to fill in the necessary details.

In particular, if $S \subset \mathbb{R}$ is an interval, then $\mathcal{F}(S)$ coincides with the space of scalar functions described in the preceding example. If $S \subset \mathbb{R}^n$ is a subset of Euclidean space, then the elements of $\mathcal{F}(S)$ are functions $f(x_1, \dots, x_n)$ depending upon the n variables corresponding to the coordinates of points $\mathbf{x} = (x_1, \dots, x_n) \in S$ in the domain. In this fashion, the set of real-valued functions defined on any domain in $\mathbb{R}^n$ is found to also form a vector space.

Another useful example is to let $S = \{x_1, \dots, x_n\} \subset \mathbb{R}$ be a finite set of real numbers. A real-valued function $f\colon S \to \mathbb{R}$ is defined by its values $f(x_1), f(x_2), \dots f(x_n)$ at the specified points. In applications, one can view such functions as indicating the *sample values* of a scalar function $f(x) \in \mathcal{F}(\mathbb{R})$ taken at the *sample points* $x_1, \dots, x_n$. For example, when measuring a physical quantity, e.g., temperature, velocity, pressure, etc., one typically only measures a finite set of sample values. The intermediate, non-recorded values between the sample points are then reconstructed through some form of interpolation — a topic

that we shall visit in depth later on. Interestingly, the sample values $f(x_i)$ can be identified with the entries f_i of a vector

$$\mathbf{f} = (\, f_1, f_2, \ldots, f_n \,)^T = (\, f(x_1), f(x_2), \ldots, f(x_n) \,)^T \in \mathbb{R}^n,$$

known as the *sample vector*. Every sampled function $f\colon \{x_1, \ldots, x_n\} \to \mathbb{R}$ corresponds to a unique vector $\mathbf{f} \in \mathbb{R}^n$ and vice versa. (However, different scalar functions $f\colon \mathbb{R} \to \mathbb{R}$ can have the same sample values.) Addition of sample functions corresponds to addition of their sample vectors, as does scalar multiplication. Thus, *the vector space of sample functions* $\mathcal{F}(S) = \mathcal{F}(\{x_1, \ldots, x_n\})$ *is the same as the vector space* $\mathbb{R}^n$! This connection between sampled functions and vectors will be the key to the finite Fourier transform, of fundamental importance in modern signal processing.

Example 2.7. The preceding construction admits yet a further generalization. We continue to let S be an arbitrary set. Let V be a vector space. The claim is that the space $\mathcal{F}(S, V)$ consisting of all V–valued functions $\mathbf{f}: S \to V$ is a vector space. In other words, we replace the particular vector space $\mathbb{R}$ in the preceding example by a general vector space, and the same conclusion holds. The operations of function addition and scalar multiplication are defined in the evident manner: $(\mathbf{f} + \mathbf{g})(x) = \mathbf{f}(x) + \mathbf{g}(x)$ and $(c\mathbf{f})(x) = c\mathbf{f}(x)$, where we are using the vector addition and scalar multiplication operations on V to induce corresponding operations on V–valued functions. The proof that $\mathcal{F}(S, V)$ satisfies all of the vector space axioms proceeds as before.

The most important example is when $S \subset \mathbb{R}^n$ is a domain in Euclidean space and $V = \mathbb{R}^m$ is itself a Euclidean space. In this case, the elements of $\mathcal{F}(S, \mathbb{R}^m)$ consist of vector-valued functions $\mathbf{f}\colon S \to \mathbb{R}^m$, so that $\mathbf{f}(\mathbf{x}) = (\, f_1(x_1, \ldots, x_n), \ldots, f_m(x_1, \ldots, x_n) \,)^T$ is a column vector consisting of m functions of n variables, all defined on a common domain S. The general construction implies that addition and scalar multiplication of vector-valued functions is done componentwise; for example

$$2 \begin{pmatrix} x^2 \\ e^x - 4 \end{pmatrix} - \begin{pmatrix} \cos x \\ x \end{pmatrix} = \begin{pmatrix} 2\,x^2 - \cos x \\ 2\,e^x - x - 8 \end{pmatrix}.$$

2.2. Subspaces.

In the preceding section, we were introduced to the most basic vector spaces that play a role in this text. Almost all of the important vector spaces arising in applications appear as particular subsets of these key examples.

Definition 2.8. A *subspace* of a vector space V is a subset $W \subset V$ which is a vector space in its own right.

Since elements of W also belong to V, the operations of vector addition and scalar multiplication for W are induced by those of V. In particular, W *must* contain the zero element of V in order to satisfy axiom (c). The verification of the vector space axioms for a subspace is particularly easy: we only need check that addition and scalar multiplication keep us within the subspace.

Proposition 2.9. *A subset $W \subset V$ of a vector space is a subspace if and only if*

(a) *for every $\mathbf{v}, \mathbf{w} \in W$, the sum $\mathbf{v} + \mathbf{w} \in W$, and*

(b) *for every $\mathbf{v} \in W$ and every $c \in \mathbb{R}$, the scalar product $c\mathbf{v} \in W$.*

Proof: The proof is essentially trivial. For example, to show commutativity, given $\mathbf{v}, \mathbf{w} \in W$, we can regard them as elements of V, in which case $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$ because V is a vector space. But the closure condition implies that the sum also belongs to W, and so the commutativity axiom also holds for elements of W. The other axioms are equally easy to validate. *Q.E.D.*

Remark: Condition (a) says that a subspace must be *closed* under addition, while (a) says it must also be closed under scalar multiplication. It will sometimes be useful to combine the two closure conditions. Thus, to prove $W \subset V$ is a subspace it suffices to check that $c\mathbf{v} + d\mathbf{w} \in W$ for every $\mathbf{v}, \mathbf{w} \in W$ and $c, d \in \mathbb{R}$.

Example 2.10. Let us list some examples of subspaces of the three-dimensional Euclidean space $\mathbb{R}^3$. In each case, we must verify the closure conditions; the first two are immediate.

(a) The trivial subspace $W = \{\mathbf{0}\}$.

(b) The entire space $W = \mathbb{R}^3$.

(c) The set of all vectors of the form $(\,x, y, 0\,)^T$, i.e., the (x, y)–coordinate plane. Note that the sum $(\,x, y, 0\,)^T + (\,\widehat{x}, \widehat{y}, 0\,)^T = (\,x + \widehat{x}, y + \widehat{y}, 0\,)^T$, and scalar multiple $c\,(\,x, y, 0\,)^T = (\,cx, cy, 0\,)^T$, of vectors in the (x, y)–plane also lie in the plane, proving closure.

(d) The set of solutions $(\,x, y, z\,)^T$ to the homogeneous linear equation

$$3x + 2y - z = 0.$$

Indeed, if $\mathbf{x} = (\,x, y, z\,)^T$ is a solution, then so is any scalar multiple $c\mathbf{x} = (\,cx, cy, cz\,)^T$ since

$$3(cx) + 2(cy) - (cz) = c(3x + 2y - z) = 0.$$

Moreover, if $\widehat{\mathbf{x}} = (\widehat{x}, \widehat{y}, \widehat{z})$ is a second solution, the sum $\mathbf{x} + \widehat{\mathbf{x}} = (\,x + \widehat{x}, y + \widehat{y}, z + \widehat{z}\,)^T$ is also a solution since

$$3(x + \widehat{x}) + 2(y + \widehat{y}) - (z + \widehat{z}) = (3x + 2y - z) + (3\,\widehat{x} + 2\,\widehat{y} - \widehat{z}) = 0.$$

Note that the solution space is a two-dimensional plane consisting of all vectors which are perpendicular (orthogonal) to the vector $(\,3, 2, -1\,)^T$.

(e) The set of all vectors lying in the plane spanned by the vectors $\mathbf{v}_1 = (\,2, -3, 0\,)^T$ and $\mathbf{v}_2 = (\,1, 0, 3\,)^T$. In other words, we consider all vectors of the form

$$\mathbf{v} = a\,\mathbf{v}_1 + b\,\mathbf{v}_2 = a \begin{pmatrix} 2 \\ -3 \\ 0 \end{pmatrix} + b \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 2a + b \\ -3a \\ 3b \end{pmatrix},$$

where $a, b \in \mathbb{R}$ are arbitrary scalars. If $\mathbf{v} = a\,\mathbf{v}_1 + b\,\mathbf{v}_2$ and $\mathbf{w} = \widehat{a}\,\mathbf{v}_1 + \widehat{b}\,\mathbf{v}_2$ are any two vectors of this form, so is

$$c\,\mathbf{v} + d\,\mathbf{w} = c(a\,\mathbf{v}_1 + b\,\mathbf{v}_2) + d(\widehat{a}\,\mathbf{v}_1 + \widehat{b}\,\mathbf{v}_2) = (a\,c + \widehat{a}\,d)\mathbf{v}_1 + (b\,c + \widehat{b}\,d)\mathbf{v}_2 = \widetilde{a}\,\mathbf{v}_1 + \widetilde{b}\,\mathbf{v}_2,$$

where $\widetilde{a} = a\,c + \widehat{a}\,d$, $\widetilde{b} = b\,c + \widehat{b}\,d$. This proves that the plane is a subspace of $\mathbb{R}^3$. The reader might already have noticed that this subspace is the same plane that was considered in item (d).

Example 2.11. The following subsets of $\mathbb{R}^3$ are *not* subspaces.

(a) The set P of all vectors of the form $(\,x, y, 1\,)^T$, i.e., the plane parallel to the $x\,y$ coordinate plane passing through $(\,0, 0, 1\,)^T$. Indeed, $\mathbf{0} \notin P$, which is the most basic requirement for a subspace. In fact, neither of the closure axioms hold for this subset.

(b) The positive octant $\mathcal{O}^+ = \{x > 0, y > 0, z > 0\}$. While the sum of two vectors in $\mathcal{O}^+$ belongs to $\mathcal{O}^+$, multiplying by negative scalars takes us outside the orthant, violating closure under scalar multiplication.

(c) The unit sphere $S^2 = \{\,x^2 + y^2 + z^2 = 1\,\}$. Again, $\mathbf{0} \notin S^2$. More generally, curved surfaces, e.g., the paraboloid $P = \{\,z = x^2 + y^2\,\}$, are not subspaces. Although $\mathbf{0} \in P$, most scalar multiples of vectors in P do not belong to P. For example, $(\,1, 1, 2\,)^T \in P$, but $2\,(\,1, 1, 2\,)^T = (\,2, 2, 4\,)^T \notin P$.

In fact, there are only four fundamentally different types of subspaces $W \subset \mathbb{R}^3$ of three-dimensional Euclidean space:

(i) The entire space $W = \mathbb{R}^3$,

(ii) a plane passing through the origin,

(iii) a line passing through the origin,

(iv) the trivial subspace $W = \{\mathbf{0}\}$.

To verify this observation, we argue as follows. If $W = \{\mathbf{0}\}$ contains only the zero vector, then we are in case (iv). Otherwise, $W \subset \mathbb{R}^3$ contains a nonzero vector $\mathbf{0} \neq \mathbf{v}_1 \in W$. But since W must contain all scalar multiples $c\,\mathbf{v}_1$ of this element, it includes the entire line in the direction of $\mathbf{v}_1$. If W contains another vector $\mathbf{v}_2$ that does not lie in the line through $\mathbf{v}_1$, then it must contain the entire plane $\{c\,\mathbf{v}_1 + d\,\mathbf{v}_2\}$ spanned by $\mathbf{v}_1, \mathbf{v}_2$. Finally, if there is a third vector $\mathbf{v}_3$ not contained in this plane, then we claim that $W = \mathbb{R}^3$. This final fact will be an immediate consequence of general results in this chapter, although the interested reader might try to prove it directly before proceeding.

Example 2.12. Let $I \subset \mathbb{R}$ be an interval, and let $\mathcal{F}(I)$ be the space of real-valued functions $f\colon I \to \mathbb{R}$. Let us look at some of the most important examples of subspaces of $\mathcal{F}(I)$. In each case, we need only verify the closure conditions to verify that the given subset is indeed a subspace.

(a) The space $\mathcal{P}^{(n)}$ of polynomials of degree $\leq n$, which we already encountered.

(b) The space $\mathcal{P}^{(\infty)} = \bigcup_{n \geq 0} \mathcal{P}^{(n)}$ consisting of all polynomials.

(c) The space $\mathrm{C}^0(I)$ of all continuous functions. Closure of this subspace relies on knowing that if $f(x)$ and $g(x)$ are continuous, then both $f(x) + g(x)$ and $c\,f(x)$ for any $c \in \mathbb{R}$ are also continuous — two basic results from calculus.

(d) More restrictively, one can consider the subspace $\mathrm{C}^n(I)$ consisting of all functions $f(x)$ that have n continuous derivatives $f'(x), f''(x), \ldots, f^{(n)}(x)$ on† I. Again, we need to know that if $f(x)$ and $g(x)$ have n continuous derivatives, so do $f(x) + g(x)$ and $cf(x)$ for any $c \in \mathbb{R}$.

(e) The space $\mathrm{C}^\infty(I) = \bigcap_{n \geq 0} \mathrm{C}^n(I)$ of infinitely differentiable or *smooth* functions is also a subspace. (The fact that this intersection is a subspace follows directly from Exercise ■.)

(f) The space $\mathcal{A}(I)$ of analytic functions on the interval I. Recall that a function $f(x)$ is called *analytic* at a point a if it is smooth, and, moreover, its Taylor series

$$f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\,f''(a)\,(x-a)^2 + \cdots = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}\,(x-a)^n \tag{2.2}$$

converges to $f(x)$ for all x sufficiently close to a. (It does not have to converge on the entire interval I.) Not every smooth function is analytic, and so $\mathcal{A}(I) \subsetneq \mathrm{C}^\infty(I)$. An explicit example is the function

$$f(x) = \begin{cases} e^{-1/x}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{2.3}$$

It can be shown that every derivative of this function at 0 exists and equals zero: $f^{(n)}(0) = 0$, $n = 0, 1, 2, \ldots$, and so the function is smooth. However, its Taylor series at $a = 0$ is $0 + 0\,x + 0\,x^2 + \cdots \equiv 0$, which converges to the zero function, not to $f(x)$. Therefore $f(x)$ is *not* analytic at $a = 0$.

(g) The set of all *mean zero* functions. The *mean* or *average* of an integrable function defined on a closed interval $I = [a, b]$ is the real number

$$\overline{f} = \frac{1}{b-a} \int_a^b f(x)\,dx. \tag{2.4}$$

In particular, f has *mean zero* if and only if $\displaystyle\int_a^b f(x)\,dx = 0$. Note that $\overline{f+g} = \overline{f} + \overline{g}$, and so the sum of two mean zero functions also has mean zero. Similarly, $\overline{cf} = c\,\overline{f}$, and any scalar multiple of a mean zero function also has mean zero.

(h) Let $x_0 \in I$ be a given point. Then the set of all functions $f(x)$ that vanish at the point, $f(x_0) = 0$, is a subspace. Indeed, if $f(x_0) = 0$ and $g(x_0) = 0$, then clearly $(f+g)(x_0) = 0$ and $c\,f(x_0) = 0$, proving closure. This example can evidently be generalized to functions that vanish at several points, or even on an entire subset.

(i) The set of all solutions $u = f(x)$ to the homogeneous linear differential equation

$$u'' + 2\,u' - 3\,u = 0.$$

Indeed, if $u = f(x)$ and $u = g(x)$ are solutions, so are $u = f(x) + g(x)$ and $u = c\,f(x)$ for any $c \in \mathbb{R}$. Note that we do *not* need to actually solve the equation to verify these claims!

† If $I = [a, b]$ is closed, we use the appropriate one-sided derivatives at its endpoints.

They follow directly from linearity; for example

$$(f+g)'' + 2(f+g)' - 3(f+g) = (f'' + 2\,f' - 3\,f) + (g'' + 2\,g' - 3\,g) = 0.$$

Warning: In the last three examples, the value 0 is essential for the indicated set of functions to be a subspace. The set of functions such that $f(x_0) = 1$, say, is not a subspace. The set of functions with a fixed nonzero mean, say $\overline{f} = 3$, is also not a subspace. Nor is the set of solutions to an inhomogeneous ordinary differential equation, say

$$u'' + 2\,u' - 3\,u = x - 3.$$

None of these subsets contain the zero function, nor do they satisfy the closure conditions.

2.3. Span and Linear Independence.

The definition of the span of a finite collection of elements of a vector space generalizes, in a natural fashion, the geometric notion of two vectors spanning a plane in $\mathbb{R}^3$. As such, it forms the first of two important, general methods for constructing subspaces of vector spaces.

Definition 2.13. Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be a finite collection of elements of a vector space V. A sum of the form

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \,\cdots\, + c_k\mathbf{v}_k = \sum_{i=1}^{k} c_i\mathbf{v}_i, \tag{2.5}$$

where the coefficients $c_1, c_2, \dots, c_k$ are any scalars, is known as a *linear combination* of the elements $\mathbf{v}_1, \dots, \mathbf{v}_k$. Their *span* is the subset $W = \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\} \subset V$ consisting of all possible linear combinations (2.5).

For example,

$$3\,\mathbf{v}_1 + \mathbf{v}_2 - 2\,\mathbf{v}_3, \quad 8\,\mathbf{v}_1 - \tfrac{1}{3}\,\mathbf{v}_3, \quad \mathbf{v}_2 = 0\,\mathbf{v}_1 + 1\,\mathbf{v}_2 + 0\,\mathbf{v}_3, \quad \text{and} \quad \mathbf{0} = 0\,\mathbf{v}_1 + 0\,\mathbf{v}_2 + 0\,\mathbf{v}_3,$$

are four different linear combinations of the three vector space elements $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in V$.

The key observation is that a span always forms a subspace.

Proposition 2.14. *The span of a collection of vectors, $W = \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, forms a subspace of the underlying vector space.*

Proof: We need to show that if

$$\mathbf{v} = c_1\mathbf{v}_1 + \,\cdots\, + c_k\mathbf{v}_k \qquad \text{and} \qquad \widehat{\mathbf{v}} = \widehat{c}_1\mathbf{v}_1 + \,\cdots\, + \widehat{c}_k\mathbf{v}_k$$

are any two linear combinations, then their sum

$$\mathbf{v} + \widehat{\mathbf{v}} = (c_1 + \widehat{c}_1)\mathbf{v}_1 + \,\cdots\, + (c_k + \widehat{c}_k)\mathbf{v}_k,$$

is also a linear combination, as is any scalar multiple

$$a\,\mathbf{v} = (a\,c_1)\mathbf{v}_1 + \,\cdots\, + (a\,c_k)\mathbf{v}_k \qquad \textit{Q.E.D.}$$

Example 2.15. Examples of subspaces spanned by vectors in $\mathbb{R}^3$:

(*i*) If $\mathbf{v}_1 \neq \mathbf{0}$ is any non-zero vector in $\mathbb{R}^3$, then its span is the line $\{\, c\,\mathbf{v}_1 \,|\, c \in \mathbb{R} \,\}$ in the direction of $\mathbf{v}_1$. If $\mathbf{v}_1 = \mathbf{0}$, then its span just consists of the origin.

(*ii*) If $\mathbf{v}_1$ and $\mathbf{v}_2$ are any two vectors in $\mathbb{R}^3$, then their span is the set of all vectors of the form $c_1\mathbf{v}_1 + c_2\mathbf{v}_2$. Typically, such a span forms a plane passing through the origin. However, if $\mathbf{v}_1$ and $\mathbf{v}_2$ are parallel, then their span is just a line. The most degenerate case is when $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{0}$, where the span is just a point — the origin.

(*iii*) If we are given three non-coplanar vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$, then their span is all of $\mathbb{R}^3$, as we shall prove below. However, if they all lie in a plane, then their span is the plane — unless they are all parallel, in which case their span is a line — or, when $\mathbf{v}_1 = \mathbf{v}_2 = \mathbf{v}_3 = \mathbf{0}$, a single point.

Thus, any subspace of $\mathbb{R}^3$ can be realized as the span of some set of vectors. Note that we can also consider the span of four or more vectors, but the range of possible subspaces is limited, as we noted above, to either a point (the origin), a line, a plane, or the entire three-dimensional space. A crucial question, that we will return to shortly, is to determine when a given vector belongs to the span of a colection of vectors.

Remark: It is entirely possible for different sets of vectors to span the *same* subspace. For instance, the pair of vectors $\mathbf{e}_1 = (\,1,0,0\,)^T$ and $\mathbf{e}_2 = (\,0,1,0\,)^T$ span the xy plane in $\mathbb{R}^3$, as do the three coplanar vectors $\mathbf{v}_1 = (\,1,-1,0\,)^T, \mathbf{v}_2 = (\,-1,2,0\,)^T, \mathbf{v}_3 = (\,2,1,0\,)^T$.

Example 2.16. Let $V = \mathcal{F}(\mathbb{R})$ denote the space of all scalar functions $f(x)$.

(*a*) The span of the three monomials $f_1(x) = 1$, $f_2(x) = x$ and $f_3(x) = x^2$ is the set of all functions of the form

$$f(x) = c_1\, f_1(x) + c_2\, f_2(x) + c_3\, f_3(x) = c_1 + c_2\, x + c_3\, x^2,$$

where c_1, c_2, c_3 are arbitrary scalars (constants). In other words, span $\{1, x, x^2\} = \mathcal{P}^{(2)}$ is the subspace of all quadratic (degree ≤ 2) polynomials. In a similar fashion, the space $\mathcal{P}^{(n)}$ of polynomials of degree $\leq n$ is spanned by the monomials $1, x, x^2, \ldots, x^n$.

(*b*) The next example plays a key role in many applications. Let $\omega \in \mathbb{R}$ be fixed. Consider the two basic trigonometric functions $f_1(x) = \cos\omega x$, $f_2(x) = \sin\omega x$ of frequency ω, and hence period $2\pi/\omega$. Their span consists of all functions of the form

$$f(x) = c_1\, f_1(x) + c_2\, f_2(x) = c_1 \cos\omega x + c_2 \sin\omega x. \tag{2.6}$$

For example, the function $\cos(\omega x + 2)$ lies in the span because, by the addition formula for the cosine,

$$\cos(\omega x + 2) = \cos 2\, \cos\omega x - \sin 2\, \sin\omega x$$

is a linear combination of $\cos\omega x$ and $\sin\omega x$.

We can express a general function in their span in the alternative *phase-amplitude form*

$$f(x) = c_1 \cos\omega x + c_2 \sin\omega x = r\cos(\omega x - \delta). \tag{2.7}$$

Expanding the right hand side, we find

$$r\cos(\omega x - \delta) \;=\; r\cos\delta\, \cos\omega x + r\sin\delta\, \sin\omega x$$

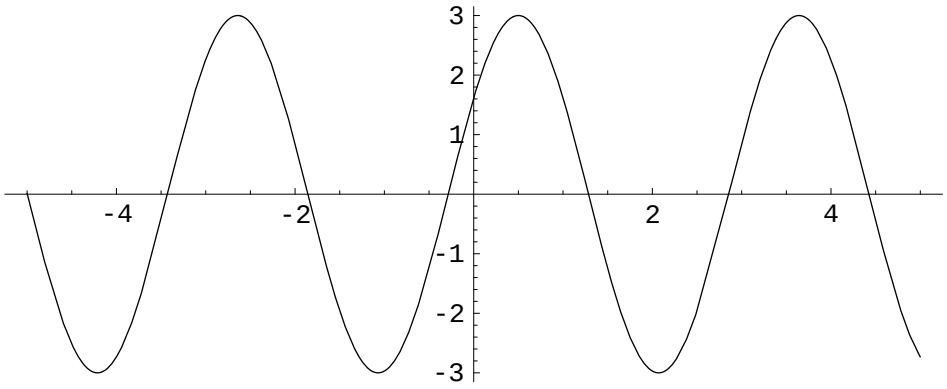


Figure 2.1. Graph of $3\cos(2x-1)$.

and hence

$$c_1 = r\cos\delta, \qquad c_2 = r\sin\delta.$$

We can view the *amplitude* $r \geq 0$ and the *phase shift* δ as the polar coordinates of point $\mathbf{c} = (c_1, c_2) \in \mathbb{R}^2$ prescribed by the coefficients. Thus, any combination of $\sin\omega x$ and $\cos\omega x$ can be rewritten as a single cosine, with a phase lag. Figure 2.1 shows the particular case $3\cos(2x-1)$ which has amplitude $r = 3$, frequency $\omega = 2$ and phase shift $\delta = 1$. The first peak appears at $x = \delta/\omega = \frac{1}{2}$.

(*c*) The space $\mathcal{T}^{(2)}$ of quadratic trigonometric polynomials is spanned by the functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos^2 x, \qquad \cos x\,\sin x, \qquad \sin^2 x.$$

Thus, the general quadratic trigonometric polynomial can be written as a linear combination

$$q(x) = c_0 + c_1\,\cos x + c_2\,\sin x + c_3\,\cos^2 x + c_4\,\cos x\,\sin x + c_5\,\sin^2 x, \tag{2.8}$$

where $c_0, \dots, c_5$ are arbitrary constants. A more useful spanning set for the same subspace is the trigonometric functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos 2x, \qquad \sin 2x. \tag{2.9}$$

Indeed, by the double angle formulas, both

$$\cos 2x = \cos^2 x - \sin^2 x, \qquad \sin 2x = 2\,\sin x\cos x,$$

have the form of a quadratic trigonometric polynomial (2.8), and hence both belong to $\mathcal{T}^{(2)}$. On the other hand, we can write

$$\cos^2 x = \tfrac{1}{2}\cos 2x + \tfrac{1}{2}, \qquad \cos x\,\sin x = \tfrac{1}{2}\sin 2x, \qquad \sin^2 x = -\tfrac{1}{2}\cos 2x + \tfrac{1}{2},$$

in terms of the functions (2.9). Therefore, the original linear combination (2.8) can be written in the alternative form

$$\begin{aligned} q(x) &= \big(c_0 + \tfrac{1}{2}c_3 + \tfrac{1}{2}c_5\big) + c_1\,\cos x + c_2\,\sin x + \big(\tfrac{1}{2}c_3 - \tfrac{1}{2}c_5\big)\,\cos 2x + \tfrac{1}{2}c_4\,\sin 2x \\ &= \widehat{c}_0 + \widehat{c}_1\,\cos x + \widehat{c}_2\,\sin x + \widehat{c}_3\,\cos 2x + \widehat{c}_4\,\sin 2x, \end{aligned} \tag{2.10}$$

and so the functions (2.9) do indeed span $\mathcal{T}^{(2)}$. It is worth noting that we first characterized $\mathcal{T}^{(2)}$ as the span of 6 functions, whereas the second characterization only required 5 functions. It turns out that 5 is the minimal number of functions needed to span $\mathcal{T}^{(2)}$, but the proof of this fact will be deferred until Chapter 3.

(*d*) The homogeneous linear ordinary differential equation

$$u'' + 2\,u' - 3\,u = 0. \tag{2.11}$$

considered in part (*i*) of Example 2.12 has two independent solutions: $f_1(x) = e^x$ and $f_2(x) = e^{-3\,x}$. (Now may be a good time for you to review the basic techniques for solving linear, constant coefficient ordinary differential equations.) The general solution to the differential equation is a linear combination

$$u = c_1\, f_1(x) + c_2\, f_2(x) = c_1\, e^x + c_2\, e^{-3\,x}.$$

Thus, the vector space of solutions to (2.11) is described as the span of these two basic solutions. The fact that there are no other solutions is not obvious, but relies on the basic existence and uniqueness theorems for linear ordinary differential equations; see Theorem 7.33 for further details.

Remark: One can also define the span of an infinite collection of elements of a vector space. To avoid convergence issues, one should only consider finite linear combinations (2.5). For example, the span of the monomials $1, x, x^2, x^3, \dots$ is the space $\mathcal{P}^{(\infty)}$ of all polynomials. (*Not* the space of convergent Taylor series.) Similarly, the span of the functions $1, \cos x, \sin x, \cos 2\,x, \sin 2\,x, \cos 3\,x, \sin 3\,x, \dots$ is the space of all *trigonometric polynomials*, to be discussed in great detail in Chapter 12.

Linear Independence and Dependence

Most of the time, all of the vectors used to form a span are essential. For example, we cannot use fewer than two vectors to span a plane in $\mathbb{R}^3$ since the span of a single vector is at most a line. However, in the more degenerate cases, some of the spanning elements are not needed. For instance, if the two vectors are parallel, then their span is a line, but only one of the vectors is really needed to define the line. Similarly, the subspace spanned by the polynomials

$$p_1(x) = x - 2, \qquad p_2(x) = x^2 - 5\,x + 4, \qquad p_3(x) = 3\,x^2 - 4\,x, \qquad p_4(x) = x^2 - 1. \tag{2.12}$$

is the vector space $\mathcal{P}^{(2)}$ of quadratic polynomials. But only three of the polynomials are really required to span $\mathcal{P}^{(2)}$. (The reason will become clear soon, but you may wish to see if you can demonstrate this on your own.) The elimination of such superfluous spanning elements is encapsulated in the following basic definition.

Definition 2.17. The vectors $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ are called *linearly dependent* if there exists a collection of scalars $c_1, \dots, c_k$, *not all zero*, such that

$$c_1 \mathbf{v}_1 + \, \cdots \, + c_k \mathbf{v}_k = \mathbf{0}. \tag{2.13}$$

Vectors which are not linearly dependent are called *linearly independent*.

The restriction that the c_i's not all simultaneously vanish is essential. Indeed, if $c_1 = \cdots = c_k = 0$, then the linear combination (2.13) is automatically zero. To check linear independence, one needs to show that the *only* linear combination that produces the zero vector (2.13) is this trivial one. In other words, $c_1 = \cdots = c_k = 0$ is the *one and only* solution to the vector equation (2.13).

Example 2.18. Some examples of linear independence and dependence:

(*a*) The vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 3 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 4 \\ 3 \end{pmatrix},$$

are linearly dependent. Indeed,

$$\mathbf{v}_1 - 2\,\mathbf{v}_2 + \mathbf{v}_3 = \mathbf{0}.$$

On the other hand, the first two vectors $\mathbf{v}_1, \mathbf{v}_2$ are linearly independent. To see this, suppose that

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \begin{pmatrix} c_1 \\ 2\,c_1 + 3\,c_2 \\ -\,c_1 + c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

For this to happen, the coefficients c_1, c_2 must satisfy the homogeneous linear system

$$c_1 = 0, \qquad 2\,c_1 + 3\,c_2 = 0, \qquad -\,c_1 + c_2 = 0,$$

which has only the trivial solution $c_1 = c_2 = 0$, proving linear independence.

(*b*) In general, any collection $\mathbf{v}_1, \ldots, \mathbf{v}_k$ that includes the zero vector, say $\mathbf{v}_1 = \mathbf{0}$, is automatically linearly dependent, since $1\,\mathbf{v}_1 + 0\,\mathbf{v}_2 + \cdots + 0\,\mathbf{v}_k = \mathbf{0}$ is a nontrivial linear combination that adds up to $\mathbf{0}$.

(*c*) The polynomials (2.12) are linearly dependent; indeed,

$$p_1(x) + p_2(x) - p_3(x) + 2\,p_4(x) \equiv 0$$

is a nontrivial linear combination that vanishes identically. On the other hand, the first three polynomials, $p_1(x), p_2(x), p_3(x)$, are linearly independent. Indeed, if the linear combination

$$c_1\,p_1(x) + c_2\,p_2(x) + c_3\,p_3(x) = (c_2 + 3\,c_3)\,x^2 + (c_1 - 5\,c_2 - 4\,c_3)\,x - 2\,c_1 + 4\,c_2 \equiv 0$$

is the zero polynomial, then its coefficients must vanish, and hence c_1, c_2, c_3 are required to solve the homogeneous linear system

$$c_2 + 3\,c_3 = 0, \qquad c_1 - 5\,c_2 - 4\,c_3 = 0, \qquad -\,2\,c_1 + 4\,c_2 = 0.$$

But this has only the trivial solution $c_1 = c_2 = c_3 = 0$, and so linear independence follows.

Remark: In the last example, we are using the basic fact that a polynomial is identically zero,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \equiv 0 \qquad \text{for all} \qquad x,$$

if and only if its coefficients all vanish: $a_0 = a_1 = \cdots = a_n = 0$. This is equivalent to the "self-evident" fact that the basic monomial functions $1, x, x^2, \ldots, x^n$ are linearly independent; see Exercise ■.

Example 2.19. The set of quadratic trigonometric functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos^2 x, \qquad \cos x \, \sin x, \qquad \sin^2 x,$$

that were used to define the vector space $\mathcal{T}^{(2)}$ of quadratic trigonometric polynomials, are, in fact, linearly dependent. This is a consequence of the basic trigonometric identity

$$\cos^2 x + \sin^2 x \equiv 1$$

which can be rewritten as a nontrivial linear combination

$$1 + 0 \cos x + 0 \sin x - \cos^2 x + 0 \cos x \, \sin x - \sin^2 x \equiv 0$$

that sums to the zero function. On the other hand, the alternative spanning set

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos 2x, \qquad \sin 2x,$$

is linearly independent, since the only identically zero linear combination

$$c_0 + c_1 \cos x + c_2 \sin x + c_3 \cos 2x + c_4 \sin 2x \equiv 0$$

is the trivial one $c_0 = \ldots = c_4 = 0$. However, the latter fact is not as obvious, and requires a bit of work to prove directly; see Exercise ■. An easier proof, based on orthogonality, will appear in Chapter 5.

Let us now focus our attention on the linear independence or dependence of a set of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ in Euclidean space. We begin by forming the $n \times k$ matrix $A = (\, \mathbf{v}_1 \, \ldots \, \mathbf{v}_k \,)$ whose *columns* are the given vectors. (The fact that we use column vectors is essential here.) The key is a very basic formula

$$A\mathbf{c} = c_1 \mathbf{v}_1 + \cdots + c_k \mathbf{v}_k, \qquad \text{where} \qquad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix}, \tag{2.14}$$

that expresses any linear combination in terms of matrix multiplication. For example,

$$\begin{pmatrix} 1 & 3 & 0 \\ -1 & 2 & 1 \\ 4 & -1 & -2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} c_1 + 3c_2 \\ -c_1 + 2c_2 + c_3 \\ 4c_1 - c_2 - 2c_3 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ -1 \\ 4 \end{pmatrix} + c_2 \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix} + c_3 \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}.$$

Formula (2.14) is an immediate consequence of the rules of matrix multiplication; see also Exercise ■c. It allows us to reformulate the notions of linear independence and span in terms of linear systems of equations. The main result is the following:

Theorem 2.20. *Let $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$ and let $A = (\, \mathbf{v}_1 \, \dots \, \mathbf{v}_k \,)$ be the corresponding $n \times k$ matrix.*

(a) *The vectors $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$ are linearly dependent if and only if there is a non-zero solution $\mathbf{c} \neq \mathbf{0}$ to the homogeneous linear system $A\,\mathbf{c} = \mathbf{0}$.*

(b) *The vectors are linearly independent if and only if the only solution to the homogeneous system $A\,\mathbf{c} = \mathbf{0}$ is the trivial one $\mathbf{c} = \mathbf{0}$.*

(c) *A vector $\mathbf{b}$ lies in the span of $\mathbf{v}_1, \dots, \mathbf{v}_k$ if and only if the linear system $A\,\mathbf{c} = \mathbf{b}$ is compatible, i.e., it has at least one solution.*

Proof: We prove the first statement, leaving the other two as exercises for the reader. The condition that $\mathbf{v}_1, \dots, \mathbf{v}_k$ be linearly dependent is that there is a nonzero vector

$$\mathbf{c} = (\, c_1, c_2, \dots, c_k \,)^T \neq \mathbf{0}$$

such that the linear combination

$$A\,\mathbf{c} = c_1\,\mathbf{v}_1 + \, \cdots \, + c_k\,\mathbf{v}_k = \mathbf{0}.$$

Therefore, linear dependence requires the existence of a nontrivial solution to the homogeneous linear system $A\,\mathbf{c} = \mathbf{0}$. *Q.E.D.*

Example 2.21. Let us determine whether the vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 4 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -4 \\ 6 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} 4 \\ 2 \\ 3 \end{pmatrix}, \tag{2.15}$$

are linearly independent or linearly dependent. We combine them as column vectors into a single matrix

$$A = \begin{pmatrix} 1 & 3 & 1 & 4 \\ 2 & 0 & -4 & 2 \\ -1 & 4 & 6 & 3 \end{pmatrix}.$$

According to Theorem 2.20, we need to figure out whether there are any nontrivial solutions to the homogeneous equation $A\,\mathbf{c} = \mathbf{0}$; this can be done by reducing A to row echelon form, which is

$$U = \begin{pmatrix} 1 & 3 & 1 & 4 \\ 0 & -6 & -6 & -6 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2.16}$$

The general solution to the homogeneous system $A\,\mathbf{c} = \mathbf{0}$ is

$$\mathbf{c} = (\, 2\,c_3 - c_4, -c_3 - c_4, c_3, c_4 \,)^T,$$

where c_3, c_4 — the free variables — are arbitrary. Any nonzero choice of c_3, c_4 will produce a nontrivial linear combination

$$(2\,c_3 - c_4)\mathbf{v}_1 + (-c_3 - c_4)\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4 = \mathbf{0}$$

that adds up to the zero vector. Therefore, the vectors (2.15) are linearly dependent.

In fact, Theorem 1.45 says that in this particular case we didn't even need to do the row reduction if we only needed to answer the question of linear dependence or linear independence. Any coefficient matrix with more columns than rows automatically has a nontrivial solution to the associated homogeneous system. This implies the following:

Lemma 2.22. *Any collection of $k > n$ vectors in $\mathbb{R}^n$ is linearly dependent.*

Warning: The converse to this lemma is *not* true. For example, the two vectors $\mathbf{v}_1 = (\,1, 2, 3\,)^T$ and $\mathbf{v}_2 = (\,-2, -4, -6\,)^T$ in $\mathbb{R}^3$ are linearly dependent since $2\,\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{0}$. For a collection of n or fewer vectors in $\mathbb{R}^n$, one does need to perform the elimination to calculate the rank of the corresponding matrix.

Lemma 2.22 is a particular case of the following general characterization of linearly independent vectors.

Proposition 2.23. *A set of k vectors in $\mathbb{R}^n$ is linearly independent if and only if the corresponding $n \times k$ matrix A has rank k. In particular, this requires $k \leq n$.*

Or, to state the result another way, the vectors are linearly independent if and only if the linear system $A\,\mathbf{c} = \mathbf{0}$ has no free variables. The proposition is an immediate corollary of Propositions 2.20 and 1.45.

Example 2.21. (*continued*) Let us now see which vectors $\mathbf{b} \in \mathbb{R}^3$ lie in the span of the vectors (2.15). This will be the case if and only if the linear system $A\,\mathbf{x} = \mathbf{b}$ has a solution. Since the resulting row echelon form (2.16) has a row of all zeros, there will be a compatibility condition on the entries of $\mathbf{b}$, and therefore not every vector lies in the span. To find the precise condition, we augment the coefficient matrix, and apply the same row operations, leading to the reduced augmented matrix

$$\left(\begin{array}{cccc|c} 1 & 3 & 1 & 4 & b_1 \\ 0 & -6 & -6 & -6 & b_2 - 2\,b_1 \\ 0 & 0 & 0 & 0 & b_3 + \frac{7}{6}\,b_2 - \frac{4}{3}\,b_1 \end{array}\right).$$

Therefore, $\mathbf{b} = (\,b_1, b_2, b_3\,)^T$ lies in the span of these four vectors if and only if

$$-\tfrac{4}{3}\,b_1 + \tfrac{7}{6}\,b_2 + b_3 = 0.$$

In other words, these four vectors only span a plane in $\mathbb{R}^3$.

The same method demonstrates that a collection of vectors will span all of $\mathbb{R}^n$ if and only if the row echelon form of the associated matrix contains no all zero rows, or, equivalently, the rank is equal to n, the number of rows in the matrix.

Proposition 2.24. *A collection of k vectors will span $\mathbb{R}^n$ if and only if their $n \times k$ matrix has rank n. In particular, this requires $k \geq n$.*

Warning: Not every collection of n or more vectors in $\mathbb{R}^n$ will span all of $\mathbb{R}^n$. A counterexample is provided by the vectors (2.15).

2.4. Bases.

In order to span a vector space or subspace, we must use a sufficient number of distinct elements. On the other hand, including too many elements in the spanning set will violate linear independence, and cause redundancies. The optimal spanning sets are those that are also linearly independent. By combining the properties of span and linear independence, we arrive at the all-important concept of a "basis".

Definition 2.25. A *basis* of a vector space V is a finite collection of elements $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$ which (a) span V, and (b) are linearly independent.

Bases are absolutely fundamental in all areas of linear algebra and linear analysis, including matrix algebra, geometry of Euclidean space, solutions to linear differential equations, both ordinary and partial, linear boundary value problems, Fourier analysis, signal and image processing, data compression, control systems, and so on.

Example 2.26. The standard basis of $\mathbb{R}^n$ consists of the n vectors

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \dots \qquad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \tag{2.17}$$

so that $\mathbf{e}_i$ is the vector with 1 in the i^{th} slot and 0's elsewhere. We already encountered these vectors as the columns of the $n \times n$ identity matrix, as in (1.39). They clearly span $\mathbb{R}^n$ since we can write any vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \, \mathbf{e}_1 + x_2 \, \mathbf{e}_2 + \, \cdots \, + x_n \, \mathbf{e}_n, \tag{2.18}$$

as a linear combination, whose coefficients are the entries of $\mathbf{x}$. Moreover, the only linear combination that gives the zero vector $\mathbf{x} = \mathbf{0}$ is the trivial one $x_1 = \cdots = x_n = 0$, and so $\mathbf{e}_1, \dots, \mathbf{e}_n$ are linearly independent.

Remark: In the three-dimensional case $\mathbb{R}^3$, a common physical notation for the standard basis is

$$\mathbf{i} = \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{j} = \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{k} = \mathbf{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{2.19}$$

There are many other possible bases for $\mathbb{R}^3$. Indeed, any three non-coplanar vectors can be used to form a basis. This is a consequence of the following general characterization of bases in $\mathbb{R}^n$.

Theorem 2.27. *Every basis of $\mathbb{R}^n$ contains exactly n vectors. A set of n vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$ is a basis if and only if the $n \times n$ matrix $A = (\, \mathbf{v}_1 \, \dots \, \mathbf{v}_n \,)$ is nonsingular.*

Proof: This is a direct consequence of Theorem 2.20. Linear independence requires that the only solution to the homogeneous system $A\,\mathbf{x} = \mathbf{0}$ is the trivial one $\mathbf{x} = \mathbf{0}$. Secondly, a vector $\mathbf{b} \in \mathbb{R}^n$ will lie in the span of $\mathbf{v}_1, \dots, \mathbf{v}_n$ if and only if the linear system $A\,\mathbf{x} = \mathbf{b}$ has a solution. For $\mathbf{v}_1, \dots, \mathbf{v}_n$ to span $\mathbb{R}^n$, this must hold for all possible right hand sides $\mathbf{b}$. Theorem 1.7 tells us that both results require that A be nonsingular, i.e., have maximal rank n. *Q.E.D.*

Thus, every basis of n-dimensional Euclidean space $\mathbb{R}^n$ contains the same number of vectors, namely n. This is a general fact, and motivates a linear algebra characterization of dimension.

Theorem 2.28. *Suppose the vector space V has a basis $\mathbf{v}_1, \dots, \mathbf{v}_n$. Then every other basis of V has the same number of elements in it. This number is called the* dimension *of V, and written* $\dim V = n$.

The proof of Theorem 2.28 rests on the following lemma.

Lemma 2.29. *Suppose $\mathbf{v}_1, \dots, \mathbf{v}_n$ span a vector space V. Then every set of $k > n$ elements $\mathbf{w}_1, \dots, \mathbf{w}_k \in V$ is linearly dependent.*

Proof: Let us write each element

$$\mathbf{w}_j = \sum_{i=1}^{n} a_{ij}\, \mathbf{v}_i, \qquad\qquad j = 1, \dots, k,$$

as a linear combination of the spanning set. Then

$$c_1\, \mathbf{w}_1 + \, \cdots \, + c_k\, \mathbf{w}_k = \sum_{i=1}^{n} \sum_{j=1}^{k} a_{ij}\, c_j \mathbf{v}_i.$$

This linear combination will be zero whenever $\mathbf{c} = (\, c_1, c_2, \dots, c_k \,)^T$ solves the homogeneous linear system

$$\sum_{j=1}^{k} a_{ij}\, c_j = 0, \qquad\qquad i = 1, \dots, n,$$

consisting of n equations in $k > n$ unknowns. Theorem 1.45 guarantees that every homogeneous system with more unknowns than equations always has a non-trivial solution $\mathbf{c} \neq \mathbf{0}$, and this immediately implies that $\mathbf{w}_1, \dots, \mathbf{w}_k$ are linearly dependent. *Q.E.D.*

Proof of Theorem 2.28: Suppose we have two bases containing a different number of elements. By definition, the smaller basis spans the vector space. But then Lemma 2.29 tell us that the elements in the larger purported basis must be linearly dependent. This contradicts our assumption that both sets are bases, and proves the theorem. *Q.E.D.*

As a direct consequence, we can now provide a precise meaning to the optimality property of bases.

Theorem 2.30. *Suppose V is an n-dimensional vector space. Then*

(a) *Every set of more than n elements of V is linearly dependent.*

(b) *No set of less than n elements spans V.*

(c) *A set of n elements forms a basis if and only if it spans V.*

(d) *A set of n elements forms a basis if and only if it is linearly independent.*

In other words, once we determine the dimension of a vector space, to check that a given collection with the correct number of elements forms a basis, we only need check one of the two defining properties: span or linear independence. Thus, n elements that span an n-dimensional vector space are automatically linearly independent and hence form a basis; vice versa, n linearly independent elements of n-dimensional vector space automatically span the space and so form a basis.

Example 2.31. The standard basis of the space $\mathcal{P}^{(n)}$ of polynomials of degree $\leq n$ is given by the $n+1$ monomials $1, x, x^2, \ldots, x^n$. (A formal proof of linear independence appears in Exercise .) We conclude that the vector space $\mathcal{P}^{(n)}$ has dimension $n+1$. Thus, any collection of $n+2$ or more polynomials of degree $\leq n$ is automatically linearly dependent. Any other basis of $\mathcal{P}^{(n)}$ must contain precisely $n+1$ polynomials. But, not every collection of $n+1$ polynomials in $\mathcal{P}^{(n)}$ is a basis — they must be linearly independent. See Exercise for details.

Remark: By definition, every vector space of dimension $1 \leq n < \infty$ has a basis. If a vector space V has no basis, it is either the trivial vector space $V = \{\mathbf{0}\}$, which by convention has dimension 0, or, by definition, its dimension is infinite. An infinite-dimensional vector space necessarily contains an infinite collection of linearly independent vectors, and hence no (finite) basis. Examples of infinite-dimensional vector spaces include most spaces of functions, such as the spaces of continuous, differentiable, or mean zero functions, as well as the space of *all* polynomials, and the space of solutions to a linear homogeneous partial differential equation. On the other hand, the solution space for a homogeneous linear ordinary differential equation turns out to be a finite-dimensional vector space. The most important example of an infinite-dimensional vector space, "Hilbert space", to be introduced in Chapter 12,is essential to modern analysis and function theory, [**122**, **126**], as well as providing the theoretical setting for all of quantum mechanics, [**100**, **104**].

Warning: There is a well-developed concept of a "complete basis" of such infinite-dimensional function spaces, essential in Fourier analysis,[**122**, **126**], but this requires additional analytical constructions that are beyond our present abilities. Thus, in this book the term "basis" *always* means a finite collection of vectors in a finite-dimensional vector space.

Lemma 2.32. *The elements $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis of V if and only if every $\mathbf{x} \in V$ can be written* uniquely *as a linear combination thereof*:

$$\mathbf{x} = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \sum_{i=1}^{n} c_i \mathbf{v}_i \tag{2.20}$$

Proof: The condition that the basis span V implies every $\mathbf{x} \in V$ can be written as some linear combination of the basis elements. Suppose we can write an element

$$\mathbf{x} = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \widehat{c}_1 \mathbf{v}_1 + \cdots + \widehat{c}_n \mathbf{v}_n$$

as two different combinations. Subtracting one from the other, we find

$$(c_1 - \widehat{c}_1)\mathbf{v}_1 + \cdots + (c_n - \widehat{c}_n)\mathbf{v}_n = \mathbf{0}.$$

Linear independence of the basis elements implies that the coefficients $c_i - \widehat{c}_i = 0$. We conclude that $c_i = \widehat{c}_i$, and hence the linear combinations are the same. *Q.E.D.*

The coefficients $(c_1, \ldots, c_n)$ in (2.20) are called the *coordinates* of the vector $\mathbf{x}$ with respect to the given basis. For the standard basis (2.17) of $\mathbb{R}^n$, the coordinates of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ are its entries — i.e., its usual Cartesian coordinates, cf. (2.18). In many applications, an inspired change of basis will lead to a better adapted coordinate system, thereby simplifying the computations.

Example 2.33. *A Wavelet Basis*. The vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \tag{2.21}$$

form a basis of $\mathbb{R}^4$. This is verified by performing Gaussian elimination on the corresponding 4×4 matrix

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

to check that it is nonsingular. This basis is a very simple example of a *wavelet basis*; the general case will be discussed in Section 13.2. Wavelets arise in modern applications to signal and digital image processing, [**43**, **128**].

How do we find the coordinates of a vector $\mathbf{x}$ relative to the basis? We need to fix the coefficients c_1, c_2, c_3, c_4 so that

$$\mathbf{x} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3 + c_4 \mathbf{v}_4.$$

We rewrite this equation in matrix form

$$\mathbf{x} = A\,\mathbf{c} \qquad \text{where} \qquad \mathbf{c} = (c_1, c_2, c_3, c_4)^T.$$

For example, solving the linear system for the vector $\mathbf{x} = (4, -2, 1, 5)^T$ by Gaussian Elimination produces the unique solution $c_1 = 2, c_2 = -1, c_3 = 3, c_4 = -2$, which are its coordinates in the wavelet basis:

$$\begin{pmatrix} 4 \\ -2 \\ 1 \\ 5 \end{pmatrix} = 2\mathbf{v}_1 - \mathbf{v}_2 + 3\mathbf{v}_3 - 2\mathbf{v}_4 = 2\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} + 3\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix} - 2\begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}.$$

In general, to find the coordinates of a vector $\mathbf{x}$ with respect to a new basis of $\mathbb{R}^n$ requires the solution of a linear system of equations, namely

$$A\,\mathbf{c} = \mathbf{x} \qquad \text{for} \qquad \mathbf{c} = A^{-1}\mathbf{x}. \tag{2.22}$$

Here $\mathbf{x} = (\,x_1, x_2, \ldots, x_n\,)^T$ are the Cartesian coordinates of $\mathbf{x}$, with respect to the standard basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$, while $\mathbf{c} = (\,c_1, c_2, \ldots, c_n\,)^T$ denotes its coordinates with respect to the new basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ formed by the columns of the coefficient matrix $A = (\,\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n\,)$. In practice, one solves for the coordinates by using Gaussian Elimination, not by matrix inversion.

Why would one want to change bases? The answer is *simplification* and *speed* — many computations and formulas become much easier, and hence faster, to perform in a basis that is adapted to the problem at hand. In signal processing, the wavelet basis is particularly appropriate for denoising, compression, and efficient storage of signals, including audio, still images, videos, medical images, geophysical images, and so on. These processes would be quite time-consuming, if not impossible in the case of video processing, to accomplish in the standard basis. Many other examples will appear throughout the text.

2.5. The Fundamental Matrix Subspaces.

Let us now return to the general study of linear systems of equations, which we write in our usual matrix form

$$A\,\mathbf{x} = \mathbf{b}\,. \tag{2.23}$$

Here A is an $m \times n$ matrix, where m is the number of equations and n the number of unknowns, i.e., the entries of $\mathbf{x} \in \mathbb{R}^n$.

Kernel and Range

There are four important vector subspaces associated with any matrix, which play a key role in the interpretation of our solution algorithm. The first two of these subspaces are defined as follows.

Definition 2.34. The *range* of an $m \times n$ matrix A is the subspace $\operatorname{rng} A \subset \mathbb{R}^m$ spanned by the columns of A. The *kernel* or *null space* of A is the subspace $\ker A \subset \mathbb{R}^n$ consisting of all vectors which are annihilated by A, so

$$\ker A = \{\, \mathbf{z} \in \mathbb{R}^n \mid\ A\,\mathbf{z} = \mathbf{0}\,\} \subset \mathbb{R}^n. \tag{2.24}$$

An alternative name for the range is the *column space* of the matrix. By definition, a vector $\mathbf{b} \in \mathbb{R}^m$ belongs to $\operatorname{rng} A$ if and only if it can be written as a linear combination,

$$\mathbf{b} = x_1\,\mathbf{v}_1 + \ \cdots\ + x_n\,\mathbf{v}_n,$$

of the columns of $A = (\,\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n\,)$. By our basic matrix multiplication formula (2.14), the right hand side of this equation equals the product $A\,\mathbf{x}$ of the matrix A with the column vector $\mathbf{x} = (\,x_1, x_2, \ldots, x_n\,)^T$, and hence $\mathbf{b} = A\,\mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n$, so

$$\operatorname{rng} A = \{\, A\,\mathbf{x} \mid\ \mathbf{x} \in \mathbb{R}^n\,\} \subset \mathbb{R}^m. \tag{2.25}$$

Therefore, *a vector* $\mathbf{b}$ *lies in the range of* A *if and only if the linear system* $A\,\mathbf{x} = \mathbf{b}$ *has a solution.* Thus, the compatibility conditions for linear systems can be re-interpreted as the conditions for a vector to lie in the range of the coefficient matrix.

A common alternative name for the kernel is the *null space* of the matrix A. The kernel of A is the set of solutions to the homogeneous system $A\,\mathbf{z} = \mathbf{0}$. The proof that $\ker A$ is a subspace requires us to verify the usual closure conditions. Suppose that $\mathbf{z}, \mathbf{w} \in \ker A$, so that $A\,\mathbf{z} = \mathbf{0} = A\,\mathbf{w}$. Then, for any scalars c, d,

$$A(c\,\mathbf{z} + d\,\mathbf{w}) = c\,A\,\mathbf{z} + d\,A\,\mathbf{w} = \mathbf{0},$$

which implies that $c\,\mathbf{z} + d\,\mathbf{w} \in \ker A$, proving that $\ker A$ is a subspace. This fact can be re-expressed as the following *superposition principle* for solutions to a homogeneous system of linear equations.

Theorem 2.35. *If* $\mathbf{z}_1, \dots, \mathbf{z}_k$ *are solutions to a homogeneous linear system* $A\,\mathbf{z} = \mathbf{0}$, *then so is any linear combination* $c_1\,\mathbf{z}_1 + \cdots + c_k\,\mathbf{z}_k$.

Warning: The set of solutions to an inhomogeneous linear system $A\,\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \neq \mathbf{0}$ is *not* a subspace.

Example 2.36. Let us compute the kernel of the matrix $A = \begin{pmatrix} 1 & -2 & 0 & 3 \\ 2 & -3 & -1 & -4 \\ 3 & -5 & -1 & -1 \end{pmatrix}$. Since we are solving the homogeneous system $A\,\mathbf{x} = \mathbf{0}$, we only need perform the elementary row operations on A itself. The resulting row echelon form $U = \begin{pmatrix} 1 & -2 & 0 & 3 \\ 0 & 1 & -1 & -10 \\ 0 & 0 & 0 & 0 \end{pmatrix}$ corresponds to the equations $x - 2y + 3w = 0$, $y - z - 10w = 0$. The free variables are z, w. The general solution to the homogeneous system is

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 2z + 17w \\ z + 10w \\ z \\ w \end{pmatrix} = z \begin{pmatrix} 2 \\ 1 \\ 1 \\ 0 \end{pmatrix} + w \begin{pmatrix} 17 \\ 10 \\ 0 \\ 1 \end{pmatrix},$$

which, for arbitrary scalars z, w, describes the most general vector in $\ker A$. Thus, the kernel of this matrix is the two-dimensional subspace of $\mathbb{R}^4$ spanned by the linearly independent vectors $(\,2, 1, 1, 0\,)^T$, $(\,17, 10, 0, 1\,)^T$.

Remark: This example is indicative of a general method for finding a basis for $\ker A$ which will be developed in more detail in the following section.

Once we know the kernel of the coefficient matrix A, i.e., the space of solutions to the homogeneous system $A\,\mathbf{z} = \mathbf{0}$, we are in a position to completely characterize the solutions to the inhomogeneous linear system (2.23).

Theorem 2.37. *The linear system* $A\,\mathbf{x} = \mathbf{b}$ *has a solution* $\mathbf{x}^\star$ *if and only if* $\mathbf{b}$ *lies in the range of* A. *If this occurs, then* $\mathbf{x}$ *is a solution to the linear system if and only if*

$$\mathbf{x} = \mathbf{x}^\star + \mathbf{z}, \tag{2.26}$$

where $\mathbf{z} \in \ker A$ *is an arbitrary element of the kernel of* A.

Proof: We already demonstrated the first part of the theorem. If $A\,\mathbf{x} = \mathbf{b} = A\,\mathbf{x}^\star$ are any two solutions, then their difference $\mathbf{z} = \mathbf{x} - \mathbf{x}^\star$ satisfies

$$A\,\mathbf{z} = A(\mathbf{x} - \mathbf{x}^\star) = A\,\mathbf{x} - A\,\mathbf{x}^\star = \mathbf{b} - \mathbf{b} = \mathbf{0},$$

and hence $\mathbf{z}$ belongs to the kernel of A. Therefore, $\mathbf{x}$ and $\mathbf{x}^\star$ are related by formula (2.26), which proves the second part of the theorem. *Q.E.D.*

Therefore, to construct the most general solution to an inhomogeneous system, we need only know one *particular solution* $\mathbf{x}^\star$, along with the general solution $\mathbf{z} \in \ker A$ to the homogeneous equation. This construction should remind the reader of the method of solution for inhomogeneous linear ordinary differential equations. Indeed, both linear algebraic systems and linear ordinary differential equations are but two particular instances of the general theory of linear systems, to be developed in Chapter 7. In particular, we can characterize the case when the linear system has a unique solution in any of the following equivalent ways.

Proposition 2.38. *Let A be an $m \times n$ matrix. Then the following conditions are equivalent:*

(*i*) $\ker A = \{\mathbf{0}\}$.
(*ii*) $\operatorname{rank} A = n$
(*iii*) *There are no free variables in the linear system $A\,\mathbf{x} = \mathbf{b}$.*
(*iv*) *The system $A\,\mathbf{x} = \mathbf{b}$ has a unique solution for each $\mathbf{b} \in \operatorname{rng} A$.*

Specializing even further to square matrices, we can characterize invertibility by looking either at its kernel or at its range.

Proposition 2.39. *If A is a square matrix, then the following three conditions are equivalent*: (*i*) *A is nonsingular;* (*ii*) $\ker A = \{\mathbf{0}\}$; (*iii*) $\operatorname{rng} A = \mathbb{R}^n$.

Example 2.40. Consider the system $A\,\mathbf{x} = \mathbf{b}$, where

$$A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -2 \\ 1 & -2 & 3 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

where the right hand side of the system will be left arbitrary. Applying our usual Gaussian Elimination procedure to the augmented matrix

$$\left(\begin{array}{ccc|c} 1 & 0 & -1 & b_1 \\ 0 & 1 & -2 & b_2 \\ 1 & -2 & 3 & b_3 \end{array}\right) \quad \text{leads to the row echelon form} \quad \left(\begin{array}{ccc|c} 1 & 0 & -1 & b_1 \\ 0 & 1 & -2 & b_2 \\ 0 & 0 & 0 & b_3 + 2\,b_2 - b_1 \end{array}\right).$$

The system has a solution if and only if the resulting compatibility condition

$$-\,b_1 + 2\,b_2 + b_3 = 0 \tag{2.27}$$

holds. This equation serves to characterize the vectors $\mathbf{b}$ that belong to the range of the matrix A, which is therefore a certain plane in $\mathbb{R}^3$ passing through the origin.

To characterize the kernel of A, we take $\mathbf{b} = \mathbf{0}$, and solve the homogeneous system $A\,\mathbf{z} = \mathbf{0}$. The row echelon form corresponds to the reduced system

$$z_1 - z_3 = 0, \qquad z_2 - 2\,z_3 = 0.$$

The free variable is z_3, and the equations are solved to give

$$z_1 = c, \qquad z_2 = 2\,c, \qquad z_3 = c,$$

where c is arbitrary. The general solution to the homogeneous system is $\mathbf{z} = (\,c, 2\,c, c\,)^T = c\,(\,1, 2, 1\,)^T$, and so the kernel is the line in the direction of the vector $(\,1, 2, 1\,)^T$.

If we take $\mathbf{b} = (\,3, 1, 1\,)^T$ — which satisfies (2.27) and hence lies in the range of A — then the general solution to the inhomogeneous system $A\,\mathbf{x} = \mathbf{b}$ is

$$x_1 = 3 + c, \qquad x_2 = 1 + 2\,c, \qquad x_3 = c,$$

where c is an arbitrary scalar. We can write the solution in the form (2.26), namely

$$\mathbf{x} = \begin{pmatrix} 3 + c \\ 1 + 2\,c \\ c \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \\ 0 \end{pmatrix} + c \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \mathbf{x}^\star + \mathbf{z},$$

where $\mathbf{x}^\star = (\,3, 1, 0\,)^T$ plays the role of the particular solution, and $\mathbf{z} = c\,(\,1, 2, 1\,)^T$ is the general element of the kernel.

The Superposition Principle

The principle of superposition lies at the heart of linearity. For homogeneous systems, superposition allows one to generate new solutions by combining known solutions. For inhomogeneous systems, superposition combines the solutions corresponding to different inhomogeneities or forcing functions. Superposition is the reason why linear systems are so important for applications and why they are so much easier to solve. We shall explain the superposition principle in the context of inhomogeneous linear algebraic systems. In Chapter 7 we shall see that the general principle applies as stated to completely general linear systems, including linear differential equations, linear boundary value problems, linear integral equations, etc.

Suppose we have found particular solutions $\mathbf{x}_1^\star$ and $\mathbf{x}_2^\star$ to two inhomogeneous linear systems

$$A\,\mathbf{x} = \mathbf{b}_1, \qquad A\,\mathbf{x} = \mathbf{b}_2,$$

that have the *same* coefficient matrix A. Consider the system

$$A\,\mathbf{x} = c_1\,\mathbf{b}_1 + c_2\,\mathbf{b}_2,$$

in which the right hand side is a linear combination or superposition of the previous two. Then a particular solution to the combined system is given by the same linear combination of the previous solutions:

$$\mathbf{x}^\star = c_1\,\mathbf{x}_1^\star + c_2\,\mathbf{x}_2^\star.$$

The proof is easy; we use the rules of matrix arithmetic to compute

$$A\,\mathbf{x}^\star = A(c_1\,\mathbf{x}_1^\star + c_2\,\mathbf{x}_2^\star) = c_1\,A\,\mathbf{x}_1^\star + c_2\,A\,\mathbf{x}_2^\star = c_1\,\mathbf{b}_1 + c_2\,\mathbf{b}_2.$$

In many applications, the inhomogeneities $\mathbf{b}_1, \mathbf{b}_2$ represent external forces, and the solutions $\mathbf{x}_1^\star, \mathbf{x}_2^\star$ represent the response of the physical apparatus to the force. The linear superposition principle says that if we know how the system responds to the individual forces, we immediately know its response to any combination thereof. The precise details of the system are irrelevant — all that is required is linearity.

Example 2.41. For example, the system

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}$$

models the mechanical response of a pair of masses connected by springs to an external force. The solution $\mathbf{x} = (\,x, y\,)^T$ represent the respective displacements of the masses, while the components of the right hand side $\mathbf{f} = (\,f, g\,)^T$ represent the respective forces applied to each mass. (See Chapter 6 for full details.) We can compute the response of the system $\mathbf{x}_1^\star = \left(\,\frac{4}{15}, -\frac{1}{15}\,\right)^T$ to a unit force $\mathbf{e}_1 = (\,1, 0\,)^T$ on the first mass, and the response $\mathbf{x}_2^\star = \left(-\frac{1}{15}, \frac{4}{15}\,\right)^T$ to a unit force $\mathbf{e}_2 = (\,0, 1\,)^T$ on the second mass. We then know the response of the system to a general force, since we can write

$$\mathbf{f} = \begin{pmatrix} f \\ g \end{pmatrix} = f\,\mathbf{e}_1 + g\,\mathbf{e}_2 = f\begin{pmatrix} 1 \\ 0 \end{pmatrix} + g\begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and hence the solution is

$$\mathbf{x} = f\,\mathbf{x}_1^\star + g\,\mathbf{x}_2^\star = f\begin{pmatrix} \frac{4}{15} \\ -\frac{1}{15} \end{pmatrix} + g\begin{pmatrix} -\frac{1}{15} \\ \frac{4}{15} \end{pmatrix} = \begin{pmatrix} \frac{4}{15}\,f - \frac{1}{15}\,g \\ -\frac{1}{15}\,f + \frac{4}{15}\,g \end{pmatrix}.$$

The preceding construction is easily extended to several inhomogeneities, and the result is a general *Superposition Principle* for inhomogeneous linear systems.

Theorem 2.42. *Suppose that we know particular solutions $\mathbf{x}_1^\star, \dots, \mathbf{x}_k^\star$ to each of the inhomogeneous linear systems*

$$A\,\mathbf{x} = \mathbf{b}_1, \qquad A\,\mathbf{x} = \mathbf{b}_2, \qquad \dots \qquad A\,\mathbf{x} = \mathbf{b}_k, \tag{2.28}$$

where $\mathbf{b}_1, \dots, \mathbf{b}_k \in \operatorname{rng} A$. Then, for any choice of scalars $c_1, \dots, c_k$, a particular solution to the combined system

$$A\,\mathbf{x} = c_1\mathbf{b}_1 + \,\cdots\, + c_k\mathbf{b}_k \tag{2.29}$$

is the same superposition

$$\mathbf{x}^\star = c_1\,\mathbf{x}_1^\star + \,\cdots\, + c_k\,\mathbf{x}_k^\star \tag{2.30}$$

of individual solutions. The general solution to (2.29) *is*

$$\mathbf{u} = \mathbf{x}^\star + \mathbf{z} = c_1\,\mathbf{x}_1^\star + \,\cdots\, + c_k\,\mathbf{x}_k^\star + \mathbf{z}, \tag{2.31}$$

where $\mathbf{z}$ is the general solution to the homogeneous equation $A\,\mathbf{z} = \mathbf{0}$.

In particular, if we know particular solutions $\mathbf{x}_1^\star, \dots, \mathbf{x}_m^\star$ to

$$A\,\mathbf{x} = \mathbf{e}_i, \qquad \text{for each} \qquad i = 1, \dots, m, \tag{2.32}$$

where $\mathbf{e}_1, \dots, \mathbf{e}_m$ are the standard basis vectors of $\mathbb{R}^m$, cf. (2.17), then we can reconstruct a particular solution $\mathbf{x}^\star$ to the general linear system $A\,\mathbf{x} = \mathbf{b}$ by first writing

$$\mathbf{b} = b_1\,\mathbf{e}_1 + \,\cdots\, + b_m\,\mathbf{e}_m$$

as a linear combination of the basis vectors, and then using superposition to form

$$\mathbf{x}^\star = b_1\,\mathbf{x}_1^\star + \,\cdots\, + b_m\,\mathbf{x}_m^\star. \tag{2.33}$$

However, for linear algebraic systems, the practical value of this insight is rather limited. Indeed, in the case when A is square and nonsingular, the superposition method is just a reformulation of the method of computing the inverse of the matrix. Indeed, the vectors $\mathbf{x}_1^\star, \dots, \mathbf{x}_n^\star$ which satisfy (2.32) are just the columns of A^{-1}, cf. (1.39), and the superposition formula (2.33) is, using (2.14), precisely the solution formula $\mathbf{x}^\star = A^{-1}\mathbf{b}$ that we abandoned in practical computations, in favor of the more efficient Gaussian elimination method. Nevertheless, the implications of this result turn out to be of great importance in the study of linear boundary value problems.

Adjoint Systems, Cokernel, and Corange

A linear system of m equations in n unknowns results in an $m \times n$ coefficient matrix A. The transposed matrix A^T will be of size $n \times m$, and forms the coefficient of an associated linear system consisting of n equations in m unknowns.

Definition 2.43. The *adjoint*[†] to a linear system $A\,\mathbf{x} = \mathbf{b}$ of m equations in n unknowns is the linear system

$$A^T\mathbf{y} = \mathbf{f} \tag{2.34}$$

of n equations in m unknowns. Here $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{f} \in \mathbb{R}^n$.

Example 2.44. Consider the linear system

$$\begin{aligned} x_1 - 3\,x_2 - 7\,x_3 + 9\,x_4 &= b_1, \\ x_2 + 5\,x_3 - 3\,x_4 &= b_2, \\ x_1 - 2\,x_2 - 2\,x_3 + 6\,x_4 &= b_3, \end{aligned} \tag{2.35}$$

of three equations in four unknowns. Its coefficient matrix is $A = \begin{pmatrix} 1 & -3 & -7 & 9 \\ 0 & 1 & 5 & -3 \\ 1 & -2 & -2 & 6 \end{pmatrix}$ has transpose $A^T = \begin{pmatrix} 1 & 0 & 1 \\ -3 & 1 & -2 \\ -7 & 5 & -2 \\ 9 & -3 & 6 \end{pmatrix}$. Thus, the adjoint system to (2.35) is the following

[†] *Warning*: Many texts misuse the term "adjoint" to describe the classical *adjugate* or cofactor matrix. These are completely unrelated, and the latter will play no role in this book.

system of four equations in three unknowns:

$$\begin{aligned} y_1 + y_3 &= f_1, \\ -3\,y_1 + y_2 - 2\,y_3 &= f_2, \\ -7\,y_1 + 5\,y_2 - 2\,y_3 &= f_3, \\ 9\,y_1 - 3\,y_2 + 6\,y_3 &= f_4. \end{aligned} \tag{2.36}$$

On the surface, there appears to be little direct connection between the solutions to a linear system and its adjoint. Nevertheless, as we shall soon see (and then in even greater depth in Sections 5.6 and 8.5) there are remarkable, but subtle interrelations between the two. These turn out to have significant consequences, not only for linear algebraic systems but to even more profound extensions to differential equations.

To this end, we use the adjoint system to define the other two fundamental subspaces associated with a coefficient matrix A.

Definition 2.45. The *corange* of an $m \times n$ matrix A is the range of its transpose,

$$\operatorname{corng} A = \operatorname{rng} A^T = \left\{ A^T\mathbf{y} \;\middle|\; \mathbf{y} \in \mathbb{R}^m \right\} \subset \mathbb{R}^n. \tag{2.37}$$

The *cokernel* or *left null space* of A is the kernel of its transpose,

$$\operatorname{coker} A = \ker A^T = \left\{ \mathbf{w} \in \mathbb{R}^m \;\middle|\; A^T\mathbf{w} = \mathbf{0} \right\} \subset \mathbb{R}^m, \tag{2.38}$$

that is, the set of solutions to the homogeneous adjoint system.

The corange coincides with the subspace of $\mathbb{R}^n$ spanned by the rows of A, and is sometimes referred to as the *row space*. As a consequence of Theorem 2.37, the adjoint system $A^T\mathbf{y} = \mathbf{f}$ has a solution if and only if $\mathbf{f} \in \operatorname{rng} A^T = \operatorname{corng} A$.

Example 2.46. To solve the linear system (2.35) appearing above, we perform Gaussian Elimination on its augmented matrix $\left(\begin{array}{cccc|c} 1 & -3 & -7 & 9 & b_1 \\ 0 & 1 & 5 & -3 & b_2 \\ 1 & -2 & -2 & 6 & b_3 \end{array}\right)$ that reduces it to the row echelon form $\left(\begin{array}{cccc|c} 1 & -3 & -7 & 9 & b_1 \\ 0 & 1 & 5 & -3 & b_2 \\ 0 & 0 & 0 & 0 & b_3 - b_2 - b_1 \end{array}\right)$. Thus, the system has a solution if and only if $\mathbf{b} \in \operatorname{rng} A$ satisfies the compatibility condition $-b_1 - b_2 + b_3 = 0$. For such vectors, the general solution is

$$\mathbf{x} = \begin{pmatrix} b_1 + 3\,b_2 - 8\,x_3 \\ b_2 - 5\,x_3 + 3\,x_4 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} b_1 + 3\,b_2 \\ b_2 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -8 \\ -5 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 3 \\ 0 \\ 1 \end{pmatrix}.$$

In the second expression, the first vector is a particular solution and the remaining terms constitute the general element of the two-dimensional kernel of A.

The solution to the adjoint system (2.36) is also obtained by Gaussian Elimination starting with its augmented matrix $\left(\begin{array}{ccc|c} 1 & 0 & 1 & f_1 \\ -3 & 1 & -2 & f_2 \\ -7 & 5 & -2 & f_3 \\ 9 & -3 & 6 & f_4 \end{array}\right)$. The resulting row echelon form is $\left(\begin{array}{ccc|c} 1 & 0 & 1 & f_1 \\ 0 & 1 & 1 & f_2 \\ 0 & 0 & 0 & f_3 - 5\,f_2 - 8\,f_1 \\ 0 & 0 & 0 & f_4 + 3\,f_2 \end{array}\right)$. Thus, there are two compatibility constraints required for a solution to the adjoint system: $-8\,f_1 - 5\,f_2 + f_3 = 0$, $3\,f_2 + f_4 = 0$. These are the conditions required for the right hand side to belong to the corange: $\mathbf{f} \in \operatorname{rng} A^T = \operatorname{corng} A$. If satisfied, the adjoint system has the general solution depending on the single free variable y_3:

$$\mathbf{y} = \begin{pmatrix} f_1 - y_3 \\ 3\,f_1 + f_2 - y_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ 3\,f_1 + f_2 \\ 0 \end{pmatrix} + y_3 \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}.$$

In the latter formula, the first term represents a particular solution, while the second is the general element of $\ker A^T = \operatorname{coker} A$.

The Fundamental Theorem of Linear Algebra

The four fundamental subspaces associated with an $m \times n$ matrix A, then, are its range, corange, kernel and cokernel. The range and cokernel are subspaces of $\mathbb{R}^m$, while the kernel and corange are subspaces of $\mathbb{R}^n$. Moreover, these subspaces are not completely arbitrary, but are, in fact, profoundly related through both their numerical and geometric properties.

The *Fundamental Theorem of Linear Algebra*[†] states that their dimensions are entirely prescribed by the rank (and size) of the matrix.

Theorem 2.47. *Let A be an $m \times n$ matrix of rank r. Then*

$$\begin{aligned} \dim \operatorname{corng} A = \dim \operatorname{rng} A = \operatorname{rank} A = \operatorname{rank} A^T = r, \\ \dim \ker A = n - r, \qquad \dim \operatorname{coker} A = m - r. \end{aligned} \tag{2.39}$$

Remark: Thus, the rank of a matrix, i.e., the number of pivots, indicates the number of linearly independent columns, which, remarkably, is always the same as the number of linearly independent rows! A matrix and its transpose have the same rank, i.e., the same number of pivots, even though their row echelon forms are quite different, and are rarely transposes of each other. Theorem 2.47 also proves our earlier contention that the rank of a matrix is an intrinsic quantity, and does not depend on which specific elementary row operations are employed during the reduction process, nor on the final row echelon form.

[†] Not to be confused with the Fundamental Theorem of Algebra, that states that every polynomial has a complex root; see Theorem 16.62.

Proof: Since the dimension of a subspace is prescribed by the number of vectors in any basis, we need to relate bases of the fundamental subspaces to the rank of the matrix. Rather than present the general argument, we will show how to construct bases for each of the subspaces in a particular instance, and thereby illustrate the method of proof. Consider the matrix

$$A = \begin{pmatrix} 2 & -1 & 1 & 2 \\ -8 & 4 & -6 & -4 \\ 4 & -2 & 3 & 2 \end{pmatrix}.$$

The row echelon form of A is obtained in the usual manner: $U = \begin{pmatrix} 2 & -1 & 1 & 2 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 0 \end{pmatrix}$. There are two pivots, and thus the rank of A is $r = 2$.

Kernel: We need to find the solutions to the homogeneous system $A\,\mathbf{x} = \mathbf{0}$. In our example, the pivots are in columns 1 and 3, and so the free variables are x_2, x_4. Using back substitution on the reduced homogeneous system $U\,\mathbf{x} = \mathbf{0}$, we find the general solution

$$\mathbf{x} = \begin{pmatrix} \frac{1}{2}\,x_2 - 2\,x_4 \\ x_2 \\ 2\,x_4 \\ x_4 \end{pmatrix} = x_2 \begin{pmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -2 \\ 0 \\ 2 \\ 1 \end{pmatrix}. \tag{2.40}$$

Note that the second and fourth entries are the corresponding free variables x_2, x_4. Therefore,

$$\mathbf{z}_1 = \begin{pmatrix} \frac{1}{2} & 1 & 0 & 0 \end{pmatrix}^T, \qquad \mathbf{z}_2 = \begin{pmatrix} -2 & 0 & 2 & 1 \end{pmatrix}^T,$$

are the basis vectors for $\ker A$. By construction, they span the kernel, and linear independence follows easily since the only way in which the linear combination (2.40) could vanish, $\mathbf{x} = \mathbf{0}$, is if both free variables vanish: $x_2 = x_4 = 0$. In general, there are $n - r$ free variables, each corresponding to one of the basis elements of the kernel, which thus implies the dimension formula for $\ker A$.

Corange: The corange is the subspace of $\mathbb{R}^n$ spanned by the rows of A. We claim that applying an elementary row operation does not alter the corange. To see this for row operations of the first type, suppose, for instance, that $\widehat{A}$ is obtained adding a times the first row of A to the second row. If $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_m$ are the rows of A, then the rows of $\widehat{A}$ are $\mathbf{r}_1, \widehat{\mathbf{r}}_2 = \mathbf{r}_2 + a\,\mathbf{r}_1, \mathbf{r}_3, \dots, \mathbf{r}_m$. If

$$\mathbf{v} = c_1\,\mathbf{r}_1 + c_2\,\mathbf{r}_2 + c_3\,\mathbf{r}_3 + \,\cdots\, + c_m\,\mathbf{r}_m$$

is any vector belonging to $\operatorname{corng} A$, then

$$\mathbf{v} = \widehat{c}_1\,\mathbf{r}_1 + c_2\,\widehat{\mathbf{r}}_2 + c_3\,\mathbf{r}_3 + \,\cdots\, + c_m\,\mathbf{r}_m, \qquad \text{where} \qquad \widehat{c}_1 = c_1 - a\,c_2,$$

is also a linear combination of the rows of the new matrix, and hence lies in $\operatorname{corng} \widehat{A}$. The converse is also valid — $\mathbf{v} \in \operatorname{corng} \widehat{A}$ implies $\mathbf{v} \in \operatorname{corng} A$ — and we conclude that elementary row operations of Type #1 do not change $\operatorname{corng} A$. The proof for the other two types of elementary row operations is even easier, and left to the reader.

Since the row echelon form U is obtained from A by a sequence of elementary row operations, we conclude that $\operatorname{corng} A = \operatorname{corng} U$. Moreover, because each nonzero row in U contains a pivot, it is not hard to see that the nonzero rows of $\operatorname{corng} U$ are linearly independent, and hence form a basis of both $\operatorname{corng} U$ and $\operatorname{corng} A$. Since there is one row per pivot, $\operatorname{corng} U = \operatorname{corng} A$ has dimension r, the number of pivots. In our example, then, a basis for $\operatorname{corng} A$ consists of the row vectors

$$\mathbf{s}_1 = (\,2 \;\; -1 \;\; 1 \;\; 2\,), \qquad \mathbf{s}_2 = (\,0 \;\; 0 \;\; -2 \;\; 4\,).$$

The reader may wish to verify their linear independence, as well as the fact that every row of A lies in their span.

Range: There are two methods for computing a basis of the range or column space. The first proves that it has dimension equal to the rank. This has the important, and remarkable consequence that the space spanned by the rows of a matrix and the space spanned by its columns always have the same dimension, even though they are, in general, subspaces of different vector spaces.

Now the range of A and the range of U are, in general, different subspaces, so we *cannot* directly use a basis for $\operatorname{rng} U$ as a basis for $\operatorname{rng} A$. However, the linear dependencies among the columns of A and U *are* the same. It is not hard to see that the columns of U that contain the pivots form a basis for $\operatorname{rng} U$. This implies that the *same* columns of A form a basis for $\operatorname{rng} A$. In particular, this implies that $\dim \operatorname{rng} A = \dim \operatorname{rng} U = r$.

In our example, the pivots lie in the first and third columns of U, and hence the first and third columns of A, namely

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -8 \\ 4 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -6 \\ 3 \end{pmatrix},$$

form a basis for $\operatorname{rng} A$. This implies that every column of A can be written uniquely as a linear combination of the first and third column, as you can validate directly.

In more detail, using our matrix multiplication formula (2.14), we see that a linear combination of columns of A is trivial,

$$c_1\,\mathbf{v}_1 + \;\cdots\; + c_n\,\mathbf{v}_n = A\,\mathbf{c} = \mathbf{0},$$

if and only if $\mathbf{c} \in \ker A$. But we know $\ker A = \ker U$, and so the *same* linear combination of columns of U, namely

$$U\,\mathbf{c} = c_1\,\mathbf{u}_1 + \;\cdots\; + c_n\,\mathbf{u}_n = \mathbf{0},$$

is also trivial. In particular, the linear independence of the pivot columns of U, labeled $\mathbf{u}_{j_1}, \dots, \mathbf{u}_{j_r}$, implies the linear independence of the same collection, $\mathbf{v}_{j_1}, \dots, \mathbf{v}_{j_r}$, of columns of A. Moreover, the fact that any other column of U can be written as a linear combination

$$\mathbf{u}_k = d_1\mathbf{u}_{j_1} + \;\cdots\; + d_r\mathbf{u}_{j_r}$$

of the pivot columns implies that the same holds for the corresponding column of A, so

$$\mathbf{v}_k = d_1\mathbf{v}_{j_1} + \;\cdots\; + d_r\mathbf{v}_{j_r}.$$

We conclude that the pivot columns of A form a basis for its range or column space.

An alternative method to find a basis for the range is to note that $\operatorname{rng} A = \operatorname{corng} A^T$. Thus, we can employ our previous algorithm to compute $\operatorname{corng} A^T$. In our example, applying Gaussian elimination to

$$A^T = \begin{pmatrix} 2 & -8 & 4 \\ -1 & 4 & -2 \\ 1 & -6 & 3 \\ 2 & -4 & 2 \end{pmatrix} \quad \text{leads to the row echelon form} \quad \widehat{U} = \begin{pmatrix} 2 & -8 & 4 \\ 0 & -2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.41}$$

Observe that the row echelon form of A^T is *not* the transpose of the row echelon form of A! However, they do have the same number of pivots since both A and A^T have the same rank. Since the pivots of A^T are in the first two columns of $\widehat{U}$, we conclude that

$$\mathbf{y}_1 = \begin{pmatrix} 2 \\ -8 \\ 4 \end{pmatrix}, \qquad \mathbf{y}_2 = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix},$$

forms an alternative basis for $\operatorname{rng} A$.

Cokernel: Finally, to determine a basis for the cokernel of the matrix, we apply the preceding algorithm for finding a basis for $\ker A^T = \operatorname{coker} A$. Since the ranks of A and A^T coincide, there are now $m - r$ free variables, which is the same as the dimension of $\ker A^T$.

In our particular example, using the reduced form (2.41), the only free variable is y_3, and the general solution to the homogeneous adjoint system $A^T\mathbf{y} = \mathbf{0}$ is

$$\mathbf{y} = \begin{pmatrix} 0 \\ \frac{1}{2}\, y_3 \\ y_3 \end{pmatrix} = y_3 \begin{pmatrix} 0 \\ \frac{1}{2} \\ 1 \end{pmatrix}.$$

We conclude that $\operatorname{coker} A$ is one-dimensional, with basis $\begin{pmatrix} 0 & \frac{1}{2} & 1 \end{pmatrix}^T$.

2.6. Graphs and Incidence Matrices.

We now present an application of linear systems to graph theory. A *graph* consists of one or more points, called *vertices*, and lines or curves connecting them, called *edges*. Edge edge connects exactly two vertices, which, for simplicity, are assumed to always be distinct, so that no edge forms a *loop* that connects a vertex to itself. However, we do permit two vertices to be connected by multiple edges. Some examples of graphs appear in Figure 2.2; the vertices are the black dots. In a planar representation of the graph, the edges may cross over each other at non-nodal points, but do not actually meet — think of a circuit where the (insulated) wires lie on top of each other, but do not touch. Thus, the first graph in Figure 2.2 has 5 vertices and 8 edges; the second has 4 vertices and 6 edges — the two central edges do not meet; the final graph has 5 vertices and 10 edges.

Graphs arise in a multitude of applications. A particular case that will be considered in depth is electrical networks, where the edges represent wires, and the vertices represent the nodes where the wires are connected. Another example is the framework for a building —

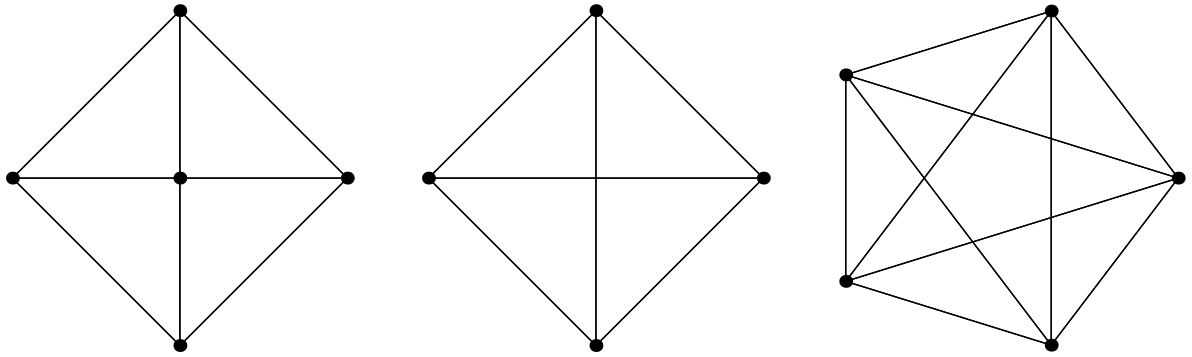


Figure 2.2. Three Different Graphs.

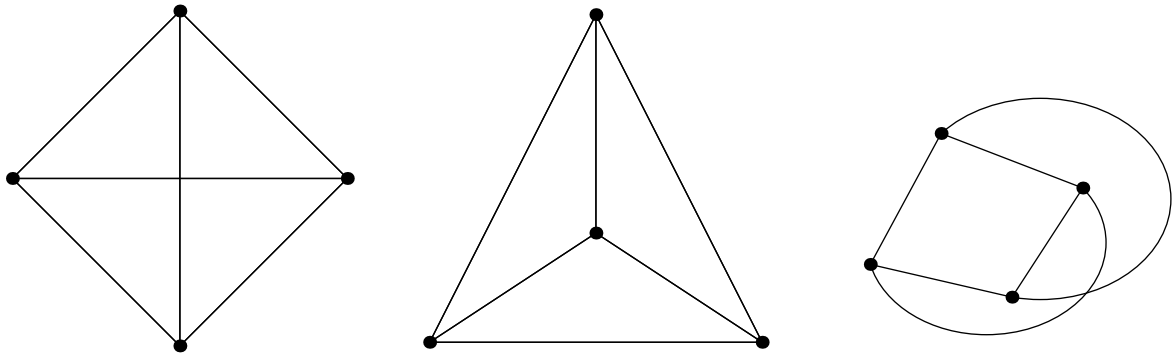


Figure 2.3. Three Versions of the Same Graph.

the edges represent the beams and the vertices the joints where the beams are connected. In each case, the graph encodes the topology — meaning interconnectedness — of the system, but not its geometry — lengths of edges, angles, etc.

Two graphs are considered to be the same if one can identify all their edges and vertices, so that they have the same connectivity properties. A good way to visualize this is to think of the graph as a collection of strings connected at the vertices. Moving the vertices and strings around without cutting or rejoining them will have no effect on the underlying graph. Consequently, there are many ways to draw a given graph; three equivalent graphs appear in Figure 2.3.

Two vertices in a graph are *adjacent* if there is an edge connecting them. Two edges are *adjacent* if they meet at a common vertex. For instance, in the graph in Figure 2.4, all vertices are adjacent; edge 1 is adjacent to all edges except edge 5. A *path* is a sequence of distinct, i.e., non-repeated, edges, with each edge adjacent to its successor. For example, in Figure 2.4, one path starts at vertex #1, then goes in order along the edges labeled as $1, 4, 3, 2$, thereby passing through vertices $1, 2, 4, 1, 3$. Note that while an edge cannot be repeated in a path, a vertex may be. A *circuit* is a path that ends up where it began. For example, the circuit consisting of edges $1, 4, 5, 2$ starts at vertex 1, then goes to vertices $2, 4, 3$ in order, and finally returns to vertex 1. The starting vertex for a circuit is not important. For example, edges $4, 5, 2, 1$ also represent the same circuit we just described. A graph is called *connected* if you can get from any vertex to any other vertex by a path,

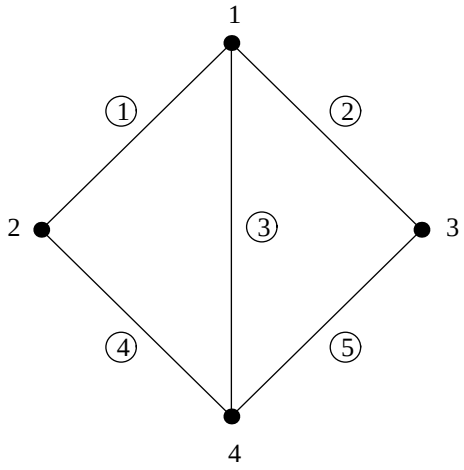


Figure 2.4. A Simple Graph.

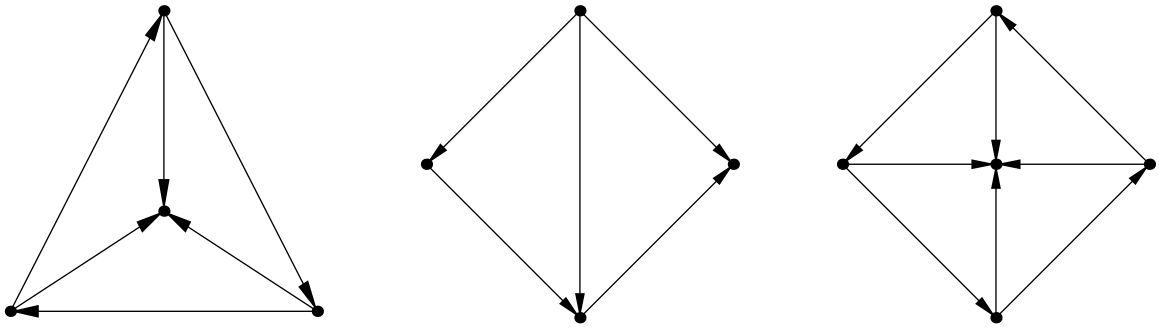


Figure 2.5. Digraphs.

which is by far the most important case for applications. We note that every graph can be decomposed into a disconnected collection of connected subgraphs.

In electrical circuits, one is interested in measuring currents and voltage drops along the wires in the network represented by the graph. Both of these quantities have a direction, and therefore we need to specify an orientation on each edge in order to quantify how the current moves along the wire. The orientation will be fixed by specifying the vertex the edge "starts" at, and the vertex it "ends" at. Once we assign a direction to an edge, a current along that wire will be positive if it moves in the same direction, i.e., goes from the starting vertex to the ending one, and negative if it moves in the opposite direction. The direction of the edge does *not* dictate the direction of the current — it just fixes what directions positive and negative values of current represent. A graph with directed edges is known as a *directed graph* or *digraph* for short. The edge directions are represented by arrows; examples of digraphs can be seen in Figure 2.5.

Consider a digraph D consisting of n vertices connected by m edges. The *incidence matrix* associated with D is an $m \times n$ matrix A whose rows are indexed by the edges and whose columns are indexed by the vertices. If edge k starts at vertex i and ends at vertex j, then row k of the incidence matrix will have a $+1$ in its (k, i) entry and -1 in its (k, j) entry; all other entries of the row are zero. Thus, our convention is that a $+1$ entry

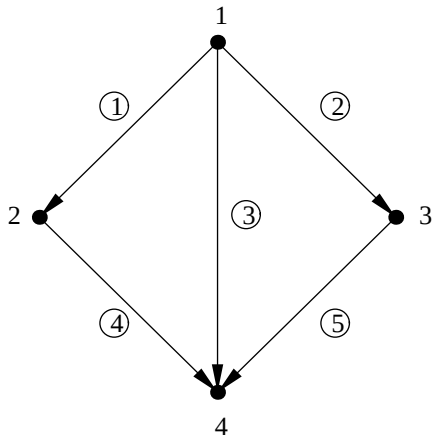


Figure 2.6. A Simple Digraph.

represents the vertex at which the edge starts and a -1 entry the vertex at which it ends.

A simple example is the digraph in Figure 2.6, which consists of five edges joined at four different vertices. Its 5×4 incidence matrix is

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{2.42}$$

Thus the first row of A tells us that the first edge starts at vertex 1 and ends at vertex 2. Similarly, row 2 says that the second edge goes from vertex 1 to vertex 3. Clearly one can completely reconstruct any digraph from its incidence matrix.

Example 2.48. The matrix

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}. \tag{2.43}$$

qualifies as an incidence matrix because each row contains a single $+1$, a single -1, and the other entries are 0. Let us construct the digraph corresponding to A. Since A has five columns, there are five vertices in the digraph, which we label by the numbers $1, 2, 3, 4, 5$. Since it has seven rows, there are 7 edges. The first row has its $+1$ in column 1 and its -1 in column 2 and so the first edge goes from vertex 1 to vertex 2. Similarly, the second edge corresponds to the second row of A and so goes from vertex 3 to vertex 1. The third row of A gives an edge from vertex 3 to vertex 2; and so on. In this manner we construct the digraph drawn in Figure 2.7.

The incidence matrix has important geometric and quantitative consequences for the graph it represents. In particular, its kernel and cokernel have topological significance. For

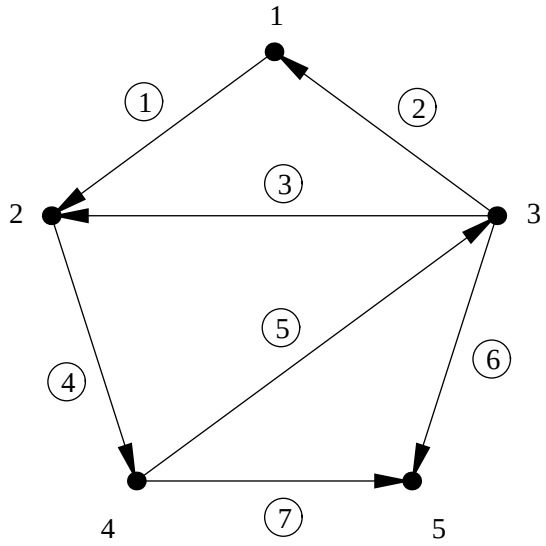


Figure 2.7. Another Digraph.

example, the kernel of the incidence matrix (2.43) is spanned by the single vector

$$\mathbf{z} = (\,1 \;\; 1 \;\; 1 \;\; 1 \;\; 1\,)^T,$$

and represents the fact that the sum of the entries in any given row of A is zero. This observation holds in general for connected digraphs.

Proposition 2.49. *If A is the incidence matrix for a connected digraph, then* $\ker A$ *is one-dimensional, with basis* $\mathbf{z} = (\,1 \;\; 1 \;\; \dots \;\; 1\,)^T$.

Proof: If edge k connects vertices i and j, then the k^{th} equation in $A\mathbf{z} = \mathbf{0}$ is $z_i = z_j$. The same equality holds, by a simple induction, if the vertices i and j are connected by a path. Therefore, if D is connected, all the entries of $\mathbf{z}$ are equal, and the result follows. *Q.E.D.*

Corollary 2.50. *If A is the incidence matrix for a connected digraph with n vertices, then* $\operatorname{rank} A = n - 1$.

Proof: This is an immediate consequence of Theorem 2.47. *Q.E.D.*

Next, let us look at the cokernel of an incidence matrix. Consider the particular example (2.42) corresponding to the digraph in Figure 2.6. We need to compute the kernel of the transposed incidence matrix

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & -1 \end{pmatrix}. \tag{2.44}$$

Solving the homogeneous system $A^T\mathbf{y} = \mathbf{0}$ by Gaussian elimination, we discover that $\operatorname{coker} A = \ker A^T$ is spanned by the two vectors

$$\mathbf{y}_1 = (\,1 \;\; 0 \;\; -1 \;\; 1 \;\; 0\,)^T, \qquad \mathbf{y}_2 = (\,0 \;\; 1 \;\; -1 \;\; 0 \;\; 1\,)^T.$$

Each of these vectors represents a *circuit* in the digraph, the nonzero entries representing the direction in which the edges are traversed. For example, $\mathbf{y}_1$ corresponds to the circuit that starts out along edge #1, then traverses edge #4 and finishes by going along edge #3 in the reverse direction, which is indicated by the minus sign in its third entry. Similarly, $\mathbf{y}_2$ represents the circuit consisting of edge #2, followed by edge #5, and then edge #3, backwards. The fact that $\mathbf{y}_1$ and $\mathbf{y}_2$ are linearly independent vectors says that the two circuits are "independent".

The general element of $\operatorname{coker} A$ is a linear combination $c_1\,\mathbf{y}_1 + c_2\,\mathbf{y}_2$. Certain values of the constants lead to other types of circuits; for example $-\mathbf{y}_1$ represents the same circuit as $\mathbf{y}_1$, but traversed in the opposite direction. Another example is

$$\mathbf{y}_1 - \mathbf{y}_2 = (\,1 \quad -1 \quad 0 \quad 1 \quad -1\,)^T,$$

which represents the square circuit going around the outside of the digraph, along edges $1, 4, 5, 2$, the fourth and second being in the reverse direction. We can view this circuit as a combination of the two triangular circuits; when we add them together the middle edge #3 is traversed once in each direction, which effectively "cancels" its contribution. (A similar cancellation occurs in the theory of line integrals; see Section A.5.) Other combinations represent "virtual" circuits; for instance, one can interpret $2\,\mathbf{y}_1 - \frac{1}{2}\,\mathbf{y}_2$ as two times around the first triangular circuit plus one half of the other triangular circuit, in the opposite direction— whatever that might mean.

Let us summarize the preceding discussion.

Theorem 2.51. *Each circuit in a digraph D is represented by a vector in the cokernel of its incidence matrix, whose entries are $+1$ if the edge is traversed in the correct direction, -1 if in the opposite direction, and 0 if the edge is not in the circuit. The dimension of the cokernel of A equals the number of independent circuits in D.*

The preceding two theorems have an important and remarkable consequence. Suppose D is a connected digraph with m edges and n vertices and A its $m \times n$ incidence matrix. Corollary 2.50 implies that A has rank $r = n - 1 = n - \dim \ker A$. On the other hand, Theorem 2.51 tells us that $\dim \operatorname{coker} A = l$ equals the number of independent circuits in D. The Fundamental Theorem 2.47 says that $r = m - l$. Equating these two different computations of the rank, we find $r = n - 1 = m - l$, or $n + l = m + 1$. This celebrated result is known as *Euler's formula* for graphs, first discovered by the extraordinarily prolific eighteenth century Swiss mathematician Leonhard Euler†.

Theorem 2.52. *If G is a connected graph, then*

$$\#\,\text{vertices} \;+\; \#\,\text{independent circuits} \;=\; \#\,\text{edges} + 1. \tag{2.45}$$

Remark: If the graph is *planar*, meaning that it can be graphed in the plane without any edges crossing over each other, then the number of independent circuits is equal to the number of "holes" in the graph, i.e., the number of distinct polygonal regions bounded

† Pronounced "Oiler"

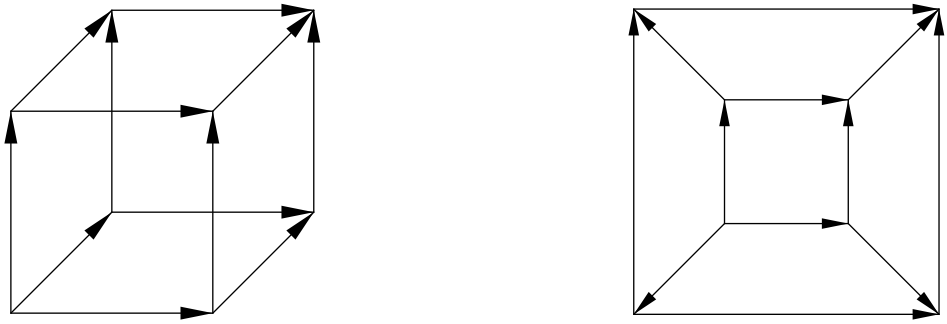


Figure 2.8. A Cubical Digraph.

by the edges of the graph. For example, the pentagonal digraph in Figure 2.7 bounds three triangles, and so has three independent circuits. For non-planar graphs, (2.45) gives a possible definition of the number of independent circuits, but one that is not entirely standard. A more detailed discussion relies on further developments in the topological properties of graphs, cf. [**33**].

Example 2.53. Consider the graph corresponding to the edges of a cube, as illustrated in Figure 2.8, where the second figure represents the same graph squashed down onto a plane. The graph has 8 vertices and 12 edges. Euler's formula (3.76) tells us that there are 5 independent circuits. These correspond to the interior square and four trapezoids in the planar version of the digraph, and hence to circuits around 5 of the 6 faces of the cube. The "missing" face does indeed define a circuit, but it can be represented as the sum of the other five circuits, and so is not independent. In Exercise ∎, the reader is asked to write out the incidence matrix for the cubical digraph and explicitly identify the basis of its kernel with the circuits.

We do not have the space to further develop the remarkable connections between graph theory and linear algebra. The interested reader is encouraged to consult a text devoted to graph theory, e.g., [**33**].

Chapter 3

Inner Products and Norms

The geometry of Euclidean space relies on the familiar properties of length and angle. The abstract concept of a norm on a vector space formalizes the geometrical notion of the length of a vector. In Euclidean geometry, the angle between two vectors is governed by their dot product, which is itself formalized by the abstract concept of an inner product. Inner products and norms lie at the heart of analysis, both linear and nonlinear, in both finite-dimensional vector spaces and infinite-dimensional function spaces. It is impossible to overemphasize their importance for both theoretical developments, practical applications in all fields, and in the design of numerical solution algorithms.

Mathematical analysis is founded on a few key inequalities. The most basic is the Cauchy–Schwarz inequality, which is valid in any inner product space. The more familiar triangle inequality for the associated norm is derived as a simple consequence. Not every norm arises from an inner product, and in the general situation, the triangle inequality becomes part of the definition. Both inequalities retain their validity in both finite-dimensional and infinite-dimensional vector spaces. Indeed, their abstract formulation helps focus on the key ideas in the proof, avoiding distracting complications resulting from the explicit formulas.

In Euclidean space $\mathbb{R}^n$, the characterization of general inner products will lead us to the extremely important class of positive definite matrices. Positive definite matrices play a key role in a variety of applications, including minimization problems, least squares, mechanical systems, electrical circuits, and the differential equations describing dynamical processes. Later, we will generalize the notion of positive definiteness to more general linear operators, governing the ordinary and partial differential equations arising in continuum mechanics and dynamics. Positive definite matrices most commonly appear in so-called Gram matrix form, consisting of the inner products between selected elements of an inner product space. In general, positive definite matrices can be completely characterized by their pivots resulting from Gaussian elimination. The associated matrix factorization can be reinterpreted as the method of completing the square for the associated quadratic form.

So far, we have confined our attention to real vector spaces. Complex numbers, vectors and functions also play an important role in applications, and so, in the final section, we formally introduce complex vector spaces. Most of the formulation proceeds in direct analogy with the real version, but the notions of inner product and norm on complex vector spaces requires some thought. Applications of complex vector spaces and their inner products are of particular importance in Fourier analysis and signal processing, and absolutely essential in modern quantum mechanics.

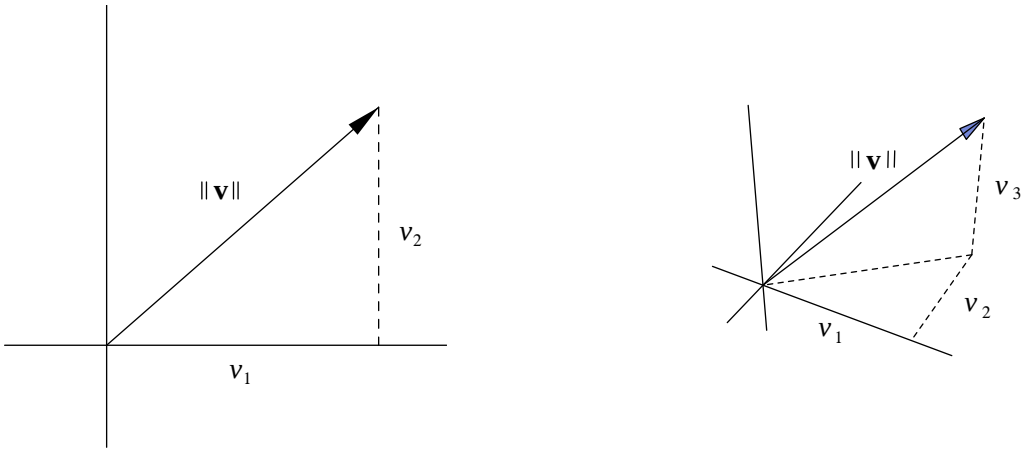


Figure 3.1. The Euclidean Norm in $\mathbb{R}^2$ and $\mathbb{R}^3$.

3.1. Inner Products.

The most basic example of an inner product is the familiar *dot product*

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \mathbf{v}\cdot\mathbf{w} = v_1\,w_1 + v_2\,w_2 + \,\cdots\, + v_n\,w_n = \sum_{i=1}^{n} v_i\,w_i, \tag{3.1}$$

between (column) vectors $\mathbf{v} = (\,v_1, v_2, \ldots, v_n\,)^T, \mathbf{w} = (\,w_1, w_2, \ldots, w_n\,)^T$ lying in the Euclidean space $\mathbb{R}^n$. An important observation is that the dot product (3.1) can be identified with the matrix product

$$\mathbf{v}\cdot\mathbf{w} = \mathbf{v}^T\mathbf{w} = (\,v_1 \;\; v_2 \;\; \ldots \;\; v_n\,)\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} \tag{3.2}$$

between a row vector $\mathbf{v}^T$ and a column vector $\mathbf{w}$.

The dot product is the cornerstone of Euclidean geometry. The key remark is that the dot product of a vector with itself,

$$\langle\, \mathbf{v}\,;\mathbf{v}\,\rangle = v_1^2 + v_2^2 + \,\cdots\, + v_n^2,$$

is the sum of the squares of its entries, and hence equal to the square of its length. Therefore, the *Euclidean norm* or *length* of a vector is found by taking the square root:

$$\|\,\mathbf{v}\,\| \;=\; \sqrt{\mathbf{v}\cdot\mathbf{v}} = \;\sqrt{v_1^2 + v_2^2 + \,\cdots\, + v_n^2}\;. \tag{3.3}$$

This formula generalizes the classical Pythagorean Theorem to n-dimensional Euclidean space; see Figure 3.1. Since each term in the sum is non-negative, the length of a vector is also non-negative, $\|\,\mathbf{v}\,\| \geq 0$. Furthermore, the only vector of length 0 is the zero vector.

The dot product and norm satisfy certain evident properties, and these serve as the basis for the abstract definition of more general inner products on real vector spaces.

Definition 3.1. An *inner product* on the real vector space V is a pairing that takes two vectors $\mathbf{v}, \mathbf{w} \in V$ and produces a real number $\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle \in \mathbb{R}$. The inner product is required to satisfy the following three axioms for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, and $c, d \in \mathbb{R}$.

(i) *Bilinearity*:

$$\begin{aligned} \langle\, c\,\mathbf{u} + d\,\mathbf{v}\,;\mathbf{w}\,\rangle &= c\,\langle\, \mathbf{u}\,;\mathbf{w}\,\rangle + d\,\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle, \\ \langle\, \mathbf{u}\,;c\,\mathbf{v} + d\,\mathbf{w}\,\rangle &= c\,\langle\, \mathbf{u}\,;\mathbf{v}\,\rangle + d\,\langle\, \mathbf{u}\,;\mathbf{w}\,\rangle. \end{aligned} \tag{3.4}$$

(ii) *Symmetry*:

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \langle\, \mathbf{w}\,;\mathbf{v}\,\rangle. \tag{3.5}$$

(iii) *Positivity*:

$$\langle\, \mathbf{v}\,;\mathbf{v}\,\rangle > 0 \qquad \text{whenever} \qquad \mathbf{v} \neq \mathbf{0}, \qquad \text{while} \qquad \langle\, \mathbf{0}\,;\mathbf{0}\,\rangle = 0. \tag{3.6}$$

A vector space equipped with an inner product is called an *inner product space*. As we shall see, a given vector space can admit many different inner products. Verification of the inner product axioms for the Euclidean dot product is straightforward, and left to the reader.

Given an inner product, the associated *norm* of a vector $\mathbf{v} \in V$ is defined as the positive square root of the inner product of the vector with itself:

$$\| \mathbf{v} \| = \sqrt{\langle\, \mathbf{v}\,;\mathbf{v}\,\rangle}\,. \tag{3.7}$$

The positivity axiom implies that $\| \mathbf{v} \| \geq 0$ is real and non-negative, and equals 0 if and only if $\mathbf{v} = \mathbf{0}$ is the zero vector.

Example 3.2. While certainly the most important inner product on $\mathbb{R}^n$, the dot product is by no means the only possibility. A simple example is provided by the weighted inner product

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = 2\,v_1\,w_1 + 5\,v_2\,w_2, \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}. \tag{3.8}$$

between vectors in $\mathbb{R}^2$. The symmetry axiom (3.5) is immediate. Moreover,

$$\begin{aligned} \langle\, c\,\mathbf{u} + d\,\mathbf{v}\,;\mathbf{w}\,\rangle &= 2\,(c\,u_1 + d\,v_1)\,w_1 + 5\,(c\,u_2 + d\,v_2)\,w_2 \\ &= (2\,c\,u_1\,w_1 + 5\,c\,u_2\,w_2) + (2\,d\,v_1\,w_1 + 5\,d\,v_2\,w_2) = c\,\langle\, \mathbf{u}\,;\mathbf{w}\,\rangle + d\,\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle, \end{aligned}$$

which verifies the first bilinearity condition; the second follows by a very similar computation. (Or, one can rely on symmetry; see Exercise ■.) Moreover,

$$\langle\, \mathbf{v}\,;\mathbf{v}\,\rangle = 2\,v_1^2 + 5\,v_2^2 \geq 0$$

is clearly strictly positive for any $\mathbf{v} \neq \mathbf{0}$ and equal to zero when $\mathbf{v} = \mathbf{0}$, which proves positivity and hence establishes (3.8) as an legitimate inner product on $\mathbb{R}^2$. The associated weighted norm is

$$\| \mathbf{v} \| \;=\; \sqrt{2\,v_1^2 + 5\,v_2^2}\,.$$

A less evident example is provided by the expression

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = v_1\,w_1 - v_1\,w_2 - v_2\,w_1 + 4\,v_2\,w_2. \tag{3.9}$$

Bilinearity is verified in the same manner as before, and symmetry is obvious. Positivity is ensured by noticing that

$$\langle \mathbf{v}\,;\mathbf{v}\,\rangle = v_1^2 - 2\,v_1\,v_2 + 4\,v_2^2 = (v_1 - v_2)^2 + 3\,v_2^2 \;>\; 0$$

is strictly positive for all nonzero $\mathbf{v} \neq \mathbf{0}$. Therefore, (3.9) defines an alternative inner product on $\mathbb{R}^2$. The associated norm

$$\|\,\mathbf{v}\,\| \;=\; \sqrt{v_1^2 - 2\,v_1\,v_2 + 4\,v_2^2}$$

defines a different notion of distance and consequential "non-Pythagorean plane geometry".

Example 3.3. Let $c_1, \ldots, c_n$ be a set of *positive* numbers. The corresponding *weighted inner product* and *weighted norm* on $\mathbb{R}^n$ are defined by

$$\langle \mathbf{v}\,;\mathbf{w}\,\rangle \;=\; \sum_{i=1}^{n} c_i\,v_i\,w_i, \qquad\qquad \|\,\mathbf{v}\,\| \;=\; \sqrt{\sum_{i=1}^{n} c_i\,v_i^2}\;. \tag{3.10}$$

The numbers $c_i > 0$ are the *weights*. The larger the weight c_i, the more the i^{th} coordinate of $\mathbf{v}$ contributes to the norm. Weighted norms are particularly important in statistics and data fitting, where one wants to emphasize certain quantities and de-emphasize others; this is done by assigning suitable weights to the different components of the data vector $\mathbf{v}$. Section 4.3 on least squares approximation methods will contain further details.

Inner Products on Function Space

Inner products and norms on function spaces will play an absolutely essential role in modern analysis, particularly Fourier analysis and the solution to boundary value problems for both ordinary and partial differential equations. Let us introduce the most important examples.

Example 3.4. Given a bounded closed interval $[a, b] \subset \mathbb{R}$, consider the vector space $\mathrm{C}^0 = \mathrm{C}^0[a, b]$ consisting of all continuous functions $f\colon [a, b] \to \mathbb{R}$. The integral

$$\langle\, f\,;g\,\rangle = \int_a^b f(x)\,g(x)\,dx \tag{3.11}$$

defines an inner product on the vector space C^0, as we shall prove below. The associated norm is, according to the basic definition (3.7),

$$\|\,f\,\| = \sqrt{\int_a^b f(x)^2\,dx}\;. \tag{3.12}$$

This quantity is known as the L^2 *norm* of the function f over the interval $[a, b]$. The L^2 norm plays the same role in infinite-dimensional function space that the Euclidean norm or length of a vector plays in the finite-dimensional Euclidean vector space $\mathbb{R}^n$.

For example, if we take $[a, b] = [0, \frac{1}{2}\pi]$, then the L^2 inner product between $f(x) = \sin x$ and $g(x) = \cos x$ is equal to

$$\langle \sin x \,;\cos x \,\rangle = \int_0^{\pi/2} \sin x \, \cos x \, dx = \frac{1}{2} \sin^2 x \, \Big|_{x=0}^{\pi/2} = \frac{1}{2} \, .$$

Similarly, the norm of the function $\sin x$ is

$$\| \sin x \| = \sqrt{\int_0^{\pi/2} (\sin x)^2 \, dx} = \sqrt{\frac{\pi}{4}} \, .$$

One must always be careful when evaluating function norms. For example, the constant function $c(x) \equiv 1$ has norm

$$\| \, 1 \, \| = \sqrt{\int_0^{\pi/2} 1^2 \, dx} = \sqrt{\frac{\pi}{2}} \, ,$$

not 1 as you might have expected. We also note that the value of the norm depends upon which interval the integral is taken over. For instance, on the longer interval $[0, \pi]$,

$$\| \, 1 \, \| = \sqrt{\int_0^{\pi} 1^2 \, dx} = \sqrt{\pi} \, .$$

Thus, when dealing with the L^2 inner product or norm, one must always be careful to specify the function space, or, equivalently, the interval on which it is being evaluated.

Let us prove that formula (3.11) does, indeed, define an inner product. First, we need to check that $\langle \, f \,; g \, \rangle$ is well-defined. This follows because the product $f(x)\, g(x)$ of two continuous functions is also continuous, and hence its integral over a bounded interval is defined and finite. The symmetry condition for the inner product is immediate:

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

because multiplication of functions is commutative. The first bilinearity axiom

$$\langle \, c\, f + d\, g \,; h \, \rangle = c \, \langle \, f \,; h \, \rangle + d \, \langle \, g \,; h \, \rangle,$$

amounts to the following elementary integral identity

$$\int_a^b \big[\, c\, f(x) + d\, g(x) \, \big] \, h(x) \, dx = c \int_a^b f(x)\, h(x) \, dx + d \int_a^b g(x)\, h(x) \, dx,$$

valid for arbitrary continuous functions f, g, h and scalars (constants) c, d. The second bilinearity axiom is proved similarly; alternatively, one can use symmetry to deduce it from the first as in Exercise ∎. Finally, positivity requires that

$$\| \, f \, \|^2 = \langle \, f \,; f \, \rangle = \int_a^b f(x)^2 \, dx \geq 0.$$

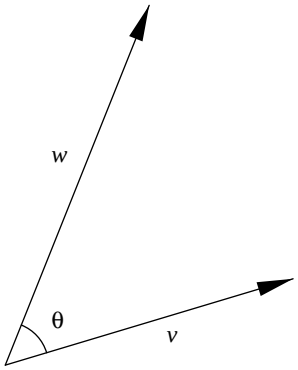


Figure 3.2. Angle Between Two Vectors.

This is clear because $f(x)^2 \geq 0$, and the integral of a nonnegative function is nonnegative. Moreover, since the function $f(x)^2$ is continuous and nonnegative, its integral will vanish, $\int_a^b f(x)^2\,dx = 0$ if and only if $f(x) \equiv 0$ is the zero function, cf. Exercise ■. This completes the demonstration.

Remark: The preceding construction applies to more general functions, but we have restricted our attention to continuous functions to avoid certain technical complications. The most general function space admitting this important inner product is known as *Hilbert space*, which forms the foundation for modern analysis, [**126**], including the rigorous theory of Fourier series, [**51**], and also lies at the heart of modern quantum mechanics, [**100**, **104**, **122**]. One does need to be extremely careful when trying to extend the inner product to more general functions. Indeed, there are nonzero, discontinuous functions with zero L^2 "norm". An example is

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & \text{otherwise}, \end{cases} \qquad \text{which satisfies} \qquad \| f \|^2 = \int_{-1}^{1} f(x)^2\,dx = 0 \tag{3.13}$$

because any function which is zero except at finitely many (or even countably many) points has zero integral. We will discuss some of the details of the Hilbert space construction in Chapters 12 and 13.

One can also define weighted inner products on the function space $C^0[a, b]$. The weights along the interval are specified by a (continuous) positive scalar function $w(x) > 0$. The corresponding *weighted inner product* and *norm* are

$$\langle\, f\,;g\,\rangle = \int_a^b f(x)\,g(x)\,w(x)\,dx, \qquad \| f \| = \sqrt{\int_a^b f(x)^2\,w(x)\,dx}\,. \tag{3.14}$$

The verification of the inner product axioms in this case is left as an exercise for the reader.

3.2. Inequalities.

Returning to the general framework of inner products on vector spaces, we now prove the most important inequality in applied mathematics. Its origins can be found in the geometric interpretation of the dot product on Euclidean space in terms of the angle between vectors.

The Cauchy–Schwarz Inequality

In two and three-dimensional Euclidean geometry, the dot product between two vectors can be geometrically characterized by the equation

$$\mathbf{v} \cdot \mathbf{w} = \| \mathbf{v} \| \, \| \mathbf{w} \| \, \cos \theta, \tag{3.15}$$

where θ measures the angle between the vectors $\mathbf{v}$ and $\mathbf{w}$, as depicted in Figure 3.2. Since

$$| \cos \theta | \leq 1,$$

the absolute value of the dot product is bounded by the product of the lengths of the vectors:

$$| \mathbf{v} \cdot \mathbf{w} | \; \leq \; \| \mathbf{v} \| \, \| \mathbf{w} \|.$$

This fundamental inequality is named after two[†] of the founders of modern analysis, Augustin Cauchy and Herman Schwarz. It holds, in fact, for *any* inner product.

Theorem 3.5. *Every inner product satisfies the* Cauchy–Schwarz inequality

$$| \langle \mathbf{v} \, ; \mathbf{w} \rangle | \; \leq \; \| \mathbf{v} \| \, \| \mathbf{w} \|, \qquad \mathbf{v}, \mathbf{w} \in V. \tag{3.16}$$

Here, $\| \mathbf{v} \|$ is the associated norm, while $| \cdot |$ denotes absolute value. Equality holds if and only if $\mathbf{v}$ and $\mathbf{w}$ are parallel[‡] vectors.

Proof: The case when $\mathbf{w} = \mathbf{0}$ is trivial, since both sides of (3.16) are equal to 0. Thus, we may suppose $\mathbf{w} \neq \mathbf{0}$. Let $t \in \mathbb{R}$ be an arbitrary scalar. Using the three basic inner product axioms, we have

$$0 \leq \| \mathbf{v} + t \, \mathbf{w} \|^2 = \langle \mathbf{v} + t \, \mathbf{w} \, ; \mathbf{v} + t \, \mathbf{w} \rangle = \| \mathbf{v} \|^2 + 2 t \, \langle \mathbf{v} \, ; \mathbf{w} \rangle + t^2 \, \| \mathbf{w} \|^2, \tag{3.17}$$

with equality holding if and only if $\mathbf{v} = - t \, \mathbf{w}$ — which requires $\mathbf{v}$ and $\mathbf{w}$ to be parallel vectors. We fix $\mathbf{v}$ and $\mathbf{w}$, and consider the right hand side of (3.17) as a quadratic function,

$$p(t) = \| \mathbf{w} \|^2 \, t^2 + 2 \, \langle \mathbf{v} \, ; \mathbf{w} \rangle \, t + \| \mathbf{v} \|^2,$$

of the scalar variable t. To get the maximum mileage out of the fact that $p(t) \geq 0$, let us look at where it assumes a minimum. This occurs when its derivative vanishes:

$$p'(t) = 2 \, \| \mathbf{w} \|^2 \, t + 2 \, \langle \mathbf{v} \, ; \mathbf{w} \rangle = 0, \qquad \text{and thus at} \qquad t = \, - \, \frac{\langle \mathbf{v} \, ; \mathbf{w} \rangle}{\| \mathbf{w} \|^2} \, .$$

[†] Russians also give credit for its discovery to their compatriot Viktor Bunyakovskii, and, indeed, many authors append his name to the inequality.

[‡] Two vectors are parallel if and only if one is a scalar multiple of the other. The zero vector is parallel to every other vector, by convention.

Substituting this particular minimizing value into (3.17), we find

$$0 \;\le\; \| \mathbf{v} \|^2 - 2\, \frac{\langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle^2}{\| \mathbf{w} \|^2} + \frac{\langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle^2}{\| \mathbf{w} \|^2} \;=\; \| \mathbf{v} \|^2 - \frac{\langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle^2}{\| \mathbf{w} \|^2}\,.$$

Rearranging this last inequality, we conclude that

$$\frac{\langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle^2}{\| \mathbf{w} \|^2} \;\le\; \| \mathbf{v} \|^2, \qquad \text{or} \qquad \langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle^2 \;\le\; \| \mathbf{v} \|^2\, \| \mathbf{w} \|^2.$$

Taking the (positive) square root of both sides of the final inequality completes the theorem's proof. *Q.E.D.*

Given any inner product on a vector space, we can use the quotient

$$\cos\theta = \frac{\langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle}{\| \mathbf{v} \|\, \| \mathbf{w} \|} \tag{3.18}$$

to *define* the "angle" between the elements $\mathbf{v}, \mathbf{w} \in V$. The Cauchy–Schwarz inequality tells us that the ratio lies between -1 and $+1$, and hence the angle θ is well-defined, and, in fact, unique if we restrict it to lie in the range $0 \le \theta \le \pi$.

For example, using the standard dot product on $\mathbb{R}^3$, the angle between the vectors $\mathbf{v} = (\,1 \;\; 0 \;\; 1\,)^T$ and $\mathbf{w} = (\,0 \;\; 1 \;\; 1\,)^T$ is given by

$$\cos\theta = \frac{1}{\sqrt{2}\cdot\sqrt{2}} = \frac{1}{2}\,,$$

and so $\theta = \frac{1}{3}\,\pi$, i.e., 60°. Similarly, the "angle" between the polynomials $p(x) = x$ and $q(x) = x^2$ defined on the interval $I = [\,0, 1\,]$ is given by

$$\cos\theta \;=\; \frac{\langle\, x \,;\, x^2 \,\rangle}{\| \,x\, \|\, \| \,x^2\, \|} \;=\; \frac{\displaystyle\int_0^1 x^3\, dx}{\displaystyle\sqrt{\int_0^1 x^2\, dx}\,\sqrt{\int_0^1 x^4\, dx}} \;=\; \frac{\frac{1}{4}}{\sqrt{\frac{1}{3}}\,\sqrt{\frac{1}{5}}} \;=\; \sqrt{\frac{15}{16}}\,,$$

so that $\theta = 0.25268\ldots$ radians.

Warning: One should not try to give this notion of angle between functions more significance than the formal definition warrants — it does not correspond to any "angular" properties of their graph. Also, the value depends on the choice of inner product and the interval upon which it is being computed. For example, if we change to the inner product on the interval $[\,-1, 1\,]$, then $\langle\, x \,;\, x^2 \,\rangle = \displaystyle\int_{-1}^1 x^3\, dx = 0$, and hence (3.18) becomes $\cos\theta = 0$, so the "angle" between x and x^2 is now $\theta = \frac{1}{2}\,\pi$.

Even in Euclidean space $\mathbb{R}^n$, the measurement of angle (and length) depends upon the choice of an underlying inner product. Different inner products lead to different angle measurements; only for the standard Euclidean dot product does angle correspond to our everyday experience.

Orthogonal Vectors

A particularly important geometrical configuration occurs when two vectors are *perpendicular*, which means that they meet at a right angle: $\theta = \frac{1}{2}\pi$ or $\frac{3}{2}\pi$, and so $\cos\theta = 0$. The angle formula (3.18) implies that the vectors $\mathbf{v}, \mathbf{w}$ are perpendicular if and only if their dot product vanishes: $\mathbf{v} \cdot \mathbf{w} = 0$. Perpendicularity also plays a key role in general inner product spaces, but, for historical reasons, has been given a different name.

Definition 3.6. Two elements $\mathbf{v}, \mathbf{w} \in V$ of an inner product space V are called *orthogonal* if their inner product $\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = 0$.

Orthogonality is a remarkably powerful tool in all applications of linear algebra, and often serves to dramatically simplify many computations. We will devote Chapter 5 to its detailed development.

Example 3.7. The vectors $\mathbf{v} = (\,1,2\,)^T$ and $\mathbf{w} = (\,6,-3\,)^T$ are orthogonal with respect to the Euclidean dot product in $\mathbb{R}^2$, since $\mathbf{v} \cdot \mathbf{w} = 1 \cdot 6 + 2 \cdot (-3) = 0$. We deduce that they meet at a 90° angle. However, these vectors are *not* orthogonal with respect to the weighted inner product (3.8):

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \left\langle \begin{pmatrix} 1 \\ 2 \end{pmatrix}; \begin{pmatrix} 6 \\ -3 \end{pmatrix} \right\rangle = 2 \cdot 1 \cdot 6 + 5 \cdot 2 \cdot (-3) = -18 \neq 0.$$

Thus, orthogonality, like angles in general, depends upon which inner product is being used.

Example 3.8. The polynomials $p(x) = x$ and $q(x) = x^2 - \frac{1}{2}$ are orthogonal with respect to the inner product $\langle\, p\,;q\,\rangle = \displaystyle\int_0^1 p(x)\,q(x)\,dx$ on the interval $[\,0,1\,]$, since

$$\langle\, x\,;\, x^2 - \tfrac{1}{2}\,\rangle = \int_0^1 x\big(\, x^2 - \tfrac{1}{2}\,\big)\,dx = \int_0^1 \big(\, x^3 - \tfrac{1}{2}\,x\,\big)\,dx = 0.$$

They fail to be orthogonal on most other intervals. For example, on the interval $[\,0,2\,]$,

$$\langle\, x\,;\, x^2 - \tfrac{1}{2}\,\rangle = \int_0^2 x\big(\, x^2 - \tfrac{1}{2}\,\big)\,dx = \int_0^2 \big(\, x^3 - \tfrac{1}{2}\,x\,\big)\,dx = 3.$$

The Triangle Inequality

The familiar triangle inequality states that the length of one side of a triangle is at most equal to the sum of the lengths of the other two sides. Referring to Figure 3.3, if the first two side are represented by vectors $\mathbf{v}$ and $\mathbf{w}$, then the third corresponds to their sum $\mathbf{v} + \mathbf{w}$, and so $\|\,\mathbf{v} + \mathbf{w}\,\| \leq \|\,\mathbf{v}\,\| + \|\,\mathbf{w}\,\|$. The triangle inequality is a direct consequence of the Cauchy–Schwarz inequality, and hence holds for *any* inner product space.

Theorem 3.9. *The norm associated with an inner product satisfies the* triangle inequality

$$\|\,\mathbf{v} + \mathbf{w}\,\| \;\leq\; \|\,\mathbf{v}\,\| + \|\,\mathbf{w}\,\| \tag{3.19}$$

for every $\mathbf{v}, \mathbf{w} \in V$. Equality holds if and only if $\mathbf{v}$ and $\mathbf{w}$ are parallel vectors.

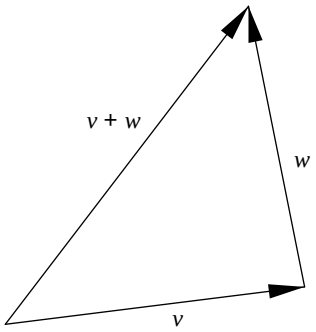


Figure 3.3. Triangle Inequality.

Proof: We compute

$$\begin{aligned} \| \mathbf{v} + \mathbf{w} \|^2 = \langle \mathbf{v} + \mathbf{w} \,; \mathbf{v} + \mathbf{w} \rangle &= \| \mathbf{v} \|^2 + 2 \langle \mathbf{v} \,; \mathbf{w} \rangle + \| \mathbf{w} \|^2 \\ &\leq \| \mathbf{v} \|^2 + 2 \| \mathbf{v} \| \, \| \mathbf{w} \| + \| \mathbf{w} \|^2 = \big(\| \mathbf{v} \| + \| \mathbf{w} \| \big)^2, \end{aligned}$$

where the inequality follows from Cauchy–Schwarz. Taking square roots of both sides and using positivity completes the proof. *Q.E.D.*

Example 3.10. The vectors $\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$ and $\mathbf{w} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}$ sum to $\mathbf{v} + \mathbf{w} = \begin{pmatrix} 3 \\ 2 \\ 2 \end{pmatrix}$. Their Euclidean norms are $\| \mathbf{v} \| = \sqrt{6}$ and $\| \mathbf{w} \| = \sqrt{13}$, while $\| \mathbf{v} + \mathbf{w} \| = \sqrt{17}$. The triangle inequality (3.19) in this case says $\sqrt{17} \leq \sqrt{6} + \sqrt{13}$, which is valid.

Example 3.11. Consider the functions $f(x) = x - 1$ and $g(x) = x^2 + 1$. Using the L^2 norm on the interval $[0, 1]$, we find

$$\| f \| = \sqrt{\int_0^1 (x-1)^2 \, dx} = \sqrt{\frac{1}{3}}, \qquad \| g \| = \sqrt{\int_0^1 (x^2+1)^2 \, dx} = \sqrt{\frac{23}{15}},$$

$$\| f + g \| = \sqrt{\int_0^1 (x^2 + x)^2 \, dx} = \sqrt{\frac{77}{60}}.$$

The triangle inequality requires $\sqrt{\frac{77}{60}} \leq \sqrt{\frac{1}{3}} + \sqrt{\frac{23}{15}}$, which is correct.

The Cauchy–Schwarz and triangle inequalities look much more impressive when writ-

ten out in full detail. For the Euclidean inner product (3.1), they are

$$\begin{aligned} \left| \sum_{i=1}^{n} v_i\, w_i \right| &\le \sqrt{\sum_{i=1}^{n} v_i^2}\ \sqrt{\sum_{i=1}^{n} w_i^2}\ , \\ \sqrt{\sum_{i=1}^{n} (v_i + w_i)^2} &\le \sqrt{\sum_{i=1}^{n} v_i^2}\ +\ \sqrt{\sum_{i=1}^{n} w_i^2}\ . \end{aligned} \tag{3.20}$$

Theorems 3.5 and 3.9 imply that these inequalities are valid for arbitrary real numbers $v_1, \dots, v_n, w_1, \dots, w_n$. For the L^2 inner product (3.12) on function space, they produce the following splendid integral inequalities:

$$\begin{aligned} \left| \int_a^b f(x)\, g(x)\, dx \right| &\le \sqrt{\int_a^b f(x)^2\, dx}\ \sqrt{\int_a^b g(x)^2\, dx}\ , \\ \sqrt{\int_a^b \big[f(x) + g(x)\big]^2\, dx} &\le \sqrt{\int_a^b f(x)^2\, dx}\ +\ \sqrt{\int_a^b g(x)^2\, dx}\ , \end{aligned} \tag{3.21}$$

which hold for arbitrary continuous (and even more general) functions. The first of these is the original Cauchy–Schwarz inequality, whose proof appeared to be quite deep when it first appeared. Only after the abstract notion of an inner product space was properly formalized did its innate simplicity and generality become evident. One can also generalize either of these sets of inequalities to weighted inner products, replacing the integration element dx by a weighted version $w(x)\, dx$, provided $w(x) > 0$.

3.3. Norms.

Every inner product gives rise to a norm that can be used to measure the magnitude or length of the elements of the underlying vector space. However, not every such norm used in analysis and applications arises from an inner product. To define a general norm on a vector space, we will extract those properties that do not directly rely on the inner product structure.

Definition 3.12. A *norm* on the vector space V assigns a real number $\| \mathbf{v} \|$ to each vector $\mathbf{v} \in V$, subject to the following axioms for all $\mathbf{v}, \mathbf{w} \in V$, and $c \in \mathbb{R}$:

- (*i*) *Positivity*: $\| \mathbf{v} \| \ge 0$, with $\| \mathbf{v} \| = 0$ if and only if $\mathbf{v} = \mathbf{0}$.
- (*ii*) *Homogeneity*: $\| c\, \mathbf{v} \| = | c |\, \| \mathbf{v} \|$.
- (*iii*) *Triangle inequality*: $\| \mathbf{v} + \mathbf{w} \| \le \| \mathbf{v} \| + \| \mathbf{w} \|$.

As we now know, every inner product gives rise to a norm. Indeed, positivity of the norm is one of the inner product axioms. The homogeneity property follows since

$$\| c\, \mathbf{v} \| \ = \ \sqrt{\langle\, c\, \mathbf{v}\, ;\, c\, \mathbf{v}\, \rangle} \ = \ \sqrt{c^2\, \langle\, \mathbf{v}\, ;\, \mathbf{v}\, \rangle} \ = \ | c |\, \sqrt{\langle\, \mathbf{v}\, ;\, \mathbf{v}\, \rangle} \ = \ | c |\, \| \mathbf{v} \|.$$

Finally, the triangle inequality for an inner product norm was established in Theorem 3.9.

Here are some important examples of norms that do not come from inner products.

Example 3.13. Let $V = \mathbb{R}^n$. The 1–*norm* of a vector $\mathbf{v} = (\,\mathbf{v}_1\ \mathbf{v}_2\ \dots\ \mathbf{v}_n\,)^T$ is defined as the sum of the absolute values of its entries:

$$\| \mathbf{v} \|_1 = | v_1 | + \ \cdots\ + | v_n |. \tag{3.22}$$

The *max* or ∞–*norm* is equal to the maximal entry (in absolute value):

$$\| \mathbf{v} \|_\infty = \sup \{ | v_1 |, \ \dots\ , | v_n | \}. \tag{3.23}$$

Verification of the positivity and homogeneity properties for these two norms is straightforward; the triangle inequality is a direct consequence of the elementary inequality

$$| a + b | \le | a | + | b |$$

for absolute values.

The Euclidean norm, 1–norm, and ∞–norm on $\mathbb{R}^n$ are just three representatives of the general *p–norm*

$$\| \mathbf{v} \|_p \ = \ \sqrt[p]{\sum_{i=1}^{n} | v_i |^p}\ . \tag{3.24}$$

This quantity defines a norm for any $1 \le p < \infty$. The ∞–norm is a limiting case of the p–norm as $p \to \infty$. Note that the Euclidean norm (3.3) is the 2–norm, and is often designated as such; it is the only p–norm which comes from an inner product. The positivity and homogeneity properties of the p–norm are straightforward. The triangle inequality, however, is not trivial; in detail, it reads

$$\sqrt[p]{\sum_{i=1}^{n} | v_i + w_i |^p} \ \le \ \sqrt[p]{\sum_{i=1}^{n} | v_i |^p} \ + \ \sqrt[p]{\sum_{i=1}^{n} | w_i |^p}\ , \tag{3.25}$$

and is known as *Minkowski's inequality*. A proof can be found in [**97**].

Example 3.14. There are analogous norms on the space $\mathrm{C}^0[a, b]$ of continuous functions on an interval $[a, b]$. Basically, one replaces the previous sums by integrals. Thus, the L^p–norm is defined as

$$\| f \|_p = \ \sqrt[p]{\int_a^b | f(x) |^p \, dx}\ . \tag{3.26}$$

In particular, the L^1 norm is given by integrating the absolute value of the function:

$$\| f \|_1 = \int_a^b | f(x) | \, dx. \tag{3.27}$$

The L^2 norm (3.12) appears as a special case, $p = 2$, and, again, is the only one arising from an inner product. The proof of the general triangle or Minkowski inequality for $p \ne 1, 2$ is again not trivial. The limiting L^∞ norm is defined by the maximum

$$\| f \|_\infty = \max \{ | f(x) | \ : \ a \le x \le b \}\,. \tag{3.28}$$

Example 3.15. Consider the polynomial $p(x) = 3x^2 - 2$ on the interval $-1 \le x \le 1$. Its L^2 norm is

$$\| p \|_2 = \sqrt{\int_{-1}^{1} (3x^2 - 2)^2 \, dx} \;=\; \sqrt{\frac{18}{5}} \;= 1.8974\ldots .$$

Its L^∞ norm is

$$\| p \|_\infty = \max \big\{ \, | \, 3x^2 - 2 \, | \;:\; -1 \le x \le 1 \, \big\} = 2,$$

with the maximum occurring at $x = 0$. Finally, its L^1 norm is

$$\begin{aligned} \| p \|_1 &= \int_{-1}^{1} | \, 3x^2 - 2 \, | \, dx \\ &= \int_{-1}^{-\sqrt{2/3}} (3x^2 - 2) \, dx + \int_{-\sqrt{2/3}}^{\sqrt{2/3}} (2 - 3x^2) \, dx + \int_{\sqrt{2/3}}^{1} (3x^2 - 2) \, dx \\ &= \left(\tfrac{4}{3}\sqrt{\tfrac{2}{3}} - 1 \right) + \tfrac{8}{3}\sqrt{\tfrac{2}{3}} + \left(\tfrac{4}{3}\sqrt{\tfrac{2}{3}} - 1 \right) = \tfrac{16}{3}\sqrt{\tfrac{2}{3}} - 2 = 2.3546\ldots . \end{aligned}$$

Every norm defines a *distance* between vector space elements, namely

$$d(\mathbf{v}, \mathbf{w}) = \| \, \mathbf{v} - \mathbf{w} \, \|. \tag{3.29}$$

For the standard dot product norm, we recover the usual notion of distance between points in Euclidean space. Other types of norms produce alternative (and sometimes quite useful) notions of distance that, nevertheless, satisfy all the familiar distance axioms. Notice that distance is symmetric, $d(\mathbf{v}, \mathbf{w}) = d(\mathbf{w}, \mathbf{v})$. Moreover, $d(\mathbf{v}, \mathbf{w}) = 0$ if and only if $\mathbf{v} = \mathbf{w}$. The triangle inequality implies that

$$d(\mathbf{v}, \mathbf{w}) \le d(\mathbf{v}, \mathbf{z}) + d(\mathbf{z}, \mathbf{w}) \tag{3.30}$$

for any triple of vectors $\mathbf{v}, \mathbf{w}, \mathbf{z}$.

Unit Vectors

Let V be a fixed normed vector space. The elements $\mathbf{u} \in V$ with unit norm $\| \, \mathbf{u} \, \| = 1$ play a special role, and are known as *unit vectors* (or functions). The following easy lemma shows how to construct a unit vector pointing in the same direction as any given nonzero vector.

Lemma 3.16. *If $\mathbf{v} \neq \mathbf{0}$ is any nonzero vector, then the vector $\mathbf{u} = \mathbf{v}/\| \, \mathbf{v} \, \|$ obtained by dividing $\mathbf{v}$ by its norm is a unit vector parallel to $\mathbf{v}$.*

Proof: We compute, making use of the homogeneity property of the norm:

$$\| \, \mathbf{u} \, \| = \left\| \, \frac{\mathbf{v}}{\| \, \mathbf{v} \, \|} \, \right\| = \frac{\| \, \mathbf{v} \, \|}{\| \, \mathbf{v} \, \|} = 1. \qquad \textit{Q.E.D.}$$

Example 3.17. The vector $\mathbf{v} = (1, -2)^T$ has length $\| \mathbf{v} \|_2 = \sqrt{5}$ with respect to the standard Euclidean norm. Therefore, the unit vector pointing in the same direction as $\mathbf{v}$ is

$$\mathbf{u} = \frac{\mathbf{v}}{\| \mathbf{v} \|_2} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix}.$$

On the other hand, for the 1 norm, $\| \mathbf{v} \|_1 = 3$, and so

$$\widetilde{\mathbf{u}} = \frac{\mathbf{v}}{\| \mathbf{v} \|_1} = \frac{1}{3} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{2}{3} \end{pmatrix}$$

is the unit vector parallel to $\mathbf{v}$ in the 1 norm. Finally, $\| \mathbf{v} \|_\infty = 2$, and hence the corresponding unit vector for the ∞ norm is

$$\widehat{\mathbf{u}} = \frac{\mathbf{v}}{\| \mathbf{v} \|_\infty} = \frac{1}{2} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -1 \end{pmatrix}.$$

Thus, the notion of unit vector will depend upon which norm is being used.

Example 3.18. Similarly, on the interval $[0, 1]$, the quadratic polynomial $p(x) = x^2 - \frac{1}{2}$ has L^2 norm

$$\| p \|_2 = \sqrt{\int_0^1 \left(x^2 - \tfrac{1}{2} \right)^2 dx} = \sqrt{\int_0^1 \left(x^4 - x^2 + \tfrac{1}{4} \right) dx} = \sqrt{\frac{7}{60}}.$$

Therefore, $u(x) = \dfrac{p(x)}{\| p \|} = \frac{\sqrt{60}}{\sqrt{7}} x^2 - \frac{\sqrt{15}}{\sqrt{7}}$ is a "unit polynomial", $\| u \|_2 = 1$, which is "parallel" to (or, more correctly, a scalar multiple of) the polynomial p. On the other hand, for the L^∞ norm,

$$\| p \|_\infty = \max \left\{ \left| x^2 - \tfrac{1}{2} \right| \;\middle|\; 0 \leq x \leq 1 \right\} = \tfrac{1}{2},$$

and hence, in this case $\widetilde{u}(x) = 2p(x) = 2x^2 - 1$ is the corresponding unit function.

The *unit sphere* for the given norm is defined as the set of all unit vectors

$$S_1 = \left\{ \| \mathbf{u} \| = 1 \right\} \subset V. \tag{3.31}$$

Thus, the unit sphere for the Euclidean norm on $\mathbb{R}^n$ is the usual round sphere

$$S_1 = \left\{ \| \mathbf{x} \|^2 = x_1^2 + x_2^2 + \cdots + x_n^2 = 1 \right\}.$$

For the ∞ norm, it is the unit cube

$$S_1 = \left\{ \mathbf{x} \in \mathbb{R}^n \;\middle|\; x_1 = \pm 1 \text{ or } x_2 = \pm 1 \text{ or } \ \ldots \ \text{ or } x_n = \pm 1 \right\}.$$

For the 1 norm, it is the unit diamond or "octahedron"

$$S_1 = \left\{ \mathbf{x} \in \mathbb{R}^n \;\middle|\; |x_1| + |x_2| + \cdots + |x_n| = 1 \right\}.$$

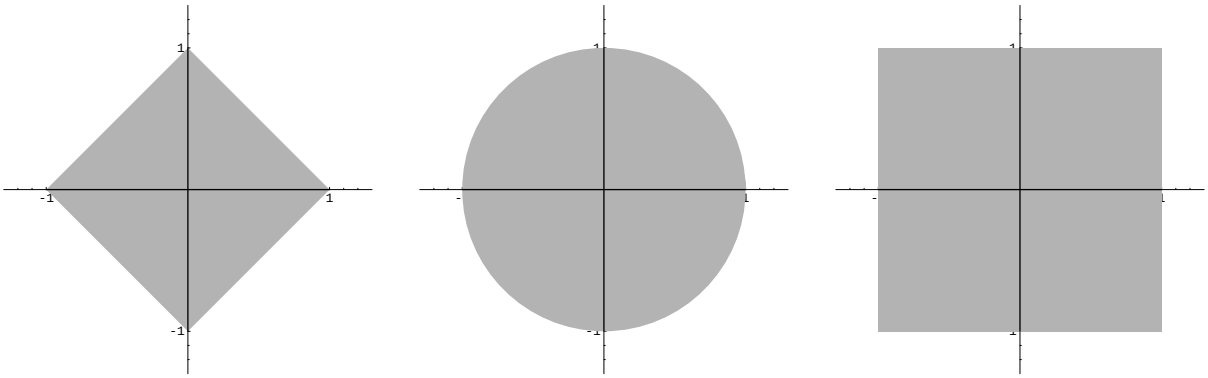


Figure 3.4. Unit Balls and Spheres for 1, 2 and ∞ Norms in $\mathbb{R}^2$.

See Figure 3.4 for the two-dimensional pictures.

In all cases, the *closed unit ball* $B_1 = \{ \| \mathbf{u} \| \leq 1 \}$ consists of all vectors of norm less than or equal to 1, and has the unit sphere as its boundary. If V is a finite-dimensional normed vector space, then the unit ball B_1 forms a *compact* subset, meaning that it is closed and bounded. This topological fact, which is *not* true in infinite-dimensional spaces, underscores the fundamental distinction between finite-dimensional vector analysis and the vastly more complicated infinite-dimensional realm.

Equivalence of Norms

While there are many different types of norms, in a finite-dimensional vector space they are all more or less equivalent. Equivalence does not mean that they assume the same value, but rather that they are, in a certain sense, always close to one another, and so for most analytical purposes can be used interchangeably. As a consequence, we may be able to simplify the analysis of a problem by choosing a suitably adapted norm.

Theorem 3.19. *Let $\| \cdot \|_1$ and $\| \cdot \|_2$ be any two norms on $\mathbb{R}^n$. Then there exist positive constants $c^\star, C^\star > 0$ such that*

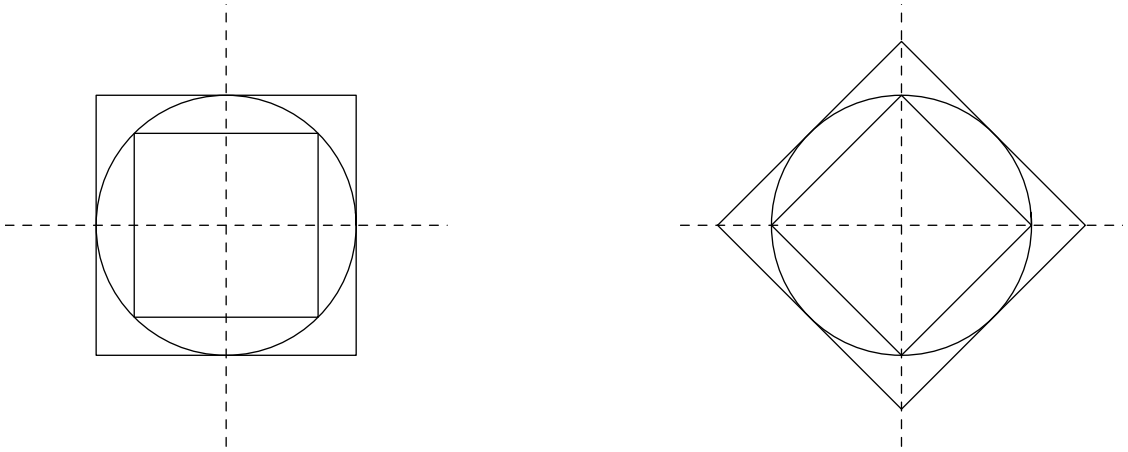
$$c^\star \, \| \mathbf{v} \|_1 \leq \| \mathbf{v} \|_2 \leq C^\star \, \| \mathbf{v} \|_1 \qquad \text{for every} \qquad \mathbf{v} \in \mathbb{R}^n. \tag{3.32}$$

Proof: We just sketch the basic idea, leaving the details to a more rigorous real analysis course, cf. [**125**, **126**]. We begin by noting that a norm defines a continuous function $f(\mathbf{v}) = \| \mathbf{v} \|$ on $\mathbb{R}^n$. (Continuity is, in fact, a consequence of the triangle inequality.) Let $S_1 = \{ \| \mathbf{u} \|_1 = 1 \}$ denote the unit sphere of the first norm. Any continuous function defined on a compact set achieves both a maximum and a minimum value. Thus, restricting the second norm function to the unit sphere S_1 of the first norm, we can set

$$c^\star = \| \mathbf{u}^\star \|_2 = \min \{ \| \mathbf{u} \|_2 \mid \mathbf{u} \in S_1 \}, \qquad C^\star = \| \mathbf{U}^\star \|_2 = \max \{ \| \mathbf{u} \|_2 \mid \mathbf{u} \in S_1 \}, \tag{3.33}$$

for certain vectors $\mathbf{u}^\star, \mathbf{U}^\star \in S_1$. Note that $0 < c^\star \leq C^\star < \infty$, with equality holding if and only if the the norms are the same. The minimum and maximum (3.33) will serve as the constants in the desired inequalities (3.32). Indeed, by definition,

$$c^\star \leq \| \mathbf{u} \|_2 \leq C^\star \qquad \text{when} \qquad \| \mathbf{u} \|_1 = 1, \tag{3.34}$$



∞ norm and 2 norm 1 norm and 2 norm

Figure 3.5. Equivalence of Norms.

and so (3.32) is valid for all $\mathbf{u} \in S_1$. To prove the inequalities in general, assume $\mathbf{v} \neq \mathbf{0}$. (The case $\mathbf{v} = \mathbf{0}$ is trivial.) Lemma 3.16 says that $\mathbf{u} = \mathbf{v}/\| \mathbf{v} \|_1 \in S_1$ is a unit vector in the first norm: $\| \mathbf{u} \|_1 = 1$. Moreover, by the homogeneity property of the norm, $\| \mathbf{u} \|_2 = \| \mathbf{v} \|_2 / \| \mathbf{v} \|_1$. Substituting into (3.34) and clearing denominators completes the proof of (3.32). *Q.E.D.*

Example 3.20. For example, consider the Euclidean norm $\| \cdot \|_2$ and the max norm $\| \cdot \|_\infty$ on $\mathbb{R}^n$. According to (3.33), the bounding constants are found by minimizing and maximizing $\| \mathbf{u} \|_\infty = \max\{ | u_1 |, \dots, | u_n | \}$ over all unit vectors $\| \mathbf{u} \|_2 = 1$ on the (round) unit sphere. Its maximal value is obtained at the poles, when $\mathbf{U}^\star = \pm \mathbf{e}_k$, with $\| \mathbf{e}_k \|_\infty = 1$. Thus, $C^\star = 1$. The minimal value is obtained when $\mathbf{u}^\star = \left(\dfrac{1}{\sqrt{n}}, \dots, \dfrac{1}{\sqrt{n}} \right)$ has all equal components, whereby $c^\star = \| \mathbf{u} \|_\infty = 1/\sqrt{n}$. Therefore,

$$\frac{1}{\sqrt{n}} \| \mathbf{v} \|_2 \leq \| \mathbf{v} \|_\infty \leq \| \mathbf{v} \|_2. \tag{3.35}$$

One can interpret these inequalities as follows. Suppose $\mathbf{v}$ is a vector lying on the unit sphere in the Euclidean norm, so $\| \mathbf{v} \|_2 = 1$. Then (3.35) tells us that its ∞ norm is bounded from above and below by $1/\sqrt{n} \leq \| \mathbf{v} \|_\infty \leq 1$. Therefore, the unit Euclidean sphere sits inside the unit sphere in the ∞ norm, and outside the sphere of radius $1/\sqrt{n}$. Figure 3.5 illustrates the two-dimensional situation.

One significant consequence of the equivalence of norms is that, in $\mathbb{R}^n$, convergence is independent of the norm. The following are all equivalent to the standard ε–δ convergence of a sequence $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \mathbf{u}^{(3)}, \dots$ of vectors in $\mathbb{R}^n$:

(*a*) the vectors converge: $\mathbf{u}^{(k)} \longrightarrow \mathbf{u}^\star$:

(*b*) the individual components all converge: $u_i^{(k)} \longrightarrow u_i^\star$ for $i = 1, \dots, n$.

(*c*) the difference in norms goes to zero: $\| \mathbf{u}^{(k)} - \mathbf{u}^\star \| \longrightarrow 0$.

The last case, called *convergence in norm*, does not depend on which norm is chosen. Indeed, the basic inequality (3.32) implies that if one norm goes to zero, so does any other

norm. An important consequence is that all norms on $\mathbb{R}^n$ induce the same topology — convergence of sequences, notions of open and closed sets, and so on. None of this is true in infinite-dimensional function space! A rigorous development of the underlying topological and analytical properties of compactness, continuity, and convergence is beyond the scope of this course. The motivated student is encouraged to consult a text in real analysis, e.g., [**125**, **126**], to find the relevant definitions, theorems and proofs.

Example 3.21. Consider the infinite-dimensional vector space $\mathrm{C}^0[0,1]$ consisting of all continuous functions on the interval $[0,1]$. The functions

$$f_n(x) = \begin{cases} 1 - n\,x, & 0 \le x \le \frac{1}{n}, \\ 0, & \frac{1}{n} \le x \le 1, \end{cases}$$

have identical L^∞ norms

$$\| f_n \|_\infty = \sup \{ \, | f_n(x) | \, | \;\; 0 \le x \le 1 \, \} = 1.$$

On the other hand, their L^2 norm

$$\| f_n \|_2 = \sqrt{\int_0^1 f_n(x)^2 \, dx} = \sqrt{\int_0^{1/n} (1 - n\,x)^2 \, dx} = \frac{1}{\sqrt{3\,n}}$$

goes to zero as $n \to \infty$. This example shows that there *is no* constant $C^\star$ such that

$$\| f \|_\infty \le C^\star \| f \|_2$$

for all $f \in \mathrm{C}^0[0,1]$. The L^∞ and L^2 norms on $\mathrm{C}^0[0,1]$ are not equivalent — there exist functions which have unit L^2 norm but arbitrarily small L^∞ norm. Similar inequivalence properties apply to all of the other standard function space norms. As a result, the topology on function space is intimately connected with the underlying choice of norm.

3.4. Positive Definite Matrices.

Let us now return to the study of inner products, and fix our attention on the finite-dimensional situation. Our immediate goal is to determine the most general inner product which can be placed on the finite-dimensional vector space $\mathbb{R}^n$. The resulting analysis will lead us to the extremely important class of positive definite matrices. Such matrices play a fundamental role in a wide variety of applications, including minimization problems, mechanics, electrical circuits, and differential equations. Moreover, their infinite-dimensional generalization to positive definite linear operators underlie all of the most important examples of boundary value problems for ordinary and partial differential equations.

Let $\langle \, \mathbf{x} \, ; \mathbf{y} \, \rangle$ denote an inner product between vectors $\mathbf{x} = (\, x_1 \; x_2 \; \dots \; x_n \,)^T$, $\mathbf{y} = (\, y_1 \; y_2 \; \dots \; y_n \,)^T$, in $\mathbb{R}^n$. Let us write the vectors in terms of the standard basis vectors:

$$\mathbf{x} = x_1 \, \mathbf{e}_1 + \, \cdots \, + x_n \, \mathbf{e}_n = \sum_{i=1}^n x_i \, \mathbf{e}_i, \qquad \mathbf{y} = y_1 \, \mathbf{e}_1 + \, \cdots \, + y_n \, \mathbf{e}_n = \sum_{j=1}^n y_j \, \mathbf{e}_j. \quad (3.36)$$

Let us carefully analyze the three basic inner product axioms, in order. We use the bilinearity of the inner product to expand

$$\langle \mathbf{x}\,;\mathbf{y}\,\rangle = \left\langle \sum_{i=1}^{n} x_i\,\mathbf{e}_i\,;\, \sum_{j=1}^{n} y_j\,\mathbf{e}_j \right\rangle = \sum_{i,j=1}^{n} x_i\,y_j \langle\, \mathbf{e}_i\,;\mathbf{e}_j\,\rangle.$$

Therefore we can write

$$\langle \mathbf{x}\,;\mathbf{y}\,\rangle = \sum_{i,j=1}^{n} k_{ij}\,x_i\,y_j = \mathbf{x}^T K\,\mathbf{y}, \tag{3.37}$$

where K denotes the $n \times n$ matrix of inner products of the basis vectors, with entries

$$k_{ij} = \langle\, \mathbf{e}_i\,;\mathbf{e}_j\,\rangle, \qquad i,j = 1,\dots,n. \tag{3.38}$$

We conclude that any inner product must be expressed in the general *bilinear form* (3.37).

The two remaining inner product axioms will impose certain conditions on the inner product matrix K. The symmetry of the inner product implies that

$$k_{ij} = \langle\, \mathbf{e}_i\,;\mathbf{e}_j\,\rangle = \langle\, \mathbf{e}_j\,;\mathbf{e}_i\,\rangle = k_{ji}, \qquad i,j = 1,\dots,n.$$

Consequently, the inner product matrix K is symmetric:

$$K = K^T.$$

Conversely, symmetry of K ensures symmetry of the bilinear form:

$$\langle \mathbf{x}\,;\mathbf{y}\,\rangle = \mathbf{x}^T K\,\mathbf{y} = (\mathbf{x}^T K\,\mathbf{y})^T = \mathbf{y}^T K^T \mathbf{x} = \mathbf{y}^T K\,\mathbf{x} = \langle\, \mathbf{y}\,;\mathbf{x}\,\rangle,$$

where the second equality follows from the fact that the quantity is a scalar, and hence equals its transpose.

The final condition for an inner product is positivity. This requires that

$$\|\,\mathbf{x}\,\|^2 = \langle\, \mathbf{x}\,;\mathbf{x}\,\rangle = \mathbf{x}^T K\,\mathbf{x} = \sum_{i,j=1}^{n} k_{ij}\,x_i\,x_j \geq 0 \qquad \text{for all} \qquad \mathbf{x} \in \mathbb{R}^n, \tag{3.39}$$

with equality if and only if $\mathbf{x} = \mathbf{0}$. The precise meaning of this positivity condition on the matrix K is not as immediately evident, and so will be encapsulated in the following very important definition.

Definition 3.22. An $n \times n$ matrix K is called *positive definite* if it is symmetric, $K^T = K$, and satisfies the positivity condition

$$\mathbf{x}^T K\,\mathbf{x} > 0 \qquad \text{for all} \qquad \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n. \tag{3.40}$$

We will sometimes write $K > 0$ to mean that K is a symmetric, positive definite matrix.

Warning: The condition $K > 0$ does *not* mean that all the entries of K are positive. There are many positive definite matrices which have some negative entries — see Example 3.24 below. Conversely, many symmetric matrices with all positive entries are not positive definite!

Remark: Although some authors allow non-symmetric matrices to be designated as positive definite, we will *only* say that a matrix is positive definite when it is symmetric. But, to underscore our convention and so as not to confuse the casual reader, we will often include the adjective "symmetric" when speaking of positive definite matrices.

Our preliminary analysis has resulted in the following characterization of inner products on a finite-dimensional vector space.

Theorem 3.23. *Every inner product on $\mathbb{R}^n$ is given by*

$$\langle \mathbf{x}\,;\mathbf{y}\,\rangle = \mathbf{x}^T K\,\mathbf{y}, \qquad \text{for} \qquad \mathbf{x},\mathbf{y} \in \mathbb{R}^n, \tag{3.41}$$

where K is a symmetric, positive definite matrix.

Given a symmetric[†] matrix K, the expression

$$q(\mathbf{x}) = \mathbf{x}^T K\,\mathbf{x} = \sum_{i,j=1}^{n} k_{ij}\,x_i\,x_j, \tag{3.42}$$

is known as a *quadratic form* on $\mathbb{R}^n$. The quadratic form is called *positive definite* if

$$q(\mathbf{x}) > 0 \qquad \text{for all} \qquad 0 \neq \mathbf{x} \in \mathbb{R}^n. \tag{3.43}$$

Thus, a quadratic form is positive definite if and only if its coefficient matrix is.

Example 3.24. Even though the symmetric matrix $K = \begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix}$ has two negative entries, it is, nevertheless, a positive definite matrix. Indeed, the corresponding quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K\,\mathbf{x} = 4x_1^2 - 4x_1x_2 + 3x_2^2 = \big(\,2x_1 - x_2\,\big)^2 + 2x_2^2 \geq 0$$

is a sum of two non-negative quantities. Moreover, $q(\mathbf{x}) = 0$ if and only if both terms are zero, which requires that $2x_1 - x_2 = 0$ and $x_2 = 0$, whereby $x_1 = 0$ also. This proves positivity for all nonzero $\mathbf{x}$, and hence $K > 0$ is indeed a positive definite matrix. The corresponding inner product on $\mathbb{R}^2$ is

$$\langle \mathbf{x}\,;\mathbf{y}\,\rangle = (\,x_1\;\,x_2\,)\begin{pmatrix} 4 & -2 \\ -2 & 3 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = 4x_1y_1 - 2x_1y_2 - 2x_2y_1 + 3x_2y_2.$$

On the other hand, despite the fact that the matrix $K = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ has all positive entries, it is *not* a positive definite matrix. Indeed, writing out

$$q(\mathbf{x}) = \mathbf{x}^T K\,\mathbf{x} = x_1^2 + 4x_1x_2 + x_2^2,$$

we find, for instance, that $q(1,-1) = -2 < 0$, violating positivity. These two simple examples should be enough to convince the reader that the problem of determining whether a given symmetric matrix is or is not positive definite is not completely elementary.

[†] Exercise ■ shows that the coefficient matrix K in any quadratic form can be taken to be symmetric without any loss of generality.

With a little practice, it is not difficult to read off the coefficient matrix K from the explicit formula for the quadratic form (3.42).

Example 3.25. Consider the quadratic form

$$q(x,y,z) = x^2 + 4\,x\,y + 6\,y^2 - 2\,x\,z + 9\,z^2$$

depending upon three variables. The corresponding coefficient matrix is

$$K = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix} \qquad \text{whereby} \qquad q(x,y,z) = \begin{pmatrix} x & y & z \end{pmatrix} \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Note that the squared terms in q contribute directly to the diagonal entries of K, while the mixed terms are split in half to give the symmetric off-diagonal entries. The reader might wish to try proving that this particular matrix is positive definite by proving positivity of the quadratic form: $q(x,y,z) > 0$ for all nonzero $\begin{pmatrix} x, y, z \end{pmatrix}^T \in \mathbb{R}^3$. Later, we will establish a systematic test for positive definiteness.

Slightly more generally, a quadratic form and its associated symmetric coefficient matrix are called *positive semi-definite* if

$$q(\mathbf{x}) = \mathbf{x}^T K\,\mathbf{x} \geq 0 \qquad \text{for all} \qquad \mathbf{x} \in \mathbb{R}^n. \tag{3.44}$$

A positive semi-definite matrix may have *null directions*, meaning non-zero vectors $\mathbf{z}$ such that $q(\mathbf{z}) = \mathbf{z}^T K\,\mathbf{z} = \mathbf{0}$. Clearly any vector $\mathbf{z} \in \ker K$ that lies in the matrix's kernel defines a null direction, but there may be others. In particular, a positive definite matrix is not allowed to have null directions, so $\ker K = \{\mathbf{0}\}$. Proposition 2.39 implies that all positive definite matrices are invertible.

Theorem 3.26. *All positive definite matrices K are non-singular.*

Example 3.27. The matrix $K = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ is positive semi-definite, but not positive definite. Indeed, the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K\,\mathbf{x} = x_1^2 - 2\,x_1\,x_2 + x_2^2 = (x_1 - x_2)^2 \geq 0$$

is a perfect square, and so clearly non-negative. However, the elements of $\ker K$, namely the scalar multiples of the vector $\begin{pmatrix} 1 & 1 \end{pmatrix}^T$, define null directions, since $q(1,1) = 0$.

Example 3.28. A general symmetric 2×2 matrix $K = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is positive definite if and only if the associated quadratic form satisfies

$$q(\mathbf{x}) = a\,x_1^2 + 2\,b\,x_1\,x_2 + c\,x_2^2 > 0 \tag{3.45}$$

for all $\mathbf{x} \neq \mathbf{0}$. Analytic geometry tells us that this is the case if and only if

$$a > 0, \qquad a\,c - b^2 > 0, \tag{3.46}$$

i.e., the quadratic form has positive leading coefficient and positive determinant (or negative discriminant). A direct proof of this elementary fact will appear shortly.

Remark: A quadratic form $q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}$ and its associated symmetric matrix K are called *negative semi-definite* if $q(\mathbf{x}) \leq 0$ for all $\mathbf{x}$ and *negative definite* if $q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$. A quadratic form is called *indefinite* if it is neither positive nor negative semi-definite; equivalently, there exist one or more points $\mathbf{x}_+$ where $q(\mathbf{x}_+) > 0$ and one or more points $\mathbf{x}_-$ where $q(\mathbf{x}_-) < 0$.

Gram Matrices

Symmetric matrices whose entries are given by inner products of elements of an inner product space play an important role. They are named after the nineteenth century Danish mathematician Jorgen Gram (not the metric mass unit).

Definition 3.29. Let V be an inner product space, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. The associated *Gram matrix*

$$K = \begin{pmatrix} \langle \mathbf{v}_1 ; \mathbf{v}_1 \rangle & \langle \mathbf{v}_1 ; \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_1 ; \mathbf{v}_n \rangle \\ \langle \mathbf{v}_2 ; \mathbf{v}_1 \rangle & \langle \mathbf{v}_2 ; \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_2 ; \mathbf{v}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_n ; \mathbf{v}_1 \rangle & \langle \mathbf{v}_n ; \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_n ; \mathbf{v}_n \rangle \end{pmatrix}. \tag{3.47}$$

is the $n \times n$ matrix whose entries are the inner products between the chosen vector space elements.

Symmetry of the inner product implies symmetry of the Gram matrix:

$$k_{ij} = \langle \mathbf{v}_i ; \mathbf{v}_j \rangle = \langle \mathbf{v}_j ; \mathbf{v}_i \rangle = k_{ji}, \qquad \text{and hence} \qquad K^T = K. \tag{3.48}$$

In fact, the most direct method for producing positive definite and semi-definite matrices is through the Gram matrix construction.

Theorem 3.30. *All Gram matrices are positive semi-definite. A Gram matrix is positive definite if and only if the elements $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$ are linearly independent.*

Proof: To prove positive (semi-)definiteness of K, we need to examine the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} = \sum_{i,j=1}^{n} k_{ij} x_i x_j.$$

Substituting the values (3.48) for the matrix entries, we find

$$q(\mathbf{x}) = \sum_{i,j=1}^{n} \langle \mathbf{v}_i ; \mathbf{v}_j \rangle x_i x_j.$$

Bilinearity of the inner product on V implies that we can assemble this summation into a single inner product

$$q(\mathbf{x}) = \left\langle \sum_{i=1}^{n} x_i \mathbf{v}_i \; ; \; \sum_{j=1}^{n} x_j \mathbf{v}_j \right\rangle = \langle \mathbf{v} ; \mathbf{v} \rangle = \| \mathbf{v} \|^2 \geq 0,$$

where

$$\mathbf{v} = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n$$

lies in the subspace of V spanned by the given vectors. This immediately proves that K is positive semi-definite.

Moreover, $q(\mathbf{x}) = \| \mathbf{v} \|^2 > 0$ as long as $\mathbf{v} \neq \mathbf{0}$. If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent, then $\mathbf{v} = \mathbf{0}$ if and only if $x_1 = \cdots = x_n = 0$, and hence, in this case, $q(\mathbf{x})$ and K are positive definite. *Q.E.D.*

Example 3.31. Consider the vectors $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 6 \end{pmatrix}$ in $\mathbb{R}^3$. For the standard Euclidean dot product, the Gram matrix is

$$K = \begin{pmatrix} \mathbf{v}_1 \cdot \mathbf{v}_1 & \mathbf{v}_1 \cdot \mathbf{v}_2 \\ \mathbf{v}_2 \cdot \mathbf{v}_1 & \mathbf{v}_2 \cdot \mathbf{v}_2 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ -3 & 45 \end{pmatrix}.$$

Positive definiteness implies that the associated quadratic form

$$q(x_1, x_2) = 6 x_1^2 - 6 x_1 x_2 + 45 x_2^2 > 0$$

is positive for all $(x_1, x_2) \neq \mathbf{0}$. This can be checked directly using the criteria in (3.46).

On the other hand, if we use the weighted inner product $\langle \mathbf{x} ; \mathbf{y} \rangle = 3 x_1 y_1 + 2 x_2 y_2 + 5 x_3 y_3$, then the corresponding Gram matrix is

$$K = \begin{pmatrix} \langle \mathbf{v}_1 ; \mathbf{v}_1 \rangle & \langle \mathbf{v}_1 ; \mathbf{v}_2 \rangle \\ \langle \mathbf{v}_2 ; \mathbf{v}_1 \rangle & \langle \mathbf{v}_2 ; \mathbf{v}_2 \rangle \end{pmatrix} = \begin{pmatrix} 16 & -21 \\ -21 & 207 \end{pmatrix},$$

which, by construction, is also positive definite.

In the case of the Euclidean dot product, the construction of the Gram matrix K can be directly implemented as follows. Given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$, let us form the $m \times n$ matrix $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \ldots \ \mathbf{v}_n)$ whose columns are the vectors in question. Owing to the identification (3.2) between the dot product and multiplication of row and column vectors, the (i, j) entry of K is given as the product

$$k_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{v}_i^T \mathbf{v}_j$$

of the i^{th} row of the transpose A^T with the j^{th} column of A. In other words, the Gram matrix

$$K = A^T A \tag{3.49}$$

is the matrix product of the transpose of A with A. For the preceding Example 3.31,

$$A = \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ -1 & 6 \end{pmatrix}, \qquad \text{and so} \qquad K = A^T A = \begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & 6 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} 6 & -3 \\ -3 & 45 \end{pmatrix}.$$

Theorem 3.30 implies that the Gram matrix (3.49) is positive definite if and only if the columns of A are linearly independent. This implies the following result.

Proposition 3.32. *Given an $m \times n$ matrix A, the following are equivalent:*

(*i*) *The $n \times n$ Gram matrix $K = A^T A$ is positive definite.*

(*ii*) *A has linearly independent columns.*

(*iii*) $\operatorname{rank} A = n \leq m$.

(*iv*) $\ker A = \{0\}$.

As noted above, Gram matrices can be based on more general inner products on more general vector spaces. Let us consider an alternative inner product on the finite-dimensional vector space $\mathbb{R}^m$. As noted in Theorem 3.23, a general inner product on $\mathbb{R}^m$ has the form

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \mathbf{v}^T C\,\mathbf{w} \qquad \text{for} \qquad \mathbf{v},\mathbf{w} \in \mathbb{R}^m, \tag{3.50}$$

where $C > 0$ is a symmetric, positive definite $m \times m$ matrix. Therefore, given n vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^m$, the entries of the corresponding Gram matrix are the products

$$k_{ij} = \langle\, \mathbf{v}_i\,;\mathbf{v}_j\,\rangle = \mathbf{v}_i^T C\,\mathbf{v}_j.$$

If we assemble the column vectors as above into an $m \times n$ matrix $A = (\,\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n\,)$, then the Gram inner products are given by multiplying the i^{th} row of A^T by the j^{th} column of the product matrix $C\,A$. Therefore, the Gram matrix based on the alternative inner product (3.50) is given by

$$K = A^T C\,A. \tag{3.51}$$

Theorem 3.30 immediately implies that K is positive definite — provided A has rank n.

Theorem 3.33. *Suppose A is an $m \times n$ matrix with linearly independent columns. Suppose $C > 0$ is any positive definite $m \times m$ matrix. Then the matrix $K = A^T C\,A$ is a positive definite $n \times n$ matrix.*

The Gram matrix K constructed in (3.51) arises in a wide range of applications, including weighted least squares approximation theory, cf. Chapter 4, the study of equilibrium of mechanical and electrical systems, cf. Chapter 6. Starting in Chapter 11, we shall look at infinite-dimensional generalizations that apply to differential equations and boundary value problems.

Example 3.34. In the majority of applications, $C = \operatorname{diag}(c_1, \ldots, c_m)$ is a diagonal positive definite matrix, which requires it to have strictly positive diagonal entries $c_i > 0$. This choice corresponds to a weighted inner product (3.10) on $\mathbb{R}^m$. For example, if we set $C = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix}$, then the weighted Gram matrix based on the vectors $\begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$, $\begin{pmatrix} 3 \\ 0 \\ 6 \end{pmatrix}$ of Example 3.31 is

$$K = A^T C\,A = \begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & 6 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 2 & 0 \\ -1 & 6 \end{pmatrix} = \begin{pmatrix} 16 & -21 \\ -21 & 207 \end{pmatrix},$$

reproducing the second part of Example 3.31.

The Gram construction also carries over to inner products on function space. Here is a particularly important example.

Example 3.35. Consider vector space $\mathrm{C}^0[0,1]$ consisting of continuous functions on the interval $0 \le x \le 1$, equipped with the L^2 inner product $\langle\, f\,;g\,\rangle = \displaystyle\int_0^1 f(x)\,g(x)\,dx$. Let us construct the Gram matrix corresponding to the elementary monomial functions $1, x, x^2$. We compute the required inner products

$$\begin{aligned}
\langle\, 1\,;1\,\rangle &= \|\,1\,\|^2 = \int_0^1 dx = 1, & \langle\, 1\,;x\,\rangle &= \int_0^1 x\,dx = \frac{1}{2}\,,\\
\langle\, x\,;x\,\rangle &= \|\,x\,\|^2 = \int_0^1 x^2\,dx = \frac{1}{3}, & \langle\, 1\,;x^2\,\rangle &= \int_0^1 x^2\,dx = \frac{1}{3}\,,\\
\langle\, x^2\,;x^2\,\rangle &= \|\,x^2\,\|^2 = \int_0^1 x^4\,dx = \frac{1}{5}, & \langle\, x\,;x^2\,\rangle &= \int_0^1 x^3\,dx = \frac{1}{4}\,.
\end{aligned}$$

Therefore, the Gram matrix is

$$K = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix}.$$

The monomial functions $1, x, x^2$ are linearly independent. Therefore, Theorem 3.30 implies that this particular matrix is positive definite.

The alert reader may recognize this Gram matrix $K = H_3$ as the 3×3 *Hilbert matrix* that we encountered in (1.67). More generally, the Gram matrix corresponding to the monomials $1, x, x^2, \ \dots\ , x^n$ has entries

$$k_{ij} = \langle\, x^i\,;x^j\,\rangle = \int_0^1 x^{i+j}\,dt = \frac{1}{i+j+1}\,, \qquad i,j = 0,\dots,n.$$

Therefore, the monomial Gram matrix $K = H_{n+1}$ is the $(n+1) \times (n+1)$ Hilbert matrix (1.67). As a consequence of Theorems 3.26 and 3.33, we have proved the following non-trivial result.

Proposition 3.36. *The $n \times n$ Hilbert matrix H_n is positive definite. In particular, H_n is a nonsingular matrix.*

Example 3.37. Let us construct the Gram matrix corresponding to the functions $1, \cos x, \sin x$ with respect to the inner product $\langle\, f\,;g\,\rangle = \displaystyle\int_{-\pi}^{\pi} f(x)\,g(x)\,dx$ on the interval $[-\pi, \pi]$. We compute the inner products

$$\begin{aligned}
\langle\, 1\,;1\,\rangle &= \|\,1\,\|^2 = \int_{-\pi}^{\pi} dx = 2\,\pi, & \langle\, 1\,;\cos x\,\rangle &= \int_{-\pi}^{\pi} \cos x\,dx = 0,\\
\langle\, \cos x\,;\cos x\,\rangle &= \|\,\cos x\,\|^2 = \int_{-\pi}^{\pi} \cos^2 x\,dx = \pi, & \langle\, 1\,;\sin x\,\rangle &= \int_{-\pi}^{\pi} \sin x\,dx = 0,\\
\langle\, \sin x\,;\sin x\,\rangle &= \|\,\sin x\,\|^2 = \int_{-\pi}^{\pi} \sin^2 x\,dx = \pi, & \langle\, \cos x\,;\sin x\,\rangle &= \int_{-\pi}^{\pi} \cos x \sin x\,dx = 0.
\end{aligned}$$

Therefore, the Gram matrix is a simple diagonal matrix $K = \begin{pmatrix} 2\pi & 0 & 0 \\ 0 & \pi & 0 \\ 0 & 0 & \pi \end{pmatrix}$. Positive definiteness of K is immediately evident.

3.5. Completing the Square.

Gram matrices furnish us with an abundant supply of positive definite matrices. However, we still do not know how to test whether a given symmetric matrix is positive definite. As we shall soon see, the secret already appears in the particular computations in Examples 3.2 and 3.24.

The student may recall the importance of the method known as "completing the square", first in the derivation of the quadratic formula for the solution to

$$q(x) = a\,x^2 + 2\,b\,x + c = 0, \tag{3.52}$$

and, later, in the integration of various types of rational functions. The key idea is to combine the first two terms in (3.52) as a perfect square, and so rewrite the quadratic function in the form

$$q(x) = a\left(x + \frac{b}{a}\right)^2 + \frac{a\,c - b^2}{a} = 0. \tag{3.53}$$

As a consequence,

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

The quadratic formula

$$x = \frac{-\,b \pm \sqrt{b^2 - a\,c}}{a}$$

follows by taking the square root of both sides and then solving for x. The intermediate step (3.53), where we eliminate the linear term, is known as *completing the square*.

We can perform the same manipulation on the corresponding homogeneous quadratic form

$$q(x_1, x_2) = a\,x_1^2 + 2\,b\,x_1\,x_2 + c\,x_2^2. \tag{3.54}$$

We write

$$q(x_1, x_2) = a\,x_1^2 + 2\,b\,x_1 x_2 + c\,x_2^2 = a\left(x_1 + \frac{b}{a}\,x_2\right)^2 + \frac{a\,c - b^2}{a}\,x_2^2 = a\,y_1^2 + \frac{a\,c - b^2}{a}\,y_2^2 \tag{3.55}$$

as a sum of squares of the new variables

$$y_1 = x_1 + \frac{b}{a}\,x_2, \qquad y_2 = x_2. \tag{3.56}$$

Since $y_1 = y_2 = 0$ if and only if $x_1 = x_2 = 0$, the final expression is positive definite if and only if both coefficients are positive:

$$a > 0, \qquad \frac{a\,c - b^2}{a} > 0.$$

This proves that conditions (3.46) are necessary and sufficient for the quadratic form (3.45) to be positive definite.

How this simple idea can be generalized to the multi-variable case will become clear if we write the quadratic form identity (3.55) in matrix form. The original quadratic form (3.54) is

$$q(\mathbf{x}) = \mathbf{x}^T K \mathbf{x}, \qquad \text{where} \qquad K = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{3.57}$$

The second quadratic form in (3.55) is

$$\widehat{q}\,(\mathbf{y}) = \mathbf{y}^T D \mathbf{y}, \qquad \text{where} \qquad D = \begin{pmatrix} a & 0 \\ 0 & \dfrac{a\,c - b^2}{a} \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \tag{3.58}$$

Anticipating the final result, the equation connecting $\mathbf{x}$ and $\mathbf{y}$ can be written in matrix form as

$$\mathbf{y} = L^T \mathbf{x} \qquad \text{or} \qquad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 + \dfrac{b}{a}\, x_2 \\ x_2 \end{pmatrix}, \qquad \text{where} \qquad L^T = \begin{pmatrix} 1 & 0 \\ \dfrac{b}{a} & 1 \end{pmatrix}.$$

Substituting into (3.58), we find

$$\mathbf{y}^T D \mathbf{y} = (L^T \mathbf{x})^T D\, (L^T \mathbf{x}) = \mathbf{x}^T L\, D\, L^T \mathbf{x} = \mathbf{x}^T K \mathbf{x}, \qquad \text{where} \qquad K = L\, D\, L^T \tag{3.59}$$

is precisely the $L\,D\,L^T$ factorization of $K = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ that appears in (1.56). We are thus led to the important conclusion that *completing the square is the same as the $L\,D\,L^T$ factorization of a symmetric matrix*, obtained through Gaussian elimination!

Recall the definition of a regular matrix as one that can be reduced to upper triangular form without any row interchanges;Theorem 1.32 says that these are the matrices admitting an $L\,D\,L^T$ factorization. The identity (3.59) is therefore valid for all regular $n \times n$ symmetric matrices, and shows how to write the associated quadratic form as a sum of squares:

$$\widehat{q}\,(\mathbf{y}) = \mathbf{y}^T D \mathbf{y} = d_1\, y_1^2 + \,\cdots\, + d_n\, y_n^2. \tag{3.60}$$

The coefficients d_i are the pivots of K. In particular, according to Exercise ■, $\widehat{q}\,(\mathbf{y}) > 0$ is positive definite if and only if all the pivots are positive: $d_i > 0$. Let us now state the main result that completely characterizes positive definite matrices.

Theorem 3.38. *A symmetric matrix K is positive definite if and only if it is regular and has all positive pivots. Consequently, K is positive definite if and only if it can be factored $K = L\,D\,L^T$, where L is special lower triangular, and D is diagonal with all positive diagonal entries.*

Example 3.39. Consider the symmetric matrix $K = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix}$. Gaussian elimination produces the factors

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{pmatrix}, \qquad L^T = \begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

in the factorization $K = L\,D\,L^T$. Since the pivots — the diagonal entries $1, 2, 6$ in D — are all positive, Theorem 3.38 implies that K is positive definite, which means that the associated quadratic form

$$q(\mathbf{x}) = x_1^2 + 4\,x_1\,x_2 - 2\,x_1\,x_3 + 6\,x_2^2 + 9\,x_3^2 > 0, \qquad \text{for all} \qquad \mathbf{x} = (\,x_1, x_2, x_3\,)^T \neq \mathbf{0}.$$

Indeed, the $L\,D\,L^T$ factorization implies that $q(\mathbf{x})$ can be explicitly written as a sum of squares:

$$q(\mathbf{x}) = y_1^2 + 2\,y_2^2 + 6\,y_3^2, \qquad \text{where} \qquad y_1 = x_1 + 2\,x_2 - x_3, \quad y_2 = x_2 + x_3, \quad y_3 = x_3,$$

are the entries of $\mathbf{y} = L^T\mathbf{x}$. Positivity of the coefficients of the y_i^2 (which are the pivots) implies that $q(\mathbf{x})$ is positive definite.

On the other hand, for the $L\,D\,L^T$ factorization

$$K = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 6 & 2 \\ 3 & 2 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -9 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix},$$

the fact that D has a negative diagonal entry, -9, implies that K is *not* positive definite — even though all its entries are positive. The associated quadratic form is

$$q(\mathbf{x}) = x_1^2 + 4\,x_1\,x_2 + 6\,x_1\,x_3 + 6\,x_2^2 + 4\,x_2\,x_3 + 8\,x_3^2$$

is not positive definite since, for instance, $q(-5, 2, 1) = -9 < 0$.

The only remaining issue is to show that an irregular matrix cannot be positive definite. For example, the quadratic form corresponding to the irregular matrix $K = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, is $q(\mathbf{x}) = 2\,x_1\,x_2$, which is clearly not positive definite, e.g., $q(1, -1) = -2$. In general, if the upper left entry $k_{11} = 0$, then it cannot serve as the first pivot, and so K is not regular. But then $q(\mathbf{e}_1) = \mathbf{e}_1^T K\,\mathbf{e}_1 = 0$, and so K is not positive definite. (It may be positive semi-definite, or, more likely, indefinite.)

Otherwise, if $k_{11} \neq 0$, then we use Gaussian elimination to make all entries lying in the first column below the pivot equal to zero. As remarked above, this is equivalent to completing the square in the initial terms of the associated quadratic form

$$\begin{aligned} q(\mathbf{x}) &= k_{11}\,x_1^2 + 2\,k_{12}\,x_1\,x_2 + \,\cdots\, + 2\,k_{1n}\,x_1\,x_n + k_{22}\,x_2^2 + \,\cdots\, + k_{nn}\,x_n^2 \\ &= k_{11}\left(x_1 + \frac{k_{12}}{k_{11}}\,x_2 + \,\cdots\, + \frac{k_{1n}}{k_{11}}\,x_n\right)^2 + \widetilde{q}(x_2, \ldots, x_n) \\ &= k_{11}\,(x_1 + l_{21}\,x_2 + \,\cdots\, + l_{n1}\,x_n)^2 + \widetilde{q}(x_2, \ldots, x_n), \end{aligned} \tag{3.61}$$

where

$$l_{21} = \frac{k_{21}}{k_{11}} = \frac{k_{12}}{k_{11}}\,, \qquad \cdots \qquad l_{n1} = \frac{k_{n1}}{k_{11}} = \frac{k_{1n}}{k_{11}}\,,$$

are precisely the multiples appearing in the first column of the lower triangular matrix L obtained from Gaussian Elimination, while

$$\widetilde{q}(x_2, \ldots, x_n) = \sum_{i,j=2}^{n} \widetilde{k}_{ij}\,x_i\,x_j$$

is a quadratic form involving one fewer variable. The entries of its symmetric coefficient matrix $\widetilde{K}$ are

$$\widetilde{k}_{ij} = \widetilde{k}_{ji} = k_{ij} - l_{j1}\, k_{1i}, \qquad \text{for} \qquad i \geq j.$$

Thus, the entries of $\widetilde{K}$ that lie on or below the diagonal are exactly the *same* as the entries appearing on or below the diagonal of K after the the first phase of the elimination process. In particular, the second pivot of K is the entry $\widetilde{k}_{22}$ that appears in the corresponding slot in $\widetilde{K}$. If $\widetilde{q}$ is not positive definite, then q cannot be positive definite. Indeed, suppose that there exist $x_2^\star, \dots, x_n^\star$, not all zero, such that $\widetilde{q}(x_2^\star, \dots, x_n^\star) \leq 0$. Setting

$$x_1^\star = -\, l_{21}\, x_2^\star - \ \cdots \ - l_{n1}\, x_n^\star,$$

makes the initial square term in (3.61) equal to 0, so $q(x_1^\star, x_2^\star, \dots, x_n^\star) = \widetilde{q}(x_2^\star, \dots, x_n^\star) \leq 0$. In particular, if the second diagonal entry $\widetilde{k}_{22} = 0$, then $\widetilde{q}$ is not positive definite, and so neither is q. Continuing this process, if any diagonal entry of the reduced matrix vanishes, then the reduced quadratic form cannot be positive definite, and so neither can q. This demonstrates that if K is irregular, then it cannot be positive definite, which completes the proof of Theorem 3.38.

The Cholesky Factorization

The identity (3.59) shows us how to write any regular quadratic form $q(\mathbf{x})$ as a sum of squares. One can push this result slightly further in the positive definite case. Since each pivot $d_i > 0$, we can write the diagonal form (3.60) as a sum of squares with unit coefficients:

$$\widehat{q}\,(\mathbf{y}) = d_1\, y_1^2 + \ \cdots \ + d_n\, y_n^2 = \big(\sqrt{d_1}\, y_1\big)^2 + \ \cdots \ + \big(\sqrt{d_n}\, y_n\big)^2 = z_1^2 + \ \cdots \ + z_n^2,$$

where $z_i = \sqrt{d_i}\, y_i$. In matrix form, we are writing

$$\widehat{q}\,(\mathbf{y}) = \mathbf{y}^T D\, \mathbf{y} = \mathbf{z}^T \mathbf{z} = \| \mathbf{z} \|^2, \quad \text{where} \quad \mathbf{z} = C\, \mathbf{y}, \quad \text{with} \quad C = \operatorname{diag}\big(\sqrt{d_1}\,, \dots, \sqrt{d_n}\,\big)$$

Since $D = C^2$, the matrix C can be thought of as a "square root" of the diagonal matrix D. Substituting back into (1.52), we deduce the *Cholesky factorization*

$$K = L\, D\, L^T = L\, C\, C^T\, L^T = M\, M^T, \qquad \text{where} \qquad M = L\, C \tag{3.62}$$

of a positive definite matrix. Note that M is a lower triangular matrix with all positive entries, namely the square roots of the pivots $m_{ii} = c_i = \sqrt{d_i}$ on its diagonal. Applying the Cholesky factorization to the corresponding quadratic form produces

$$q(\mathbf{x}) = \mathbf{x}^T K\, \mathbf{x} = \mathbf{x}^T M\, M^T \mathbf{x} = \mathbf{z}^T \mathbf{z} = \| \mathbf{z} \|^2, \qquad \text{where} \qquad \mathbf{z} = M^T \mathbf{x}. \tag{3.63}$$

One can interpret this as a change of variables from $\mathbf{x}$ to $\mathbf{z}$ that converts an arbitrary inner product norm, as defined by the square root of the positive definite quadratic form $q(\mathbf{x})$, into the standard Euclidean norm $\| \mathbf{z} \|$.

Example 3.40. For the matrix $K = \begin{pmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{pmatrix}$ considered in Example 3.39, the Cholesky formula (3.62) gives $K = M\,M^T$, where

$$M = LC = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{6} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & \sqrt{2} & 0 \\ -1 & \sqrt{2} & \sqrt{6} \end{pmatrix}.$$

The associated quadratic function can then be written as a sum of pure squares:

$$q(\mathbf{x}) = x_1^2 + 4\,x_1\,x_2 - 2\,x_1\,x_3 + 6\,x_2^2 + 9\,x_3^2 = z_1^2 + z_2^2 + z_3^2,$$

where $\mathbf{z} = M^T\mathbf{x}$, or, explicitly, $z_1 = x_1 + 2\,x_2 - x_3$, $z_2 = \sqrt{2}\,x_2 + \sqrt{2}\,x_3$, $z_3 = \sqrt{6}\,x_3$..

3.6. Complex Vector Spaces.

Although physical applications ultimately require real answers, complex numbers and complex vector spaces assume an extremely useful, if not essential role in the intervening analysis. Particularly in the description of periodic phenomena, complex numbers and complex exponentials assume a central role, dramatically simplifying complicated trigonometric formulae. Complex variable methods are essential in fluid mechanics, electrical engineering, Fourier analysis, potential theory, electromagnetism, and so on. In quantum mechanics, complex numbers are ubiquitous. The basic physical quantities are complex wave functions. Moreover, the Schrödinger equation, which is the basic equation governing quantum systems, is a complex partial differential equation with complex-valued solutions.

In this section, we survey the basic facts about complex numbers and complex vector spaces. Most of the constructions are entirely analogous to the real case, and will not be dwelled on at length. The one exception is the complex version of an inner product, which does introduce some novelties not found in its simpler real counterpart. Complex analysis (integration and differentiation of complex functions) and its applications to fluid flows, potential theory, waves and other areas of mathematics, physics and engineering, will be the subject of Chapter 16.

Complex Numbers

Recall that a complex number is an expression of the form $z = x + \mathrm{i}\,y$, where x, y are real and[†] $\mathrm{i} = \sqrt{-1}$. We call $x = \operatorname{Re} z$ the *real part* of z and $y = \operatorname{Im} z$ the *imaginary part*. (*Note*: The imaginary part is the real number y, *not* $\mathrm{i}\,y$.) A real number x is merely a complex number with zero imaginary part: $\operatorname{Im} z = 0$. Complex addition and multiplication are based on simple adaptations of the rules of real arithmetic to include the identity $\mathrm{i}^{\,2} = -1$, and so

$$\begin{aligned} (x + \mathrm{i}\,y) + (u + \mathrm{i}\,v) &= (x + u) + \mathrm{i}\,(y + v), \\ (x + \mathrm{i}\,y) \cdot (u + \mathrm{i}\,v) &= (x\,u - y\,v) + \mathrm{i}\,(x\,v + y\,u). \end{aligned} \tag{3.64}$$

[†] Electrical engineers prefer to use j to indicate the imaginary unit.

Complex numbers enjoy all the usual laws of real addition and multiplication, *including commutativity*: $z\,w = w\,z$.

We can identity a complex number $x + \mathrm{i}\,y$ with a vector $(\,x, y\,)^T \in \mathbb{R}^2$ in the real plane. Complex addition (3.64) corresponds to vector addition, but complex multiplication does *not* have a readily identifiable vector counterpart.

Another important operation on complex numbers is that of complex conjugation.

Definition 3.41. The *complex conjugate* of $z = x + \mathrm{i}\,y$ is $\overline{z} = x - \mathrm{i}\,y$, whereby $\operatorname{Re}\overline{z} = \operatorname{Re} z$, $\operatorname{Im}\overline{z} = -\operatorname{Im} z$.

Geometrically, the operation of complex conjugation coincides with reflection of the corresponding vector through the real axis, as illustrated in Figure 3.6. In particular $\overline{z} = z$ if and only if z is real. Note that

$$\operatorname{Re} z = \frac{z + \overline{z}}{2}\,, \qquad \operatorname{Im} z = \frac{z - \overline{z}}{2\,\mathrm{i}}\,. \tag{3.65}$$

Complex conjugation is compatible with complex arithmetic:

$$\overline{z + w} = \overline{z} + \overline{w}, \qquad \overline{z\,w} = \overline{z}\,\overline{w}.$$

In particular, the product of a complex number and its conjugate

$$z\,\overline{z} = (x + \mathrm{i}\,y)\,(x - \mathrm{i}\,y) = x^2 + y^2 \tag{3.66}$$

is real and non-negative. Its square root is known as the *modulus* of the complex number $z = x + \mathrm{i}\,y$, and written

$$|\,z\,| = \sqrt{x^2 + y^2}\,. \tag{3.67}$$

Note that $|\,z\,| \geq 0$, with $|\,z\,| = 0$ if and only if $z = 0$. The modulus $|\,z\,|$ generalizes the absolute value of a real number, and coincides with the standard Euclidean norm in the (x, y)–plane. This implies the validity of the triangle inequality

$$|\,z + w\,| \leq |\,z\,| + |\,w\,|. \tag{3.68}$$

Equation (3.66) can be rewritten in terms of the modulus as

$$z\,\overline{z} = |\,z\,|^2. \tag{3.69}$$

Rearranging the factors, we deduce the formula for the reciprocal of a nonzero complex number:

$$\frac{1}{z} = \frac{\overline{z}}{|\,z\,|^2}\,, \qquad z \neq 0, \qquad \text{or, equivalently} \qquad \frac{1}{x + \mathrm{i}\,y} = \frac{x - \mathrm{i}\,y}{x^2 + y^2}\,. \tag{3.70}$$

The general formula for complex division

$$\frac{w}{z} = \frac{w\,\overline{z}}{|\,z\,|^2} \qquad \text{or, equivalently} \qquad \frac{u + \mathrm{i}\,v}{x + \mathrm{i}\,y} = \frac{(x\,u + y\,v) + \mathrm{i}\,(x\,v - y\,u)}{x^2 + y^2}\,, \tag{3.71}$$

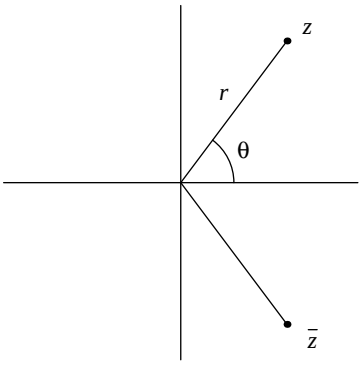


Figure 3.6. Complex Numbers.

is an immediate consequence.

The modulus of a complex number,

$$r = |\, z \,| = \sqrt{x^2 + y^2}\,,$$

is one component of its polar coordinate representation

$$x = r\cos\theta, \qquad y = r\sin\theta \qquad \text{or} \qquad z = r(\cos\theta + \mathrm{i}\,\sin\theta). \tag{3.72}$$

The polar angle, which measures the angle that the line connecting z to the origin makes with the horizontal axis, is known as the *phase*, and written

$$\mathrm{ph}\, z = \theta. \tag{3.73}$$

The more common term is the *argument*, and written $\arg z = \mathrm{ph}\, z$. For various reasons, and to avoid confusion with the argument of a function, we have chosen to use "phase" throughout this text. As such, the phase is only defined up to an integer multiple of 2π. We note that the modulus and phase of a product of complex numbers can be readily computed:

$$|\, z\, w \,| = |\, z \,|\, |\, w \,|, \qquad \mathrm{ph}\,(z\, w) = \mathrm{ph}\, z + \mathrm{ph}\, w. \tag{3.74}$$

On the other hand, complex conjugation preserves the modulus, but negates the phase:

$$|\, \overline{z} \,| = |\, z \,|, \qquad \mathrm{ph}\, \overline{z} = -\,\mathrm{ph}\, z. \tag{3.75}$$

One of the most important formulas in all of mathematics is *Euler's formula*

$$e^{\,\mathrm{i}\,\theta} = \cos\theta + \mathrm{i}\,\sin\theta, \tag{3.76}$$

relating the complex exponential with the real sine and cosine functions. This basic identity has a variety of mathematical justifications; see Exercise ■ for one that is based on comparing power series. Euler's formula (3.76) can be used to compactly rewrite the polar form (3.72) of a complex number as

$$z = r\, e^{\,\mathrm{i}\,\theta} \qquad \text{where} \qquad r = |\, z \,|, \qquad \theta = \mathrm{ph}\, z. \tag{3.77}$$

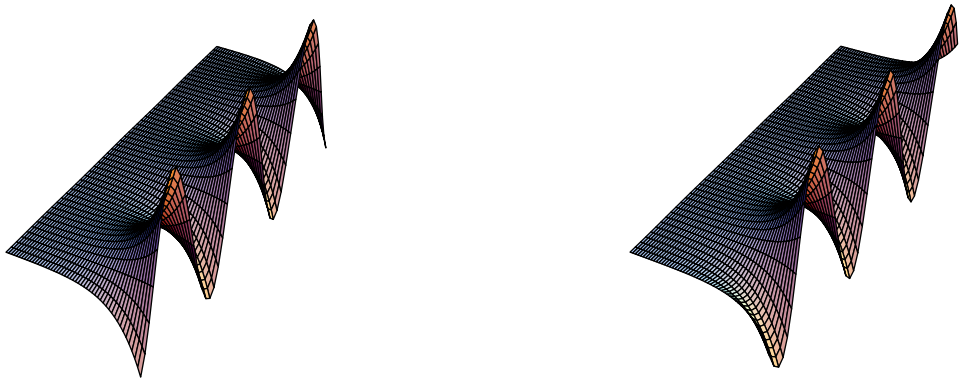


Figure 3.7. Real and Imaginary Parts of e^z.

The complex conjugate identity

$$e^{-\mathrm{i}\,\theta} = \cos(-\theta) + \mathrm{i}\,\sin(-\theta) = \cos\theta - \mathrm{i}\,\sin\theta = \overline{e^{\mathrm{i}\,\theta}}\,,$$

permits us to express the basic trigonometric functions in terms of complex exponentials:

$$\cos\theta = \frac{e^{\mathrm{i}\,\theta} + e^{-\mathrm{i}\,\theta}}{2}\,, \qquad \sin\theta = \frac{e^{\mathrm{i}\,\theta} - e^{-\mathrm{i}\,\theta}}{2\,\mathrm{i}}\,. \tag{3.78}$$

These formulae are very useful when working with trigonometric identities and integrals.

The exponential of a general complex number is easily derived from the basic Euler formula and the standard properties of the exponential function — which carry over unaltered to the complex domain; thus,

$$e^z = e^{x+\mathrm{i}\,y} = e^x\, e^{\mathrm{i}\,y} = e^x \cos y + \mathrm{i}\, e^x \sin y. \tag{3.79}$$

Graphs of the real and imaginary parts of the complex exponential appear in Figure 3.7. Note that $e^{2\,\pi\,\mathrm{i}} = 1$, and hence the exponential function is periodic

$$e^{z+2\,\pi\,\mathrm{i}} = e^z \tag{3.80}$$

with imaginary period $2\,\pi\,\mathrm{i}$ — a reflection of the periodicity of the trigonometric functions in Euler's formula.

Complex Vector Spaces and Inner Products

A *complex vector space* is defined in exactly the same manner as its real cousin, cf. Definition 2.1, the only difference being that we replace real scalars $\mathbb{R}$ by complex scalars $\mathbb{C}$. The most basic example is the n-dimensional complex vector space $\mathbb{C}^n$ consisting of all column vectors $\mathbf{z} = (\,z_1, z_2, \ldots, z_n\,)^T$ that have n complex entries: $z_1, \ldots, z_n \in \mathbb{C}$. Verification of each of the vector space axioms is a straightforward exercise.

We can write any complex vector $\mathbf{z} = \mathbf{x} + \mathrm{i}\,\mathbf{y} \in \mathbb{C}^n$ as a linear combination of two real vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Its complex conjugate $\overline{\mathbf{z}} = \mathbf{x} - \mathrm{i}\,\mathbf{y}$ is obtained by taking the complex

conjugates of its individual entries. Thus, for example, if

$$\mathbf{z} = \begin{pmatrix} 1 + 2\,\mathrm{i} \\ -3 \\ 5\,\mathrm{i} \end{pmatrix}, \qquad \text{then} \qquad \overline{\mathbf{z}} = \begin{pmatrix} 1 - 2\,\mathrm{i} \\ -3 \\ -5\,\mathrm{i} \end{pmatrix}.$$

In particular, $\mathbf{z} \in \mathbb{R}^n \subset \mathbb{C}^n$ is a real vector if and only if $\mathbf{z} = \overline{\mathbf{z}}$.

Most of the vector space concepts we developed in the real domain, including span, linear independence, basis, and dimension, can be straightforwardly extended to the complex regime. The one exception is the concept of an inner product, which requires a little thought. In analysis, the most important applications of inner products and norms are based on the associated inequalities: Cauchy–Schwarz and triangle. But there is no natural ordering of the complex numbers, and so one *cannot* make any sense of a complex inequality like $z < w$. Inequalities only make sense in the real domain, and so the norm of a complex vector should still be a positive, real number.

With this in mind, the naïve idea of simply summing the squares of the entries of a complex vector will *not* define a norm on $\mathbb{C}^n$, since the result will typically be complex. Moreover, this would give some nonzero complex vectors, e.g., $(\,1 \quad \mathrm{i}\,)^T$, a zero "norm", violating positivity[†].

The correct definition is modeled on the definition of the modulus

$$|\,z\,| = \sqrt{z\,\overline{z}}$$

of a complex scalar $z \in \mathbb{C}$. If, in analogy with the real definition (3.7), the quantity inside the square root is to represent the inner product of z with itself, then we should define the dot product between two complex numbers to be

$$z \cdot w = z\,\overline{w}, \qquad \text{so that} \qquad z \cdot z = z\,\overline{z} = |\,z\,|^2.$$

If $z = x + \mathrm{i}\,y$ and $w = u + \mathrm{i}\,v$, then

$$z \cdot w = z\,\overline{w} = (x + \mathrm{i}\,y)\,(u - \mathrm{i}\,v) = (x\,u + y\,v) + \mathrm{i}\,(y\,u - x\,v). \tag{3.81}$$

Thus, the dot product of two complex numbers is, in general, complex. The real part of $z \cdot w$ is, in fact, the Euclidean dot product between the corresponding vectors in $\mathbb{R}^2$, while the imaginary part is, interestingly, their scalar cross-product, cf. (cross2).

The vector version of this construction is named after the nineteenth century French mathematician Charles Hermite, and called the *Hermitian dot product* on $\mathbb{C}^n$. It has the explicit formula

$$\mathbf{z} \cdot \mathbf{w} = \mathbf{z}^T\,\overline{\mathbf{w}} = z_1\,\overline{w}_1 + z_2\,\overline{w}_2 + \,\cdots\, + z_n\,\overline{w}_n, \quad \text{for} \quad \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix}. \tag{3.82}$$

[†] On the other hand, in relativity, the Minkowski "norm" is also not always positive, and indeed the vectors with zero norm play a critical role as they lie on the light cone emanating from the origin, [**106**].

Pay attention to the fact that we must apply complex conjugation to all the entries of the second vector. For example, if $\mathbf{z} = \begin{pmatrix} 1+\mathrm{i} \\ 3+2\,\mathrm{i} \end{pmatrix}$, $\mathbf{w} = \begin{pmatrix} 1+2\,\mathrm{i} \\ \mathrm{i} \end{pmatrix}$, then

$$\mathbf{z} \cdot \mathbf{w} = (1+\mathrm{i})(1-2\,\mathrm{i}) + (3+2\,\mathrm{i})(-\mathrm{i}) = 5 - 4\,\mathrm{i}\,.$$

On the other hand,

$$\mathbf{w} \cdot \mathbf{z} = (1+2\,\mathrm{i})(1-\mathrm{i}) + \mathrm{i}\,(3-2\,\mathrm{i}) = 5 + 4\,\mathrm{i}\,.$$

Therefore, the Hermitian dot product is *not* symmetric. Reversing the order of the vectors results in complex conjugation of the dot product:

$$\mathbf{w} \cdot \mathbf{z} \;=\; \overline{\mathbf{z} \cdot \mathbf{w}}.$$

But this extra complication does have the effect that the induced norm, namely

$$0 \le \| \mathbf{z} \| = \sqrt{\mathbf{z} \cdot \mathbf{z}} = \sqrt{\mathbf{z}^T \overline{\mathbf{z}}} = \sqrt{|\, z_1 \,|^2 + \,\cdots\, + |\, z_n \,|^2}\,, \tag{3.83}$$

is strictly positive for all $\mathbf{0} \neq \mathbf{z} \in \mathbb{C}^n$. For example, if

$$\mathbf{z} = \begin{pmatrix} 1+3\,\mathrm{i} \\ -2\,\mathrm{i} \\ -5 \end{pmatrix}, \qquad \text{then} \qquad \| \mathbf{z} \| = \sqrt{|\, 1+3\,\mathrm{i} \,|^2 + |\, -2\,\mathrm{i} \,|^2 + |\, -5 \,|^2} = \sqrt{39}\,.$$

The Hermitian dot product is well behaved under complex vector addition:

$$(\mathbf{z} + \widehat{\mathbf{z}}) \cdot \mathbf{w} = \mathbf{z} \cdot \mathbf{w} + \widehat{\mathbf{z}} \cdot \mathbf{w}, \qquad\qquad \mathbf{z} \cdot (\mathbf{w} + \widehat{\mathbf{w}}) = \mathbf{z} \cdot \mathbf{w} + \mathbf{z} \cdot \widehat{\mathbf{w}}.$$

However, while complex scalar multiples can be extracted from the first vector without alteration, when they multiply the second vector, they emerge as complex conjugates:

$$(c\,\mathbf{z}) \cdot \mathbf{w} = c\,(\mathbf{z} \cdot \mathbf{w}), \qquad\qquad \mathbf{z} \cdot (c\,\mathbf{w}) = \overline{c}\,(\mathbf{z} \cdot \mathbf{w}), \qquad\qquad c \in \mathbb{C}.$$

Thus, the Hermitian dot product is not bilinear in the strict sense, but satisfies something that, for lack of a better name, is known as *sesqui-linearity*.

The general definition of an inner product on a complex vector space is modeled on the preceding properties of the Hermitian dot product.

Definition 3.42. An *inner product* on the complex vector space V is a pairing that takes two vectors $\mathbf{v}, \mathbf{w} \in V$ and produces a complex number $\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle \in \mathbb{C}$, subject to the following requirements for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, and $c, d \in \mathbb{C}$.

(i) *Sesqui-linearity*:

$$\begin{aligned} \langle\, c\,\mathbf{u} + d\,\mathbf{v}\,;\mathbf{w}\,\rangle &= c\,\langle\, \mathbf{u}\,;\mathbf{w}\,\rangle + d\,\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle, \\ \langle\, \mathbf{u}\,;c\,\mathbf{v} + d\,\mathbf{w}\,\rangle &= \overline{c}\,\langle\, \mathbf{u}\,;\mathbf{v}\,\rangle + \overline{d}\,\langle\, \mathbf{u}\,;\mathbf{w}\,\rangle. \end{aligned} \tag{3.84}$$

(ii) *Conjugate Symmetry*:

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \overline{\langle\, \mathbf{w}\,;\mathbf{v}\,\rangle}. \tag{3.85}$$

(iii) *Positivity*:

$$\| \mathbf{v} \|^2 = \langle\, \mathbf{v}\,;\mathbf{v}\,\rangle \ge 0, \quad \text{and} \quad \langle\, \mathbf{v}\,;\mathbf{v}\,\rangle = 0 \quad \text{if and only if} \quad \mathbf{v} = \mathbf{0}. \tag{3.86}$$

Thus, when dealing with a complex inner product space, one must pay careful attention to the complex conjugate that appears when the second argument in the inner product is multiplied by a complex scalar, as well as the complex conjugate that appears when switching the order of the two arguments.

Theorem 3.43. *The Cauchy–Schwarz inequality,*

$$|\langle \mathbf{v}\,;\mathbf{w}\rangle| \le \|\mathbf{v}\|\,\|\mathbf{w}\|, \qquad \mathbf{v},\mathbf{w}\in V.$$

with $|\cdot|$ now denoting the complex modulus, and the triangle inequality

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

hold for any complex inner product space.

The proof of this result is practically the same as in the real case, and the details are left to the reader.

Example 3.44. The vectors $\mathbf{v} = (1+\mathrm{i}, 2\,\mathrm{i}, -3)^T$, $\mathbf{w} = (2-\mathrm{i}, 1, 2+2\,\mathrm{i})^T$, satisfy

$$\|\mathbf{v}\| = \sqrt{2+4+9} = \sqrt{15}, \qquad \|\mathbf{w}\| = \sqrt{5+1+8} = \sqrt{14},$$
$$\mathbf{v}\cdot\mathbf{w} = (1+\mathrm{i})(2+\mathrm{i}) + 2\,\mathrm{i} + (-3)(2-2\,\mathrm{i}) = -5+11\,\mathrm{i}.$$

Thus, the Cauchy–Schwarz inequality reads

$$|\langle \mathbf{v}\,;\mathbf{w}\rangle| = |-5+11\,\mathrm{i}| = \sqrt{146} \le \sqrt{210} = \sqrt{15}\,\sqrt{14} = \|\mathbf{v}\|\,\|\mathbf{w}\|.$$

Similarly, the triangle inequality tells us that

$$\|\mathbf{v}+\mathbf{w}\| = \|(3, 1+2\,\mathrm{i}, -1+2\,\mathrm{i})^T\| = \sqrt{9+5+5} = \sqrt{19} \le \sqrt{15}+\sqrt{14} = \|\mathbf{v}\|+\|\mathbf{w}\|.$$

Example 3.45. Let $\mathrm{C}^0 = \mathrm{C}^0[-\pi,\pi]$ denote the complex vector space consisting of all complex valued continuous functions $f(x) = u(x) + \mathrm{i}\,v(x)$ depending upon the *real* variable $-\pi \le x \le \pi$. The Hermitian L^2 inner product is defined as

$$\langle f\,;g\rangle = \int_{-\pi}^{\pi} f(x)\,\overline{g(x)}\,dx\,, \tag{3.87}$$

with corresponding norm

$$\|f\| = \sqrt{\int_{-\pi}^{\pi} |f(x)|^2\,dx} = \sqrt{\int_{-\pi}^{\pi} \big[\,u(x)^2+v(x)^2\,\big]\,dx}\,. \tag{3.88}$$

The reader should check that (3.87) satisfies the basic Hermitian inner product axioms.

For example, if k, l are integers, then the inner product of the complex exponential functions $e^{\mathrm{i}\,kx}$ and $e^{\mathrm{i}\,lx}$ is

$$\langle e^{\mathrm{i}\,kx}\,;e^{\mathrm{i}\,lx}\rangle = \int_{-\pi}^{\pi} e^{\mathrm{i}\,kx}e^{-\mathrm{i}\,lx}\,dx = \int_{-\pi}^{\pi} e^{\mathrm{i}\,(k-l)x}\,dx = \begin{cases} 2\pi, & k=l,\\ \dfrac{e^{\mathrm{i}\,(k-l)x}}{\mathrm{i}\,(k-l)}\,\Big|_{x=-\pi}^{\pi} = 0, & k\ne l. \end{cases}$$

We conclude that when $k \ne l$, the complex exponentials $e^{\mathrm{i}\,kx}$ and $e^{\mathrm{i}\,lx}$ are orthogonal, since their inner product is zero. This key example will be of fundamental significance in the complex version of Fourier analysis.

Chapter 4

Minimization and Least Squares Approximation

Because Nature strives to be efficient, many systems arising in applications are founded on a minimization principle. For example, in a mechanical system, the stable equilibrium positions minimize the potential energy. The basic geometrical problem of minimizing distance also appears in many contexts. For example, in optics and relativity, light rays follow the paths of minimal distance — the geodesics on the curved space-time. In data analysis, the most fundamental method for fitting a function to a set of sampled data points is to minimize the least squares error, which serves as a measurement of the overall deviation between the sample data and the function. The least squares paradigm carries over to a wide range of applied mathematical systems. In particular, it underlies the theory of Fourier series, in itself of inestimable importance in mathematics, physics and engineering. Solutions to many of the important boundary value problems arising in mathematical physics and engineering are also characterized by an underlying minimization principle. Moreover, the finite element numerical solution method relies on the associated minimization principle. Optimization is ubiquitous in control theory, engineering design and manufacturing, linear programming, econometrics, and most other fields of analysis.

This chapter introduces and solves the most basic minimization problem — that of a quadratic function of several variables. The minimizer is found by solving an associated linear system. The solution to the quadratic minimization problem leads directly to a broad range of applications, including least squares fitting of data, interpolation, and approximation of functions. Applications to equilibrium mechanics will form the focus of Chapter 6. Applications to the numerical solution of differential equations in numerical analysis will appear starting in Chapter 11. More general nonlinear minimization problems, which, as usual, require a thorough analysis of the linear situation, will be deferred until Section 19.3.

4.1. Minimization Problems.

Let us begin by introducing three important minimization problems — one physical, one analytical, and one geometrical.

Equilibrium Mechanics

A fundamental principle of mechanics is that systems in equilibrium minimize potential energy. For example, a ball in a bowl will roll downhill until it reaches the bottom, where it minimizes its potential energy due to gravity. Similarly, a pendulum will swing back and forth unless it is at the bottom of its arc, where potential energy is minimized. Actually, the pendulum has a second equilibrium position at the top of the arc, but this

is an *unstable* equilibrium, meaning that any tiny movement will knock it off balance. Therefore, a better way of stating the principle is that *stable equilibria* are where the mechanical system minimizes potential energy. For the ball rolling on a curved surface, the local minima — the bottoms of valleys — are the stable equilibria, while the local maxima — the tops of hills — are unstable. This basic idea is fundamental to the understanding and analysis of the equilibrium configurations of a wide range of physical systems, including masses and springs, structures, electrical circuits, and even continuum models of solid mechanics and elasticity, fluid mechanics, electromagnetism, thermodynamics, statistical mechanics, and so on.

Solution of Equations

Suppose we wish to solve a system of equations

$$f_1(\mathbf{x}) = 0, \qquad f_2(\mathbf{x}) = 0, \qquad \dots \qquad f_m(\mathbf{x}) = 0, \tag{4.1}$$

where $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$. This system can be converted into a minimization problem in the following seemingly silly manner. Define

$$p(\mathbf{x}) = \big[\, f_1(\mathbf{x}) \,\big]^2 + \, \cdots \, + \big[\, f_m(\mathbf{x}) \,\big]^2 = \| \, \mathbf{f}(\mathbf{x}) \, \|^2, \tag{4.2}$$

where $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^m$. Clearly, $p(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. Moreover, $p(\mathbf{x}^\star) = 0$ if and only if each summand is zero, and hence $\mathbf{x}^\star$ is a solution to (4.1). Therefore, the minimum value of $p(\mathbf{x})$ is zero, and the minimum is achieved if and only if $\mathbf{x} = \mathbf{x}^\star$ solves the system (4.1).

The most important case is when we have a linear system

$$A\,\mathbf{x} = \mathbf{b} \tag{4.3}$$

consisting of m equations in n unknowns. In this case, the solutions may be obtained by minimizing the function

$$p(\mathbf{x}) = \| \, A\,\mathbf{x} - \mathbf{b} \, \|^2. \tag{4.4}$$

Of course, it is not clear that we have gained much, since we already know how to solve $A\,\mathbf{x} = \mathbf{b}$ by Gaussian elimination. However, this rather simple artifice has profound consequences.

Suppose that the system (4.3) does *not* have a solution, i.e., $\mathbf{b}$ does not lie in the range of the matrix A. This situation is very typical when there are more equations than unknowns: $m > n$. Such problems arise in data fitting, when the measured data points are all supposed to lie on a straight line, say, but rarely do so exactly, due to experimental error. Although we know there is no exact solution to the system, we might still try to find the vector $\mathbf{x}^\star$ that comes as close to solving the system as possible. One way to measure closeness is by looking at the magnitude of the residual vector $\mathbf{r} = A\,\mathbf{x} - \mathbf{b}$, i.e., the difference between the left and right hand sides of the system. The smaller $\| \, \mathbf{r} \, \| = \| \, A\,\mathbf{x} - \mathbf{b} \, \|$, the better the attempted solution. The vector $\mathbf{x}^\star$ that minimizes the function (4.4) is known as the *least squares solution* to the linear system. We note that if the linear system (4.3) happens to have a actual solution, with $A\,\mathbf{x}^\star = \mathbf{b}$, then $\mathbf{x}^\star$ qualifies as the least squares solution too, since in this case $p(\mathbf{x}^\star) = 0$ achieves its absolute minimum.

Thus, the least squares solutions naturally generalize traditional solutions. While not the only possible method, least squares is is easiest to analyze and solve, and hence, typically, the method of choice for fitting functions to experimental data and performing statistical analysis.

The Closest Point

The following minimization problem arises in elementary geometry. Given a point $\mathbf{b} \in \mathbb{R}^m$ and a subset $V \subset \mathbb{R}^m$, find the point $\mathbf{v}^\star \in V$ that is closest to $\mathbf{b}$. In other words, we seek to minimize the distance $d(\mathbf{b}, \mathbf{v}) = \| \mathbf{v} - \mathbf{b} \|$ over all possible $\mathbf{v} \in V$.

The simplest situation occurs when V is a subspace of $\mathbb{R}^m$. In this case, the closest point problem can be reformulated as a least squares minimization problem. Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ be a basis for V. The general element $\mathbf{v} \in V$ is a linear combination of the basis vectors. Applying our handy matrix multiplication formula (2.14), we can write the subspace elements in the form

$$\mathbf{v} = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = A\,\mathbf{x},$$

where $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ is the $m \times n$ matrix formed by the (column) basis vectors. Note that we can identify $V = \operatorname{rng} A$ with the range of A, i.e., the subspace spanned by its columns. Consequently, the closest point in V to $\mathbf{b}$ is found by minimizing

$$\| \mathbf{v} - \mathbf{b} \|^2 = \| A\,\mathbf{x} - \mathbf{b} \|^2$$

over all possible $\mathbf{x} \in \mathbb{R}^n$. This is exactly the same as the least squares function (4.4)! Thus, if $\mathbf{x}^\star$ is the least squares solution to the system $A\,\mathbf{x} = \mathbf{b}$, then $\mathbf{v}^\star = A\,\mathbf{x}^\star$ is the closest point to $\mathbf{b}$ belonging to $V = \operatorname{rng} A$. In this way, we have established a fundamental connection between least squares solutions to linear systems and the geometrical problem of minimizing distances to subspaces.

All three of the preceding minimization problems are solved by the same underlying mathematical construction, which will be described in detail in Section 4.3.

Remark: We will concentrate on minimization problems. Maximizing a function $f(\mathbf{x})$ is the same as minimizing its negative $-f(\mathbf{x})$, and so can be easily handled by the same methods.

4.2. Minimization of Quadratic Functions.

The simplest algebraic equations are the linear systems; these must be thoroughly understood before venturing into the far more complicated nonlinear realm. For minimization problems, the starting point is the minimization of a quadratic function. (Linear functions do not have minima — think of the function $f(x) = b\,x + c$ whose graph is a straight line.) In this section, we shall see how the problem of minimizing a general quadratic function of n variables can be solved by linear algebra techniques.

Let us begin by reviewing the very simplest example — minimizing a scalar quadratic function

$$p(x) = a\,x^2 + 2\,b\,x + c. \tag{4.5}$$

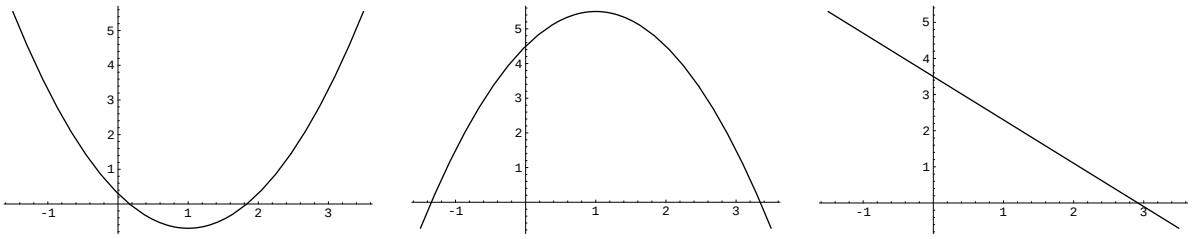


Figure 4.1. Parabolas.

If $a > 0$, then the graph of p is a parabola pointing upwards, and so there exists a unique minimum value. If $a < 0$, the parabola points downwards, and there is no minimum (although there is a maximum). If $a = 0$, the graph is a straight line, and there is neither minimum nor maximum — except in the trivial case when $b = 0$ also, and the function is constant, with every x qualifying as a minimum and a maximum. The three nontrivial possibilities are sketched in Figure 4.1.

In the case $a > 0$, the minimum can be found by calculus. The *critical points* of a function, which are candidates for minima (and maxima), are found by setting its derivative to zero. In this case, differentiating, and solving

$$p'(x) = 2\,a\,x + 2\,b = 0,$$

we conclude that the only possible minimum value occurs at

$$x^\star = -\,\frac{b}{a}\,, \qquad \text{where} \qquad p(x^\star) = c - \frac{b^2}{a}\,. \tag{4.6}$$

Of course, one must check that this critical point is indeed a minimum, and not a maximum or inflection point. The second derivative test will show that $p''(x^\star) = 2\,a > 0$, and so $x^\star$ is at least a local minimum.

A more instructive approach to this problem — and one that only requires elementary algebra — is to "complete the square". As was done in (3.53), we rewrite

$$p(x) = a\left(x + \frac{b}{a}\right)^2 + \frac{a\,c - b^2}{a}\,. \tag{4.7}$$

If $a > 0$, then the first term is always ≥ 0, and moreover equals 0 only at $x^\star = -\,b/a$, reproducing (4.6). The second term is constant, and so unaffected by the value of x. We conclude that $p(x)$ is minimized when the squared term in (4.7) vanishes. Thus, the simple algebraic identity (4.7) immediately proves that the global minimum of p is at $x^\star = -\,b/a$, and, moreover its minimal value $p(x^\star) = (a\,c - b^2)/a$ is the constant term.

Now that we have the scalar case firmly in hand, let us turn to the more difficult problem of minimizing quadratic functions that depend on several variables. Thus, we seek to minimize a (real) *quadratic function*

$$p(\mathbf{x}) = p(x_1, \ldots, x_n) = \sum_{i,j\,=1}^{n} k_{ij}\,x_i\,x_j - 2\sum_{i\,=1}^{n} f_i\,x_i + c, \tag{4.8}$$

depending on n variables $\mathbf{x} = (\,x_1, x_2, \ldots, x_n\,)^T \in \mathbb{R}^n$. The coefficients k_{ij}, f_i and c are all assumed to be real; moreover, according to Exercise ■, we can assume, without loss of generality, that the coefficients of the quadratic terms are symmetric: $k_{ij} = k_{ji}$. Note that $p(\mathbf{x})$ is slightly more general than a quadratic form (3.42) in that it also contains linear and constant terms. We shall rewrite the quadratic function (4.8) in a more convenient matrix notation:

$$p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c, \tag{4.9}$$

where $K = (k_{ij})$ is a symmetric $n \times n$ matrix, $\mathbf{f}$ is a constant vector, and c is a constant scalar. We shall adapt our method of completing the square to find its minimizer.

We first note that in the simple scalar case (4.5), we needed to impose the condition that the quadratic coefficient a is *positive* in order to obtain a (unique) minimum. The corresponding condition for the multivariable case is that the quadratic coefficient matrix K be *positive definite*. This key assumption enables us to prove a very general minimization theorem.

Theorem 4.1. *If $K > 0$ is a symmetric, positive definite matrix, then the quadratic function* (4.9) *has a unique minimizer, which is the solution to the linear system*

$$K\mathbf{x} = \mathbf{f}, \qquad \text{namely} \qquad \mathbf{x}^\star = K^{-1}\mathbf{f}. \tag{4.10}$$

The minimum value of $p(\mathbf{x})$ is equal to any of the following expressions:

$$p(\mathbf{x}^\star) = p(K^{-1}\mathbf{f}) = c - \mathbf{f}^T K^{-1}\mathbf{f} = c - \mathbf{f}^T\mathbf{x}^\star = c - (\mathbf{x}^\star)^T K\,\mathbf{x}^\star. \tag{4.11}$$

Proof: Suppose $\mathbf{x}^\star = K^{-1}\mathbf{f}$ is the (unique — why?) solution to (4.10). Then, for any $\mathbf{x} \in \mathbb{R}^n$, we can write

$$\begin{aligned} p(\mathbf{x}) &= \mathbf{x}^T K\,\mathbf{x} - 2\mathbf{x}^T\mathbf{f} + c = \mathbf{x}^T K\,\mathbf{x} - 2\mathbf{x}^T K\,\mathbf{x}^\star + c \\ &= (\mathbf{x} - \mathbf{x}^\star)^T K(\mathbf{x} - \mathbf{x}^\star) + \big[\, c - (\mathbf{x}^\star)^T K\,\mathbf{x}^\star \,\big], \end{aligned} \tag{4.12}$$

where we used the symmetry of $K = K^T$ to identify $\mathbf{x}^T K\,\mathbf{x}^\star = (\mathbf{x}^\star)^T K\,\mathbf{x}$. The second term in the final formula does not depend on $\mathbf{x}$. Moreover, the first term has the form $\mathbf{y}^T K\,\mathbf{y}$ where $\mathbf{y} = \mathbf{x} - \mathbf{x}^\star$. Since we assumed that K is positive definite, $\mathbf{y}^T K\,\mathbf{y} \geq 0$ and vanishes if and only if $\mathbf{y} = \mathbf{x} - \mathbf{x}^\star = \mathbf{0}$, which achieves its minimum. Therefore, the minimum of $p(\mathbf{x})$ occurs at $\mathbf{x} = \mathbf{x}^\star$. The minimum value of $p(\mathbf{x})$ is equal to the constant term. The alternative expressions in (4.11) follow from simple substitutions. *Q.E.D.*

Example 4.2. Let us illustrate the result with a simple example. Consider the problem of minimizing the quadratic function

$$p(x_1, x_2) = 4x_1^2 - 2x_1x_2 + 3x_2^2 + 3x_1 - 2x_2 + 1$$

over all (real) x_1, x_2. We first write p in the matrix form (4.9), so

$$p(x_1, x_2) = (\,x_1 \quad x_2\,) \begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 2\,(\,x_1 \quad x_2\,) \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix} + 1,$$

whereby

$$K = \begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix}. \tag{4.13}$$

(Pay attention to the overall factor of -2 preceding the linear terms.) According to the theorem, to find the minimum, we must solve the linear system

$$\begin{pmatrix} 4 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -\frac{3}{2} \\ 1 \end{pmatrix}. \tag{4.14}$$

Applying our Gaussian elimination algorithm, only one operation is required to place the coefficient matrix in upper triangular form:

$$\left(\begin{array}{cc|c} 4 & -1 & -\frac{3}{2} \\ -1 & 3 & 1 \end{array} \right) \quad \mapsto \quad \left(\begin{array}{cc|c} 4 & -1 & -\frac{3}{2} \\ 0 & \frac{11}{4} & \frac{5}{8} \end{array} \right).$$

Note that the coefficient matrix is regular (no row interchanges are required) and its two pivots, namely 4, $\frac{11}{4}$, are both positive; this proves that $K > 0$ and hence $p(x_1, x_2)$ really does have a minimum, obtained by applying Back Substitution to the reduced system:

$$\mathbf{x}^\star = \begin{pmatrix} x_1^\star \\ x_2^\star \end{pmatrix} = \begin{pmatrix} -\frac{7}{22} \\ \frac{5}{22} \end{pmatrix} \approx \begin{pmatrix} -.318182 \\ .227273 \end{pmatrix}.$$

The quickest way to compute the minimal value

$$p(\mathbf{x}^\star) = p\left(-\tfrac{7}{22}, \tfrac{5}{22}\right) = \tfrac{13}{44} \approx .295455$$

is to use the second formula in (4.11).

It is instructive to compare the algebraic solution method with the minimization procedure you learned in multi-variable calculus. The *critical points* of $p(x_1, x_2)$ are found by setting both partial derivatives equal to zero:

$$\frac{\partial p}{\partial x_1} = 8\,x_1 - 2\,x_2 + 3 = 0, \qquad \frac{\partial p}{\partial x_2} = -2\,x_1 + 6\,x_2 - 2 = 0.$$

If we divide by an overall factor of 2, these are precisely the *same* linear equations we already constructed in (4.14). Thus, not surprisingly, the calculus approach leads to the same critical point. To check whether a critical point is a local minimum, we need to test the second derivative. In the case of a function of several variables, this requires analyzing the *Hessian matrix*, which is the symmetric matrix of second order partial derivatives

$$H = \begin{pmatrix} \dfrac{\partial^2 p}{\partial x_1^2} & \dfrac{\partial^2 p}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 p}{\partial x_1 \partial x_2} & \dfrac{\partial^2 p}{\partial x_2^2} \end{pmatrix} = \begin{pmatrix} 8 & -2 \\ -2 & 6 \end{pmatrix} = 2\,K,$$

which is exactly twice the quadratic coefficient matrix (4.13). If the Hessian matrix is positive definite — which we already know in this case — then the critical point is indeed

a (local) minimum. Thus, the calculus and algebraic approaches to this minimization problem lead (not surprisingly) to identical results. However, the algebraic method is *more* powerful, because it immedaitely produces the *unique*, *global* minimum, whereas, without extra work (e.g., proving convexity of the function), calculus can only guarantee that the critical point is a local minimum, [**9**]. The reader can find the full story on minimization of nonlinear functions, which is, in fact based on the algebraic theory of positive definite matrices, in Section 19.3.

The most efficient method for producing a minimum of a quadratic function $p(\mathbf{x})$ on $\mathbb{R}^n$, then, is to first write out the symmetric coefficient matrix K and the vector $\mathbf{f}$. Solving the system $K\,\mathbf{x} = \mathbf{f}$ will produce the minimizer $\mathbf{x}^\star$ *provided* $K > 0$ — which should be checked during the course of the procedure by making sure no row interchanges are used and all the pivots are positive. If these conditions are not met then (with one minor exception — see below) one immediately concludes that there is no minimizer.

Example 4.3. Let us minimize the quadratic function

$$p(x, y, z) = x^2 + 2\,x\,y + x\,z + 2\,y^2 + y\,z + 2\,z^2 + 6\,y - 7\,z + 5.$$

This has the matrix form (4.9) with

$$K = \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 1 & 2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 2 \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} 0 \\ -3 \\ \frac{7}{2} \end{pmatrix}, \qquad c = 5.$$

Gaussian elimination produces the $L\,D\,L^T$ factorization

$$K = \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 1 & 2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{3}{4} \end{pmatrix} \begin{pmatrix} 1 & 1 & \frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The pivots, i.e., the diagonal entries of D, are all positive, and hence K is positive definite. Theorem 4.1 then guarantees that $p(x, y, z)$ has a unique minimizer, which is found by solving the linear system $K\,\mathbf{x} = \mathbf{f}$. The solution is then quickly obtained by forward and back substitution:

$$x^\star = 2, \qquad y^\star = -3, \qquad z^\star = 2, \qquad \text{with} \qquad p(x^\star, y^\star, z^\star) = p(2, -3, 2) = -11.$$

Theorem 4.1 solves the general quadratic minimization problem when the quadratic coefficient matrix is positive definite. If K is not positive definite, then the quadratic function (4.9) does not have a minimum, apart from one exceptional situation.

Theorem 4.4. *If $K > 0$ is positive definite, then the quadratic function $p(\mathbf{x}) = \mathbf{x}^T K\,\mathbf{x} - 2\,\mathbf{x}^T \mathbf{f} + c$ has a unique global minimizer $\mathbf{x}^\star$ satisfying $K\,\mathbf{x}^\star = \mathbf{f}$. If $K \geq 0$ is positive semi-definite, and $\mathbf{f} \in \operatorname{rng} K$, then every solution to $K\,\mathbf{x}^\star = \mathbf{f}$ is a global minimum of $p(\mathbf{x})$. However, in the semi-definite case, the minimum is not unique since $p(\mathbf{x}^\star + \mathbf{z}) = p(\mathbf{x}^\star)$ for any null vector $\mathbf{z} \in \ker K$. In all other cases, there is no global minimum, and $p(\mathbf{x})$ can assume arbitrarily large negative values.*

Proof: The first part is just a restatement of Theorem 4.1. The second part is proved by a similar computation, and uses the fact that a positive semi-definite but not definite matrix has a nontrivial kernel. If K is not positive semi-definite, then one can find a vector $\mathbf{y}$ such that $a = \mathbf{y}^T K \mathbf{y} < 0$. If we set $\mathbf{x} = t\mathbf{y}$, then $p(\mathbf{x}) = p(t\mathbf{y}) = at^2 + 2bt + c$, with $b = \mathbf{y}^T \mathbf{f}$. Since $a < 0$, by choosing $|\, t \,| \gg 0$ sufficiently large, one can arrange that $p(t\mathbf{y}) \ll 0$ is an arbitrarily large negative quantity. The one remaining case — when K is positive semi-definite, but $\mathbf{f} \notin \operatorname{rng} K$ — is left until Exercise . *Q.E.D.*

4.3. Least Squares and the Closest Point.

We are now in a position to solve the basic geometric problem of finding the element in a subspace that is closest to a given point in Euclidean space.

Problem: Let V be a subspace of $\mathbb{R}^m$. Given $\mathbf{b} \in \mathbb{R}^m$, find $\mathbf{v}^\star \in V$ which minimizes $\| \mathbf{v} - \mathbf{b} \|$ over all possible $\mathbf{v} \in V$.

The minimal distance $\| \mathbf{v}^\star - \mathbf{b} \|$ to the closest point is called the *distance* from the point $\mathbf{b}$ to the subspace V. Of course, if $\mathbf{b} \in V$ lies in the subspace, then the answer is easy: the closest point is $\mathbf{v}^\star = \mathbf{b}$ itself. The distance from $\mathbf{b}$ to the subspace is zero. Thus, the problem only becomes interesting when $\mathbf{b} \notin V$.

Remark: Initially, you may assume that $\| \cdot \|$ denotes the usual Euclidean norm, and so the distance corresponds to the usual Euclidean length. But it will be no more difficult to solve the closest point problem for *any* norm that arises from an inner product: $\| \mathbf{v} \| = \sqrt{\langle \mathbf{v} \,;\, \mathbf{v} \rangle}$. In fact, requiring that $V \subset \mathbb{R}^m$ is not crucial either; the same method works when V is a finite-dimensional subspace of *any* inner product space.

However, the methods do not apply to more general norms not coming from inner products, e.g., the 1 norm or ∞ norm. These are much harder to handle, and, in such cases, the closest point problem is a *nonlinear* minimization problem whose solution requires the more sophisticated methods of Section 19.3.

When solving the closest point problem, the goal is to minimize the distance

$$\| \mathbf{v} - \mathbf{b} \|^2 = \| \mathbf{v} \|^2 - 2 \langle \mathbf{v} \,;\, \mathbf{b} \rangle + \| \mathbf{b} \|^2, \tag{4.15}$$

over all possible $\mathbf{v}$ belonging to the subspace $V \subset \mathbb{R}^m$. Let us assume that we know a basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ of V, with $n = \dim V$. Then the most general vector in V is a linear combination

$$\mathbf{v} = x_1 \mathbf{v}_1 + \, \cdots \, + x_n \mathbf{v}_n \tag{4.16}$$

of the basis vectors. We substitute the formula (4.16) for $\mathbf{v}$ into the distance function (4.15). As we shall see, the resulting expression is a quadratic function of the coefficients $\mathbf{x} = (\, x_1, x_2, \dots, x_n \,)^T$, and so the minimum is provided by Theorem 4.1.

First, the quadratic terms come from expanding

$$\| \mathbf{v} \|^2 = \langle \mathbf{v} \,;\, \mathbf{v} \rangle = \langle \, x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n \,;\, x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n \, \rangle = \sum_{i,j=1}^{n} x_i x_j \, \langle \, \mathbf{v}_i \,;\, \mathbf{v}_j \, \rangle. \tag{4.17}$$

Therefore,

$$\| \mathbf{v} \|^2 = \sum_{i,j=1}^{n} k_{ij}\, x_i\, x_j = \mathbf{x}^T K \mathbf{x},$$

where K is the symmetric $n \times n$ Gram matrix whose (i,j) entry is the inner product

$$k_{ij} = \langle\, \mathbf{v}_i \,;\, \mathbf{v}_j \,\rangle, \tag{4.18}$$

between the basis vectors of our subspace. Similarly,

$$\langle\, \mathbf{v} \,;\, \mathbf{b} \,\rangle = \langle\, x_1 \mathbf{v}_1 + \, \cdots \, + x_n \mathbf{v}_n \,;\, \mathbf{b} \,\rangle = \sum_{i=1}^{n} x_i \,\langle\, \mathbf{v}_i \,;\, \mathbf{b} \,\rangle,$$

and so

$$\langle\, \mathbf{v} \,;\, \mathbf{b} \,\rangle = \sum_{i=1}^{n} x_i\, f_i = \mathbf{x}^T \mathbf{f},$$

where $\mathbf{f} \in \mathbb{R}^n$ is the vector whose i^{th} entry is the inner product

$$f_i = \langle\, \mathbf{v}_i \,;\, \mathbf{b} \,\rangle \tag{4.19}$$

between the point and the subspace basis elements. We conclude that the squared distance function (4.15) reduces to the quadratic function

$$p(\mathbf{x}) = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c = \sum_{i,j=1}^{n} k_{ij}\, x_i\, x_j - 2 \sum_{i=1}^{n} f_i\, x_i + c, \tag{4.20}$$

in which K and $\mathbf{f}$ are given in (4.18), (4.19), while $c = \| \mathbf{b} \|^2$.

Since we assumed that the basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent, Proposition 3.32 assures us that the associated Gram matrix $K = A^T A$ is positive definite. Therefore, we may directly apply our basic minimization Theorem 4.1 to solve the closest point problem.

Theorem 4.5. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis for the subspace $V \subset \mathbb{R}^m$. Given $\mathbf{b} \in \mathbb{R}^m$, the closest point $\mathbf{v}^\star = x_1^\star \mathbf{v}_1 + \, \cdots \, + x_n^\star \mathbf{v}_n \in V$ is prescribed by the solution $\mathbf{x}^\star = K^{-1}\mathbf{f}$ to the linear system*

$$K\,\mathbf{x} = \mathbf{f}, \tag{4.21}$$

where K and $\mathbf{f}$ are given in (4.18), (4.19). *The distance between the point and the subspace is*

$$\| \mathbf{v}^\star - \mathbf{b} \| = \sqrt{\| \mathbf{b} \|^2 - \mathbf{f}^T \mathbf{x}^\star}\ . \tag{4.22}$$

When using the standard Euclidean inner product and norm on $\mathbb{R}^n$ to measure distance, the entries of the Gram matrix K and the vector $\mathbf{f}$ are given by dot products:

$$k_{ij} = \mathbf{v}_i \cdot \mathbf{v}_j = \mathbf{v}_i^T \mathbf{v}_j, \qquad f_i = \mathbf{v}_i \cdot \mathbf{b} = \mathbf{v}_i^T \mathbf{b}.$$

As in (3.49), both sets of equations can be combined into a single matrix equation. If $A = (\mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_n)$ denotes the $m \times n$ matrix formed by the basis vectors, then

$$K = A^T A, \qquad \mathbf{f} = A^T \mathbf{b}, \qquad c = \| \mathbf{b} \|^2. \tag{4.23}$$

A direct derivation of these equations is instructive. Since, by formula (2.14),

$$\mathbf{v} = x_1 \mathbf{v}_1 + \ \cdots \ + x_n \mathbf{v}_n = A \mathbf{x},$$

we find

$$\begin{aligned} \| \mathbf{v} - \mathbf{b} \|^2 = \| A\mathbf{x} - \mathbf{b} \|^2 &= (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) = (\mathbf{x}^T A^T - \mathbf{b}^T)(A\mathbf{x} - \mathbf{b}) \\ &= \mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b} = \mathbf{x}^T K \mathbf{x} - 2\mathbf{x}^T \mathbf{f} + c, \end{aligned}$$

thereby justifying (4.23). (In the next to last equality, we equate the scalar quantities $\mathbf{b}^T A \mathbf{x} = (\mathbf{b}^T A \mathbf{x})^T = \mathbf{x}^T A^T \mathbf{b}$.)

If, instead of the Euclidean inner product, we adopt an alternative inner product $\langle \mathbf{v} \, ; \mathbf{w} \rangle = \mathbf{v}^T C \mathbf{w}$ prescribed by a positive definite matrix $C > 0$, then the same computations produce

$$K = A^T C A, \qquad \mathbf{f} = A^T C \mathbf{b}, \qquad c = \| \mathbf{b} \|^2. \tag{4.24}$$

The weighted Gram matrix formula was previously derived in (3.51).

Example 4.6. Let $V \subset \mathbb{R}^3$ be the plane spanned by $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 2 \\ -3 \\ -1 \end{pmatrix}$. Our goal is to find the point $\mathbf{v}^\star \in V$ that is closest to $\mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, where distance is measured in the usual Euclidean norm. We combine the basis vectors to form the matrix $A = \begin{pmatrix} 1 & 2 \\ 2 & -3 \\ -1 & -1 \end{pmatrix}$. According to (4.23), the positive definite Gram matrix and associated vector are

$$K = A^T A = \begin{pmatrix} 6 & -3 \\ -3 & 14 \end{pmatrix}, \qquad \mathbf{f} = A^T \mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

(Or, alternatively, these can be computed directly by taking inner products, as in (4.18), (4.19).) We solve the linear system $K\mathbf{x} = \mathbf{f}$ for $\mathbf{x}^\star = K^{-1}\mathbf{f} = \left(\frac{4}{15}, \frac{1}{5} \right)^T$. Theorem 4.5 implies that the closest point is

$$\mathbf{v}^\star = x_1^\star \mathbf{v}_1 + x_2^\star \mathbf{v}_2 = A\mathbf{x}^\star = \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{15} \\ -\frac{7}{15} \end{pmatrix} \approx \begin{pmatrix} .6667 \\ -.0667 \\ -.4667 \end{pmatrix}.$$

The distance from the point $\mathbf{b}$ to the plane is $\| \mathbf{v}^\star - \mathbf{b} \| = \frac{1}{\sqrt{3}} \approx .5774$.

Suppose, on the other hand, that distance is measured in the weighted norm $\| \mathbf{v} \| = v_1^2 + \frac{1}{2} v_2^2 + \frac{1}{3} v_3^2$ corresponding to the diagonal matrix $C = \operatorname{diag}(1, \frac{1}{2}, \frac{1}{3})$. In this case, we form the weighted Gram matrix and vector (4.24):

$$K = A^T C A = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & -3 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} \frac{10}{3} & -\frac{2}{3} \\ -\frac{2}{3} & \frac{53}{6} \end{pmatrix},$$

$$\mathbf{f} = A^T C \mathbf{b} = \begin{pmatrix} 1 & 2 & -1 \\ 2 & -3 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

and so

$$\mathbf{x}^\star = K^{-1} \mathbf{f} \approx \begin{pmatrix} .3506 \\ .2529 \end{pmatrix}, \qquad \mathbf{v}^\star = A \mathbf{x}^\star \approx \begin{pmatrix} .8563 \\ -.0575 \\ -.6034 \end{pmatrix}.$$

In this case, the distance between the point and the subspace is measured in the weighted norm: $\| \mathbf{v}^\star - \mathbf{b} \| \approx .3790$.

Remark: The solution to the closest point problem given in Theorem 4.5 applies, as stated, to the more general case when $V \subset W$ is a finite-dimensional subspace of a general inner product space W. The underlying inner product space W can even be infinite-dimensional, which it is when dealing with least squares approximations in function space, to be described at the end of this chapter, and in Fourier analysis.

Least Squares

As we first observed in Section 4.1, the solution to the closest point problem also solves the basic least squares minimization problem! Let us officially define the notion of a (classical) least squares solution to a linear system.

Definition 4.7. The *least squares solution* to a linear system of equations

$$A \mathbf{x} = \mathbf{b} \tag{4.25}$$

is the vector $\mathbf{x}^\star \in \mathbb{R}^n$ that minimizes the Euclidean norm $\| A \mathbf{x} - \mathbf{b} \|$.

Remark: Later, we will generalize the least squares method to more general weighted norms coming from inner products. However, for the time being we restrict our attention to the Euclidean version.

If the system (4.25) actually has a solution, then it is automatically the least squares solution. Thus, the concept of least squares solution is new only when the system does not have a solution, i.e., $\mathbf{b}$ does not lie in the range of A. We also want the least squares solution to be unique. As with an ordinary solution, this happens if and only if $\ker A = \{\mathbf{0}\}$, or, equivalently, the columns of A are linearly independent, or, equivalently, $\operatorname{rank} A = n$.

As before, to make the connection with the closest point problem, we identify the subspace $V = \operatorname{rng} A \subset \mathbb{R}^m$ as the range or column space of the matrix A. If the columns

of A are linearly independent, then they form a basis for the range V. Since every element of the range can be written as $\mathbf{v} = A\,\mathbf{x}$, minimizing $\| A\,\mathbf{x} - \mathbf{b} \|$ is the same as minimizing the distance $\| \mathbf{v} - \mathbf{b} \|$ between the point and the subspace. The least squares solution $\mathbf{x}^\star$ to the minimization problem gives the closest point $\mathbf{v}^\star = A\,\mathbf{x}^\star$ in $V = \operatorname{rng} A$. Therefore, the least squares solution follows from Theorem 4.5. In the Euclidean case, we state the result more explicitly by using (4.23) to write out the linear system (4.21) and the minimal distance (4.22).

Theorem 4.8. *Assume* $\ker A = \{\mathbf{0}\}$. *Set* $K = A^T A$ *and* $\mathbf{f} = A^T \mathbf{b}$. *Then the least squares solution to* $A\,\mathbf{x} = \mathbf{b}$ *is the unique solution to the* normal equations

$$K\,\mathbf{x} = \mathbf{f} \qquad \text{or} \qquad (A^T A)\,\mathbf{x} = A^T \mathbf{b}, \tag{4.26}$$

namely

$$\mathbf{x}^\star = (A^T A)^{-1} A^T \mathbf{b}. \tag{4.27}$$

The least squares error is

$$\| A\,\mathbf{x}^\star - \mathbf{b} \|^2 = \| \mathbf{b} \|^2 - \mathbf{f}^T \mathbf{x}^\star = \| \mathbf{b} \|^2 - \mathbf{b}^T A\,(A^T A)^{-1} A^T\,\mathbf{b}. \tag{4.28}$$

Note that the normal equations (4.26) can be simply obtained by multiplying the original system $A\,\mathbf{x} = \mathbf{b}$ on both sides by A^T. In particular, if A is square and invertible, then $(A^T A)^{-1} = A^{-1}(A^T)^{-1}$, and so (4.27) reduces to $\mathbf{x} = A^{-1}\mathbf{b}$, while the two terms in the error formula (4.28) cancel out, producing 0 error. In the rectangular case — when this is *not* allowed — formula (4.27) gives a new formula for the solution to (4.25) when $\mathbf{b} \in \operatorname{rng} A$.

Example 4.9. Consider the linear system

$$\begin{aligned} x_1 + 2\,x_2 &= 1, \\ 3\,x_1 - x_2 + x_3 &= 0, \\ -\,x_1 + 2\,x_2 + x_3 &= -1, \\ x_1 - x_2 - 2\,x_3 &= 2, \\ 2\,x_1 + x_2 - x_3 &= 2, \end{aligned}$$

consisting of 5 equations in 3 unknowns. The coefficient matrix and right hand side are

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & -1 & -2 \\ 2 & 1 & -1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \\ -1 \\ 2 \\ 2 \end{pmatrix}.$$

A direct application of Gaussian elimination shows that $\mathbf{b} \notin \operatorname{rng} A$, and so the system is incompatible — it has no solution. Of course, to apply the least squares method, one is not required to check this in advance. If the system has a solution, it is the least squares solution too, and the least squares method will find it.

To form the normal equations (4.26), we compute

$$K = A^T A = \begin{pmatrix} 16 & -2 & -2 \\ -2 & 11 & 2 \\ -2 & 2 & 7 \end{pmatrix}, \qquad \mathbf{f} = A^T \mathbf{b} = \begin{pmatrix} 8 \\ 0 \\ -7 \end{pmatrix}.$$

Solving the 3×3 system $K\,\mathbf{x} = \mathbf{f}$ by Gaussian elimination, we find

$$\mathbf{x} = K^{-1}\mathbf{f} \approx (\, .4119, .2482, -.9532\,)^T,$$

to be the least squares solution to the system. The least squares error is

$$\| \,\mathbf{b} - A\,\mathbf{x}^\star \,\| \;\approx\; \| \,(\, -.0917, .0342, .131, .0701, .0252\,)^T \,\| \;\approx\; .1799,$$

which is reasonably small — indicating that the system is, roughly speaking, not too incompatible.

Remark: If $\ker A \neq \{\mathbf{0}\}$, then the least squares solution to $A\,\mathbf{x} = \mathbf{b}$ is not unique, cf. Exercise ∎. When you ask MATLAB to solve such a linear system (when A is not square) then it gives you the least squares solution that has the minimum Euclidean norm.

4.4. Data Fitting and Interpolation.

One of the most important applications of the least squares minimization process is to the fitting of data points. Suppose we are running an experiment in which we measure a certain time-dependent physical quantity. At time t_i we make the measurement y_i, and thereby obtain a set of, say, m data points

$$(t_1, y_1), \qquad (t_2, y_2), \qquad \dots \qquad (t_m, y_m). \tag{4.29}$$

Suppose our theory indicates that the data points are supposed to all lie on a single line

$$y = \alpha + \beta t, \tag{4.30}$$

whose precise form — meaning its coefficients α, β — is to be determined. For example, a police car is interested in clocking the speed of a vehicle using measurements of its relative distance at several times. Assuming that the vehicle is traveling at constant speed, its position at time t will have the linear form (4.30), with β, the velocity, and α, the initial position, to be determined. Experimental error will almost inevitably make this impossible to achieve exactly, and so the problem is to find the straight line (4.30) which "best fits" the measured data.

The *error* between the measured value y_i and the sample value predicted by the function (4.30) at $t = t_i$ is

$$e_i = y_i - (\alpha + \beta t_i), \qquad i = 1, \dots, m.$$

We can write this system in matrix form as

$$\mathbf{e} = \mathbf{y} - A\,\mathbf{x},$$

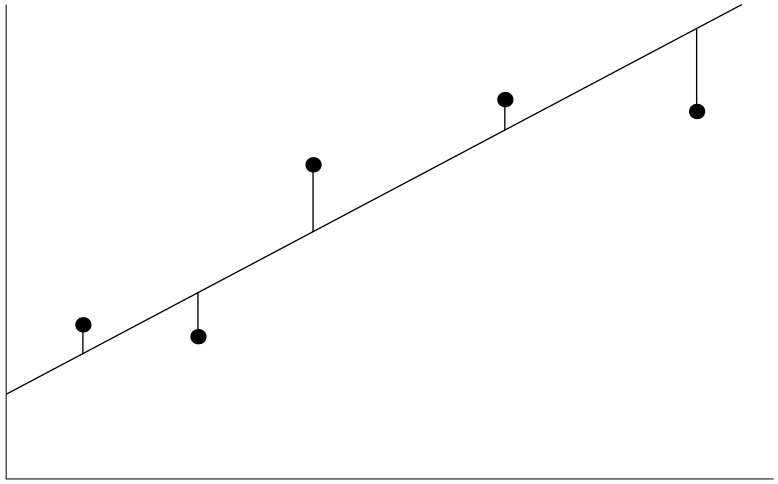


Figure 4.2. Least Squares Approximation of Data by a Straight Line.

where

$$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \qquad \text{while} \qquad A = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \tag{4.31}$$

We call $\mathbf{e}$ the *error vector* and $\mathbf{y}$ the *data vector*. The coefficients α, β of our desired function (4.30) are the unknowns, forming the entries of the column vector $\mathbf{x}$.

If we could fit the data exactly, so $y_i = \alpha + \beta t_i$ for all i, then each $e_i = 0$, and we could solve $A\,\mathbf{x} = \mathbf{y}$. In matrix language, the data points all lie on a straight line if and only if $\mathbf{y} \in \operatorname{rng} A$. If the data points are not all collinear, then we seek the straight line that minimizes the total squared error or Euclidean norm

$$\text{Error} = \| \mathbf{e} \| = \sqrt{e_1^2 + \cdots + e_m^2}\,.$$

Pictorially, referring to Figure 4.2, the errors are the vertical distances from the points to the line, and we are seeking to minimize the square root of the sum of the squares of the individual errors†, hence the term *least squares*. In vector language, we are looking for the coefficient vector $\mathbf{x} = (\,\alpha, \beta\,)^T$ which minimizes the Euclidean norm of the error vector

$$\| \mathbf{e} \| = \| A\,\mathbf{x} - \mathbf{y} \|. \tag{4.32}$$

† This choice of minimization may strike the reader as a little odd. Why not just minimize the sum of the absolute value of the errors, i.e., the 1 norm $\| \mathbf{e} \|_1 = | e_1 | + \cdots + | e_n |$ of the error vector, or minimize the maximal error, i.e., the ∞ norm $\| \mathbf{e} \|_\infty = \max\{ | e_1 |, \cdots, | e_n | \}$? Or, even better, why minimize the vertical distance to the line? Maybe the perpendicular distance from each data point to the line, as computed in Exercise ■, would be a better measure of error. The answer is that, although all of these alternative minimization criteria are interesting and potentially useful, they all lead to *nonlinear* minimization problems, and are much harder to solve! The least squares minimization problem can be solved by linear algebra, whereas the others lead to nonlinear minimization problems. Moreover, one needs to be properly understand the linear solution before moving on to the more treacherous nonlinear situation, cf. Section 19.3.

Thus, we are precisely in the situation of characterizing the least squares solution to the system $A\,\mathbf{x} = \mathbf{y}$ that was covered in the preceding subsection.

Theorem 4.8 prescribes the solution to this least squares minimization problem. We form the normal equations

$$(A^T A)\,\mathbf{x} = A^T \mathbf{y}, \qquad \text{with solution} \qquad \mathbf{x}^\star = (A^T A)^{-1} A^T \mathbf{y}. \tag{4.33}$$

Invertibility of the Gram matrix $K = A^T A$ relies on the assumption that the matrix A have linearly independent columns. This requires that its columns be linearly independent, and so not all the t_i are equal, i.e., we must measure the data at at least two distinct times. Note that this restriction does not preclude measuring some of the data at the same time, e.g., by repeating the experiment. However, choosing *all* the t_i's to be the same is a silly data fitting problem. (Why?)

For the particular matrices (4.31), we compute

$$\begin{aligned} A^T A &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_m \end{pmatrix} \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{pmatrix} = \begin{pmatrix} m & \sum t_i \\ \sum t_i & \sum (t_i)^2 \end{pmatrix} = m \begin{pmatrix} 1 & \overline{t} \\ \overline{t} & \overline{t^2} \end{pmatrix}, \\ A^T \mathbf{y} &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_m \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum t_i\, y_i \end{pmatrix} = m \begin{pmatrix} \overline{y} \\ \overline{t\,y} \end{pmatrix}, \end{aligned} \tag{4.34}$$

where the overbars, namely

$$\overline{t} = \frac{1}{m} \sum_{i=1}^{m} t_i, \qquad \overline{y} = \frac{1}{m} \sum_{i=1}^{m} y_i, \qquad \overline{t^2} = \frac{1}{m} \sum_{i=1}^{m} t_i^2, \qquad \overline{t\,y} = \frac{1}{m} \sum_{i=1}^{m} t_i\, y_i, \tag{4.35}$$

denote the *average* sample values of the indicated variables.

Warning: The average of a product is *not* equal to the product of the averages! In particular,

$$\overline{t^2} \;\neq\; (\,\overline{t}\,)^2, \qquad\qquad \overline{t\,y} \;\neq\; \overline{t}\,\overline{y}.$$

Substituting (4.34) into the normal equations (4.33), and canceling the common factor of m, we find that we have only to solve a pair of linear equations

$$\alpha + \overline{t}\,\beta = \overline{y}, \qquad\qquad \overline{t}\,\alpha + \overline{t^2}\,\beta = \overline{t\,y}.$$

The solution is

$$\alpha = \overline{y} - \overline{t}\,\beta, \qquad\qquad \beta = \frac{\overline{t\,y} - \overline{t}\,\overline{y}}{\overline{t^2} - (\,\overline{t}\,)^2} = \frac{\sum (t_i - \overline{t}\,)\, y_i}{\sum (t_i - \overline{t}\,)^2}\,. \tag{4.36}$$

Therefore, the best (in the least squares sense) straight line that fits the given data is

$$y = \beta\,(t - \overline{t}\,) + \overline{y},$$

where the line's slope β is given in (4.36).

Example 4.10. Suppose the data points are given by the table

t_i	0	1	3	6
y_i	2	3	7	12

Then

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 6 \end{pmatrix}, \qquad A^T = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 6 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 7 \\ 12 \end{pmatrix}.$$

Therefore

$$A^T A = \begin{pmatrix} 4 & 10 \\ 10 & 46 \end{pmatrix}, \qquad A^T \mathbf{y} = \begin{pmatrix} 24 \\ 96 \end{pmatrix}.$$

The normal equations (4.33) reduce to

$$4\alpha + 10\beta = 24, \qquad 10\alpha + 46\beta = 96, \qquad \text{so} \qquad \alpha = \tfrac{12}{7}, \qquad \beta = \tfrac{12}{7}.$$

Therefore, the best least squares fit to the data is the straight line

$$y = \tfrac{12}{7} + \tfrac{12}{7}\, t.$$

Alternatively, one can compute this formula directly from (4.36).

Example 4.11. Suppose we are given a sample of an unknown radioactive isotope. At time t_i we measure, using a Geiger counter, the amount m_i of radioactive material in the sample. The problem is to determine the initial amount of material and the isotope's half life. If the measurements were exact, we would have $m(t) = m_0 e^{\beta t}$, where $m_0 = m(0)$ is the initial mass, and $\beta < 0$ the decay rate. The half life is given by $t^\star = \dfrac{\log 2}{\beta}$; see Example 8.1 for additional information.

As it stands this is not a linear least squares problem, but it can be converted to that form by taking logarithms:

$$y(t) = \log m(t) = \log m_0 + \beta t = \alpha + \beta t.$$

We can thus do a linear least squares fit on the logarithms $y_i = \log m_i$ of the radioactive mass data at the measurement times t_i to determine the best values for β and $\alpha = \log m_0$.

Polynomial Approximation and Interpolation

The basic least squares philosophy has a variety of different extensions, all interesting and all useful. First, we can replace the affine function (4.30) by a quadratic function

$$y = \alpha + \beta t + \gamma t^2, \tag{4.37}$$

In this case, we are looking for the parabola that best fits the data. For example, Newton's theory of gravitation says that (in the absence of air resistance) a falling object obeys the

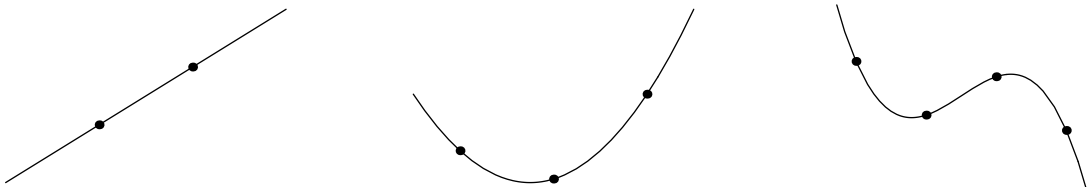


Figure 4.3. Interpolating Polynomials.

parabolic law (4.37), where $\alpha = h_0$ is the initial height, $\beta = v_0$ is the initial velocity, and $\gamma = \frac{1}{2} g m$ is one half the weight of the object. Suppose we observe a falling body, and measure its height y_i at times t_i. Then we can approximate its initial height, initial velocity and weight by finding the parabola (4.37) that best fits the data. Again, we characterize the least squares fit by minimizing the sum of the squares of errors $e_i = y_i - y(t_i)$.

The method can evidently be extended to a completely general polynomial function

$$y(t) = \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n \tag{4.38}$$

of degree n. The total least squares error between the data and the sample values of the function is equal to

$$\| \mathbf{e} \|^2 = \sum_{i=1}^{m} \big[\, y_i - y(t_i) \,\big]^2 = \| \mathbf{y} - A \mathbf{x} \|^2, \tag{4.39}$$

where

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^n \\ 1 & t_2 & t_2^2 & \dots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \dots & t_m^n \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}. \tag{4.40}$$

In particular, if $m = n + 1$, then A is square, and so, assuming A is invertible, we can solve $A \mathbf{x} = \mathbf{y}$ exactly. In other words, there is no error, and the solution is an *interpolating polynomial*, meaning that it fits the data exactly. A proof of the following result can be found in Exercise ∎.

Lemma 4.12. *If $t_1, \dots, t_{n+1}$ are distinct, $t_i \neq t_j$, then the $(n+1) \times (n+1)$ interpolation matrix* (4.40) *is nonsingular.*

This result immediately implies the basic existence theorem for interpolating polynomials.

Theorem 4.13. *Let $t_1, \dots, t_{n+1}$ be distinct sample points. Then, for any prescribed data $y_1, \dots, y_{n+1}$, there exists a unique degree n interpolating polynomial* (4.38) *with sample values $y(t_i) = y_i$ for all $i = 1, \dots, n+1$.*

Thus, two points will determine a unique interpolating line, three points a unique interpolating parabola, four points an interpolating cubic, and so on. Examples are illustrated in Figure 4.3.

Example 4.14. The basic ideas of interpolation and least squares fitting of data can be applied to approximate complicated mathematical functions by much simpler polynomials. Such approximation schemes are used in all numerical computations — when you ask your computer or calculator to compute e^t or $\cos t$ or any other function, it only knows how to add, subtract, multiply and divide, and so must rely on an approximation scheme based on polynomials† In the "dark ages" before computers, one would consult precomputed tables of values of the function at particular data points. If one needed a value at a nontabulated point, then some form of polynomial interpolation would typically be used to accurately approximate the intermediate value.

For example, suppose we want to compute reasonably accurate values for the exponential function e^t for values of t lying in the interval $0 \le t \le 1$ by using a quadratic polynomial

$$p(t) = \alpha + \beta t + \gamma t^2. \tag{4.41}$$

If we choose 3 points, say $t_1 = 0, t_2 = .5, t_3 = 1$, then there is a unique quadratic polynomial (4.41) that interpolates e^t at the data points, i.e.,

$$p(t_i) = e^{t_i} \qquad \text{for} \qquad i = 1, 2, 3.$$

In this case, the coefficient matrix (4.40), namely

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .5 & .25 \\ 1 & 1 & 1 \end{pmatrix},$$

is invertible. Therefore, we can exactly solve the interpolation equations $A\,\mathbf{x} = \mathbf{y}$, where

$$\mathbf{y} = \begin{pmatrix} e^{t_1} \\ e^{t_2} \\ e^{t_3} \end{pmatrix} = \begin{pmatrix} 1 \\ 1.64872 \\ 2.71828 \end{pmatrix}$$

is the data vector. The solution

$$\mathbf{x} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 1. \\ .876603 \\ .841679 \end{pmatrix}$$

yields the interpolating polynomial

$$p(t) = 1 + .876603\,t + .841679\,t^2. \tag{4.42}$$

It is the unique quadratic polynomial that agrees with e^t at the three specified data points. See Figure 4.4 for a comparison of the graphs; the first graph shows e^t, the second $p(t)$, and

† Actually, one could also allow interpolation and approximation by rational functions, a subject known as *Padé approximation theory*. See [**12**] for details.

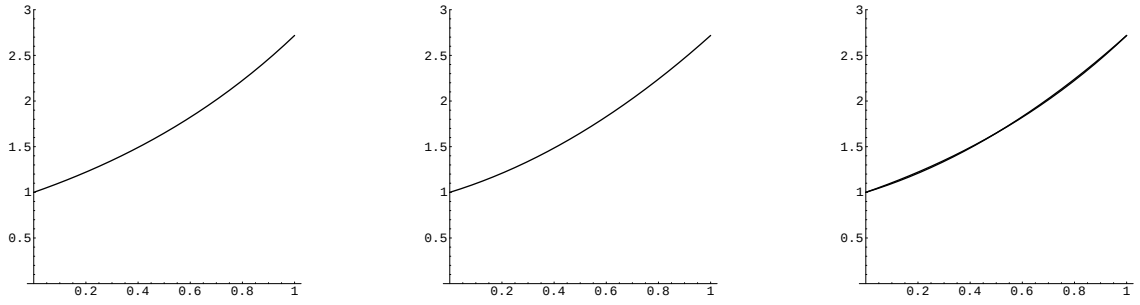


Figure 4.4. Quadratic Interpolating Polynomial for e^t.

the third lays the two graphs on top of each other. Even with such a simple interpolation scheme, the two functions are quite close. The L^∞ norm of the difference is

$$\| e^t - p(t) \|_\infty = \max \big\{ \, | e^t - p(t) | \, \big| \, \ 0 \le t \le 1 \, \big\} \approx .01442,$$

with the maximum error occurring at $t \approx .796$.

There is, in fact, an explicit formula for the interpolating polynomial that is named after the influential eighteenth century Italo–French mathematician Joseph–Louis Lagrange. It relies on the basic superposition principle for solving inhomogeneous systems — Theorem 2.42. Specifically, if we know the solutions $\mathbf{x}_1, \dots, \mathbf{x}_{n+1}$ to the particular interpolation systems

$$A \, \mathbf{x}_k = \mathbf{e}_k, \qquad k = 1, \dots, n+1, \tag{4.43}$$

where $\mathbf{e}_1, \dots, \mathbf{e}_{n+1}$ are the standard basis vectors of $\mathbb{R}^{n+1}$, then the solution to

$$A \, \mathbf{x} = \mathbf{y} = y_1 \, \mathbf{e}_1 + \, \cdots \, + y_{n+1} \, \mathbf{e}_{n+1}$$

is given by the superposition formula

$$\mathbf{x} = y_1 \, \mathbf{x}_1 + \, \cdots \, + y_{n+1} \, \mathbf{x}_{n+1}.$$

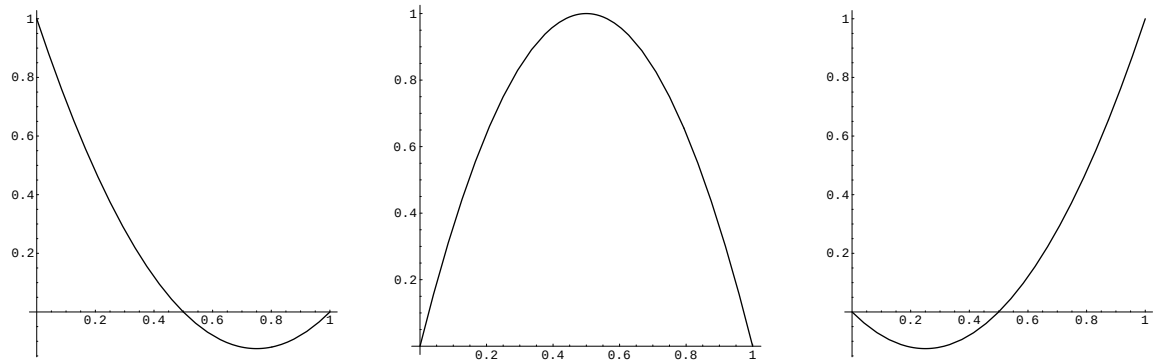
The particular interpolation equation (4.43) corresponds to interpolation data $\mathbf{y} = \mathbf{e}_k$, meaning that $y_k = 1$, while $y_i = 0$ at all points t_i with $i \neq k$. If we can find the $n+1$ particular interpolating polynomials that realize this very special data, we can use superposition to construct the general interpolating polynomial. It turns out that there is a simple explicit formula for the basic interpolating polynomials.

Theorem 4.15. *Given distinct values $t_1, \dots, t_{n+1}$, the k^{th}* Lagrange interpolating polynomial *is the degree n polynomial given by*

$$L_k(t) = \frac{(t - t_1) \, \cdots \, (t - t_{k-1})(t - t_{k+1}) \, \cdots \, (t - t_{n+1})}{(t_k - t_1) \, \cdots \, (t_k - t_{k-1})(t_k - t_{k+1}) \, \cdots \, (t_k - t_{n+1})}, \qquad k = 1, \dots, n+1. \tag{4.44}$$

It is the unique polynomial of degree n that satisfies

$$L_k(t_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases} \qquad i, k = 1, \dots, n+1. \tag{4.45}$$



$L_1(t)$ $L_2(t)$ $L_3(t)$

Figure 4.5. Lagrange Interpolating Polynomials for the Points $0, .5, 1$.

Proof: The uniqueness of the Lagrange interpolating polynomial is an immediate consequence of Theorem 4.13. To show that (4.44) is the correct formula, we note that when $t = t_i$, $i \neq k$, the factor $(t - t_i)$ in the numerator of $L_k(t)$ vanishes, while when $t = t_k$ the numerator and denominator are equal. *Q.E.D.*

Theorem 4.16. *If $t_1, \dots, t_{n+1}$ are distinct, then the degree n polynomial that interpolates the associated data $y_1, \dots, y_{n+1}$ is*

$$p(t) = y_1 L_1(t) + \cdots + y_{n+1} L_{n+1}(t). \tag{4.46}$$

Proof: We merely compute

$$p(t_k) = y_1 L_1(t_k) + \cdots + y_k L_k(t) + \cdots + y_{n+1} L_{n+1}(t_k) = y_k,$$

where, according to (4.45), every summand except the k^{th} is zero. *Q.E.D.*

Example 4.17. For example, the three quadratic Lagrange interpolating polynomials for the values $t_1 = 0, t_2 = \frac{1}{2}, t_3 = 1$ used to interpolate e^t in Example 4.14 are

$$\begin{aligned} L_1(t) &= \frac{(t - \frac{1}{2})(t - 1)}{(0 - \frac{1}{2})(0 - 1)} = 2t^2 - 3t + 1, \\ L_2(t) &= \frac{(t - 0)(t - 1)}{(\frac{1}{2} - 0)(\frac{1}{2} - 1)} = -4t^2 + 4t, \\ L_3(t) &= \frac{(t - 0)(t - \frac{1}{2})}{(1 - 0)(1 - \frac{1}{2})} = 2t^2 - t. \end{aligned} \tag{4.47}$$

Thus, one can rewrite the quadratic interpolant (4.42) to e^t as

$$\begin{aligned} y(t) &= L_1(t) + e^{1/2} L_2(t) + e\, L_3(t) \\ &= (2t^2 - 3t + 1) + 1.64872(-4t^2 + 4t) + 2.71828(2t^2 - t). \end{aligned}$$

We stress that this is the *same* interpolating polynomial — we have merely rewritten it in the more transparent Lagrange form.

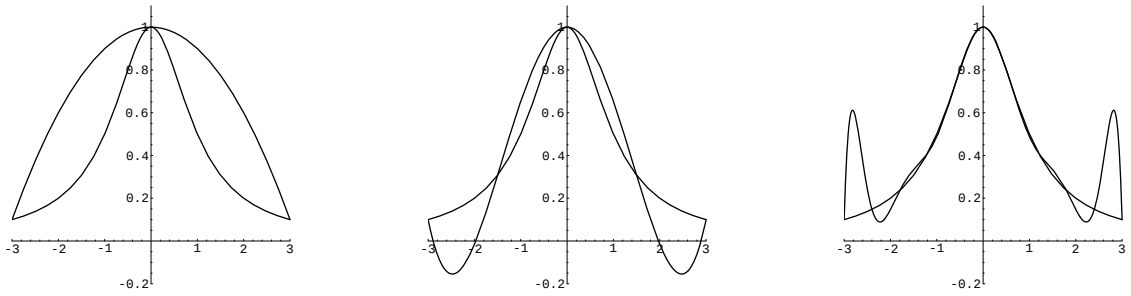


Figure 4.6. Degree 2, 4 and 10 Interpolating Polynomials for $1/(1+t^2)$.

One might expect that the higher the degree, the more accurate the interpolating polynomial. This expectation turns out, unfortunately, not to be uniformly valid. While low degree interpolating polynomials are usually reasonable approximants to functions, high degree interpolants are more expensive to compute, and, moreover, can be rather badly behaved, particularly near the ends of the interval. For example, Figure 4.6 displays the degree $2, 4$ and 10 interpolating polynomials for the function $1/(1+t^2)$ on the interval $-3 \le t \le 3$ using equally spaced data points. Note the rather poor approximation of the function near the endpoints of the interval. Higher degree interpolants fare even worse, although the bad behavior becomes more and more concentrated near the ends of the interval. As a consequence, high degree polynomial interpolation tends not to be used in practical applications. Better alternatives rely on least squares approximants by low degree polynomials, to be described next, and interpolation by piecewise cubic splines, a topic that will be discussed in depth in Chapter 11.

If we have $m > n+1$ data points, then, usually, there is no degree n polynomial that fits all the data, and so one must switch over to a least squares approximation. The first requirement is that the associated $m \times (n+1)$ interpolation matrix (4.40) has rank $n+1$; this follows from Lemma 4.12 provided at least $n+1$ of the values $t_1, \dots, t_m$ are distinct. Thus, given data at $m \ge n+1$ different sample points $t_1, \dots, t_m$, we can uniquely determine the best least squares polynomial of degree n that fits the data by solving the normal equations (4.33).

Example 4.18. If we use more than three data points, but still require a quadratic polynomial, then we cannot interpolate exactly, and must use a least squares approximant. Let us return to the problem of approximating the exponential function e^t. For instance, using five equally spaced sample points $t_1 = 0, t_2 = .25, t_3 = .5, t_4 = .75, t_5 = 1$, the coefficient matrix and sampled data vector (4.40) are

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & .25 & .0625 \\ 1 & .5 & .25 \\ 1 & .75 & .5625 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 1.28403 \\ 1.64872 \\ 2.11700 \\ 2.71828 \end{pmatrix}.$$

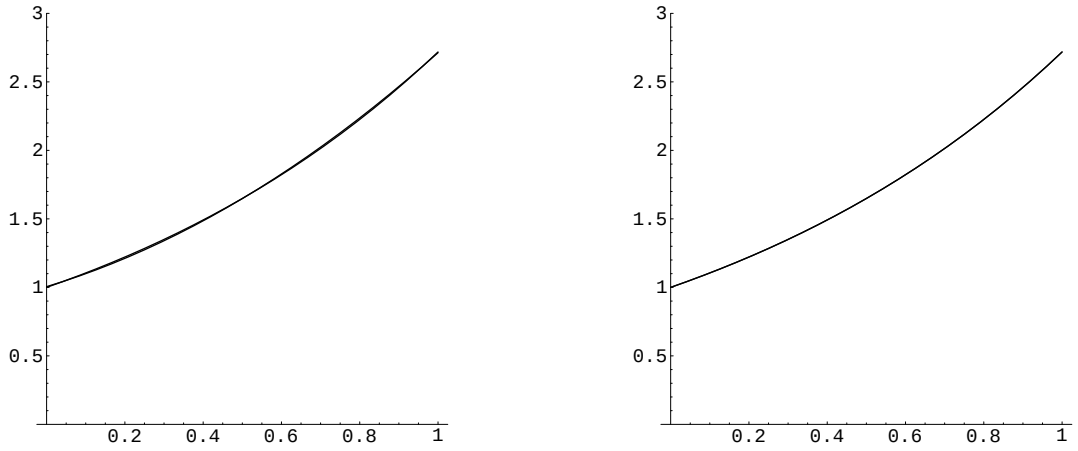


Figure 4.7. Quadratic Approximating Polynomial and Quartic Interpolating Polynomial for e^t.

The solution to the normal equations (4.26), with

$$K = A^T A = \begin{pmatrix} 5. & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.38281 \end{pmatrix}, \qquad \mathbf{f} = A^T \mathbf{y} = \begin{pmatrix} 8.76803 \\ 5.45140 \\ 4.40153 \end{pmatrix},$$

is

$$\mathbf{x} = K^{-1}\mathbf{f} = (\, 1.00514, .864277, .843538 \,)^T .$$

This leads to the modified approximating quadratic polynomial

$$p_2(t) = 1.00514 + .864277\, t + .843538\, t^2.$$

On the other hand, the quartic interpolating polynomial

$$p_4(t) = .069416\, t^4 + .140276\, t^3 + .509787\, t^2 + .998803\, t + 1$$

is found directly from the data values as above. The quadratic polynomial has a maximal error of $\approx .011$ — slightly better than the quadratic interpolant — while the quartic has a significantly smaller maximal error: $\approx .0000527$. (In this case, high degree interpolants are not ill behaved.) See Figure 4.7 for a comparison of the graphs, and Example 4.21 below for further discussion.

Approximation and Interpolation by General Functions

There is nothing special about polynomial functions in the preceding approximation scheme. For example, suppose we were interested in finding the best 2π-periodic trigonometric approximation

$$y = \alpha_1 \cos t + \alpha_2 \sin t$$

to a given set of data. Again, the least squares error takes the same form $\| \mathbf{y} - A\mathbf{x} \|^2$ as in (4.39), where

$$A = \begin{pmatrix} \cos t_1 & \sin t_1 \\ \cos t_2 & \sin t_2 \\ \vdots & \vdots \\ \cos t_m & \sin t_m \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

The key is that the unspecified parameters — in this case α_1, α_2 — occur *linearly* in the approximating function. Thus, the most general case is to approximate the data (4.29) by a linear combination

$$y(t) = \alpha_1 h_1(t) + \alpha_2 h_2(t) + \, \cdots \, + \alpha_n h_n(t),$$

of prescribed, linearly independent functions $h_1(x), \ldots, h_n(x)$. The least squares error is, as always, given by

$$\text{Error} = \sqrt{\sum_{i=1}^{m} \big(y_i - y(t_i) \big)^2} \; = \; \| \mathbf{y} - A\mathbf{x} \|,$$

where the coefficient matrix and vector of unknown coefficients are

$$A = \begin{pmatrix} h_1(t_1) & h_2(t_1) & \ldots & h_n(t_1) \\ h_1(t_2) & h_2(t_2) & \ldots & h_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(t_m) & h_2(t_m) & \ldots & h_n(t_m) \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}. \tag{4.48}$$

Thus, the columns of A are the sampled values of the functions. If A is square and nonsingular, then we can find an interpolating function of the prescribed form by solving the linear system

$$A\mathbf{x} = \mathbf{y}. \tag{4.49}$$

A particularly important case is provided by the $2n + 1$ trigonometric functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos 2x, \qquad \sin 2x, \qquad \ldots \qquad \cos nx, \qquad \sin nx.$$

Interpolation on $2n + 1$ equally spaced data points on the interval $[\, 0, 2\pi \,]$ leads to the discrete Fourier transform, of profound significance in signal processing, data transmission, and compression, [**27**]. Trigonometric interpolation and the discrete Fourier transform will be the focus of Section 13.1.

If there are more than n data points, then we cannot, in general, interpolate exactly, and must content ourselves with a least squares approximation. The least squares solution to the interpolation equations (4.49) is found by solving the associated normal equations $K\mathbf{x} = \mathbf{f}$, where the (i, j) entry of $K = A^T A$ is m times the average value of the product of $h_i(t)$ and $h_j(t)$, namely

$$k_{ij} = m \, \overline{h_i(t) \, h_j(t)} = \sum_{\kappa=1}^{m} h_i(t_\kappa) \, h_j(t_\kappa), \tag{4.50}$$

whereas the i^{th} entry of $\mathbf{f} = A^T\mathbf{y}$ is m times the average

$$f_i = m\,\overline{h_i(t)\,y} = \sum_{\kappa=1}^{m} h_i(t_\kappa)\,y_\kappa. \tag{4.51}$$

The one key question is whether the columns of A are linearly independent; this is more subtle than the polynomial case covered by Lemma 4.12, and requires the sampled function vectors to be linearly independent, which in general is different than requiring the functions themselves to be linearly independent. See Exercise ■ for a few details on the distinction between these two notions of linear independence.

If the parameters do not occur linearly in the functional formula, then we cannot use a linear analysis to find the least squares solution. For example, a direct linear least squares approach does not suffice to find the frequency ω, the amplitude r, *and* the phase δ of a general trigonometric approximation:

$$y = c_1\cos\omega t + c_2\sin\omega t = r\cos(\omega t + \delta).$$

Approximating data by such a function constitutes a *nonlinear* minimization problem, and must be solved by the more sophisticated techniques presented in Section 19.3.

Weighted Least Squares

Another generalization is to introduce weights in the measurement of the least squares error. Suppose some of the data is known to be more reliable or more significant than others. For example, measurements at an earlier time may be more accurate, or more critical to the data fitting problem, than measurements at later time. In that situation, we should penalize any errors at the earlier times and downplay errors in the later data.

In general, this requires the introduction of a positive weight $c_i > 0$ associated to each data point (t_i, y_i); the larger the weight, the more important the error. For a straight line approximation $y = \alpha + \beta t$, the *weighted least squares* error is defined as

$$\text{Error} = \sqrt{\sum_{i=1}^{m} c_i\, e_i^2} \;=\; \sqrt{\sum_{i=1}^{m} c_i\,\big[\,y_i - (\alpha + \beta t_i)\,\big]^2}\;.$$

Let us rewrite this formula in matrix form. Let $C = \operatorname{diag}(c_1, \ldots, c_m)$ denote the diagonal *weight matrix*. Note that $C > 0$ is positive definite, since all the weights are positive. The least squares error

$$\text{Error} = \sqrt{\mathbf{e}^T C\,\mathbf{e}} = \|\,\mathbf{e}\,\|$$

is then the norm of the error vector $\mathbf{e}$ with respect to the weighted inner product

$$\langle\,\mathbf{v}\,;\mathbf{w}\,\rangle = \mathbf{v}^T C\,\mathbf{w} \tag{4.52}$$

induced by the matrix C. Since $\mathbf{e} = \mathbf{y} - A\,\mathbf{x}$,

$$\begin{aligned}\|\,\mathbf{e}\,\|^2 = \|\,A\,\mathbf{x} - \mathbf{y}\,\|^2 &= (A\,\mathbf{x} - \mathbf{y})^T C\,(A\,\mathbf{x} - \mathbf{y})\\ &= \mathbf{x}^T A^T C\,A\,\mathbf{x} - 2\mathbf{x}^T A^T C\,\mathbf{y} + \mathbf{y}^T C\,\mathbf{y} = \mathbf{x}^T K\,\mathbf{x} - 2\mathbf{x}^T\mathbf{f} + c,\end{aligned} \tag{4.53}$$

where

$$K = A^T C\, A, \qquad \mathbf{f} = A^T C\, \mathbf{y}, \qquad c = \mathbf{y}^T C\, \mathbf{y} = \| \mathbf{y} \|^2.$$

Note that K is the Gram matrix derived in (3.51), whose entries

$$k_{ij} = \langle \mathbf{v}_i \,;\, \mathbf{v}_j \rangle = \mathbf{v}_i^T C\, \mathbf{v}_j$$

are the weighted inner products between the column vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ of A. Theorem 3.33 immediately implies that K is positive definite — provided A has linearly independent columns or, equivalently, has rank $n \leq m$.

Theorem 4.19. *Suppose A is an $m \times n$ matrix with linearly independent columns. Suppose $C > 0$ is any positive definite $m \times m$ matrix. Then, the quadratic function* (4.53) *giving the weighted least squares error has a unique minimizer, which is the solution to the* weighted normal equations

$$A^T C\, A\, \mathbf{x} = A^T C\, \mathbf{y}, \qquad \text{so that} \qquad \mathbf{x} = (A^T C\, A)^{-1} A^T C\, \mathbf{y}. \tag{4.54}$$

In other words, the weighted least squares solution is obtained by multiplying both sides of the original system $A\,\mathbf{x} = \mathbf{y}$ by the matrix $A^T C$. The derivation of this result allows $C > 0$ to be *any* positive definite matrix. In applications, the off-diagonal entries of C can be used to weight cross-correlation terms in the data.

Example 4.20. In Example 4.10 we fit the data

t_i	0	1	3	6
y_i	2	3	7	12
c_i	3	2	$\frac{1}{2}$	$\frac{1}{4}$

with an unweighted least squares line. Now we shall assign the weights for the error at each sample point listed in the last row of the table, so that errors in the first two data values carry more weight. To find the weighted least squares line $y = \alpha + \beta t$ that best fits the data, we compute

$$A^T C\, A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 6 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \\ 1 & 6 \end{pmatrix} = \begin{pmatrix} \frac{23}{4} & 5 \\ 5 & \frac{31}{2} \end{pmatrix},$$

$$A^T C\, \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 6 \end{pmatrix} \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 7 \\ 12 \end{pmatrix} = \begin{pmatrix} \frac{37}{2} \\ \frac{69}{2} \end{pmatrix}.$$

Thus, the weighted normal equations (4.54) reduce to

$$\tfrac{23}{4}\,\alpha + 5\,\beta = \tfrac{37}{2}, \qquad 5\,\alpha + \tfrac{31}{2}\,\beta = \tfrac{69}{2}, \qquad \text{so} \qquad \alpha = 1.7817, \qquad \beta = 1.6511.$$

Therefore, the least squares fit to the data under the given weights is $y = 1.7817 + 1.6511\,t$.

Least Squares Approximation in Function Spaces

So far, while we have used least squares minimization to interpolate and approximate known, complicated functions by simpler polynomials, we have only worried about the errors committed at a discrete, preassigned set of sample points. A more uniform approach would be to take into account the errors committed at *all* points in the interval of interest. This can be accomplished by replacing the discrete, finite-dimensional vector space norm on sample vectors by a continuous, infinite-dimensional function space norm in order to specify the least squares error that must be minimized over the entire interval.

More specifically, we let $V = \mathrm{C}^0[a,b]$ denote the space of continuous functions on the bounded interval $[a,b]$ with L^2 inner product

$$\langle\, f\,;g\,\rangle = \int_a^b f(t)\,g(t)\,dt. \tag{4.55}$$

Let $\mathcal{P}^{(n)}$ denote the subspace consisting of all polynomials of degree $\leq n$. For simplicity, we employ the standard monomial basis $1, t, t^2, \ldots, t^n$. We will be approximating a general function $f(t) \in \mathrm{C}^0[a,b]$ by a polynomial

$$p(t) = \alpha_1 + \alpha_2\,t + \,\cdots\, + \alpha_{n+1}\,t^n \in \mathcal{P}^{(n)} \tag{4.56}$$

of degree at most n. The *error function* $e(t) = f(t) - p(t)$ measures the discrepancy between the function and its approximating polynomial at each t. Instead of summing the squares of the errors at a finite set of sample points, we go to a continuous limit that integrates the squared errors of all points in the interval. Thus, the approximating polynomial will be characterized as the one that minimizes the L^2 *least squares error*

$$\text{Error} = \|\,e\,\| = \|\,p - f\,\| = \sqrt{\int_a^b [\,p(t) - f(t)\,]^2\,dt}\;. \tag{4.57}$$

To solve the minimization problem, we begin by substituting (4.56) and expanding, as in (4.17):

$$\|\,p - f\,\|^2 = \left\| \sum_{i=1}^{n+1} \alpha_i\,t^{i-1} - f(t) \right\|^2 = \sum_{i,j=1}^{n+1} \alpha_i\,\alpha_j\,\langle\, t^{i-1}\,;t^j\,\rangle - 2\sum_{i=1}^{n+1} \alpha_i\,\langle\, t^{i-1}\,;f(t)\,\rangle + \|\,f(t)\,\|^2.$$

As a result, we are led to minimize the same kind of quadratic function

$$\mathbf{x}^T K\,\mathbf{x} - 2\mathbf{x}^T\mathbf{f} + c, \tag{4.58}$$

where $\mathbf{x} = \big(\,\alpha_1, \alpha_2, \ldots, \alpha_{n+1}\,\big)^T$ is the vector containing the unknown coefficients in the minimizing polynomial, while

$$k_{ij} = \langle\, t^{i-1}\,;t^{j-1}\,\rangle = \int_a^b t^{i+j-2}\,dt, \qquad f_i = \langle\, t^{i-1}\,;f\,\rangle = \int_a^b t^{i-1} f(t)\,dt, \tag{4.59}$$

are, as before, the Gram matrix K consisting of inner products between basis monomials along with the vector $\mathbf{f}$ of inner products between the monomials and the right hand side. The coefficients of the least squares minimizing polynomial are thus found by solving the associated normal equations $K\,\mathbf{x} = \mathbf{f}$.

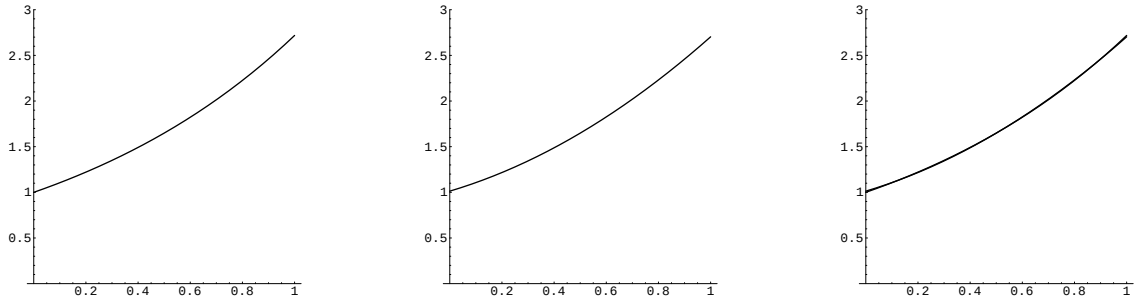


Figure 4.8. Quadratic Least Squares Approximation of e^t.

Example 4.21. Let us return to the problem of approximating the exponential function $f(t) = e^t$ on the interval $0 \le t \le 1$. We consider the subspace $\mathcal{P}^{(2)}$ consisting of all quadratic polynomials

$$p(t) = \alpha + \beta t + \gamma t^2.$$

Using the monomial basis $1, t, t^2$, the normal equations are

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} e-1 \\ 1 \\ e-2 \end{pmatrix}.$$

The coefficient matrix is the Gram matrix K consisting of the inner products

$$\langle\, t^i \,;\, t^j \,\rangle = \int_0^1 t^{i+j}\, dt = \frac{1}{i+j+1}$$

between basis monomials, while the right hand side is the vector of inner products

$$\langle\, e^t \,;\, t^i \,\rangle = \int_0^1 t^i\, e^t\, dt.$$

The solution is computed to be

$$\alpha = 39\, e - 105 \simeq 1.012991, \quad \beta = -216\, e + 588 \simeq .851125, \quad \gamma = 210\, e - 570 \simeq .839184,$$

leading to the least squares quadratic approximant

$$p^\star(t) = 1.012991 + .851125\, t + .839184\, t^2, \tag{4.60}$$

that is plotted in Figure 4.8 The least squares error is

$$\| e^t - p^\star(t) \| \simeq .00527593.$$

The maximal error is measured by the L^∞ norm of the difference,

$$\| e^t - p^\star(t) \|_\infty = \max \left\{ \;| e^t - p^\star(t) | \;\Big|\; 0 \le t \le 1 \right\} \simeq .014981815,$$

with the maximum occurring at $t = 1$. Thus, the simple quadratic polynomial (4.60) will give a reasonable approximation to the first two decimal places in e^t on the entire interval $[0, 1]$. A more accurate approximation can be made by taking a higher degree polynomial, or by decreasing the length of the interval.

Remark: Although the least squares polynomial (4.60) minimizes the L^2 norm of the error, it does slightly worse with the L^∞ norm than the previous sample-based minimizer (4.42). The problem of finding the quadratic polynomial that minimizes the L^∞ norm is more difficult, and must be solved by nonlinear minimization methods.

Remark: As noted in Example 3.35, the Gram matrix for the simple monomial basis is the $n \times n$ Hilbert matrix (1.67). The ill conditioned nature of the Hilbert matrix, and the consequential difficulty in accurately solving the normal equations, complicates the practical numerical implementation of high degree least squares polynomial approximations. A better approach, based on an alternative orthogonal polynomial basis, will be discussed in in the ensuing Chapter.

Chapter 5

Orthogonality

Orthogonality is the mathematical formalization of the geometrical property of perpendicularity — suitably adapted to general inner product spaces. In finite-dimensional spaces, bases that consist of mutually orthogonal elements play an essential role in the theory, in applications, and in practical numerical algorithms. Many computations become dramatically simpler and less prone to numerical instabilities when performed in orthogonal systems. In infinite-dimensional function space, orthogonality unlocks the secrets of Fourier analysis and its manifold applications, and underlies basic series solution methods for partial differential equations. Indeed, many large scale modern applications, including signal processing and computer vision, would be impractical, if not completely infeasible were it not for the dramatic simplifying power of orthogonality. As we will later discover, orthogonal systems naturally arise as eigenvector and eigenfunction bases for symmetric matrices and self-adjoint boundary value problems for both ordinary and partial differential equations, and so play a major role in both finite-dimensional and infinite-dimensional analysis and applications.

Orthogonality is motivated by geometry, and the methods have significant geometrical consequences. Orthogonal matrices play an essential role in the geometry of Euclidean space, computer graphics, animation, and three-dimensional image analysis, to be discussed in Chapter 7. The orthogonal projection of a point onto a subspace is the closest point or least squares minimizer. Moreover, when written in terms of an orthogonal basis for the subspace, the normal equations underlying least squares analysis have an elegant explicit solution formula. Yet another important fact is that the four fundamental subspaces of a matrix form mutually orthogonal pairs. The orthogonality property leads directly to a new characterization of the compatibility conditions for linear systems known as the Fredholm alternative.

The duly famous Gram–Schmidt process will convert an arbitrary basis of an inner product space into an orthogonal basis. As such, it forms one of the key algorithms of linear analysis, in both finite-dimensional vector spaces and also function space where it leads to the classical orthogonal polynomials and other systems of orthogonal functions. In Euclidean space, the Gram–Schmidt process can be re-interpreted as a new kind of matrix factorization, in which a nonsingular matrix $A = QR$ is written as the product of an orthogonal matrix Q and an upper triangular matrix R. The QR factorization underlies one of the primary numerical algorithms for computing eigenvalues, to be presented in Section 10.6.

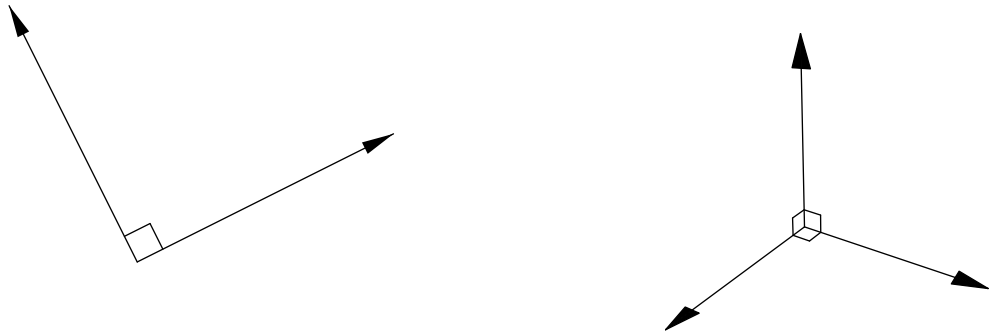


Figure 5.1. Orthogonal Bases in $\mathbb{R}^2$ and $\mathbb{R}^3$.

5.1. Orthogonal Bases.

Let V be a fixed real[†] inner product space. Recall that two elements $\mathbf{v}, \mathbf{w} \in V$ are called *orthogonal* if their inner product vanishes: $\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = 0$. In the case of vectors in Euclidean space, this means that they meet at a right angle, as sketched in Figure 5.1.

A particularly important configuration is when V admits a basis consisting of mutually orthogonal elements.

Definition 5.1. A basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ of V is called *orthogonal* if $\langle\, \mathbf{u}_i\,;\mathbf{u}_j\,\rangle = 0$ for all $i \neq j$. The basis is called *orthonormal* if, in addition, each vector has unit length: $\| \mathbf{u}_i \| = 1$, for all $i = 1, \dots, n$.

For the Euclidean space $\mathbb{R}^n$ equipped with the standard dot product, the simplest example of an orthonormal basis is the standard basis

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \qquad \dots \qquad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Orthogonality follows because $\mathbf{e}_i \cdot \mathbf{e}_j = 0$, for $i \neq j$, while $\| \mathbf{e}_i \| = 1$ implies normality.

Since a basis cannot contain the zero vector, there is an easy way to convert an orthogonal basis to an orthonormal basis. Namely, one replaces each basis vector by a unit vector pointing in the same direction, as in Lemma 3.16.

[†] The methods can be adapted more or less straightforwardly to complex inner product spaces. The main complication, as noted in Section 3.6, is that we need to be careful with the order of vectors appearing in the non-symmetric complex inner products. In this chapter, we will write all inner product formulas in the proper order so that they retain their validity in complex vector spaces.

Lemma 5.2. *If $\mathbf{v}_1, \dots, \mathbf{v}_n$ is any orthogonal basis, then the normalized vectors $\mathbf{u}_i = \mathbf{v}_i / \| \mathbf{v}_i \|$ form an orthonormal basis.*

Example 5.3. The vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 5 \\ -2 \\ 1 \end{pmatrix},$$

are easily seen to form a basis of $\mathbb{R}^3$. Moreover, they are mutually perpendicular, $\mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_1 \cdot \mathbf{v}_3 = \mathbf{v}_2 \cdot \mathbf{v}_3 = 0$, and so form an orthogonal basis with respect to the standard dot product on $\mathbb{R}^3$. When we divide each orthogonal basis vector by its length, the result is the orthonormal basis

$$\mathbf{u}_1 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{30}} \begin{pmatrix} 5 \\ -2 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{5}{\sqrt{30}} \\ -\frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{pmatrix},$$

satisfying $\mathbf{u}_1 \cdot \mathbf{u}_2 = \mathbf{u}_1 \cdot \mathbf{u}_3 = \mathbf{u}_2 \cdot \mathbf{u}_3 = 0$ and $\| \mathbf{u}_1 \| = \| \mathbf{u}_2 \| = \| \mathbf{u}_3 \| = 1$. The appearance of square roots in the elements of an orthonormal basis is fairly typical.

A useful observation is that any orthogonal collection of nonzero vectors is automatically linearly independent.

Proposition 5.4. *If $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ are nonzero, mutually orthogonal, so $\langle \mathbf{v}_i \, ; \mathbf{v}_j \rangle = 0$ for all $i \neq j$, then they are linearly independent.*

Proof: Suppose

$$c_1 \mathbf{v}_1 + \, \cdots \, + c_k \mathbf{v}_k = \mathbf{0}.$$

Let us take the inner product of the equation with any $\mathbf{v}_i$. Using linearity of the inner product and orthogonality of the elements, we compute

$$0 = \langle c_1 \mathbf{v}_1 + \, \cdots \, + c_k \mathbf{v}_k \, ; \mathbf{v}_i \rangle = c_1 \langle \mathbf{v}_1 \, ; \mathbf{v}_i \rangle + \, \cdots \, + c_k \langle \mathbf{v}_k \, ; \mathbf{v}_i \rangle = c_i \langle \mathbf{v}_i \, ; \mathbf{v}_i \rangle = c_i \| \mathbf{v}_i \|^2.$$

Therefore, provided $\mathbf{v}_i \neq \mathbf{0}$, we conclude that the coefficient $c_i = 0$. Since this holds for all $i = 1, \dots, k$, linear independence of $\mathbf{v}_1, \dots, \mathbf{v}_k$ follows. *Q.E.D.*

As a direct corollary, we infer that any orthogonal collection of nonzero vectors is automatically a basis for its span.

Proposition 5.5. *Suppose $\mathbf{v}_1, \dots, \mathbf{v}_n \in V$ are mutually orthogonal nonzero elements of an inner product space V. Then $\mathbf{v}_1, \dots, \mathbf{v}_n$ form an orthogonal basis for their span $W = \operatorname{span} \{ \mathbf{v}_1, \dots, \mathbf{v}_n \} \subset V$, which is therefore a subspace of dimension $n = \dim W$. In particular, if $\dim V = n$, then they form a orthogonal basis for V.*

Orthogonality is also of great significance for function spaces.

Example 5.6. Consider the vector space $\mathcal{P}^{(2)}$ consisting of all quadratic polynomials $p(x) = \alpha + \beta x + \gamma x^2$, equipped with the L^2 inner product and norm

$$\langle p\,; q\rangle = \int_0^1 p(x)\,q(x)\,dx, \qquad \| p \| = \sqrt{\langle p\,; p\rangle} \;=\; \sqrt{\int_0^1 p(x)^2\,dx}\;.$$

The standard monomials $1, x, x^2$ do *not* form an orthogonal basis. Indeed,

$$\langle 1\,; x\rangle = \tfrac{1}{2}\,, \qquad \langle 1\,; x^2\rangle = \tfrac{1}{3}\,, \qquad \langle x\,; x^2\rangle = \tfrac{1}{4}\,.$$

One orthogonal basis of $\mathcal{P}^{(2)}$ is provided by following polynomials:

$$p_1(x) = 1, \qquad p_2(x) = x - \tfrac{1}{2}\,, \qquad p_3(x) = x^2 - x + \tfrac{1}{6}\,. \tag{5.1}$$

Indeed, one easily verifies that $\langle p_1\,; p_2\rangle = \langle p_1\,; p_3\rangle = \langle p_2\,; p_3\rangle = 0$, while

$$\| p_1 \| = 1, \qquad \| p_2 \| = \frac{1}{\sqrt{12}} = \frac{1}{2\sqrt{3}}\,, \qquad \| p_3 \| = \frac{1}{\sqrt{180}} = \frac{1}{6\sqrt{5}}\,. \tag{5.2}$$

The corresponding orthonormal basis is found by dividing each orthogonal basis element by its norm:

$$u_1(x) = 1, \qquad u_2(x) = \sqrt{3}\,(2x - 1)\,, \qquad u_3(x) = \sqrt{5}\,\big(6x^2 - 6x + 1\big)\,. \tag{5.3}$$

In Section 5.4 below, we will learn how to construct such orthogonal systems of polynomials.

Computations in Orthogonal Bases

What are the advantages of orthogonal and orthonormal bases? Once one has a basis of a vector space, a key issue is how to express other elements as linear combinations of the basis elements — that is, to find their *coordinates* in the prescribed basis. In general, this is not an easy problem, since it requires solving a system of linear equations, (2.22). In high dimensional situations arising in applications, computing the solution may require a considerable, if not infeasible amount of time and effort.

However, if the basis is orthogonal, or, even better, orthonormal, then the change of basis computation requires almost no work. This is the crucial insight underlying the efficacy of both discrete and continuous Fourier methods, large data least squares approximations, signal and image processing, and a multitude of other crucial applications.

Theorem 5.7. *Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be an orthonormal basis for an inner product space V. Then one can write any element $\mathbf{v} \in V$ as a linear combination*

$$\mathbf{v} = c_1\,\mathbf{u}_1 + \;\cdots\; + c_n\,\mathbf{u}_n, \tag{5.4}$$

in which the coordinates

$$c_i = \langle\, \mathbf{v}\,; \mathbf{u}_i\,\rangle, \qquad i = 1, \dots, n, \tag{5.5}$$

are explicitly given as inner products. Moreover, the norm

$$\| \mathbf{v} \| = \sqrt{c_1^2 + \;\cdots\; + c_n^2} \;=\; \sqrt{\sum_{i=1}^{n} \langle\, \mathbf{v}\,; \mathbf{u}_i\,\rangle^{\,2}} \tag{5.6}$$

is the square root of the sum of the squares of its coordinates.

Proof: Let us compute the inner product of (5.4) with one of the basis vectors. Using the orthonormality conditions

$$\langle \mathbf{u}_i \,;\, \mathbf{u}_j \rangle = \begin{cases} 0 & i \neq j, \\ 1 & i = j, \end{cases} \tag{5.7}$$

and bilinearity of the inner product, we find

$$\langle \mathbf{v} \,;\, \mathbf{u}_i \rangle = \left\langle \sum_{j=1}^{n} c_j \, \mathbf{u}_j \,;\, \mathbf{u}_i \right\rangle = \sum_{j=1}^{n} c_j \, \langle \mathbf{u}_j \,;\, \mathbf{u}_i \rangle = c_i \, \| \mathbf{u}_i \|^2 = c_i.$$

To prove formula (5.6), we similarly expand

$$\| \mathbf{v} \|^2 = \langle \mathbf{v} \,;\, \mathbf{v} \rangle = \sum_{i,j=1}^{n} c_i \, c_j \, \langle \mathbf{u}_i \,;\, \mathbf{u}_j \rangle = \sum_{i=1}^{n} c_i^2,$$

again making use of the orthonormality of the basis elements. *Q.E.D.*

Example 5.8. Let us rewrite the vector $\mathbf{v} = (\,1, 1, 1\,)^T$ in terms of the orthonormal basis

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} \frac{5}{\sqrt{30}} \\ -\frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{30}} \end{pmatrix},$$

constructed in Example 5.3. Computing the dot products

$$\mathbf{v} \cdot \mathbf{u}_1 = \tfrac{2}{\sqrt{6}}\,, \qquad \mathbf{v} \cdot \mathbf{u}_2 = \tfrac{3}{\sqrt{5}}\,, \qquad \mathbf{v} \cdot \mathbf{u}_3 = \tfrac{4}{\sqrt{30}}\,,$$

we conclude that

$$\mathbf{v} = \tfrac{2}{\sqrt{6}}\, \mathbf{u}_1 + \tfrac{3}{\sqrt{5}}\, \mathbf{u}_2 + \tfrac{4}{\sqrt{30}}\, \mathbf{u}_3,$$

as the reader can validate. Needless to say, a direct computation based on solving the associated linear system, as in Chapter 2, is more tedious.

While passage from an orthogonal basis to its orthonormal version is elementary — one simply divides each basis element by its norm — we shall often find it more convenient to work directly with the unnormalized version. The next result provides the corresponding formula expressing a vector in terms of an orthogonal, but not necessarily orthonormal basis. The proof proceeds exactly as in the orthonormal case, and details are left to the reader.

Theorem 5.9. *If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form an orthogonal basis, then the corresponding coordinates of a vector*

$$\mathbf{v} = a_1 \, \mathbf{v}_1 + \, \cdots \, + a_n \, \mathbf{v}_n \qquad \textit{are given by} \qquad a_i = \frac{\langle \mathbf{v} \,;\, \mathbf{v}_i \rangle}{\| \mathbf{v}_i \|^2}\,. \tag{5.8}$$

In this case, the norm can be computed via the formula

$$\| \mathbf{v} \|^2 = \sum_{i=1}^{n} a_i^2 \, \| \mathbf{v}_i \|^2 = \sum_{i=1}^{n} \left(\frac{\langle \mathbf{v} \, ; \mathbf{v}_i \rangle}{\| \mathbf{v}_i \|} \right)^2 . \tag{5.9}$$

Equation (5.8), along with its orthonormal simplification (5.5), is one of the most important and useful formulas we shall establish. Applications will appear repeatedly throughout the remainder of the text.

Example 5.10. The wavelet basis

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \tag{5.10}$$

introduced in Example 2.33 is, in fact, an orthogonal basis of $\mathbb{R}^4$. The norms are

$$\| \mathbf{v}_1 \| = 2, \qquad \| \mathbf{v}_2 \| = 2, \qquad \| \mathbf{v}_3 \| = \sqrt{2}, \qquad \| \mathbf{v}_4 \| = \sqrt{2}.$$

Therefore, using (5.8), we can readily express any vector as a linear combination of the wavelet basis vectors. For example,

$$\mathbf{v} = \begin{pmatrix} 4 \\ -2 \\ 1 \\ 5 \end{pmatrix} = 2\,\mathbf{v}_1 - \mathbf{v}_2 + 3\,\mathbf{v}_3 - 2\,\mathbf{v}_4,$$

where the wavelet basis coordinates are computed directly by

$$\frac{\langle \mathbf{v} \, ; \mathbf{v}_1 \rangle}{\| \mathbf{v}_1 \|^2} = \frac{8}{4} = 2\,, \quad \frac{\langle \mathbf{v} \, ; \mathbf{v}_2 \rangle}{\| \mathbf{v}_2 \|^2} = \frac{-4}{4} = -1, \quad \frac{\langle \mathbf{v} \, ; \mathbf{v}_3 \rangle}{\| \mathbf{v}_3 \|^2} = \frac{6}{2} = 3 \quad \frac{\langle \mathbf{v} \, ; \mathbf{v}_4 \rangle}{\| \mathbf{v}_4 \|^2} = \frac{-4}{2} = -2\,.$$

This is clearly a lot quicker than solving the linear system, as we did in Example 2.33. Finally, we note that

$$46 = \| \mathbf{v} \|^2 = 2^2 \, \| \mathbf{v}_1 \|^2 + (-1)^2 \, \| \mathbf{v}_2 \|^2 + 3^2 \, \| \mathbf{v}_3 \|^2 + (-2)^2 \, \| \mathbf{v}_4 \|^2 = 4 \cdot 4 + 1 \cdot 4 + 9 \cdot 2 + 4 \cdot 2,$$

in conformity with (5.9).

Example 5.11. The same formulae are equally valid for orthogonal bases in function spaces. For example, to express a quadratic polynomial

$$p(x) = c_1 \, p_1(x) + c_2 \, p_2(x) + c_3 \, p_3(x) = c_1 + c_2 \left(x - \tfrac{1}{2} \right) + c_3 \left(x^2 - x + \tfrac{1}{6} \right)$$

in terms of the orthogonal basis (5.1), we merely compute the inner product integrals

$$c_1 = \frac{\langle p \, ; p_1 \rangle}{\| p_1 \|^2} = \int_0^1 p(x) \, dx, \qquad c_2 = \frac{\langle p \, ; p_2 \rangle}{\| p_2 \|^2} = 12 \int_0^1 p(x) \big(x - \tfrac{1}{2} \big) \, dx,$$

$$c_3 = \frac{\langle p \, ; p_2 \rangle}{\| p_2 \|^2} = 180 \int_0^1 p(x) \big(x^2 - x + \tfrac{1}{6} \big) \, dx.$$

Thus, for example,

$$p(x) = x^2 + x + 1 = \tfrac{11}{6} + 2\left(x - \tfrac{1}{2}\right) + \left(x^2 - x + \tfrac{1}{6}\right),$$

as is easily checked.

Example 5.12. Perhaps the most important example of an orthogonal basis is provided by the basic trigonometric functions. Let $\mathcal{T}^{(n)}$ denote the vector space consisting of all *trigonometric polynomials*

$$T(x) = \sum_{0 \le j+k \le n} a_{jk} (\sin x)^j (\cos x)^k \tag{5.11}$$

of *degree* $\le n$. The constituent monomials $(\sin x)^j (\cos x)^k$ obviously span $\mathcal{T}^{(n)}$, but they do not form a basis owing to identities stemming from the basic trigonometric formula $\cos^2 x + \sin^2 x = 1$; see Example 2.19 for additional details. Exercise ∎ introduced a more convenient spanning set consisting of the $2n + 1$ functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos 2x, \qquad \sin 2x, \qquad \dots \qquad \cos nx, \qquad \sin nx. \tag{5.12}$$

Let us prove that these functions form an orthogonal basis of $\mathcal{T}^{(n)}$ with respect to the L^2 inner product and norm:

$$\langle f\,; g \rangle = \int_{-\pi}^{\pi} f(x)\, g(x)\, dx, \qquad\qquad \| f \|^2 = \int_{-\pi}^{\pi} f(x)^2\, dx. \tag{5.13}$$

The elementary integration formulae

$$\int_{-\pi}^{\pi} \cos kx \, \cos lx \, dx = \begin{cases} 0, & k \ne l, \\ 2\pi, & k = l = 0, \\ \pi, & k = l \ne 0, \end{cases} \qquad \int_{-\pi}^{\pi} \sin kx \, \sin lx \, dx = \begin{cases} 0, & k \ne l, \\ \pi, & k = l \ne 0, \end{cases}$$
$$\int_{-\pi}^{\pi} \cos kx \, \sin lx \, dx = 0, \tag{5.14}$$

which are valid for all nonnegative integers $k, l \ge 0$, imply the orthogonality relations

$$\begin{aligned} &\langle \cos kx\,; \cos lx \rangle = \langle \sin kx\,; \sin lx \rangle = 0, \qquad k \ne l, \qquad \langle \cos kx\,; \sin lx \rangle = 0, \\ &\qquad \| \cos kx \| = \| \sin kx \| = \sqrt{\pi}\,, \qquad k \ne 0, \qquad\qquad \| 1 \| = \sqrt{2\pi}\,. \end{aligned} \tag{5.15}$$

Proposition 5.5 now assures us that the functions (5.12) form a basis for $\mathcal{T}^{(n)}$. One key consequence is that $\dim \mathcal{T}^{(n)} = 2n + 1$ — a fact that is not so easy to establish directly. Orthogonality of the trigonometric functions (5.12) means that we can compute the coefficients $a_0, \dots, a_n, b_1, \dots, b_n$ of any trigonometric polynomial

$$p(x) = a_0 + \sum_{k=1}^{n} \left(a_k \cos kx + b_k \sin kx \right) \tag{5.16}$$

by an explicit integration formula. Namely,

$$a_0 = \frac{\langle\, f\,;1\,\rangle}{\|\,1\,\|^2} = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} f(x)\,dx, \qquad a_k = \frac{\langle\, f\,;\cos k\,x\,\rangle}{\|\cos k\,x\,\|^2} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos k\,x\,dx,$$
$$b_k = \frac{\langle\, f\,;\sin k\,x\,\rangle}{\|\sin k\,x\,\|^2} = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin k\,x\,dx, \qquad k \geq 1. \tag{5.17}$$

These formulae willplay an essential role in the theory and applications of Fourier series; see Chapter 12.

5.2. The Gram–Schmidt Process.

Once one becomes convinced of the utility of orthogonal and orthonormal bases, the natural question follows: How can we construct them? A practical algorithm was first discovered by Laplace in the eighteenth century. Today the algorithm is known as the *Gram–Schmidt process*, after its rediscovery by Jorgen Gram, who we already met in Chapter 3, and Erhard Schmidt, a nineteenth century German mathematician. It forms one of the premier algorithms of applied and computational linear algebra.

Let V denote a finite-dimensional inner product space. (To begin with, the reader can assume V us a subspace of $\mathbb{R}^m$ with the standard Euclidean dot product, although the algorithm will be formulated in complete generality.) We assume that we already know some basis $\mathbf{w}_1, \dots, \mathbf{w}_n$ of V, where $n = \dim V$. Our goal is to use this information to construct an orthogonal basis $\mathbf{v}_1, \dots, \mathbf{v}_n$.

We will construct the orthogonal basis elements one by one. Since initially we are not worrying about normality, there are no conditions on the first orthogonal basis element $\mathbf{v}_1$ and so there is no harm in choosing

$$\mathbf{v}_1 = \mathbf{w}_1.$$

Note that $\mathbf{v}_1 \neq \mathbf{0}$ since $\mathbf{w}_1$ appears in the original basis. The second basis vector must be orthogonal to the first: $\langle\, \mathbf{v}_2\,;\mathbf{v}_1\,\rangle = 0$. Let us try to arrange this by subtracting a suitable multiple of $\mathbf{v}_1$, and set

$$\mathbf{v}_2 = \mathbf{w}_2 - c\,\mathbf{v}_1,$$

where c is a scalar to be determined. The orthogonality condition

$$0 = \langle\, \mathbf{v}_2\,;\mathbf{v}_1\,\rangle = \langle\, \mathbf{w}_2\,;\mathbf{v}_1\,\rangle - c\,\langle\, \mathbf{v}_1\,;\mathbf{v}_1\,\rangle = \langle\, \mathbf{w}_2\,;\mathbf{v}_1\,\rangle - c\,\|\,\mathbf{v}_1\,\|^2$$

requires that $c = \dfrac{\langle\, \mathbf{w}_2\,;\mathbf{v}_1\,\rangle}{\|\,\mathbf{v}_1\,\|^2}$, and therefore

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\langle\, \mathbf{w}_2\,;\mathbf{v}_1\,\rangle}{\|\,\mathbf{v}_1\,\|^2}\,\mathbf{v}_1. \tag{5.18}$$

Linear independence of $\mathbf{v}_1 = \mathbf{w}_1$ and $\mathbf{w}_2$ ensures that $\mathbf{v}_2 \neq \mathbf{0}$. (Check!)

Next, we construct

$$\mathbf{v}_3 = \mathbf{w}_3 - c_1\,\mathbf{v}_1 - c_2\,\mathbf{v}_2$$

by subtracting suitable multiples of the first two orthogonal basis elements from $\mathbf{w}_3$. We want $\mathbf{v}_3$ to be orthogonal to both $\mathbf{v}_1$ and $\mathbf{v}_2$. Since we already arranged that $\langle \mathbf{v}_1 ; \mathbf{v}_2 \rangle = 0$, this requires

$$0 = \langle \mathbf{v}_3 ; \mathbf{v}_1 \rangle = \langle \mathbf{w}_3 ; \mathbf{v}_1 \rangle - c_1 \langle \mathbf{v}_1 ; \mathbf{v}_1 \rangle, \qquad 0 = \langle \mathbf{v}_3 ; \mathbf{v}_2 \rangle = \langle \mathbf{w}_3 ; \mathbf{v}_2 \rangle - c_2 \langle \mathbf{v}_2 ; \mathbf{v}_2 \rangle,$$

and hence

$$c_1 = \frac{\langle \mathbf{w}_3 ; \mathbf{v}_1 \rangle}{\| \mathbf{v}_1 \|^2}, \qquad c_2 = \frac{\langle \mathbf{w}_3 ; \mathbf{v}_2 \rangle}{\| \mathbf{v}_2 \|^2}.$$

Therefore, the next orthogonal basis vector is given by the formula

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\langle \mathbf{w}_3 ; \mathbf{v}_1 \rangle}{\| \mathbf{v}_1 \|^2} \mathbf{v}_1 - \frac{\langle \mathbf{w}_3 ; \mathbf{v}_2 \rangle}{\| \mathbf{v}_2 \|^2} \mathbf{v}_2.$$

Continuing in the same manner, suppose we have already constructed the mutually orthogonal vectors $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$ as linear combinations of $\mathbf{w}_1, \dots, \mathbf{w}_{k-1}$. The next orthogonal basis element $\mathbf{v}_k$ will be obtained from $\mathbf{w}_k$ by subtracting off a suitable linear combination of the previous orthogonal basis elements:

$$\mathbf{v}_k = \mathbf{w}_k - c_1 \mathbf{v}_1 - \cdots - c_{k-1} \mathbf{v}_{k-1}.$$

Since $\mathbf{v}_1, \dots, \mathbf{v}_{k-1}$ are already orthogonal, the orthogonality constraint

$$0 = \langle \mathbf{v}_k ; \mathbf{v}_j \rangle = \langle \mathbf{w}_k ; \mathbf{v}_j \rangle - c_j \langle \mathbf{v}_j ; \mathbf{v}_j \rangle$$

requires

$$c_j = \frac{\langle \mathbf{w}_k ; \mathbf{v}_j \rangle}{\| \mathbf{v}_j \|^2} \qquad \text{for} \qquad j = 1, \dots, k-1. \tag{5.19}$$

In this fashion, we establish the general *Gram–Schmidt formula*

$$\mathbf{v}_k = \mathbf{w}_k - \sum_{j=1}^{k-1} \frac{\langle \mathbf{w}_k ; \mathbf{v}_j \rangle}{\| \mathbf{v}_j \|^2} \mathbf{v}_j, \qquad k = 1, \dots, n. \tag{5.20}$$

The Gram–Schmidt process (5.20) defines a recursive procedure for constructing the orthogonal basis vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. If we are actually after an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_n$, we merely normalize the resulting orthogonal basis vectors, setting $\mathbf{u}_k = \mathbf{v}_k / \| \mathbf{v}_k \|$ for $k = 1, \dots, n$.

Example 5.13. The vectors

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \qquad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \qquad \mathbf{w}_3 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix}, \tag{5.21}$$

are readily seen to form a basis† of $\mathbb{R}^3$. To construct an orthogonal basis (with respect to the standard dot product) using the Gram–Schmidt procedure, we begin by setting $\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$. The next basis vector is

$$\mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2}\, \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} - \frac{-1}{3} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix}.$$

The last orthogonal basis vector is

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2}\, \mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\| \mathbf{v}_2 \|^2}\, \mathbf{v}_2 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} - \frac{-3}{3} \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} - \frac{7}{\frac{14}{3}} \begin{pmatrix} \frac{4}{3} \\ \frac{1}{3} \\ \frac{5}{3} \end{pmatrix} = \begin{pmatrix} 1 \\ -\frac{3}{2} \\ -\frac{1}{2} \end{pmatrix}.$$

The reader can easily validate the orthogonality of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.

An orthonormal basis is obtained by dividing each vector by its length. Since

$$\| \mathbf{v}_1 \| = \sqrt{3}, \qquad \| \mathbf{v}_2 \| = \sqrt{\frac{14}{3}}, \qquad \| \mathbf{v}_3 \| = \sqrt{\frac{7}{2}}.$$

we produce the corresponding orthonormal basis vectors

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} \frac{4}{\sqrt{42}} \\ \frac{1}{\sqrt{42}} \\ \frac{5}{\sqrt{42}} \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} \frac{2}{\sqrt{14}} \\ -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{14}} \end{pmatrix}. \tag{5.22}$$

Example 5.14. Here is a typical sort of problem: Find an orthonormal basis (with respect to the dot product) for the subspace $V \subset \mathbb{R}^4$ consisting of all vectors which are orthogonal to the vector $\mathbf{a} = (1, 2, -1, -3)^T$. Now, a vector $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ is orthogonal to $\mathbf{a}$ if and only if

$$\mathbf{x} \cdot \mathbf{a} = x_1 + 2x_2 - x_3 - 3x_4 = 0.$$

Solving this homogeneous linear system by the usual method, we find that the free variables are x_2, x_3, x_4, and so a (non-orthogonal) basis for the subspace is

$$\mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{w}_3 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

† This will, in fact, be a consequence of the successful completion of the Gram–Schmidt algorithm and does not need to be checked in advance. If the given vectors were not linearly independent, then eventually one of the Gram–Schmidt vectors would vanish, and the process will break down.

To obtain an orthogonal basis, we apply the Gram–Schmidt process. First, $\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix}$. The next element is $\mathbf{v}_2 = \mathbf{w}_2 - \dfrac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2}\, \mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \dfrac{-2}{5} \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ 1 \\ 0 \end{pmatrix}$. The last element of our orthogonal basis is

$$\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2}\, \mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\| \mathbf{v}_2 \|^2}\, \mathbf{v}_2 = \begin{pmatrix} 3 \\ 0 \\ 0 \\ 1 \end{pmatrix} - \frac{-6}{5} \begin{pmatrix} -2 \\ 1 \\ 0 \\ 0 \end{pmatrix} - \frac{\frac{3}{5}}{\frac{6}{5}} \begin{pmatrix} \frac{1}{5} \\ \frac{2}{5} \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \\ -\frac{1}{2} \\ 1 \end{pmatrix}.$$

An orthonormal basis can then be obtained by dividing each $\mathbf{v}_i$ by its length:

$$\mathbf{u}_1 = \begin{pmatrix} -\frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{30}} \\ \frac{5}{\sqrt{30}} \\ 0 \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} \frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \\ -\frac{1}{\sqrt{10}} \\ \frac{2}{\sqrt{10}} \end{pmatrix}. \tag{5.23}$$

The Gram–Schmidt procedure has one final important consequence. By definition, every finite-dimensional vector space admits a basis. Given an inner product, the Gram–Schmidt process enables one to construct an orthogonal and even orthonormal basis of the space. Therefore, we have, in fact, implemented a constructive proof of the existence of orthogonal and orthonormal bases of finite-dimensional inner product spaces. Indeed, the construction shows that there are many different orthogonal and hence orthonormal bases.

Theorem 5.15. *A finite-dimensional inner product space has an orthonormal basis.*

Modifications of the Gram–Schmidt Process

With the basic Gram–Schmidt algorithm now in hand, it is worth looking at a couple of reformulations that have both practical and theoretical uses. The first is an alternative approach that can be used to directly construct the orthonormal basis vectors $\mathbf{u}_1, \dots, \mathbf{u}_n$ from the basis $\mathbf{w}_1, \dots, \mathbf{w}_n$.

We begin by replacing each orthogonal basis vector in the basic Gram–Schmidt formula (5.20) by its normalized version $\mathbf{u}_j = \mathbf{v}_j / \| \mathbf{v}_j \|$. As a result, we find that the original basis vectors can be expressed in terms of the orthonormal basis via a "triangular" system

$$\begin{aligned} \mathbf{w}_1 &= r_{11}\, \mathbf{u}_1, \\ \mathbf{w}_2 &= r_{12}\, \mathbf{u}_1 + r_{22}\, \mathbf{u}_2, \\ \mathbf{w}_3 &= r_{13}\, \mathbf{u}_1 + r_{23}\, \mathbf{u}_2 + r_{33}\, \mathbf{u}_3, \\ \vdots \qquad & \qquad \vdots \qquad \vdots \qquad \ddots \\ \mathbf{w}_n &= r_{1n}\, \mathbf{u}_1 + r_{2n}\, \mathbf{u}_2 + \cdots + r_{nn}\, \mathbf{u}_n. \end{aligned} \tag{5.24}$$

The coefficients r_{ij} can, in fact, be directly computed without using the intermediate derivation. Indeed, taking the inner product of the j^{th} equation with the orthonormal basis vector $\mathbf{u}_j$, we find, in view of the orthonormality constraints (5.7),

$$\langle \mathbf{w}_j \,;\mathbf{u}_i \rangle = \langle r_{1j}\,\mathbf{u}_1 + \,\cdots\, + r_{jj}\,\mathbf{u}_j \,;\mathbf{u}_i \rangle = r_{1j}\,\langle \mathbf{u}_1 \,;\mathbf{u}_i \rangle + \,\cdots\, + r_{jj}\,\langle \mathbf{u}_n \,;\mathbf{u}_i \rangle = r_{ij},$$

and hence

$$r_{ij} = \langle \mathbf{w}_j \,;\mathbf{u}_i \rangle. \tag{5.25}$$

On the other hand, according to (5.6),

$$\| \mathbf{w}_j \|^2 = \| r_{1j}\,\mathbf{u}_1 + \,\cdots\, + r_{jj}\,\mathbf{u}_j \|^2 = r_{1j}^2 + \,\cdots\, + r_{j-1,j}^2 + r_{jj}^2. \tag{5.26}$$

The pair of equations (5.25), (5.26) can be rearranged to devise a recursive procedure to compute the orthonormal basis. At stage j, we assume that we have already constructed $\mathbf{u}_1, \dots, \mathbf{u}_{j-1}$. We then compute[†]

$$r_{ij} = \langle \mathbf{w}_j \,;\mathbf{u}_i \rangle, \qquad \text{for each} \qquad i = 1, \dots, j-1. \tag{5.27}$$

We obtain the next orthonormal basis vector $\mathbf{u}_j$ by the formulae

$$r_{jj} = \sqrt{\| \mathbf{w}_j \|^2 - r_{1j}^2 - \,\cdots\, - r_{j-1,j}^2}\,, \qquad \mathbf{u}_j = \frac{\mathbf{w}_j - r_{1j}\,\mathbf{u}_1 - \,\cdots\, - r_{j-1,j}\,\mathbf{u}_{j-1}}{r_{jj}}\,. \tag{5.28}$$

Running through the formulae (5.27), (5.28) for $j = 1, \dots, n$ leads to the *same* orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ as the previous version of the Gram–Schmidt process.

Example 5.16. Let us apply the revised algorithm to the vectors

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}, \qquad \mathbf{w}_2 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \qquad \mathbf{w}_3 = \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix},$$

of Example 5.13. To begin, we set

$$r_{11} = \| \mathbf{w}_1 \| = \sqrt{3}\,, \qquad \mathbf{u}_1 = \frac{\mathbf{w}_1}{r_{11}} = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{3}} \end{pmatrix}.$$

The next step is to compute

$$r_{12} = \langle \mathbf{w}_2 \,;\mathbf{u}_1 \rangle = -\frac{1}{\sqrt{3}}\,, \quad r_{22} = \sqrt{\| \mathbf{w}_2 \|^2 - r_{12}^2} = \sqrt{\frac{14}{3}}\,, \quad \mathbf{u}_2 = \frac{\mathbf{w}_2 - r_{12}\mathbf{u}_1}{r_{22}} = \begin{pmatrix} \frac{4}{\sqrt{42}} \\ \frac{1}{\sqrt{42}} \\ \frac{5}{\sqrt{42}} \end{pmatrix}.$$

[†] When $j = 1$, there is nothing to do.

The final step yields

$$r_{13} = \langle\, \mathbf{w}_3 \,;\, \mathbf{u}_1 \,\rangle = -\sqrt{3}\,, \qquad r_{23} = \langle\, \mathbf{w}_3 \,;\, \mathbf{u}_2 \,\rangle = \sqrt{\frac{21}{2}}\,,$$

$$r_{33} = \sqrt{\| \mathbf{w}_3 \|^2 - r_{13}^2 - r_{23}^2} = \sqrt{\frac{7}{2}}\,, \qquad \mathbf{u}_3 = \frac{\mathbf{w}_3 - r_{13}\,\mathbf{u}_1 - r_{23}\,\mathbf{u}_2}{r_{33}} = \begin{pmatrix} \frac{2}{\sqrt{14}} \\ -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{14}} \end{pmatrix}.$$

As advertised, the result is the same orthonormal basis vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ found in Example 5.13.

For hand computations, the orthogonal version (5.20) of the Gram–Schmidt process is slightly easier — even if one does ultimately want an orthonormal basis — since it avoids the square roots that are ubiquitous in the orthonormal version (5.27), (5.28). On the other hand, for numerical implementation on a computer, the orthonormal version is a bit faster, as it involves fewer arithmetic operations.

However, in practical, large scale computations, both versions of the Gram–Schmidt process suffer from a serious flaw. They are subject to numerical instabilities, and so round-off errors may seriously corrupt the computations, producing inaccurate, non-orthogonal vectors. Fortunately, there is a simple rearrangement of the calculation that obviates this difficulty and leads to a numerically robust algorithm that is used in practice. The idea is to treat the vectors simultaneously rather than sequentially, making full use of the orthonormal basis vectors as they arise. More specifically, the algorithm begins as before — we take $\mathbf{u}_1 = \mathbf{w}_1/\| \mathbf{w}_1 \|$. We then subtract off the appropriate multiples of $\mathbf{u}_1$ from all of the remaining basis vectors so as to arrange their orthogonality to $\mathbf{u}_1$. This is accomplished by setting

$$\mathbf{w}_k^{(2)} = \mathbf{w}_k - \langle\, \mathbf{w}_k \,;\, \mathbf{u}_1 \,\rangle\, \mathbf{u}_1, \qquad \text{for} \qquad k = 2, \dots, n.$$

The second orthonormal basis vector $\mathbf{u}_2 = \mathbf{w}_2^{(2)}/\| \mathbf{w}_2^{(2)} \|$ is then obtained by normalizing. We next modify the remaining vectors $\mathbf{w}_3^{(2)}, \dots, \mathbf{w}_n^{(2)}$ to produce vectors

$$\mathbf{w}_k^{(3)} = \mathbf{w}_k^{(2)} - \langle\, \mathbf{w}_k^{(2)} \,;\, \mathbf{u}_2 \,\rangle\, \mathbf{u}_2, \qquad k = 3, \dots, n,$$

that are orthogonal to both $\mathbf{u}_1$ and $\mathbf{u}_2$. Then $\mathbf{u}_3 = \mathbf{w}_3^{(3)}/\| \mathbf{w}_3^{(3)} \|$ is taken as the next orthonormal basis element, and so on. The full algorithm starts with the initial basis vectors $\mathbf{w}_j = \mathbf{w}_k^{(1)}$, $k = 1, \dots, n$, and then recursively computes

$$\mathbf{u}_j = \frac{\mathbf{w}_j^{(j)}}{\| \mathbf{w}_j^{(j)} \|}\,, \qquad \mathbf{w}_k^{(j+1)} = \mathbf{w}_k^{(j)} - \langle\, \mathbf{w}_k^{(j)} \,;\, \mathbf{u}_j \,\rangle\, \mathbf{u}_j, \qquad \begin{matrix} j = 1, \dots n, \\ k = j+1, \dots, n. \end{matrix} \tag{5.29}$$

(In the final phase, when $j = n$, the second formula is no longer relevant.) The result is a numerically stable computation of the *same* orthonormal basis vectors $\mathbf{u}_1, \dots, \mathbf{u}_n$.

Example 5.17. Let us apply the stable Gram–Schmidt process (5.29) to the basis vectors

$$\mathbf{w}_1^{(1)} = \mathbf{w}_1 = \begin{pmatrix} 2 \\ 2 \\ -1 \end{pmatrix}, \qquad \mathbf{w}_2^{(1)} = \mathbf{w}_2 = \begin{pmatrix} 0 \\ 4 \\ -1 \end{pmatrix}, \qquad \mathbf{w}_3^{(1)} = \mathbf{w}_3 = \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix}.$$

The first orthonormal basis vector is $\mathbf{u}_1 = \dfrac{\mathbf{w}_1^{(1)}}{\| \mathbf{w}_1^{(1)} \|} = \begin{pmatrix} \frac{2}{3} \\ \frac{2}{3} \\ -\frac{1}{3} \end{pmatrix}$. Next, we compute

$$\mathbf{w}_2^{(2)} = \mathbf{w}_2^{(1)} - \langle\, \mathbf{w}_2^{(1)} \,;\, \mathbf{u}_1 \,\rangle\, \mathbf{u}_1 = \begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix}, \qquad \mathbf{w}_3^{(2)} = \mathbf{w}_3^{(1)} - \langle\, \mathbf{w}_3^{(1)} \,;\, \mathbf{u}_1 \,\rangle\, \mathbf{u}_1 = \begin{pmatrix} -1 \\ 0 \\ -2 \end{pmatrix}.$$

The second orthonormal basis vector is $\mathbf{u}_2 = \dfrac{\mathbf{w}_2^{(2)}}{\| \mathbf{w}_2^{(2)} \|} = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$. Finally,

$$\mathbf{w}_3^{(3)} = \mathbf{w}_3^{(2)} - \langle\, \mathbf{w}_3^{(2)} \,;\, \mathbf{u}_2 \,\rangle\, \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ -2 \end{pmatrix}, \qquad \mathbf{u}_3 = \frac{\mathbf{w}_3^{(3)}}{\| \mathbf{w}_3^{(3)} \|} = \begin{pmatrix} -\frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \\ -\frac{2\sqrt{2}}{3} \end{pmatrix}.$$

The resulting vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ form the desired orthonormal basis.

5.3. Orthogonal Matrices.

Matrices whose columns form an orthonormal basis of $\mathbb{R}^n$ relative to the standard Euclidean dot product have a distinguished role. Such "orthogonal matrices" appear in a wide range of applications in geometry, physics, quantum mechanics, partial differential equations, symmetry theory, and special functions. Rotational motions of bodies in three-dimensional space are described by orthogonal matrices, and hence they lie at the foundations of rigid body mechanics, including satellite and underwater vehicle motions, as well as three-dimensional computer graphics and animation. Furthermore, orthogonal matrices are an essential ingredient in one of the most important methods of numerical linear algebra: the QR algorithm for computing eigenvalues of matrices, to be presented in Section 10.6.

Definition 5.18. A square matrix Q is called an *orthogonal matrix* if it satisfies

$$Q^T Q = \mathrm{I}\,. \tag{5.30}$$

The orthogonality condition implies that one can easily invert an orthogonal matrix:

$$Q^{-1} = Q^T. \tag{5.31}$$

In fact the two conditions are equivalent, and hence a matrix is orthogonal if and only if its inverse is equal to its transpose. The second important characterization of orthogonal matrices relates them directly to orthonormal bases.

Proposition 5.19. *A matrix Q is orthogonal if and only if its columns form an orthonormal basis with respect to the Euclidean dot product on $\mathbb{R}^n$.*

Proof: Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the columns of Q. Then $\mathbf{u}_1^T, \dots, \mathbf{u}_n^T$ are the rows of the transposed matrix Q^T. The $(i,j)^{\text{th}}$ entry of the product $Q^T Q = \mathrm{I}$ is given as the product of the i^{th} row of Q^T times the j^{th} column of Q. Thus, $\mathbf{u}_i \cdot \mathbf{u}_j = \mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$ which are precisely the conditions (5.7) for $\mathbf{u}_1, \dots, \mathbf{u}_n$ to form an orthonormal basis. *Q.E.D.*

Warning: Technically, we should be referring to an "orthonormal" matrix, not an "orthogonal" matrix. But the terminology is so standard throughout mathematics that we have no choice but to adopt it here. There is no commonly accepted term for a matrix whose columns form an orthogonal but not orthonormal basis.

Example 5.20. A 2×2 matrix $Q = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is orthogonal if and only if its columns $\mathbf{u}_1 = \begin{pmatrix} a \\ c \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} b \\ d \end{pmatrix}$, form an orthonormal basis of $\mathbb{R}^2$. Equivalently, the requirement

$$Q^T Q = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & a\,c + b\,d \\ a\,c + b\,d & b^2 + d^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

implies that its entries must satisfy the algebraic equations

$$a^2 + c^2 = 1, \qquad a\,c + b\,d = 0, \qquad b^2 + d^2 = 1.$$

The first and last equations say the points $(a, c)^T$ and $(b, d)^T$ lie on the unit circle in $\mathbb{R}^2$, and so

$$a = \cos\theta, \qquad c = \sin\theta, \qquad b = \cos\psi, \qquad d = \sin\psi,$$

for some choice of angles θ, ψ. The remaining orthogonality condition is

$$0 = a\,c + b\,d = \cos\theta\, \cos\psi + \sin\theta\, \sin\psi = \cos(\theta - \psi).$$

This implies that θ and ψ differ by a right angle: $\psi = \theta \pm \frac{1}{2}\pi$. The $\pm$ sign leads to two cases:

$$b = -\sin\theta, \quad d = \cos\theta, \qquad \text{or} \qquad b = \sin\theta, \quad d = -\cos\theta.$$

As a result, every 2×2 orthogonal matrix has one of two possible forms

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}, \qquad \text{where} \qquad 0 \leq \theta < 2\,\pi. \tag{5.32}$$

The corresponding orthonormal bases are illustrated in Figure 5.2. Note that the former is a right-handed basis which can be obtained from the standard basis $\mathbf{e}_1, \mathbf{e}_2$ by a rotation through angle θ, while the latter has the opposite, reflected orientation.

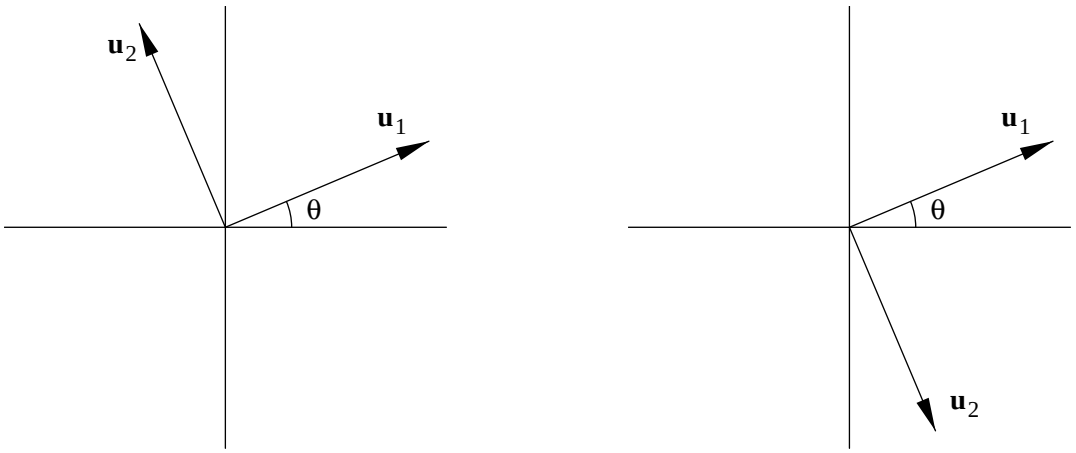


Figure 5.2. Orthonormal Bases in $\mathbb{R}^2$.

Example 5.21. A 3×3 orthogonal matrix $Q = (\,\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3\,)$ is prescribed by 3 mutually perpendicular vectors of unit length in $\mathbb{R}^3$. For instance, the orthonormal basis constructed in (5.22) corresponds to the orthogonal matrix $Q = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{42}} & -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & \frac{5}{\sqrt{42}} & -\frac{1}{\sqrt{14}} \end{pmatrix}$. A complete list of 3×3 orthogonal matrices can be found in Exercises ■ and ■.

Lemma 5.22. *An orthogonal matrix has determinant* $\det Q = \pm 1$.

Proof: Taking the determinant of (5.30) gives

$$1 = \det \mathrm{I} = \det(Q^T Q) = \det Q^T \det Q = (\det Q)^2,$$

which immediately proves the lemma. *Q.E.D.*

An orthogonal matrix is called *proper* if it has determinant $+1$. Geometrically, the columns of a proper orthogonal matrices form a right-handed basis of $\mathbb{R}^n$, as defined in Exercise ■. An *improper* orthogonal matrix, with determinant -1, corresponds to a left-handed basis that lives in a mirror image world.

Proposition 5.23. *The product of two orthogonal matrices is also orthogonal.*

Proof: If $Q_1^T Q_1 = \mathrm{I} = Q_2^T Q_2$, then $(Q_1 Q_2)^T (Q_1 Q_2) = Q_2^T Q_1^T Q_1 Q_2 = Q_2^T Q_2 = \mathrm{I}$, and so $Q_1 Q_2$ is also orthogonal. *Q.E.D.*

This property says that the set of all orthogonal matrices forms a group[†], known as the *orthogonal group*. The orthogonal group lies at the foundation of everyday Euclidean geometry.

[†] The precise mathematical definition of a group can be found in Exercise ■. Although they will not play a significant role in this text, groups are the mathematical formalization of symmetry and, as such, form one of the most fundamental concepts in advanced mathematics and its applications, particularly quantum mechanics and modern theoretical physics. Indeed, according to the mathematician Felix Klein, cf. [**152**], all geometry is based on group theory.

The QR Factorization

The Gram–Schmidt procedure for orthonormalizing bases of $\mathbb{R}^n$ can be reinterpreted as a matrix factorization. This is more subtle than the LU factorization that resulted from Gaussian elimination, but is of comparable importance, and is used in a broad range of applications in mathematics, physics, engineering and numerical analysis.

Let $\mathbf{w}_1, \dots, \mathbf{w}_n$ be a basis of $\mathbb{R}^n$, and let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the corresponding orthonormal basis that results from any one of the three implementations of the Gram–Schmidt process. We assemble both sets of column vectors to form nonsingular $n \times n$ matrices

$$A = (\mathbf{w}_1 \quad \mathbf{w}_2 \quad \dots \quad \mathbf{w}_n), \qquad\qquad Q = (\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_n).$$

Since the $\mathbf{u}_i$ form an orthonormal basis, Q is an orthogonal matrix. In view of the matrix multiplication formula (2.14), the Gram–Schmidt equations (5.24) can be recast into an equivalent matrix form:

$$A = QR, \qquad \text{where} \qquad R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{nn} \end{pmatrix} \tag{5.33}$$

is an upper triangular matrix, whose entries are the previously computed coefficients (5.27), (5.28). Since the Gram–Schmidt process works on any basis, the only requirement on the matrix A is that its columns form a basis of $\mathbb{R}^n$, and hence A can be any nonsingular matrix. We have therefore established the celebrated *QR factorization* of nonsingular matrices.

Theorem 5.24. *Any nonsingular matrix A can be factorized, $A = QR$, into the product of an orthogonal matrix Q and an upper triangular matrix R. The factorization is unique if all the diagonal entries of R are assumed to be positive.*

The proof of uniqueness is left to Exercise ■.

Example 5.25. The columns of the matrix $A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & -2 \\ -1 & 2 & 3 \end{pmatrix}$ are the same as the basis vectors considered in Example 5.16. The orthonormal basis (5.22) constructed using the Gram–Schmidt algorithm leads to the orthogonal and upper triangular matrices

$$Q = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{42}} & -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & \frac{5}{\sqrt{42}} & -\frac{1}{\sqrt{14}} \end{pmatrix}, \qquad R = \begin{pmatrix} \sqrt{3} & -\frac{1}{\sqrt{3}} & -\sqrt{3} \\ 0 & \frac{\sqrt{14}}{\sqrt{3}} & \frac{\sqrt{21}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}.$$

The reader may wish to verify that, indeed, $A = QR$.

While any of the three implementations of the Gram–Schmidt algorithm will produce the QR factorization of a given matrix $A = (\mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n)$, the stable version, as encoded

QR Factorization of a Matrix A

```
start
    for j = 1 to n
        set r_jj = sqrt(a_1j^2 + ... + a_nj^2)
        if r_jj = 0, stop; print "A has linearly dependent columns"
            else for i = 1 to n
                set a_ij = a_ij / r_jj
            next i
        for k = j + 1 to n
            set r_jk = a_1j a_1k + ... + a_nj a_nk
            for i = 1 to n
                set a_ik = a_ik - a_ij r_jk
            next i
        next k
    next j
end
```

in equations (5.29), is the one to use in practical computations, as it is the least likely to fail due to numerical artifacts arising from round-off errors. The accompanying pseudocode program reformulates the algorithm purely in terms of the matrix entries a_{ij} of A. During the course of the algorithm, the entries of the matrix A are successively overwritten; the final result is the orthogonal matrix Q appearing in place of A. The entries r_{ij} of R must be stored separately.

Example 5.26. Let us factorize the matrix

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

using the numerically stable QR algorithm. As in the program, we work directly on the matrix A, gradually changing it into orthogonal form. In the first loop, we set $r_{11} = \sqrt{5}$ to be the norm of the first column vector of A. We then normalize the first column by dividing by r_{11}; the resulting matrix is $\begin{pmatrix} \frac{2}{\sqrt{5}} & 1 & 0 & 0 \\ \frac{1}{\sqrt{5}} & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$. The next entries $r_{12} =$

$\frac{4}{\sqrt{5}}$, $r_{13} = \frac{1}{\sqrt{5}}$, $r_{14} = 0$, are obtained by taking the dot products of the first column with the other three columns. For $j = 1, 2, 3$, we subtract r_{1j} times the first column from the j^{th} column; the result $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{5} & -\frac{2}{5} & 0 \\ \frac{1}{\sqrt{5}} & \frac{6}{5} & \frac{4}{5} & 0 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$ is a matrix whose first column is normalized to have unit length, and whose second, third and fourth columns are orthogonal to it. In the next loop, we normalize the second column by dividing by its norm $r_{22} = \frac{\sqrt{14}}{\sqrt{5}}$, and so obtain the matrix $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & -\frac{2}{5} & 0 \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & \frac{4}{5} & 0 \\ 0 & \frac{5}{\sqrt{70}} & 2 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$. We then take dot products of the second column with the remaining two columns to produce $r_{23} = \frac{16}{\sqrt{70}}$, $r_{24} = \frac{\sqrt{5}}{\sqrt{14}}$. Subtracting these multiples of the second column from the third and fourth columns, we obtain $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{7} & \frac{3}{14} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{7} & -\frac{3}{7} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{7} & \frac{9}{14} \\ 0 & 0 & 1 & 2 \end{pmatrix}$, which now has its first two columns orthonormalized, and orthogonal to the last two columns. We then normalize the third column by dividing by $r_{33} = \frac{\sqrt{15}}{\sqrt{7}}$, and so $\begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{\sqrt{105}} & \frac{3}{14} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{\sqrt{105}} & -\frac{3}{7} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{\sqrt{105}} & \frac{9}{14} \\ 0 & 0 & \frac{7}{\sqrt{105}} & 2 \end{pmatrix}$. Finally, we subtract $r_{34} = \frac{20}{\sqrt{105}}$ times the third column from the fourth column. Dividing the resulting fourth column by its norm $r_{44} = \frac{\sqrt{5}}{\sqrt{6}}$ results in the final formulas

$$Q = \begin{pmatrix} \frac{2}{\sqrt{5}} & -\frac{3}{\sqrt{70}} & \frac{2}{\sqrt{105}} & -\frac{1}{\sqrt{30}} \\ \frac{1}{\sqrt{5}} & \frac{6}{\sqrt{70}} & -\frac{4}{\sqrt{105}} & \frac{2}{\sqrt{30}} \\ 0 & \frac{5}{\sqrt{70}} & \frac{6}{\sqrt{105}} & -\frac{3}{\sqrt{30}} \\ 0 & 0 & \frac{7}{\sqrt{105}} & \frac{4}{\sqrt{30}} \end{pmatrix}, \qquad R = \begin{pmatrix} \sqrt{5} & \frac{4}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ 0 & \frac{\sqrt{14}}{\sqrt{5}} & \frac{16}{\sqrt{70}} & \frac{\sqrt{5}}{\sqrt{14}} \\ 0 & 0 & \frac{\sqrt{15}}{\sqrt{7}} & \frac{20}{\sqrt{105}} \\ 0 & 0 & 0 & \frac{\sqrt{5}}{\sqrt{6}} \end{pmatrix},$$

for the $A = Q\,R$ factorization.

The $Q\,R$ factorization can be used as an alternative to Gaussian elimination to solve linear systems. Indeed, the system

$$A\,\mathbf{x} = \mathbf{b} \qquad \text{becomes} \qquad Q\,R\,\mathbf{x} = \mathbf{b}, \qquad \text{and hence} \qquad R\,\mathbf{x} = Q^T\,\mathbf{b}, \tag{5.34}$$

since $Q^{-1} = Q^T$ is an orthogonal matrix. Since R is upper triangular, the latter system

can be solved for $\mathbf{x}$ by back substitution. The resulting algorithm, while more expensive to compute, does offer some numerical advantages over traditional Gaussian elimination as it is less prone to inaccuracies resulting from ill-conditioned coefficient matrices.

Example 5.27. Let us apply the $A = Q\,R$ factorization

$$\begin{pmatrix} 1 & 1 & 2 \\ 1 & 0 & -2 \\ -1 & 2 & 3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{4}{\sqrt{42}} & \frac{2}{\sqrt{14}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{42}} & -\frac{3}{\sqrt{14}} \\ -\frac{1}{\sqrt{3}} & \frac{5}{\sqrt{42}} & -\frac{1}{\sqrt{14}} \end{pmatrix} \begin{pmatrix} \sqrt{3} & -\frac{1}{\sqrt{3}} & -\sqrt{3} \\ 0 & \frac{\sqrt{14}}{\sqrt{3}} & \frac{\sqrt{21}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix},$$

that we found in Example 5.25 to solve the linear system $A\,\mathbf{x} = (\,0, -4, 5\,)^T$. We first compute

$$Q^T\mathbf{b} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{4}{\sqrt{42}} & \frac{1}{\sqrt{42}} & \frac{5}{\sqrt{42}} \\ \frac{2}{\sqrt{14}} & -\frac{3}{\sqrt{14}} & -\frac{1}{\sqrt{14}} \end{pmatrix} \begin{pmatrix} 0 \\ -4 \\ 5 \end{pmatrix} = \begin{pmatrix} -3\sqrt{3} \\ \frac{\sqrt{21}}{\sqrt{2}} \\ \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}.$$

We then solve the upper triangular system

$$R\mathbf{x} = \begin{pmatrix} \sqrt{3} & -\frac{1}{\sqrt{3}} & -\sqrt{3} \\ 0 & \frac{\sqrt{14}}{\sqrt{3}} & \frac{\sqrt{21}}{\sqrt{2}} \\ 0 & 0 & \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -3\sqrt{3} \\ \frac{\sqrt{21}}{\sqrt{2}} \\ \frac{\sqrt{7}}{\sqrt{2}} \end{pmatrix}$$

by back substitution, leading to the solution $\mathbf{x} = (\,-2, 0, 1\,)^T$.

5.4. Orthogonal Polynomials.

Orthogonal and orthonormal bases play, if anything, an even more essential role in the analysis on function spaces. Unlike the Euclidean space $\mathbb{R}^n$, most obvious bases of a (finite dimensional) function space are typically not orthogonal with respect to any natural inner product. Thus, the computation of an orthonormal basis of functions is a critical step towards simplifying the subsequent analysis. The Gram–Schmidt process can be applied in the same manner, and leads to the classical orthogonal polynomials that arise in approximation and interpolation theory. Other orthogonal systems of functions play starring roles in Fourier analysis and its generalizations, in quantum mechanics, in the solution of partial differential equations by separation of variables, and a host of other applications.

In this section, we concentrate on orthogonal polynomials. Orthogonal systems of trigonometric functions will appear in Chapters 12 and 13. Orthogonal systems of special functions, including Bessel functions and spherical harmonics, are used in the solution to linear partial differential equations in Chapters 17 and 18.

The Legendre Polynomials

We shall construct an orthonormal basis for the vector space $\mathcal{P}^{(n)}$ consisting of all polynomials of degree $\leq n$. For definiteness the construction will be based on the particular

L^2 inner product

$$\langle\, p\,;q\,\rangle = \int_{-1}^{1} p(t)\, q(t)\, dt. \tag{5.35}$$

The method will work for any other bounded interval, but choosing $[\,-1,1\,]$ will lead us to a particularly important case. We shall apply the Gram–Schmidt orthogonalization process to the elementary, but non-orthogonal monomial basis $1,\, t,\, t^2,\, \dots\, t^n$. Because

$$\langle\, t^k\,;t^l\,\rangle = \int_{-1}^{1} t^{k+l}\, dt = \begin{cases} \dfrac{2}{k+l+1}, & k+l \text{ even}, \\ 0, & k+l \text{ odd}, \end{cases} \tag{5.36}$$

odd degree monomials are orthogonal to even degree monomials, but that is all. Let $q_0(t), q_1(t), \dots, q_n(t)$ denote the orthogonal polynomials that result from applying the Gram–Schmidt process to the non-orthogonal monomial basis $1, t, t^2, \dots, t^n$. We begin by setting

$$q_0(t) = 1, \qquad\qquad \| \,q_0\,\|^2 = \int_{-1}^{1} q_0(t)^2\, dt = 2.$$

According to (5.18), the next orthogonal basis polynomial is

$$q_1(t) = t \,-\, \frac{\langle\, t\,;q_0\,\rangle}{\| \,q_0\,\|^2}\, q_0(t) = t, \qquad\qquad \| \,q_1\,\|^2 = \tfrac{2}{3}\,.$$

In general, the Gram–Schmidt formula (5.20) says we should define

$$q_k(t) \;=\; t^k \;-\; \sum_{j\,=\,0}^{k-1} \frac{\langle\, t^k\,;q_j\,\rangle}{\| \,q_j\,\|^2}\, q_j(t) \qquad \text{for} \qquad k = 1, 2, \dots\,.$$

We can then recursively compute the next few polynomials

$$\begin{aligned} q_2(t) &= t^2 - \tfrac{1}{3}, & \| \,q_2\,\|^2 &= \tfrac{8}{45}\,, \\ q_3(t) &= t^3 - \tfrac{3}{5}\,t, & \| \,q_3\,\|^2 &= \tfrac{8}{175}\,, \\ q_4(t) &= t^4 - \tfrac{6}{7}\,t^2 + \tfrac{3}{35}\,, & \| \,q_4\,\|^2 &= \tfrac{128}{11025}\,, \end{aligned} \tag{5.37}$$

and so on. The reader can verify that they satisfy the orthogonality conditions

$$\langle\, q_i\,;q_j\,\rangle = \int_{-1}^{1} q_i(t)\, q_j(t)\, dt = 0, \qquad\qquad i \neq j.$$

The resulting polynomials $q_0, q_1, q_2, \dots$ are known as the *monic*[†] *Legendre polynomials*, in honor of the 18^{th} century French mathematician Adrien–Marie Legendre who used them to study Newtonian gravitation. Since the first n of them, namely $q_0, \dots, q_{n-1}$ span the

[†] A polynomial is called *monic* if its leading coefficient is equal to 1.

subspace $\mathcal{P}^{(n-1)}$ of polynomials of degree $\leq n-1$, the next one, q_n, is the unique monic polynomial that is orthogonal to every polynomial of degree $\leq n-1$:

$$\langle\, t^k \,;\, q_n \,\rangle = 0, \qquad\qquad k = 0, \ldots, n-1. \tag{5.38}$$

Since the monic Legendre polynomials form a basis for the space of polynomials, one can uniquely rewrite any polynomial of degree n as a linear combination:

$$p(t) = c_0\, q_0(t) + c_1\, q_1(t) + \,\cdots\, + c_n\, q_n(t). \tag{5.39}$$

In view of the general orthogonality formula (5.8), the coefficients are simply given by inner products

$$c_k = \frac{\langle\, p \,;\, q_k \,\rangle}{\| \, q_k \, \|^2} = \frac{1}{\| \, q_k \, \|^2} \int_{-1}^{1} p(t)\, q_k(t)\, dt, \qquad\qquad k = 0, \ldots, n. \tag{5.40}$$

For example,

$$t^4 = q_4(t) + \tfrac{6}{7}\, q_2(t) + \tfrac{1}{5}\, q_0(t) = (t^4 - \tfrac{6}{7}\, t^2 + \tfrac{3}{35}) + \tfrac{6}{7}\, (t^2 - \tfrac{1}{3}) + \tfrac{1}{5}.$$

The coefficients can either be obtained directly, or via (5.40); for example,

$$c_4 = \frac{11025}{128} \int_{-1}^{1} t^4\, q_4(t)\, dt = 1, \qquad\qquad c_3 = \frac{175}{8} \int_{-1}^{1} t^4\, q_3(t)\, dt = 0.$$

The classical *Legendre polynomials* are certain scalar multiples, namely

$$P_k(t) = \frac{(2\,k)!}{2^k\, (k!)^2}\, q_k(t), \qquad\qquad k = 0, 1, 2, \ldots, \tag{5.41}$$

of the orthogonal basis polynomials. The multiple is fixed by the requirement that

$$P_k(1) = 1, \tag{5.42}$$

which is not so important here, but does play a role in other applications. The first few classical Legendre polynomials are

$$\begin{aligned}
&P_0(t) = 1, && \| \, P_0 \, \|^2 = 2,\\
&P_1(t) = t, && \| \, P_1 \, \|^2 = \tfrac{2}{3}\,,\\
&P_2(t) = \tfrac{3}{2}\, t^2 - \tfrac{1}{2}\,, && \| \, P_2 \, \|^2 = \tfrac{2}{5},\\
&P_3(t) = \tfrac{5}{2}\, t^3 - \tfrac{3}{2}\, t, && \| \, P_3 \, \|^2 = \tfrac{2}{7},\\
&P_4(t) = \tfrac{35}{8}\, t^4 - \tfrac{15}{4}\, t^2 + \tfrac{3}{8}\,, && \| \, P_4 \, \|^2 = \tfrac{2}{9},\\
&P_5(t) = \tfrac{63}{8}\, t^5 - \tfrac{35}{4}\, t^3 + \tfrac{15}{8}\, t. && \| \, P_5 \, \|^2 = \tfrac{2}{11},\\
&P_6(t) = \tfrac{231}{16}\, t^6 - \tfrac{315}{16}\, t^4 + \tfrac{105}{16}\, t^2 - \tfrac{5}{16}\,, && \| \, P_6 \, \|^2 = \tfrac{2}{13}\,,
\end{aligned}$$

and are graphed in Figure 5.3. There is, in fact, an explicit formula for the Legendre polynomials, due to the early nineteenth century Portuguese mathematician Olinde Rodrigues.

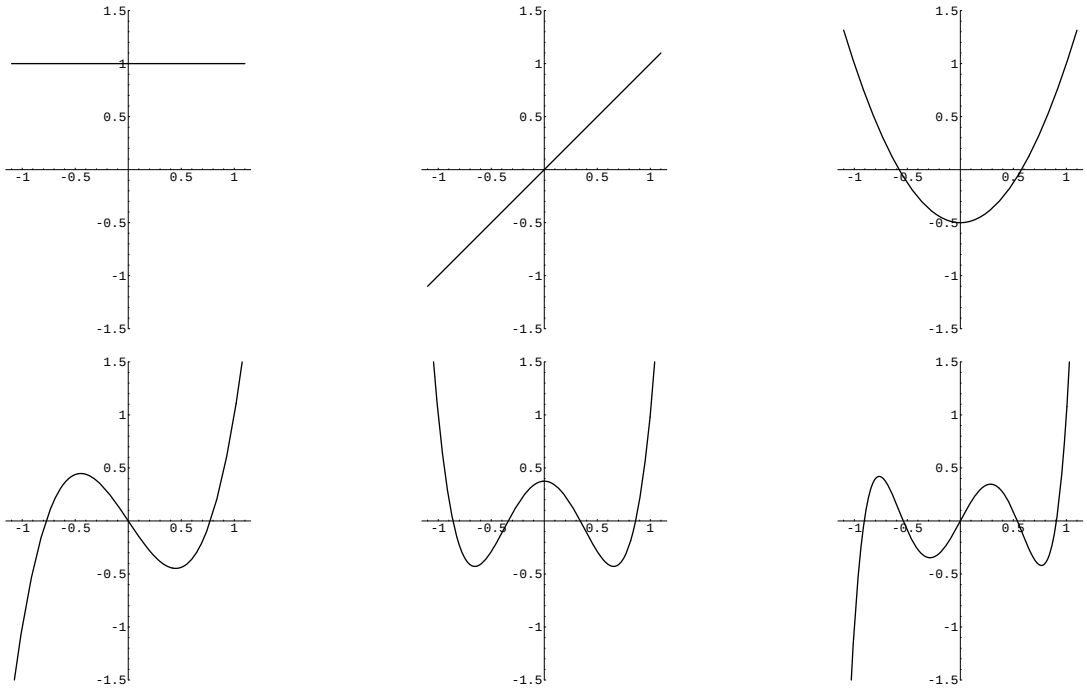


Figure 5.3. The Legendre Polynomials $P_0(t), \dots, P_5(t)$.

Theorem 5.28. *The* Rodrigues formula *for the classical Legendre polynomials is*

$$P_k(t) = \frac{1}{2^k\, k!}\, \frac{d^k}{dt^k}\, (t^2-1)^k, \qquad \| P_k \| = \sqrt{\frac{2}{2\,k+1}}\,, \qquad k = 0, 1, 2, \dots\,. \tag{5.43}$$

Thus, for example,

$$P_4(t) = \frac{1}{16 \cdot 4!}\, \frac{d^4}{dt^4}\, (t^2-1)^4 = \frac{1}{384}\, \frac{d^4}{dt^4}\, (t^2-1)^4 = \tfrac{35}{8}\, t^4 - \tfrac{15}{4}\, t^2 + \tfrac{3}{8}\,.$$

Proof: Let

$$R_{j,k}(t) = \frac{d^j}{dt^j}\, (t^2-1)^k, \tag{5.44}$$

which is evidently a polynomial of degree $2\,k-j$. In particular, the Rodrigues formula (5.43) claims that $P_k(t)$ is a multiple of $R_{k,k}(t)$. Note that

$$\frac{d}{dt}\, R_{j,k}(t) = R_{j+1,k}(t). \tag{5.45}$$

Moreover,

$$R_{j,k}(1) = 0 = R_{j,k}(-1) \qquad \text{whenever} \qquad j < k, \tag{5.46}$$

since, by the product rule, differentiating $(t^2-1)^k$ a total of $j < k$ times still leaves at least one factor of t^2-1 in each summand, which therefore vanishes at $t = \pm 1$.

Lemma 5.29. *If $j \le k$, then the polynomial $R_{j,k}(t)$ is orthogonal to all polynomials of degree $\le j-1$.*

Proof: In other words,

$$\langle\, t^i\,;R_{j,k}\,\rangle = \int_{-1}^{1} t^i\, R_{j,k}(t)\, dt = 0, \qquad \text{for all} \qquad 0 \le i < j \le k. \tag{5.47}$$

Since $j > 0$, we use (5.45) to write $R_{j,k}(t) = R'_{j-1,k}(t)$. Integrating by parts,

$$\begin{aligned}\langle\, t^i\,;R_{j,k}\,\rangle &= \int_{-1}^{1} t^i\, R'_{j-1,k}(t)\, dt \\ &= i t^i\, R_{j-1,k}(t)\,\Big|_{t=-1}^{1} - i \int_{-1}^{1} t^{i-1}\, R_{j-1,k}(t)\, dt = -i\,\langle\, t^{i-1}\,;R_{j-1,k}\,\rangle,\end{aligned}$$

where the boundary terms vanish owing to (5.46). We then repeat the process, and eventually

$$\begin{aligned}\langle\, t^i\,;R_{j,k}\,\rangle &= -i\,\langle\, t^{i-1}\,;R_{j-1,k}\,\rangle \\ &= i(i-1)\,\langle\, t^{i-2}\,;R_{j-2,k}\,\rangle = \,\cdots\, = (-1)^i\, i\,(i-1)\,\cdots\, 3\cdot 2\,\langle\, 1\,;R_{j-i,k}\,\rangle \\ &= (-1)^i\, i!\int_{-1}^{1} R_{j-i,k}(t)\, dt = (-1)^i\, i!\, R_{j-i-1,k}(t)\,\Big|_{t=-1}^{1} = 0,\end{aligned}$$

by (5.46), and since $j > i$. *Q.E.D.*

In particular, $R_{k,k}(t)$ is a polynomial of degree k which is orthogonal to every polynomial of degree $\le k-1$. By our earlier remarks, this implies that it is a constant multiple,

$$R_{k,k}(t) = c_k\, P_k(t)$$

of the k^{th} Legendre polynomial. To determine c_k, we need only compare the leading terms:

$$R_{k,k}(t) = \frac{d^k}{dt^k}\,(t^2-1)^k = \frac{d^k}{dt^k}\,(t^{2\,k} + \cdots) = \frac{(2\,k)!}{(k!)^2}\,t^k + \cdots, \quad \text{while} \quad P_k(t) = \frac{(2\,k)!}{2^k\, k!}\,t^{2\,k} + \cdots.$$

We conclude that $c_k = 2^k\, k!$, which proves (5.43). The proof of the formula for $\| P_k \|$ can be found in Exercise ∎. *Q.E.D.*

The Legendre polynomials play an important role in many aspects of applied mathematics, including numerical analysis, least squares approximation of functions, and solution of partial differential equations.

Other Systems of Orthogonal Polynomials

The standard Legendre polynomials form an orthogonal system with respect to the L^2 inner product on the interval $[-1, 1]$. Dealing with any other interval, or, more generally, a weighted inner product between functions on an interval, leads to a different, suitably adapted collection of orthogonal polynomials. In all cases, applying the Gram–Schmidt process to the standard monomials $1, t, t^2, t^3, \ldots$ will produce the desired orthogonal system.

Example 5.30. In this example, we construct orthogonal polynomials for the weighted inner product

$$\langle\, f\,;g\,\rangle = \int_0^\infty f(t)\, g(t)\, e^{-\,t}\, dt \tag{5.48}$$

on the interval $[\,0, \infty\,)$. A straightforward integration by parts proves that

$$\int_0^\infty t^k\, e^{-\,t}\, dt = k!, \qquad \text{and hence} \qquad \langle\, t^i\,;t^j\,\rangle = (i+j)! \qquad \|\, t^i\,\|^2 = (2\,i)! \tag{5.49}$$

We apply the Gram–Schmidt process to construct a system of orthogonal polynomials for this inner product. The first few are

$$\begin{aligned}
&q_0(t) = 1, && \|\, q_0\,\|^2 = 1,\\
&q_1(t) = t \,-\, \frac{\langle\, t\,;q_0\,\rangle}{\|\, q_0\,\|^2}\, q_0(t) = t - 1, && \|\, q_1\,\|^2 = 1,\\
&q_2(t) = t^2 \,-\, \frac{\langle\, t^2\,;q_0\,\rangle}{\|\, q_0\,\|^2}\, q_0(t) \,-\, \frac{\langle\, t^2\,;q_1\,\rangle}{\|\, q_1\,\|^2}\, q_1(t) = t^2 - 4\,t + 2, && \|\, q_2\,\|^2 = 4,\\
&q_3(t) = t^3 - 9\,t^2 + 18\,t - 6, && \|\, q_3\,\|^2 = 36\,.
\end{aligned}$$

The resulting orthogonal polynomials are known as the (monic) *Laguerre polynomials*, named after the nineteenth century French mathematician Edmond Laguerre.

In some cases, a change of variables may be used to relate systems of orthogonal polynomials and thereby circumvent the Gram–Schmidt computation. Suppose, for instance, that our goal is to construct an orthogonal system of polynomials for the L^2 inner product $\langle\!\langle\, f\,;g\,\rangle\!\rangle = \displaystyle\int_a^b f(t)\, g(t)\, dt$ on the interval $[\,a,b\,]$. The key remark is that we can map the interval $[\,-1,1\,]$ to $[\,a,b\,]$ by a simple linear change of variables of the form $s = \alpha + \beta\, t$. Specifically,

$$s = \frac{2\,t - b - a}{b - a} \qquad \text{will change} \qquad a \le t \le b \qquad \text{to} \qquad -\,1 \le s \le 1. \tag{5.50}$$

The map changes functions $F(s), G(s)$, defined for $-1 \le s \le 1$, into the functions

$$f(t) = F\left(\, \frac{2\,t - b - a}{b - a}\,\right), \qquad g(t) = G\left(\, \frac{2\,t - b - a}{b - a}\,\right), \tag{5.51}$$

defined for $a \le t \le b$. Moreover, interpreting (5.50) as a change of variables for the integrals, we have $ds = \dfrac{2}{b-a}\, dt$, and so the inner products are related by

$$\begin{aligned}
\langle\, f\,;g\,\rangle = \int_a^b f(t)\, g(t)\, dt &= \int_a^b F\left(\, \frac{2\,t - b - a}{b - a}\,\right)\, G\left(\, \frac{2\,t - b - a}{b - a}\,\right) dt\\
&= \int_{-1}^1 F(s)\, G(s)\, \frac{b-a}{2}\, ds = \frac{b-a}{2}\, \langle\, F\,;G\,\rangle,
\end{aligned} \tag{5.52}$$

where the final L^2 inner product is over the interval $[-1, 1]$. In particular, the change of variables maintains orthogonality, while rescaling the norms:

$$\langle\, f\,;g\,\rangle = 0 \quad \text{if and only if} \quad \langle\, F\,;G\,\rangle = 0, \qquad \| f \| = \sqrt{\frac{b-a}{2}}\ \| F \|. \tag{5.53}$$

Moreover, if $F(s)$ is a polynomial of degree n in s, then $f(t)$ is a polynomial of degree n in t and vice versa. Applying these observations to the Legendre polynomials, we immediately deduce the following.

Proposition 5.31. *The transformed Legendre polynomials*

$$\widetilde{P}_k(t) = P_k\left(\frac{2t-b-a}{b-a}\right), \qquad k = 0, 1, 2, \dots, \qquad \| \widetilde{P}_k \| = \sqrt{\frac{b-a}{2\,k+1}}\,, \tag{5.54}$$

form an orthogonal system of polynomials with respect to the L^2 *inner product on the interval* $[a, b]$.

Example 5.32. As an example, consider the L^2 inner product $\langle\!\langle\, f\,;g\,\rangle\!\rangle = \displaystyle\int_0^1 f(t)\,g(t)\,dt$ on the interval $[0, 1]$. The map $s = 2t - 1$ will change $0 \le t \le 1$ to $-1 \le s \le 1$. According to Proposition 5.31, this change of variables will convert the Legendre polynomials $P_k(s)$ into an orthogonal system of polynomials

$$\widetilde{P}_k(t) = P_k(2t-1), \qquad \text{with corresponding } \mathrm{L}^2 \text{ norms} \qquad \| \widetilde{P}_k \| = \sqrt{\frac{1}{2\,k+1}}\,.$$

on the interval $[0, 1]$. The first few are

$$\begin{aligned}
&\widetilde{P}_0(t) = 1, && \widetilde{P}_3(t) = 20\,t^3 - 30\,t^2 + 12\,t - 1,\\
&\widetilde{P}_1(t) = 2\,t - 1, && \widetilde{P}_4(t) = 70\,t^4 - 140\,t^3 + 90\,t^2 - 20\,t + 1,\\
&\widetilde{P}_2(t) = 6\,t^2 - 6\,t + 1, && \widetilde{P}_5(t) = \tfrac{63}{8}\,t^5 - \tfrac{35}{4}\,t^3 + \tfrac{15}{8}\,t.
\end{aligned} \tag{5.55}$$

One can, as an alternative, derive these formulae through a direct application of the Gram–Schmidt process.

5.5. Orthogonal Projections and Least Squares.

In Chapter 4, we introduced, solved and learned the significance of the problem of finding the point on a prescribed subspace that lies closest to a given point. In this section, we shall discover an important geometrical interpretation of our solution: the closest point is the *orthogonal projection* of the point onto the subspace. Furthermore, if we adopt an orthogonal, or, even better, orthonormal basis for the subspace, then the closest point can be constructed through a very elegant, explicit formula. In this manner, orthogonality allows us to effectively bypass the normal equations and solution formulae that were so laboriously computed in Chapter 4. The resulting orthogonal projection formulae have important practical consequences for the solution of a wide range of least squares minimization problems.

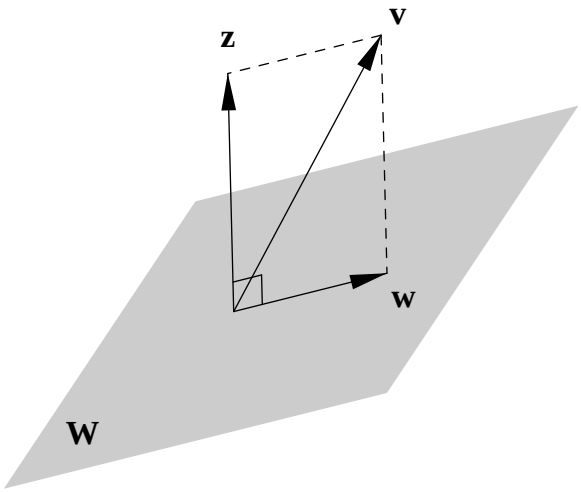


Figure 5.4. The Orthogonal Projection of a Vector onto a Subspace.

Orthogonal Projection

We begin by characterizing the orthogonal projection of a vector onto a subspace. Throughout this section, we will consider a prescribed finite-dimensional subspace $W \subset V$ of a real inner product space V. While the subspace is necessarily finite-dimensional, the inner product space itself may be infinite-dimensional. Initially, though, you may wish to concentrate on $V = \mathbb{R}^m$ with the ordinary Euclidean dot product, which is the easiest case to visualize as it coincides with our geometric intuition, as in Figure 5.4.

A vector $\mathbf{z} \in V$ is said to be *orthogonal* to the subspace W if it is orthogonal to every vector in W, so $\langle \mathbf{z} ; \mathbf{w} \rangle = 0$ for all $\mathbf{w} \in W$. Given a basis $\mathbf{w}_1, \dots, \mathbf{w}_n$ for W, we note that $\mathbf{z}$ is orthogonal to W if and only if it is orthogonal to every basis vector: $\langle \mathbf{z} ; \mathbf{w}_i \rangle = 0$ for $i = 1, \dots, n$. Indeed, any other vector in W has the form $\mathbf{w} = c_1 \mathbf{w}_1 + \cdots + c_n \mathbf{w}_n$ and hence, by linearity, $\langle \mathbf{z} ; \mathbf{w} \rangle = c_1 \langle \mathbf{z} ; \mathbf{w}_1 \rangle + \cdots + c_n \langle \mathbf{z} ; \mathbf{w}_n \rangle = 0$, as required.

Definition 5.33. The *orthogonal projection* of $\mathbf{v}$ onto the subspace W is the element $\mathbf{w} \in W$ that makes the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ orthogonal to W.

As we shall see, the orthogonal projection is unique. The explicit construction is greatly simplified by taking a orthonormal basis of the subspace, which, if necessary, can be arranged by applying the Gram–Schmidt process to a known basis. (A direct construction of the orthogonal projection in terms of a general basis appears in Exercise ■.)

Theorem 5.34. *Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be an orthonormal basis for the subspace $W \subset V$. Then the orthogonal projection of a vector $\mathbf{v} \in V$ onto W is*

$$\mathbf{w} = c_1 \mathbf{u}_1 + \cdots + c_n \mathbf{u}_n \qquad \text{where} \qquad c_i = \langle \mathbf{v} ; \mathbf{u}_i \rangle, \qquad i = 1, \dots, n. \tag{5.56}$$

Proof: First, since $\mathbf{u}_1, \dots, \mathbf{u}_n$ form a basis of the subspace, the orthogonal projection element $\mathbf{w} = c_1 \mathbf{u}_1 + \cdots + c_n \mathbf{u}_n$ must be some linear combination thereof. Definition 5.33

requires that the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ be orthogonal to W. It suffices to check orthogonality to the basis vectors of W. By our orthonormality assumption, for each $1 \le i \le n$,

$$\begin{aligned} 0 = \langle\, \mathbf{z} \,;\, \mathbf{u}_i \,\rangle &= \langle\, \mathbf{v} \,;\, \mathbf{u}_i \,\rangle - \langle\, \mathbf{w} \,;\, \mathbf{u}_i \,\rangle = \langle\, \mathbf{v} \,;\, \mathbf{u}_i \,\rangle - \langle\, c_1\, \mathbf{u}_1 + \cdots + c_n\, \mathbf{u}_n \,;\, \mathbf{u}_i \,\rangle \\ &= \langle\, \mathbf{v} \,;\, \mathbf{u}_i \,\rangle - c_1 \langle\, \mathbf{u}_1 \,;\, \mathbf{u}_i \,\rangle - \ \cdots\ - c_n \langle\, \mathbf{u}_n \,;\, \mathbf{u}_i \,\rangle = \langle\, \mathbf{v} \,;\, \mathbf{u}_i \,\rangle - c_i. \end{aligned}$$

We deduce that the coefficients $c_i = \langle\, \mathbf{v} \,;\, \mathbf{u}_i \,\rangle$ of the orthogonal projection $\mathbf{w}$ are uniquely prescribed by the orthogonality requirement. *Q.E.D.*

More generally, if we employ an orthogonal basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ for the subspace W, then the same argument demonstrates that the orthogonal projection of $\mathbf{v}$ onto W is given by

$$\mathbf{w} = a_1 \mathbf{v}_1 + \ \cdots\ + a_n \mathbf{v}_n, \qquad \text{where} \qquad a_i = \frac{\langle\, \mathbf{v} \,;\, \mathbf{v}_i \,\rangle}{\| \,\mathbf{v}_i\, \|^2}, \qquad i = 1, \dots, n. \tag{5.57}$$

Of course, we could equally well replace the orthogonal basis by the orthonormal basis obtained by dividing each vector by its length: $\mathbf{u}_i = \mathbf{v}_i / \| \,\mathbf{v}_i\, \|$. The reader should be able to prove that the two formulae (5.56), (5.57) for the orthogonal projection yield the same vector $\mathbf{w}$.

Example 5.35. Consider the plane $W \subset \mathbb{R}^3$ spanned by the orthogonal vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

According to formula (5.57), the orthogonal projection of $\mathbf{v} = (\,1, 0, 0\,)^T$ onto W is

$$\mathbf{w} = \frac{\langle\, \mathbf{v} \,;\, \mathbf{v}_1 \,\rangle}{\| \,\mathbf{v}_1\, \|^2}\, \mathbf{v}_1 + \frac{\langle\, \mathbf{v} \,;\, \mathbf{v}_2 \,\rangle}{\| \,\mathbf{v}_2\, \|^2}\, \mathbf{v}_2 = \frac{1}{6} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}.$$

Alternatively, we can replace $\mathbf{v}_1, \mathbf{v}_2$ by the orthonormal basis

$$\mathbf{u}_1 = \frac{\mathbf{v}_1}{\| \,\mathbf{v}_1\, \|} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix}, \qquad \mathbf{u}_2 = \frac{\mathbf{v}_2}{\| \,\mathbf{v}_2\, \|} = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}.$$

Then, using the orthonormal version (5.56),

$$\mathbf{w} = \langle\, \mathbf{v} \,;\, \mathbf{u}_1 \,\rangle\, \mathbf{u}_1 + \langle\, \mathbf{v} \,;\, \mathbf{u}_2 \,\rangle\, \mathbf{u}_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{pmatrix} + \frac{1}{\sqrt{3}} \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{pmatrix}.$$

The answer is, of course, the same. As the reader may notice, while the theoretical formula is simpler when written in an orthonormal basis, for hand computations the orthogonal basis version avoids dealing with square roots. (Of course, when performing the computation on a computer, this is not a significant issue.)

An intriguing observation is that the coefficients in the orthogonal projection formulae (5.56) and (5.57) coincide with the formulae (5.5), (5.8) for writing a vector in terms of an orthonormal or orthogonal basis. Indeed, if $\mathbf{v}$ were an element of W, then it would coincide with its orthogonal projection, $\mathbf{w} = \mathbf{v}$ (why?). As a result, the orthogonal projection formulae include the orthogonal basis formulae as a special case.

It is also worth noting that the *same* formulae occur in the Gram–Schmidt algorithm, (5.19). This observation leads to a useful geometric interpretation for the Gram–Schmidt construction. For each $k = 1, \dots, n$, let

$$V_k = \operatorname{span}\{\mathbf{w}_1, \dots, \mathbf{w}_k\} = \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\} = \operatorname{span}\{\mathbf{u}_1, \dots, \mathbf{u}_k\} \tag{5.58}$$

denote the k-dimensional subspace spanned by the first k basis elements. The basic Gram–Schmidt formula (5.20) can be rewritten in the form $\mathbf{v}_k = \mathbf{w}_k - \mathbf{y}_k$, where $\mathbf{y}_k$ is the orthogonal projection of $\mathbf{w}_k$ onto the subspace V_{k-1}. The resulting vector $\mathbf{v}_k$ is, by construction, orthogonal to the subspace, and hence orthogonal to all of the previous basis elements, which serves to rejustify the Gram–Schmidt construction.

Orthogonal Least Squares

Now we make an important connection: The orthogonal projection of a vector onto a subspace is also the least squares vector — the closest point in the subspace!

Theorem 5.36. *Let $W \subset V$ be a finite-dimensional subspace of an inner product space. Given a vector $\mathbf{v} \in V$, the closest point or least squares minimizer $\mathbf{w} \in W$ is the same as the orthogonal projection of $\mathbf{v}$ onto W.*

Proof: Let $\mathbf{w} \in W$ be the orthogonal projection of $\mathbf{v}$ onto the subspace, which requires that the difference $\mathbf{z} = \mathbf{v} - \mathbf{w}$ be orthogonal to W. Suppose $\widetilde{\mathbf{w}} \in W$ is any other vector in the subspace. Then,

$$\| \mathbf{v} - \widetilde{\mathbf{w}} \|^2 = \| \mathbf{w} + \mathbf{z} - \widetilde{\mathbf{w}} \|^2 = \| \mathbf{w} - \widetilde{\mathbf{w}} \|^2 + 2 \langle \mathbf{w} - \widetilde{\mathbf{w}} \,;\, \mathbf{z} \rangle + \| \mathbf{z} \|^2 = \| \mathbf{w} - \widetilde{\mathbf{w}} \|^2 + \| \mathbf{z} \|^2.$$

The inner product term $\langle \mathbf{w} - \widetilde{\mathbf{w}} \,;\, \mathbf{z} \rangle = 0$ vanishes because $\mathbf{z}$ is orthogonal to every vector in W, including $\mathbf{w} - \widetilde{\mathbf{w}}$. Since $\mathbf{z} = \mathbf{v} - \mathbf{w}$ is uniquely prescribed by the vector $\mathbf{v}$, the second term $\| \mathbf{z} \|^2$ does not change with the choice of the point $\widetilde{\mathbf{w}} \in W$. Therefore, $\| \mathbf{v} - \widetilde{\mathbf{w}} \|^2$ will be minimized if and only if $\| \mathbf{w} - \widetilde{\mathbf{w}} \|^2$ is minimized. Since $\widetilde{\mathbf{w}} \in W$ is allowed to be any element of the subspace W, the minimal value $\| \mathbf{w} - \widetilde{\mathbf{w}} \|^2 = 0$ occurs when $\widetilde{\mathbf{w}} = \mathbf{w}$. Thus, the closest point $\widetilde{\mathbf{w}}$ coincides with the orthogonal projection $\mathbf{w}$. *Q.E.D.*

In particular, if we are supplied with an orthonormal or orthogonal basis of our subspace, then we can compute the closest least squares point $\mathbf{w} \in W$ to $\mathbf{v}$ using our orthogonal projection formulae (5.56) or (5.57). In this way, orthogonal bases have a very dramatic simplifying effect on the least squares approximation formulae. They completely avoid the construction of and solution to the much more complicated normal equations.

Example 5.37. Consider the least squares problem of finding the closest point $\mathbf{w}$ to the vector $\mathbf{v} = (\,1, 2, 2, 1\,)^T$ in the three-dimensional subspace spanned by the orthogonal[†]

[†] We use the ordinary Euclidean norm on $\mathbb{R}^4$ throughout this example.

vectors $\mathbf{v}_1 = (\,1,-1,2,0\,)^T$, $\mathbf{v}_2 = (\,0,2,1,-2\,)^T$, $\mathbf{v}_3 = (\,1,1,0,1\,)^T$. Since the spanning vectors are orthogonal (but not orthonormal), we can use the orthogonal projection formula (5.57) to find the linear combination $\mathbf{w} = a_1\,\mathbf{v}_1 + a_2\,\mathbf{v}_2 + a_3\,\mathbf{v}_3$. Thus,

$$a_1 = \frac{\langle\,\mathbf{v}\,;\mathbf{v}_1\,\rangle}{\|\,\mathbf{v}_1\,\|^2} = \frac{3}{6} = \frac{1}{2}\,, \qquad a_2 = \frac{\langle\,\mathbf{v}\,;\mathbf{v}_2\,\rangle}{\|\,\mathbf{v}_2\,\|^2} = \frac{4}{9}\,, \qquad a_3 = \frac{\langle\,\mathbf{v}\,;\mathbf{v}_3\,\rangle}{\|\,\mathbf{v}_3\,\|^2} = \frac{4}{3}\,,$$

and so $\mathbf{w} = \frac{1}{2}\,\mathbf{v}_1 + \frac{4}{9}\,\mathbf{v}_2 + \frac{4}{3}\,\mathbf{v}_3 = \left(\,\frac{11}{6}, \frac{31}{18}, \frac{13}{9}, \frac{4}{9}\,\right)^T$ is the closest point to $\mathbf{v}$ in the subspace.

Even when we only know a non-orthogonal basis for the subspace, it may still be a good strategy to first use Gram–Schmidt to replace it by an orthogonal or even orthonormal basis, and then apply the orthogonal projection formulae (5.56), (5.57) to calculate the least squares point. Not only does this simplify the final computation, it will often avoid the ill-conditioning and numerical inaccuracies that sometimes afflict the direct solution to the normal equations (4.26). The following example illustrates this alternative procedure.

Example 5.38. Let us return to the problem, solved in Example 4.6, of finding the closest point on plane V spanned by $\mathbf{w}_1 = (\,1,2,-1\,)^T$, $\mathbf{w}_2 = (\,2,-3,-1\,)^T$ to the point $\mathbf{b} = (\,1,0,0\,)^T$. We proceed now by first using the Gram–Schmidt process to compute an orthogonal basis

$$\mathbf{v}_1 = \mathbf{w}_1 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_2 = \mathbf{w}_2 - \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\|\,\mathbf{v}_1\,\|^2}\,\mathbf{w}_1 = \begin{pmatrix} \frac{5}{2} \\ -2 \\ -\frac{3}{2} \end{pmatrix},$$

for our subspace. Therefore, applying the orthogonal projection formula (5.57), the closest point is

$$\mathbf{v}^\star = \frac{\mathbf{b} \cdot \mathbf{v}_1}{\|\,\mathbf{v}_1\,\|^2}\,\mathbf{v}_1 + \frac{\mathbf{b} \cdot \mathbf{v}_2}{\|\,\mathbf{v}_2\,\|^2}\,\mathbf{v}_2 = \begin{pmatrix} \frac{2}{3} \\ -\frac{1}{15} \\ -\frac{7}{15} \end{pmatrix},$$

reconfirming our earlier result. By this device, we have managed to circumvent the tedious solving of linear equations.

Let us revisit the problem, described in Section 4.4, of approximating experimental data by a least squares minimization procedure. The required calculations are significantly simplified by the introduction of an orthogonal basis of the least squares subspace. Given sample points $t_1, \ldots, t_m$, let

$$\mathbf{t}_k = \left(\,t_1^k, t_2^k, \ldots, t_m^k\,\right)^T, \qquad k = 0, 1, 2, \ldots$$

be the vectors obtained by sampling the monomial t^k. More generally, sampling a polynomial

$$y = p(t) = \alpha_0 + \alpha_1\,t + \,\cdots\, + \alpha_n\,t^n \tag{5.59}$$

results in the self-same linear combination

$$\mathbf{p} = (\,p(t_1), \ldots, p(t_n)\,)^T = \alpha_0\,\mathbf{t}_0 + \alpha_1\,\mathbf{t}_1 + \,\cdots\, + \alpha_n\,\mathbf{t}_n \tag{5.60}$$

of monomial sample vectors. We conclude that the sampled polynomial vectors form a subspace $W = \text{span}\,\{\,\mathbf{t}_0, \dots, \mathbf{t}_n\,\} \subset \mathbb{R}^m$ spanned by the monomial sample vectors.

Let $\mathbf{y} = (\,y_1, y_2, \dots, y_m\,)^T$ denote data measured at the sample points. The polynomial least squares approximation to the given data is, by definition, the polynomial $y = p(t)$ whose corresponding sample vector $\mathbf{p} \in W$ is the closest point or, equivalently, the orthogonal projection of the data vector $\mathbf{y}$ onto the subspace W. The sample vectors $\mathbf{t}_0, \dots, \mathbf{t}_n$ are not orthogonal, and so the direct approach requires solving the normal equations (4.33) in order to find the desired polynomial least squares coefficients $\alpha_0, \dots, \alpha_n$.

An alternative method is to first use the Gram–Schmidt procedure to construct an orthogonal basis for the subspace W, from which the least squares coefficients are found by simply taking appropriate inner products. Let us adopt the rescaled version

$$\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \frac{1}{m}\,\sum_{i=1}^{m} v_i\, w_i = \overline{v\, w} \tag{5.61}$$

of the standard dot product† on $\mathbb{R}^m$. If $\mathbf{v}, \mathbf{w}$ represent the sample vectors corresponding to the functions $v(t), w(t)$, then their inner product $\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle$ is equal to the average value of the product function $v(t)\, w(t)$ on the m sample points. In particular, the inner product between our "monomial" basis vectors corresponding to sampling t^k and t^l is

$$\langle\, \mathbf{t}_k\,;\mathbf{t}_l\,\rangle = \frac{1}{m}\,\sum_{i=1}^{m} t_i^k\, t_i^l = \frac{1}{m}\,\sum_{i=1}^{m} t_i^{k+l} = \overline{t^{k+l}}, \tag{5.62}$$

which is the averaged sample value of the monomial t^{k+l}.

To keep the formulae reasonably simple, let us further assume‡ that the sample points are evenly spaced and symmetric about 0. The second requirement means that if t_i is a sample point, so is $-\,t_i$. An example would be the seven sample points $-3, -2, -1, 0, 1, 2, 3$. As a consequence of these two assumptions, the averaged sample values of the odd powers of t vanish: $\overline{t^{2i+1}} = 0$. Hence, by (5.62), the sample vectors $\mathbf{t}_k$ and $\mathbf{t}_l$ are orthogonal whenever $k + l$ is odd.

Applying the Gram–Schmidt algorithm to $\mathbf{t}_0, \mathbf{t}_1, \mathbf{t}_2, \dots$ produces the orthogonal basis vectors $\mathbf{q}_0, \mathbf{q}_1, \mathbf{q}_2, \dots$. Each $\mathbf{q}_k = (\,q_k(t_1), \dots, q_k(t_m)\,)^T$ can be interpreted as the sample vector for a certain interpolating polynomial $q_k(t)$ of degree k. The first few polynomials $q_k(t)$, their corresponding orthogonal sample vectors, along with their squared norms,

† For weighted least squares, we would adopt an appropriately weighted inner product.

‡ The method works without these particular assumptions, but the formulas become more unwieldy; see Exercise ■.

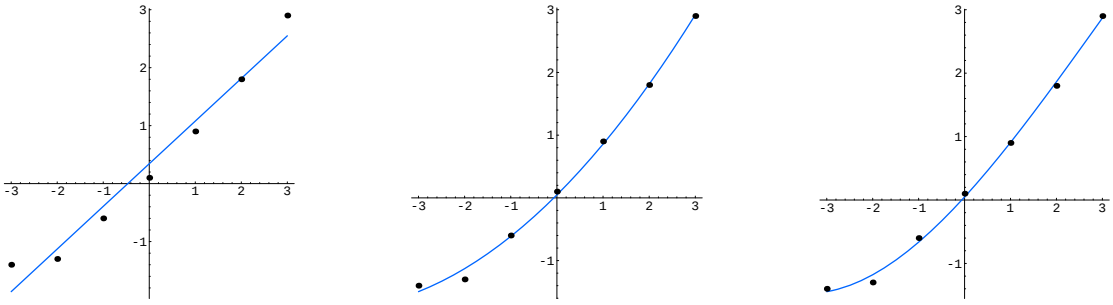


Figure 5.5. Least Squares Data Approximations.

$\| \mathbf{q}_k \|^2 = \overline{q_k(t)^2}$, follow:

$$
\begin{aligned}
&q_0(t) = 1, && \mathbf{q}_0 = \mathbf{t}_0, && \| \mathbf{q}_0 \|^2 = 1,\\
&q_1(t) = t, && \mathbf{q}_1 = \mathbf{t}_1, && \| \mathbf{q}_1 \|^2 = \overline{t^2},\\
&q_2(t) = t^2 - \overline{t^2}\,, && \mathbf{q}_2 = \mathbf{t}_2 - \overline{t^2}\,\mathbf{t}_0, && \| \mathbf{q}_2 \|^2 = \overline{t^4} - \big(\overline{t^2}\big)^2,\\
&q_3(t) = t^3 - \frac{\overline{t^4}}{\overline{t^2}}\, t\,, && \mathbf{q}_3 = \mathbf{t}_3 - \frac{\overline{t^4}}{\overline{t^2}}\,\mathbf{t}_1, && \| \mathbf{q}_3 \|^2 = \overline{t^6} - \frac{\big(\overline{t^4}\big)^2}{\overline{t^2}}\,.
\end{aligned}
\tag{5.63}
$$

With these in hand, the least squares approximating polynomial of degree n to the given data vector $\mathbf{y}$ is given by a linear combination

$$
p(t) = a_0\, q_0(t) + a_1\, q_1(t) + a_2\, q_2(t) + \, \cdots \, + a_n\, q_n(t). \tag{5.64}
$$

The required coefficients are obtained directly through the orthogonality formulae (5.57), and so

$$
a_k = \frac{\langle\, \mathbf{q}_k \,;\, \mathbf{y} \,\rangle}{\| \mathbf{q}_k \|^2} = \frac{\overline{q_k\, y}}{\overline{q_k^2}}\,. \tag{5.65}
$$

An additional advantage of the orthogonal basis approach, beyond the fact that one can write down explicit formulas for the coefficients, is that the *same* coefficients a_j appear in all the least squares formulae, and hence one can readily increase the degree, and, presumably, the accuracy, of the approximating polynomial without having to recompute any of the lower degree terms. For instance, if a quadratic approximant $a_0 + a_1\, q_1(t) + a_2\, q_2(t)$ looks insufficiently close, one can add in the cubic term $a_3\, q_3(t)$ with a_3 given by (5.65) for $k = 3$, *without* having to recompute the quadratic coefficients a_0, a_1, a_2. This simplification is *not* valid when using the non-orthogonal basis elements, where the lower order coefficients will change whenever the degree of the approximating polynomial is increased.

.

Example 5.39. Consider the following tabulated sample values:

t_i	-3	-2	-1	0	1	2	3
y_i	-1.4	-1.3	$-.6$	$.1$	$.9$	1.8	2.9

To compute polynomial least squares fits of degrees $1, 2$ and 3, we begin by computing the polynomials (5.63), which for the given sample points t_i are

$$\begin{aligned} &q_0(t) = 1, \qquad q_1(t) = t, \qquad q_2(t) = t^2 - 4, \qquad q_3(t) = t^3 - 7t\,, \\ &\| \mathbf{q}_0 \|^2 = 1, \qquad \| \mathbf{q}_1 \|^2 = 4, \qquad \| \mathbf{q}_2 \|^2 = 12, \qquad \| \mathbf{q}_3 \|^2 = \tfrac{216}{7}. \end{aligned}$$

Thus, to four decimal places, the coefficients for the least squares approximation (5.64) are

$$\begin{aligned} &a_0 = \langle\, \mathbf{q}_0 \,;\, \mathbf{y} \,\rangle = 0.3429, \qquad &&a_1 = \tfrac{1}{4} \langle\, \mathbf{q}_1 \,;\, \mathbf{y} \,\rangle = 0.7357, \\ &a_2 = \tfrac{1}{12} \langle\, \mathbf{q}_2 \,;\, \mathbf{y} \,\rangle = 0.0738, \qquad &&a_3 = \tfrac{7}{216} \langle\, \mathbf{q}_3 \,;\, \mathbf{y} \,\rangle = -0.0083. \end{aligned}$$

To obtain the best linear approximation, we use

$$p_1(t) = a_0\, q_0(t) + a_1\, q_1(t) = 0.3429 + 0.7357\, t,$$

with a least squares error of 0.7081. Similarly, the quadratic and cubic least squares approximations are

$$\begin{aligned} p_2(t) &= 0.3429 + 0.7357\, t + 0.0738\, (t^2 - 4), \\ p_3(t) &= 0.3429 + 0.7357\, t + 0.0738\, (t^2 - 4) - 0.0083\, (t^3 - 7t), \end{aligned}$$

with respective least squares errors 0.2093 and 0.1697 at the sample points. A plot of the three approximations appears in Figure 5.5. The cubic term does not significantly increase the accuracy of the approximation, and so this data probably comes from sampling a quadratic function.

Orthogonal Polynomials and Least Squares

In a similar fashion, the orthogonality of Legendre polynomials and more general orthogonal functions serves to simplify the construction of least squares approximants in function space. As an example, let us reconsider the problem, from Chapter 4, of approximating e^t by a polynomial of degree n. For the interval $-1 \le t \le 1$, we write the best least squares approximant as a linear combination of Legendre polynomials,

$$p(t) = a_0\, P_0(t) + a_1\, P_1(t) + \,\cdots\, + a_n\, P_n(t) = a_0 + a_1\, t + a_2 \left(\tfrac{3}{2}\, t^2 - \tfrac{1}{2}\right) + \,\cdots\,. \tag{5.66}$$

Since the Legendre polynomials form an orthogonal basis, the least squares coefficients can be immediately computed by the inner product formula (5.57), so

$$a_k = \frac{\langle\, e^t \,;\, P_k \,\rangle}{\| P_k \|^2} = \frac{2\,k + 1}{2} \int_{-1}^{1} e^t\, P_k(t)\, dt.$$

For example, the quadratic approximant is obtained from the first three terms in (5.66), where

$$a_0 = \frac{1}{2} \int_{-1}^{1} e^t\, dt = \frac{1}{2} \left(e - \frac{1}{e}\right) \simeq 1.175201, \qquad a_1 = \frac{3}{2} \int_{-1}^{1} t\, e^t\, dt = \frac{3}{e} \simeq 1.103638,$$

$$a_2 = \frac{5}{2} \int_{-1}^{1} \left(\tfrac{3}{2}\, t^2 - \tfrac{1}{2}\right) e^t\, dt = \frac{5}{2} \left(e - \frac{7}{e}\right) \simeq .357814.$$

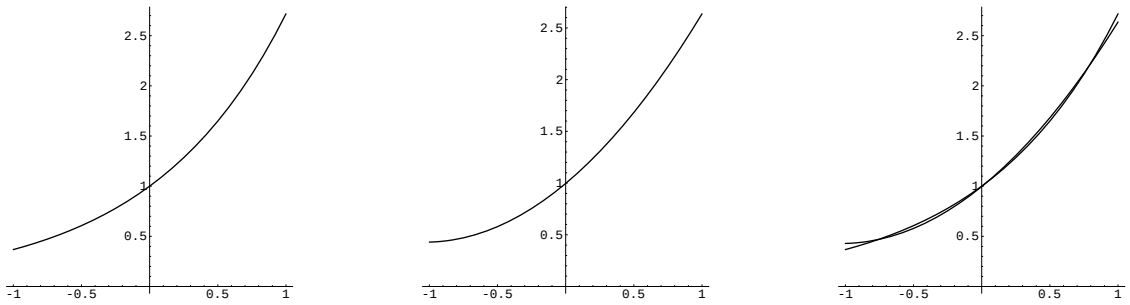


Figure 5.6. Quadratic Least Squares Approximation to e^t.

Therefore

$$e^t \;\approx\; 1.175201 + 1.103638\,t + .357814\,\left(\tfrac{3}{2}\,t^2 - \tfrac{1}{2}\right) \tag{5.67}$$

gives the quadratic least squares approximation to e^t on $[-1, 1]$. Graphs appear in Figure 5.6; the first graph shows e^t, the second the quadratic approximant (5.67), and the third lays the two graphs on top of each other.

As in the discrete case, there are two major advantages of the orthogonal Legendre approach over the direct approach presented in Example 4.21. First, we do not need to solve any linear systems of equations. Indeed, the coefficient matrix for polynomial least squares approximation based on the monomial basis is some variant of the notoriously ill-conditioned Hilbert matrix, (1.67), and the computation of an accurate solution is particularly tricky. Our precomputation of an orthogonal system of polynomials has successfully circumvented the dangerous Hilbert matrix system.

The second advantage was already mentioned in the preceding subsection. Unlike the direct approach, the coefficients a_k do *not* change if we desire to go to higher accuracy by increasing the degree of the approximating polynomial. For instance, in the first case, if the quadratic approximation (5.67) is not accurate enough, we can add in a cubic correction $a_3\,P_3(t) = a_3\,\left(\frac{5}{2}\,t^3 - \frac{3}{2}\,t\right)$, where we compute the required coefficient by

$$a_3 = \frac{7}{2}\int_{-1}^{1} \left(\tfrac{5}{2}\,t^3 - \tfrac{3}{2}\,t\right) e^t\,dt = \frac{7}{2}\left(37\,e - \frac{5}{e}\right) \simeq .070456.$$

We do not need to recompute the coefficients a_0, a_1, a_2. The successive Legendre polynomial coefficients decrease fairly rapidly:

$$\begin{aligned} &a_0 \simeq 1.175201, \qquad a_1 \simeq 1.103638, \qquad a_2 \simeq .357814, \qquad a_3 \simeq .070456,\\ &a_4 \simeq .009965, \qquad a_5 \simeq .001100, \qquad a_6 \simeq .000099, \end{aligned}$$

leading to greater and greater accuracy in the least squares approximation. An explanation will appear in Chapter 12.

If we switch to another norm, then we need to construct an associated set of orthogonal polynomials to apply the method. For instance, the polynomial least squares approximation of degree n to a function $f(t)$ with respect to the L^2 norm on $[0, 1]$ has the form $a_0 + a_1\,\widetilde{P}_1(t) + a_2\,\widetilde{P}_2(t) + \,\cdots\, + a_n\,\widetilde{P}_n(t)$, where $\widetilde{P}_1(t)$ are the rescaled Legendre

polynomials (5.55), and, by orthogonality,

$$a_k = \frac{\langle\, f\,;\widetilde{P}_k\,\rangle}{\|\,\widetilde{P}_k\,\|^2} = (2\,k+1)\,\int_0^1 f(t)\,\widetilde{P}_k(t)\,dt.$$

For the particular function e^t, we find

$$a_0 = \int_0^1 e^t\,dt = e-1\ \simeq\ 1.718282, \qquad a_1 = 3\int_0^1 (2t-1)\,e^t\,dt = 3(3-e)\ \simeq\ .845155,$$

$$a_2 = 5\int_0^1 (6t^2-6t+1)e^t\,dt = 5(7\,e-19)\ \simeq\ .139864.$$

Thus, the best quadratic least squares approximation is

$$\begin{aligned} p_2^\star(t) &= 1.718282 + .845155\,(2t-1) + .139864\,(6t^2-6t+1)\\ &= 1.012991 + .851125\,t + .839184\,t^2. \end{aligned}$$

It is worth emphasizing that this is the *same* approximating polynomial as we computed in (4.60). The use of an orthogonal system of polynomials merely streamlines the computation.

5.6. Orthogonal Subspaces.

We now extend the notion of orthogonality from individual elements to subspaces of an inner product space V.

Definition 5.40. Two subspaces $W, Z \subset V$ are called *orthogonal* if every vector in W is orthogonal to every vector in Z.

In other words, W and Z are orthogonal subspaces if and only if $\langle\,\mathbf{w}\,;\mathbf{z}\,\rangle = 0$ for every $\mathbf{w}\in W, \mathbf{z}\in Z$. In practice, one only needs to check orthogonality of basis elements: If $\mathbf{w}_1,\dots,\mathbf{w}_k$ is a basis for W and $\mathbf{z}_1,\dots,\mathbf{z}_l$ a basis for Z, then W and Z are orthogonal if and only if $\langle\,\mathbf{w}_i\,;\mathbf{z}_j\,\rangle = 0$ for all $i=1,\dots,k$ and $j=1,\dots,l$.

Example 5.41. The plane $W\subset\mathbb{R}^3$ defined by the equation $2x-y+3z=0$ is orthogonal, with respect to the dot product, to the line Z spanned by its normal vector $\mathbf{n} = (\,2,-1,3\,)^T$. Indeed, every $\mathbf{w} = (\,x,y,z\,)^T \in W$ satisfies the orthogonality condition $\mathbf{w}\cdot\mathbf{n} = 2x-y+3z=0$, which is just the equation for the plane.

Example 5.42. Let W be the span of $\mathbf{w}_1 = (\,1,-2,0,1\,)^T, \mathbf{w}_2 = (\,3,1,2,1\,)^T$, and Z the span of $\mathbf{z}_1 = (\,3,2,0,1\,)^T, \mathbf{z}_2 = (\,1,0,-1,-1\,)^T$. Then all $\mathbf{w}_i\cdot\mathbf{z}_j = 0$, and hence W and Z are orthogonal subspaces of $\mathbb{R}^4$ under the Euclidean dot product.

Definition 5.43. The *orthogonal complement* to a subspace $W\subset V$, denoted $W^\perp$, is defined as the set of all vectors which are orthogonal to W, so

$$W^\perp = \{\,\mathbf{v}\in V\mid\ \langle\,\mathbf{v}\,;\mathbf{w}\,\rangle = 0\ \text{ for all }\ \mathbf{w}\in W\,\}. \tag{5.68}$$

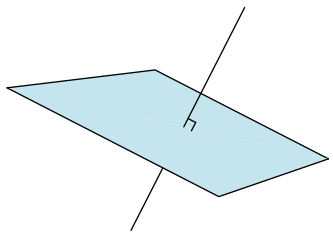


Figure 5.7. Orthogonal Complement to a Line.

One easily checks that the orthogonal complement $W^\perp$ to a subspace $W \subset V$ is also a subspace. Moreover, $W \cap W^\perp = \{\mathbf{0}\}$. (Why?) Note that the orthogonal complement to a subspace will depend upon which inner product is being used. In the remainder of the chapter, we will concentrate exclusively on the Euclidean inner product.

Example 5.44. Let $W = \{\, (t, 2t, 3t)^T \mid t \in \mathbb{R} \,\}$ be the line (one-dimensional subspace) in the direction of the vector $\mathbf{w}_1 = (1, 2, 3)^T \in \mathbb{R}^3$. The orthogonal complement $W^\perp$ will be the plane passing through the origin having normal vector $\mathbf{w}_1$, as sketched in Figure 5.7. In other words, $\mathbf{z} = (x, y, z)^T \in W^\perp$ if and only if

$$\mathbf{z} \cdot \mathbf{w}_1 = x + 2y + 3z = 0. \tag{5.69}$$

Thus $W^\perp$ is characterized as the solution space to the homogeneous linear equation (5.69), or, equivalently, the kernel of the 1×3 matrix $A = \mathbf{w}_1^T = (\,1 \;\; 2 \;\; 3\,)$. We can write the general solution to the equation in the form

$$\mathbf{z} = \begin{pmatrix} -2y - 3z \\ y \\ z \end{pmatrix} = y \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} -3 \\ 0 \\ 1 \end{pmatrix} = y\,\mathbf{z}_1 + z\,\mathbf{z}_2,$$

where y, z are the free variables. The indicated solution vectors $\mathbf{z}_1 = (-2, 1, 0)^T$, $\mathbf{z}_2 = (-3, 0, 1)^T$, form a (non-orthogonal) basis for the orthogonal complement $W^\perp$.

Proposition 5.45. *Suppose that $W \subset V$ is a finite-dimensional subspace of an inner product space. Then every vector $\mathbf{v} \in V$ can be uniquely decomposed into $\mathbf{v} = \mathbf{w} + \mathbf{z}$ where $\mathbf{w} \in W$ and $\mathbf{z} \in W^\perp$.*

Proof: We let $\mathbf{w} \in W$ be the orthogonal projection of $\mathbf{v}$ onto W. Then $\mathbf{z} = \mathbf{v} - \mathbf{w}$ is, by definition, orthogonal to W and hence belongs to $W^\perp$. Note that $\mathbf{z}$ can be viewed as the orthogonal projection of $\mathbf{v}$ onto the complementary subspace $W^\perp$. If we are given two such decompositions, $\mathbf{v} = \mathbf{w} + \mathbf{z} = \widetilde{\mathbf{w}} + \widetilde{\mathbf{z}}$, then $\mathbf{w} - \widetilde{\mathbf{w}} = \widetilde{\mathbf{z}} - \mathbf{z}$. The left hand side of this equation lies in W while the right hand side belongs to $W^\perp$. But, as we already noted, the only vector that belongs to both W and $W^\perp$ is the zero vector. Thus, $\mathbf{w} = \widetilde{\mathbf{w}}$ and $\mathbf{z} = \widetilde{\mathbf{z}}$, which proves uniqueness. *Q.E.D.*

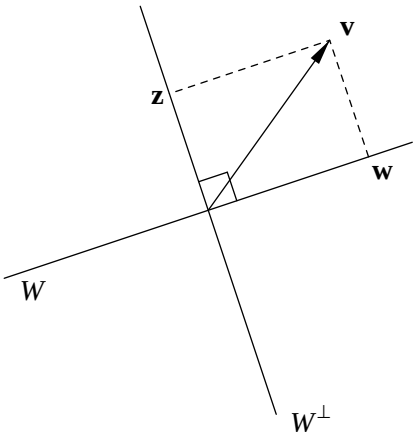


Figure 5.8. Orthogonal Decomposition of a Vector.

As a direct consequence of Exercise ∎, we conclude that a subspace and its orthogonal complement have complementary dimensions:

Proposition 5.46. *If* $\dim W = m$ *and* $\dim V = n$, *then* $\dim W^\perp = n - m$.

Example 5.47. Return to the situation described in Example 5.44. Let us decompose the vector $\mathbf{v} = (\,1, 0, 0\,)^T \in \mathbb{R}^3$ into a sum $\mathbf{v} = \mathbf{w} + \mathbf{z}$ of a vector $\mathbf{w}$ lying in the line W and a vector $\mathbf{z}$ belonging to its orthogonal plane $W^\perp$, defined by (5.69). Each is obtained by an orthogonal projection onto the subspace in question, but we only need to compute one of the two directly since the second can be obtained by subtracting from $\mathbf{v}$.

Orthogonal projection onto a one-dimensional subspace is easy since any basis is, trivially, an orthogonal basis. Thus, the projection of $\mathbf{v}$ onto the line spanned by $\mathbf{w}_1 = (\,1, 2, 3\,)^T$ is $\mathbf{w} = \| \mathbf{w}_1 \|^{-2} \langle\, \mathbf{v} \,;\, \mathbf{w}_1 \,\rangle\, \mathbf{w}_1 = \left(\, \frac{1}{14}, \frac{2}{14}, \frac{3}{14} \,\right)^T$. The component in $W^\perp$ is then obtained by subtraction: $\mathbf{z} = \mathbf{v} - \mathbf{w} = \left(\, \frac{13}{14}, -\frac{2}{14}, -\frac{3}{14} \,\right)^T$. Alternatively, one can obtain $\mathbf{z}$ directly by orthogonal projection onto the plane $W^\perp$. You need to be careful: the basis derived in Example 5.44 is not orthogonal, and so you will need to set up and solve the normal equations to find the closest point $\mathbf{z}$. Or, you can first convert the basis into an orthogonal basis by a single Gram–Schmidt step, and then use the orthogonal projection formula (5.57). All three methods lead to the same vector $\mathbf{z} \in W^\perp$.

Example 5.48. Let $W \subset \mathbb{R}^4$ be the two-dimensional subspace spanned by the orthogonal vectors $\mathbf{w}_1 = (\,1, 1, 0, 1\,)^T$ and $\mathbf{w}_2 = (\,1, 1, 1, -2\,)^T$. Its orthogonal complement $W^\perp$ (with respect to the Euclidean dot product) is the set of all vectors $\mathbf{v} = (\,x, y, z, w\,)^T$ that satisfy the linear system

$$\mathbf{v} \cdot \mathbf{w}_1 = x + y + w = 0, \qquad \mathbf{v} \cdot \mathbf{w}_2 = x + y + z - 2\,w = 0.$$

Applying the usual algorithm — the free variables are y and w — we find that the solution space is spanned by $\mathbf{z}_1 = (\,-1, 1, 0, 0\,)^T, \mathbf{z}_2 = (\,-1, 0, 3, 1\,)^T$, which form a non-orthogonal basis for $W^\perp$.

The orthogonal basis $\mathbf{y}_1 = \mathbf{z}_1 = (\,-1, 1, 0, 0\,)^T$ and $\mathbf{y}_2 = \mathbf{z}_2 - \frac{1}{2}\,\mathbf{z}_1 = \left(\, -\frac{1}{2}, -\frac{1}{2}, 3, 1 \,\right)^T$ for $W^\perp$ is obtained by a single Gram–Schmidt step. To decompose the vector $\mathbf{v} =$

$(\,1,0,0,0\,)^T = \mathbf{w} + \mathbf{z}$, say, we compute the two orthogonal projections: $\mathbf{w} = \frac{1}{3}\,\mathbf{w}_1 + \frac{1}{7}\,\mathbf{w}_2 = \left(\,\frac{10}{21}, \frac{10}{21}, \frac{1}{7}, \frac{1}{21}\,\right)^T \in W$, and $\mathbf{z} = -\frac{1}{2}\,\mathbf{y}_1 - \frac{1}{21}\,\mathbf{y}_2 = \left(\,\frac{11}{21}, -\frac{10}{21}, -\frac{1}{7}, -\frac{1}{21}\,\right)^T \in W^\perp$. Or you can easily obtain $\mathbf{z} = \mathbf{v} - \mathbf{w}$ by subtraction.

Proposition 5.49. *If W is a finite-dimensional subspace of an inner product space, then $(W^\perp)^\perp = W$.*

This result is an immediate corollary of the orthogonal decomposition Proposition 5.45. *Warning*: Propositions 5.45 and 5.49 are *not* necessarily true for infinite-dimensional vector spaces. In general, if $\dim W = \infty$, one can only assert that $W \subseteq (W^\perp)^\perp$. For example, it can be shown that, [**125**], on any bounded interval $[\,a,b\,]$ the orthogonal complement to the subspace of all polynomials $\mathcal{P}^{(\infty)} \subset \mathrm{C}^0[\,a,b\,]$ with respect to the L^2 inner product is trivial: $(\mathcal{P}^{(\infty)})^\perp = \{0\}$. This means that the only continuous function which satisfies the *moment equations*

$$\langle\, x^n\,;f(x)\,\rangle = \int_a^b x^n f(x)\,dx = 0, \qquad \text{for all} \qquad n = 0,1,2,\ldots$$

is the zero function $f(x) \equiv 0$. But the orthogonal complement of $\{0\}$ is the entire space, and so $((\mathcal{P}^{(\infty)})^\perp)^\perp = \mathrm{C}^0[\,a,b\,] \neq \mathcal{P}^{(\infty)}$.

The difference is that, in infinite-dimensional function space, a proper subspace $W \subsetneq V$ can be *dense*[†], whereas in finite dimensions, every proper subspace is a "thin" subset that only occupies an infinitesimal fraction of the entire vector space. This seeming paradox underlies the success of numerical methods, such as the finite element method, in approximating functions by elements of a subspace.

Orthogonality of the Fundamental Matrix Subspaces and the Fredholm Alternative

In Chapter 2, we introduced the four fundamental subspaces associated with an $m \times n$ matrix A. According to the fundamental Theorem 2.47, the first two, the kernel or null space and the corange or row space, are subspaces of $\mathbb{R}^n$ having complementary dimensions. The second two, the cokernel or left null space and the range or column space, are subspaces of $\mathbb{R}^m$, also of complementary dimensions. In fact, more than this is true — the subspace pairs are orthogonal complements with respect to the standard Euclidean dot product!

Theorem 5.50. *Let A be an $m \times n$ matrix of rank r. Then its kernel and corange are orthogonal complements as subspaces of $\mathbb{R}^n$, of respective dimensions $n - r$ and r, while its cokernel and range are orthogonal complements in $\mathbb{R}^m$, of respective dimensions $m - r$ and r:*

$$\ker A = (\operatorname{corng} A)^\perp \subset \mathbb{R}^n, \qquad \operatorname{coker} A = (\operatorname{rng} A)^\perp \subset \mathbb{R}^m. \tag{5.70}$$

Figure 5.9 illustrates the geometric configuration of (5.70).

[†] In general, a subset $W \subset V$ of a normed vector space is *dense* if, for every $\mathbf{v} \in V$, there are elements $\mathbf{w} \in W$ that are arbitrarily close, $\| \mathbf{v} - \mathbf{w} \| < \varepsilon$ for every $\varepsilon > 0$. The Weierstrass approximation theorem, [**126**], tells us that the polynomials form a dense subspace of the space of continuous functions, and underlies the proof of the result mentioned in the preceding paragraph.

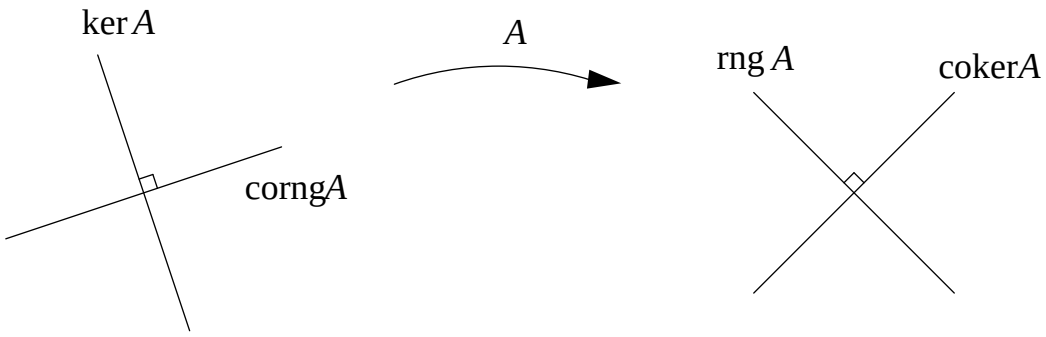


Figure 5.9. The Fundamental Matrix Subspaces.

Proof: A vector $\mathbf{x} \in \mathbb{R}^n$ lies in $\ker A$ if and only if $A\,\mathbf{x} = \mathbf{0}$. According to the rules of matrix multiplication, the i^{th} entry of $A\,\mathbf{x}$ equals the product of the i^{th} row $\mathbf{r}_i^T$ of A and $\mathbf{x}$. But this product vanishes, $\mathbf{r}_i^T\,\mathbf{x} = \mathbf{r}_i \cdot \mathbf{x} = 0$, if and only if $\mathbf{x}$ is orthogonal to $\mathbf{r}_i$. Therefore, $\mathbf{x} \in \ker A$ if and only if $\mathbf{x}$ is orthogonal to all the rows of A. Since the rows span $\operatorname{corng} A = \operatorname{rng} A^T$, this is equivalent to the statement that $\mathbf{x}$ lies in the orthogonal complement $(\operatorname{corng} A)^{\perp}$, which proves the first statement. The proof for the range and cokernel follows from the same argument applied to the transposed matrix A^T. *Q.E.D.*

Combining Theorems 2.47 and 5.50, we deduce the following important characterization of compatible linear systems, known as the *Fredholm alternative*. The Swedish mathematician Ivar Fredholm's main interest was in solving linear integral equations, but his compatibility criterion is also applicable to linear matrix systems, as well as linear differential equations, linear variational problems, and many other linear systems.

Theorem 5.51. *The linear system $A\,\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b}$ is orthogonal to the cokernel of A.*

Indeed, the linear system has a solution if and only if the right hand side $\mathbf{b} \in \operatorname{rng} A$ belongs to the range of A, which, by (5.70), requires that $\mathbf{b}$ be orthogonal to the cokernel $\operatorname{coker} A$. Therefore, the compatibility conditions for the linear system $A\,\mathbf{x} = \mathbf{b}$ can be written in the form

$$\mathbf{y} \cdot \mathbf{b} = 0 \qquad \text{for every } \mathbf{y} \text{ satisfying} \qquad A^T\mathbf{y} = \mathbf{0}. \tag{5.71}$$

Or, to state in another way, the vector $\mathbf{b}$ is a linear combination of the columns of A if and only if it is orthogonal to every vector $\mathbf{y}$ in the cokernel of A. In practice, one only needs to check orthogonality of $\mathbf{b}$ with respect to a basis $\mathbf{y}_1, \dots, \mathbf{y}_{m-r}$ of the cokernel, leading to a system of $m - r$ compatibility constraints, where $r = \operatorname{rank} A$ denotes the rank of the coefficient matrix. We note that $m - r$ is also the number of all zero rows in the row echelon form of A, and hence yields precisely the same number of constraints on the right hand side $\mathbf{b}$.

Example 5.52. In Example 2.40, we analyzed the linear system $A\,\mathbf{x} = \mathbf{b}$ with coefficient matrix $A = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -2 \\ 1 & -2 & 3 \end{pmatrix}$. Using direct Gaussian elimination, we were led to

a single compatibility condition, namely $-b_1+2b_2+b_3=0$, required for the system to have a solution. We now understand the meaning behind this equation: it is telling us that the right hand side $\mathbf{b}$ must be orthogonal to the cokernel of A. The cokernel is determined by solving the homogeneous adjoint system $A^T\mathbf{y}=\mathbf{0}$, and is the line spanned by the vector $\mathbf{y}_1=(-1,2,1)^T$. Thus, the compatibility condition requires that $\mathbf{b}$ be orthogonal to $\mathbf{y}_1$, in accordance with the Fredholm Theorem 5.51.

Example 5.53. Let us determine the compatibility conditions for the linear system

$$x_1-x_2+3x_3=b_1,\quad -x_1+2x_2-4x_3=b_2,\quad 2x_1+3x_2+x_3=b_3,\quad x_1+2x_3=b_4,$$

by computing the cokernel of its coefficient matrix $A=\begin{pmatrix} 1 & -1 & 3\\ -1 & 2 & -4\\ 2 & 3 & 1\\ 1 & 0 & 2\end{pmatrix}$. To this end, we need to solve the homogeneous adjoint system $A^T\mathbf{y}=\mathbf{0}$, namely

$$y_1-y_2+2y_3+y_4=0,\qquad -y_1+2y_2+3y_3=0,\qquad 3y_1-4y_2+y_3+2y_4=0.$$

Using Gaussian elimination, we find the general solution

$$\mathbf{y}=y_3\,(-7,-5,1,0)^T+y_4\,(-2,-1,0,1)^T$$

is a linear combination (whose coefficients are the free variables) of the two basis vectors for $\operatorname{coker}A$. Thus, the compatibility conditions are obtained by taking their dot products with the right hand side of the original system:

$$-7b_1-5b_2+b_3=0,\qquad -2b_1-b_2+b_4=0.$$

The reader can check that these are indeed the same compatibility conditions that result from a direct Gaussian elimination on the augmented matrix $\big(A\mid\mathbf{b}\big)$.

We are now very close to a full understanding of the fascinating geometry that lurks behind the simple algebraic operation of multiplying a vector $\mathbf{x}\in\mathbb{R}^n$ by an $m\times n$ matrix, resulting in a vector $\mathbf{b}=A\mathbf{x}\in\mathbb{R}^m$. Since the kernel and corange of A are orthogonal complementary subspaces in the domain space $\mathbb{R}^n$, Proposition 5.46 tells us that we can uniquely decompose $\mathbf{x}=\mathbf{w}+\mathbf{z}$ where $\mathbf{w}\in\operatorname{corng}A$, while $\mathbf{z}\in\ker A$. Since $A\mathbf{z}=\mathbf{0}$, we have

$$\mathbf{b}=A\mathbf{x}=A(\mathbf{w}+\mathbf{z})=A\mathbf{w}.$$

Therefore, we can regard multiplication by A as a combination of two operations:

(i) The first is an orthogonal projection onto the subspace $\operatorname{corng}A$ taking $\mathbf{x}$ to $\mathbf{w}$.

(ii) The second takes a vector in $\operatorname{corng}A\subset\mathbb{R}^n$ to a vector in $\operatorname{rng}A\subset\mathbb{R}^m$, taking the orthogonal projection $\mathbf{w}$ to the image vector $\mathbf{b}=A\mathbf{w}=A\mathbf{x}$.

Moreover, if A has rank r then, according to Theorem 2.47, both $\operatorname{rng}A$ and $\operatorname{corng}A$ are r-dimensional subspaces, albeit of different vector spaces. Each vector $\mathbf{b}\in\operatorname{rng}A$ corresponds to a unique vector $\mathbf{w}\in\operatorname{corng}A$. Indeed, if $\mathbf{w},\widetilde{\mathbf{w}}\in\operatorname{corng}A$ satisfy $\mathbf{b}=A\mathbf{w}=A\widetilde{\mathbf{w}}$, then $A(\mathbf{w}-\widetilde{\mathbf{w}})=\mathbf{0}$ and hence $\mathbf{w}-\widetilde{\mathbf{w}}\in\ker A$. But, since they are complementary subspaces, the only vector that belongs to both the kernel and the corange is the zero vector, and hence $\mathbf{w}=\widetilde{\mathbf{w}}$. In this manner, we have proved the first part of the following result; the second is left as Exercise ∎.

Proposition 5.54. *Multiplication by an $m \times n$ matrix A of rank r defines a one-to-one correspondence between the r-dimensional subspaces* $\operatorname{corng} A \subset \mathbb{R}^n$ *and* $\operatorname{rng} A \subset \mathbb{R}^m$. *Moreover, if $\mathbf{v}_1, \dots, \mathbf{v}_r$ forms a basis of* $\operatorname{corng} A$ *then their images $A\mathbf{v}_1, \dots, A\mathbf{v}_r$ form a basis for* $\operatorname{rng} A$.

In summary, the linear system $A\mathbf{x} = \mathbf{b}$ has a solution if and only if $\mathbf{b} \in \operatorname{rng} A$, or, equivalently, is orthogonal to every vector $\mathbf{y} \in \operatorname{coker} A$. If the compatibility conditions hold, then the system has a *unique* solution $\mathbf{w} \in \operatorname{corng} A$ that, by the definition of the corange or row space, is a linear combination of the *rows* of A. The general solution to the system is $\mathbf{x} = \mathbf{w} + \mathbf{z}$ where $\mathbf{w}$ is the particular solution belonging to the corange, while $\mathbf{z} \in \ker A$ is an arbitrary element of the kernel.

Theorem 5.55. *A compatible linear system $A\mathbf{x} = \mathbf{b}$ with $\mathbf{b} \in \operatorname{rng} A = (\operatorname{coker} A)^\perp$ has a unique solution $\mathbf{w} \in \operatorname{corng} A$ with $A\mathbf{w} = \mathbf{b}$. The general solution is $\mathbf{x} = \mathbf{w} + \mathbf{z}$ where $\mathbf{z} \in \ker A$. The particular solution is distinguished by the fact that it has minimum Euclidean norm $\| \mathbf{w} \|$ among all possible solutions.*

Indeed, since the corange and kernel are orthogonal subspaces, the norm of a general solution $\mathbf{x} = \mathbf{w} + \mathbf{z}$ is

$$\| \mathbf{x} \|^2 = \| \mathbf{w} + \mathbf{z} \|^2 = \| \mathbf{w} \|^2 + 2\, \mathbf{w} \cdot \mathbf{z} + \| \mathbf{z} \|^2 = \| \mathbf{w} \|^2 + \| \mathbf{z} \|^2 \geq \| \mathbf{w} \|^2,$$

with equality if and only if $\mathbf{z} = \mathbf{0}$.

Example 5.56. Consider the linear system

$$\begin{pmatrix} 1 & -1 & 2 & -2 \\ 0 & 1 & -2 & 1 \\ 1 & 3 & -5 & 2 \\ 5 & -1 & 9 & -6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 4 \\ 6 \end{pmatrix}.$$

Applying the standard Gaussian elimination algorithm, we discover that the coefficient matrix has rank 3, and the kernel is spanned by the single vector $\mathbf{z}_1 = (\, 1, -1, 0, 1 \,)^T$. The system itself is compatible; indeed, the right hand side is orthogonal to the basis cokernel vector $(\, 2, 24, -7, 1 \,)^T$, and so satisfies the Fredholm alternative.

The general solution to the linear system is $\mathbf{x} = (\, t, 3 - t, 1, t \,)^T$ where $t = w$ is the free variable. We decompose the solution $\mathbf{x} = \mathbf{w} + \mathbf{z}$ into a vector $\mathbf{w}$ in the corange and an element $\mathbf{z}$ in the kernel. The easiest way to do this is to first compute its orthogonal projection $\mathbf{z} = \| \mathbf{z}_1 \|^{-2} \mathbf{x} \cdot \mathbf{z}_1 \, \mathbf{z}_1 = (\, t - 1, 1 - t, 0, t - 1 \,)^T$ of the solution $\mathbf{x}$ onto the one-dimensional kernel. We conclude that $\mathbf{w} = \mathbf{x} - \mathbf{z} = (\, 1, 2, 1, 1 \,)^T \in \operatorname{corng} A$ is the unique solution belonging to the corange of the coefficient matrix, i.e., the only solution that can be written as a linear combination of its row vectors, or, equivalently, the only solution which is orthogonal to the kernel. The reader should check this by finding the coefficients in the linear combination, or, equivalently, writing $\mathbf{w} = A^T \mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^4$.

In this example, the analysis was simplified by the fact that the kernel was one-dimensional, and hence the orthogonal projection was relatively easy to compute. In more complicated situations, to determine the decomposition $\mathbf{x} = \mathbf{w} + \mathbf{z}$ one needs to solve the

normal equations (4.26) in order to find the orthogonal projection or least squares point in the subspace; alternatively, one can first determine an orthogonal basis for the subspace, and then apply the orthogonal (or orthonormal) projection formula (5.57). Of course, once one of the constituents $\mathbf{w}, \mathbf{z}$ has been found, the other can be simply obtained by subtraction from $\mathbf{x}$.

Chapter 6

Equilibrium

In this chapter, we turn to some interesting applications of linear algebra to the analysis of mechanical structures and electrical circuits. We will discover that there are remarkable analogies between electrical and mechanical systems. Both fit into a very general mathematical framework which, when suitably formulated, will also apply in the continuous realm, and ultimately governs the equilibria of systems arising throughout physics and engineering. The one difference is that discrete structures and circuits are governed by linear algebraic equations on finite-dimensional vector spaces, whereas continuous media are modeled by differential equations and boundary value problems on infinite-dimensional function spaces.

We begin by analyzing in detail a linear chain of masses interconnected by springs and constrained to move only in the longitudinal direction. Our general mathematical framework is already manifest in this rather simple mechanical structure. Next, we consider simple electrical circuits consisting of resistors and current sources interconnected by a network of wires. Finally, we treat small (so as to remain in a linear regime) displacements of two and three-dimensional structures constructed out of elastic bars. In all cases, we only consider the equilibrium configurations; dynamical processes for each of the physical systems will be taken up in Chapter 9.

In the mechanical and electrical systems treated in the present chapter, the linear system governing the equilibrium configuration has the same structure: the coefficient matrix is of general positive (semi-)definite Gram form. The positive definite cases correspond to stable structures and circuits, which can support any external forcing, and possess a unique stable equilibrium solution that can be characterized by a minimization principle. On the other hand, the positive semi-definite cases correspond to unstable structures and circuits that cannot remain in equilibrium except for very special configurations of external forces. In the case of mechanical structures, the instabilities are of two types: rigid motions, under which the structure maintains its overall absence of any applied force.

6.1. Springs and Masses.

A *mass–spring chain* consists of n masses m_1, m_2, $\dots$ m_n arranged in a straight line. Each mass is connected to its immediate neighbor(s) by a spring. Moreover, the mass–spring chain may be connected at one or both ends by a spring to a solid support. At first, for specificity, let us look at the case when both ends of the chain are attached, as illustrated in Figure 6.1. To be definite, we assume that the masses are arranged in a vertical line, and order them from top to bottom. On occasion, we may refer to the top support as mass m_0 and the bottom support as mass m_{n+1}. For simplicity, we will only

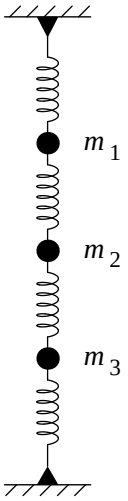


Figure 6.1. A Mass–Spring Chain with Fixed Ends.

allow the masses to move in the vertical direction — one-dimensional motion. (Section 6.3 deals with the more complicated cases of two- and three-dimensional motion.)

If we subject some or all of the masses to an external force, e.g., gravity, then the system will move[†] to a new equilibrium position. The motion of the i^{th} mass is measured by its *displacement* u_i from its original position, which, since we are only allowing vertical motion, is a scalar. Referring to Figure 6.1, we use the convention that $u_i > 0$ if the mass has moved downwards, and $u_i < 0$ if it has moved upwards. The problem is to determine the new equilibrium configuration of the chain under the prescribed forcing, that is, to set up and solve a system of equations for the displacements $u_1, \dots, u_n$.

Let e_j denote the *elongation* of the j^{th} spring, which connects mass m_{j-1} to mass m_j. By "elongation", we mean how far the spring has been stretched, so that $e_j > 0$ if the spring is longer than its reference length, while $e_j < 0$ if the spring has been compressed. The elongations can be determined directly from the displacements according to the geometric formula

$$e_j = u_j - u_{j-1}, \qquad j = 2, \dots, n, \tag{6.1}$$

while

$$e_1 = u_1, \qquad e_{n+1} = -u_n, \tag{6.2}$$

since the top and bottom supports are fixed. We write the elongation equations (6.1), (6.2) in matrix form

$$\mathbf{e} = A\,\mathbf{u}, \tag{6.3}$$

[†] The differential equations governing its dynamical behavior will be the subject of Chapter 9. Damping or frictional effects will cause the system to eventually settle down into a stable equilibrium configuration.

where $\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n+1} \end{pmatrix}$ is the *elongation vector*, $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$ is the *displacement vector*, and the coefficient matrix

$$A = \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \\ & & & & & -1 \end{pmatrix} \tag{6.4}$$

has size $(n+1) \times n$, with only the non-zero entries being indicated. The matrix A is known as the *reduced incidence matrix*† for the mass–spring chain. It effectively encodes the underlying geometry of the mass–spring chain, including the boundary conditions at the top and the bottom.

The next step is to connect the elongation e_j experienced by the j^{th} spring to its internal force y_j. This is the basic *constitutive assumption*, that relates geometry to kinematics. In the present case, we shall assume that the springs are not stretched (or compressed) particularly far, and so obey *Hooke's Law*

$$y_j = c_j\, e_j, \tag{6.5}$$

named after the prolific seventeenth century English scientist and inventor Robert Hooke. The constant $c_j > 0$ measures the spring's *stiffness*. Hooke's Law says that force is proportional to elongation — the more you stretch a spring, the more internal force it experiences. A hard spring will have a large stiffness and so takes a large force to stretch, whereas a soft spring will have a small, but still positive, stiffness. We write (6.5) in matrix form

$$\mathbf{y} = C\,\mathbf{e}, \tag{6.6}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n+1} \end{pmatrix}, \qquad C = \begin{pmatrix} c_1 & & & \\ & c_2 & & \\ & & \ddots & \\ & & & c_{n+1} \end{pmatrix}$$

Note particularly that $C > 0$ is a diagonal, positive definite matrix.

Finally, the forces must balance if the system is to remain in equilibrium. Let f_i denote the external force on the i^{th} mass m_i. We also measure force in the downwards direction, so $f_i > 0$ means the force is pulling the i^{th} mass downwards. (In particular, gravity would induce a positive force on each mass.) The i^{th} mass is immediately below

† The connection with the incidence matrix of a graph will become evident in Section 6.2.

the i^{th} spring and above the $(i+1)^{\text{st}}$ spring. If the i^{th} spring is stretched, it will exert an upwards force on m_i, while if the $(i+1)^{\text{st}}$ spring is stretched, it will pull m_i downwards. Therefore, the balance of forces on m_i requires that

$$f_i = y_i - y_{i+1}. \tag{6.7}$$

The matrix form of the force balance law is

$$\mathbf{f} = A^T \mathbf{y} \tag{6.8}$$

where $\mathbf{f} = (f_1, \dots, f_n)^T$. The remarkable, and very general fact is that the force balance coefficient matrix

$$A^T = \begin{pmatrix} 1 & -1 & & & & & \\ & 1 & -1 & & & & \\ & & 1 & -1 & & & \\ & & & 1 & -1 & & \\ & & & & \ddots & \ddots & \\ & & & & & 1 & -1 \end{pmatrix} \tag{6.9}$$

is the *transpose* of the reduced incidence matrix (6.4) for the chain. This connection between geometry and force balance turns out to be very general, and is the reason underlying the positivity of the final coefficient matrix in the resulting system of equilibrium equations.

Summarizing, we have

$$\mathbf{e} = A\,\mathbf{u}, \qquad \mathbf{y} = C\,\mathbf{e}, \qquad \mathbf{f} = A^T\,\mathbf{y}. \tag{6.10}$$

These equations can be combined into a single linear system

$$K\,\mathbf{u} = \mathbf{f}, \qquad \text{where} \qquad K = A^T C\,A \tag{6.11}$$

is called the *stiffness matrix* associated with the entire mass–spring chain. The stiffness matrix K has the form of a Gram matrix (3.51) for the weighted inner product $\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle = \mathbf{v}^T C\,\mathbf{w}$ induced by the diagonal matrix of spring stiffnesses. Theorem 3.33 tells us that since A has linearly independent columns (which should be checked), and $C > 0$ is positive definite, then the stiffness matrix $K > 0$ is automatically positive definite. In particular, Theorem 3.38 guarantees that K is an invertible matrix, and hence the linear system (6.11) has a unique solution $\mathbf{u} = K^{-1}\mathbf{f}$. We can therefore conclude that the mass–spring chain assumes a unique equilibrium position.

In fact, in the particular case considered here,

$$K = \begin{pmatrix} c_1 + c_2 & -c_2 & & & & & \\ -c_2 & c_2 + c_3 & -c_3 & & & & \\ & -c_3 & c_3 + c_4 & -c_4 & & & \\ & & -c_4 & c_4 + c_5 & -c_5 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -c_{n-1} & c_{n-1} + c_n & -c_n \\ & & & & & -c_n & c_n + c_{n+1} \end{pmatrix} \tag{6.12}$$

has a very simple symmetric, tridiagonal form. As such, we can apply our tridiagonal solution algorithm of Section 1.7 to rapidly solve the system.

Example 6.1. Let us consider the particular case of $n = 3$ masses connected by identical springs with unit spring constant. Thus, $c_1 = c_2 = c_3 = c_4 = 1$ and $C = \operatorname{diag}(1,1,1,1) = \mathrm{I}$ is the 4×4 identity matrix. The 3×3 stiffness matrix is then

$$K = A^T A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

A straightforward Gaussian elimination produces the $K = L D L^T$ factorization

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

With this in hand, we can solve the basic equilibrium equations $K\mathbf{u} = \mathbf{f}$ by our basic forward and back substitution algorithm.

Suppose, for example, we pull the middle mass downwards with a unit force, so $f_2 = 1$ while $f_1 = f_3 = 0$. Then $\mathbf{f} = (\,0, 1, 0\,)^T$, and the solution to the equilibrium equations (6.11) is $\mathbf{u} = \left(\,\frac{1}{2}, 1, \frac{1}{2}\,\right)^T$, whose entries prescribe the mass displacements. Observe that all three masses have moved down, with the middle mass moving twice as far as the other two. The corresponding spring elongations and internal forces are obtained by matrix multiplication

$$\mathbf{y} = \mathbf{e} = A\mathbf{u} = \left(\,\tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, -\tfrac{1}{2}\,\right)^T.$$

Thus the top two springs are elongated, while the bottom two are compressed, all by an equal amount.

Similarly, if all the masses are equal, $m_1 = m_2 = m_3 = m$, then the solution under a constant downwards gravitational force $\mathbf{f} = (\,mg, mg, mg\,)^T$ of magnitude g is

$$\mathbf{u} = K^{-1} \begin{pmatrix} mg \\ mg \\ mg \end{pmatrix} = \begin{pmatrix} \frac{3}{2}\,mg \\ 2\,mg \\ \frac{3}{2}\,mg \end{pmatrix},$$

and

$$\mathbf{y} = \mathbf{e} = A\mathbf{u} = \left(\,\tfrac{3}{2}\,mg, \quad \tfrac{1}{2}\,mg, \quad -\tfrac{1}{2}\,mg, \quad -\tfrac{3}{2}\,mg\,\right)^T.$$

Now, the middle mass has only moved 33% farther than the others, whereas the top and bottom spring are experiencing three times as much elongation/compression as the middle springs.

An important observation is that we *cannot* determine the internal forces $\mathbf{y}$ or elongations $\mathbf{e}$ directly from the force balance law (6.8) because the transposed matrix A^T is not square, and so the system $\mathbf{f} = A^T\mathbf{y}$ does not have a unique solution. We must first

determine the displacements $\mathbf{u}$ using the full equilibrium equations (6.11), and then use the resulting displacements to reconstruct the elongations and internal forces. This situation is referred to as *statically indeterminate*.

Remark: Even though we construct $K = A^T C\, A$ and then factor it as $K = L\, D\, L^T$, there is no direct algorithm to get from A and C to L and D, which, typically, are matrices of a different size.

The behavior of the system will depend upon both the forcing and the boundary conditions. Suppose, by way of contrast, that we only fix the top of the chain to a support, and leave the bottom mass hanging freely, as in Figure 6.2. The geometric relation between the displacements and the elongations has the same form (6.3) as before, but the reduced incidence matrix is slightly altered:

$$A = \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \end{pmatrix}. \tag{6.13}$$

This matrix has size $n \times n$ and is obtained from the preceding example (6.4) by eliminating the last row corresponding to the missing bottom spring. The constitutive equations are still governed by Hooke's law $\mathbf{y} = C\,\mathbf{e}$, as in (6.6), with $C = \operatorname{diag}(c_1, \dots, c_n)$ the $n \times n$ diagonal matrix of spring stiffnesses. Finally, the force balance equations are also found to have the same general form $\mathbf{f} = A^T \mathbf{y}$ as in (6.8), but with the transpose of the revised incidence matrix (6.13). In conclusion, the equilibrium equations $K\,\mathbf{x} = \mathbf{f}$ have an identical form (6.11), based on the revised stiffness matrix

$$K = A^T C A = \begin{pmatrix} c_1 + c_2 & -c_2 & & & & & \\ -c_2 & c_2 + c_3 & -c_3 & & & & \\ & -c_3 & c_3 + c_4 & -c_4 & & & \\ & & -c_4 & c_4 + c_5 & -c_5 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -c_{n-1} & c_{n-1} + c_n & -c_n \\ & & & & & -c_n & c_n \end{pmatrix} \tag{6.14}$$

Note that only the bottom right entry is different from the fixed end version (6.12). In contrast to the chain with two fixed ends, this system is called *statically determinate* because the incidence matrix A is square and nonsingular. This means that it is possible to solve the force balance law (6.8) directly for the internal forces $\mathbf{y} = A^{-1}\mathbf{f}$ without having to solve the full equilibrium equations for the displacements.

Example 6.2. For a three mass chain with one free end and equal unit spring

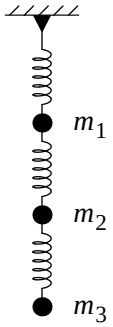


Figure 6.2. A Mass–Spring Chain with One Free End.

constants $c_1 = c_2 = c_3 = 1$, the stiffness matrix is

$$K = A^T A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Pulling the middle mass downwards with a unit force, whereby $\mathbf{f} = (\,0, 1, 0\,)^T$, results in the displacements

$$\mathbf{u} = K^{-1}\mathbf{f} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \qquad \text{so that} \qquad \mathbf{y} = \mathbf{e} = A\,\mathbf{u} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

In this configuration, the bottom two masses have moved equal amounts, and twice as far as the top mass. Because we are only pulling on the middle mass, the lower-most spring hangs free and experiences no elongation, whereas the top two springs are stretched by the same amount.

Similarly, for a chain of equal masses subject to a constant downwards gravitational force $\mathbf{f} = (\,mg, mg, mg\,)^T$, the equilibrium position is

$$\mathbf{u} = K^{-1} \begin{pmatrix} mg \\ mg \\ mg \end{pmatrix} = \begin{pmatrix} 3mg \\ 5mg \\ 6mg \end{pmatrix}, \qquad \text{and} \qquad \mathbf{y} = \mathbf{e} = A\,\mathbf{u} = \begin{pmatrix} 3mg \\ 2mg \\ mg \end{pmatrix}.$$

Note how much further the masses have moved now that the restraining influence of the bottom support has been removed. The top spring is experiencing the most elongation, and is thus the most likely to break, because it must support all three masses.

The Minimization Principle

According to Theorem 4.1, when the coefficient matrix of the linear system governing a mass–spring chain is positive definite, the unique equilibrium solution can be characterized by a minimization principle. The quadratic function to be minimized has a physical interpretation: it is the potential energy of the system. Nature is parsimonious when it comes to energy: physical systems seek out equilibrium configurations that minimize energy. This

general minimization principle can often be advantageously used in the construction of mathematical models, as well as in their solution, both analytical and numerical.

The energy function to be minimized can be determined directly from physical principles. For a mass–spring chain, the potential energy of the i^{th} mass equals the product of the applied force times its displacement: $-f_i\, u_i$. The minus sign is the result of our convention that a positive displacement $u_i > 0$ means that the mass has moved down, and hence decreased its potential energy. Thus, the total potential energy due to external forcing on all the masses in the chain is

$$-\sum_{i=1}^{n} f_i\, u_i \;=\; -\,\mathbf{u}^T\mathbf{f}.$$

Next, we calculate the internal energy of the system. The potential energy in a single spring elongated by an amount e is obtained by integrating the internal force, $y = ce$, leading to

$$\int_0^e y\, de = \int_0^e c\, e\, de = \tfrac{1}{2}\, c\, e^2.$$

Totalling the contributions from each spring, we find the internal spring energy to be

$$\frac{1}{2}\sum_{i=1}^{n} c_i\, e_i^2 = \tfrac{1}{2}\,\mathbf{e}^T C\,\mathbf{e} = \tfrac{1}{2}\,\mathbf{u}^T A^T C A\,\mathbf{u} = \tfrac{1}{2}\,\mathbf{u}^T K\,\mathbf{u},$$

where we used the incidence equation $\mathbf{e} = A\,\mathbf{u}$ relating elongation and displacement. Therefore, the total potential energy is

$$p(\mathbf{u}) = \tfrac{1}{2}\,\mathbf{u}^T K\,\mathbf{u} - \mathbf{u}^T\mathbf{f}. \tag{6.15}$$

Since $K > 0$, Theorem 4.1 implies that this quadratic function has a unique minimizer that satisfies the equilibrium equation $K\,\mathbf{u} = \mathbf{f}$.

Example 6.3. For a three mass chain with two fixed ends described in Example 6.1, the potential energy function (6.15) has the explicit form

$$\begin{aligned} p(\mathbf{u}) &= \frac{1}{2}\,(\,u_1\;u_2\;u_3\,)\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} - (\,u_1\;u_2\;u_3\,)\begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \\ &= u_1^2 - u_1\,u_2 + u_2^2 - u_2\,u_3 + u_3^2 - f_1\,u_1 - f_2\,u_2 - f_3\,u_3, \end{aligned}$$

where $\mathbf{f} = (\,f_1, f_2, f_3\,)^T$ is the external forcing. The minimizer of this particular quadratic function gives the equilibrium displacements $\mathbf{u} = (\,u_1, u_2, u_3\,)^T$ of the three masses.

6.2. Electrical Networks.

An electrical *network* consists of a collection of wires that are joined together at their ends. The junctions where one or more wires are connected are called *nodes*. Abstractly, we can view any such electrical network as a graph, the wires being the edges and the

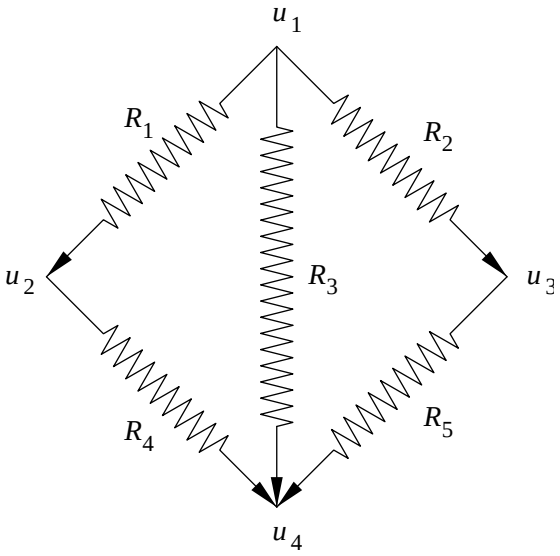


Figure 6.3. A Simple Electrical Network.

nodes the vertices. To begin with we assume that there are no electrical devices (batteries, inductors, capacitors, etc.) in the network and so the the only impediment to current flowing through the network is each wire's resistance. (If desired, we may add resistors to the network to increase the resistance along the wires.) As we shall see, resistance (or, rather, its reciprocal) plays a very similar role to spring stiffness.

We shall introduce current sources into the network at one or more of the nodes, and would like to determine how the induced current flows through the wires in the network. The basic equilibrium equations for the currents are the consequence of three fundamental laws of electricity.

Voltage is defined as the electromotive force that moves electrons through a wire. is induced by a drop in the voltage potential along the wire. The voltage in a wire is induced by the difference in the voltage potentials at the two ends, just as the gravitational force on a mass is induced by a difference in gravitational potential. To quantify voltage, we need to assign an orientation to the wire. Then a positive voltage means the electrons move in the assigned direction, while under a negative voltage they move in reverse. The original choice of orientation is arbitrary, but once assigned will pin down the sign conventions used by voltages, currents, etc. To this end, we draw a digraph to represent the network, and each edge or wire is assigned a direction that indicates its starting and ending vertices or nodes. A simple example is illustrated in Figure 6.3, and contains five wires joined at four different nodes. The arrows indicate the orientations of the wires, while the wavy lines are the standard electrical symbols for resistance.

In an electrical network, each node will have a voltage potential, denoted u_i. If wire k starts at node i and ends at node j, under its assigned orientation, then its voltage v_k equals the potential difference at its ends:

$$v_k = u_i - u_j. \tag{6.16}$$

Note that $v_k > 0$ if $u_i > u_j$, and so the electrons go from the starting node i to the ending

node j, in accordance with our choice of orientation. In our particular illustrative example,

$$v_1 = u_1 - u_2, \qquad v_2 = u_1 - u_3, \qquad v_3 = u_1 - u_4, \qquad v_4 = u_2 - u_4, \qquad v_5 = u_3 - u_4.$$

Let us rewrite this system in matrix form

$$\mathbf{v} = A\,\mathbf{u}, \tag{6.17}$$

where, for our particular example,

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{6.18}$$

The alert reader will recognize this matrix as the *incidence matrix* (2.42) for the digraph defined by the circuit; see (2.42). This is true in general — *the voltages along the wires of an electrical network are related to the potentials at the nodes by a linear system of the form* (6.17)*, where A is the incidence matrix of the network digraph.* The rows of the incidence matrix are indexed by the wires; the columns are indexed by the nodes. Each row of the matrix A has a single $+1$ in the column indexed by the starting node, and a single -1 in the column of the ending node.

Kirchhoff's Voltage Law states that the sum of the voltages around each closed loop in the network is zero. For example, in the circuit under consideration, around the left-hand triangle we have

$$v_1 + v_4 - v_3 = (u_1 - u_2) + (u_2 - u_4) - (u_1 - u_4) = 0.$$

Note that v_3 appears with a minus sign since we must traverse wire #3 in the opposite direction to its assigned orientation when going around the loop in the counterclockwise direction. The voltage law is a direct consequence of (6.17). Indeed, as discussed in Section 2.6, the loops can be identified with vectors $\ell \in \operatorname{coker} A = \ker A^T$ in the cokernel of the incidence matrix, and so

$$\ell \cdot \mathbf{v} = \ell^T \mathbf{v} = \ell^T A\,\mathbf{u} = 0. \tag{6.19}$$

Therefore, orthogonality of the voltage vector $\mathbf{v}$ to the loop vector ℓ is the mathematical formulation of the zero-loop relation.

Given a prescribed set of voltages $\mathbf{v}$ along the wires, can one find corresponding voltage potentials $\mathbf{u}$ at the nodes? To answer this question, we need to solve $\mathbf{v} = A\,\mathbf{u}$, which requires $\mathbf{v} \in \operatorname{rng} A$. According to the Fredholm Alternative Theorem 5.51, the necessary and sufficient condition for this to hold is that $\mathbf{v}$ be orthogonal to $\operatorname{coker} A$. Theorem 2.51 says that the cokernel of an incidence matrix is spanned by the loop vectors, and so $\mathbf{v}$ is a possible set of voltages if and only if $\mathbf{v}$ is orthogonal to all the loop vectors $\ell \in \operatorname{coker} A$, i.e., the Voltage Law is necessary and sufficient for the given voltages to be physically realizable in the network.

Kirchhoff's Laws are related to the topology of the circuit — how the different wires are connected together. *Ohm's Law* is a constitutive relation, indicating what the wires

are made of. The resistance along a wire, including any added resistors, prescribes the relation between voltage and current or the rate of flow of electric charge. The law reads

$$v_k = R_k\, y_k, \tag{6.20}$$

where v_k is the voltage and y_k (often denoted I_k in the engineering literature) denotes the current along wire k. Thus, for a fixed voltage, the larger the resistance of the wire, the smaller the current that flows through it. The direction of the current is also prescribed by our choice of orientation of the wire, so that $y_k > 0$ if the current is flowing from the starting to the ending node. We combine the individual equations (6.20) into a matrix form

$$\mathbf{v} = R\,\mathbf{y}, \tag{6.21}$$

where the *resistance matrix* $R = \operatorname{diag}(R_1, \dots, R_n) > 0$ is diagonal and positive definite. We shall, in analogy with (6.6), replace (6.21) by the inverse relationship

$$\mathbf{y} = C\,\mathbf{v}, \tag{6.22}$$

where $C = R^{-1}$ is the *conductance matrix*, again diagonal, positive definite, whose entries are the *conductances* $c_k = 1/R_k$ of the wires. For the particular circuit in Figure 6.3,

$$C = \begin{pmatrix} c_1 & & & & \\ & c_2 & & & \\ & & c_3 & & \\ & & & c_4 & \\ & & & & c_5 \end{pmatrix} = \begin{pmatrix} 1/R_1 & & & & \\ & 1/R_2 & & & \\ & & 1/R_3 & & \\ & & & 1/R_4 & \\ & & & & 1/R_5 \end{pmatrix}. \tag{6.23}$$

Finally, we stipulate that electric current is not allowed to accumulate at any node, i.e., every electron that arrives at a node must leave along one of the wires. Let $y_k, y_l, \dots, y_m$ denote the currents along all the wires $k, l, \dots, m$ that meet at node i in the network, and f_i an external current source, if any, applied at node i. *Kirchhoff's Current Law* requires that the net current into the node, namely

$$\pm\, y_k \,\pm\, y_l \,\pm\, \cdots \,\pm\, y_m + f_i = 0, \tag{6.24}$$

must be zero. Each $\pm$ sign is determined by the orientation of the wire, with $-$ if node i is a starting node or $+$ if it is an ending node.

In our particular example, suppose that we send a 1 amp current source into the first node. Then Kirchhoff's Current Law requires

$$y_1 + y_2 + y_3 = 1, \qquad -y_1 + y_4 = 0, \qquad -y_2 + y_5 = 0, \qquad -y_3 - y_4 - y_5 = 0.$$

Since we have solved (6.24) for the currents, the signs in front of the y_i have been reversed, with $+$ now indicating a starting node and $-$ an ending node. The matrix form of this system is

$$A^T \mathbf{y} = \mathbf{f}, \tag{6.25}$$

where $\mathbf{y} = (\, y_1, y_2, y_3, y_4, y_5 \,)^T$ are the currents along the five wires, and $\mathbf{f} = (\, 1, 0, 0, 0 \,)^T$ represents the current sources at the four nodes. The coefficient matrix

$$A^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & -1 \end{pmatrix}, \tag{6.26}$$

is the *transpose* of the incidence matrix (6.18). As in the mass–spring chain, this is a general fact, and is an immediate result of Kirchhoff's two laws. *The coefficient matrix for the current law is the transpose of the incidence matrix for the voltage law.*

Let us assemble the full system of equilibrium equations:

$$\mathbf{v} = A\,\mathbf{u}, \qquad \mathbf{y} = C\,\mathbf{v}, \qquad \mathbf{f} = A^T\mathbf{y}. \tag{6.27}$$

Remarkably, we arrive at a system of linear relations that has an identical form to the mass–spring chain system (6.10). As before, they combine into a single linear system

$$K\,\mathbf{u} = \mathbf{f}, \qquad \text{where} \qquad K = A^T C\,A \tag{6.28}$$

is the *resistivity matrix* associated with the given network. In our particular example, combining (6.18), (6.23), (6.26) produces the resistivity matrix

$$K = A^T C\,A = \begin{pmatrix} c_1 + c_2 + c_3 & -c_1 & -c_2 & -c_3 \\ -c_1 & c_1 + c_4 & 0 & -c_4 \\ -c_2 & 0 & c_2 + c_5 & -c_5 \\ -c_3 & -c_4 & -c_5 & c_3 + c_4 + c_5 \end{pmatrix} \tag{6.29}$$

depending on the conductances of the five wires in the network.

Remark: There is a simple pattern to the resistivity matrix, evident in (6.29). The diagonal entries k_{ii} equal the sum of the conductances of all the wires having node i at one end. The non-zero off-diagonal entries k_{ij}, $i \neq j$, equal $-c_k$, the conductance of the wire[†] joining node i to node j, while $k_{ij} = 0$ if there is no wire joining the two nodes.

Consider the case when all the wires in our network have equal unit resistance, and so $c_k = 1/R_k = 1$ for $k = 1, \ldots, 5$. Then the resistivity matrix is

$$K = \begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}. \tag{6.30}$$

However, trying to solve the system (6.28) runs into an immediate difficulty: *there is no solution*! The matrix (6.30) is *not* positive definite — it has zero determinant, and so is not invertible. Moreover, the particular current source vector $\mathbf{f} = (\, 1, 0, 0, 0 \,)^T$ does not lie in the range of K. Something is clearly amiss.

[†] This assumes that there is only one wire joining the two nodes.

Before getting discouraged, let us sit back and use a little physical intuition. We are trying to put a 1 amp current into the network at node 1. Where can the electrons go? The answer is nowhere — they are trapped in the circuit and, as they accumulate, something drastic will happen — sparks will fly! This is clearly an unstable situation, and so the fact that the equilibrium equations do not have a solution is trying to tell us that the physical system cannot remain in a steady state. The physics rescues the mathematics, or, vice versa, the mathematics elucidates the underlying physical processes!

In order to achieve a steady state in an electrical network, we must remove as much current as we put in. In other words, the sum of all the current sources must vanish:

$$f_1 + f_2 + \cdots + f_n = 0.$$

For example, if we feed a 1 amp current into node 1, then we must extract a total of 1 amp's worth of current from the other nodes. If we extract a 1 amp current from node 4, the modified current source vector $\mathbf{f} = (\,1, 0, 0, -1\,)^T$ does indeed lie in the range of K (check!) and the equilibrium system (6.28) has a solution. Fine ...

But we are not out of the woods yet. As we know, if a linear system has a singular square coefficient matrix, then either it has no solutions — the case we already rejected — or it has infinitely many solutions — the case we are considering now. In the particular network under consideration, the general solution to the linear system

$$\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

is found by Gaussian elimination:

$$\mathbf{u} = \begin{pmatrix} \frac{1}{2} + t \\ \frac{1}{4} + t \\ \frac{1}{4} + t \\ t \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \\ 0 \end{pmatrix} + t \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \tag{6.31}$$

where $t = u_4$ is the free variable. The nodal voltage potentials

$$u_1 = \tfrac{1}{2} + t, \qquad u_2 = \tfrac{1}{4} + t, \qquad u_3 = \tfrac{1}{4} + t, \qquad u_4 = t,$$

depend on a free parameter t.

The ambiguity arises because we have not specified a baseline value for the voltage potentials. Indeed, voltage potential is a mathematical abstraction that cannot be measured directly; only relative potential differences have physical import. To eliminate the ambiguity, one needs to assign a base potential level. (A similar ambiguity arises in the specification of gravitational potential.) In terrestrial electricity, the Earth is assumed to be at a zero voltage potential. Specifying a particular node to have zero potential is physically equivalent to grounding that node. Grounding one of the nodes, e.g., setting $u_4 = t = 0$, will then uniquely specify all the other voltage potentials, resulting in a unique solution $u_1 = \frac{1}{2}$, $u_2 = \frac{1}{4}$, $u_3 = \frac{1}{4}$, $u_4 = 0$, to the system.

On the other hand, even without specification of a baseline potential level, the corresponding voltages and currents along the wires are uniquely specified. In our example, computing $\mathbf{y} = \mathbf{v} = A\,\mathbf{u}$ gives

$$y_1 = v_1 = \tfrac{1}{4}\,, \qquad y_2 = v_2 = \tfrac{1}{4}\,, \qquad y_3 = v_3 = \tfrac{1}{2}\,, \qquad y_4 = v_4 = \tfrac{1}{4}\,, \qquad y_5 = v_5 = \tfrac{1}{4}\,,$$

independent of the value of t in (6.31). Thus, the nonuniqueness of the voltage potential solution $\mathbf{u}$ is not an essential difficulty. All physical quantities that we can measure — currents and voltages — *are* uniquely specified by the solution to the equilibrium system.

Remark: Although they have no real physical meaning, we *cannot* dispense with the nonmeasurable (and non-unique) voltage potentials $\mathbf{u}$. Most circuits are *statically indeterminate* since their incidence matrix is rectangular and not invertible, and so the linear system $A^T\mathbf{y} = \mathbf{f}$ *cannot* be solved directly for the currents in terms of the voltage sources — it does not have a unique solution. Only by first solving the full equilibrium system (6.28) for the potentials, and then using the relation $\mathbf{y} = CA\,\mathbf{u}$ between the potentials and the currents, can we determine the actual values of the currents in our network.

Let us analyze what is going on in the context of our general mathematical framework. Proposition 3.32 says that the resistivity matrix $K = A^TCA$ is positive definite (and hence nonsingular) provided A has linearly independent columns, or, equivalently, $\ker A = \{\mathbf{0}\}$. But Proposition 2.49 says that the incidence matrix A of a directed graph *never* has a trivial kernel. Therefore, the resistivity matrix K is only positive semi-definite, and hence singular. If the network is connected, then $\ker A = \ker K = \operatorname{coker} K$ is one-dimensional, spanned by the vector $\mathbf{z} = (\,1, 1, 1, \ldots, 1\,)^T$. According to the Fredholm Alternative Theorem 5.51, the fundamental network equation $K\,\mathbf{u} = \mathbf{f}$ has a solution if and only if $\mathbf{f}$ is orthogonal to $\operatorname{coker} K$, and so the current source vector must satisfy

$$\mathbf{f} \cdot \mathbf{z} = f_1 + f_2 + \cdots + f_n = 0, \tag{6.32}$$

as we already observed. Therefore, the linear algebra reconfirms our physical intuition: a connected network admits an equilibrium configuration, obtained by solving (6.28), if and only if the nodal current sources add up to zero, i.e., there is no net influx of current into the network.

Grounding one of the nodes is equivalent to nullifying the value of its voltage potential: $u_i = 0$. This variable is now fixed, and can be safely eliminated from our system. To accomplish this, we let $A^\star$ denote the $m \times (n-1)$ matrix obtained by deleting the i^{th} column from A. For example, if we ground node number 4 in our sample network, then we erase the fourth column of the incidence matrix (6.18), leading to the *reduced incidence matrix*

$$A^\star = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{6.33}$$

The key observation is that $A^\star$ has trivial kernel, $\ker A^\star = \{\mathbf{0}\}$, and therefore the reduced network resistivity matrix

$$K^\star = (A^\star)^T C\, A^\star = \begin{pmatrix} c_1 + c_2 + c_3 & -c_1 & -c_2 \\ -c_1 & c_1 + c_4 & 0 \\ -c_2 & 0 & c_2 + c_5 \end{pmatrix}. \tag{6.34}$$

is positive definite. Note that we can obtain $K^\star$ directly from K by deleting both its fourth row and fourth column. Let $\mathbf{f}^\star = (\,1, 0, 0\,)^T$ denote the reduced current source vector obtained by deleting the fourth entry from $\mathbf{f}$. Then the reduced linear system is

$$K^\star \mathbf{u}^\star = \mathbf{f}^\star, \qquad \text{where} \qquad \mathbf{u}^\star = (\,u_1, u_2, u_3\,)^T, \tag{6.35}$$

is the reduced voltage potential vector. Positive definiteness of $K^\star$ implies that (6.35) has a unique solution $\mathbf{u}^\star$, from which we can reconstruct the voltages $\mathbf{v} = A^\star \mathbf{u}^\star$ and currents $\mathbf{y} = C\,\mathbf{v} = CA^\star \mathbf{u}^\star$ along the wires. In our example, if all the wires have unit resistance, then the reduced system (6.35) is

$$\begin{pmatrix} 3 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

and has unique solution $\mathbf{u}^\star = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}^T$. The voltage potentials are

$$u_1 = \tfrac{1}{2}, \qquad u_2 = \tfrac{1}{4}, \qquad u_3 = \tfrac{1}{4}, \qquad u_4 = 0,$$

and correspond to the earlier solution (6.31) when $t = 0$. The corresponding voltages and currents along the wires are the same as before.

So far, we have only considered the effect of current sources at the nodes. Suppose now that the circuit contains one or more batteries. Each battery serves as a voltage source along one of the wires, and we let b_k denote the voltage of a battery connected to wire k. The quantity b_k comes with a sign, indicated by the battery's positive and negative terminals. Our convention is that $b_k > 0$ if the current from the battery runs in the same direction as our chosen orientation of the wire. The battery voltage modifies the voltage balance equation (6.16):

$$v_k = u_i - u_j + b_k.$$

The corresponding matrix form (6.17) becomes

$$\mathbf{v} = A\,\mathbf{u} + \mathbf{b}, \tag{6.36}$$

where $\mathbf{b} = (\,b_1, b_2, \ldots, b_m\,)^T$ is the *battery vector* whose entries are indexed by the wires. (If there is no battery on wire k, the corresponding entry is $b_k = 0$.) The remaining two equations are as before, so $\mathbf{y} = C\,\mathbf{v}$ are the currents in the wires, and, in the absence of external current sources, Kirchhoff's Current Law implies $A^T\mathbf{y} = \mathbf{0}$. Using the modified formula (6.36) for the voltages, these combine into the following equilibrium system

$$K^\star \mathbf{u} = A^T C\, A\,\mathbf{u} = -\,A^T C\,\mathbf{b}. \tag{6.37}$$

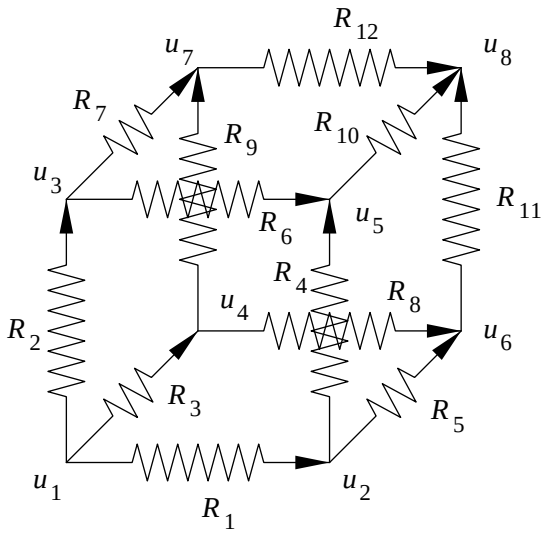


Figure 6.4. Cubical Electrical Network with a Battery.

Thus, interestingly, the voltage potentials satisfy the normal *weighted least squares* equations (4.54) corresponding to the system $A\,\mathbf{u} = -\,\mathbf{b}$, with weights given by the conductances in the individual wires in the circuit. It is a remarkable fact that Nature solves a least squares problem in order to make the weighted norm of the voltages $\mathbf{v}$ as small as possible.

Furthermore, the batteries have exactly the same effect on the voltage potentials as if we imposed the current source vector

$$\mathbf{f} = -\,A^T C\,\mathbf{b}. \tag{6.38}$$

Namely, the effect of the battery of voltage b_k on wire k is the exactly the same as introducing an additional current sources of $-\,c_k\,b_k$ at the starting node and $c_k\,b_k$ at the ending node. Note that the induced current vector $\mathbf{f} \in \operatorname{rng} K$ continues to satisfy the network constraint (6.32). Vice versa, a given system of current sources $\mathbf{f}$ has the same effect as any collection of batteries $\mathbf{b}$ that satisfies (6.38).

Unlike a current source, a circuit with a battery always admits a solution for the voltage potentials and currents. Although the currents are uniquely determined, the voltage potentials are not. As before, to eliminate the ambiguity, we can ground one of the nodes and use the reduced incidence matrix $A^\star$ and reduced current source vector $\mathbf{f}^\star$ obtained by eliminating the column/entry corresponding to the grounded node.

Example 6.4. Consider an electrical network running along the sides of a cube, where each wire contains a 2 ohm resistor and there is a 9 volt battery source on one wire. The problem is to determine how much current flows through the wire directly opposite the battery. Orienting the wires and numbering them as indicated in Figure 6.4, the incidence

matrix is

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

We connect the battery along wire #1 and measure the resulting current along wire #12. To avoid the ambiguity in the voltage potentials, we ground the last node and erase the final column from A to obtain the reduced incidence matrix $A^\star$. Since the resistance matrix R has all 2's along the diagonal, the conductance matrix is $C = \frac{1}{2}\,\mathrm{I}$. Therefore the network resistivity matrix is

$$K^\star = (A^\star)^T C A^\star = \tfrac{1}{2}\,(A^\star)^T A^\star = \frac{1}{2} \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 & -1 & -1 & 0 \\ -1 & 0 & 3 & 0 & -1 & 0 & -1 \\ -1 & 0 & 0 & 3 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 & 3 & 0 & 0 \\ 0 & -1 & 0 & -1 & 0 & 3 & 0 \\ 0 & 0 & -1 & -1 & 0 & 0 & 3 \end{pmatrix}.$$

The current source corresponding to the battery $\mathbf{b} = (\,9, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\,)^T$ along the first wire is

$$\mathbf{f}^\star = -\,(A^\star)^T C\,\mathbf{b} = \tfrac{1}{2}\,\,(\,-9, 9, 0, 0, 0, 0, 0\,)^T.$$

Solving the resulting linear system by Gaussian elimination, the voltage potentials are

$$\mathbf{u}^\star = (K^\star)^{-1}\mathbf{f}^\star = \big(\,-3, \tfrac{9}{4}, -\tfrac{9}{8}, -\tfrac{9}{8}, \tfrac{3}{8}, \tfrac{3}{8}, -\tfrac{3}{4}\,\big)^T.$$

Thus, the induced currents along the sides of the cube are

$$\mathbf{y} = C\,\mathbf{v} = C\,(A^\star\mathbf{u}^\star + \mathbf{b}) = \big(\,\tfrac{15}{8}, -\tfrac{15}{16}, -\tfrac{15}{16}, \tfrac{15}{16}, \tfrac{15}{16}, -\tfrac{3}{4}, -\tfrac{3}{16}, -\tfrac{3}{4}, -\tfrac{3}{16}, \tfrac{3}{16}, \tfrac{3}{16}, -\tfrac{3}{8}\,\big)^T.$$

In particular, the current on the wire that is opposite the battery is $y_{12} = -\frac{3}{8}$, flowing in the opposite direction to its orientation. The most current flows through the battery wire, while wires $7, 9, 10$ and 11 transmit the least current.

The Minimization Principle and the Electrical–Mechanical Analogy

As with a mass–spring chain, the current flows in such a resistive electrical network can be characterized by a minimization principle. The *power* in a wire is defined as the

product of its current y and voltage v,

$$P = y\,v = R\,y^2 = c\,v^2, \tag{6.39}$$

where R is the resistance, $c = 1/R$ the conductance, and we are using Ohm's Law (6.20) to relate voltage and current. Physically, the power tells us the rate at which electrical energy is converted into heat or energy by the resistance along the wire.

Summing over all the wires in the network, the total power is the dot product

$$\begin{aligned} P = \sum_{k=1}^{m} y_k\, v_k = \mathbf{y}^T\mathbf{v} = \mathbf{v}^T C\,\mathbf{v} = (A\,\mathbf{u} + \mathbf{b})^T C\,(A\,\mathbf{u} + \mathbf{b}) \\ = \mathbf{u}^T A^T C\,A\,\mathbf{u} + 2\,\mathbf{u}^T A^T C\,\mathbf{b} + \mathbf{b}^T C\,\mathbf{b}. \end{aligned}$$

The resulting quadratic function can be written in the usual form[†]

$$\tfrac{1}{2}\,P = p(\mathbf{u}) = \tfrac{1}{2}\,\mathbf{u}^T K\,\mathbf{u} - \mathbf{u}^T\mathbf{f} + c, \tag{6.40}$$

where $K = A^T C\,A$ is the network resistivity matrix, while $\mathbf{f} = -\,A^T C\,\mathbf{b}$ are the equivalent current sources at the nodes (6.38) that correspond to the batteries. The last term $c = \frac{1}{2}\,\mathbf{b}^T C\,\mathbf{b}$ is one half the internal power in the battery, and does not depend upon the currents/voltages in the wires. In deriving (6.40), we have ignored any additional external current sources at the nodes. By an analogous argument, a current source will contribute to the linear terms in the power in the same fashion, and so the linear terms $\mathbf{u}^T\mathbf{f}$ represent the effect of both batteries and external current sources.

In general, the resistivity matrix K is only positive semi-definite, and so the quadratic power function (6.40) does not, in general, have a minimizer. As argued above, to ensure equilibrium, we need to ground one or more of the nodes. The resulting reduced form

$$p(\mathbf{u}^\star) = \tfrac{1}{2}\,(\mathbf{u}^\star)^T K^\star \mathbf{u}^\star - (\mathbf{u}^\star)^T \mathbf{f}^\star,$$

for the power now has a positive definite coefficient matrix $K^\star > 0$. The minimizer of the power function is the solution $\mathbf{u}^\star$ to the reduced linear system (6.35). Therefore, the network adjusts itself to *minimize the power or total energy loss*! Just as with mechanical systems, Nature solves a minimization problem in an effort to conserve energy.

We have discovered the remarkable correspondence between the equilibrium equations for electrical networks (6.10), and those of mass–spring chains (6.27). This *Electrical–Mechanical Correspondence* is summarized in the following table. In the following section, we will see that the analogy extends to more general structures. In Chapter 15, we will discover that it continues to apply in the continuous regime, and subsumes solid mechanics, fluid mechanics, electrostatics, and many other physical systems in a common mathematical framework!

[†] For alternating currencts, there is no annoying factor of 2 in the formula for the power, and the analogy is more direct.

Structures	Variables	Networks
Displacements	$\mathbf{u}$	Voltages
Elongations[‡]	$\mathbf{v} = A\,\mathbf{u}$	Voltage drops
Spring stiffnesses	C	Conductivities
Internal Forces	$\mathbf{y} = C\,\mathbf{v}$	Currents
External forcing	$\mathbf{f} = A^T\mathbf{y}$	Current sources
Stiffness matrix	$K = A^T C\,A$	Resistivity matrix
Potential energy	$p(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T K\mathbf{u} - \mathbf{u}^T\mathbf{f}$	$\frac{1}{2}\times$ Power
Prestressed bars/springs	$\mathbf{v} = A\,\mathbf{u} + \mathbf{b}$	Batteries

6.3. Structures in Equilibrium.

A *structure* (sometimes known as a *truss*) is a mathematical idealization of a framework for a building. Think of a skyscraper when just the I-beams are connected together — before the walls, floors, ceilings, roof and ornamentation are added. An ideal structure is constructed of elastic bars connected at *joints*. By a *bar*, we mean a straight, rigid rod that can be (slightly) elongated, but not bent. (Beams, which are allowed to bend, are more complicated, and we defer their treatment until Section 11.4.) When a bar is stretched, it obeys Hooke's law (at least in the linear regime we are currently modeling) and so, for all practical purposes, behaves like a spring with a very large stiffness. As a result, a structure can be regarded as a two- or three-dimensional generalization of a mass–spring chain.

The joints will allow the bar to rotate in any direction. Of course, this is an idealization; in a building, the rivets and bolts will prevent rotation to a significant degree. However, under moderate stress — for example, if the wind is blowing through our skyscraper, the rivets and bolts can only be expected to keep the structure connected, and the rotational motion will provide stresses on the bolts which must be taken into account when designing the structure. Of course, under extreme stress, the structure will fall apart — a disaster that its designers must avoid. The purpose of this section is to derive conditions that will guarantee that a structure is rigidly stable under moderate forcing, or, alternatively, understand the mechanisms that might lead to its collapse.

Bars

The first order of business is to understand how an individual bar reacts to motion. We have already encountered the basic idea in our treatment of springs. The key complication here is that the ends of the bar/spring are not restricted to a single direction of motion, but can move in either two or three-dimensional space. We use $d = 2$ or 3 to denote the dimension of the underlying space. (When $d = 1$, the truss reduces to a mass–spring chain.)

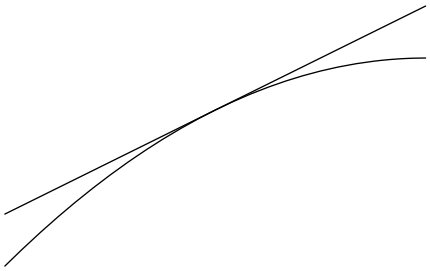


Figure 6.5. Tangent Line Approximation.

Consider an unstressed bar with one end at position $\mathbf{a}_1 \in \mathbb{R}^d$ and the other end at position $\mathbf{a}_2 \in \mathbb{R}^d$. In $d = 2$ dimensions, we write $\mathbf{a}_i = (\, a_i, b_i \,)^T$, while in $d = 3$-dimensional space $\mathbf{a}_i = (\, a_i, b_i, c_i \,)^T$. The length of the bar is $L = \| \, \mathbf{a}_1 - \mathbf{a}_2 \, \|$, where we use the standard Euclidean norm to measure distance on $\mathbb{R}^d$ throughout this section.

Suppose we move the ends of the bar a little, sending $\mathbf{a}_i$ to $\mathbf{b}_i = \mathbf{a}_i + \varepsilon \, \mathbf{u}_i$ and simultaneously $\mathbf{a}_j$ to $\mathbf{b}_j = \mathbf{a}_j + \varepsilon \, \mathbf{u}_j$. The unit vectors $\mathbf{u}_i, \mathbf{u}_j \in \mathbb{R}^d$ indicate the respective direction of displacement of the two ends, and we think of $\varepsilon > 0$, the magnitude of the displacement, as small. How much has this motion stretched the bar? The length of the displaced bar is

$$\begin{aligned} L + e = \| \, \mathbf{b}_i - \mathbf{b}_j \, \| &= \| \, (\mathbf{a}_i + \varepsilon \, \mathbf{u}_i) - (\mathbf{a}_j + \varepsilon \, \mathbf{u}_j) \, \| = \| \, (\mathbf{a}_i - \mathbf{a}_j) + \varepsilon \, (\mathbf{u}_i - \mathbf{u}_j) \, \| \\ &= \sqrt{\| \, \mathbf{a}_i - \mathbf{a}_j \, \|^2 + 2 \, \varepsilon \, (\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j) + \varepsilon^2 \, \| \, \mathbf{u}_i - \mathbf{u}_j \, \|^2} \; . \end{aligned} \tag{6.41}$$

The difference between the new length and the original length, namely

$$e \;=\; \sqrt{\| \, \mathbf{a}_i - \mathbf{a}_j \, \|^2 + 2 \, \varepsilon \, (\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j) + \varepsilon^2 \, \| \, \mathbf{u}_i - \mathbf{u}_j \, \|^2} \;-\; \| \, \mathbf{a}_i - \mathbf{a}_j \, \|, \tag{6.42}$$

is, by definition, the bar's *elongation*.

If the underlying dimension d is 2 or more, the elongation e is a *nonlinear* function of the displacement vectors $\mathbf{u}_i, \mathbf{u}_j$. Thus, an exact, geometrical treatment of structures in equilibrium requires dealing with nonlinear systems of equations. For example, the design of robotic mechanisms, [**111**], requires dealing with the fully nonlinear equations. However, in many practical situations, the displacements are fairly small, so $\varepsilon \ll 1$. For example, when a building moves, the lengths of bars are in meters, but the displacements are, barring catastrophes, typically in centimeters if not millimeters. In such situations, we can replace the geometrically exact elongation by a much simpler linear approximation.

The most basic linear approximation to a nonlinear function $g(\varepsilon)$ near $\varepsilon = 0$ is given by its tangent line or linear Taylor polynomial

$$g(\varepsilon) \;\approx\; g(0) + g'(0) \, \varepsilon, \tag{6.43}$$

as in Figure 6.5. In the case of small displacements of a bar, the elongation (6.42) is a square root function of the particular form

$$g(\varepsilon) = \sqrt{a^2 + 2 \, \varepsilon \, b + \varepsilon^2 \, c^2} \; - a,$$

where

$$a = \| \mathbf{a}_i - \mathbf{a}_j \|, \qquad b = (\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j), \qquad c = \| \mathbf{u}_i - \mathbf{u}_j \|,$$

are independent of ε. Since $g(0) = 0$ and $g'(0) = \dfrac{b}{a}$, the linear approximation (6.43) has the form

$$\sqrt{a^2 + 2\,\varepsilon\, b + \varepsilon^2\, c^2} \; - a \;\approx\; \varepsilon\, \frac{b}{a} \qquad \text{for} \qquad \varepsilon \ll 1.$$

In this manner, we arrive at the linear approximation to the bar's elongation

$$e \;\approx\; \varepsilon\, \frac{(\mathbf{a}_i - \mathbf{a}_j) \cdot (\mathbf{u}_i - \mathbf{u}_j)}{\| \mathbf{a}_i - \mathbf{a}_j \|} \;=\; \mathbf{n} \cdot (\varepsilon\, \mathbf{u}_i - \varepsilon\, \mathbf{u}_j), \qquad \text{where} \qquad \mathbf{n} = \frac{(\mathbf{a}_i - \mathbf{a}_j)}{\| \mathbf{a}_i - \mathbf{a}_j \|}$$

is the unit vector, $\| \mathbf{n} \| = 1$, that points in the direction of the bar from node j to node i.

The overall small factor of ε was merely a device used to derive the linear approximation. It can now be safely discarded, so that the displacement of the i^{th} node is now $\mathbf{u}_i$ instead of $\varepsilon\, \mathbf{u}_i$, and we assume $\| \mathbf{u}_i \|$ is small. If bar k connects node i to node j, then its (approximate) elongation is equal to

$$e_k = \mathbf{n}_k \cdot (\mathbf{u}_i - \mathbf{u}_j) = \mathbf{n}_k \cdot \mathbf{u}_i - \mathbf{n}_k \cdot \mathbf{u}_j, \qquad \text{where} \qquad \mathbf{n}_k = \frac{\mathbf{a}_i - \mathbf{a}_j}{\| \mathbf{a}_i - \mathbf{a}_j \|}. \tag{6.44}$$

The elongation e_k is the sum of two terms: the first, $\mathbf{n}_k \cdot \mathbf{u}_i$, is the component of the displacement vector for node i in the direction of the unit vector $\mathbf{n}_k$ that points along the bar *towards* node i, whereas the second, $-\mathbf{n}_k \cdot \mathbf{u}_j$, is the component of the displacement vector for node j in the direction of the unit vector $-\mathbf{n}_k$ that points in the opposite direction along the bar *towards* node j. Their sum gives the total elongation of the bar.

We assemble all the linear equations (6.44) relating nodal displacements to bar elongations in matrix form

$$\mathbf{e} = A\, \mathbf{u}. \tag{6.45}$$

Here $\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix} \in \mathbb{R}^m$ is the vector of elongations, while $\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} \in \mathbb{R}^{dn}$ is the vector of displacements. Each $\mathbf{u}_i \in \mathbb{R}^d$ is itself a column vector with d entries, and so $\mathbf{u}$ has a total of dn entries. For example, in the planar case $d = 2$, we have $\mathbf{u}_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$ since each node's displacement has both an x and y component, and so

$$\mathbf{u} \;=\; \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} \;=\; \begin{pmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_n \\ y_n \end{pmatrix} \;\in\; \mathbb{R}^{2\,n}.$$

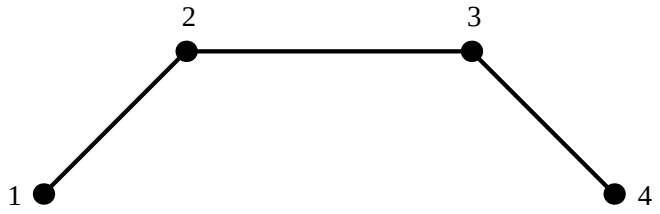


Figure 6.6. Three Bar Planar Structure.

In three dimensions, $d = 3$, we have $\mathbf{u}_i = (\, x_i, y_i, z_i \,)^T$, and so each node will contribute three components to the displacement vector

$$\mathbf{u} = (\, x_1, y_1, z_1, x_2, y_2, z_2, \ \dots \ , x_n, y_n, z_n \,)^T \ \in \ \mathbb{R}^{3\,n}.$$

The *incidence matrix* A connecting the displacements and elongations will be of size $m \times d\,n$. The k^{th} row of A will have (at most) $2\,d$ nonzero entries. The entries in the d slots corresponding to node i will be the components of the (transposed) unit bar vector $\mathbf{n}_k^T$ pointing towards node i, as given in (6.44), while the entries in the d slots corresponding to node j will be the components of its negative $-\,\mathbf{n}_k^T$, which is the unit bar vector pointing towards node j. All other entries are 0. The constructions are best appreciated by working through an explicit example.

Example 6.5. Consider the planar structure pictured in Figure 6.6. The four nodes are at positions

$$\mathbf{a}_1 = (0,0)^T, \qquad \mathbf{a}_2 = (1,1)^T, \qquad \mathbf{a}_3 = (3,1)^T, \qquad \mathbf{a}_4 = (4,0)^T,$$

so the two side bars are at 45° angles and the center bar is horizontal. Applying our algorithm, the associated incidence matrix is

$$A = \left(\begin{array}{cc|cc|cc|cc} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{array}\right). \tag{6.46}$$

The three rows of A refer to the three bars in our structure. The columns come in pairs, as indicated by the vertical lines in the matrix: the first two columns refer to the x and y displacements of the first node; the third and fourth columns refer to the second node, and so on. The first two entries of the first row of A indicate the unit vector

$$\mathbf{n}_1 \ = \ \frac{\mathbf{a}_1 - \mathbf{a}_2}{\| \, \mathbf{a}_1 - \mathbf{a}_2 \, \|} \ = \ \left(-\tfrac{1}{\sqrt{2}}, -\tfrac{1}{\sqrt{2}} \right)^T$$

that points along the first bar towards the first node, while the third and fourth entries have the opposite signs, and form the unit vector

$$-\,\mathbf{n}_1 \ = \ \frac{\mathbf{a}_2 - \mathbf{a}_1}{\| \, \mathbf{a}_2 - \mathbf{a}_1 \, \|} \ = \ \left(\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}} \right)^T$$

along the same bar that points in the opposite direction — towards the second node. The remaining entries are zero because the first bar only connects the first two nodes. Similarly, the unit vector along the second bar pointing towards node 2 is

$$\mathbf{n}_2 \;=\; \frac{\mathbf{a}_2 - \mathbf{a}_3}{\| \mathbf{a}_2 - \mathbf{a}_3 \|} \;=\; (-1, 0)^T,$$

and this gives the third and fourth entries of the second row of A; the fifth and sixth entries are their negatives, corresponding to the unit vector $-\mathbf{n}_2$ pointing towards node 3. The last row is constructed from the unit vector in the direction of bar #3 in the same fashion.

Remark: Interestingly, the incidence matrix for a structure only depends on the directions of the bars and not their lengths. This is analogous to the fact that the incidence matrix for an electrical network only depends on the connectivity properties of the wires and not on their overall lengths. Indeed, one can regard the incidence matrix for a structure as a kind of d–dimensional generalization of the incidence matrix for a directed graph.

The next phase of our procedure is to introduce the constitutive relations for the bars in our structure that determine their internal forces or stresses. As we remarked at the beginning of the section, each bar is viewed as a very strong spring, subject to a linear Hooke's law equation

$$y_k = c_k\, e_k \tag{6.47}$$

that relates its elongation e_k to its internal force y_k. The bar stiffness $c_k > 0$ is a positive scalar, and so $y_k > 0$ if the bar is in tension, while $y_k < 0$ if the bar is compressed. In this approximation, there is no bending and the bars will only experience external forcing at the nodes. We write (6.47) in matrix form

$$\mathbf{y} = C\,\mathbf{e},$$

where $C = \operatorname{diag}(c_1, \dots, c_m) > 0$ is a diagonal, positive definite matrix.

Finally, we need to balance the forces at each node in order to achieve equilibrium. If bar k terminates at node i, then it exerts a force $-y_k\,\mathbf{n}_k$ on the node, where $\mathbf{n}_k$ is the unit vector pointing towards the node in the direction of the bar, as in (6.44). The minus sign comes from physics: if the bar is under tension, so $y_k > 0$, then it is trying to contract back to its unstressed state, and so will pull the node towards it — in the opposite direction to $\mathbf{n}_k$ — while a bar in compression will push the node away. In addition, we may have an externally applied force vector, denoted by $\mathbf{f}_i$, on node i, which might be some combination of gravity, weights, mechanical forces, and so on. (In this admittedly simplified model, external forces only act on the nodes.) Force balance at equilibrium requires that the sum of all the forces, external and internal, at each node cancel; thus,

$$\mathbf{f}_i \;+\; \sum_k\, (-\,y_k\,\mathbf{n}_k) = 0, \qquad \text{or} \qquad \sum_k\, y_k\,\mathbf{n}_k = \mathbf{f}_i,$$

where the sum is over all the bars that are attached to node i. The matrix form of the force balance equations is (and this should no longer come as a surprise)

$$\mathbf{f} = A^T\mathbf{y}, \tag{6.48}$$

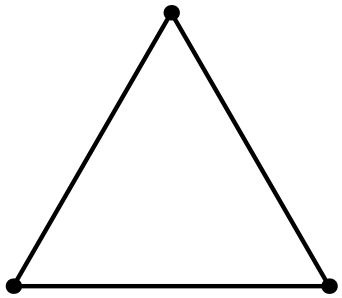


Figure 6.7. A Triangular Structure.

where A^T is the transpose of the incidence matrix, and $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n)^T \in \mathbb{R}^{dn}$ is the vector containing all external force on the nodes. Putting everything together, (6.45), (6.47), (6.48), i.e.,

$$\mathbf{e} = A\,\mathbf{u}, \qquad \mathbf{y} = C\,\mathbf{e}, \qquad \mathbf{f} = A^T\mathbf{y},$$

we once again are lead to our familiar linear system of equations

$$K\,\mathbf{u} = \mathbf{f}, \qquad \text{where} \qquad K = A^T C\,A. \tag{6.49}$$

The stiffness matrix K is a positive (semi-)definite Gram matrix (3.51) associated with the weighted inner product on the space of elongations prescribed by the diagonal matrix C.

As we know, the stiffness matrix for our structure will be positive definite, $K > 0$, if and only if the incidence matrix has trivial kernel: $\ker A = \{\mathbf{0}\}$. The preceding example, and indeed all of these constructed so far, will not have this property, for the same reason as in an electrical network — because we have not tied down (or "grounded") our structure anywhere. In essence, we are considering a structure floating in outer space, which is free to move around without changing its shape. As we will see, each possible rigid motion of the structure will correspond to an element of the kernel of its incidence matrix, and thereby prevent positive definiteness of the structure matrix K.

Example 6.6. Consider a planar space station in the shape of a unit equilateral triangle, as in Figure 6.7. Placing the nodes at positions

$$\mathbf{a}_1 = \left(\tfrac{1}{2}, \quad \tfrac{\sqrt{3}}{2}\right)^T, \qquad \mathbf{a}_2 = (1, 0)^T, \qquad \mathbf{a}_3 = (0, 0)^T,$$

we use the preceding algorithm to compute the incidence matrix

$$A = \left(\begin{array}{cc|cc|cc} \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 & -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} & -\frac{\sqrt{3}}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \end{array}\right),$$

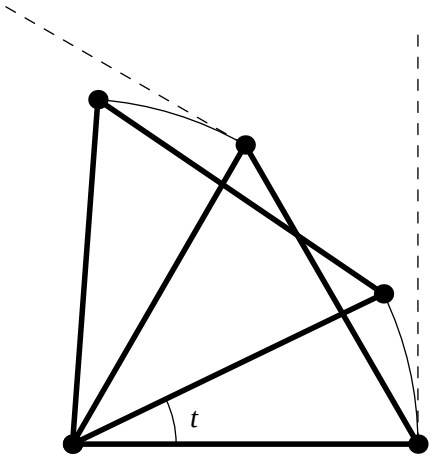


Figure 6.8. Rotating a Space Station.

whose rows are indexed by the bars, and whose columns are indexed in pairs by the three nodes. The kernel of A is three-dimensional, with basis

$$\mathbf{z}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{z}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathbf{z}_3 = \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}. \tag{6.50}$$

These three displacement vectors correspond to three different planar rigid motions: the first two correspond to translations, and the third to a rotation.

The translations are easy to discern. Translating the space station in a horizontal direction means that we move all three nodes the same amount, and so the displacements are $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}_3 = \mathbf{a}$ for some fixed vector $\mathbf{a}$. In particular, a rigid unit horizontal translation has $\mathbf{a} = \mathbf{e}_1 = (\,1, 0\,)^T$, and corresponds to the first kernel basis vector. Similarly, a unit vertical translation of all three nodes corresponds to $\mathbf{a} = \mathbf{e}_2 = (\,0, 1\,)^T$, and corresponds to the second kernel basis vector. Any other translation is a linear combination of these two. Translations do not alter the lengths of any of the bars, and so do not induce any stress in the structure.

The rotations are a little more subtle, owing to the linear approximation that we used to compute the elongations. Referring to Figure 6.8, rotating the space station through a small angle ε around the node $\mathbf{a}_3 = (\,0, 0\,)^T$ will move the other two nodes to positions

$$\mathbf{b}_1 = \begin{pmatrix} \frac{1}{2}\cos\varepsilon - \frac{\sqrt{3}}{2}\sin\varepsilon \\ \frac{1}{2}\sin\varepsilon + \frac{\sqrt{3}}{2}\cos\varepsilon \end{pmatrix}, \qquad \mathbf{b}_2 = \begin{pmatrix} \cos\varepsilon \\ \sin\varepsilon \end{pmatrix}, \qquad \mathbf{b}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{6.51}$$

However, the corresponding displacements

$$\mathbf{u}_1 = \mathbf{b}_1 - \mathbf{a}_1 = \begin{pmatrix} \frac{1}{2}(\cos\varepsilon - 1) - \frac{\sqrt{3}}{2}\sin\varepsilon \\ \frac{1}{2}\sin\varepsilon + \frac{\sqrt{3}}{2}(\cos\varepsilon - 1) \end{pmatrix},$$
$$\mathbf{u}_2 = \mathbf{b}_2 - \mathbf{a}_2 = \begin{pmatrix} \cos\varepsilon - 1 \\ \sin\varepsilon \end{pmatrix}, \qquad \mathbf{u}_3 = \mathbf{b}_3 - \mathbf{a}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{6.52}$$

do *not* combine into a vector that belongs to $\ker A$. The problem is that, under a rotation, the nodes move along circles, while the kernel displacements $\mathbf{u} = \varepsilon\,\mathbf{z} \in \ker A$ correspond to straight line motion! In order to maintain consistency, we must adopt a similar linear approximation of the nonlinear circular motion of the nodes. Thus, we replace the nonlinear displacements $\mathbf{u}_j(\varepsilon)$ in (6.52) by their linear tangent approximations[†] $\varepsilon\,\mathbf{u}_j'(0)$, and so

$$\mathbf{u}_1 \;\approx\; \varepsilon \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix}, \qquad \mathbf{u}_2 \;\approx\; \varepsilon \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The resulting displacements *do* combine to produce the displacement vector

$$\mathbf{u} \;=\; \varepsilon \begin{pmatrix} -\frac{\sqrt{3}}{2} & \frac{1}{2} & 0 & 1 & 0 & 0 \end{pmatrix}^T \;=\; \varepsilon\,\mathbf{z}_3$$

that moves the space station in the direction of the third element of the kernel of the incidence matrix! Thus, as claimed, $\mathbf{z}_3$ represents the linear approximation to a rigid rotation around the first node.

Remarkably, the rotations around the other two nodes, although distinct nonlinear motions, can be linearly approximated by particular combinations of the three kernel basis elements $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, and so already appear in our description of $\ker A$. For example, the displacement vector

$$\mathbf{u} = \varepsilon \left(\tfrac{\sqrt{3}}{2}\,\mathbf{z}_1 + \tfrac{1}{2}\,\mathbf{z}_2 - \mathbf{z}_3 \right) = \varepsilon \begin{pmatrix} 0 & 0 & \frac{\sqrt{3}}{2} & -\frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} \end{pmatrix}^T \tag{6.53}$$

represents the linear approximation to a rigid rotation around the first node. We conclude that the three-dimensional kernel of the incidence matrix represents the sum total of all possible rigid motions of the space station, or, more correctly, their linear approximations.

Which types of forces will maintain the space station in equilibrium? This will happen if and only if we can solve the force balance equations

$$A^T\mathbf{y} = \mathbf{f} \tag{6.54}$$

for the internal forces $\mathbf{y}$. The Fredholm Alternative Theorem 5.51 implies that the system (6.54) has a solution if and only if $\mathbf{f}$ is orthogonal to $\operatorname{coker} A^T = \ker A$. Therefore, $\mathbf{f} =$

[†] Note that $\mathbf{u}_j(0) = \mathbf{0}$.

$(\, f_1 \;\, g_1 \;\, f_2 \;\, g_2 \;\, f_3 \;\, g_3 \,)^T$ must be orthogonal to the basis vectors (6.50), and so must satisfy the three linear constraints

$$\mathbf{z}_1 \cdot \mathbf{f} = f_1 + f_2 + f_3 = 0, \qquad \mathbf{z}_2 \cdot \mathbf{f} = g_1 + g_2 + g_3 = 0, \qquad \mathbf{z}_3 \cdot \mathbf{f} = \tfrac{\sqrt{3}}{2}\, f_1 + \tfrac{1}{2}\, g_1 + g_3 = 0.$$

The first requires that there is no net horizontal force on the space station. The second requires no net vertical force. The last constraint requires that the *moment* of the forces around the first node vanishes. The vanishing of the force moments around each of the other two nodes follows, since the associated kernel vectors can be expressed as linear combinations of the three basis elements. The physical requirements are clear. If there is a net horizontal or vertical force, the space station will rigidly translate in that direction; if there is a non-zero force moment, the station will rigidly rotate. In any event, unless the force balance equations are satisfied, the space station cannot remain in equilibrium. A freely floating space station is in an unstable configuration and can easily be set into motion.

Since there are three independent rigid motions, we need to impose three constraints on the structure in order to stabilize it. "Grounding" one of the nodes, i.e., preventing it from moving by attaching it to a fixed support, will serve to eliminate the two translational instabilities. For example, setting $\mathbf{u}_3 = \mathbf{0}$ has the effect of fixing the third node of the space station to a support. With this specification, we can eliminate the variables associated with that node, and thereby delete the corresponding columns of the incidence matrix — leaving the *reduced incidence matrix*

$$A^\star = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} & \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The kernel of $A^\star$ is now only one-dimensional, spanned by the single vector

$$\mathbf{z}_3^\star = \begin{pmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} & 0 & 1 \end{pmatrix}^T,$$

which corresponds to (the linear approximation of) the rotations around the fixed node. To prevent the structure from rotating, we can also fix the second node, by further requiring $\mathbf{u}_2 = \mathbf{0}$. This allows us to eliminate the third and fourth columns of the incidence matrix and the resulting "doubly reduced" incidence matrix

$$A^{\star\star} = \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & 0 \end{pmatrix}.$$

Now $\ker A^{\star\star} = \{\mathbf{0}\}$ is trivial, and hence the corresponding reduced stiffness matrix

$$K^{\star\star} = (A^{\star\star})^T A^{\star\star} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{pmatrix}$$

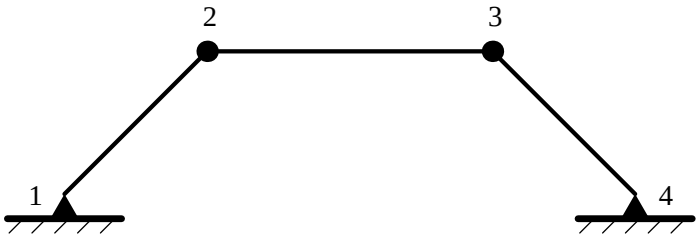


Figure 6.9. Three Bar Structure with Fixed Supports.

is positive definite. The space station with two fixed nodes is a stable structure, which can now support an arbitrary external forcing. (Forces on the fixed nodes now have no effect since they are no longer allowed to move.)

In general, a planar structure without any fixed nodes will have at least a three-dimensional kernel, corresponding to the rigid planar motions of translations and (linear approximations to) rotations. To stabilize the structure, one must fix two (non-coincident) nodes. A three-dimensional structure that is not tied to any fixed supports will admit 6 independent rigid motions in its kernel. Three of these correspond to rigid translations in the three coordinate directions, while the other three correspond to linear approximations to the rigid rotations around the three coordinate axes. To eliminate the rigid motion instabilities of the structure, one needs to fix three non-collinear nodes; details can be found in the exercises.

Even after attaching a sufficient number of nodes to fixed supports so as to eliminate all possible rigid motions, there may still remain nonzero vectors in the kernel of the reduced incidence matrix of the structure. These indicate additional instabilities in which the shape of the structure can deform without any applied force. Such non-rigid motions are known as *mechanisms* of the structure. Since a mechanism moves the nodes without elongating any of the bars, it does not induce any internal forces. A structure that admits a mechanism is unstable — even very tiny external forces may provoke a large motion.

Example 6.7. Consider the three bar structure of Example 6.5, but now with its two ends attached to supports, as pictured in Figure 6.9. Since we are fixing nodes 1 and 4, setting $\mathbf{u}_1 = \mathbf{u}_4 = \mathbf{0}$, we should remove the first two and last column pairs from the incidence matrix (6.46), leading to the reduced incidence matrix

$$A^\star = \left(\begin{array}{cc|cc} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{array} \right).$$

The structure no longer admits any rigid motions. However, the kernel of $A^\star$ is one-dimensional, spanned by reduced displacement vector $\mathbf{z}^\star = (\,1 \;\; -1 \;\; 1 \;\; 1\,)^T$, which corresponds to the unstable mechanism that displaces the second node in the direction $\mathbf{u}_2 = (\,1 \;\; -1\,)^T$ and the third node in the direction $\mathbf{u}_3 = (\,1 \;\; 1\,)^T$. Geometrically, then,

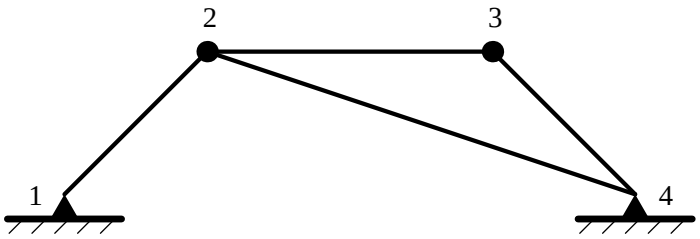


Figure 6.10. Reinforced Planar Structure.

$\mathbf{z}^\star$ represents the displacement where node 2 moves down and to the left at a 45° angle, while node 3 moves simultaneously up and to the left at a 45° angle. This mechanism does not alter the lengths of the three bars (at least in our linear approximation regime) and so requires no net force to be set into motion.

As with the rigid motions of the space station, an external forcing vector $\mathbf{f}^\star$ will maintain equilibrium only when it lies in the corange of $A^\star$, and hence must be orthogonal to all the mechanisms in $\ker A^\star = (\operatorname{corng} A^\star)^\perp$. Thus, the nodal forces $\mathbf{f}_2 = (\, f_2, g_2 \,)^T$ and $\mathbf{f}_3 = (\, f_3, g_3 \,)^T$ must satisfy the balance law

$$\mathbf{z}^\star \cdot \mathbf{f}^\star = f_2 - g_2 + f_3 + g_3 = 0.$$

If this fails, the equilibrium equation has no solution, and the structure will move. For example, a uniform horizontal force $f_2 = f_3 = 1$, $g_2 = g_3 = 0$, will induce the mechanism, whereas a uniform vertical force, $f_2 = f_3 = 0$, $g_2 = g_3 = 1$, will maintain equilibrium. In the latter case, the solution to the equilibrium equations

$$K^\star \mathbf{u}^\star = \mathbf{f}^\star, \qquad \text{where} \qquad K^\star = (A^\star)^T A^\star = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} & -1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ -1 & 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

is indeterminate, since we can add in any element of $\ker K^\star = \ker A^\star$, so

$$\mathbf{u}^\star = (\, -3 \quad 5 \quad -2 \quad 0 \,)^T + t\,(\, 1 \quad -1 \quad 1 \quad 1 \,)^T.$$

In other words, the equilibrium position is not unique, since the structure can still be displaced in the direction of the unstable mechanism while maintaining the overall force balance. On the other hand, the elongations and internal forces

$$\mathbf{y} = \mathbf{e} = A^\star\, \mathbf{u}^\star = (\, -\sqrt{2} \quad -1 \quad -\sqrt{2} \,)^T,$$

are well-defined, indicating that, under our stabilizing uniform vertical force, all three bars are compressed, with the two diagonal bars experiencing 41.4% more compression than the horizontal bar.

Remark: Just like the rigid rotations, the mechanisms described here are linear approximations to the actual nonlinear motions. In a physical structure, the vertices will move along curves whose tangents at the initial configuration are the directions indicated by the mechanism vector. In certain cases, a structure can admit a linear mechanism, but one that cannot be physically realized due to the nonlinear constraints imposed by the geometrical configurations of the bars. Nevertheless, such a structure is at best borderline stable, and should not be used in any real-world applications that rely on stability of the structure.

We can always stabilize a structure by first fixing nodes to eliminate rigid motions, and then adding in extra bars to prevent mechanisms. In the preceding example, suppose we attach an additional bar connecting nodes 2 and 4, leading to the reinforced structure in Figure 6.10. The revised incidence matrix is

$$A = \left(\begin{array}{cc|cc|cc|cc} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & 0 & -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} & 0 & 0 & \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \end{array}\right).$$

and is obtained from (6.46) by appending another row representing the added bar. When nodes 1 and 4 are fixed, the reduced incidence matrix

$$A^\star = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} & 0 & 0 \end{pmatrix}$$

has trivial kernel, $\ker A^\star = \{\mathbf{0}\}$, and hence the structure is stable. It admits no mechanisms, and can support any configuration of forces (within reason — very large forces will take us outside the linear regime described by the model, and the structure may be crushed!).

This particular case is *statically determinate* owing to the fact that the incidence matrix is square and nonsingular, which implies that one can solve the force balance equations (6.54) directly for the internal forces. For instance, a uniform downwards vertical force $f_2 = f_3 = 0$, $g_2 = g_3 = -1$, e.g., gravity, will produce the internal forces

$$y_1 = -\sqrt{2}, \qquad y_2 = -1, \qquad y_3 = -\sqrt{2}, \qquad y_4 = 0$$

indicating that bars 1, 2 and 3 are experiencing compression, while, interestingly, the reinforcing bar 4 remains unchanged in length and hence experiences no internal force. Assuming the bars are all of the same material, and taking the elastic constant to be 1, so

$C = \mathrm{I}$, then the reduced stiffness matrix is

$$K^\star = (A^\star)^T A^\star = \begin{pmatrix} \frac{12}{5} & \frac{1}{5} & -1 & 0 \\ \frac{1}{5} & \frac{3}{5} & 0 & 0 \\ -1 & 0 & \frac{3}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

The solution to the reduced equilibrium equations is

$$\mathbf{u}^\star = \begin{pmatrix} -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} & -\frac{7}{2} \end{pmatrix}^T, \qquad \text{so} \qquad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{2} & -\frac{3}{2} \end{pmatrix}^T, \quad \mathbf{u}_3 = \begin{pmatrix} -\frac{3}{2} & -\frac{7}{2} \end{pmatrix}^T.$$

give the displacements of the two nodes under the applied force. Both are moving down and to the left, with node 3 moving relatively farther owing to its lack of reinforcement.

Suppose we reinforce the structure yet further by adding in a bar connecting nodes 1 and 3. The resulting reduced incidence matrix

$$A^\star = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} & 0 & 0 \\ 0 & 0 & \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{pmatrix}$$

again has trivial kernel, $\ker A^\star = \{\mathbf{0}\}$, and hence the structure is stable. Indeed, adding in extra bars to a stable structure cannot cause it to lose stability. (In matrix language, appending additional rows to a matrix cannot increase the size of its kernel, cf. Exercise .) Since the incidence matrix is rectangular, the structure is now *statically indeterminate* and we cannot determine the internal forces without first solving the full equilibrium equations (6.49) for the displacements. The stiffness matrix is

$$K^\star = (A^\star)^T A^\star = \begin{pmatrix} \frac{12}{5} & \frac{1}{5} & -1 & 0 \\ \frac{1}{5} & \frac{3}{5} & 0 & 0 \\ -1 & 0 & \frac{12}{5} & -\frac{1}{5} \\ 0 & 0 & -\frac{1}{5} & \frac{3}{5} \end{pmatrix}.$$

For the same uniform vertical force, the displacement $\mathbf{u}^\star = (K^\star)^{-1}\mathbf{f}^\star$ is

$$\mathbf{u}^\star = \begin{pmatrix} \frac{1}{10} & -\frac{17}{10} & -\frac{1}{10} & -\frac{17}{10} \end{pmatrix}^T,$$

so that the free nodes now move symmetrically down and towards the center of the structure. The internal forces on the bars are

$$y_1 = -\tfrac{4}{5}\sqrt{2}, \qquad y_2 = -\tfrac{1}{5}, \qquad y_3 = -\tfrac{4}{5}\sqrt{2}, \qquad y_4 = -\sqrt{\tfrac{2}{5}}, \qquad y_5 = -\sqrt{\tfrac{2}{5}}.$$

All five bars are now experiencing compression, the two outside bars being the most stressed, the reinforcing bars slightly more than half that, while the center bar feels less

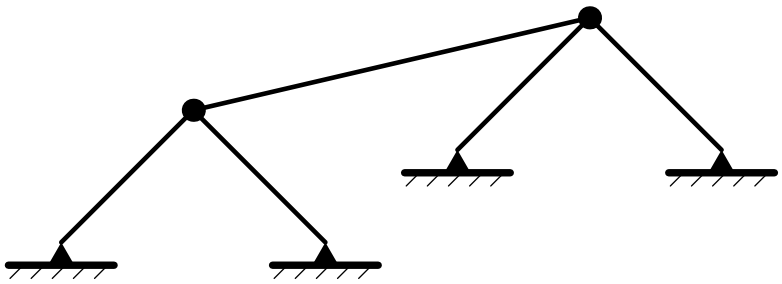


Figure 6.11. A Swing Set.

than a fifth the stress that the outside bars experience. This simple computation should already indicate to the practicing construction engineer which of the bars in our structure are more likely to collapse under an applied external force. By comparison, the reader can investigate what happens under a uniform horizontal force.

Summarizing our discussion, we have established the following fundamental result characterizing the stability and equilibrium of structures.

Theorem 6.8. *A structure is stable, and will maintain an equilibrium under arbitrary external forcing, if and only if its reduced incidence matrix $A^\star$ has linearly independent columns, or, equivalently,* $\ker A^\star = \{\mathbf{0}\}$. *More generally, an external force $\mathbf{f}^\star$ on a structure will maintain equilibrium if and only if $\mathbf{f}^\star \in (\ker A^\star)^\perp$, which means that the external force is orthogonal to all rigid motions and all mechanisms admitted by the structure.*

Example 6.9. A swing set is to be constructed, consisting of two diagonal supports at each end and a horizontal cross bar. Is this configuration stable, i.e., can a child swing on it without it collapsing? The movable joints are at positions

$$\mathbf{a}_1 = (\,1,1,3\,)^T, \qquad \mathbf{a}_2 = (\,4,1,3\,)^T.$$

The four fixed supports are at positions

$$\mathbf{a}_3 = (\,0,0,0\,)^T, \qquad \mathbf{a}_4 = (\,0,2,0\,)^T, \qquad \mathbf{a}_5 = (\,5,0,0\,)^T, \qquad \mathbf{a}_6 = (\,5,2,0\,)^T.$$

The reduced incidence matrix for the structure is calculated in the usual manner:

$$A^\star = \left(\begin{array}{ccc|ccc} \frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \end{array}\right).$$

For instance, the first three entries contained in the first row refer to the unit vector $\mathbf{n}_1 = \dfrac{\mathbf{a}_1 - \mathbf{a}_3}{\| \mathbf{a}_1 - \mathbf{a}_3 \|}$ in the direction of the bar going from $\mathbf{a}_3$ to $\mathbf{a}_1$. Suppose the three bars

have the same stiffness, and so (taking $c_1 = \cdots = c_5 = 1$) the reduced stiffness matrix for the structure is

$$K^\star = (A^\star)^T A^\star = \begin{pmatrix} \frac{13}{11} & 0 & \frac{6}{11} & -1 & 0 & 0 \\ 0 & \frac{2}{11} & 0 & 0 & 0 & 0 \\ \frac{6}{11} & 0 & \frac{18}{11} & 0 & 0 & 0 \\ -1 & 0 & 0 & \frac{13}{11} & 0 & -\frac{6}{11} \\ 0 & 0 & 0 & 0 & \frac{2}{11} & 0 \\ 0 & 0 & 0 & -\frac{6}{11} & 0 & \frac{18}{11} \end{pmatrix}$$

We find $\ker K^\star = \ker A^\star$ is one-dimensional, with basis

$$\mathbf{z}^\star = (\,3 \;\; 0 \;\; -1 \;\; 3 \;\; 0 \;\; 1\,)^T,$$

which indicates a mechanism that causes the swing set to collapse: the first node moves up and to the right, while the second node moves down and to the right, the horizontal motion being three times as large as the vertical. The structure can support forces $\mathbf{f}_1 = (\,f_1, g_1, h_1\,)^T$, $\mathbf{f}_2 = (\,f_2, g_2, h_2\,)^T$, if and only if the combined force vector $\mathbf{f}^\star$ is orthogonal to the mechanism vector $\mathbf{z}^\star$, and so

$$3\,(f_1 + f_2) - h_1 + h_2 = 0.$$

Thus, as long as the net horizontal force is in the y direction and the vertical forces on the two joints are equal, the structure will maintain its shape. Otherwise, a reinforcing bar, say from $\mathbf{a}_1$ to $\mathbf{a}_6$ (although this will interfere with the swinging!) or a pair of bars from the nodes to two new ground supports, will be required to completely stabilize the swing.

For a uniform downwards unit vertical force, $\mathbf{f} = (\,0, 0, -1, 0, 0, -1\,)^T$, a particular solution to (6.11) is

$$\mathbf{u}^\star = \left(\, \tfrac{13}{6} \;\; 0 \;\; -\tfrac{4}{3} \;\; \tfrac{11}{6} \;\; 0 \;\; 0 \,\right)^T$$

and the general solution $\mathbf{u} = \mathbf{u}^\star + t\,\mathbf{z}^\star$ is obtained by adding in an arbitrary element of the kernel. The resulting forces/elongations are uniquely determined,

$$\mathbf{y} = \mathbf{e} = A^\star\,\mathbf{u} = A^\star\,\mathbf{u}^\star = \left(-\tfrac{\sqrt{11}}{6} \;\; -\tfrac{\sqrt{11}}{6} \;\; -\tfrac{1}{3} \;\; -\tfrac{\sqrt{11}}{6} \;\; -\tfrac{\sqrt{11}}{6} \right)^T,$$

so that every bar is compressed, the middle one experiencing slightly more than half the stress of the outer supports.

If we stabilize the structure by adding in two vertical supports at the nodes, then the

new reduced incidence matrix

$$A^\star = \begin{pmatrix} \frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ \frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & \frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{11}} & -\frac{1}{\sqrt{11}} & \frac{3}{\sqrt{11}} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

has trivial kernel, indicating stabilization of the structure. The reduced stiffness matrix

$$K^\star = \begin{pmatrix} \frac{13}{11} & 0 & \frac{6}{11} & -1 & 0 & 0 \\ 0 & \frac{2}{11} & 0 & 0 & 0 & 0 \\ \frac{6}{11} & 0 & \frac{29}{11} & 0 & 0 & 0 \\ -1 & 0 & 0 & \frac{13}{11} & 0 & -\frac{6}{11} \\ 0 & 0 & 0 & 0 & \frac{2}{11} & 0 \\ 0 & 0 & 0 & -\frac{6}{11} & 0 & \frac{29}{11} \end{pmatrix}$$

is only slightly different than before, but this is enough to make it positive definite, $K^\star > 0$, and so allow arbitrary external forcing without collapse. Under the uniform vertical force, the internal forces are

$$\mathbf{y} = \mathbf{e} = A^\star \mathbf{u} = \begin{pmatrix} -\frac{\sqrt{11}}{10} & -\frac{\sqrt{11}}{10} & -\frac{1}{5} & -\frac{\sqrt{11}}{10} & -\frac{\sqrt{11}}{10} & -\frac{2}{5} & -\frac{2}{5} \end{pmatrix}^T.$$

Note the overall reductions in stress in the original bars; the two new vertical bars are now experiencing the largest amount of stress.

Chapter 7

Linear Functions and Linear Systems

We began this book by learning how to systematically solve systems of linear algebraic equations. This "elementary" problem formed our launching pad for developing the fundamentals of linear algebra. In its initial form, matrices and vectors were the primary focus of our study, but the theory was developed in a sufficiently general and abstract form that it can be immediately applied to many other important situations — particularly infinite-dimensional function spaces. Indeed, applied mathematics deals, not just with algebraic equations, but also differential equations, difference equations, integral equations, integro-differential equations, differential delay equations, control systems, and many, many other types of systems — not all of which, unfortunately, can be adequately developed in this introductory text. It is now time to assemble what we have learned about linear matrix systems and place the results in a suitably general framework that will lead to insight into the fundamental principles that govern completely general linear problems.

The most basic underlying object of linear systems theory is the vector space, and we have already seen that the elements of vector spaces can be vectors, or functions, or even vector-valued functions. The seminal ideas of span, linear independence, basis and dimension are equally applicable and equally vital in more general contexts, particularly function spaces. Just as vectors in Euclidean space are prototypes of general elements of vector spaces, matrices are also prototypes of much more general objects, known as *linear functions*. Linear functions are also known as linear maps or linear operators, particularly when we deal with function spaces, and include linear differential operators, linear integral operators, evaluation of a function or its derivative at a point, and many other basic operations on functions. Generalized functions, such as the delta function to be introduced in Chapter 11, are, in fact, properly formulated as linear operators on a suitable space of functions. As such, linear maps form the simplest class of functions on vector spaces. Nonlinear functions can often be closely approximated by linear functions, generalizing the calculus approximation of a function by its tangent line. As a result, linear functions must be thoroughly understood before any serious progress can be made in the vastly more complicated nonlinear world.

In geometry, linear functions are interpreted as linear transformations of space (or space-time), and, as such, lie at the foundations of motion of bodies, computer graphics and games, and the mathematical formulation of symmetry. Most basic geometrical transformations, including rotations, scalings, reflections, projections, shears and so on, are governed by linear transformations. However, translations require a slight generalization, known as an *affine function*. Linear operators on infinite-dimensional function spaces are the basic objects of quantum mechanics. Each quantum mechanical observable

(mass, energy, momentum) is formulated as a linear operator on an infinite-dimensional Hilbert space — the space of wave functions or states of the system. The dynamics of the quantum mechanical system is governed by the linear Schrödinger equation, [**100**, **104**]. It is remarkable that quantum mechanics is an entirely linear theory, whereas classical and relativistic mechanics are inherently nonlinear! The holy grail of modern physics — the unification of general relativity and quantum mechanics — is to resolve the apparent incompatibility of the microscopic and macroscopic physical regimes.

A *linear system* is just an equation satisfied by a linear function. The most basic linear system is a system of linear algebraic equations. Linear systems also include linear differential equations, linear boundary value problems, linear partial differential equations, and many, many others in a common conceptual framework. The fundamental idea of linear superposition and the relation between the solutions to inhomogeneous and homogeneous systems underlie the structure of the solution space of all linear systems. You have no doubt encountered many of these ideas in your first course on ordinary differential equations; they have also already appeared in our development of the theory underlying the solution of linear algebraic systems. The second part of this book will be devoted to solution techniques for particular classes of linear systems arising in applied mathematics.

7.1. Linear Functions.

We begin our study of linear functions with the basic definition. For simplicity, we shall concentrate on real linear functions between real vector spaces. Extending the concepts and constructions to complex linear functions on complex vector spaces is not difficult, and will be dealt with later.

Definition 7.1. Let V and W be real vector spaces. A function $L\colon V \to W$ is called *linear* if it obeys two basic rules:

$$L[\mathbf{v} + \mathbf{w}] = L[\mathbf{v}] + L[\mathbf{w}], \qquad L[c\,\mathbf{v}] = c\,L[\mathbf{v}], \tag{7.1}$$

We will call V the *domain space* and W the *target space*[†] for L.

In particular, setting $c = 0$ in the second condition implies that a linear function always maps the zero element in V to the zero element in W, so

$$L[\mathbf{0}] = \mathbf{0}. \tag{7.2}$$

We can readily combine the two defining conditions into a single rule

$$L[c\,\mathbf{v} + d\,\mathbf{w}] = c\,L[\mathbf{v}] + d\,L[\mathbf{w}], \qquad \text{for all} \qquad \mathbf{v}, \mathbf{w} \in V, \quad c, d \in \mathbb{R}, \tag{7.3}$$

that characterizes linearity of a function L. An easy induction proves that a linear function respects linear combinations, so

$$L[c_1\mathbf{v}_1 + \,\cdots\, + c_k\mathbf{v}_k] = c_1\,L[\mathbf{v}_1] + \,\cdots\, + c_k\,L[\mathbf{v}_k] \tag{7.4}$$

[†] The term "target" is used here to avoid later confusion with the range of L, which, in general, is a subspace of the target vector space W.

for any $c_1, \dots, c_k \in \mathbb{R}$ and $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$.

The interchangeable terms *linear map*, *linear operator* and, when $V = W$, *linear transformation* are all commonly used as alternatives to "linear function", depending on the circumstances and taste of the author. The term "linear operator" is particularly useful when the underlying vector space is a function space, so as to avoid confusing the two different uses of the word "function". As usual, we will sometimes refer to the elements of a vector space as "vectors" even though they might be functions or matrices or something else, depending upon the vector space being considered.

Example 7.2. The simplest linear function is the zero function $L[\mathbf{v}] \equiv \mathbf{0}$ which maps every element $\mathbf{v} \in V$ to the zero vector in W. Note that, in view of (7.2), this is the *only* constant linear function. A nonzero constant function is *not*, despite its evident simplicity, linear. Another simple but important linear function is the identity function $\mathrm{I} = \mathrm{I}_V : V \to V$ which leaves every vector unchanged: $\mathrm{I}[\mathbf{v}] = \mathbf{v}$. Slightly more generally, the operation of scalar multiplication $M_a[\mathbf{v}] = a\,\mathbf{v}$ by a fixed scalar $a \in \mathbb{R}$ defines a linear function from V to itself.

Example 7.3. Suppose $V = \mathbb{R}$. We claim that every linear function $L : \mathbb{R} \to \mathbb{R}$ has the form

$$y = L(x) = a\,x,$$

for some constant a. Indeed, writing $x \in \mathbb{R}$ as a scalar product $x = x \cdot 1$, and using the second property in (7.1), we find

$$L(x) = L(x \cdot 1) = x \cdot L(1) = a\,x, \qquad \text{where} \qquad a = L(1).$$

Therefore, the only scalar linear functions are those whose graph is a straight line passing through the origin.

Warning: Even though the graph of the function

$$y = a\,x + b, \tag{7.5}$$

is a straight line, this is *not* a linear function — unless $b = 0$ so the line goes through the origin. The correct name for a function of the form (7.5) is an *affine function*; see Definition 7.20 below.

Example 7.4. Let $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$. Let A be an $m \times n$ matrix. Then the function $L[\mathbf{v}] = A\,\mathbf{v}$ given by matrix multiplication is easily seen to be a linear function. Indeed, the requirements (7.1) reduce to the basic distributivity and scalar multiplication properties of matrix multiplication:

$$A(\mathbf{v} + \mathbf{w}) = A\,\mathbf{v} + A\,\mathbf{w}, \qquad A(c\,\mathbf{v}) = c\,A\,\mathbf{v}, \qquad \text{for all} \qquad \mathbf{v}, \mathbf{w} \in \mathbb{R}^n, \quad c \in \mathbb{R}.$$

In fact, *every* linear function between two Euclidean spaces has this form.

Theorem 7.5. *Every linear function $L : \mathbb{R}^n \to \mathbb{R}^m$ is given by matrix multiplication, $L[\mathbf{v}] = A\,\mathbf{v}$, where A is an $m \times n$ matrix.*

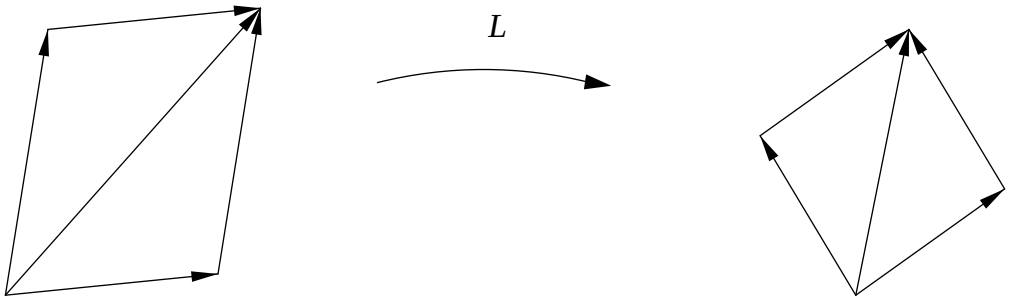


Figure 7.1. Linear Function on Euclidean Space.

Warning: Pay attention to the order of m and n. While A has size $m \times n$, the linear function L goes *from* $\mathbb{R}^n$ *to* $\mathbb{R}^m$.

Proof: The key idea is to look at what the linear function does to the basis vectors. Let $\mathbf{e}_1, \dots, \mathbf{e}_n$ be the standard basis of $\mathbb{R}^n$, and let $\widehat{\mathbf{e}}_1, \dots, \widehat{\mathbf{e}}_m$ be the standard basis of $\mathbb{R}^m$. (We temporarily place hats on the latter to avoid confusing the two.) Since $L[\mathbf{e}_j] \in \mathbb{R}^m$, we can write it as a linear combination of the latter basis vectors:

$$L[\mathbf{e}_j] = \mathbf{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} = a_{1j}\,\widehat{\mathbf{e}}_1 + a_{2j}\,\widehat{\mathbf{e}}_2 + \cdots + a_{mj}\,\widehat{\mathbf{e}}_m, \qquad j = 1, \dots, n. \tag{7.6}$$

Let us construct the $m \times n$ matrix

$$A = (\,\mathbf{a}_1\ \mathbf{a}_2\ \dots\ \mathbf{a}_n\,) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \tag{7.7}$$

whose columns are the image vectors (7.6). Using (7.4), we then compute the effect of L on a general vector $\mathbf{v} = (\,v_1, v_2, \dots, v_n\,)^T \in \mathbb{R}^n$:

$$L[\mathbf{v}] = L[v_1\,\mathbf{e}_1 + \cdots + v_n\,\mathbf{e}_n] = v_1\,L[\mathbf{e}_1] + \cdots + v_n\,L[\mathbf{e}_n] = v_1\,\mathbf{a}_1 + \cdots + v_n\,\mathbf{a}_n = A\mathbf{v}.$$

The final equality follows from our basic formula (2.14) connecting matrix multiplication and linear combinations. We conclude that the vector $L[\mathbf{v}]$ coincides with the vector $A\mathbf{v}$ obtained by multiplying $\mathbf{v}$ by the coefficient matrix A. *Q.E.D.*

The proof of Theorem 7.5 shows us how to construct the matrix representative of a given linear function $L\colon \mathbb{R}^n \to \mathbb{R}^m$. We merely assemble the image column vectors $\mathbf{a}_1 = L[\mathbf{e}_1], \dots, \mathbf{a}_n = L[\mathbf{e}_n]$ into an $m \times n$ matrix A.

Example 7.6. In the case of a function from $\mathbb{R}^n$ to itself, the two basic linearity conditions (7.1) have a simple geometrical interpretation. Since vector addition is the

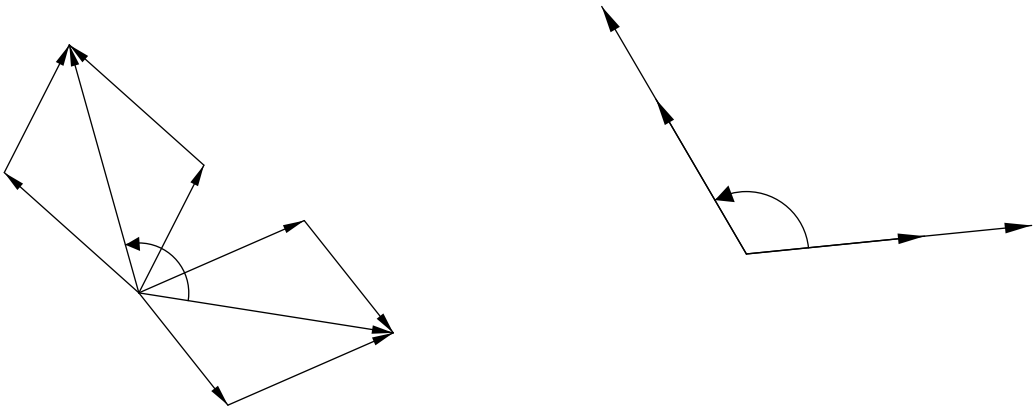


Figure 7.2. Linearity of Rotations.

same as completing the parallelogram indicated in Figure 7.1, the first linearity condition requires that L map parallelograms to parallelograms. The second linearity condition says that if we stretch a vector by a factor c, then its image under L must also be stretched by the same amount. Thus, one can often detect linearity by simply looking at the geometry of the function.

As a specific example, consider the function $R_\theta \colon \mathbb{R}^2 \to \mathbb{R}^2$ that rotates the vectors in the plane around the origin by a specified angle θ. This geometric transformation clearly preserves parallelograms, as well as stretching — see Figure 7.2 — and hence defines a linear function. In order to find its matrix representative, we need to find out where the basis vectors $\mathbf{e}_1, \mathbf{e}_2$ are mapped. Referring to Figure 7.3, we have

$$R_\theta[\,\mathbf{e}_1\,] = \cos\theta\,\mathbf{e}_1 + \sin\theta\,\mathbf{e}_2 = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}, \qquad R_\theta[\,\mathbf{e}_2\,] = -\sin\theta\,\mathbf{e}_1 + \cos\theta\,\mathbf{e}_2 = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}.$$

According to the general recipe (7.7), we assemble these two column vectors to obtain the matrix form of the rotation transformation, and so

$$R_\theta[\,\mathbf{v}\,] = A_\theta\,\mathbf{v}, \qquad \text{where} \qquad A_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{7.8}$$

Therefore, rotating a vector $\mathbf{v} = \begin{pmatrix} x \\ y \end{pmatrix}$ through angle θ gives the vector

$$\widehat{\mathbf{v}} = R_\theta[\,\mathbf{v}\,] = A_\theta\,\mathbf{v} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{pmatrix}$$

with coordinates

$$\widehat{x} = x\cos\theta - y\sin\theta, \qquad \widehat{y} = x\sin\theta + y\cos\theta.$$

These formulae can be proved directly, but, in fact, are a consequence of the underlying linearity of rotations.

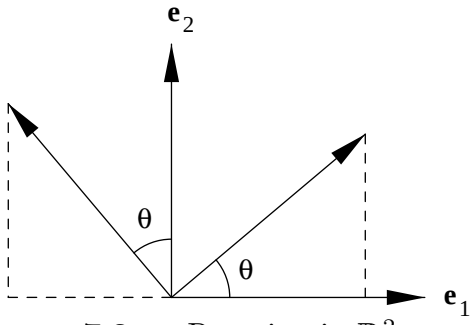


Figure 7.3. Rotation in $\mathbb{R}^2$.

Linear Operators

So far, we have concentrated on linear functions on Euclidean space, and discovered that they are all represented by matrices. For function spaces, there is a much wider variety of linear operators available, and a complete classification is out of the question. Let us look at some of the main representative examples that arise in applications.

Example 7.7. (*i*) Recall that $\mathrm{C}^0[a,b]$ denotes the vector space consisting of all continuous functions on the interval $[a,b]$. Evaluation of the function at a point, $L[f] = f(x_0)$, defines a linear operator $L\colon \mathrm{C}^0[a,b] \to \mathbb{R}$, because

$$L[cf + dg] = c\,f(x_0) + d\,g(x_0) = c\,L[f] + d\,L[g]$$

for any functions $f, g \in \mathrm{C}^0[a,b]$ and scalars (constants) c, d.

(*ii*) Another real-valued linear function is the integration operator

$$I[f] = \int_a^b f(x)\,dx. \tag{7.9}$$

Linearity of I is an immediate consequence of the basic integration identity

$$\int_a^b \big[\, c\,f(x) + d\,g(x) \,\big]\,dx = c\int_a^b f(x)\,dx \;+\; d\int_a^b g(x)\,dx,$$

which is valid for arbitrary integrable — which includes continuous — functions f, g and scalars c, d.

(*iii*) We have already seen that multiplication of functions by a fixed scalar $f(x) \mapsto c\,f(x)$ defines a linear map $M_c\colon \mathrm{C}^0[a,b] \to \mathrm{C}^0[a,b]$; the particular case $c = 1$ reduces to the identity transformation $\mathrm{I} = M_1$. More generally, if $a(x) \in \mathrm{C}^0[a,b]$ is a fixed continuous function, then the operation $M_a[f(x)] = a(x)\,f(x)$ of multiplication by a also defines a linear transformation $M_a\colon \mathrm{C}^0[a,b] \to \mathrm{C}^0[a,b]$.

(*iv*) Another important linear transformation is the indefinite integral

$$J[f] = \int_a^x f(y)\,dy. \tag{7.10}$$

According to the Fundamental Theorem of Calculus, the integral of a continuous function is continuously differentiable; therefore, $J\colon \mathrm{C}^0[a,b] \to \mathrm{C}^1[a,b]$ defines a linear operator from the space of continuous functions to the space of continuously differentiable functions.

(v) Vice versa, differentiation of functions is also a linear operation. To be precise, since not every continuous function can be differentiated, we take the domain space to be the vector space $\mathrm{C}^1[a,b]$ of continuously differentiable functions on the interval $[a,b]$. The derivative operator

$$D[f] = f' \tag{7.11}$$

defines a linear operator $D\colon \mathrm{C}^1[a,b] \to \mathrm{C}^0[a,b]$. This follows from the elementary differentiation formula

$$D[cf + dg] = (cf + dg)' = cf' + dg' = c\,D[f] + d\,D[g],$$

valid whenever c, d are constant.

The Space of Linear Functions

Given vector spaces V, W, we use $\mathcal{L}(V,W)$ to denote the set of all† linear functions $L\colon V \to W$. We claim that $\mathcal{L}(V,W)$ is itself a vector space. We add two linear functions $L, M \in \mathcal{L}(V,W)$ in the same way we add general functions: $(L+M)[\mathbf{v}] = L[\mathbf{v}] + M[\mathbf{v}]$. You should check that $L+M$ satisfies the linear function axioms (7.1) provided L and M do. Similarly, multiplication of a linear function by a scalar $c \in \mathbb{R}$ is defined so that $(cL)[\mathbf{v}] = cL[\mathbf{v}]$, again producing a linear function. The verification that $\mathcal{L}(V,W)$ satisfies the basic vector space axioms is left to the reader.

In particular, if $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, then Theorem 7.5 implies that we can identify $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ with the space $\mathcal{M}_{m\times n}$ of all $m \times n$ matrices. Addition of linear functions corresponds to matrix addition, while scalar multiplication coincides with the usual scalar multiplication of matrices. Therefore, the space of all $m \times n$ matrices forms a vector space — a fact we already knew. A basis for $\mathcal{M}_{m\times n}$ is given by the $m\,n$ matrices E_{ij}, $1 \le i \le m, 1 \le j \le n$, which have a single 1 in the (i,j) position and zeros everywhere else. Therefore, the dimension of $\mathcal{M}_{m\times n}$ is $m\,n$. Note that E_{ij} corresponds to the specific linear transformation mapping $\mathbf{e}_j \in \mathbb{R}^n$ to $\widehat{\mathbf{e}}_i \in \mathbb{R}^m$ and every other $\mathbf{e}_k \in \mathbb{R}^n$ to zero.

Example 7.8. The space of linear transformations of the plane, $\mathcal{L}(\mathbb{R}^2, \mathbb{R}^2)$ is identified with the space $\mathcal{M}_{2\times 2}$ of 2×2 matrices $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The standard basis of $\mathcal{M}_{2\times 2}$ consists of the $4 = 2 \cdot 2$ matrices

$$E_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad E_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad E_{22} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Indeed, we can uniquely write any other matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\,E_{11} + b\,E_{12} + c\,E_{21} + d\,E_{22},$$

as a linear combination of these four basis matrices.

† In infinite-dimensional situations, one usually imposes additional restrictions, e.g., continuity or boundedness of the linear operators. We can safely relegate these more subtle distinctions to a more advanced treatment of the subject. See [**122**] for a full discussion of the rather sophisticated analytical details, which do play an important role in serious quantum mechanical applications.

A particularly important case is when the target space of the linear functions is $\mathbb{R}$.

Definition 7.9. The *dual space* to a vector space V is defined as the vector space $V^* = \mathcal{L}(V, \mathbb{R})$ consisting of all real-valued linear functions $L\colon V \to \mathbb{R}$.

If $V = \mathbb{R}^n$, then every linear function $L\colon \mathbb{R}^n \to \mathbb{R}$ is given by multiplication by a $1 \times n$ matrix, i.e., a row vector. Explicitly,

$$L[\mathbf{v}] = \mathbf{a}\,\mathbf{v} = a_1 v_1 + \cdots + a_n v_n, \qquad \text{where} \qquad \mathbf{a} = (\,a_1\ a_2\ \ldots\ a_n\,), \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}.$$

Therefore, we can identify the dual space $(\mathbb{R}^n)^*$ with the space of *row* vectors with n entries. In light of this observation, the distinction between row vectors and column vectors is now seen to be much more sophisticated than mere semantics or notation. Row vectors should be viewed as real-valued linear functions — the *dual* objects to column vectors.

The *standard dual basis* $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ of $(\mathbb{R}^n)^*$ consists of the standard row basis vectors, namely $\boldsymbol{\varepsilon}_j$ is the row vector with 1 in the j^{th} slot and zeros elsewhere. The j^{th} dual basis element defines the linear function

$$E_j[\mathbf{v}] = \boldsymbol{\varepsilon}_j\,\mathbf{v} = v_j,$$

that picks off the j^{th} coordinate of $\mathbf{v}$ — with respect to the original basis $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Thus, the dimension of $V = \mathbb{R}^n$ and its dual $(\mathbb{R}^n)^*$ are both equal to n.

An inner product structure provides a mechanism for identifying a vector space and its dual. However, it should be borne in mind that this identification will depend upon the choice of inner product.

Theorem 7.10. *Let V be a finite-dimensional real inner product space. Then every linear function $L\colon V \to \mathbb{R}$ is given by an inner product*

$$L[\mathbf{v}] = \langle\, \mathbf{a}\,;\mathbf{v}\,\rangle \tag{7.12}$$

with a unique vector $\mathbf{a} \in V$. The correspondence between L and $\mathbf{a}$ allows us to identify $V^* \simeq V$.

Proof: Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis of V. (If necessary, we can use the Gram–Schmidt process to generate such a basis.) If we write $\mathbf{v} = x_1\mathbf{u}_1 + \cdots + x_n\mathbf{u}_n$, then, by linearity,

$$L[\mathbf{v}] = x_1 L[\mathbf{u}_1] + \cdots + x_n L[\mathbf{u}_n] = a_1 x_1 + \cdots + a_n x_n,$$

where $a_i = L[\mathbf{u}_i]$. On the other hand, if we write $\mathbf{a} = a_1\,\mathbf{u}_1 + \cdots + a_n\,\mathbf{u}_n$, then, by orthonormality of the basis,

$$\langle\, \mathbf{a}\,;\mathbf{v}\,\rangle = \sum_{i,j=1}^{n} a_i\, x_j \langle\, \mathbf{u}_i\,;\mathbf{u}_j\,\rangle = a_1 x_1 + \cdots + a_n x_n.$$

Thus equation (7.12) holds, which completes the proof. *Q.E.D.*

Remark: In the particular case when $V = \mathbb{R}^n$ is endowed with the standard dot product, then Theorem 7.10 identifies a row vector representing a linear function with the corresponding column vector obtained by transposition $\mathbf{a} \mapsto \mathbf{a}^T$. Thus, the naïve identification of a row and a column vector is, in fact, an indication of a much more subtle phenomenon that relies on the identification of $\mathbb{R}^n$ with its dual based on the Euclidean inner product. Alternative inner products will lead to alternative, more complicated, identifications of row and column vectors; see Exercise ■ for details.

Important: Theorem 7.10 is *not* true if V is infinite-dimensional. This fact will have important repercussions for the analysis of the differential equations of continuum mechanics, which will lead us immediately into the much deeper waters of generalized function theory. Details will be deferred until Section 11.2.

Composition of Linear Functions

Besides adding and multiplying by scalars, one can also compose linear functions.

Lemma 7.11. *Let V, W, Z be vector spaces. If $L: V \to W$ and $M: W \to Z$ are linear functions, then the composite function $M \circ L: V \to Z$, defined by $(M \circ L)[\mathbf{v}] = M[L[\mathbf{v}]]$ is linear.*

Proof: This is straightforward:

$$\begin{aligned}(M \circ L)[c\,\mathbf{v} + d\,\mathbf{w}] &= M[L[c\,\mathbf{v} + d\,\mathbf{w}]] = M[c\,L[\mathbf{v}] + d\,L[\mathbf{w}]] \\ &= c\,M[L[\mathbf{v}]] + d\,M[L[\mathbf{w}]] = c\,(M \circ L)[\mathbf{v}] + d\,(M \circ L)[\mathbf{w}],\end{aligned}$$

where we used, successively, the linearity of L and then of M. *Q.E.D.*

For example, if $L[\mathbf{v}] = A\,\mathbf{v}$ maps $\mathbb{R}^n$ to $\mathbb{R}^m$, and $M[\mathbf{w}] = B\,\mathbf{w}$ maps $\mathbb{R}^m$ to $\mathbb{R}^l$, so that A is an $m \times n$ matrix and B is a $l \times m$ matrix, then

$$(M \circ L)[\mathbf{v}] = M[L[\mathbf{v}]] = B(A\,\mathbf{v}) = (B\,A)\,\mathbf{v},$$

and hence the composition $M \circ L: \mathbb{R}^n \to \mathbb{R}^l$ corresponds to the $l \times n$ product matrix BA. In other words, on Euclidean space, composition of linear functions is the same as matrix multiplication!

As with matrix multiplication, composition of (linear) functions is not commutative. In general the order of the constituents makes a difference.

Example 7.12. Composing two rotations gives another rotation: $R_\varphi \circ R_\theta = R_{\varphi+\theta}$. In other words, if we first rotate by angle θ and then by angle φ, the net result is rotation by angle $\varphi + \theta$. On the matrix level of (7.8), this implies that $A_\varphi \cdot A_\theta = A_{\varphi+\theta}$, or, explicitly,

$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} \cos(\varphi+\theta) & -\sin(\varphi+\theta) \\ \sin(\varphi+\theta) & \cos(\varphi+\theta) \end{pmatrix}.$$

Multiplying out the left hand side, we deduce the well-known trigonometric addition formulae

$$\cos(\varphi+\theta) = \cos\varphi\,\cos\theta - \sin\varphi\,\sin\theta, \qquad \sin(\varphi+\theta) = \cos\varphi\,\sin\theta + \sin\varphi\,\cos\theta.$$

In fact, this computation constitutes a *bona fide* proof of these two identities!

Example 7.13. One can build up more sophisticated linear operators on function space by adding and composing simpler ones. In particular, the linear higher order derivative operators are obtained by composing the derivative operator D, defined in (7.11), with itself. For example,

$$D^2[f] = D \circ D[f] = D[f'] = f''$$

is the second derivative operator. One needs to exercise some care about the domain of definition, since not every function is differentiable. In general,

$$D^k[f] = f^{(k)}(x) \qquad \text{defines a linear operator} \qquad D^k \colon \mathrm{C}^n[a,b] \longrightarrow \mathrm{C}^{n-k}[a,b]$$

for any $n \geq k$.

If we compose D^k with the linear operation of multiplication by a fixed function $a(x) \in \mathrm{C}^{n-k}[a,b]$ we obtain the linear operator $f(x) \mapsto a D^k[f] = a(x)\, f^{(k)}(x)$. Finally, a general *linear ordinary differential operator* of order n

$$L = a_n(x)\, D^n + a_{n-1}(x)\, D^{n-1} + \cdots + a_1(x)\, D + a_0(x) \tag{7.13}$$

is obtained by summing such linear operators. If the coefficient functions $a_0(x), \dots, a_n(x)$ are continuous, then

$$L[u] = a_n(x)\, \frac{d^n u}{dx^n} + a_{n-1}(x)\, \frac{d^{n-1} u}{dx^{n-1}} + \cdots + a_1(x)\, \frac{du}{dx} + a_0(x) u \tag{7.14}$$

defines a linear operator from $\mathrm{C}^n[a,b]$ to $\mathrm{C}^0[a,b]$. The most important case — but certainly not the only one arising in applications — is when the coefficients $a_i(x) = c_i$ of L are all constant.

Inverses

The inverse of a linear function is defined in direct analogy with the Definition 1.13 of the inverse of a (square) matrix.

Definition 7.14. Let $L \colon V \to W$ be a linear function. If $M \colon W \to V$ is a linear function such that both composite functions

$$L \circ M = \mathrm{I}_W, \qquad\qquad M \circ L = \mathrm{I}_V, \tag{7.15}$$

are equal to the identity function, then we call M the *inverse* of L and write $M = L^{-1}$.

The two conditions (7.15) require

$$L[M[\mathbf{w}]] = \mathbf{w} \qquad \text{for all} \qquad \mathbf{w} \in W, \qquad \text{and} \qquad M[L[\mathbf{v}]] = \mathbf{v} \qquad \text{for all} \qquad \mathbf{v} \in V.$$

Of course, if $M = L^{-1}$ is the inverse of L, then $L = M^{-1}$ is the inverse of M since the conditions are symmetric.

If $V = \mathbb{R}^n$, $W = \mathbb{R}^m$, so that L and M are given by matrix multiplication, by A and B respectively, then the conditions (7.15) reduce to the usual conditions

$$A\, B = \mathrm{I}, \qquad\qquad B\, A = \mathrm{I},$$

for matrix inversion, cf. (1.33). Therefore $B = A^{-1}$ is the inverse matrix. In particular, for L to have an inverse, we need $m = n$ and its coefficient matrix A to be square and nonsingular.

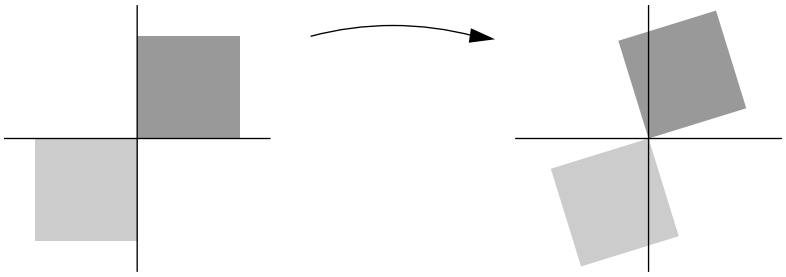


Figure 7.4. Rotation.

Example 7.15. The Fundamental Theorem of Calculus says, roughly, that differentiation $D[f] = f'$ and (indefinite) integration $J[f] = \int_a^x f(y)\,dy$ are "inverse" operations. More precisely, the derivative of the indefinite integral of f is equal to f, and hence

$$D[J[f(x)]] = \frac{d}{dx}\int_a^x f(y)\,dy = f(x).$$

In other words, the composition

$$D \circ J = \mathrm{I}_{\mathrm{C}^0[a,b]}$$

defines the identity operator on the function space $\mathrm{C}^0[a,b]$. On the other hand, if we integrate the derivative of a continuously differentiable function $f \in \mathrm{C}^1[a,b]$, we obtain

$$J[D[f(x)]] = J[f'(x)] = \int_a^x f'(y)\,dy = f(x) - f(a).$$

Therefore

$$J[D[f(x)]] = f(x) - f(a), \qquad \text{and so} \qquad J \circ D \neq \mathrm{I}_{\mathrm{C}^1[a,b]}$$

is *not* the identity operator. Therefore, differentiation, D, is a left inverse for integration, J, but not a right inverse!

This perhaps surprising phenomenon could not be anticipated from the finite-dimensional matrix theory. Indeed, if a matrix A has a left inverse B, then B is automatically a right inverse too, and we write $B = A^{-1}$ as the inverse of A. On an infinite-dimensional vector space, a linear operator may possess one inverse without necessarily the other. However, if both a left and a right inverse exist they must be equal; see Exercise ■.

If we restrict D to the subspace $V = \{\, f \,|\, f(a) = 0 \,\} \subset \mathrm{C}^1[a,b]$ consisting of all continuously differentiable functions that vanish at the left hand endpoint, then $J\colon \mathrm{C}^0[a,b] \to V$, and $D\colon V \to \mathrm{C}^0[a,b]$ are, by the preceding argument, inverse linear operators: $D \circ J = \mathrm{I}_{\mathrm{C}^0[a,b]}$, and $J \circ D = \mathrm{I}_V$. Note that $V \subsetneq \mathrm{C}^1[a,b] \subsetneq \mathrm{C}^0[a,b]$. Thus, we discover the curious and disconcerting infinite-dimensional phenomenon that J defines a one-to-one, invertible, linear map from a vector space $\mathrm{C}^0[a,b]$ to a proper subspace $V \subsetneq \mathrm{C}^0[a,b]$. This paradoxical situation *cannot* occur in finite dimensions. A linear map on a finite-dimensional vector space can only be invertible when the domain and target spaces have the same dimension, and hence its matrix is necessarily square!

7.2. Linear Transformations.

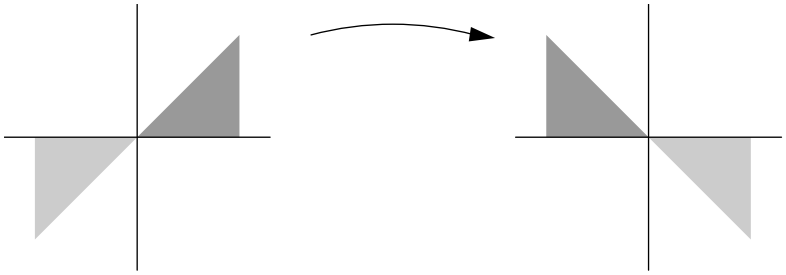


Figure 7.5. Reflection through the y axis.

A linear function $L\colon \mathbb{R}^n \to \mathbb{R}^n$ that maps n-dimensional Euclidean space to itself defines a *linear transformation*. As such, it can be assigned a geometrical interpretation that leads to further insight into the nature and scope of linear functions. The transformation L maps a point $\mathbf{x} \in \mathbb{R}^n$ to its image point $L[\mathbf{x}] = A\,\mathbf{x}$, where A is its $n \times n$ matrix representative. Many of the basic maps that appear in geometry, in computer graphics and computer gaming, in deformations of elastic bodies, in symmetry and crystallography, and in Einstein's special relativity, are defined by linear transformations. The two-, three- and four-dimensional (viewing time as a fourth dimension) cases are of particular importance.

Most of the important classes linear transformations already appear in the two-dimensional case. Every linear function $L\colon \mathbb{R}^2 \to \mathbb{R}^2$ has the form

$$L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a\,x + b\,y \\ c\,x + d\,y \end{pmatrix}, \qquad \text{where} \qquad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{7.16}$$

is an arbitrary 2×2 matrix. We have already encountered the rotation matrices

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{7.17}$$

whose effect is to rotate every vector in $\mathbb{R}^2$ through an angle θ; in Figure 7.4 we illustrate the effect on a couple of square regions in the plane. Planar rotation matrices coincide with the 2×2 proper orthogonal matrices, meaning matrices Q that satisfy

$$Q^T Q = \mathrm{I}\,, \qquad \det Q = +1. \tag{7.18}$$

The improper orthogonal matrices, i.e., those with determinant -1, define reflections. For example, the matrix

$$A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{corresponds to the linear transformation} \quad L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -x \\ y \end{pmatrix}, \tag{7.19}$$

which reflects the plane through the y axis; see Figure 7.5. It can be visualized by thinking of the y axis as a mirror. Another simple example is the improper orthogonal matrix

$$R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad \text{The corresponding linear transformation} \quad L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} y \\ x \end{pmatrix} \tag{7.20}$$

is a reflection through the diagonal line $y = x$, as illustrated in Figure 7.6.

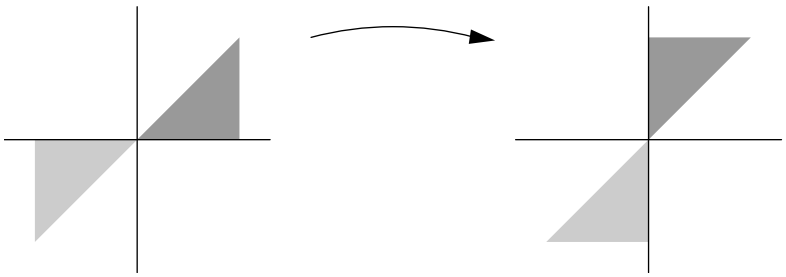


Figure 7.6. Reflection through the Diagonal.

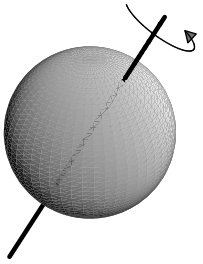


Figure 7.7. A Three–Dimensional Rotation.

A similar bipartite classification of orthogonal matrices carries over to three-dimensional (and even higher dimensional) space. The proper orthogonal matrices correspond to rotations and the improper to reflections, or, more generally, reflections combined with rotations. For example, the proper orthogonal matrix $\begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$ corresponds to a rotation through an angle θ around the z–axis, while $\begin{pmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{pmatrix}$ corresponds to a rotation through an angle φ around the y–axis. In general, a proper orthogonal matrix $Q = (\, \mathbf{u}_1 \ \mathbf{u}_2 \ \mathbf{u}_3 \,)$ with columns $\mathbf{u}_i = Q\mathbf{e}_i$ corresponds to the rotation in which the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are rotated to new positions given by the orthonormal basis $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$. It can be shown — see Exercise ■ — that *every* 3×3 orthogonal matrix corresponds to a rotation around a line through the origin in $\mathbb{R}^3$ — the axis of the rotation, as sketched in Figure 7.7.

Since the product of two (proper) orthogonal matrices is also (proper) orthogonal, this implies that the composition of two rotations is also a rotation. Unlike the planar case, the order in which the rotations are performed is important! Multiplication of $n \times n$ orthogonal matrices is *not* commutative for $n \geq 3$. For example, rotating first around the z–axis and then rotating around the y–axis does *not* have the same effect as first rotating around the y–axis and then rotating first around the z–axis. If you don't believe this, try it out with a solid object, e.g., this book, and rotate through 90°, say, around each axis; the final configuration of the book will depend upon the order in which you do the

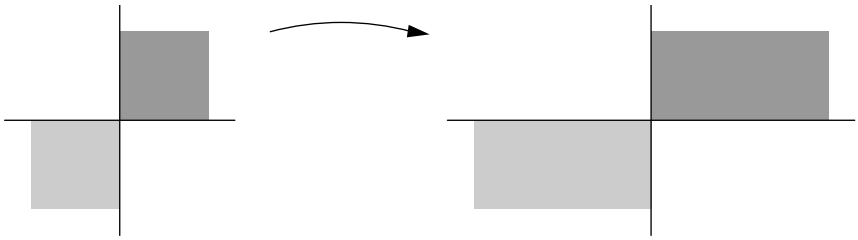


Figure 7.8. Stretch along the x–axis.

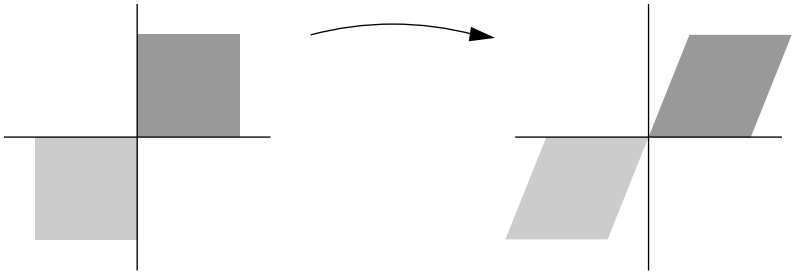


Figure 7.9. Shear in the x direction.

rotations. Then prove this mathematically by showing that the two rotation matrices do not commute.

Other important linear transformations arise from elementary matrices. First, the elementary matrices corresponding to the third type of row operations — multiplying a row by a scalar — correspond to simple stretching transformations. For example, if

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \text{then the linear transformation} \qquad L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2\,x \\ y \end{pmatrix}$$

has the effect of stretching along the x axis by a factor of 2; see Figure 7.8. A matrix with a negative diagonal entry corresponds to a reflection followed by a stretch. For example, the elementary matrix (7.19) gives an example of a pure reflection, while the more general elementary matrix

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

can be written as the product of a reflection through the y axis followed by a stretch along the x axis. In this case, the order of these operations is immaterial.

For 2×2 matrices, there is only one type of row interchange matrix, namely the matrix (7.20) that yields a reflection through the diagonal $y = x$. The elementary matrices of Type #1 correspond to *shearing transformations* of the plane. For example, the matrix

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \qquad \text{represents the linear transformation} \qquad L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x + 2\,y \\ y \end{pmatrix},$$

which has the effect of shearing the plane along the x–axis. The constant 2 will be called the *shear factor*, which can be either positive or negative. Each point moves parallel to the x axis by an amount proportional to its (signed) distance from the axis; see Figure 7.9.

Similarly, the elementary matrix

$$\begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} \qquad \text{represents the linear transformation} \qquad L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y - 3\,x \end{pmatrix},$$

which represents a shear along the y axis. Shears map rectangles to parallelograms; distances are altered, but areas are unchanged.

All of the preceding linear maps are invertible, and so represented by nonsingular matrices. Besides the zero map/matrix, which sends every point $\mathbf{x} \in \mathbb{R}^2$ to the origin, the simplest singular map is

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{corresponding to the linear transformation} \qquad L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix},$$

which is merely the orthogonal projection of the vector $(\,x, y\,)^T$ onto the x–axis. Other rank one matrices represent various kinds of projections from the plane to a line through the origin; see Exercise ■ for details.

A similar classification of linear maps appears in higher dimensions. The linear transformations constructed from elementary matrices can be built up from the following four basic types:

(i) A stretch in a single coordinate direction.

(ii) A reflection through a coordinate plane.

(iii) A reflection through a diagonal plane,

(iv) A shear along a coordinate axis.

Moreover, we already proved that every nonsingular matrix can be written as a product of elementary matrices; see (1.41). This has the remarkable consequence that *every* linear transformation can be constructed from a sequence of elementary stretches, reflections, and shears. In addition, there is one further, non-invertible type of basic linear transformation:

(v) An orthogonal projection onto a lower dimensional subspace.

All possible linear transformations of $\mathbb{R}^n$ can be built up, albeit non-uniquely, as a combination of these five basic types.

Example 7.16. Consider the matrix $A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}$ corresponding to a plane rotation through $\theta = 30^\circ$, cf. (7.17). Rotations are *not* elementary linear transformations. To express this particular rotation as a product of elementary matrices, we need to perform a Gauss-Jordan row reduction to reduce it to the identity matrix. Let us indicate the basic steps:

$$E_1 = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\sqrt{3}} & 1 \end{pmatrix}, \qquad E_1\,A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix},$$

$$E_2 = \begin{pmatrix} 1 & 0 \\ 0 & \frac{\sqrt{3}}{2} \end{pmatrix}, \qquad E_2\,E_1\,A = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ 0 & 1 \end{pmatrix},$$

$$E_3 = \begin{pmatrix} \frac{2}{\sqrt{3}} & 0 \\ 0 & 1 \end{pmatrix}, \qquad E_3 E_2 E_1 A = \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} \\ 0 & 1 \end{pmatrix},$$
$$E_4 = \begin{pmatrix} 1 & \frac{1}{\sqrt{3}} \\ 0 & 1 \end{pmatrix}, \qquad E_4 E_3 E_2 E_1 A = \mathrm{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and hence

$$\begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} = A = E_1^{-1} E_2^{-1} E_3^{-1} E_4^{-1} = \begin{pmatrix} 1 & 0 \\ \frac{1}{\sqrt{3}} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{3}}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{\sqrt{3}} \\ 0 & 1 \end{pmatrix}.$$

Therefore, a 30° rotation can be effected by performing the following composition of elementary transformations in the prescribed order:

(1) First, a shear in the x–direction with shear factor $-\frac{1}{\sqrt{3}}$,

(2) Then a stretch in the direction of the x–axis by a factor of $\frac{\sqrt{3}}{2}$,

(3) Then a stretch (or, rather, a contraction) in the y-direction by the reciprocal factor $\frac{2}{\sqrt{3}}$,

(4) Finally, a shear in the direction of the y–axis with shear factor $\frac{1}{\sqrt{3}}$.

The fact that the combination of these special transformations results in a pure rotation is surprising and non-obvious. Similar decompositions can be systematically found for higher dimensional linear transformations.

Change of Basis

Sometimes a linear transformation represents an elementary geometrical transformation, but this is not evident because the matrix happens to be written in the wrong coordinates. The characterization of linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ as multiplication by $m \times n$ matrices in Theorem 7.5 relies on using the standard bases of the domain and target spaces. In many cases, the standard basis is not particularly well adapted to the linear transformation, and one can often gain more insight by adopting a more suitable basis. Therefore, we need to understand how to write a given linear transformation in a new basis.

The following general result says that, in *any* basis, a linear function on finite-dimensional vector spaces can be realized by matrix multiplication of the coordinates. But the particular matrix representative will depend upon the choice of basis.

Theorem 7.17. *Let $L\colon V \to W$ be a linear function. Suppose V has basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ and W has basis $\mathbf{w}_1, \dots, \mathbf{w}_m$. We can write*

$$\mathbf{v} = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n \in V, \qquad \mathbf{w} = y_1 \mathbf{w}_1 + \cdots + y_m \mathbf{w}_m \in W,$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ are the coordinates of $\mathbf{v}$ relative to the chosen basis on V and $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$ are those of $\mathbf{w}$ relative to its basis. Then the linear function $\mathbf{w} = L[\mathbf{v}]$ is given in these coordinates by multiplication, $\mathbf{y} = B\,\mathbf{x}$, by an $m \times n$ matrix B.

Proof: We mimic the proof of Theorem 7.5, replacing the standard basis vectors by more general basis vectors. In other words, we should apply L to the basis vectors of V and express the result as a linear combination of the basis vectors in W. Specifically, we write $L[\mathbf{v}_j] = \sum_{i=1}^{m} b_{ij}\,\mathbf{w}_i$. The coefficients b_{ij} form the entries of the desired coefficient matrix. Indeed, by linearity

$$L[\mathbf{v}] = L[x_1\,\mathbf{v}_1 + \,\cdots\, + x_n\,\mathbf{v}_n] = x_1\,L[\mathbf{v}_1] + \,\cdots\, + x_n\,L[\mathbf{v}_n] = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} b_{ij}\,x_j\right)\mathbf{w}_i,$$

and so $y_i = \sum_{j=1}^{n} b_{ij}\,x_j$ as claimed. *Q.E.D.*

Suppose that the linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^m$ is represented by a certain $m \times n$ matrix A relative to the standard bases $\mathbf{e}_1, \dots, \mathbf{e}_n$ and $\widehat{\mathbf{e}}_1, \dots, \widehat{\mathbf{e}}_m$ of the domain and target spaces. If we introduce new bases for $\mathbb{R}^n$ and $\mathbb{R}^m$ then the *same* linear transformation may have a completely different matrix representation. Therefore, different matrices may represent the same underlying linear transformation, with respect to different bases.

Example 7.18. Consider the linear transformation $L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x - y \\ 2\,x + 4\,y \end{pmatrix}$ which we write in the standard, Cartesian coordinates x, y on $\mathbb{R}^2$. The corresponding coefficient matrix $A = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix}$ is the matrix representation of L — relative to the standard basis $\mathbf{e}_1, \mathbf{e}_2$ of $\mathbb{R}^2$. This means that

$$L[\mathbf{e}_1] = \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \mathbf{e}_1 + 2\,\mathbf{e}_2, \qquad L[\mathbf{e}_2] = \begin{pmatrix} -1 \\ 4 \end{pmatrix} = -\,\mathbf{e}_1 + 4\,\mathbf{e}_2.$$

Let us see what happens if we replace the standard basis by the alternative basis

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}.$$

What is the corresponding matrix formulation of the same linear transformation? According to the recipe of Theorem 7.17, we must compute

$$L[\mathbf{v}_1] = \begin{pmatrix} 2 \\ -2 \end{pmatrix} = 2\,\mathbf{v}_1, \qquad L[\mathbf{v}_2] = \begin{pmatrix} 3 \\ -6 \end{pmatrix} = 3\,\mathbf{v}_2.$$

The linear transformation acts by stretching in the direction $\mathbf{v}_1$ by a factor of 2 and simultaneously stretching in the direction $\mathbf{v}_2$ by a factor of 3. Therefore, the matrix form of L with respect to this new basis is the diagonal matrix $D = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. In general,

$$L[a\,\mathbf{v}_1 + b\,\mathbf{v}_2] = 2\,a\,\mathbf{v}_1 + 3\,b\,\mathbf{v}_2,$$

whose effect is to multiply the new basis coordinates $\mathbf{a} = (\,a, b\,)^T$ by the diagonal matrix D. Both A and D represent the *same* linear transformation — the former in the standard

basis and the latter in the new basis. The simple geometry of this linear transformation is thereby exposed through the inspired choice of an adapted basis. The secret behind the choice of such well-adapted bases will be revealed in Chapter 8.

How does one effect a change of basis in general? According to formula (2.22), if $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a new basis of $\mathbb{R}^n$, then the coordinates $\mathbf{y} = (\, y_1, y_2, \dots, y_n \,)^T$ of a vector

$$\mathbf{x} = y_1 \mathbf{v}_1 + y_2 \mathbf{v}_2 + \, \cdots \, + y_n \mathbf{v}_n$$

are found by solving the linear system

$$S\,\mathbf{y} = \mathbf{x}, \qquad \text{where} \qquad S = (\, \mathbf{v}_1 \; \mathbf{v}_2 \; \dots \; \mathbf{v}_n \,) \tag{7.21}$$

is the nonsingular $n \times n$ matrix whose columns are the basis vectors.

Consider first a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^n$ from $\mathbb{R}^n$ to itself. When written in terms of the standard basis, $L[\mathbf{x}] = A\,\mathbf{x}$ has a certain $n \times n$ coefficient matrix A. To change to the new basis $\mathbf{v}_1, \dots, \mathbf{v}_n$, we use (7.21) to rewrite the standard $\mathbf{x}$ coordinates in terms of the new $\mathbf{y}$ coordinates. We also need to write the target vector $\mathbf{f} = A\,\mathbf{x}$ in terms of the new coordinates, which requires $\mathbf{f} = S\,\mathbf{g}$. Therefore, the new target coordinates are expressed in terms of the new domain coordinates via

$$\mathbf{g} = S^{-1}\mathbf{f} = S^{-1}A\,\mathbf{x} = S^{-1}A\,S\,\mathbf{y} = B\,\mathbf{y}.$$

Therefore, in the new basis, the matrix form of our linear transformation is

$$B = S^{-1}A\,S. \tag{7.22}$$

Two matrices A and B which are related by such an equation for some nonsingular matrix S are called *similar*. Similar matrices represent the *same* linear transformation, but relative to *different* bases of the underlying vector space $\mathbb{R}^n$.

Returning to the preceding example, we assemble the new basis vectors to form the change of basis matrix $S = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$, and verify that

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

reconfirming our earlier computation.

More generally, a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^m$ is represented by an $m \times n$ matrix A with respect to the standard bases on both the domain and target spaces. What happens if we introduce a new basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ on the domain space $\mathbb{R}^n$ *and* a new basis $\mathbf{w}_1, \dots, \mathbf{w}_m$ on the target space $\mathbb{R}^m$? Arguing as above, we conclude that the matrix representative of L with respect to these new bases is given by

$$B = T^{-1}\,A\,S, \tag{7.23}$$

where $S = (\, \mathbf{v}_1 \; \mathbf{v}_2 \; \dots \; \mathbf{v}_n \,)$ is the domain basis matrix, while $T = (\, \mathbf{w}_1 \; \mathbf{w}_2 \; \dots \; \mathbf{w}_m \,)$ is the range basis matrix.

In particular, suppose that the linear transformation has rank

$$r = \dim \operatorname{rng} L = \dim \operatorname{corng} L.$$

Let us choose a basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ of $\mathbb{R}^n$ such that $\mathbf{v}_1, \dots, \mathbf{v}_r$ form a basis of $\operatorname{corng} L$ while $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ form a basis for $\ker L = (\operatorname{corng} L)^\perp$. According to Proposition 5.54, the image vectors $\mathbf{w}_1 = L[\mathbf{v}_1], \dots, \mathbf{w}_r = L[\mathbf{v}_r]$ form a basis for $\operatorname{rng} L$, while $L[\mathbf{v}_{r+1}] = \cdots = L[\mathbf{v}_n] = \mathbf{0}$. We further choose a basis $\mathbf{w}_{r+1}, \dots, \mathbf{w}_m$ for $\operatorname{coker} L = (\operatorname{rng} L)^\perp$, and note that the combination $\mathbf{w}_1, \dots, \mathbf{w}_m$ forms a basis for $\mathbb{R}^m$. The matrix form of L relative to these two adapted bases is simply

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}. \tag{7.24}$$

In this matrix, the first r rows have a single 1 in the diagonal slot, indicating that the first r basis vectors of the domain space are mapped to the first r basis vectors of the target space while the last $m - r$ rows are all zero, indicating that the last $n - r$ basis vectors in the domain are all mapped to $\mathbf{0}$. Thus, by a suitable choice of bases on both the domain and target spaces, any linear transformation has an extremely simple *canonical form*.

Example 7.19. According to the illustrative example following Theorem 2.47, the matrix

$$A = \begin{pmatrix} 2 & -1 & 1 & 2 \\ -8 & 4 & -6 & -4 \\ 4 & -2 & 3 & 2 \end{pmatrix}$$

has rank 2. Based on the calculations, we choose the domain space basis

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 2 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 0 \\ -2 \\ 4 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} \frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} -2 \\ 0 \\ 2 \\ 1 \end{pmatrix},$$

noting that $\mathbf{v}_1, \mathbf{v}_2$ are a basis for the row space $\operatorname{corng} A$, while $\mathbf{v}_3, \mathbf{v}_4$ are a basis for $\ker A$. For our basis of the target space, we first compute $\mathbf{w}_1 = A\mathbf{v}_1$ and $\mathbf{w}_2 = A\mathbf{v}_2$, which form a basis for $\operatorname{rng} A$. We supplement these by the single basis vector $\mathbf{w}_3$ for $\operatorname{coker} A$, and so

$$\mathbf{w}_1 = \begin{pmatrix} 10 \\ -34 \\ 17 \end{pmatrix}, \qquad \mathbf{w}_2 = \begin{pmatrix} 6 \\ -4 \\ 2 \end{pmatrix}, \qquad \mathbf{w}_3 = \begin{pmatrix} 0 \\ \frac{1}{2} \\ 1 \end{pmatrix},$$

In terms of these two bases, the canonical matrix form of the linear function is

$$B = T^{-1}AS = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where the bases are assembled to form the matrices

$$S = \begin{pmatrix} 2 & 0 & \frac{1}{2} & -2 \\ -1 & 0 & 1 & 0 \\ 1 & -2 & 0 & 2 \\ 2 & 4 & 0 & 1 \end{pmatrix}, \qquad T = \begin{pmatrix} 10 & 6 & 0 \\ -34 & -4 & \frac{1}{2} \\ 17 & 2 & 1 \end{pmatrix}.$$

7.3. Affine Transformations and Isometries.

Not every transformation of importance in geometrical applications arises as a linear function. A simple example is a *translation*, where all the points in $\mathbb{R}^n$ are moved in the same direction by a common distance. The function that accomplishes this is

$$T[\mathbf{x}] = \mathbf{x} + \mathbf{a}, \qquad \mathbf{x} \in \mathbb{R}^n, \tag{7.25}$$

where $\mathbf{a} \in \mathbb{R}^n$ is a fixed vector that determines the direction and the distance that the points are translated. Except in the trivial case $\mathbf{a} = \mathbf{0}$, the translation T is *not* a linear function because

$$T[\mathbf{x} + \mathbf{y}] = \mathbf{x} + \mathbf{y} + \mathbf{a} \neq T[\mathbf{x}] + T[\mathbf{y}] = \mathbf{x} + \mathbf{y} + 2\,\mathbf{a}.$$

Or, even more simply, one notes that $T[\mathbf{0}] = \mathbf{a} \neq \mathbf{0}$.

Combining translations and linear functions leads us to an important class of geometrical transformations.

Definition 7.20. A function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ of the form

$$F[\mathbf{x}] = A\,\mathbf{x} + \mathbf{b}, \tag{7.26}$$

where A is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^n$ a fixed vector, is called an *affine function*.

For example, every affine function from $\mathbb{R}$ to itself has the form

$$f(x) = \alpha\, x + \beta. \tag{7.27}$$

As mentioned earlier, even though the graph of $f(x)$ is a straight line, f is *not* a linear function — unless $\beta = 0$, and the line goes through the origin. Thus, to be technically correct, we should refer to (7.27) as an *affine scalar function*.

Example 7.21. The affine function

$$F(x, y) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -y + 1 \\ x - 2 \end{pmatrix}$$

has the effect of first rotating the plane $\mathbb{R}^2$ by 90° about the origin, and then translating by the vector $(\,1, -2\,)^T$. The reader may enjoy proving that this combination has the same effect as just rotating the plane through an angle of 90° centered at the point $\left(\,\frac{3}{4}, -\frac{1}{2}\,\right)$. See Exercise ∎.

The composition of two affine functions is again an affine function. Specifically, given $F[\mathbf{x}] = A\,\mathbf{x} + \mathbf{a}$, $G[\mathbf{y}] = B\,\mathbf{y} + \mathbf{b}$, then

$$\begin{aligned}(G \circ F)[\mathbf{x}] &= G[F[\mathbf{x}]] = G[A\,\mathbf{x} + \mathbf{a}] \\ &= B\,(A\,\mathbf{x} + \mathbf{a}) + \mathbf{b} = C\,\mathbf{x} + \mathbf{c},\end{aligned} \qquad \text{where} \quad C = B\,A, \quad \mathbf{c} = B\,\mathbf{a} + \mathbf{b}. \tag{7.28}$$

Note that the coefficient matrix of the composition is the product of the coefficient matrices, but the resulting vector of translation is *not* the sum the two translation vectors!

Isometry

A transformation that preserves distance is known as a *rigid motion*, or, more abstractly, as an *isometry*. We already encountered the basic rigid motions in Chapter 6 — they are the translations and the rotations.

Definition 7.22. A function $F\colon V \to V$ is called an *isometry* on a normed vector space if it preserves the distance:

$$d(F[\mathbf{v}], F[\mathbf{w}]) = d(\mathbf{v}, \mathbf{w}) \qquad \text{for all} \qquad \mathbf{v}, \mathbf{w} \in V. \tag{7.29}$$

Since the distance between points is just the norm of the vector between them, $d(\mathbf{v}, \mathbf{w}) = \| \mathbf{v} - \mathbf{w} \|$, cf. (3.29), the isometry condition (7.29) can be restated as

$$\big\| F[\mathbf{v}] - F[\mathbf{w}] \big\| = \| \mathbf{v} - \mathbf{w} \| \qquad \text{for all} \qquad \mathbf{v}, \mathbf{w} \in V. \tag{7.30}$$

Clearly, any translation

$$T[\mathbf{v}] = \mathbf{v} + \mathbf{a}, \qquad \text{where} \qquad \mathbf{a} \in V \quad \text{is a fixed vector}$$

defines an isometry since $T[\mathbf{v}] - T[\mathbf{w}] = \mathbf{v} - \mathbf{w}$. A linear transformation $L\colon V \to V$ defines an isometry if and only if

$$\big\| L[\mathbf{v}] \big\| = \| \mathbf{v} \| \qquad \text{for all} \qquad \mathbf{v} \in V, \tag{7.31}$$

because, by linearity, $L[\mathbf{v}] - L[\mathbf{w}] = L[\mathbf{v} - \mathbf{w}]$. More generally, an affine transformation $F[\mathbf{v}] = L[\mathbf{v}] + \mathbf{a}$ is an isometry if and only if its linear part $L[\mathbf{v}]$ is.

For the standard Euclidean norm on $V = \mathbb{R}^n$, the linear isometries consist of rotations and reflections. Both are characterized by orthogonal matrices, the rotations having determinant $+1$, while the reflections have determinant -1.

Proposition 7.23. *A linear transformation $L[\mathbf{x}] = Q\,\mathbf{v}$ defines a Euclidean isometry of $\mathbb{R}^n$ if and only if Q is an orthogonal matrix.*

Proof: The linear isometry condition (7.31) requires that

$$\big\| Q\,\mathbf{x} \big\|^2 = (Q\,\mathbf{x})^T Q\,\mathbf{x} = \mathbf{x}^T Q^T Q\,\mathbf{x} = \mathbf{x}^T \mathbf{x} = \| \mathbf{x} \|^2 \qquad \text{for all} \qquad \mathbf{x} \in \mathbb{R}^n. \tag{7.32}$$

Clearly this holds if and only if $Q^T Q = \mathrm{I}$, which is precisely the condition (5.30) that Q be an orthogonal matrix. *Q.E.D.*

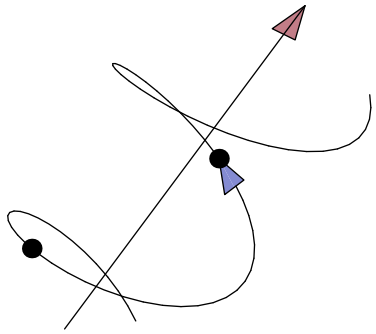


Figure 7.10. A Screw.

Remark: It can be proved, [**153**], that the most general Euclidean isometry of $\mathbb{R}^n$ is an affine transformation $F[\mathbf{x}] = Q\,\mathbf{x} + \mathbf{a}$ where Q is an orthogonal matrix and $\mathbf{a}$ is a constant vector. Therefore, every Euclidean isometry is a combination of translations, rotations and reflections. The *proper isometries* correspond to the rotations, with $\det Q = 1$, and can be realized as physical motions; improper isometries, with $\det Q = -1$, are then obtained by reflection in a mirror.

The isometries of $\mathbb{R}^2$ and $\mathbb{R}^3$ are fundamental to the understanding of how objects move in three-dimensional space. Basic computer graphics and animation require efficient implementation of rigid isometries in three-dimensional space, coupled with appropriate (nonlinear) perspective maps prescribing the projection of three-dimensional objects onto a two-dimensional viewing screen.

There are three basic types of proper affine isometries. First are the translations $F[\mathbf{x}] = \mathbf{x}+\mathbf{a}$ in a fixed direction $\mathbf{a}$. Second are the rotations. For example, $F[\mathbf{x}] = Q\,\mathbf{x}$ with $\det Q = 1$ represent rotations around the origin, while the more general case $F[\mathbf{x}] = Q(\mathbf{x}-\mathbf{b}) + \mathbf{b} = Q\,\mathbf{x} + (\,\mathrm{I} - Q)\mathbf{b}$ is a rotation around the point $\mathbf{b}$. Finally, the *screw motions* are affine maps of the form $F[\mathbf{x}] = Q\,\mathbf{x}+\mathbf{a}$ where the orthogonal matrix Q represents a rotation through an angle θ around a fixed axis $\mathbf{a}$, which is also the direction of the translation term; see Figure 7.10. For example, $F\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ a \end{pmatrix}$ represents a vertical screw along the the z–axis through an angle θ by an distance a. As its name implies, a screw represents the motion of a point on the head of a screw. It can be proved, cf. Exercise ∎, that every proper isometry of $\mathbb{R}^3$ is either a translation, a rotation, or a screw.

7.4. Linear Systems.

The abstract notion of a linear system serves to unify, in a common conceptual framework, linear systems of algebraic equations, linear ordinary differential equations, linear

partial differential equations, linear boundary value problems, and a wide variety of other linear problems in mathematics and its applications. The idea is simply to replace matrix multiplication by a general linear function. Many of the structural results we learned in the matrix context have, when suitably formulated, direct counterparts in these more general situations, thereby shedding some light on the nature of their solutions.

Definition 7.24. A *linear system* is an equation of the form

$$L[\mathbf{u}] = \mathbf{f}, \tag{7.33}$$

in which $L\colon V \to W$ is a linear function between vector spaces, the right hand side $\mathbf{f} \in W$ is an element of the target space, while the desired solution $\mathbf{u} \in V$ belongs to the domain space. The system is *homogeneous* if $\mathbf{f} = \mathbf{0}$; otherwise, it is called *inhomogeneous*.

Example 7.25. If $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, then, according to Theorem 7.5, every linear function $L\colon \mathbb{R}^n \to \mathbb{R}^m$ is given by matrix multiplication: $L[\mathbf{u}] = A\,\mathbf{u}$. Therefore, in this particular case, every linear system is a matrix system, namely $A\,\mathbf{u} = \mathbf{f}$.

Example 7.26. A *linear ordinary differential equation* takes the form $L[u] = f$, where L is an n^{th} order linear differential operator of the form (7.13), and the right hand side is, say, a continuous function. Written out, the differential equation takes the familiar form

$$L[u] = a_n(x)\,\frac{d^n u}{dx^n} + a_{n-1}(x)\,\frac{d^{n-1}u}{dx^{n-1}} + \,\cdots\, + a_1(x)\,\frac{du}{dx} + a_0(x)u = f(x). \tag{7.34}$$

You should already have some familiarity with solving the constant coefficient case. Appendix C describes a method for constructing series representations for the solutions to more general, non-constant coefficient equations.

Example 7.27. Let $K(x, y)$ be a function of two variables which is continuous for all $a \le x, y \le b$. Then the integral

$$I_K[u] = \int_a^b K(x, y)\,u(y)\,dy$$

defines a linear operator $I_K\colon \mathrm{C}^0[a, b] \to \mathrm{C}^0[a, b]$, known as an *integral transform*. Important examples include the Fourier and Laplace transforms, to be discussed in Chapter 13. Finding the inverse transform requires solving a *linear integral equation* $I_K[u] = f$, which has the explicit form

$$\int_a^b K(x, y)\,u(y)\,dy = f(x).$$

Example 7.28. One can combine linear maps to form more complicated, "mixed" types of linear systems. For example, consider a typical initial value problem

$$u'' + u' - 2\,u = x, \qquad u(0) = 1, \qquad u'(0) = -1, \tag{7.35}$$

for a scalar unknown function $u(x)$. The differential equation can be written as a linear system

$$L[u] = x, \qquad \text{where} \qquad L[u] = (D^2 + D - 2)[u] = u'' + u' - 2\,u$$

is a linear, constant coefficient differential operator. If we further define

$$M[u] = \begin{pmatrix} L[u] \\ u(0) \\ u'(0) \end{pmatrix} = \begin{pmatrix} u''(x) + u'(x) - 2\,u(x) \\ u(0) \\ u'(0) \end{pmatrix},$$

then M defines a linear map whose domain is the space C^2 of twice continuously differentiable functions, and whose range is the vector space V consisting of all triples[†] $\mathbf{v} = \begin{pmatrix} f(x) \\ a \\ b \end{pmatrix}$, where $f \in \mathrm{C}^0$ is a continuous function and $a, b \in \mathbb{R}$ are real constants. You should convince yourself that V is indeed a vector space under the evident addition and scalar multiplication operations. In this way, we can write the initial value problem (7.35) in linear systems form as $M[u] = \mathbf{f}$, where $\mathbf{f} = (\,x, 1, -1\,)^T$.

A similar construction applies to linear boundary value problems. For example, the boundary value problem

$$u'' + u = e^x, \qquad u(0) = 1, \qquad u(1) = 2,$$

is in the form of a linear system

$$M[u] = \mathbf{f}, \qquad \text{where} \qquad M[u] = \begin{pmatrix} u''(x) + u(x) \\ u(0) \\ u(1) \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} e^x \\ 1 \\ 2 \end{pmatrix}.$$

Note that $M\colon \mathrm{C}^2 \to V$ defines a linear map having the preceding domain and target spaces.

The Superposition Principle

Before attempting to tackle general inhomogeneous linear systems, it will help to look first at the homogeneous version. The most important fact is that homogeneous linear systems admit a superposition principle, that allows one to construct new solutions from known solutions. As we learned, the word "superposition" refers to taking linear combinations of solutions.

Consider a general homogeneous linear system

$$L[\mathbf{z}] = \mathbf{0} \tag{7.36}$$

where L is a linear function. If we are given two solutions, say $\mathbf{z}_1$ and $\mathbf{z}_2$, meaning that

$$L[\mathbf{z}_1] = \mathbf{0}, \qquad L[\mathbf{z}_2] = \mathbf{0},$$

then their sum $\mathbf{z}_1 + \mathbf{z}_2$ is automatically a solution, since, in view of the linearity of L,

$$L[\mathbf{z}_1 + \mathbf{z}_2] = L[\mathbf{z}_1] + L[\mathbf{z}_2] = \mathbf{0} + \mathbf{0} = \mathbf{0}.$$

[†] This is a particular case of the general Cartesian product construction between vector spaces, with $V = \mathrm{C}^0 \times \mathbb{R}^2$. See Exercise ■ for details.

Similarly, given a solution $\mathbf{z}$ and any scalar c, the scalar multiple $c\,\mathbf{z}$ is automatically a solution, since

$$L[c\,\mathbf{z}] = c\,L[\mathbf{z}] = c\,\mathbf{0} = \mathbf{0}.$$

Combining these two elementary observations, we can now state the general *superposition principle*. The proof is an immediate consequence of formula (7.4).

Theorem 7.29. *If $\mathbf{z}_1, \dots, \mathbf{z}_k$ are all solutions to the same homogeneous linear system $L[\mathbf{z}] = \mathbf{0}$, and $c_1, \dots, c_k$ are any scalars, then the linear combination $c_1\,\mathbf{z}_1 + \cdots + c_k\,\mathbf{z}_k$ is also a solution.*

As with matrices, we call the solution space to the homogeneous linear system (7.36) the *kernel* of the linear function L. Theorem 7.29 implies that the kernel always forms a subspace.

Proposition 7.30. *If $L\colon V \to W$ is a linear function, then its* kernel

$$\ker L = \{\, \mathbf{z} \in V \mid \; L[\mathbf{z}] = 0 \,\} \subset V \tag{7.37}$$

forms a subspace of the domain space V.

As we know, in the case of linear matrix systems, the kernel can be explicitly determined by the basic Gaussian elimination algorithm. For more general linear operators, one must develop appropriate solution techniques for solving the homogeneous linear system. Here is a simple example from the theory of linear, constant coefficient ordinary differential equations.

Example 7.31. Consider the second order linear differential operator

$$L = D^2 - 2\,D - 3, \tag{7.38}$$

which maps the function $u(x)$ to the function

$$L[u] = (D^2 - 2\,D - 3)[u] = u'' - 2\,u' - 3\,u.$$

The associated homogeneous system takes the form of a homogeneous, linear, second order ordinary differential equation

$$L[u] = u'' - 2\,u' - 3\,u = 0. \tag{7.39}$$

In accordance with the standard solution method, we plug the exponential *ansatz*[†]

$$u = e^{\lambda\,x}$$

[†] The German word *ansatz* (plural *ansätze*) refers to the method of finding a solution to a complicated equation by guessing the solution's form in advance. Typically, one is not clever enough to guess the precise solution, and so the ansatz will have one or more free parameters — in this case the constant exponent λ — that, with some luck, can be rigged up to fulfill the requirements imposed by the equation. Thus, a reasonable English translation of "ansatz" is "inspired guess".

into the equation. The result is

$$L[e^{\lambda x}] = D^2[e^{\lambda x}] - 2\,D[e^{\lambda x}] - 3\,e^{\lambda x} = (\lambda^2 - 2\,\lambda - 3)e^{\lambda x},$$

and therefore, $e^{\lambda x}$ is a solution if and only if λ satisfies the *characteristic equation*

$$0 = \lambda^2 - 2\,\lambda - 3 = (\lambda - 3)(\lambda + 1).$$

The two roots are $\lambda_1 = 3$, $\lambda_2 = -1$, and hence

$$u_1(x) = e^{3\,x}, \qquad\qquad u_2(x) = e^{-\,x}, \tag{7.40}$$

are two linearly independent solutions of (7.39). According to the general superposition principle, every linear combination

$$u(x) = c_1\,u_1(x) + c_2\,u_2(x) = c_1\,e^{3\,x} + c_2\,e^{-\,x}$$

of these two basic solutions is also a solution, for any choice of constants c_1, c_2. In fact, this two-parameter family constitutes the most general solution to the ordinary differential equation (7.39). Thus, the kernel of the second order differential operator (7.38) is two-dimensional, with basis given by the independent exponential solutions (7.40).

In general, the solution space to an n^{th} order homogeneous linear ordinary differential equation

$$L[u] = a_n(x)\,\frac{d^n u}{dx^n} + a_{n-1}(x)\,\frac{d^{n-1}u}{dx^{n-1}} + \,\cdots\, + a_1(x)\,\frac{du}{dx} + a_0(x)u = 0 \tag{7.41}$$

forms a subspace of the vector space $\mathrm{C}^n[a,b]$ of n times continuously differentiable functions, since it is just the kernel of a linear differential operator $L\colon \mathrm{C}^n[a,b] \to \mathrm{C}^0[a,b]$. This implies that linear combinations of solutions are also solutions. To determine the number of solutions, or, more precisely, the dimension of the solution space, we need to impose some mild restrictions on the differential operator.

Definition 7.32. The differential operator L is called *nonsingular* on an open interval $[a,b]$ if all its coefficients $a_n(x), \dots, a_0(x) \in \mathrm{C}^0[a,b]$ are continuous functions and its *leading coefficient* does not vanish: $a_n(x) \neq 0$ for all $a < x < b$.

The basic existence and uniqueness result governing nonsingular homogeneous linear ordinary differential equations can be formulated as a characterization of the dimension of the solution space.

Theorem 7.33. *The kernel of a nonsingular n^{th} order ordinary differential operator forms an n-dimensional subspace* $\ker L \subset \mathrm{C}^n[a,b]$.

A proof of this result can be found in Section 20.1. The fact that the kernel has dimension n means that it has a basis consisting of n linearly independent solutions $u_1(x), \dots, u_n(x) \in \mathrm{C}^n[a,b]$ such that the general solution to the homogeneous differential equation (7.41) is given by a linear combination

$$u(x) = c_1\,u_1(x) + \,\cdots\, + c_n\,u_n(x),$$

where $c_1, \dots, c_n$ are arbitrary constants. Therefore, once we find n linearly independent solutions of an n^{th} order homogeneous linear ordinary differential equation, we can immediately write down its most general solution.

The condition that the leading coefficient $a_n(x)$ does not vanish is essential. Points where $a_n(x) = 0$ are known as *singular points*. They arise in many applications, but must be treated separately and with care; see Appendix C. (Of course, if the coefficients are constant then there is nothing to worry about — either the leading coefficient is nonzero, $a_n \neq 0$, or the operator is, in fact, of lower order than advertised.)

Example 7.34. A second order *Euler differential equation* takes the form

$$E[u] = a\,x^2\,u'' + b\,x\,u' + c\,u = 0, \tag{7.42}$$

where $0 \neq a, b, c$ are constants, and $E = a\,x^2\,D^2 + b\,x\,D + c$ is a second order, non-constant coefficient differential operator. Instead of the exponential solution ansatz used in the constant coefficient case, Euler equations are solved by using a power ansatz

$$u(x) = x^r$$

with unknown exponent r. Substituting into the differential equation, we find

$$E[\,x^r\,] = a\,r\,(r-1)\,x^r + b\,r\,x^r + c\,x^r = [\,a\,r\,(r-1) + b\,r + c\,]\,x^r = 0,$$

and hence x^r is a solution if and only if r satisfies the *characteristic equation*

$$a\,r\,(r-1) + b\,r + c = a\,r^2 + (b-a)\,r + c = 0. \tag{7.43}$$

If the characteristic equation has two distinct real roots, $r_1 \neq r_2$, then there are two linearly independent solutions $u_1(x) = x^{r_1}$ and $u_2(x) = x^{r_2}$, and the general (real) solution to (7.42) has the form

$$u(x) = c_1\,|\,x\,|^{r_1} + c_2\,|\,x\,|^{r_2}. \tag{7.44}$$

(The absolute values are usually needed to ensure that the solutions remain real when $x < 0$ is negative.) The other cases — repeated roots and complex roots — will be discussed below.

The Euler equation has a singular point at $x = 0$, where its leading coefficient vanishes. Theorem 7.33 assures us that the differential equation has a two-dimensional solution space on any interval not containing the singular point. However, predicting the number of solutions which remain continuously differentiable at $x = 0$ is not as immediate, since it depends on the values of the exponents r_1 and r_2. For instance, the case

$$x^2\,u'' - 3\,x\,u' + 3\,u = 0 \qquad \text{has solution} \qquad u = c_1\,x + c_2\,x^3,$$

which forms a two-dimensional subspace of $\mathrm{C}^0(\mathbb{R})$. However,

$$x^2\,u'' + x\,u' - u = 0 \qquad \text{has solution} \qquad u = c_1\,x + \frac{c_2}{x}\,,$$

and only the multiples of the first solution x are continuous at $x = 0$. Therefore, the solutions that are continuous everywhere form only a one-dimensional subspace of $\mathrm{C}^0(\mathbb{R})$. Finally,

$$x^2\,u'' + 5\,x\,u' + 3\,u = 0 \qquad \text{has solution} \qquad u = \frac{c_1}{x} + \frac{c_2}{x^3}\,,$$

and there are no nontrivial solutions $u \neq 0$ that are continuous at $x = 0$.

Example 7.35. Considerr the *Laplace equation*

$$\Delta[u] = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{7.45}$$

for a function $u(x, y)$ defined on a domain $\Omega \subset \mathbb{R}^2$. The Laplace equation is *the* most important partial differential equation, and its applications range over almost all fields of mathemtics, physics and engineering, including complex analysis, geometry, fluid mechanics, electromagnetism, elasticity, thermodynamics, and quantum mechanics. It is a homogeneous linear partial differential equation corresponding to the partial differential operator $\Delta = \partial_x^2 + \partial_y^2$ known as the *Laplacian* operator. Linearity can either be proved directly, or by noting that Δ is built up from the basic linear partial derivative operators ∂_x, ∂_y by the processes of composition and addition, as in Exercise ■.

Unlike the case of a linear ordinary differential equation, there are an infinite number of linearly independent solutions to the Laplace equation. Examples include the trigonometric/exponential solutions

$$e^{\omega\,x}\cos\omega\,y, \qquad e^{\omega\,x}\sin\omega\,y, \qquad e^{\omega\,y}\cos\omega\,x, \qquad e^{\omega\,y}\sin\omega\,y,$$

where ω is *any* real constant. There are also infinitely many independent polynomial solutions, the first few of which are

$$1, \qquad x, \qquad y, \qquad x^2 - y^2, \qquad x\,y, \qquad x^3 - 3\,x\,y^2, \qquad \dots$$

The reader might enjoy finding some more polynomial solutions and trying to spot the pattern. (The answer will appear shortly.) As usual, we can build up more complicated solutions by taking general linear combinations of these particular ones. In fact, it will be shown that the most general solution to the Laplace equation can be written as a convergent infinite series in the basic polynomial solutions. Later, in Chapters 15 and 16, we will learn how to construct these and many other solutions to the planar Laplace equation.

Inhomogeneous Systems

Now we turn our attention to an inhomogeneous linear system

$$L[\mathbf{u}] = \mathbf{f}. \tag{7.46}$$

Unless $\mathbf{f} = \mathbf{0}$, the solution space to (7.46) is *not* a subspace. (Why?) The key question is existence — is there a solution to the system? In the homogeneous case, existence is not an issue, since $\mathbf{0}$ is always a solution to $L[\mathbf{z}] = \mathbf{0}$. The key question for homogeneous

systems is uniqueness — whether $\ker L = \{\mathbf{0}\}$, in which case $\mathbf{0}$ is the only solution, or whether there are nontrivial solutions $\mathbf{0} \neq \mathbf{z} \in \ker L$.

In the matrix case, the compatibility of an inhomogeneous system $A\,\mathbf{x} = \mathbf{b}$ — which was required for the existence of a solution — led to the general definition of the range of a matrix, which we copy verbatim for linear functions.

Definition 7.36. The *range* of a linear function $L\colon V \to W$ is the subspace

$$\operatorname{rng} L = \{\, L[\mathbf{v}] \;\mid\; \mathbf{v} \in V \,\} \;\subset\; W.$$

The proof that $\operatorname{rng} L$ is a subspace is straightforward. If $\mathbf{f} = L[\mathbf{v}]$ and $\mathbf{g} = L[\mathbf{w}]$ are any two elements of the range, so is any linear combination, since, by linearity

$$c\,\mathbf{f} + d\,\mathbf{g} = c\,L[\mathbf{v}] + d\,L[\mathbf{w}] = L[c\,\mathbf{v} + d\,\mathbf{w}] \in \operatorname{rng} L.$$

For example, if $L[\mathbf{v}] = A\,\mathbf{v}$ is given by multiplication by an $m \times n$ matrix, then its range is the subspace $\operatorname{rng} L = \operatorname{rng} A \subset \mathbb{R}^m$ spanned by the columns of A — the *column space* of the coefficient matrix. When L is a linear differential operator, or more general linear operator, characterizing its range can be a much more challenging problem.

The fundamental theorem regarding solutions to inhomogeneous linear equations exactly mimics our earlier result, Theorem 2.37, in the particular case of matrix systems.

Theorem 7.37. *Let $L\colon V \to W$ be a linear function. Let $\mathbf{f} \in W$. Then the inhomogeneous linear system*

$$L[\mathbf{u}] = \mathbf{f} \tag{7.47}$$

has a solution if and only if $\mathbf{f} \in \operatorname{rng} L$. In this case, the general solution to the system has the form

$$\mathbf{u} = \mathbf{u}^\star + \mathbf{z} \tag{7.48}$$

where $\mathbf{u}^\star$ is a particular solution, so $L[\mathbf{u}^\star] = \mathbf{f}$, and $\mathbf{z}$ is a general element of $\ker L$, i.e., the general solution to the corresponding homogeneous system

$$L[\mathbf{z}] = \mathbf{0}. \tag{7.49}$$

Proof: We merely repeat the proof of Theorem 2.37. The existence condition $\mathbf{f} \in \operatorname{rng} L$ is an immediate consequence of the definition of the range. Suppose $\mathbf{u}^\star$ is a particular solution to (7.47). If $\mathbf{z}$ is a solution to (7.49), then, by linearity,

$$L[\mathbf{u}^\star + \mathbf{z}] = L[\mathbf{u}^\star] + L[\mathbf{z}] = \mathbf{f} + \mathbf{0} = \mathbf{f},$$

and hence $\mathbf{u}^\star + \mathbf{z}$ is also a solution to (7.47). To show that every solution has this form, let $\mathbf{u}$ be a second solution, so that $L[\mathbf{u}] = \mathbf{f}$. Then

$$L[\mathbf{u} - \mathbf{u}^\star] = L[\mathbf{u}] - L[\mathbf{u}^\star] = \mathbf{f} - \mathbf{f} = \mathbf{0}.$$

Therefore $\mathbf{u} - \mathbf{u}^\star = \mathbf{z} \in \ker L$ is a solution to (7.49). *Q.E.D.*

Remark: In physical systems, the inhomogeneity $\mathbf{f}$ typically corresponds to an external forcing function. The solution $\mathbf{z}$ to the homogeneous system represents the system's natural, unforced motion. Therefore, the decomposition formula (7.48) states that a linear system responds to an external force as a combination of its own internal motion and a specific motion $\mathbf{u}^\star$ induced by the forcing. Examples of this important principle appear throughout the book.

Corollary 7.38. *The inhomogeneous linear system* (7.47) *has a* unique *solution if and only if* $\mathbf{f} \in \operatorname{rng} L$ *and* $\ker L = \{\mathbf{0}\}$.

Therefore, to prove that a linear system has a unique solution, we first need to prove an *existence result* that there is at least one solution, which requires the right hand side $\mathbf{f}$ to lie in the range of the operator L, and then a *uniqueness result*, that the only solution to the homogeneous system $L[\mathbf{z}] = \mathbf{0}$ is the trivial zero solution $\mathbf{z} = \mathbf{0}$. Consequently, if an inhomogeneous system $L[\mathbf{u}] = \mathbf{f}$ has a unique solution, then any other inhomogeneous system $L[\mathbf{u}] = \mathbf{g}$ that is defined by the *same* linear function also has a unique solution for every $\mathbf{g} \in \operatorname{rng} L$.

Example 7.39. Consider the inhomogeneous linear second order differential equation

$$u'' + u' - 2\,u = x.$$

Note that this can be written in the linear system form

$$L[u] = x, \qquad \text{where} \qquad L = D^2 + D - 2$$

is a linear second order differential operator. The kernel of the differential operator L is found by solving the associated homogeneous linear equation

$$L[z] = z'' + z' - 2\,z = 0.$$

Applying the usual solution method, we find that the homogeneous differential equation has a two-dimensional solution space, with basis functions

$$z_1(x) = e^{-2\,x}, \qquad z_2(x) = e^{x}.$$

Therefore, the general element of $\ker L$ is a linear combination

$$z(x) = c_1\,z_1(x) + c_2\,z_2(x) = c_1\,e^{-2\,x} + c_2\,e^{x}.$$

To find a particular solution to the inhomogeneous differential equation, we rely on the method of undetermined coefficients†. We introduce the solution ansatz $u = a\,x + b$, and compute

$$L[u] = L[a\,x + b] = -2\,a\,x - 2\,b + a = x.$$

† One could also employ the method of variation of parameters, although in general the undetermined coefficient method, when applicable, is the more straightforward of the two. Details of the two methods can be found, for instance, in [**24**].

Equating the two expressions, we conclude that $a = -\frac{1}{2}$, $b = -\frac{1}{4}$, and hence

$$u^\star(x) = -\tfrac{1}{2}\,x - \tfrac{1}{4}$$

is a particular solution to the inhomogeneous differential equation. Theorem 7.37 then says that the general solution is

$$u(x) = u^\star(x) + z(x) = -\tfrac{1}{2}\,x - \tfrac{1}{4} + c_1\,e^{-2\,x} + c_2\,e^{x}.$$

Example 7.40. By inspection, we see that

$$u(x,y) = -\tfrac{1}{2}\,\sin(x+y)$$

is a solution to the particular *Poisson equation*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \sin(x+y). \tag{7.50}$$

Theorem 7.37 implies that *every* solution to this inhomogeneous version of the Laplace equation takes the form

$$u(x,y) = -\tfrac{1}{2}\,\sin(x+y) + z(x,y),$$

where $z(x,y)$ is an arbitrary solution to the homogeneous Laplace equation (7.45).

Example 7.41. The problem is to solve the linear boundary value problem

$$u'' + u = x, \qquad u(0) = 0, \qquad u(\pi) = 0. \tag{7.51}$$

The first step is to solve the differential equation. To this end, we find that $\cos x$ and $\sin x$ form a basis for the solution space to the corresponding homogeneous differential equation $z'' + z = 0$. The method of undetermined coefficients then produces the particular solution $u^\star(x) = x$ to the inhomogeneous differential equation, and so the general solution is

$$u(x) = x + c_1\cos x + c_2\sin x. \tag{7.52}$$

The next step is to see whether any solutions also satisfy the boundary conditions. Plugging formula (7.52) into the boundary conditions gives

$$u(0) = c_1 = 0, \qquad u(\pi) = \pi - c_1 = 0.$$

However, these two conditions are incompatible, and so there is *no* solution to the linear system (7.51). The function $f(x) = x$ does not lie in the range of the differential operator $L[u] = u'' + u$ when u is subjected to the boundary conditions.

On the other hand, if we change the inhomogeneity, the boundary value problem

$$u'' + u = x - \tfrac{1}{2}\,\pi, \qquad u(0) = 0, \qquad u(\pi) = 0. \tag{7.53}$$

does admit a solution, but the solution fails to be unique. Applying the preceding solution method, we find that the function

$$u(x) = x - \tfrac{1}{2}\,\pi + \tfrac{1}{2}\,\pi\cos x + c\,\sin x$$

solves the system for any choice of constant c. Note that $z(x) = \sin x$ forms a basis for the kernel or solution space of the homogeneous boundary value problem

$$z'' + z = 0, \qquad z(0) = 0, \qquad z(\pi) = 0.$$

Incidentally, if we slightly modify the interval of definition, considering

$$u'' + u = f(x), \qquad u(0) = 0, \qquad u\left(\tfrac{1}{2}\,\pi\right) = 0, \tag{7.54}$$

then the system is compatible for *any* inhomogeneity $f(x)$, and the solution to the boundary value problem is unique. For example, if $f(x) = x$, then the unique solution is

$$u(x) = x - \tfrac{1}{2}\,\pi \sin x\,. \tag{7.55}$$

This example highlights some major differences between boundary value problems and initial value problems for ordinary differential equations. For nonsingular initial value problems, there is a unique solution for any set of initial conditions. For boundary value problems, the structure of the solution space — either a unique solution for all inhomogeneities, or no solution, or infinitely many solutions, depending on the right hand side — has more of the flavor of a linear matrix system. An interesting question is how to characterize the inhomogeneities $f(x)$ that admit a solution, i.e., lie in the range of the operator. We will return to this question in Chapter 11.

Superposition Principles for Inhomogeneous Systems

The *superposition principle* for inhomogeneous linear systems allows us to combine different inhomogeneities — provided we do not change the underlying linear operator. The result is a straightforward generalization of the matrix version described in Theorem 2.42.

Theorem 7.42. *Let $L\colon V \to W$ be a prescribed linear function. Suppose that, for each $i = 1, \dots, k$, we know a particular solution $\mathbf{u}_i^\star$ to the inhomogeneous linear system $L[\mathbf{u}] = \mathbf{f}_i$ for some $\mathbf{f}_i \in \operatorname{rng} L$. Given scalars $c_1, \dots, c_k$, a particular solution to the combined inhomogeneous system*

$$L[\mathbf{u}] = c_1\,\mathbf{f}_1 + \,\cdots\, + c_k\,\mathbf{f}_k \tag{7.56}$$

is the same linear combination $\mathbf{u}^\star = c_1\,\mathbf{u}_1^\star + \cdots + c_k\,\mathbf{u}_k^\star$ of particular solutions. The general solution to the inhomogeneous system (7.56) *is*

$$\mathbf{u} = \mathbf{u}^\star + \mathbf{z} = c_1\,\mathbf{u}_1^\star + \,\cdots\, + c_k\,\mathbf{u}_k^\star + \mathbf{z},$$

where $\mathbf{z} \in \ker L$ is the general solution to the associated homogeneous system $L[\mathbf{z}] = \mathbf{0}$.

The proof is an easy consequence of linearity, and left to the reader. In physical terms, the superposition principle can be interpreted as follows. If we know the response of a linear physical system to several different external forces, represented by $\mathbf{f}_1, \dots, \mathbf{f}_k$, then the response of the system to a linear combination of these forces is just the identical linear combination of the individual responses. The homogeneous solution $\mathbf{z}$ represents an internal motion that the system acquires independent of any external forcing. Superposition requires linearity of the system, and so is always applicable in quantum mechanics,

which is a linear theory. But, in classical and relativistic mechanics superposition only applies in a linear approximation corresponding to small motions/displacements/etc. The nonlinear regime is much more unpredictable, and combinations of external forces may lead to unexpected results.

Example 7.43. We already know that a particular solution to the linear differential equation

$$u'' + u = x \qquad \text{is} \qquad u_1^\star = x.$$

The method of undetermined coefficients is used to solve the inhomogeneous equation

$$u'' + u = \cos x.$$

Since $\cos x$ and $\sin x$ are already solutions to the homogeneous equation, we must use the solution ansatz $u = a\,x\cos x + b\,x\sin x$, which, when substituted into the differential equation, produces the particular solution

$$u_2^\star = -\tfrac{1}{2}\,x\sin x.$$

Therefore, by the superposition principle, the combination inhomogeneous system

$$u'' + u = 3\,x - 2\cos x \quad \text{has a particular solution} \quad u^\star = 3\,u_1^\star - 2\,u_2^\star = 3\,x + x\sin x.$$

The general solution is obtained by appending the general solution to the homogeneous equation: $u = 3\,x + x\sin x + c_1\cos x + c_2\sin x$.

Example 7.44. Consider the boundary value problem

$$u'' + u = x, \qquad u(0) = 2, \qquad u\big(\tfrac{1}{2}\,\pi\big) = -1, \tag{7.57}$$

which is a modification of (7.54) with inhomogeneous boundary conditions. The superposition principle applies here, and allows us to decouple the inhomogeneity due to the forcing from the inhomogeneity due to the boundary conditions. We already solved the boundary value problem with homogeneous boundary conditions; see (7.55). On the other hand, the unforced boundary value problem

$$u'' + u = 0, \qquad u(0) = 2, \qquad u\big(\tfrac{1}{2}\,\pi\big) = -1, \tag{7.58}$$

has unique solution

$$u(x) = 2\cos x - \sin x. \tag{7.59}$$

Therefore, the solution to the combined problem (7.57) is the sum of these two:

$$u(x) = x + 2\cos x - \big(1 + \tfrac{1}{2}\,\pi\big)\,\sin x\,.$$

The solution is unique because the corresponding homogeneous boundary value problem

$$z'' + z = 0, \qquad z(0) = 0, \qquad z\big(\tfrac{1}{2}\,\pi\big) = 0,$$

has only the trivial solution $z(x) \equiv 0$. Incidentally, the solution (7.59) can itself be decomposed as a linear combination of the solutions $\cos x$ and $\sin x$ to a pair of yet more elementary boundary value problems with just one inhomogeneous boundary condition; namely, $u(0) = 1$, $u\big(\tfrac{1}{2}\,\pi\big) = 0$, and, respectively, $u(0) = 0$, $u\big(\tfrac{1}{2}\,\pi\big) = 1$.

Complex Solutions to Real Systems

The easiest way to obtain solutions to a linear, homogeneous, constant coefficient ordinary differential equation is through an exponential ansatz, which effectively reduces it to the algebraic characteristic equation. Complex roots of the characteristic equation yield complex exponential solutions. But, if the equation is real, then the real and imaginary parts of the complex solutions are automatically real solutions. This solution technique is a particular case of a general principle for producing real solutions to real linear systems from, typically, simpler complex solutions. To work, the method requires some additional structure on the vector spaces involved.

Definition 7.45. A complex vector space V is called *conjugated* if it admits an operation of complex conjugation taking $\mathbf{u} \in V$ to $\overline{\mathbf{u}}$ that is compatible with scalar multiplication. In other words, if $\mathbf{u} \in V$ and $\lambda \in \mathbb{C}$, then we require $\overline{\lambda\,\mathbf{u}} = \overline{\lambda}\,\overline{\mathbf{u}}$.

The simplest example of a conjugated vector space is $\mathbb{C}^n$. The complex conjugate of a vector is obtained by conjugating all its entries. Thus we have

$$\begin{aligned} \mathbf{u} &= \mathbf{v} + \mathrm{i}\,\mathbf{w}, \\ \overline{\mathbf{u}} &= \mathbf{v} - \mathrm{i}\,\mathbf{w}, \end{aligned} \qquad \text{where} \qquad \mathbf{v} = \operatorname{Re}\mathbf{u} = \frac{\mathbf{u} + \overline{\mathbf{u}}}{2}, \qquad \mathbf{w} = \operatorname{Im}\mathbf{u} = \frac{\mathbf{u} - \overline{\mathbf{u}}}{2\,\mathrm{i}}, \tag{7.60}$$

are the real and imaginary parts of $\mathbf{u} \in \mathbb{C}^n$. For example, if

$$\mathbf{u} = \begin{pmatrix} 1 - 2\,\mathrm{i} \\ 3\,\mathrm{i} \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix} + \mathrm{i} \begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix}, \qquad \text{then} \qquad \overline{\mathbf{u}} = \begin{pmatrix} 1 + 2\,\mathrm{i} \\ -3\,\mathrm{i} \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 5 \end{pmatrix} - \mathrm{i} \begin{pmatrix} -2 \\ 3 \\ 0 \end{pmatrix}.$$

The same definition of real and imaginary part carries over to general conjugated vector spaces. A subspace $V \subset \mathbb{C}^n$ is conjugated if and only if $\overline{\mathbf{u}} \in V$ whenever $\mathbf{u} \in V$. Another prototypical example of a conjugated vector space is the space of complex-valued functions $f(x) = r(x) + \mathrm{i}\,s(x)$ defined on the interval $a \le x \le b$. The complex conjugate function is $\overline{f}(x) = r(x) - \mathrm{i}\,s(x)$. Thus, the complex conjugate of

$$e^{(1+3\,\mathrm{i})x} = e^x \cos 3x + \mathrm{i}\,e^x \sin 3x \quad \text{is} \quad \overline{e^{(1+3\,\mathrm{i})x}} = e^{(1-3\,\mathrm{i})x} = e^x \cos 3x - \mathrm{i}\,e^x \sin 3x.$$

An element $\mathbf{v} \in V$ of a conjugated vector space is called *real* if $\overline{\mathbf{v}} = \mathbf{v}$. One easily checks that the real and imaginary parts of a general element, as defined by (7.60), are both real elements.

Definition 7.46. A linear operator $L\colon V \to W$ between conjugated vector spaces is called *real* if it commutes with complex conjugation:

$$L[\overline{\mathbf{u}}] = \overline{L[\mathbf{u}]}. \tag{7.61}$$

For example, the linear function $F\colon \mathbb{C}^n \to \mathbb{C}^m$ given by matrix multiplication, $F(\mathbf{u}) = A\,\mathbf{u}$, is real if and only if A is a real matrix. Similarly, a differential operator (7.13) is real if its coefficients are real-valued functions.

Theorem 7.47. *If $L[\mathbf{u}] = \mathbf{0}$ is a real homogeneous linear system and $\mathbf{u} = \mathbf{v} + \mathrm{i}\,\mathbf{w}$ is a complex solution, then its complex conjugate $\overline{\mathbf{u}} = \mathbf{v} - \mathrm{i}\,\mathbf{w}$ is also a solution. Moreover, both the real and imaginary parts, $\mathbf{v}$ and $\mathbf{w}$, of a complex solution are real solutions.*

Proof: First note that, by reality,

$$L[\overline{\mathbf{u}}] = \overline{L[\mathbf{u}]} = \mathbf{0} \qquad \text{whenever} \qquad L[\mathbf{u}] = \mathbf{0},$$

and hence the complex conjugate $\overline{\mathbf{u}}$ of any solution is also a solution. Therefore, by linear superposition, $\mathbf{v} = \operatorname{Re} \mathbf{u} = \frac{1}{2}\,(\mathbf{u} + \overline{\mathbf{u}})$ and $\mathbf{w} = \operatorname{Im} \mathbf{u} = \frac{1}{2\,\mathrm{i}}\,(\mathbf{u} - \overline{\mathbf{u}})$ are also solutions. *Q.E.D.*

Example 7.48. The real linear matrix system

$$\begin{pmatrix} 2 & -1 & 3 & 0 \\ -2 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

has a complex solution

$$\mathbf{u} = \begin{pmatrix} -1 - 3\,\mathrm{i} \\ 1 \\ 1 + 2\,\mathrm{i} \\ -2 - 4\,\mathrm{i} \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -2 \end{pmatrix} + \mathrm{i} \begin{pmatrix} -3 \\ 0 \\ 2 \\ -4 \end{pmatrix}.$$

Since the coefficient matrix is real, the real and imaginary parts,

$$\mathbf{v} = (\,-1, 1, 1, -2\,)^T, \qquad \mathbf{w} = (\,-3, 0, 2, -4\,)^T,$$

are both solutions of the system.

On the other hand, the complex linear system

$$\begin{pmatrix} 2 & -2\,\mathrm{i} & \mathrm{i} & 0 \\ 1 + \mathrm{i} & 0 & -2 - \mathrm{i} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

has the complex solution

$$\mathbf{u} = \begin{pmatrix} 1 - \mathrm{i} \\ -\,\mathrm{i} \\ 2 \\ 2 + 2\,\mathrm{i} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \\ 2 \end{pmatrix} + \mathrm{i} \begin{pmatrix} -1 \\ -1 \\ 0 \\ 2 \end{pmatrix}.$$

However, neither the real nor the imaginary part is a solution to the system.

Example 7.49. Consider the real ordinary differential equation

$$u'' + 2\,u' + 5\,u = 0.$$

To solve it, as in Example 7.31, we use the exponential ansatz $u = e^{\lambda\, x}$, leading to the characteristic equation

$$\lambda^2 + 2\,\lambda + 5 = 0.$$

There are two roots,

$$\lambda_1 = -1 + 2\,\mathrm{i}\,, \qquad \lambda_2 = -1 - 2\,\mathrm{i}\,,$$

leading, via Euler's formula (3.76), to the complex solutions

$$u_1(x) = e^{(-1+2\,\mathrm{i}\,)x} = e^{-x}\cos 2x + \mathrm{i}\,e^{-x}\sin 2x,$$
$$u_2(x) = e^{(-1-2\,\mathrm{i}\,)x} = e^{-x}\cos 2x - \mathrm{i}\,e^{-x}\sin 2x.$$

The complex conjugate of the first solution is the second, in accordance with Theorem 7.47. Moreover, the real and imaginary parts of the two solutions

$$v(x) = e^{-x}\cos 2x, \qquad w(x) = e^{-x}\sin 2x,$$

are individual real solutions. The general solution is a linear combination

$$u(x) = c_1\, e^{-x}\cos 2x + c_2\, e^{-x}\sin 2x,$$

of the two linearly independent real solutions.

Example 7.50. Consider the second order Euler differential equation

$$L[u] = x^2\,u'' + 7x\,u' + 13\,u = 0.$$

The roots of the associated characteristic equation

$$r\,(r-1) + 7r + 13 = r^2 + 6r + 13 = 0$$

are complex: $r = -3 \pm 2\,\mathrm{i}\,$, and the resulting solutions $x^r = x^{-3\pm 2\,\mathrm{i}}$ are complex conjugate powers. Using Euler's formula (3.76), we write them in real and imaginary form, e.g.,

$$x^{-3+2\,\mathrm{i}} = x^{-3}\,e^{2\,\mathrm{i}\,\log x} = x^{-3}\cos(2\,\log x) + \mathrm{i}\,x^{-3}\sin(2\,\log x).$$

Again, by Theorem 7.47, the real and imaginary parts of the complex solution are by themselves real solutions to the equation. Therefore, the general real solution is

$$u(x) = c_1\,x^{-3}\cos(2\,\log x) + c_2\,x^{-3}\sin(2\,\log x).$$

Example 7.51. The complex monomial

$$u(x,y) = (x + \mathrm{i}\,y)^n$$

is a solution to the Laplace equation (7.45) because, by the chain rule,

$$\frac{\partial^2 u}{\partial x^2} = n(n-1)(x+\mathrm{i}\,y)^{n-2}, \qquad \frac{\partial^2 u}{\partial y^2} = n(n-1)\,\mathrm{i}^{\,2}\,(x+\mathrm{i}\,y)^{n-2} = -\,n(n-1)(x+\mathrm{i}\,y)^{n-2},$$

and hence $u_{xx} + u_{yy} = 0$. Since the Laplace operator is real, Theorem 7.47 implies that the real and imaginary parts of this complex solution are real solutions. The resulting real solutions are known as *harmonic polynomials*.

To find the explicit formulae for the harmonic polynomials, we use the Binomial Formula and the fact that $\mathrm{i}^2 = -1$, $\mathrm{i}^3 = -\mathrm{i}$, $\mathrm{i}^4 = 1$, etc., to expand

$$\begin{aligned}(x + \mathrm{i}\, y)^n &= x^n + n x^{n-1}(\mathrm{i}\, y) + \binom{n}{2} x^{n-2}(\mathrm{i}\, y)^2 + \binom{n}{3} x^{n-3}(\mathrm{i}\, y)^3 + \cdots \\ &= x^n + \mathrm{i}\, n x^{n-1} y - \binom{n}{2} x^{n-2} y^2 - \mathrm{i} \binom{n}{3} x^{n-3} y^3 + \cdots,\end{aligned}$$

in which we use the standard notation

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \tag{7.62}$$

for the *binomial coefficients*. Separating the real and imaginary terms, we find

$$\begin{aligned}\operatorname{Re}(x + \mathrm{i}\, y)^n &= x^n - \binom{n}{2} x^{n-2} y^2 + \binom{n}{4} x^{n-4} y^4 + \cdots, \\ \operatorname{Im}(x + \mathrm{i}\, y)^n &= n x^{n-1} y - \binom{n}{3} x^{n-3} y^3 + \binom{n}{5} x^{n-5} y^5 + \cdots.\end{aligned} \tag{7.63}$$

The first few of these harmonic polynomials were described in Example 7.35. In fact, it can be proved that every polynomial solution to the Laplace equation is a linear combination of the fundamental real harmonic polynomials; see Chapter 16 for full details.

7.5. Adjoints.

In Sections 2.5 and 5.6, we discovered the importance of the adjoint system $A^T \mathbf{y} = \mathbf{f}$ in the analysis of systems of linear equations $A\,\mathbf{x} = \mathbf{b}$. Two of the four fundamental matrix subspaces are based on the transposed matrix. While the $m \times n$ matrix A defines a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$, its transpose, A^T, has size $n \times m$ and hence characterizes a linear function in the *reverse* direction, from $\mathbb{R}^m$ to $\mathbb{R}^n$.

As with most fundamental concepts for linear matrix systems, the adjoint system and transpose operation on the coefficient matrix are the prototypes of a much more general construction that is valid for general linear functions. However, it is not as obvious how to "transpose" a more general linear operator $L[u]$, e.g., a differential operator acting on function space. In this section, we shall introduce the concept of the *adjoint* of a linear function that generalizes the transpose operation on matrices. Unfortunately, most of the interesting examples must be deferred until we develop additional analytical tools, starting in Chapter 11.

The adjoint (and transpose) relies on an inner product structure on both the domain and target spaces. For simplicity, we restrict our attention to real inner product spaces, leaving the complex version to the interested reader. Thus, we begin with a linear function $L\colon V \to W$ that maps an inner product space V to a second inner product space W. We distinguish the inner products on V and W (which may be different even when V and W are the same vector space) by using a single angle bracket

$$\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle \qquad \text{to denote the inner product between} \qquad \mathbf{v}, \widetilde{\mathbf{v}} \in V,$$

and a double angle bracket

$$\langle\!\langle\, \mathbf{w}\,;\widetilde{\mathbf{w}}\,\rangle\!\rangle \qquad \text{to denote the inner product between} \qquad \mathbf{w},\widetilde{\mathbf{w}}\in W.$$

With the prescription of inner products on both the domain and target spaces, the abstract definition of the adjoint of a linear function can be formulated.

Definition 7.52. Let V,W be inner product spaces, and let $L\colon V\to W$ be a linear function. The *adjoint* of L is the function $L^*\colon W\to V$ that satisfies

$$\langle\!\langle\, L[\mathbf{v}]\,;\mathbf{w}\,\rangle\!\rangle = \langle\, \mathbf{v}\,;L^*[\mathbf{w}]\,\rangle \qquad \text{for all} \qquad \mathbf{v}\in V,\quad \mathbf{w}\in W. \tag{7.64}$$

Note that the adjoint function goes in the *opposite* direction to L, just like the transposed matrix. Also, the left hand side of equation (7.64) indicates the inner product on W, while the right hand side is the inner product on V — which is where the respective vectors live. In infinite-dimensional situations, the adjoint may not exist. But if it does, then it is uniquely determined by (7.64); see Exercise ■.

Remark: Technically, (7.64) only defines the "formal adjoint" of L. For the infinite-dimensional function spaces arising in analysis, a true adjoint must satisfy certain additional requirements, [**122**]. However, we will suppress all such advanced analytical complications in our introductory treatment of the subject.

Lemma 7.53. *The adjoint of a linear function is a linear function.*

Proof: Given $\mathbf{v}\in V$, $\mathbf{w},\mathbf{z}\in W$, and scalars $c,d\in\mathbb{R}$, we find

$$\begin{aligned}\langle\, \mathbf{v}\,;L^*[c\,\mathbf{w}+d\,\mathbf{z}]\,\rangle &= \langle\!\langle\, L[\mathbf{v}]\,;c\,\mathbf{w}+d\,\mathbf{z}\,\rangle\!\rangle = c\,\langle\!\langle\, L[\mathbf{v}]\,;\mathbf{w}\,\rangle\!\rangle + d\,\langle\!\langle\, L[\mathbf{v}]\,;\mathbf{z}\,\rangle\!\rangle\\ &= c\,\langle\, \mathbf{v}\,;L^*[\mathbf{w}]\,\rangle + d\,\langle\, \mathbf{v}\,;L^*[\mathbf{z}]\,\rangle = \langle\, \mathbf{v}\,;c\,L^*[\mathbf{w}]+d\,L^*[\mathbf{z}]\,\rangle.\end{aligned}$$

Since this holds for all $\mathbf{v}\in V$, we must have

$$L^*[c\,\mathbf{w}+d\,\mathbf{z}] = c\,L^*[\mathbf{w}]+d\,L^*[\mathbf{z}],$$

proving linearity. *Q.E.D.*

The proof of the next result is left as an exercise.

Lemma 7.54. *The adjoint of the adjoint of L is just $L=(L^*)^*$.*

Example 7.55. Let us first show how the defining equation (7.64) for the adjoint leads directly to the transpose of a matrix. Let $L\colon\mathbb{R}^n\to\mathbb{R}^m$ be the linear function $L[\mathbf{v}]=A\,\mathbf{v}$ defined by multiplication by the $m\times n$ matrix A. Then $L^*\colon\mathbb{R}^m\to\mathbb{R}^n$ is linear, and so is represented by matrix multiplication, $L^*[\mathbf{w}]=A^*\,\mathbf{w}$, by an $n\times m$ matrix A^*. We impose the ordinary Euclidean dot products

$$\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle = \mathbf{v}^T\,\widetilde{\mathbf{v}}, \qquad \mathbf{v},\widetilde{\mathbf{v}}\in\mathbb{R}^n, \qquad\qquad \langle\!\langle\, \mathbf{w}\,;\widetilde{\mathbf{w}}\,\rangle\!\rangle = \mathbf{w}^T\,\widetilde{\mathbf{w}}, \qquad \mathbf{w},\widetilde{\mathbf{w}}\in\mathbb{R}^m,$$

as our inner products on both $\mathbb{R}^n$ and $\mathbb{R}^m$. Evaluation of both sides of the adjoint equation (7.64) gives

$$\begin{aligned} \langle\!\langle\, L[\mathbf{v}]\,;\mathbf{w}\,\rangle\!\rangle &= \langle\!\langle\, A\,\mathbf{v}\,;\mathbf{w}\,\rangle\!\rangle = (A\,\mathbf{v})^T\mathbf{w} = \mathbf{v}^T A^T\,\mathbf{w}, \\ \langle\,\mathbf{v}\,;L^*[\mathbf{w}]\,\rangle &= \langle\,\mathbf{v}\,;A^*\,\mathbf{w}\,\rangle = \mathbf{v}^T A^*\,\mathbf{w}. \end{aligned} \tag{7.65}$$

Since these must agree for all $\mathbf{v}, \mathbf{w}$, cf. Exercise ■, the matrix A^* representing L^* is equal to the transposed matrix A^T. Therefore, *the adjoint of a matrix with respect to the Euclidean inner product is its transpose*: $A^* = A^T$.

Example 7.56. Let us now adopt different, weighted inner products on the domain and target spaces for the linear map $L\colon \mathbb{R}^n \to \mathbb{R}^m$ given by $L[\mathbf{v}] = A\,\mathbf{v}$. Suppose that

the inner product on the domain space $\mathbb{R}^n$ is given by $\langle\,\mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle = \mathbf{v}^T M\,\widetilde{\mathbf{v}}$, while
the inner product on the target space $\mathbb{R}^m$ is given by $\langle\!\langle\, \mathbf{w}\,;\widetilde{\mathbf{w}}\,\rangle\!\rangle = \mathbf{w}^T C\,\widetilde{\mathbf{w}}$,

where $M > 0$ and $C > 0$ are positive definite matrices of respective sizes $m \times m$ and $n \times n$. Then, in place of (7.65), we have

$$\langle\!\langle\, A\,\mathbf{v}\,;\mathbf{w}\,\rangle\!\rangle = (A\,\mathbf{v})^T C\,\mathbf{w} = \mathbf{v}^T A^T C\,\mathbf{w}, \qquad\qquad \langle\,\mathbf{v}\,;A^*\,\mathbf{w}\,\rangle = \mathbf{v}^T M\,A^*\,\mathbf{w}.$$

Equating these expressions, we deduce that $A^T C = M\,A^*$. Therefore the *weighted adjoint* of the matrix A is given by the more complicated formula

$$A^* = M^{-1}\,A^T C. \tag{7.66}$$

In applications, M plays the role of the mass matrix, and explicitly appears in the dynamical systems to be solved in Chapter 9. In particular, suppose A is square, defining a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$. If we adopt the same inner product $\langle\,\mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle = \mathbf{v}^T C\,\widetilde{\mathbf{v}}$ on both the domain and target spaces $\mathbb{R}^n$, then the adjoint matrix $A^* = C^{-1}\,A^T C$ is similar to the transpose.

Everything that we learned about transposes can be reinterpreted in the more general language of adjoints. The next result generalizes the fact, (1.49), that the transpose of the product of two matrices is the product of the transposes, in the reverse order.

Lemma 7.57. *If $L\colon V \to W$ and $M\colon W \to Z$ have respective adjoints $L^*\colon W \to V$ and $M^*\colon Z \to W$, then the composite linear function $M \circ L\colon V \to Z$ has adjoint $(M \circ L)^* = L^* \circ M^*$, which maps Z to V.*

Proof: Let $\langle\,\mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle, \langle\!\langle\, \mathbf{w}\,;\widetilde{\mathbf{w}}\,\rangle\!\rangle, \langle\!\langle\!\langle\, \mathbf{z}\,;\widetilde{\mathbf{z}}\,\rangle\!\rangle\!\rangle$, denote, respectively, the inner products on V, W, Z. For $\mathbf{v} \in V$, $\mathbf{z} \in Z$, we compute using the definition (7.64),

$$\begin{aligned} \langle\,\mathbf{v}\,;(M \circ L)^*[\mathbf{z}]\,\rangle &= \langle\!\langle\!\langle\, M \circ L[\mathbf{v}]\,;\mathbf{z}\,\rangle\!\rangle\!\rangle = \langle\!\langle\!\langle\, M[L[\mathbf{v}]]\,;\mathbf{z}\,\rangle\!\rangle\!\rangle \\ &= \langle\!\langle\, L[\mathbf{v}]\,;M^*[\mathbf{z}]\,\rangle\!\rangle = \langle\,\mathbf{v}\,;L^*[M^*[\mathbf{z}]]\,\rangle = \langle\,\mathbf{v}\,;(L^* \circ M^*)[\mathbf{z}]\,\rangle. \end{aligned}$$

Since this holds for all $\mathbf{v}$ and $\mathbf{z}$, the identification follows. *Q.E.D.*

In this chapter, we have only looked at adjoints in the finite-dimensional situation, when the linear functions are given by matrix multiplication. The equally important case of adjoints of linear operators on function spaces, e.g., differential operators appearing in boundary value problems, will be a principal focus of Section 11.3.

Self–Adjoint and Positive Definite Linear Functions

Throughout this section V will be a fixed inner product space. We can generalize the notions of symmetric and positive definite matrices to linear operators on V in a natural fashion. The analog of a symmetric matrix is a self-adjoint linear function.

Definition 7.58. A linear function $K: V \to V$ is called *self-adjoint* if $K^* = K$. A self-adjoint linear function is *positive definite* if

$$\langle\, \mathbf{v}\,;K[\,\mathbf{v}\,]\,\rangle > 0 \qquad \text{for all} \qquad \mathbf{0} \neq \mathbf{v} \in V. \tag{7.67}$$

In particular, if $K > 0$ then $\ker K = \{\mathbf{0}\}$, and so the positive definite linear system $K[\,\mathbf{u}\,] = \mathbf{f}$ with $\mathbf{f} \in \operatorname{rng} K$ has a unique solution. The next result generalizes our basic observation that the Gram matrices $K = A^T A$, cf. (3.49), are symmetric and positive (semi-)definite.

Theorem 7.59. *Let $L: V \to W$ be a linear map between inner product spaces with adjoint $L^*: W \to V$. Then the composite map $K = L^* \circ L: V \to V$ is self-adjoint. Moreover, K is positive definite if and only if* $\ker L = \{\mathbf{0}\}$.

Proof: First, by Lemmas 7.57 and 7.54,

$$K^* = (L^* \circ L)^* = L^* \circ (L^*)^* = L^* \circ L = K,$$

proving self-adjointness. Furthermore, for $\mathbf{v} \in V$, the inner product

$$\langle\, \mathbf{v}\,;K[\,\mathbf{v}\,]\,\rangle = \langle\, \mathbf{v}\,;L^*[\,L[\,\mathbf{v}\,]\,]\,\rangle = \langle\, L[\,\mathbf{v}\,]\,;L[\,\mathbf{v}\,]\,\rangle = \|\, L[\,\mathbf{v}\,]\,\|^2 > 0$$

is strictly positive provided $L[\,\mathbf{v}\,] \neq \mathbf{0}$. Thus, if $\ker L = \{\mathbf{0}\}$, then the positivity condition (7.67) holds, and conversely. *Q.E.D.*

Consider the case of a linear function $L: \mathbb{R}^n \to \mathbb{R}^m$ that is represented by the $m \times n$ matrix A. For the Euclidean dot product on the two spaces, the adjoint L^* is represented by the transpose A^T, and hence the map $K = L^* \circ L$ has matrix representation $A^T A$. Therefore, in this case Theorem 7.59 reduces to our earlier Proposition 3.32 governing the positive definiteness of the Gram matrix product $A^T A$. If we change the inner product on the target space space to $\langle\!\langle\, \mathbf{w}\,;\widetilde{\mathbf{w}}\,\rangle\!\rangle = \mathbf{w}^T C\,\widetilde{\mathbf{w}}$, then L^* is represented by $A^T C$, and hence $K = L^* \circ L$ has matrix form $A^T C\, A$, which is the general symmetric, positive definite Gram matrix constructed in (3.51) that played a key role in our development of the equations of equilibrium in Chapter 6. Finally, if we also use the alternative inner product $\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle = \mathbf{v}^T M\,\widetilde{\mathbf{v}}$ on the domain space $\mathbb{R}^n$, then, according to (7.66), the adjoint of L has matrix form

$$A^* = M^{-1}\, A^T C, \qquad \text{and therefore} \qquad K = A^* A = M^{-1}\, A^T C\, A \tag{7.68}$$

is a self-adjoint, positive definite matrix with respect to the weighted inner product on $\mathbb{R}^n$ prescribed by the positive definite matrix M. In this case, the positive definite, self-adjoint operator K is no longer represented by a symmetric matrix. So, we did not quite tell the truth when we said we would only allow symmetric matrices to be positive definite — we really meant only self-adjoint matrices. The general case will be important in our discussion of the vibrations of mass/spring chains that have unequal masses. Extensions of these constructions to differential operators underlies the analysis of the static and dynamic differential equations of continuum mechanics, to be studied in Chapters 11–18.

Minimization

In Chapter 4, we learned that the solution to a matrix system $K\mathbf{u} = \mathbf{f}$, with positive definite coefficient matrix $K > 0$, can be characterized as the unique minimizer for the quadratic function $p(\mathbf{u}) = \frac{1}{2}\,\mathbf{u}^T K\mathbf{u} - \mathbf{u}^T\mathbf{f}$. There is an analogous minimization principle that characterizes the solutions to linear systems defined by positive definite linear operators. This general result is of tremendous importance in analysis of boundary value problems for differential equations and also underlies the finite element numerical solution algorithms. Details will appear in the subsequent chapters.

Theorem 7.60. *Let $K\colon V \to V$ be a positive definite operator on an inner product space V. If $\mathbf{f} \in \operatorname{rng} K$, then the quadratic function*

$$p(\mathbf{u}) = \tfrac{1}{2}\,\langle\, \mathbf{u}\,;K[\mathbf{u}]\,\rangle - \langle\, \mathbf{u}\,;\mathbf{f}\,\rangle \tag{7.69}$$

has a unique minimizer, which is the solution $\mathbf{u} = \mathbf{u}^\star$ to the linear system $K[\mathbf{u}] = \mathbf{f}$.

Proof: The proof mimics that of its matrix counterpart in Theorem 4.1. Since $\mathbf{f} = K[\mathbf{u}^\star]$, we can write

$$p(\mathbf{u}) = \tfrac{1}{2}\,\langle\, \mathbf{u}\,;K[\mathbf{u}]\,\rangle - \langle\, \mathbf{u}\,;K[\mathbf{u}^\star]\,\rangle = \tfrac{1}{2}\,\langle\, \mathbf{u}-\mathbf{u}^\star\,;K[\mathbf{u}-\mathbf{u}^\star]\,\rangle - \tfrac{1}{2}\,\langle\, \mathbf{u}^\star\,;K[\mathbf{u}^\star]\,\rangle. \tag{7.70}$$

where we used linearity, along with the fact that K is self-adjoint to identify the terms $\langle\, \mathbf{u}\,;K[\mathbf{u}^\star]\,\rangle = \langle\, \mathbf{u}^\star\,;K[\mathbf{u}]\,\rangle$. Since $K > 0$ is positive definite, the first term on the right hand side of (7.70) is always ≥ 0; moreover it equals its minimal value 0 if and only if $\mathbf{u} = \mathbf{u}^\star$. On the other hand, the second term does not depend upon $\mathbf{u}$ at all, and hence is a constant. Therefore, to minimize $p(\mathbf{u})$ we must make the first term as small as possible, which is accomplished by setting $\mathbf{u} = \mathbf{u}^\star$. *Q.E.D.*

Remark: For linear functions given by matrix multiplication, positive definiteness automatically implies invertibility, and so the linear system $K\mathbf{u} = \mathbf{f}$ has a solution for every right hand side. This is no longer necessarily true when K is a positive definite operator on an infinite-dimensional function space. Therefore, the existence of a solution or minimizer is a significant issue. And, in fact, many modern analytical existence results rely on such minimization principles.

Theorem 7.61. *Suppose $L\colon V \to W$ is a linear map between inner product spaces with $\ker L = \{\mathbf{0}\}$ and adjoint map $L^*\colon W \to V$. Let $K = L^* \circ L\colon V \to V$ be the associated positive definite operator. If $\mathbf{f} \in \operatorname{rng} K$, then the quadratic function*

$$p(\mathbf{u}) = \tfrac{1}{2}\,\|\, L[\mathbf{u}]\,\|^2 - \langle\, \mathbf{u}\,;\mathbf{f}\,\rangle \tag{7.71}$$

has a unique minimizer $\mathbf{u}^\star$*, which is the solution to the linear system* $K[\,\mathbf{u}^\star\,] = \mathbf{f}$.

Proof: It suffices to note that the quadratic term in (7.69) can be written in the alternative form

$$\langle\, \mathbf{u}\,;K[\,\mathbf{u}\,]\,\rangle = \langle\, \mathbf{u}\,;L^*[\,L[\,\mathbf{u}\,]\,]\,\rangle = \langle\, L[\,\mathbf{u}\,]\,;L[\,\mathbf{u}\,]\,\rangle = \|\, L[\,\mathbf{u}\,]\,\|^2.$$

Thus, (7.71) reduces to the quadratic function of the form (7.69) with $K = L^* \circ L$, and so Theorem 7.61 is an immediate consequence of Theorem 7.60. *Q.E.D.*

Warning: In (7.71), the first term $\|\, L[\,\mathbf{u}\,]\,\|^2$ is computed using the norm based on the inner product on W, while the second term $\langle\, \mathbf{u}\,;\mathbf{f}\,\rangle$ employs the inner product on V.

Example 7.62. For a generalized positive definite matrix (7.68), the quadratic function (7.71) is computed with respect to the alternative inner product $\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle = \mathbf{v}^T M\,\widetilde{\mathbf{v}}$, so

$$p(\mathbf{u}) = \tfrac{1}{2}\,(A\,\mathbf{u})^T C\,A\,\mathbf{u} - \mathbf{u}^T M\,\mathbf{f} = \tfrac{1}{2}\,\mathbf{u}^T(A^T C\,A)\mathbf{u} - \mathbf{u}^T(M\,\mathbf{f}).$$

Theorem 7.61 tells us that the minimizer of the quadratic function is the solution to

$$A^T C\,A\,\mathbf{u} = M\,\mathbf{f}, \qquad \text{or} \qquad K\,\mathbf{u} = M^{-1}\,A^T C\,A\,\mathbf{u} = \mathbf{f}.$$

This also follows from our earlier finite-dimensional minimization Theorem 4.1.

This section is a preview of things to come, but the full implications will require us to develop more analytical expertise. In Chapters 11, 15 and 18 , we will find that the most important minimization principles for characterizing solutions to the linear boundary value problems of physics and engineering all arise through this general, abstract construction.

Chapter 8

Eigenvalues

So far, our applications have concentrated on statics: unchanging equilibrium configurations of physical systems — mass/spring chains, circuits, and structures — that are modeled by linear systems of algebraic equations. It is now time to allow motion in our universe. In general, a *dynamical system* refers to the (differential) equations governing the temporal behavior of some physical system: mechanical, electrical, chemical, fluid, etc. Our immediate goal is to understand the behavior of the simplest class of linear dynamical systems — frirst order autonomous linear systems of ordinary differential equations. As always, complete analysis of the linear situation is an essential prerequisite to making progress in the more complicated nonlinear realm.

We begin with a very quick review of the scalar case, whose solutions are exponential functions. Substituting a similar exponential solution ansatz[†] into the system leads us immediately to the equations defining the eigenvalues and eigenvectors of the coefficient matrix. Eigenvalues and eigenvectors are of absolutely fundamental importance in both the mathematical theory and a very wide range of applications, including iterative systems and numerical solution methods. Thus, to continue we need to gain a proper grounding in their basic theory and computation.

The present chapter develops the most important properties of eigenvalues and eigenvectors; the applications to dynamical systems will appear in Chapter 9, while applications to iterative systems and numerical methods is the topic of Chapter 10. Extensions of the eigenvalue concept to differential operators acting on infinite-dimensional function space, of essential importance for solving linear partial differential equations modelling continuous dynamical systems, will be covered in later chapters.Each square matrix has a collection of one or more complex scalars called eigenvalues and associated vectors, called eigenvectors. Roughly speaking, the eigenvectors indicate directions of pure stretch and the eigenvalues the amount of stretching. Most matrices are complete, meaning that their (complex) eigenvectors form a basis of the underlying vector space. When written in the eigenvector basis, the matrix assumes a very simple diagonal form, and the analysis of its properties becomes extremely simple. A particularly important class are the symmetric matrices, whose eigenvectors form an orthogonal basis of $\mathbb{R}^n$; in fact, this is by far the most common way for orthogonal bases to appear. Incomplete matrices are trickier, and we relegate them and their associated non-diagonal Jordan canonical form to the final section. The numerical computation of eigenvalues and eigenvectors is a challenging issue, and must be

[†] See the footnote in Chapter 7 for an explanation of the term "ansatz" or inspired guess.

be deferred until Section 10.6. Unless you are prepared to consult that section now, in order to solve the computer-based problems in this chapter, you will need to make use of a program that can accurately compute eigenvalues and eigenvectors of matrices.

A non-square matrix A does not have eigenvalues; however, we have already made extensive use of the associated square Gram matrix $K = A^T A$. The square roots of the eigenvalues of K serve to define the singular values of A. Singular values and principal component analysis are now used in an increasingly broad range of modern applications, including statistical analysis, image processing, semantics, language and speech recognition, and learning theory. The singular values are used to define the condition number of a matrix, that indicates the degree of difficulty of accurately solving the associated linear system.

8.1. Simple Dynamical Systems.

The purpose of this section is to motivate the concepts of eigenvalue and eigenvector of square matrices by attempting to solve the simplest class of dynamical systems — first order linear systems of ordinary differential equations. We begin with a review of the scalar case, introducing basic notions of stability in preparation for the general version, to be treated in depth in Chapter 9. We use the exponential form of the scalar solution as a template for a possible solution in the vector case, and this immediately leads us to the fundamental eigenvalue/eigenvector equation. Readers who are uninterested in such motivations are advised skip ahead to Section 8.2.

Scalar Ordinary Differential Equations

Eigenvalues first appear when attempting to solve linear systems of ordinary differential equations. In order to motivate the construction, we begin by reviewing the scalar case. Consider the elementary ordinary differential equation

$$\frac{du}{dt} = a\,u. \tag{8.1}$$

Here $a \in \mathbb{R}$ is a real constant, while the unknown $u(t)$ is a scalar function. As you learned in first year calculus, the general solution to (8.1) is an exponential function

$$u(t) = c\,e^{a\,t}. \tag{8.2}$$

The integration constant c is uniquely determined by a single initial condition

$$u(t_0) = b \tag{8.3}$$

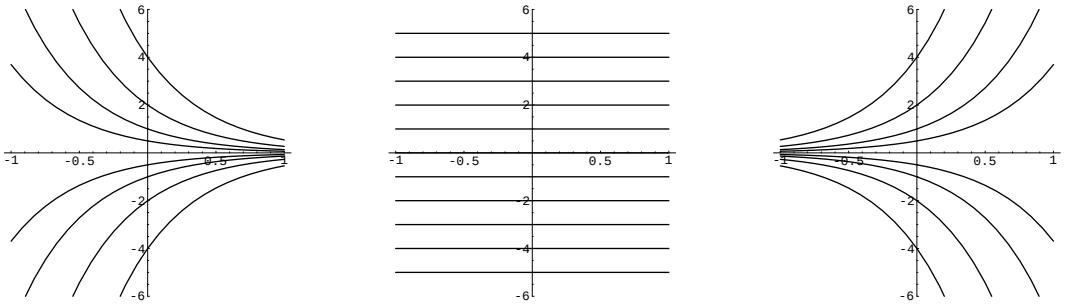
imposed at an initial time t_0. Substituting $t = t_0$ into the solution formula (8.2),

$$u(t_0) = c\,e^{a\,t_0} = b, \qquad \text{and so} \qquad c = b\,e^{-a\,t_0}.$$

We conclude that

$$u(t) = b\,e^{a(t-t_0)}. \tag{8.4}$$

is the unique solution to the scalar initial value problem (8.1), (8.3).



$a < 0$ $\qquad$ $a = 0$ $\qquad$ $a > 0$

Figure 8.1. Solutions to $\dot{u} = a\,u$.

Example 8.1. The radioactive decay of an isotope, say Uranium 238, is governed by the differential equation

$$\frac{du}{dt} = -\,\gamma\,u. \tag{8.5}$$

Here $u(t)$ denotes the amount of the isotope remaining at time t, and the coefficient $\gamma > 0$ governs the decay rate. The solution is given by an exponentially decaying function $u(t) = c\,e^{-\gamma\,t}$, where $c = u(0)$ is the initial amount of radioactive material.

The *half-life* $t^\star$ is the time it takes for half of a sample to decay, that is when $u(t^\star) = \frac{1}{2}\,u(0)$. To determine $t^\star$, we solve the algebraic equation

$$e^{-\gamma\,t^\star} = \tfrac{1}{2}, \qquad \text{so that} \qquad t^\star = \frac{\log 2}{\gamma}\,. \tag{8.6}$$

At each integer multiple $n\,t^\star$ of the half-life, exactly half of the isotope has decayed, i.e., $u(n\,t^\star) = 2^{-n}\,u(0)$.

Let us make some elementary, but pertinent observations about this simple linear dynamical system. First of all, since the equation is homogeneous, the zero function $u(t) \equiv 0$ (corresponding to $c = 0$) is a constant solution, known as an *equilibrium solution* or *fixed point*, since it does not depend on t. If the coefficient $a > 0$ is positive, then the solutions (8.2) are exponentially growing (in absolute value) as $t \to +\infty$. This implies that the zero equilibrium solution is *unstable*. The initial condition $u(t_0) = 0$ produces the zero solution, but if we make a tiny error (either physical, numerical, or mathematical) in the initial data, say $u(t_0) = \varepsilon$, then the solution $u(t) = \varepsilon\,e^{a(t-t_0)}$ will eventually get very far away from equilibrium. More generally, any two solutions with very close, but not equal, initial data, will eventually become arbitrarily far apart: $|\,u_1(t) - u_2(t)\,| \to \infty$ as $t \to \infty$. One consequence is the inherent difficulty in accurately computing the long time behavior of the solution, since small numerical errors will eventually have very large effects.

On the other hand, if $a < 0$, the solutions are exponentially decaying in time. In this case, the zero solution is *stable*, since a small error in the initial data will have a negligible effect on the solution. In fact, the zero solution is *globally asymptotically stable*. The phrase "asymptotically stable" implies that solutions that start out near zero eventually return; more specifically, if $u(t_0) = \varepsilon$ is small, then $u(t) \to 0$ as $t \to \infty$. The adjective "globally" implies that this happens no matter how large the initial data is. In fact, for

a linear system, the stability (or instability) of an equilibrium solution is always a global phenomenon.

The borderline case is when $a = 0$. Then all the solutions to (8.1) are constant. In this case, the zero solution is *stable* — indeed, globally stable — but not asymptotically stable. The solution to the initial value problem $u(t_0) = \varepsilon$ is $u(t) \equiv \varepsilon$. Therefore, a solution that starts out near equilibrium will remain near, but will not asymptotically return. The three qualitatively different possibilities are illustrated in Figure 8.1.

First Order Dynamical Systems

The simplest class of *dynamical systems* consist of n first order ordinary differential equations for n unknown functions

$$\frac{du_1}{dt} = f_1(t, u_1, \dots, u_n), \qquad \dots \qquad \frac{du_n}{dt} = f_n(t, u_1, \dots, u_n),$$

which depend on a scalar variable $t \in \mathbb{R}$, which we usually view as time. We will often write the system in the equivalent vector form

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(t, \mathbf{u}). \tag{8.7}$$

The vector-valued solution $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))^T$ serves to parametrize a curve in $\mathbb{R}^n$, called a *solution trajectory*. A dynamical system is called *autonomous* if the time variable t does not appear explicitly on the right hand side, and so has the system has the form

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}). \tag{8.8}$$

Dynamical systems of ordinary differential equations appear in an astonishing variety of applications, and have been the focus of intense research activity since the early days of calculus.

We shall concentrate most of our attention on the very simplest case: a homogeneous, linear, autonomous dynamical system

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}, \tag{8.9}$$

in which A is a constant $n \times n$ matrix. In full detail, the system consists of n linear ordinary differential equations

$$\begin{aligned}
\frac{du_1}{dt} &= a_{11}\,u_1 + a_{12}\,u_2 + \,\cdots\, + a_{1n}\,u_n,\\
\frac{du_2}{dt} &= a_{21}\,u_1 + a_{22}\,u_2 + \,\cdots\, + a_{2n}\,u_n,\\
&\vdots \qquad\qquad\qquad\qquad \vdots\\
\frac{du_n}{dt} &= a_{n1}\,u_1 + a_{n2}\,u_2 + \,\cdots\, + a_{nn}\,u_n,
\end{aligned} \tag{8.10}$$

involving n unknown functions $u_1(t), u_2(t), \dots, u_n(t)$. In the autonomous case, the coefficients a_{ij} are assumed to be (real) constants. We seek not only to develop basic solution

techniques for such dynamical systems, but to also understand their behavior from both a qualitative and quantitative standpoint.

Drawing our inspiration from the exponential solution formula (8.2) in the scalar case, let us investigate whether the vector system has any solutions of a similar exponential form

$$\mathbf{u}(t) = e^{\lambda t}\,\mathbf{v}, \tag{8.11}$$

in which λ is a constant scalar, so $e^{\lambda t}$ is a scalar function of t, while $\mathbf{v} \in \mathbb{R}^n$ is a constant vector. In other words, the components $u_i(t) = v_i\, e^{\lambda t}$ of our desired solution are assumed to be constant multiples of the *same* exponential function. Since $\mathbf{v}$ is assumed to be constant, the derivative of $\mathbf{u}(t)$ is easily found:

$$\frac{d\mathbf{u}}{dt} = \frac{d}{dt}\left(e^{\lambda t}\,\mathbf{v}\right) = \lambda\, e^{\lambda t}\,\mathbf{v}.$$

On the other hand, since $e^{\lambda t}$ is a scalar, it commutes with matrix multiplication, and so

$$A\,\mathbf{u} = A\, e^{\lambda t}\,\mathbf{v} = e^{\lambda t} A\,\mathbf{v}.$$

Therefore, $\mathbf{u}(t)$ will solve the system (8.9) if and only if

$$\lambda\, e^{\lambda t}\,\mathbf{v} = e^{\lambda t} A\,\mathbf{v},$$

or, canceling the common scalar factor $e^{\lambda t}$,

$$\lambda\,\mathbf{v} = A\,\mathbf{v}.$$

The result is a system of algebraic equations relating the vector $\mathbf{v}$ and the scalar λ. Analysis of this system and its ramifications will be the topic of the remainder of this chapter. After gaining a complete understanding, we will return to the solution of linear dynamical systems in Chapter 9.

8.2. Eigenvalues and Eigenvectors.

We inaugurate our discussion of eigenvalues and eigenvectors with the fundamental definition.

Definition 8.2. Let A be an $n \times n$ matrix. A scalar λ is called an *eigenvalue* of A if there is a *non-zero* vector $\mathbf{v} \neq \mathbf{0}$, called an *eigenvector*, such that

$$A\,\mathbf{v} = \lambda\,\mathbf{v}. \tag{8.12}$$

Thus, the matrix A effectively stretches the eigenvector $\mathbf{v}$ by an amount specified by the eigenvalue λ. In this manner, the eigenvectors specify the directions of pure stretch for the linear transformation defined by the matrix A.

Remark: The odd-looking terms "eigenvalue" and "eigenvector" are hybrid German–English words. In the original German, they are *Eigenwert* and *Eigenvektor*, which can be fully translated as "proper value" and "proper vector". For some reason, the half-translated terms have acquired a certain charm, and are now standard. The alternative English terms *characteristic value* and *characteristic vector* can be found in some (mostly older) texts. Oddly, the term *characteristic equation*, to be defined below, is still used.

The requirement that the eigenvector $\mathbf{v}$ be nonzero is important, since $\mathbf{v} = \mathbf{0}$ is a trivial solution to the eigenvalue equation (8.12) for *any* scalar λ. Moreover, as far as solving linear ordinary differential equations goes, the zero vector $\mathbf{v} = \mathbf{0}$ only gives the trivial zero solution $\mathbf{u}(t) \equiv \mathbf{0}$.

The eigenvalue equation (8.12) is a system of linear equations for the entries of the eigenvector $\mathbf{v}$ — provided the eigenvalue λ is specified in advance — but is "mildly" nonlinear as a combined system for λ and $\mathbf{v}$. Gaussian elimination per se will not solve the problem, and we are in need of a new idea. Let us begin by rewriting the equation in the form

$$(A - \lambda \, \mathrm{I}\,)\mathbf{v} = \mathbf{0}, \tag{8.13}$$

where I is the identity matrix of the correct size[†]. Now, for given λ, equation (8.13) is a homogeneous linear system for $\mathbf{v}$, and always has the trivial zero solution $\mathbf{v} = \mathbf{0}$. But we are specifically seeking a nonzero solution! According to Theorem 1.45, a homogeneous linear system has a nonzero solution $\mathbf{v} \neq \mathbf{0}$ if and only if its coefficient matrix, which in this case is $A - \lambda \, \mathrm{I}\,$, is singular. This observation is the key to resolving the eigenvector equation.

Theorem 8.3. *A scalar λ is an eigenvalue of the $n \times n$ matrix A if and only if the matrix $A - \lambda \, \mathrm{I}\,$ is singular, i.e., of rank $< n$. The corresponding eigenvectors are the nonzero solutions to the eigenvalue equation $(A - \lambda \, \mathrm{I}\,)\mathbf{v} = \mathbf{0}$.*

We know a number of ways to characterize singular matrices, including the determinantal criterion given in Theorem 1.50. Therefore, the following result is an immediate corollary of Theorem 8.3.

Proposition 8.4. *A scalar λ is an eigenvalue of the matrix A if and only if λ is a solution to the* characteristic equation

$$\det(A - \lambda \, \mathrm{I}\,) = 0. \tag{8.14}$$

In practice, when finding eigenvalues and eigenvectors by hand, one first solves the characteristic equation (8.14). Then, for each eigenvalue λ one uses standard linear algebra methods, i.e., Gaussian elimination, to solve the corresponding linear system (8.13) for the eigenvector $\mathbf{v}$.

Example 8.5. Consider the 2×2 matrix

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

We compute the determinant in the characteristic equation using (1.34):

$$\det(A - \lambda \, \mathrm{I}\,) = \det \begin{pmatrix} 3 - \lambda & 1 \\ 1 & 3 - \lambda \end{pmatrix} = (3 - \lambda)^2 - 1 = \lambda^2 - 6\,\lambda + 8.$$

[†] Note that it is not legal to write (8.13) in the form $(A - \lambda)\mathbf{v} = \mathbf{0}$ since we do not know how to subtract a scalar λ from a matrix A. Worse, if you type $A - \lambda$ in MATLAB, it will subtract λ from *all* the entries of A, which is *not* what we are after!

The characteristic equation is a quadratic polynomial equation, and can be solved by factorization:

$$\lambda^2 - 6\,\lambda + 8 = (\lambda - 4)\,(\lambda - 2) = 0.$$

We conclude that A has two eigenvalues: $\lambda_1 = 4$ and $\lambda_2 = 2$.

For each eigenvalue, the corresponding eigenvectors are found by solving the associated homogeneous linear system (8.13). For the first eigenvalue, the corresponding eigenvector equation is

$$(A - 4\,\mathrm{I})\,\mathbf{v} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \text{or} \qquad \begin{aligned} -x + y &= 0, \\ x - y &= 0. \end{aligned}$$

The general solution is

$$x = y = a, \qquad \text{so} \qquad \mathbf{v} = \begin{pmatrix} a \\ a \end{pmatrix} = a \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where a is an arbitrary scalar. Only the nonzero solutions[†] count as eigenvectors, and so the eigenvectors for the eigenvalue $\lambda_1 = 4$ must have $a \neq 0$, i.e., they are all nonzero scalar multiples of the basic eigenvector $\mathbf{v}_1 = (\,1, 1\,)^T$.

Remark: In general, if $\mathbf{v}$ is an eigenvector of A for the eigenvalue λ, then so is any nonzero scalar multiple of $\mathbf{v}$. In practice, we only distinguish linearly independent eigenvectors. Thus, in this example, we shall say "$\mathbf{v}_1 = (\,1, 1\,)^T$ is *the* eigenvector corresponding to the eigenvalue $\lambda_1 = 4$", when we really mean that the eigenvectors for $\lambda_1 = 4$ consist of all nonzero scalar multiples of $\mathbf{v}_1$.

Similarly, for the second eigenvalue $\lambda_2 = 2$, the eigenvector equation is

$$(A - 2\,\mathrm{I})\,\mathbf{v} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The solution $(\,-a, a\,)^T = a\,(\,-1, 1\,)^T$ is the set of scalar multiples of the eigenvector $\mathbf{v}_2 = (\,-1, 1\,)^T$. Therefore, the complete list of eigenvalues and eigenvectors (up to scalar multiple) is

$$\lambda_1 = 4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = 2, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Example 8.6. Consider the 3×3 matrix

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

[†] If, at this stage, you end up with a linear system with only the trivial zero solution, you've done something wrong! Either you don't have a correct eigenvalue — maybe you made a mistake setting up and/or solving the characteristic equation — or you've made an error solving the homogeneous eigenvector system.

Using the formula (1.82) for a 3×3 determinant, we compute the characteristic equation

$$\begin{aligned} 0 = \det(A - \lambda\, \mathrm{I}\,) &= \det \begin{pmatrix} -\lambda & -1 & -1 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & 2-\lambda \end{pmatrix} \\ &= (-\lambda)(2-\lambda)^2 + (-1)\cdot 1 \cdot 1 + (-1)\cdot 1 \cdot 1 - \\ &\qquad - 1 \cdot (2-\lambda)(-1) - 1 \cdot 1 \cdot (-\lambda) - (2-\lambda)\cdot 1 \cdot (-1) \\ &= -\lambda^3 + 4\lambda^2 - 5\lambda + 2. \end{aligned}$$

The resulting cubic polynomial can be factorized:

$$-\lambda^3 + 4\lambda^2 - 5\lambda + 2 = -(\lambda - 1)^2 (\lambda - 2) = 0.$$

Most 3×3 matrices have three different eigenvalues, but this particular one has only two: $\lambda_1 = 1$, which is called a double eigenvalue since it is a double root of the characteristic equation, along with a simple eigenvalue $\lambda_2 = 2$.

The eigenvector equation (8.13) for the double eigenvalue $\lambda_1 = 1$ is

$$(A - \mathrm{I}\,)\mathbf{v} = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The general solution to this homogeneous linear system

$$\mathbf{v} = \begin{pmatrix} -a-b \\ a \\ b \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

depends upon two free variables, $y = a$, $z = b$. Any nonzero solution forms a valid eigenvector for the eigenvalue $\lambda_1 = 1$, and so the general eigenvector is any non-zero linear combination of the two "basis eigenvectors" $\mathbf{v}_1 = (\,-1, 1, 0\,)^T$, $\widehat{\mathbf{v}}_1 = (\,-1, 0, 1\,)^T$.

On the other hand, the eigenvector equation for the simple eigenvalue $\lambda_2 = 2$ is

$$(A - 2\,\mathrm{I}\,)\mathbf{v} = \begin{pmatrix} -2 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The general solution

$$\mathbf{v} = \begin{pmatrix} -a \\ a \\ a \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$$

consists of all scalar multiple of the eigenvector $\mathbf{v}_2 = (\,-1, 1, 1\,)^T$.

In summary, the eigenvalues and (basis) eigenvectors for this matrix are

$$\begin{aligned} &\lambda_1 = 1, && \mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, && \widehat{\mathbf{v}}_1 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \\ &\lambda_2 = 2, && \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned} \tag{8.15}$$

In general, given an eigenvalue λ, the corresponding *eigenspace* $V_\lambda \subset \mathbb{R}^n$ is the subspace spanned by all its eigenvectors. Equivalently, the eigenspace is the kernel

$$V_\lambda = \ker(A - \lambda\, \mathrm{I}\,). \tag{8.16}$$

In particular, λ is an eigenvalue if and only if $V_\lambda \neq \{\mathbf{0}\}$ is a nontrivial subspace, and then every nonzero element of V_λ is a corresponding eigenvector. The most economical way to indicate each eigenspace is by writing out a basis, as in (8.15).

Example 8.7. The characteristic equation of the matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$ is

$$0 = \det(A - \lambda\, \mathrm{I}\,) = -\lambda^3 + \lambda^2 + 5\,\lambda + 3 = -\,(\lambda + 1)^2\,(\lambda - 3).$$

Again, there is a double eigenvalue $\lambda_1 = -1$ and a simple eigenvalue $\lambda_2 = 3$. However, in this case the matrix

$$A - \lambda_1\, \mathrm{I} = A + \mathrm{I} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 0 & 2 \end{pmatrix}$$

has only a one-dimensional kernel, spanned by $(\,2, -1, -2\,)^T$. Thus, even though λ_1 is a double eigenvalue, it only admits a one-dimensional eigenspace. The list of eigenvalues and eigenvectors is, in a sense, incomplete:

$$\lambda_1 = -1, \qquad \mathbf{v}_1 = \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix}, \qquad\qquad \lambda_2 = 3, \qquad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}.$$

Example 8.8. Finally, consider the matrix $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & -2 \\ 2 & 2 & -1 \end{pmatrix}$. The characteristic equation is

$$0 = \det(A - \lambda\, \mathrm{I}\,) = -\lambda^3 + \lambda^2 - 3\,\lambda - 5 = -\,(\lambda + 1)\,(\lambda^2 - 2\,\lambda + 5).$$

The linear factor yields the eigenvalue -1. The quadratic factor leads to two complex roots, $1 + 2\,\mathrm{i}$ and $1 - 2\,\mathrm{i}$, which can be obtained via the quadratic formula. Hence A has one real and two complex eigenvalues:

$$\lambda_1 = -1, \qquad\qquad \lambda_2 = 1 + 2\,\mathrm{i}\,, \qquad\qquad \lambda_3 = 1 - 2\,\mathrm{i}\,.$$

Complex eigenvalues are as important as real eigenvalues, and we need to be able to handle them too. To find the corresponding eigenvectors, which will also be complex, we need to solve the usual eigenvalue equation (8.13), which is now a complex homogeneous linear system. For example, the eigenvector(s) for $\lambda_2 = 1 + 2\,\mathrm{i}$ are found by solving

$$(A - (1 + 2\,\mathrm{i})\,\mathrm{I})\mathbf{v} = \begin{pmatrix} -2\,\mathrm{i} & 2 & 0 \\ 0 & -2\,\mathrm{i} & -2 \\ 2 & 2 & -2 - 2\,\mathrm{i} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This linear system can be solved by Gaussian elimination (with complex pivots). A simpler approach is to work directly: the first equation $-2\,\mathrm{i}\,x + 2y = 0$ tells us that $y = \mathrm{i}\,x$, while the second equation $-2\,\mathrm{i}\,y - 2z = 0$ says $z = -\mathrm{i}\,y = x$. If we trust our calculations so far, we do not need to solve the final equation $2x + 2y + (-2 - 2\,\mathrm{i})z = 0$, since we know that the coefficient matrix is singular and hence it must be a consequence of the first two equations. (However, it does serve as a useful check on our work.) So, the general solution $\mathbf{v} = (\,x, \mathrm{i}\,x, x\,)^T$ is an arbitrary constant multiple of the complex eigenvector $\mathbf{v}_2 = (\,1, \mathrm{i}\,, 1\,)^T$.

Summarizing, the matrix under consideration has three complex eigenvalues and three corresponding eigenvectors, each unique up to (complex) scalar multiple:

$$\begin{aligned} &\lambda_1 = -1, && \lambda_2 = 1 + 2\,\mathrm{i}\,, && \lambda_3 = 1 - 2\,\mathrm{i}\,, \\ &\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, && \mathbf{v}_2 = \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix}, && \mathbf{v}_3 = \begin{pmatrix} 1 \\ -\mathrm{i} \\ 1 \end{pmatrix}. \end{aligned}$$

Note that the third complex eigenvalue is the complex conjugate of the second, and the eigenvectors are similarly related. This is indicative of a general fact for real matrices:

Proposition 8.9. *If A is a real matrix with a complex eigenvalue $\lambda = \mu + \mathrm{i}\,\nu$ and corresponding complex eigenvector $\mathbf{v} = \mathbf{x} + \mathrm{i}\,\mathbf{y}$, then the complex conjugate $\overline{\lambda} = \mu - \mathrm{i}\,\nu$ is also an eigenvalue with complex conjugate eigenvector $\overline{\mathbf{v}} = \mathbf{x} - \mathrm{i}\,\mathbf{y}$.*

Proof: First take complex conjugates of the eigenvalue equation (8.12)

$$\overline{A}\,\overline{\mathbf{v}} = \overline{A\,\mathbf{v}} = \overline{\lambda\,\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{v}}.$$

Using the fact that a real matrix is unaffected by conjugation, so $\overline{A} = A$, we conclude

$$A\,\overline{\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{v}}, \tag{8.17}$$

which is the eigenvalue equation for the eigenvalue $\overline{\lambda}$ and eigenvector $\overline{\mathbf{v}}$. *Q.E.D.*

As a consequence, when dealing with real matrices, one only needs to compute the eigenvectors for *one* of each complex conjugate pair of eigenvalues. This observation effectively halves the amount of work in the unfortunate event that we are confronted with complex eigenvalues.

Remark: The reader may recall that we said one should never use determinants in practical computations. So why have we reverted to using determinants to find eigenvalues? The truthful answer is that the practical computation of eigenvalues and eigenvectors *never* resorts to the characteristic equation! The method is fraught with numerical traps and inefficiencies when (*a*) computing the determinant leading to the characteristic equation, then (*b*) solving the resulting polynomial equation, which is itself a nontrivial numerical problem, [**30**], and, finally, (*c*) solving each of the resulting linear eigenvector systems. Indeed, if we only know an approximation $\widetilde{\lambda}$ to the true eigenvalue λ, the approximate eigenvector system $(A - \widetilde{\lambda})\mathbf{v} = \mathbf{0}$ has a nonsingular coefficient matrix, and hence only admits the trivial solution — which does not even qualify as an eigenvector! Nevertheless, the characteristic equation does give us important theoretical insight into the structure of the eigenvalues of a matrix, and can be used on small, e.g., 2×2 and 3×3, matrices, when exact arithmetic is employed. Numerical algorithms for computing eigenvalues and eigenvectors are based on completely different ideas, and will be discussed in Section 10.6.

Basic Properties of Eigenvalues

If A is an $n \times n$ matrix, then its *characteristic polynomial* is

$$p_A(\lambda) = \det(A - \lambda\, \mathrm{I}\,) = c_n\, \lambda^n + c_{n-1}\, \lambda^{n-1} + \, \cdots \, + c_1\, \lambda + c_0. \tag{8.18}$$

The fact that $p_A(\lambda)$ is a polynomial of degree n is a consequence of the general determinantal formula (1.81). Indeed, every term is plus or minus a product of matrix entries containing one from each row and one from each column. The term corresponding to the identity permutation is obtained by multiplying the the diagonal entries together, which, in this case, is

$$(a_{11}-\lambda)\,(a_{22}-\lambda)\, \cdots \, (a_{nn}-\lambda) = (-1)^n\lambda^n + (-1)^{n-1}\big(\, a_{11} + a_{22} + \, \cdots \, + a_{nn}\,\big)\, \lambda^{n-1} + \cdots \,, \tag{8.19}$$

All of the other terms have at most $n-2$ diagonal factors $a_{ii} - \lambda$, and so are polynomials of degree $\leq n-2$ in λ. Thus, (8.19) is the only summand containing the monomials λ^n and λ^{n-1}, and so their respective coefficients are

$$c_n = (-1)^n, \qquad c_{n-1} = (-1)^{n-1}(a_{11} + a_{22} + \, \cdots \, + a_{nn}) = (-1)^{n-1}\, \mathrm{tr}\, A, \tag{8.20}$$

where $\mathrm{tr}\, A$, the sum of its diagonal entries, is called the *trace* of the matrix A. The other coefficients $c_{n-2}, \dots, c_1$ in (8.18) are more complicated combinations of the entries of A. However, setting $\lambda = 0$ implies $p_A(0) = \det A = c_0$, and hence the constant term equals the determinant of the matrix. In particular, if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a 2×2 matrix, its characteristic polynomial has the form

$$\begin{aligned} p_A(\lambda) = \det(A - \lambda\, \mathrm{I}\,) &= \det \begin{pmatrix} a - \lambda & b \\ c & d - \lambda \end{pmatrix} \\ &= \lambda^2 - (a+d)\lambda + (a\,d - b\,c) = \lambda^2 - (\mathrm{tr}\, A)\lambda + (\det A). \end{aligned} \tag{8.21}$$

As a result of these considerations, the characteristic equation of an $n \times n$ matrix A is a polynomial equation of degree n, namely $p_A(\lambda) = 0$. According to the Fundamental

Theorem of Algebra (see Corollary 16.63) every (complex) polynomial of degree n can be completely factored:

$$p_A(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n). \tag{8.22}$$

The complex numbers $\lambda_1, \ldots, \lambda_n$, some of which may be repeated, are the *roots* of the characteristic equation $p_A(\lambda) = 0$, and hence the eigenvalues of the matrix A. Therefore, we immediately conclude:

Theorem 8.10. *An $n \times n$ matrix A has at least one and at most n distinct complex eigenvalues.*

Most $n \times n$ matrices — meaning those for which the characteristic polynomial factors into n *distinct* factors — have *exactly* n complex eigenvalues. More generally, an eigenvalue λ_j is said to have *multiplicity* m if the factor $(\lambda - \lambda_j)$ appears exactly m times in the factorization (8.22) of the characteristic polynomial. An eigenvalue is *simple* if it has multiplicity 1. In particular, A has n distinct eigenvalues if and only if all its eigenvalues are simple. In all cases, when the eigenvalues are counted in accordance with their multiplicity, every $n \times n$ matrix has a total of n possibly repeated eigenvalues.

An example of a matrix with just one eigenvalue, of multiplicity n, is the $n \times n$ identity matrix I, whose only eigenvalue is $\lambda = 1$. In this case, *every* nonzero vector in $\mathbb{R}^n$ is an eigenvector of the identity matrix, and so the eigenspace is all of $\mathbb{R}^n$. At the other extreme, the "bidiagonal" *Jordan block matrix*

$$J_\lambda = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \tag{8.23}$$

also has only one eigenvalue, λ, again of multiplicity n. But in this case, J_λ has only one eigenvector (up to scalar multiple), which is the standard basis vector $\mathbf{e}_n$, and so its eigenspace is one-dimensional.

Remark: If λ is a complex eigenvalue of multiplicity k for the real matrix A, then its complex conjugate $\overline{\lambda}$ also has multiplicity k. This is because complex conjugate roots of a real polynomial necessarily appear with identical multiplicities.

Remark: If $n \leq 4$, then one can, in fact, write down an explicit formula for the solution to a polynomial equation of degree n, and hence explicit (but not particularly helpful) formulae for the eigenvalues of general $2 \times 2, 3 \times 3$ and 4×4 matrices. As soon as $n \geq 5$, there is no explicit formula (at least in terms of radicals), and so one must usually resort to numerical approximations. This remarkable and deep algebraic result was proved by the young Norwegian mathematician Nils Hendrik Abel in the early part of the nineteenth century, [**57**].

If we explicitly multiply out the factored product (8.22) and equate the result to the characteristic polynomial (8.18), we find that its coefficients $c_0, c_1, \dots c_{n-1}$ can be written as certain polynomials of the roots, known as the *elementary symmetric polynomials*. The first and last are of particular importance:

$$c_0 = \lambda_1 \lambda_2 \cdots \lambda_n, \qquad c_{n-1} = (-1)^{n-1} (\lambda_1 + \lambda_2 + \cdots + \lambda_n). \tag{8.24}$$

Comparison with our previous formulae for the coefficients c_0 and c_{n-1} leads us to the following useful result.

Proposition 8.11. *The sum of the eigenvalues of a matrix equals its trace*:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{tr} A = a_{11} + a_{22} + \cdots + a_{nn}. \tag{8.25}$$

The product of the eigenvalues equals its determinant:

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det A. \tag{8.26}$$

Remark: For repeated eigenvalues, one must add or multiply them in the formulae (8.25), (8.26) according to their multiplicity.

Example 8.12. The matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$ considered in Example 8.7 has trace and determinant

$$\operatorname{tr} A = 1, \qquad \det A = 3.$$

These fix, respectively, the coefficient of λ^2 and the constant term in the characteristic equation. This matrix has two distinct eigenvalues, -1, which is a double eigenvalue, and 3, which is simple. For this particular matrix, formulae (8.25), (8.26) become

$$1 = \operatorname{tr} A = (-1) + (-1) + 3, \qquad 3 = \det A = (-1)(-1)\, 3.$$

8.3. Eigenvector Bases and Diagonalization.

Most of the vector space bases that play a distinguished role in applications consist of eigenvectors of a particular matrix. In this section, we show that the eigenvectors for any "complete" matrix automatically form a basis for $\mathbb{R}^n$ or, in the complex case, $\mathbb{C}^n$. In the following subsection, we use the eigenvector basis to rewrite the linear transformation determined by the matrix in a simple diagonal form.

The first task is to show that eigenvectors corresponding to distinct eigenvalues are automatically linearly independent.

Lemma 8.13. *If $\lambda_1, \dots, \lambda_k$ are* distinct *eigenvalues of the same matrix A, then the corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ are linearly independent.*

Proof: We use induction on the number of eigenvalues. The case $k = 1$ is immediate since an eigenvector cannot be zero. Assume that we know the result for $k-1$ eigenvalues. Suppose we have a linear combination

$$c_1 \mathbf{v}_1 + \cdots + c_{k-1} \mathbf{v}_{k-1} + c_k \mathbf{v}_k = \mathbf{0} \tag{8.27}$$

which vanishes. Let us multiply this equation by the matrix A:

$$\begin{aligned} A\big(c_1 \mathbf{v}_1 + \cdots + c_{k-1} \mathbf{v}_{k-1} + c_k \mathbf{v}_k \big) &= c_1 A\mathbf{v}_1 + \cdots + c_{k-1} A\mathbf{v}_{k-1} + c_k A\mathbf{v}_k \\ &= c_1 \lambda_1 \mathbf{v}_1 + \cdots + c_{k-1} \lambda_{k-1} \mathbf{v}_{k-1} + c_k \lambda_k \mathbf{v}_k = \mathbf{0}. \end{aligned}$$

On the other hand, if we just multiply the original equation by λ_k, we also have

$$c_1 \lambda_k \mathbf{v}_1 + \cdots + c_{k-1} \lambda_k \mathbf{v}_{k-1} + c_k \lambda_k \mathbf{v}_k = \mathbf{0}.$$

Subtracting this from the previous equation, the final terms cancel and we are left with the equation

$$c_1(\lambda_1 - \lambda_k)\mathbf{v}_1 + \cdots + c_{k-1}(\lambda_{k-1} - \lambda_k)\mathbf{v}_{k-1} = \mathbf{0}.$$

This is a vanishing linear combination of the first $k-1$ eigenvectors, and so, by our induction hypothesis, can only happen if all the coefficients are zero:

$$c_1(\lambda_1 - \lambda_k) = 0, \qquad \dots \qquad c_{k-1}(\lambda_{k-1} - \lambda_k) = 0.$$

The eigenvalues were assumed to be distinct, so $\lambda_j \neq \lambda_k$ when $j \neq k$; consequently, $c_1 = \cdots = c_{k-1} = 0$. Substituting these values back into (8.27), we find $c_k \mathbf{v}_k = \mathbf{0}$, and so $c_k = 0$ also, since the eigenvector $\mathbf{v}_k \neq \mathbf{0}$. Thus we have proved that (8.27) holds if and only if $c_1 = \cdots = c_k = 0$, which implies the linear independence of the eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_k$. This completes the induction step. *Q.E.D.*

The most important consequence of this result is stated in the corollary.

Theorem 8.14. *If the $n \times n$ real matrix A has n distinct real eigenvalues $\lambda_1, \dots, \lambda_n$, then the corresponding real eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a basis of $\mathbb{R}^n$. If A (which may now be either a real or a complex matrix) has n distinct complex eigenvalues, then its eigenvectors form a basis of $\mathbb{C}^n$.*

If a matrix has multiple eigenvalues, then there may or may not be an eigenvector basis of $\mathbb{R}^n$ (or $\mathbb{C}^n$). The matrix in Example 8.6 has an eigenvector basis, whereas the matrix in Example 8.7 does not. In general, it can be proved that the dimension of the eigenspace is less than or equal to the multiplicity of the eigenvalue. In particular, a simple eigenvalue has a one-dimensional eigenspace, and hence, up to scalar multiple, only one associated eigenvector.

Definition 8.15. An eigenvalue λ of a matrix A is called *complete* if its eigenspace $V_\lambda = \ker(A - \lambda \, \mathrm{I})$ has the same dimension as its multiplicity. The matrix A is *complete* if all its eigenvalues are.

Remark: The multiplicity of an eigenvalue λ_i is sometimes referred to as its *algebraic multiplicity*. The dimension of the eigenspace V_λ is called the *geometric multiplicity*, and so completeness requires that the two multiplicities are equal. The word "complete" is not completely standard; other common terms for such matrices are *perfect*, *semi-simple* and, as discussed below, *diagonalizable*.

Note that a simple eigenvalue is automatically complete, and so only multiple eigenvalues can cause the incompleteness of a matrix.

Theorem 8.16. *An $n \times n$ real or complex matrix A is complete if and only if its eigenvectors span $\mathbb{C}^n$. In particular, any $n \times n$ matrix that has n distinct eigenvalues is complete.*

A $n \times n$ matrix is *incomplete* if it does not have n linearly independent complex eigenvectors. Most matrices, including those with all simple eigenvalues, are complete. Incomplete matrices are more tricky to deal with, and we relegate most of the messy details to Section 8.6.

Remark: We already noted that complex eigenvectors of a real matrix always appear in conjugate pairs: $\mathbf{v} = \mathbf{x} \pm \mathrm{i}\,\mathbf{y}$. It can be shown that the real and imaginary parts of these vectors form a real basis for $\mathbb{R}^n$. (See Exercise ■ for the underlying principle.) For instance, in Example 8.8, the complex eigenvectors are $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \pm \mathrm{i} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. The vectors $\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, consisting of the real eigenvector and the real and imaginary parts of the complex eigenvectors, form a basis for $\mathbb{R}^3$.

Diagonalization

Every $n \times n$ matrix A represents a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^n$, namely the function $L[\mathbf{u}] = A\mathbf{u}$ given by matrix multiplication. As we learned in Section 7.2, the matrix representing a linear transformation depends upon the choice basis of $\mathbb{R}^n$. Some bases give a particular simple matrix representation.

For example, the linear transformation $L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x - y \\ 2x + 4y \end{pmatrix}$ studied in Example 7.18 is represented by the matrix $A = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix}$ — when expressed in terms of the standard basis of $\mathbb{R}^2$. In terms of the alternative basis $\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$, the linear transformation is represented by the diagonal matrix $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$, which indicates its simple stretching action on the new basis: $A\mathbf{v}_1 = 2\mathbf{v}_1$ and $A\mathbf{v}_2 = 3\mathbf{v}_2$. Now we can understand the reason for this simplification. *The new basis consists of the two eigenvectors of the*

matrix A. This observation is indicative of a general fact: representing a linear transformation in terms of an eigenvector basis has the effect of replacing its matrix representative by a simple diagonal form. The effect is to *diagonalize* the original coefficient matrix.

According to (7.22), if $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a basis of $\mathbb{R}^n$, then the corresponding matrix representative of the linear transformation $L[\mathbf{v}] = A\mathbf{v}$ is given by the similar matrix $B = S^{-1}A\,S$, where $S = (\,\mathbf{v}_1\ \mathbf{v}_2\ \dots\ \mathbf{v}_n\,)^T$ is the matrix whose columns are the basis vectors. In the preceding example, $S = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$, and we find that $S^{-1}A\,S = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ is a diagonal matrix.

Definition 8.17. A square matrix A is called *diagonalizable* if there exists a nonsingular matrix S and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ such that

$$S^{-1}A\,S = \Lambda. \tag{8.28}$$

A diagonal matrix represents a linear transformation that simultaneously stretches[†] in the direction of the basis vectors. Thus, every diagonalizable matrix represents a elementary combination of (complex) stretching transformations.

To understand the diagonalization equation (8.28), we rewrite it in the equivalent form

$$A\,S = S\,\Lambda. \tag{8.29}$$

Using the basic property (1.11) of matrix multiplication, one easily sees that the k^{th} column of this $n \times n$ matrix equation is given by

$$A\mathbf{v}_k = \lambda_k \mathbf{v}_k.$$

Therefore, the columns of S are necessarily eigenvectors, and the entries of the diagonal matrix Λ are the corresponding eigenvalues! And, as a result, a diagonalizable matrix A must have n linearly independent eigenvectors, i.e., an eigenvector basis, to form the columns of the diagonalizing matrix S. Since the diagonal form Λ contains the eigenvalues along its diagonal, it is uniquely determined up to a permutation of its entries.

Now, as we know, not every matrix has an eigenvector basis. Moreover, even when it exists, the eigenvector basis may be complex, in which case S is a complex matrix, and the entries of the diagonal matrix Λ are the complex eigenvalues. Thus, we should distinguish between complete matrices that are diagonalizable over the complex numbers and the more restrictive class of matrices which can be diagonalized by a real matrix S.

Theorem 8.18. *A matrix is complex diagonalizable if and only if it is complete. A matrix is real diagonalizable if and only if it is complete and has all real eigenvalues.*

[†] A negative diagonal entry represents the combination of a reflection and stretch. Complex entries correspond to a complex stretching transformation. See Section 7.2 for details.

Example 8.19. The 3×3 matrix $A = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ considered in Example 8.5 has eigenvector basis

$$\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

We assemble these to form the eigenvector matrix

$$S = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \qquad \text{and so} \qquad S^{-1} = \begin{pmatrix} -1 & 0 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The diagonalization equation (8.28) becomes

$$S^{-1} A\, S = \begin{pmatrix} -1 & 0 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \Lambda,$$

with the eigenvalues of A appearing on the diagonal of Λ, in the same order as the eigenvectors.

Remark: If a matrix is not complete, then it cannot be diagonalized. Incomplete matrices represent generalized shearing transformations, and will be the subject of the following subsection. A simple example is a matrix of the form $\begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$ for $c \neq 0$, which represents a shear in the direction of the x axis.

8.4. Eigenvalues of Symmetric Matrices.

Fortunately, the matrices that arise in most applications are complete and, in fact, possess some additional structure that ameliorates the calculation of their eigenvalues and eigenvectors. The most important class are the symmetric, including positive definite, matrices. In fact, not only are the eigenvalues of a symmetric matrix necessarily real, the eigenvectors always form an *orthogonal basis* of the underlying Euclidean space. In such situations, we can tap into the dramatic power of orthogonal bases that we developed in Chapter 5. In fact, this is by far the most common way for orthogonal bases to appear — as the eigenvector bases of symmetric matrices.

Theorem 8.20. *If $A = A^T$ be a real symmetric $n \times n$ matrix, Then*

(a) *All the eigenvalues of A are real.*

(b) *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

(c) *There is an orthonormal basis of $\mathbb{R}^n$ consisting of n eigenvectors of A.*

In particular, all symmetric matrices are complete.

Remark: Orthogonality is with respect to the standard dot product on $\mathbb{R}^n$. As we noted in Section 7.5, the transpose or adjoint operation is intimately connected with the dot product. A corresponding result holds for self-adjoint linear transformations on general inner product spaces; see Exercise ■ for details.

Example 8.21. The 2×2 matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ considered in Example 8.5 is symmetric, and so has real eigenvalues $\lambda_1 = 4$ and $\lambda_2 = 2$. You can easily check that the corresponding eigenvectors $\mathbf{v}_1 = (\,1 \quad 1\,)^T$ and $\mathbf{v}_2 = (\,-1 \quad 1\,)^T$ are orthogonal: $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, and hence form on orthogonal basis of $\mathbb{R}^2$. An orthonormal basis is provided by the unit eigenvectors

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \tag{8.30}$$

obtained by dividing each eigenvector by its length: $\mathbf{u}_k = \mathbf{v}_k / \| \mathbf{v}_k \|$.

Proof of Theorem 8.20: First recall that if $A = A^T$ is real, symmetric, then

$$(A\,\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\,\mathbf{w}) \qquad \text{for all} \qquad \mathbf{v}, \mathbf{w} \in \mathbb{C}^n, \tag{8.31}$$

where we use the Euclidean dot product for real vectors and, more generally, the Hermitian dot product $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \overline{\mathbf{w}}$ when they are complex. (See Exercise ■.)

To prove property (a), suppose λ is a complex eigenvalue with complex eigenvector $\mathbf{v} \in \mathbb{C}^n$. Consider the Hermitian dot product of the complex vectors $A\,\mathbf{v}$ and $\mathbf{v}$:

$$(A\,\mathbf{v}) \cdot \mathbf{v} = (\lambda\,\mathbf{v}) \cdot \mathbf{v} = \lambda \,\| \mathbf{v} \|^2.$$

On the other hand, by (8.31),

$$(A\,\mathbf{v}) \cdot \mathbf{v} = \mathbf{v} \cdot (A\,\mathbf{v}) = \mathbf{v} \cdot (\lambda\,\mathbf{v}) = \mathbf{v}^T\, \overline{\lambda\,\mathbf{v}} = \overline{\lambda} \,\| \mathbf{v} \|^2.$$

Equating these two expressions, we deduce

$$\overline{\lambda} \,\| \mathbf{v} \|^2 = \lambda \,\| \mathbf{v} \|^2.$$

Since $\mathbf{v}$ is an eigenvector, it is nonzero, $\mathbf{v} \neq \mathbf{0}$, and so $\overline{\lambda} = \lambda$. This proves that the eigenvalue λ is real.

To prove (b), suppose

$$A\,\mathbf{v} = \lambda\,\mathbf{v}, \qquad A\,\mathbf{w} = \mu\,\mathbf{w},$$

where $\lambda \neq \mu$ are distinct real eigenvalues. Then, again by (8.31),

$$\lambda\,\mathbf{v} \cdot \mathbf{w} = (A\,\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\,\mathbf{w}) = \mathbf{v} \cdot (\mu\,\mathbf{w}) = \mu\,\mathbf{v} \cdot \mathbf{w},$$

and hence

$$(\lambda - \mu)\,\mathbf{v} \cdot \mathbf{w} = 0.$$

Since $\lambda \neq \mu$, this implies that $\mathbf{v} \cdot \mathbf{w} = 0$ and hence the eigenvectors $\mathbf{v}, \mathbf{w}$ are orthogonal.

Finally, the proof of (c) is easy if all the eigenvalues of A are distinct. Theorem 8.14 implies that the eigenvectors form a basis of $\mathbb{R}^n$, and part (b) proves they are orthogonal. (An alternative proof starts with orthogonality, and then applies Proposition 5.4 to prove that the eigenvectors form a basis.) To obtain an orthonormal basis, we merely divide the eigenvectors by their lengths: $\mathbf{u}_k = \mathbf{v}_k / \| \mathbf{v}_k \|$, as in Lemma 5.2.

To prove (c) in general, we proceed by induction on the size n of the matrix A. The case of a 1×1 matrix is trivial. (Why?) Let A have size $n \times n$. We know that A has at least one eigenvalue, λ_1, which is necessarily real. Let $\mathbf{v}_1$ be the associated eigenvector. Let

$$V^\perp = \{ \, \mathbf{w} \in \mathbb{R}^n \mid \mathbf{v}_1 \cdot \mathbf{w} = 0 \, \}$$

denote the orthogonal complement to the eigenspace V_{λ_1} — the set of all vectors orthogonal to the first eigenvector. Proposition 5.46 implies that $\dim V^\perp = n - 1$, and so we may identify $V^\perp \simeq \mathbb{R}^{n-1}$ by introducing an orthonormal basis. Moreover, $A\, V^\perp \subset V^\perp$, and the restriction of A to $V^\perp$ remains symmetric. Thus, we can use our induction hypothesis to construct an orthonormal basis of $V^\perp$ consisting of eigenvectors $\mathbf{u}_2, \dots, \mathbf{u}_n \in V^\perp$. Throwing in $\mathbf{u}_1 = \mathbf{v}_1 / \| \mathbf{v}_1 \|$ completes the orthonormal basis of $\mathbb{R}^n$. and thus completes the proof. *Q.E.D.*

Example 8.22. Consider the symmetric matrix $A = \begin{pmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \\ 2 & 2 & -1 \end{pmatrix}$. A straightforward computation produces its eigenvalues and eigenvectors:

$$\lambda_1 = 9, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \qquad \lambda_2 = 3, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \lambda_3 = -3, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}.$$

As the reader can check, the eigenvectors form an orthogonal basis of $\mathbb{R}^3$. The orthonormal eigenvector basis promised by Theorem 8.20 is obtained by dividing each eigenvector by its norm:

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \end{pmatrix}.$$

Finally, we can characterize positive definite matrices by their eigenvalues.

Theorem 8.23. *A symmetric matrix $K = K^T$ is positive definite if and only if all of its eigenvalues are strictly positive.*

Proof: First, if $K > 0$, then, by definition, $\mathbf{x}^T K \mathbf{x} > 0$ for all nonzero vectors $\mathbf{x} \in \mathbb{R}^n$. In particular, if $\mathbf{x} = \mathbf{v}$ is an eigenvector with (necessarily real) eigenvalue λ, then

$$0 < \mathbf{v}^T K \mathbf{v} = \mathbf{v}^T (\lambda \mathbf{v}) = \lambda \, \| \mathbf{v} \|^2, \tag{8.32}$$

which immediately proves that $\lambda > 0$. Conversely, suppose K has all positive eigenvalues. Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the orthonormal eigenvector basis of $\mathbb{R}^n$ guaranteed by Theorem 8.20, with $K\,\mathbf{u}_j = \lambda_j\,\mathbf{u}_j$. Then, writing

$$\mathbf{x} = c_1\,\mathbf{u}_1 + \,\cdots\, + c_n\,\mathbf{u}_n, \qquad \text{we have} \qquad K\,\mathbf{x} = c_1\,\lambda_1\,\mathbf{u}_1 + \,\cdots\, + c_n\,\lambda_n\,\mathbf{u}_n.$$

Therefore, using the orthonormality of the eigenvectors,

$$\mathbf{x}^T K\,\mathbf{x} = (c_1\,\mathbf{u}_1 + \,\cdots\, + c_n\,\mathbf{u}_n)\cdot(c_1\,\lambda_1\,\mathbf{u}_1 + \,\cdots\, + c_n\,\lambda_n\,\mathbf{u}_n) = \lambda_1\,c_1^2 + \,\cdots\, + \lambda_n\,c_n^2 \geq 0,$$

Moreover, the result is strictly positive for $\mathbf{x} \neq \mathbf{0}$ since not all the coefficients $c_1, \dots, c_n$ can be zero. This proves that that K is positive definite. *Q.E.D.*

Remark: The same proof shows that K is positive semi-definite if and only if all its eigenvalues satisfy $\lambda \geq 0$. A positive semi-definite matrix that is not positive definite admits a zero eigenvalue and one or more *null eigenvectors*, i.e., solutions to $K\,\mathbf{v} = \mathbf{0}$. Every nonzero element $\mathbf{0} \neq \mathbf{v} \in \ker K$ of the kernel is a null eigenvector.

Example 8.24. Consider the symmetric matrix $K = \begin{pmatrix} 8 & 0 & 1 \\ 0 & 8 & 1 \\ 1 & 1 & 7 \end{pmatrix}$. Its characteristic equation is

$$\det(K - \lambda\,I) = -\lambda^3 + 23\,\lambda^2 - 174\,\lambda + 432 = -(\lambda - 9)(\lambda - 8)(\lambda - 6),$$

and so its eigenvalues are $9, 8$ and 6. Since they are all positive, we conclude that K is a positive definite matrix. The associated eigenvectors are

$$\lambda_1 = 9, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \lambda_2 = 8, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \lambda_3 = 6, \qquad \mathbf{v}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

Note that the eigenvectors form an orthogonal basis of $\mathbb{R}^3$, as guaranteed by Theorem 8.20. We can construct an orthonormal eigenvector basis

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{pmatrix},$$

by dividing each eigenvector by its norm.

The Spectral Theorem

Since a real, symmetric matrix admits an eigenvector basis, it is diagonalizable. Moreover, since we can arrange that the eigenvectors form an orthonormal basis, the diagonalizing matrix takes a particularly simple form. Recall that an $n \times n$ matrix Q is called *orthogonal* if and only if its columns form an orthonormal basis of $\mathbb{R}^n$. Alternatively, one characterizes orthogonal matrices by the condition $Q^{-1} = Q^T$, as per Definition 5.18.

Therefore, when we use the orthonormal eigenvector basis in the diaognalization formula (8.28), the result is the important *Spectral Theorem* that governs the diagonalization of symmetric matrices.

Theorem 8.25. *Let A be a real, symmetric matrix. Then there exists an orthogonal matrix Q such that*

$$A = Q\,\Lambda\,Q^{-1} = Q\,\Lambda\,Q^T, \tag{8.33}$$

where Λ is a real diagonal matrix. The eigenvalues of A appear on the diagonal of Λ, while the eigenvectors are the corresponding columns of Q.

Remark: The term "spectrum" refers to the eigenvalues of a matrix or, more generally, a linear operator. The terminology is motivated by physics. The spectral energy lines of atoms, molecules and nuclei are characterized as the eigenvalues of the governing quantum mechanical linear operators, [**100**, **104**].

Warning: The spectral factorization $A = Q\,\Lambda\,Q^T$ and the Gaussian factorization $A = L\,D\,L^T$ of a regular symmetric matrix, cf. (1.52), are completely different. In particular, the eigenvalues are *not* the pivots: $\Lambda \neq D$.

The *spectral decomposition* (8.33) provides us with an alternative means of diagonalizing the associated quadratic form $q(\mathbf{x}) = \mathbf{x}^T A\,\mathbf{x}$, i.e., of completing the square. We write

$$q(\mathbf{x}) = \mathbf{x}^T A\,\mathbf{x} = \mathbf{x}^T\,Q\,\Lambda\,Q^T\,\mathbf{x} = \mathbf{y}^T \Lambda\,\mathbf{y} = \sum_{i=1}^{n} \lambda_i\,y_i^2, \tag{8.34}$$

where $\mathbf{y} = Q^T\mathbf{x} = Q^{-1}\mathbf{x}$ are the coordinates of $\mathbf{x}$ with respect to the orthonormal eigenvalue basis of A, cf. (7.21). In particular, $q(\mathbf{x}) > 0$ for all nonzero $\mathbf{x}$ — which means A is positive definite — if and only if each eigenvalue $\lambda_i > 0$ is strictly positive, reconfirming Theorem 8.23.

Example 8.26. For the 2×2 matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ considered in Example 8.21, the orthonormal eigenvectors (8.30) produce the diagonalizing orthogonal rotation matrix $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$. The reader can check the spectral factorization

$$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} = A = Q\,\Lambda\,Q^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

According to (8.34), the associated quadratic form is diagonalized as

$$q(\mathbf{x}) = 3\,x_1^2 + 2\,x_1\,x_2 + 3\,x_2^2 = 4\,y_1^2 + 2\,y_2^2,$$

where $\mathbf{y} = Q^T\mathbf{x}$, i.e., $y_1 = \dfrac{x_1 + x_2}{\sqrt{2}}$, $y_2 = \dfrac{-x_1 + x_2}{\sqrt{2}}$.

We note that you can choose Q to be a proper orthogonal matrix, so $\det Q = 1$, since an improper orthogonal matrix can be made proper by multiplying one of its columns by -1, which does not affect its status as an eigenvector matrix. Since a proper orthogonal matrix Q represents a rigid rotation of $\mathbb{R}^n$, the diagonalization of a symmetric matrix can be interpreted as a rotation of the coordinate system in which the orthogonal eigenvectors line

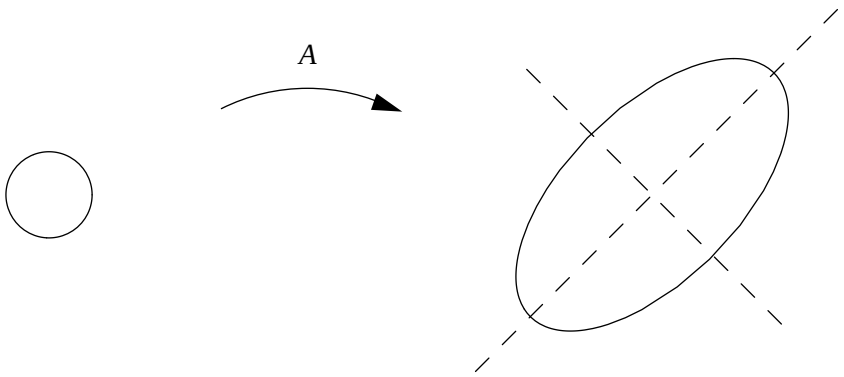


Figure 8.2. Stretching a Circle into an Ellipse.

up along the coordinate axes. Therefore, a linear transformation $L(\mathbf{x}) = A\,\mathbf{x}$ represented by a positive definite matrix $A > 0$ can be regarded as a combination of stretches along a mutually orthogonal set of directions. In elasticity, the stress tensor of a deformed body is represented by a positive definite matrix. Its eigenvalues are known as the *principal stretches* and its eigenvectors the principal directions of stretch of the elastic deformation.

A good way to visualize this is to consider the effect of the linear transformation on the unit (Euclidean) sphere

$$S_1 = \{\, \| \mathbf{x} \| = 1 \,\}.$$

Stretching the sphere in orthogonal directions will map it into an ellipsoid $E = L[\,S_1\,]$ whose axes are aligned with the directions of stretch. Explicitly, if the linear transformation is giveny by $\mathbf{y} = A\,\mathbf{x}$, then

$$E = L[S_1] = \{\, A\,\mathbf{x} \,|\, \| \mathbf{x} \| = 1 \,\} = \left\{\, \mathbf{y} \;\middle|\; \| A^{-1}\mathbf{y} \| = 1 \,\right\}. \tag{8.35}$$

For example, the matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ considered in the preceding example represents the linear transformation

$$\widetilde{x} = 3\,x + y, \qquad \widetilde{y} = x + 3\,y.$$

Therefore, the unit circle $x^2 + y^2 = 1$ will be mapped to the ellipse

$$\left(\frac{3\,\widetilde{x} - \widetilde{y}}{8} \right)^2 + \left(\frac{-\,\widetilde{x} + 3\,\widetilde{y}}{8} \right)^2 = \tfrac{5}{32}\,\widetilde{x}^2 - \tfrac{3}{16}\,\widetilde{x}\,\widetilde{y} + \tfrac{5}{32}\,\widetilde{y}^2 = 1,$$

whose principal axes line up with the eigenvectors $\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$, $\mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$; see Figure 8.2. The eigenvalues, $4, 2$, prescribe the ellipse's *semi-axes*.

Optimization Principles for Eigenvalues

As we learned in Chapter 4, the solution to a linear system with positive definite coefficient matrix can be characterized by a minimization principle. Thus, it should come as no surprise that eigenvalues of positive definite, and even more general symmetric matrices,

can also be characterized by some sort of optimization procedure. A number of basic numerical algorithms for computing eigenvalues, of both matrices and, later on, differential operators are based on such optimization principles.

First consider the relatively simple case of a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. We assume that the diagonal entries, which are the *same* as the eigenvalues, appear in decreasing order,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \tag{8.36}$$

so λ_1 is the largest eigenvalue, while λ_n is the smallest. The effect of Λ on a vector $\mathbf{y} \in \mathbb{R}^n$ is to multiply its entries by the diagonal eigenvalues: $\Lambda\,\mathbf{y} = (\,\lambda_1\,y_1, \lambda_2\,y_2, \dots, \lambda_n\,y_n\,)^T$. In other words, the linear transformation represented by the coefficient matrix Λ has the effect of stretching[†] in the i^{th} coordinate direction by the factor λ_i. In particular, the maximal stretch occurs in the first direction, with factor λ_1, while the minimal stretch occurs in the last direction, with factor λ_n. The germ of the optimization principles for characterizing the extreme eigenvalues is contained in this geometrical observation.

Let us turn our attention to the associated quadratic form

$$q(\mathbf{y}) = \mathbf{y}^T \Lambda\,\mathbf{y} = \lambda_1\,y_1^2 + \lambda_2\,y_2^2 + \cdots + \lambda_n\,y_n^2. \tag{8.37}$$

Note that $q(t\,\mathbf{e}_1) = \lambda_1\,t^2$, and hence if $\lambda_1 > 0$, then $q(\mathbf{y})$ has no maximum; On the other hand, if $0 \geq \lambda_i$ for all i, then $q(\mathbf{y}) \leq 0$, and its maximal value is $q(\mathbf{0}) = 0$. Thus, in either case, a strict maximization of $q(\mathbf{y})$ is not of much help.

Suppose, however, that we try to maximize $q(\mathbf{y})$ when $\mathbf{y}$ is restricted to be a unit vector (in the Euclidean norm):

$$\| \mathbf{y} \|^2 = y_1^2 + \cdots + y_n^2 = 1.$$

In view of our ordering convention (8.36) and the positivity of each y_i^2,

$$q(\mathbf{y}) = \lambda_1\,y_1^2 + \lambda_2\,y_2^2 + \cdots + \lambda_n\,y_n^2 \leq \lambda_1\,y_1^2 + \lambda_1\,y_2^2 + \cdots + \lambda_1\,y_n^2 = \lambda_1\big(\,y_1^2 + \cdots + y_n^2\,\big) = \lambda_1.$$

Moreover, $q(\mathbf{e}_1) = \lambda_1$, and therefore the maximal value of $q(\mathbf{y})$ over all unit vectors *is* the largest eigenvalue of Λ:

$$\lambda_1 = \max\{\, q(\mathbf{y}) \mid \| \mathbf{y} \| = 1 \,\}.$$

By the same reasoning, $q(\mathbf{y})$ also has a minimal value

$$\lambda_n = \min\{\, q(\mathbf{y}) \mid \| \mathbf{y} \| = 1 \,\}$$

equal to the smallest eigenvalue. Thus, we can represent the two extreme eigenvalues by optimization principles, albeit of a slightly different character than we treated in Chapter 4.

Now suppose A is any symmetric matrix. We use the spectral decomposition (8.33) to diagonalize the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T A\,\mathbf{x} = \mathbf{x}^T\,Q\,\Lambda\,Q^T\,\mathbf{x} = \mathbf{y}^T \Lambda\,\mathbf{y}, \qquad \text{where} \qquad \mathbf{y} = Q^T\mathbf{x} = Q^{-1}\mathbf{x},$$

[†] If $\lambda_i < 0$, then the effect is to stretch and reflect.

as in (8.34). According to the preceding discussion, the minimum of $\mathbf{y}^T \Lambda \mathbf{y}$ over all unit vectors $\| \mathbf{y} \| = 1$ is the smallest eigenvalue λ_1 of Λ, which is the *same* as the smallest eigenvalue of A. Moreover, since Q is an orthogonal matrix, Proposition 7.23 tell us that it maps unit vectors to unit vectors:

$$1 = \| \mathbf{y} \| = \| Q^T \mathbf{x} \| = \| \mathbf{x} \|.$$

Thus, minimizing $q(\mathbf{x})$ over all unit vectors $\mathbf{y} = Q^T \mathbf{x}$ is the same as minimizing over all unit vectors $\mathbf{x}$. Similar reasoning applies to the smallest eigenvalue λ_n. In this fashion, we have established the basic optimization principles for the largest and smallest eigenvalues of a symmetric matrix.

Theorem 8.27. *If A is a symmetric matrix, then*

$$\lambda_1 = \max \left\{ \, \mathbf{x}^T A \, \mathbf{x} \;\big|\; \| \mathbf{x} \| = 1 \, \right\}, \qquad \lambda_n = \min \left\{ \, \mathbf{x}^T A \, \mathbf{x} \;\big|\; \| \mathbf{x} \| = 1 \, \right\}, \tag{8.38}$$

are, respectively its largest and smallest eigenvalues.

The maximal value is achieved when we set $\mathbf{x} = \pm \mathbf{u}_1$ to be a unit eigenvector corresponding to the largest eigenvalue; similarly, the minimal value is at $\mathbf{x} = \pm \mathbf{u}_n$.

Remark: In multivariable calculus, the eigenvalue λ plays the role of a "Lagrange multiplier" for the constrained optimization problem, cf. [**9**].

Example 8.28. The problem is to maximize the value of the quadratic form

$$q(x, y) = 3\,x^2 + 2\,x\,y + 3\,y^2$$

for all x, y lying on the unit circle $x^2 + y^2 = 1$. This maximization problem is precisely of form (8.38). The coefficient matrix is $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$, whose eigenvalues are, according to Example 8.5, $\lambda_1 = 4$ and $\lambda_2 = 2$. Theorem 8.27 implies that the maximal value for the quadratic form on the unit circle is the largest eigenvalue, and hence equal to 4, while its minimal value is the smallest eigenvalue, and hence equal to 2. Indeed, if we evaluate $q(x, y)$ on the unit eigenvectors, we obtain $q\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) = 2$, $q\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) = 4$, while $2 \leq q(x, y) \leq 4$ for all x, y such that $x^2 + y^2 = 1$.

In practical applications, the restriction of the quadratic form to unit vectors may not be particularly convenient. One can, however, easily rephrase the eigenvalue optimization principles in a form that utilizes general vectors. If $\mathbf{v} \neq \mathbf{0}$ is any nonzero vector, then $\mathbf{x} = \mathbf{v}/\| \mathbf{v} \|$ is a unit vector. Substituting this expression for $\mathbf{x}$ in the quadratic form $q(\mathbf{x}) = \mathbf{x}^T A \, \mathbf{x}$ leads to the following optimization principles for the extreme eigenvalues of a symmetric matrix:

$$\lambda_1 = \max \left\{ \, \frac{\mathbf{v}^T A \, \mathbf{v}}{\| \mathbf{v} \|^2} \;\Big|\; \mathbf{v} \neq \mathbf{0} \, \right\}, \qquad \lambda_n = \min \left\{ \, \frac{\mathbf{v}^T A \, \mathbf{v}}{\| \mathbf{v} \|^2} \;\Big|\; \mathbf{v} \neq \mathbf{0} \, \right\}. \tag{8.39}$$

Thus, we replace optimization of a quadratic polynomial over the unit sphere by optimization of a rational function over all of $\mathbb{R}^n \setminus \{\mathbf{0}\}$. Referring back to Example 8.28, the

maximum value of

$$r(x,y) = \frac{3x^2 + 2xy + 3y^2}{x^2 + y^2} \qquad \text{for all} \qquad \begin{pmatrix} x \\ y \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

is equal to 4, the same maximal eigenvalue of the corresponding coefficient matrix.

What about characterizing one of the intermediate eigenvalues? Then we need to be a little more sophisticated in designing the optimization principle. To motivate the construction, look first at the diagonal case. If we restrict the quadratic form (8.37) to vectors $\widetilde{\mathbf{y}} = (\,0, y_2, \ldots, y_n\,)^T$ whose first component is zero, we obtain

$$q(\widetilde{\mathbf{y}}) = q(0, y_2, \ldots, y_n) = \lambda_2\, y_2^2 + \,\cdots\, + \lambda_n\, y_n^2.$$

The maximum value of $q(\widetilde{\mathbf{y}})$ over all such $\widetilde{\mathbf{y}}$ of norm 1 is, by the same reasoning, the second largest eigenvalue λ_2. Moreover, we can characterize such vectors geometrically by noting that they are orthogonal to the first standard basis vector, $\widetilde{\mathbf{y}} \cdot \mathbf{e}_1 = 0$, which also happens to be the eigenvector of Λ corresponding to the eigenvalue λ_1. Similarly, if we want to find the j^{th} largest eigenvalue λ_j, we maximize $q(\mathbf{y})$ over all unit vectors $\widehat{\mathbf{y}}$ whose first $j-1$ components vanish, $y_1 = \cdots = y_{j-1} = 0$, or, stated geometrically, over all $\widetilde{\mathbf{y}}$ such that $\| \widehat{\mathbf{y}} \| = 1$ and $\widehat{\mathbf{y}} \cdot \mathbf{e}_1 = \cdots = \widehat{\mathbf{y}} \cdot \mathbf{e}_{j-1} = 0$, i.e., over all vectors orthogonal to the first $j-1$ eigenvectors of Λ.

A similar reasoning based on the Spectral Theorem 8.25 and the orthogonality of eigenvectors of symmetric matrices, leads to the general result.

Theorem 8.29. *Let A be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding orthogonal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then the maximal value of the quadratic form $\mathbf{x}^T A\,\mathbf{x}$ over all unit vectors which are orthogonal to the first $j-1$ eigenvectors is the j^{th} eigenvalue:*

$$\lambda_j = \max \left\{ \, \mathbf{x}^T A\,\mathbf{x} \;\big|\; \| \mathbf{x} \| = 1, \;\; \mathbf{x} \cdot \mathbf{v}_1 = \cdots = \mathbf{x} \cdot \mathbf{v}_{j-1} = 0 \, \right\}. \tag{8.40}$$

Thus, at least in principle, one can compute the eigenvalues and eigenvectors recursively. First, find the largest eigenvalue λ_1 by the basic maximization principle (8.38) and its associated eigenvector $\mathbf{v}_1$ by solving the eigenvector system (8.13). The next largest eigenvalue λ_2 is then characterized by the constrained minimization principle (8.40), and so on. However, this is not a very practical algorithm for numerical computations.

8.5. Singular Values.

We have already indicated the centrality of the eigenvalues and eigenvectors of a square matrix for both theory and applications. Much more evidence to this effect will appear in the following chapters. Alas, rectangular matrices have no eigenvalues (why?), and so, at first glance, do not appear to possess any quantities of comparable significance. However, our earlier treatment of both least squares minimization problems as well as the equilibrium equations for structures and circuits made essential use of the symmetric, positive semi-definite *square* Gram matrix $K = A^T A$ — which can be naturally formed even when A is non-square. We anticipate that the eigenvalues of K might play a comparably important role. Since they are not easily related to the eigenvalues of A — which, in the truly rectangular case, don't even exist — we shall endow them with a new name.

Definition 8.30. The *singular values* $\sigma_1, \dots, \sigma_n$ of an $m \times n$ matrix A are the square roots, $\sigma_i = \sqrt{\lambda_i}$, of the eigenvalues of the Gram matrix $K = A^T A$. The corresponding eigenvectors of K are known as the *singular vectors* of A.

Since K is positive semi-definite, its eigenvalues are always non-negative, $\lambda_i \geq 0$, and hence the singular values of A are also all non-negative[†], $\sigma_i \geq 0$ — no matter whether A itself has positive, negative, or even complex eigenvalues, or is rectangular and has no eigenvalues at all. However, for symmetric matrices, there is a direct connection between the two quantities:

Proposition 8.31. *If $A = A^T$ is a symmetric matrix, its singular values are the absolute values of its eigenvalues: $\sigma_i = |\, \lambda_i \,|$; its singular vectors coincide with the associated eigenvectors.*

Proof: When A is symmetric, $K = A^T A = A^2$. So, if $A\mathbf{v} = \lambda \mathbf{v}$, then $K\mathbf{v} = A^2 \mathbf{v} = \lambda^2 \mathbf{v}$. Thus, every eigenvector $\mathbf{v}$ of A is also an eigenvector of K with eigenvalue λ^2. Therefore, the eigenvector basis of A is an eigenvector basis for K, and hence forms a complete system of singular vectors for A also. *Q.E.D.*

The standard convention is to label the singular values in *decreasing* order, so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. Thus, σ_1 will always denote the largest or *dominant* singular value. If $A^T A$ has repeated eigenvalues, the singular values of A are repeated with the same multiplicities.

Example 8.32. Let $A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}$. The associated Gram matrix is $K = A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$, with eigenvalues $\lambda_1 = 40$ and $\lambda_2 = 10$. Thus, the singular values of A are $\sigma_1 = \sqrt{40} \approx 6.3246\ldots$ and $\sigma_2 = \sqrt{10} \approx 3.1623\ldots$. Note that these are not the same as the eigenvalues of A, namely $\lambda_1 = \frac{1}{2}(3 + \sqrt{89}) \approx 6.2170\ldots$, $\lambda_2 = \frac{1}{2}(3 - \sqrt{89}) \approx -3.2170$.

A rectangular matrix Σ will be called *diagonal* if its only nonzero entries are on the main diagonal starting in the upper left hand corner, and so $\sigma_{ij} = 0$ for $i \neq j$. An example is the matrix

$$\Sigma = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

whose only nonzero entries are in the diagonal $(1,1)$ and $(2,2)$ positions. (Its last diagonal entry happens to be 0.)

The generalization of the spectral factorization to non-symmetric matrices is known as the *singular value decomposition*, commonly abbreviated as SVD. Unlike the spectral decomposition, the singular value decomposition applies to arbitrary real rectangular matrices.

[†] *Warning*: Some authors, [**121**], only designate the nonzero σ_i's as singular values.

Theorem 8.33. *Any real $m \times n$ matrix A can be factorized*

$$A = P\,\Sigma\,Q^T \tag{8.41}$$

into the product of an $m \times m$ orthogonal matrix P, the $m \times n$ diagonal matrix Σ that has the first $l = \min\{m, n\}$ singular values of A as its diagonal entries, and an $n \times n$ orthogonal matrix Q^T.

Proof: Writing the factorization (8.41) as $AQ = P\,\Sigma$, and looking at the columns of the resulting matrix equation, we find the systems

$$A\,\mathbf{u}_i = \sigma_i\,\mathbf{v}_i, \qquad i = 1, \dots, n, \tag{8.42}$$

relating the orthonormal columns of $Q = (\,\mathbf{u}_1\ \mathbf{u}_2\ \dots\ \mathbf{u}_n\,)$ to the orthonormal columns of $P = (\,\mathbf{v}_1\ \mathbf{v}_2\ \dots\ \mathbf{v}_m\,)$. The scalars σ_i in (8.42) are the diagonal entries of Σ or, if $m < i \leq n$, equal to 0. The fact that P and Q are both orthogonal matrices means that their column vectors form orthonormal bases for, respectively, $\mathbb{R}^m$ and $\mathbb{R}^n$ under the Euclidean dot product. In this manner, the singular values indicate how far the linear transformation represented by the matrix A stretches a distinguished set of orthonormal basis vectors.

To construct the required bases, we prescribe $\mathbf{u}_1, \dots, \mathbf{u}_n$ to be the orthonormal eigenvector basis of the Gram matrix $K = A^T A$; thus

$$A^T A\,\mathbf{u}_j = K\,\mathbf{u}_j = \lambda_j\,\mathbf{u}_j = \sigma_j^2\,\mathbf{u}_j.$$

We claim that the image vectors $\mathbf{w}_i = A\,\mathbf{u}_i$ are automatically orthogonal. Indeed, in view of the orthonormality of the $\mathbf{u}_i$,

$$\mathbf{w}_i \cdot \mathbf{w}_j = \mathbf{w}_i^T \mathbf{w}_j = (A\,\mathbf{u}_i)^T A\,\mathbf{u}_j = \mathbf{u}_i^T A^T A\,\mathbf{u}_j = \lambda_j\,\mathbf{u}_i^T\,\mathbf{u}_j = \lambda_j\,\mathbf{u}_i \cdot \mathbf{u}_j = \begin{cases} 0, & i \neq j, \\ \sigma_i^2, & i = j. \end{cases} \tag{8.43}$$

Consequently, $\mathbf{w}_1, \dots, \mathbf{w}_n$ form an orthogonal system of vectors of respective norms

$$\| \mathbf{w}_i \| = \sqrt{\mathbf{w}_i \cdot \mathbf{w}_i\ } = \sigma_i.$$

Since $\mathbf{u}_1, \dots, \mathbf{u}_n$ form a basis of $\mathbb{R}^n$, their images $\mathbf{w}_1 = A\,\mathbf{u}_1, \dots, \mathbf{w}_n = A\,\mathbf{u}_n$ span $\operatorname{rng} A$. Suppose that A has r non-zero singular values, so $\sigma_{r+1} = \cdots = \sigma_n = 0$. Then the corresponding image vectors $\mathbf{w}_1, \dots, \mathbf{w}_r$ are non-zero, mutually orthogonal vectors, and hence form an orthogonal basis for $\operatorname{rng} A$. Since the dimension of $\operatorname{rng} A$ is equal to its rank, this implies that the number of non-zero singular values is $r = \operatorname{rank} A$. The corresponding unit vectors

$$\mathbf{v}_i = \frac{\mathbf{w}_i}{\sigma_i} = \frac{A\,\mathbf{u}_i}{\sigma_i}\,, \qquad i = 1, \dots, r, \tag{8.44}$$

are an orthonormal basis for $\operatorname{rng} A$. Let us further select an orthonormal basis $\mathbf{v}_{r+1}, \dots, \mathbf{v}_m$ for its orthogonal complement $\operatorname{coker} A = (\operatorname{rng} A)^\perp$. The combined set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_m$ clearly forms an orthonormal basis of $\mathbb{R}^m$, and satisfies (8.42). In this manner, the resulting orthonormal bases $\mathbf{u}_1, \dots, \mathbf{u}_n$ and $\mathbf{v}_1, \dots, \mathbf{v}_m$ form the respective columns of the orthogonal matrices Q, P in the singular value decomposition (8.41). *Q.E.D.*

Warning: If $m < n$, then only the first m singular values appear along the diagonal of Σ. It follows from the proof that the remaining $n - m$ singular values are all zero.

Example 8.34. For the matrix $A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}$ considered in Example 8.32, the orthonormal eigenvector basis of $K = A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$ is given by $\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ and $\mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$. Thus, $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$. On the other hand, according to (8.44),

$$\mathbf{v}_1 = \frac{A\,\mathbf{u}_1}{\sigma_1} = \frac{1}{\sqrt{40}} \begin{pmatrix} 4\sqrt{2} \\ 2\sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}, \qquad \mathbf{v}_2 = \frac{A\,\mathbf{u}_2}{\sigma_2} = \frac{1}{\sqrt{10}} \begin{pmatrix} \sqrt{2} \\ -2\sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix},$$

and thus $P = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}$. You may wish to validate the resulting singular value decomposition

$$A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = P\,\Sigma\,Q^T.$$

As their name suggests, the singular values can be used to detect singular matrices. Indeed, the singular value decomposition tells us some interesting new geometrical information about the action of the matrix, filling in further details in the discussion begun in Section 2.5 and continued in Section 5.6. The next result follows directly from the proof of Theorem 8.33.

Theorem 8.35. *Let $\sigma_1, \dots, \sigma_r > 0$ be the non-zero singular values of the $m \times n$ matrix A. Let $\mathbf{v}_1, \dots, \mathbf{v}_m$ and $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the orthonormal bases of, respectively, $\mathbb{R}^m$ and $\mathbb{R}^n$ provided by the columns of P and Q in its singular value decomposition $A = P\,\Sigma\,Q^T$. Then*

(*i*) $r = \operatorname{rank} A$,

(*ii*) $\mathbf{u}_1, \dots, \mathbf{u}_r$ *form an orthonormal basis for* $\operatorname{corng} A$,

(*iii*) $\mathbf{u}_{r+1}, \dots, \mathbf{u}_n$ *form an orthonormal basis for* $\ker A$,

(*iv*) $\mathbf{v}_1, \dots, \mathbf{v}_r$ *form an orthonormal basis for* $\operatorname{rng} A$,

(*v*) $\mathbf{v}_{r+1}, \dots, \mathbf{v}_m$ *form an orthonormal basis for* $\operatorname{coker} A$.

We already noted in Section 5.6 that the linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^m$ defined by matrix multiplication by A can be interpreted as a projection from $\mathbb{R}^n$ to $\operatorname{corng} A$ followed by an invertible map from $\operatorname{corng} A$ to $\operatorname{rng} A$. The singular value decompostion tells us that not only is the latter map invertible, it is simply a combination of stretches in the r mutually orthogonal singular directions, whose magnitudes equal the nonzero singular values. In this way, we have at last reached a complete understanding of the subtle geometry underlying the simple operation of matrix multiplication.

An alternative useful interpretation is to view the two orthogonal matrices in (8.41) as defining rigid rotations or reflections. Therefore, in all cases, a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ is composed of three ingredients:

(i) A rotation/reflection of the domain space $\mathbb{R}^n$ prescribed by Q^T, followed by

(ii) a simple stretching map of the coordinate vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ of domain space, mapping $\mathbf{e}_i$ to $\sigma_i \mathbf{e}_i$ in the target space $\mathbb{R}^m$, followed by

(iii) a rotation/reflection of the target space prescribed by P.

In fact, in most cases we can choose both P and Q to be proper orthogonal matrices representing rotations; see Exercise ■.

Condition Number

The singular values not only provide a nice geometric interpretation of the action of the matrix, they also play a key role in modern computational algorithms. The relative magnitudes of the singular values can be used to distinguish well-behaved linear systems from ill-conditioned systems which are much trickier to solve accurately. A square matrix with a zero singular value is singular; a matrix with one or more very small singular values is considered to be close to singular, and hence ill-conditioned in the sense that it is hard to invert numerically. Such ill-conditioning is traditionally quantified as follows.

Definition 8.36. The *condition number* of an $n \times n$ matrix is the ratio between its largest and smallest singular value: $\kappa(A) = \sigma_1/\sigma_n$.

A matrix with a very large condition number is said to be *ill-conditioned*; in practice, this occurs when the condition number is larger than the reciprocal of the machine's precision, e.g., 10^6 for single precision arithmetic. As the name implies, it is much harder to solve a linear system $A\mathbf{x} = \mathbf{b}$ when its coefficient matrix is ill-conditioned. In the extreme case when A has one or more zero singular values, so $\sigma_n = 0$, its condition number is infinite, and the linear system is singular, with either no solution or infinitely many solutions.

The computation of the rank of a matrix is a significant numerical challenge. Making tiny numerical errors in the entries can have a dramatic effect on the rank; for example, the matrix $A = \begin{pmatrix} 1 & 1 & -1 \\ 2 & 2 & -2 \\ 3 & 3 & -3 \end{pmatrix}$ has rank $r = 1$, but a tiny change, say to $\widetilde{A} = \begin{pmatrix} 1.00001 & 1 & -1 \\ 2 & 2.00001 & -2 \\ 3 & 3 & -3.00001 \end{pmatrix}$, will produce a nonsingular matrix of rank 3. The latter matrix, however, is very close to singular, and this is highlighted by their respective singular values. For the first matrix, they are $\sigma_1 = \sqrt{42} \approx 6.48$, $\sigma_2 = \sigma_3 = 0$, reconfirming that A has rank 1, whereas for $\widetilde{A}$ we find $\sigma_1 \approx 6.48075, \sigma_2, \sigma_3 \approx .000001$. The fact that the second and third singular values are very small indicates that $\widetilde{A}$ is very close to a matrix of rank 1 and should be viewed as a numerical perturbation of such a matrix. thus, in practical terms, one assigns a threshhold for singular values, and treats any small sijngular value below the threshold as zero.

This idea underlies the method of *Principal Component Analysis* in modern statistics, data analysis, imaging and many other fields, [**87**]. The largest singular values and associated singular vectors indicate the principal components of the matrix, while small singular values indicate unimportant directions. In such applications, the columns of the matrix A represent the data vectors, normalized to have mean $\mathbf{0}$, or, equivalently, so that the row sums of A are all 0; see Exercise ∎. The corresponding Gram matrix $K = A^T A$ can be identified as the *covariance matrix* associated with the data. The principal components indicate the directions of correlations and clustering to be found in the data. Classification of patterns in images, sounds, sematics, and many other areas are being successfully analyzed by this approach.

The Pseudoinverse

With the singular value decomposition in hand, we are able to introduce a generalization of the inverse of a matrix that applies to cases when the matrix in question is singular or even rectangular. We begin with the diagonal case. Let Σ be an $m \times n$ diagonal matrix with r nonzero diagonal entries $\sigma_1, \dots, \sigma_r$. We define the *pseudoinverse* of Σ to be the $n \times m$ diagonal matrix Σ^+ whose nonzero diagonal entries are the reciprocals $1/\sigma_1, \dots, 1/\sigma_r$. For example, if

$$\Sigma = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \text{then} \qquad \Sigma^+ = \begin{pmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In particular, if Σ is a nonsingular square diagonal matrix, then its pseudoinverse and ordinary inverse are the same: $\Sigma^+ = \Sigma^{-1}$.

Definition 8.37. The *pseudoinverse* of an $m \times n$ matrix A with singular value decomposition $A = P\,\Sigma\,Q^T$ is the $n \times m$ matrix $A^+ = Q\,\Sigma^+\,P^T$.

Note that the latter equation is the singular value decomposition of the pseudoinverse A^+, and hence its nonzero singular values are the reciprocals of the nonzero singular values of A. If A is a non-singular square matrix, then its pseudoinverse agrees with its ordinary inverse, since

$$A^{-1} = (P\,\Sigma\,Q^T)^{-1} = Q^{-T}\,\Sigma^{-1}\,P^{-1} = Q\,\Sigma^+\,P^T = A^+,$$

where we used the fact that the inverse of an orthogonal matrix is equal to its transpose.

If A is square and nonsingular, then, as we know, the solution to the linear system $A\,\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x}^\star = A^{-1}\mathbf{b}$. For a general coefficient matrix, the vector $\mathbf{x}^\star = A^+\,\mathbf{b}$ obtained by applying the pseudoinverse to the right hand side plays a distinguished role — it is the *least squares soution* to the system! In this manner, the pseudoinverse provides us with a direct route to least squares solutions to systems of linear equations.

Theorem 8.38. *Consider the linear system $A\,\mathbf{x} = \mathbf{b}$. Let $\mathbf{x}^\star = A^+\,\mathbf{b}$, where A^+ is the pseudoinverse of A. If $\ker A = \{\mathbf{0}\}$, then $\mathbf{x}^\star$ is the least squares solution to the system. If, more generally, $\ker A \neq \{\mathbf{0}\}$, then $\mathbf{x}^\star$ is the least squares solution of minimal Euclidean norm among all vectors that minimize the least squares error $\| A\,\mathbf{x} - \mathbf{b} \|$.*

Proof: To show that $\mathbf{x}^\star = A^+\mathbf{b}$ is the least squares solution to the system, we must check that it satisfies the normal equations $A^TA\mathbf{x}^\star = A^T\mathbf{b}$. Using the definition of the pseudoinverse and the singular value decomposition (8.41), we find

$$\begin{aligned} A^TA\mathbf{x}^\star &= A^TAA^+\mathbf{b} = (P\Sigma Q^T)^T(P\Sigma Q^T)(Q\Sigma^+P^T)\mathbf{b} \\ &= Q\Sigma^T\Sigma\Sigma^+P^T\mathbf{b} = Q\Sigma^TP^T\mathbf{b} = A^T\mathbf{b}, \end{aligned}$$

where the next to last equality is left as Exercise ∎ for the reader. This proves that $\mathbf{x}^\star$ solves the normal equations, and hence minimizes the least squares error[†].

Thus, when $\operatorname{rank} A = n$, the vector $\mathbf{x}^\star$ is the unique least squares aolution to the system. More generally, if A has rank $r < n$, so, $\ker A \neq \{\mathbf{0}\}$, only the first r singular values are nonzero, and therefore the last $n-r$ rows of Σ^+ are all zero. This implies that the last $n-r$ entries of the vector $\mathbf{c} = \Sigma^+P^T\mathbf{b}$ are also all zero, so $\mathbf{c} = (\,c_1, \dots c_r, 0, \dots, 0\,)^T$. We conclude that

$$\mathbf{x}^\star = A^+\mathbf{b} = Q\,\Sigma^+\,P^T\mathbf{b} = Q\,\mathbf{c} = c_1\,\mathbf{u}_1 + \,\cdots\, + c_r\,\mathbf{u}_r$$

is a linear combination of the first r singular vectors, and hence, by Theorem 8.35, $\mathbf{x}^\star \in \operatorname{corng} A$. The most general least squares solution has the form $\mathbf{x} = \mathbf{x}^\star + \mathbf{z}$ where $\mathbf{z} \in \ker A$, and the fact that $\|\,\mathbf{x}\,\|^\star$ is minimized follows as in Theorem 5.55. *Q.E.D.*

When forming the pseudoinverse, we see see that very small singular values lead to very large entries in Σ^+, which will cause numerical difficulties when computing the least squares solution $\mathbf{x}^\star = A^+\mathbf{b}$ to the linear system. A common and effective computational strategy to avoid the effects of small singular values is to replace the corresponding diagonal entries of the pseudoinverse Σ^+ by 0. This has the effect of regularizing ill-conditioned matrices that are very close to singular — rather than solve the system directly for $\mathbf{x} = A^{-1}\mathbf{b}$, one tends to use the suitably regularized pseudoinverse.

Finally, we note that practical numerical algorithms for computing singular values and the singular value decomposition can be found in [**121**, **66**]

8.6. Incomplete Matrices and the Jordan Canonical Form.

Unfortunately, not all matrices are complete. Matrices that do not have an eigenvector basis are considerably less convenient to deal with. However, as they occasionally appear in applications, it is worth learning how to handle them. We shall show how to supplement the eigenvectors in order to obtain a basis in which the matrix assumes a simple, but now non-diagonal form. The resulting construction is named after the nineteenth century French mathematician Camille Jordan (and not Wilhelm Jordan of Gauss–Jordan fame).

Throughout this section, A will be an $n \times n$ matrix, with either real or complex entries. We let $\lambda_1, \dots, \lambda_k$ denote the distinct eigenvalues of A. We recall that Theorem 8.10 guarantees that every matrix has at least one (complex) eigenvalue, so $k \geq 1$. Moreover, we are assuming that $k < n$, as otherwise A would be complete.

[†] In Chapter 4, this was proved under the assumption that $\ker A = \{\mathbf{0}\}$. You are asked to establish the general case in Exercise ∎.

Definition 8.39. Let A be a square matrix. A *Jordan chain* of length j is a sequence of non-zero vectors $w_1, \dots, w_j \in \mathbb{C}^m$ that satisfies

$$A\,\mathbf{w}_1 = \lambda\,\mathbf{w}_1, \qquad A\,\mathbf{w}_i = \lambda\,\mathbf{w}_i + \mathbf{w}_{i-1}, \qquad i = 2, \dots, j, \tag{8.45}$$

where λ is an eigenvalue of A.

Note that the initial vector $\mathbf{w}_1$ in a Jordan chain is a genuine eigenvector. The others, $\mathbf{w}_2, \dots, \mathbf{w}_j$, are *generalized eigenvectors*, in accordance with the following definition.

Definition 8.40. A nonzero vector $\mathbf{w} \neq \mathbf{0}$ that satisfies

$$(A - \lambda\,\mathrm{I}\,)^k\,\mathbf{w} = \mathbf{0} \tag{8.46}$$

for some $k > 0$ and $\lambda \in \mathbb{C}$ is called a *generalized eigenvector* of the matrix A.

Note that every ordinary eigenvector is automatically a generalized eigenvector, since we can just take $k = 1$ in (8.46), but the converse is not necessarily valid. We shall call the minimal value of k for which (8.46) holds the *index* of the generalized eigenvector. Thus, an ordinary eigenvector is a generalized eigenvector of index 1. Since $A - \lambda\,\mathrm{I}$ is nonsingular whenever λ is not an eigenvalue of A, its k^{th} power $(A - \lambda\,\mathrm{I}\,)^k$ is also nonsingular. Therefore, generalized eigenvectors can only exist when λ is an ordinary eigenvalue of A — there are no additional "generalized eigenvalues".

Lemma 8.41. *The i^{th} vector $\mathbf{w}_i$ in a Jordan chain* (8.45) *is a generalized eigenvector of index i.*

Proof: By definition, $(A - \lambda\,\mathrm{I}\,)\,\mathbf{w}_1 = \mathbf{0}$, and so $\mathbf{w}_1$ is an eigenvector. Next, we have $(A - \lambda\,\mathrm{I}\,)\,\mathbf{w}_2 = \mathbf{w}_1$, and so $(A - \lambda\,\mathrm{I}\,)^2\,\mathbf{w}_2 = (A - \lambda\,\mathrm{I}\,)\,\mathbf{w}_1 = \mathbf{0}$. Thus, $\mathbf{w}_2$ a generalized eigenvector of index 2. A simple induction proves that $(A - \lambda\,\mathrm{I}\,)^i\,\mathbf{w}_i = \mathbf{0}$. *Q.E.D.*

Example 8.42. Consider the 3×3 Jordan block $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$. The only eigenvalue is $\lambda = 2$, and $A - 2\,\mathrm{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$. We claim that the standard basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ form a Jordan chain. Indeed, $A\,\mathbf{e}_1 = 2\,\mathbf{e}_1$, and hence $\mathbf{e}_1 \in \ker(A - 2\,\mathrm{I}\,)$ is a genuine eigenvector. Furthermore, $A\,\mathbf{e}_2 = 2\,\mathbf{e}_2 + \mathbf{e}_1$, and $A\,\mathbf{e}_3 = 2\,\mathbf{e}_3 + \mathbf{e}_2$, as you can easily check. Thus, $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ satisfy the Jordan chain equations for the eigenvalue $\lambda = 2$. Note that $\mathbf{e}_2$ lies in the kernel of $(A - 2\,\mathrm{I}\,)^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, and so is a generalized eigenvector of index 2. Indeed, every vector of the form $\mathbf{w} = a\,\mathbf{e}_1 + b\,\mathbf{e}_2$ with $b \neq 0$ is a generalized eigenvector of index 2. (When $b = 0$, $a \neq 0$, the vector $\mathbf{w} = a\,\mathbf{e}_1$ is an ordinary eigenvector of index 1.) Finally, $(A - 2\,\mathrm{I}\,)^3 = \mathrm{O}$, and so every vector $\mathbf{v} \in \mathbb{R}^3$, including $\mathbf{e}_3$, is a generalized eigenvector of index 3 (or less).

Given a matrix A, a basis of $\mathbb{R}^n$ or $\mathbb{C}^n$ is called a *Jordan basis* if it consists of one or more nonoverlapping Jordan chains. Thus, for the Jordan matrix in Example 8.42, the standard basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is, in fact, a Jordan basis. An eigenvector basis qualifies as a Jordan basis, since each eigenvector belongs to a Jordan chain of length 1. Jordan bases are the desired extension of eigenvector bases, valid for all square matrices.

Theorem 8.43. *Every $n \times n$ matrix admits a Jordan basis of $\mathbb{C}^n$. The first elements of the Jordan chains form a maximal system of linearly independent eigenvectors. Moreover, the number of generalized eigenvectors in the Jordan basis that are in Jordan chains associated with an eigenvalue λ is the same as the eigenvalue's multiplicity.*

Example 8.44. Consider the matrix $A = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}$. With some work, its characteristic equation is found to be

$$p_A(\lambda) = \det(A - \lambda\, \mathrm{I}\,) = \lambda^5 + \lambda^4 - 5\,\lambda^3 - \lambda^2 + 8\,\lambda - 4 = (\lambda - 1)^3(\lambda + 2)^2 = 0,$$

and hence A has two eigenvalues: 1, which is a triple eigenvalue, and -2, which is double. Solving the associated homogeneous system $(A - \lambda\, \mathrm{I}\,)\mathbf{v} = \mathbf{0}$, we discover that, up to constant multiple, there are only two eigenvectors: $\mathbf{v}_1 = (\,0, 0, 0, -1, 1\,)^T$ for $\lambda_1 = 1$ and, anticipating our final numbering, $\mathbf{v}_4 = (\,-1, 1, 1, -2, 0\,)^T$ for $\lambda_2 = -2$. Thus, A is far from complete.

To construct a Jordan basis, we first note that since A has 2 linearly independent eigenvectors, the Jordan basis will contain two Jordan chains; the one associated with the triple eigenvalue $\lambda_1 = 1$ has length 3, while $\lambda_2 = -2$ admits a Jordan chain of length 2. To construct the former, we need to first solve the system $(A - \mathrm{I}\,)\mathbf{w} = \mathbf{v}_1$. Note that the coefficient matrix is singular — it must be since 1 is an eigenvalue — and the general solution is $\mathbf{w} = \mathbf{v}_2 + t\,\mathbf{v}_1$ where $\mathbf{v}_2 = (\,0, 1, 0, 0, -1\,)^T$, and t is the free variable. The appearance of an arbitrary multiple of the eigenvector $\mathbf{v}_1$ in the solution is not unexpected; indeed, the kernel of $A - \mathrm{I}\,$ is the eigenspace for $\lambda_1 = 1$. We can choose any solution, e.g., $\mathbf{v}_2$ as the second element in the Jordan chain. To find the last element of the chain, we solve $(A - \mathrm{I}\,)\mathbf{w} = \mathbf{v}_2$ to find $\mathbf{w} = \mathbf{v}_3 + t\,\mathbf{v}_1$ where $\mathbf{v}_3 = (\,0, 0, 0, 1, 0\,)^T$ can be used as the Jordan chain element. Similarly, to construct the Jordan chain for the second eigenvalue, we solve $(A + 2\,\mathrm{I}\,)\mathbf{w} = \mathbf{v}_4$ and find $\mathbf{w} = \mathbf{v}_5 + t\,\mathbf{v}_4$ where $\mathbf{v}_5 = (\,-1, 0, 0, -2, 1\,)^T$. Thus, the desired Jordan basis is

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -2 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_5 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ -2 \\ 1 \end{pmatrix},$$

with $A\,\mathbf{v}_1 = \mathbf{v}_1$, $A\,\mathbf{v}_2 = \mathbf{v}_1 + \mathbf{v}_2$, $A\,\mathbf{v}_3 = \mathbf{v}_2 + \mathbf{v}_2$, $A\,\mathbf{v}_4 = -2\,\mathbf{v}_4$, $A\,\mathbf{v}_5 = \mathbf{v}_4 - 2\,\mathbf{v}_5$.

To prove Theorem 8.43, we begin with a simple lemma.

Lemma 8.45. *If $\mathbf{v}_1, \dots, \mathbf{v}_n$ forms a Jordan basis for the matrix A, it also forms a Jordan basis for $B = A - c\,\mathrm{I}$, for any scalar c.*

Proof: We note that the eigenvalues of B are of the form $\lambda - c$, where λ is an eigenvalue of A. Moreover, given a Jordan chain $\mathbf{w}_1, \dots, \mathbf{w}_j$ of A, we have

$$B\,\mathbf{w}_1 = (\lambda - c)\,\mathbf{w}_1, \qquad B\,\mathbf{w}_i = (\lambda - c)\,\mathbf{w}_i + \mathbf{w}_{i-1}, \qquad i = 2, \dots, j,$$

so $\mathbf{w}_1, \dots, \mathbf{w}_j$ is also a Jordan chain for B corresponding to the eigenvalue $\lambda - c$. *Q.E.D.*

The proof of Theorem 8.43 will be done by induction on the size n of the matrix. The case $n = 1$ is trivial, since any nonzero element of $\mathbb{C}$ is a Jordan basis for a 1×1 matrix $A = (a)$. To perform the induction step, we assume that the result is valid for all matrices of size $\leq n - 1$. Let A be an $n \times n$ matrix. According to Theorem 8.10, A has at least one complex eigenvalue λ. Let $B = A - \lambda\,\mathrm{I}$. Since λ is an eigenvalue of A, we know that 0 is an eigenvalue of B. This means that $\ker B \neq \{\mathbf{0}\}$, and so $r = \operatorname{rank} B < n$. Moreover, by Lemma 8.45, any Jordan basis of B is also a Jordan basis for A, and so we can concentrate all our attention on the singular matrix B from now on.

We note that $W = \operatorname{rng} B \subset \mathbb{C}^n$ is an invariant subspace, i.e., $B\,\mathbf{w} \in W$ whenever $\mathbf{w} \in W$, cf. Exercise ∎. Moreover, since B is singular, $\dim W = r = \operatorname{rank} B < n$. Thus, by fixing a basis of W, we can realize the restriction $B\colon W \to W$ as multiplication by an $r \times r$ matrix. The fact that $r < n$ allows us to invoke the induction hypothesis, and deduce the existence of a Jordan basis $\mathbf{w}_1, \dots, \mathbf{w}_r \in W \subset \mathbb{C}^n$ for the action of B on the subspace W. Our goal is to complete this collection to a full Jordan basis on $\mathbb{C}^n$.

To this end, we append two additional kinds of vectors. Suppose that the Jordan basis of W contains k null Jordan chains associated with the zero eigenvalue. Each null Jordan chain consists of vectors $\mathbf{w}_1, \dots, \mathbf{w}_j \in W$ satisfying

$$B\,\mathbf{w}_1 = \mathbf{0}, \qquad B\,\mathbf{w}_2 = \mathbf{w}_1, \qquad \dots \qquad B\,\mathbf{w}_j = \mathbf{w}_{j-1}. \tag{8.47}$$

The number of null Jordan chains is equal to the number of linearly independent null eigenvectors of B in $W = \operatorname{rng} B$, that is $k = \dim(\ker B \cap \operatorname{rng} B)$. To each null Jordan chain (8.47), we append a vector $\mathbf{w}_{j+1} \in \mathbb{C}^n$ such that

$$B\,\mathbf{w}_{j+1} = \mathbf{w}_j; \tag{8.48}$$

the existence of $\mathbf{w}_{j+1}$ comes from our condition that $\mathbf{w}_j \in \operatorname{rng} B$. Appending (8.48) to (8.47), we deduce that $\mathbf{w}_1, \dots, \mathbf{w}_{j+1} \in \mathbb{C}^n$ forms a null Jordan chain, of length $j + 1$, for B. Having extended all the null Jordan chains in W, the resulting collection contains $r + k$ vectors in $\mathbb{C}^n$ arranged in nonoverlapping Jordan chains. To complete to a basis, we include $n - r - k$ additional linearly independent null vectors $\mathbf{z}_1, \dots, \mathbf{z}_{n-r-k} \in \ker B \setminus \operatorname{rng} B$ that lie outside its range. Since $B\,\mathbf{z}_j = \mathbf{0}$, each $\mathbf{z}_j$ forms a null Jordan chain of length 1. We claim that the complete collection consisting of the non-null Jordan chains in W, the k extended null chains, and the additional null vectors $\mathbf{z}_1, \dots, \mathbf{z}_{n-r-k}$, forms the desired Jordan basis. By construction, it consists of nonoverlapping Jordan chains. The only remaining technical issue is proving that the vectors are linear independent, which is left as a challenge for the reader in Exercise ∎.

Just as an eigenvector basis diagonalizes a complete matrix, a Jordan basis provides a particularly simple form for an incomplete matrix, known as the *Jordan canonical form*.

Definition 8.46. A $n \times n$ matrix of the form[†]

$$J_{\lambda,n} = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \tag{8.49}$$

in which λ is a real or complex number, is known as a *Jordan block*.

In particular, a 1×1 Jordan block is merely a scalar $J_{\lambda,1} = \lambda$. Since every matrix has at least one (complex) eigenvector the Jordan block matrices have the least possible number of eigenvectors.

Lemma 8.47. *The $n \times n$ Jordan block matrix $J_{\lambda,n}$ has a single eigenvalue, λ, and a single independent eigenvector, $\mathbf{e}_1$. The standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ form a Jordan chain for $J_{\lambda,n}$.*

Definition 8.48. A *Jordan matrix* is a square matrix of block diagonal form

$$J = \operatorname{diag}(J_{\lambda_1,n_1}, J_{\lambda_2,n_2}, \dots, J_{\lambda_k,n_k}) = \begin{pmatrix} J_{\lambda_1,n_1} & & & \\ & J_{\lambda_2,n_2} & & \\ & & \ddots & \\ & & & J_{\lambda_k,n_k} \end{pmatrix}, \tag{8.50}$$

in which one or more Jordan blocks, not necessarily of the same size, lie along the diagonal, while all off-diagonal blocks are zero.

Note that the only non-zero entries in a Jordan matrix are those on the diagonal, which can have any complex value, and those on the superdiagonal, which are either 1 or 0. The positions of the superdiagonal 1's uniquely prescribes the Jordan blocks.

For example, the 6×6 matrices

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

are all Jordan matrices; the first is a diagonal matrix, consisting of 6 distinct 1×1 Jordan blocks; the second has a 4×4 Jordan block followed by a 2×2 block that happen to have the same diagonal entries; the last has three 2×2 Jordan blocks.

[†] All non-displayed entries are zero.

As a simple corollary of Lemma 8.47 combined with the matrix's block structure, as in Exercise ∎, we obtain a complete classification of the eigenvectors and eigenvalues of a Jordan matrix.

Lemma 8.49. *The Jordan matrix* (8.50) *has eigenvalues* $\lambda_1, \dots, \lambda_k$. *The standard basis vectors* $\mathbf{e}_1, \dots, \mathbf{e}_n$ *form a Jordan basis; the Jordan chains are labeled by the Jordan blocks.*

Thus, in the preceding examples of Jordan matrices, the first has three double eigenvalues, $1, 2, 3$, and corresponding linearly independent eigenvectors $\mathbf{e}_1, \mathbf{e}_6; \mathbf{e}_2, \mathbf{e}_5; \mathbf{e}_3, \mathbf{e}_4$, each of which belongs to a Jordan chain of length 1. The second matrix has only one eigenvalue, -1, but two Jordan chains, namely $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ and $\mathbf{e}_5, \mathbf{e}_6$. The last has eigenvalues $0, 1, 2$ and three Jordan chains, namely $\mathbf{e}_1, \mathbf{e}_2$, and $\mathbf{e}_3, \mathbf{e}_4$, and $\mathbf{e}_5, \mathbf{e}_6$. In particular, the only complete Jordan matrices are the diagonal matrices, all of whose Jordan blocks are of size 1×1.

Theorem 8.50. *Let* A *be an* $n \times n$ *real or complex matrix. Let* $S = (\, \mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n \,)$ *be the matrix whose columns are a Jordan basis of* A. *Then* S *places* A *in* Jordan canonical form

$$S^{-1} A S = J = \operatorname{diag}(J_{\lambda_1, n_1}, J_{\lambda_2, n_2}, \dots, J_{\lambda_k, n_k}). \tag{8.51}$$

The diagonal entries of the similar Jordan matrix J *are the eigenvalues of* A. *In particular,* A *is complete* (*diagonalizable*) *if and only if every Jordan block is of size* 1×1 *or, equivalently, all Jordan chains are of length* 1. *The Jordan canonical form of* A *is uniquely determined up to a permutation of the diagonal Jordan blocks.*

For instance, the matrix $A = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}$ considered in Example 8.44 has the following Jordan basis matrix and Jordan canonical form

$$S = \begin{pmatrix} 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 1 & -2 & -2 \\ 1 & -1 & 0 & 0 & 1 \end{pmatrix}, \qquad J = S^{-1} A S = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix}.$$

Chapter 9

Linear Dynamical Systems

The term *dynamical system* refers to the (differential) equations governing the time-varying behavior of some physical system. A system consisting of discrete units, e.g., the vibrations of a mass–spring chain, a more general structure, or an electrical circuit, will be modelled by a system of ordinary differential equations. In this chapter, we will analyze the simplest class of dynamical systems: first and second order linear systems of ordinary differential equations. Dynamics of continuous media — fluids, solids and gases — are modelled by partial differential equations, and will form the focus of the later chapters.

In Chapter 8 we first motivated and then developed the required mathematical tools — eigenvalues and eigenvectors — for the analysis of such linear systems. For a first order linear system, the "eigensolutions" describe the basic modes of exponential growth, decay, or periodic behavior. The stability of the equilibrium solution is almost entirely determined by the eigenvalues of the coefficient matrix, which highlights their critical role in real-world applications. Many of the basic phenomena already make an appearance in the two-dimensional case, and we devote Section 9.3 to a complete description of the planar linear systems. In the Section 9.4, we re-interpret the solution of a first order systems in terms of the matrix exponential, which is a direct analog of the usual scalar exponential function. Matrix exponentials are particularly effective for solving inhomogeneous or forced linear systems, and have useful applications in geometry and computer graphics.

Mechanical or electrical systems are modeled by second order linear equations. In the absence of damping or frictional effects, the eigensolutions constitute the system's normal or internal modes, each periodically vibrating with its associated fundamental frequency — the square root of the eigenvalue. The general motion is obtained by linear superposition, and is, in general, no longer periodic. To the eye, the resulting "quasi-periodic" motion may seem quite erratic — even though mathematically it is merely the superposition of a finite number of simple periodic motions. When subjected to periodic forcing, the system usually remains in a quasi-periodic motion that superimposes a periodic response to the forcing onto the internal vibrations. However, attempting to force the system at one of its internal frequencies can lead to resonance, where the vibrations become larger and larger, eventually resulting in catastrophic breakdown. Frictional effects will damp out the quasiperiodic vibrations, and help mitigate the effects of resonance.

9.1. Basic Solution Methods.

Now we have accumulated enough experience in the theory and computation of eigenvalues and eigenvectors to be able to analyze dynamical systems governed by linear, homogeneous, constant coefficient ordinary differential equations. Our initial focus will be

on systems

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u} \tag{9.1}$$

consisting of n first order linear ordinary differential equations in n unknowns $\mathbf{u}(t) = (\,u_1(t), \dots, u_n(t)\,)^T \in \mathbb{R}^n$. The coefficient matrix A, of size $n \times n$, is assumed to be a constant real matrix — although extensions to complex systems are not difficult. Nonautonomous systems, in which $A(t)$ depends on the time, are more difficult and we shall not attempt to solve them in this book.

As we saw in Section 8.1, the vector-valued exponential function $\mathbf{u}(t) = e^{\lambda t}\,\mathbf{v}$ is a (non-zero) solution to (9.1) if and only if

$$A\,\mathbf{v} = \lambda\,\mathbf{v}$$

and hence, assuming $\mathbf{v} \neq \mathbf{0}$, the scalar λ must an eigenvalue of A and $\mathbf{v}$ the corresponding eigenvector. Linear superposition can be employed to combine the resulting exponential eigensolutions to the system. In particular, if the coefficient matrix A is complete, then it admits n linearly independent eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, which, along with their associated eigenvalues $\lambda_1, \dots, \lambda_n$ will produce n distinct exponential solutions

$$\mathbf{u}_1(t) = e^{\lambda_1 t}\,\mathbf{v}_1, \qquad \dots \qquad \mathbf{u}_n(t) = e^{\lambda_n t}\,\mathbf{v}_n. \tag{9.2}$$

Since the system (9.1) is linear and homogeneous, for any constant scalars $c_1, \dots, c_n$, the linear combination

$$\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \,\cdots\, + c_n\mathbf{u}_n(t) = c_1\,e^{\lambda_1 t}\,\mathbf{v}_1 + \,\cdots\, + c_n\,e^{\lambda_n t}\,\mathbf{v}_n, \tag{9.3}$$

of the exponential eigensolutions is also a solution.

Are there any other solutions? The answer is no — in fact (9.3) represents the most general solution to the system. This result is a consequence of the basic existence and uniqueness theorem for linear systems of ordinary differential equations, which we discuss next.

Example 9.1. Consider the linear system

$$\frac{du}{dt} = 3\,u + v, \qquad \frac{dv}{dt} = u + 3\,v. \tag{9.4}$$

We first write the system in matrix form

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}, \quad \text{with unknown } \mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} \text{ and coefficient matrix } A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

In Example 8.5, we found the the eigenvalues and eigenvectors of A to be

$$\lambda_1 = 4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = 2, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We use the eigenvalues and eigenvectors to construct the two particular exponential solutions

$$\mathbf{u}_1(t) = e^{4\,t}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} e^{4\,t} \\ e^{4t} \end{pmatrix}, \qquad \mathbf{u}_2(t) = e^{2\,t}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\,e^{2\,t} \\ e^{2t} \end{pmatrix}.$$

According to the preceding remark, to be justified below, the general solution to (9.4) is then given as a linear combination

$$\mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = c_1\, e^{4\,t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\, e^{2\,t} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1\, e^{4\,t} - c_2\, e^{2\,t} \\ c_1\, e^{4t} + c_2\, e^{2t} \end{pmatrix}, \tag{9.5}$$

where c_1, c_2 are arbitrary constants.

The Phase Plane

Many fundamental physical phenomena are modeled by second order ordinary differential equations. The simplest scalar version is a linear, homogeneous equation

$$\frac{d^2u}{dt^2} + \alpha\, \frac{du}{dt} + \beta\, u = 0, \tag{9.6}$$

in which α, β are prescribed constants. In your first course on ordinary differential equations, you learned how to solve such equations; the basic method is reviewed in the following example; see also Example 7.31.

Example 9.2. Consider the second order equation

$$\frac{d^2u}{dt^2} + \frac{du}{dt} - 6\, u = 0. \tag{9.7}$$

To solve the equation, we substitute an exponential formula or ansatz[†] $u(t) = e^{\lambda\, t}$ into the equation. The result is the *characteristic equation* for the unspecified exponent λ:

$$\lambda^2 + \lambda - 6 = 0, \qquad \text{with solutions} \qquad \lambda_1 = 2, \qquad \lambda_2 = -3.$$

We conclude that $e^{2\,t}$ and $e^{-3\,t}$ form a basis for the two-dimensional solution space to (9.7), and so the general solution can be written as a linear combination

$$u(t) = c_1\, e^{2\,t} + c_2\, e^{-3\,t},$$

where c_1, c_2 are arbitrary constants. (See Theorem 7.33 for a justification.)

There is a standard trick to convert any second order scalar equation, e.g., (9.6), into a first order system. One introduces the variables[‡]

$$u_1 = u, \qquad\qquad u_2 = \dot{u} = \frac{du}{dt}\,. \tag{9.8}$$

In view of (9.6), these variables satisfy

$$\frac{du_1}{dt} = \frac{du}{dt} = u_2, \qquad\qquad \frac{du_2}{dt} = \frac{d^2u}{dt^2} = -\,\beta\, u - \alpha\, \frac{du}{dt} = -\,\beta\, u_1 - \alpha\, u_2.$$

[†] See the footnote on p. 344 for an explanation of this term.

[‡] We will often use dots as a shorthand notation for time derivatives.

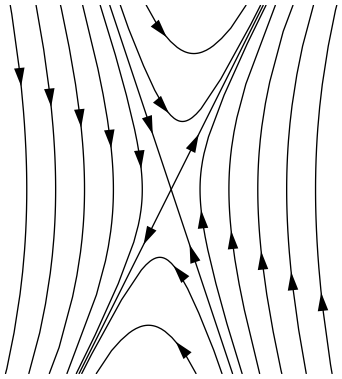


Figure 9.1. Phase Plane Trajectories for $\dot{u}_1 = u_2$, $\dot{u}_2 = 6\,u_1 - u_2$.

In this manner, the second order equation is converted into the equivalent first order system

$$\dot{\mathbf{u}} = A\,\mathbf{u}, \qquad \text{where} \qquad \mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 1 \\ -\beta & -\alpha \end{pmatrix}. \tag{9.9}$$

The $(u_1, u_2) = (u, \dot{u})$ plane is referred to as the *phase plane*. The solutions $\mathbf{u}(t)$ to (9.9) parametrize curves in the phase plane — the solution *trajectories* or *orbits*. In particular, the equilibrium solution $\mathbf{u}(t) \equiv \mathbf{0}$ remains fixed at the origin, and so its trajectory is a single point All other solutions describe genuine curves. The collection of all possible solution trajectories is called the *phase portrait* of the system. An important fact is that, for a (constant coefficient) first order system, *the phase plane trajectories never cross*. This striking property, which is also valid for nonlinear systems, is a consequence of the uniqueness properties of solutions, and will be discussed in detail in Section 20.2. Thus, the phase portrait consists of a family of non-intersecting curves and equilibrium points that fill out the entire phase plane. The direction of motion along the trajectory is indicated by a small arrow. The one feature that is not so easily pictured in the phase portrait is the speed at which the solution moves along the phase curves — this would require a more complicated three-dimensional plot in which the third axis indicates time.

It is not hard to verify that every solution $u(t)$ to the second order equation yields a solution $\mathbf{u}(t) = \big(\, u(t), \dot{u}(t) \,\big)^T$ to the phase plane system (9.9). Vice versa, if $\mathbf{u}(t) = (\, u_1(t), u_2(t) \,)^T$ is any solution to the system (9.9), then its first component $u(t) = u_1(t)$ defines a solution to the original scalar equation (9.6). We conclude that the scalar equation and its phase plane version are completely equivalent; solving one will immediately lead to a solution of the other.

Example 9.3. For the second order equation (9.7), the equivalent phase plane system is

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 1 \\ 6 & -1 \end{pmatrix} \mathbf{u}, \qquad \text{or, in full detail,} \qquad \begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= 6\,u_1 - u_2. \end{aligned} \tag{9.10}$$

Our identification (9.8) of the phase plane variables tells us that the solution to the system

(9.10) is given by

$$u_1(t) = u(t) = c_1\, e^{2\,t} + c_2\, e^{-3\,t},$$
$$u_2(t) = \frac{du}{dt} = 2\,c_1\, e^{2\,t} - 3\,c_2\, e^{-3\,t},$$

and hence

$$\mathbf{u}(t) = \begin{pmatrix} c_1\, e^{2\,t} + c_2\, e^{-3\,t} \\ 2\,c_1\, e^{2\,t} - 3\,c_2\, e^{-3\,t} \end{pmatrix} = c_1 \begin{pmatrix} e^{2\,t} \\ 2\,e^{2\,t} \end{pmatrix} + c_2 \begin{pmatrix} e^{-3\,t} \\ -3\,e^{-3\,t} \end{pmatrix}.$$

A plot of the phase plane trajectories $\mathbf{u}(t)$ for various choices of the constants c_1, c_2 appears in Figure 9.1. The horizontal axis represents the solution $u_1 = u(t)$ whereas the vertical axis represents is derivative $u_2 = \dot{u}(t)$. With some practice, one learns to understand the temporal behavior of the solution from studying its phase plane trajectory. Many more examples will appear in Section 9.3 below.

Existence and Uniqueness

Before proceeding further, it will help to briefly summarize the existence and uniqueness properties for solutions to linear systems of ordinary differential equations. These are direct consequences of the general existence and uniqueness theorem for nonlinear systems of ordinary differential equations, to be discussed in Section 20.1. Even though we will only study the constant coefficient case in detail in this text, the results are equally applicable to homogeneous linear systems with variable coefficients, and so, but only in this subsection, we allow the coefficient matrix to depend continuously on t.

The key fact is that a system of n first order ordinary differential equations requires n initial conditions — one for each variable — in order to specify its solution uniquely. More specifically:

Theorem 9.4. *Let $A(t)$ be an $n \times n$ matrix of continuous functions on the interval $a < t < b$. Given an initial time $a < t_0 < b$ and an initial vector $\mathbf{b} \in \mathbb{R}^n$, the* initial value problem

$$\frac{d\mathbf{u}}{dt} = A(t)\,\mathbf{u}, \qquad \mathbf{u}(t_0) = \mathbf{b}, \tag{9.11}$$

admits a unique solution $\mathbf{u}(t)$ which is defined for all $a < t < b$.

In particular, an initial value problem for a constant coefficient system (9.1) admits a unique solution $\mathbf{u}(t)$ that is defined for all $-\infty < t < \infty$. Uniqueness of solutions implies that, for such homogeneous systems, the solution with zero initial conditions $\mathbf{u}(t_0) = \mathbf{0}$ is the trivial zero solution $\mathbf{u}(t) \equiv \mathbf{0}$. Uniqueness has the important consequence that linear independence needs only be checked at a single point.

Lemma 9.5. *The solutions $\mathbf{u}_1(t), \dots, \mathbf{u}_k(t)$ to a first order homogeneous linear system $\dot{\mathbf{u}} = A(t)\,\mathbf{u}$ are linearly independent functions if and only if their initial values $\mathbf{u}_1(t_0), \dots, \mathbf{u}_k(t_0)$ are linearly independent vectors in $\mathbb{R}^n$.*

Proof: If the solutions are linearly dependent, then

$$\mathbf{u}(t) = c_1\,\mathbf{u}_1(t) + \,\cdots\, + c_k\,\mathbf{u}_k(t) \equiv \mathbf{0} \tag{9.12}$$

for some constant scalars $c_1, \ldots, c_k$ not all zero. The equation holds, in particular, at $t = t_0$,

$$\mathbf{u}(t_0) = c_1\,\mathbf{u}_1(t_0) + \,\cdots\, + c_k\,\mathbf{u}_k(t_0) = \mathbf{0}, \tag{9.13}$$

proving linear dependence of the initial vectors. Conversely, if the initial values are linearly dependent, then (9.13) hold for some $c_1, \ldots, c_k$ not all zero. Linear superposition implies that the corresponding linear combination $\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \,\cdots\, + c_k\mathbf{u}_k(t)$ is a solution to the system, with zero initial condition. By uniqueness, $\mathbf{u}(t) \equiv \mathbf{0}$ for all t, and so (9.12) holds, proving linear dependence of the solutions. *Q.E.D.*

Warning: This result is *not* true if the functions are not solutions to a *first order* system! For example, $\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$, $\mathbf{u}_2(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$, are linearly independent vector-valued functions, but, at time $t = 0$, the vectors $\mathbf{u}_1(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{u}_2(0)$ are linearly dependent. Even worse, $\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$, $\mathbf{u}_2(t) = \begin{pmatrix} t \\ t^2 \end{pmatrix}$, define linearly dependent vectors at every fixed value of t, but as vector-valued functions they are, nonetheless, linearly independent. In view of Lemma 9.5, neither pair of functions can be solutions to a common linear ordinary differential equation.

The next result tells us how many different solutions we need in order to construct the general solution by linear superposition.

Theorem 9.6. *Let $\mathbf{u}_1(t), \ldots, \mathbf{u}_n(t)$ be n linearly independent solutions to the homogeneous system of n first order linear ordinary differential equations $\dot{\mathbf{u}} = A(t)\,\mathbf{u}$. then the general solution is a linear combination $\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \,\cdots\, + c_n\mathbf{u}_n(t)$ depending on n arbitrary constants $c_1, \ldots, c_n$.*

Proof: If we have n linearly independent solutions, then Lemma 9.5 implies that, at the initial time t_0, the vectors $\mathbf{u}_1(t_0), \ldots, \mathbf{u}_n(t_0)$ are linearly independent, and hence form a basis for $\mathbb{R}^n$. This means that we can express any initial condition

$$\mathbf{u}(t_0) = \mathbf{b} = c_1\mathbf{u}_1(t_0) + \,\cdots\, + c_n\mathbf{u}_n(t_0)$$

as a linear combination of the initial vectors. Superposition and uniqueness of solutions implies that the corresponding solution to the initial value problem (9.11) is given by the same linear combination

$$\mathbf{u}(t) = \mathbf{b} = c_1\mathbf{u}_1(t) + \,\cdots\, + c_n\mathbf{u}_n(t).$$

We conclude that every solution to the ordinary differential equation can be written in the prescribed form. *Q.E.D.*

Complete Systems

Now we have assembled the basic ingredients that will enable us to construct the complete solution to most first order homogeneous, linear, constant coefficient systems of ordinary differential equations. For a system of n equations, the goal is to find n linearly independent solutions. Each eigenvalue and eigenvector leads to an exponential solution of the form $e^{\lambda t}\,\mathbf{v}$. The solutions will be linearly independent if and only if the eigenvectors are — this will follow easily from Lemma 9.5. Thus, if the $n \times n$ matrix admits an eigenvector basis, i.e., it is complete, then we have the requisite number of solutions, and hence have solved the differential equation.

Theorem 9.7. *If the $n \times n$ matrix A is complete, then the general (complex) solution to the constant coefficient linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$ is given by*

$$\mathbf{u}(t) = c_1\, e^{\lambda_1 t}\,\mathbf{v}_1 + \,\cdots\, + c_n\, e^{\lambda_n t}\,\mathbf{v}_n, \tag{9.14}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the eigenvector basis, $\lambda_1, \ldots, \lambda_n$ the corresponding eigenvalues, and $c_1, \ldots, c_n$ arbitrary constants, uniquely specified by the initial conditions $\mathbf{u}(t_0) = \mathbf{b}$.

Proof: Since the eigenvectors are linearly independent, the solutions define linearly independent vectors $\mathbf{u}_k(0) = \mathbf{v}_k$ at time $t = 0$. Thus, Lemma 9.5 implies that the functions $\mathbf{u}_k(t)$ are, indeed, linearly independent. Therefore, the result is an immediate consequence of Theorem 9.6. *Q.E.D.*

Example 9.8. Let us solve the initial value problem

$$\begin{aligned} \dot{u}_1 &= -2\,u_1 + u_2, & u_1(0) &= 3, \\ \dot{u}_2 &= 2\,u_1 - 3\,u_2 & u_2(0) &= 0. \end{aligned}$$

The coefficient matrix of the system is $A = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix}$. A straightforward computation produces the following eigenvalues and eigenvectors of A:

$$\lambda_1 = -4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \qquad\qquad \lambda_2 = -1, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The corresponding exponential solutions $\mathbf{u}_1(t) = e^{-4t}\begin{pmatrix} 1 \\ -2 \end{pmatrix}, \mathbf{u}_2(t) = e^{-t}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ form a basis for the two-dimensional solution space. The general solution is an arbitrary linear combination

$$\mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = c_1\, e^{-4t}\begin{pmatrix} 1 \\ -2 \end{pmatrix} + c_2\, e^{-t}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1\, e^{-4t} + c_2\, e^{-t} \\ -2\,c_1\, e^{-4t} + c_2\, e^{-t} \end{pmatrix},$$

where c_1, c_2 are constant scalars. Once we have the general solution in hand, the final step is to determine the values of c_1, c_2 so as to satisfy the initial conditions. Evaluating the solution at $t = 0$, we find we need to solve the linear system

$$c_1 + c_2 = 3, \qquad\qquad -2\,c_1 + c_2 = 0,$$

for $c_1 = 1$, $c_2 = 2$. Thus, the (unique) solution to the initial value problem is

$$u_1(t) = e^{-4t} + 2\,e^{-t}, \qquad u_2(t) = -2\,e^{-4t} + 2\,e^{-t}. \tag{9.15}$$

Note that both components of the solution decay exponentially fast to 0 as $t \to \infty$.

Example 9.9. Consider the linear system

$$\dot{u}_1 = u_1 + 2\,u_2, \qquad \dot{u}_2 = u_2 - 2\,u_3, \qquad \dot{u}_3 = 2\,u_1 + 2\,u_2 - u_3.$$

The coefficient matrix is

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

In Example 8.8 we computed the eigenvalues and eigenvectors:

$$\lambda_1 = -1, \qquad \lambda_2 = 1 + 2\,\mathrm{i}\,, \qquad \lambda_3 = 1 - 2\,\mathrm{i}\,,$$
$$\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -\mathrm{i} \\ 1 \end{pmatrix}.$$

The first leads to a real solution, but the second and third lead to complex solutions to our real system of equations, e.g., $\widehat{\mathbf{u}}_2(t) = e^{(1+2\,\mathrm{i}\,)\,t}\,(\,1,\,\mathrm{i}\,,1\,)^T$. While this is a perfectly valid complex solution, it is not so convenient to work with if, as in most applications, we require real-valued functions. Since the underlying linear system is real, the general reality principle of Theorem 7.47 says that any complex solution can be broken up into its real and imaginary parts, each of which is a *real* solution. Applying Euler's formula (3.76) to the complex exponential, we find

$$\widehat{\mathbf{u}}_2(t) = e^{(1+2i)t} \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix} = \big(\, e^t \cos 2t + \mathrm{i}\, e^t \sin 2t \,\big) \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix} = \begin{pmatrix} e^t \cos 2t \\ -e^t \sin 2t \\ e^t \cos 2t \end{pmatrix} + \mathrm{i} \begin{pmatrix} e^t \sin 2t \\ e^t \cos 2t \\ e^t \sin 2t \end{pmatrix},$$

which yields two real vector-valued solutions to the system, as you can readily check. In this manner, we have produced three linearly independent real solutions to our system:

$$\mathbf{u}_1(t) = \begin{pmatrix} -e^{-t} \\ e^{-t} \\ e^{-t} \end{pmatrix}, \qquad \mathbf{u}_2(t) = \begin{pmatrix} e^t \cos 2t \\ -e^t \sin 2t \\ e^t \cos 2t \end{pmatrix}, \qquad \mathbf{u}_3(t) = \begin{pmatrix} e^t \sin 2t \\ e^t \cos 2t \\ e^t \sin 2t \end{pmatrix}.$$

Theorem 9.6 tells us that the general solution is a linear combination of the 3 independent solutions:

$$\mathbf{u}(t) = c_1\,\mathbf{u}_1(t) + c_2\,\mathbf{u}_2(t) + c_3\,\mathbf{u}_3(t) = \begin{pmatrix} -c_1\,e^{-t} + c_2\,e^t \cos 2t + c_3\,e^t \sin 2t \\ c_1\,e^{-t} - c_2\,e^t \sin 2t + c_3\,e^t \cos 2t \\ c_1\,e^{-t} + c_2\,e^t \cos 2t + c_3\,e^t \sin 2t \end{pmatrix}.$$

The constants c_1, c_2, c_3 are uniquely prescribed by imposing initial conditions; for example, the solution with $\mathbf{u}(0) = (\,2, -1, -2\,)^T$ requires $c_1 = -2, c_2 = 0, c_3 = 1$, and so the

solution's components are $u_1(t) = 2\,e^{-t} + e^{t}\sin 2t$, $u_2(t) = -2\,e^{-t} + e^{t}\cos 2t$, $u_3(t) = -2\,e^{-t} + e^{t}\sin 2t$.

Incidentally, the third complex solution also produces two real solutions, but these reproduce the ones we have already listed. In fact, since $\lambda_3 = \overline{\lambda_2}$ is the complex conjugate of the eigenvalue λ_2, its eigenvector $\mathbf{v}_3 = \overline{\mathbf{v}_2}$ is also the complex conjugate of the eigenvector $\mathbf{v}_2$, and, finally, the solutions are also related by complex conjugation:

$$\widehat{\mathbf{u}}_3(t) = e^{(1-2i)t}\begin{pmatrix} 1 \\ -\,\mathrm{i} \\ 1 \end{pmatrix} = \begin{pmatrix} e^{t}\cos 2t \\ -\,e^{t}\sin 2t \\ e^{t}\cos 2t \end{pmatrix} - \,\mathrm{i}\begin{pmatrix} e^{t}\sin 2t \\ e^{t}\cos 2t \\ e^{t}\sin 2t \end{pmatrix} = \overline{\widehat{\mathbf{u}}_2(t)}\,.$$

In general, when dealing with complex eigenvalues of real systems, you only need to look at one eigenvalue from each complex conjugate pair to find a complete system of real solutions.

The General Case

If the matrix A is not complete, then the formulae for the solutions are a little more intricate, and involve polynomials as well as (complex) exponentials. When dealing with an incomplete matrix, we do not have sufficient eigenvectors to construct all the solutions, and so make use of its Jordan basis. Let us first describe the solutions associated with a Jordan chain.

Lemma 9.10. *Suppose $\mathbf{w}_1, \dots, \mathbf{w}_k$ form a Jordan chain of length k for the eigenvalue λ of the matrix A. Then there are k linearly independent solutions to the corresponding first order system $\dot{\mathbf{u}} = A\,\mathbf{u}$ having the form*

$$\mathbf{u}_1(t) = e^{\lambda t}\mathbf{w}_1, \qquad \mathbf{u}_2(t) = e^{\lambda t}(t\,\mathbf{w}_1 + \mathbf{w}_2), \qquad \mathbf{u}_3(t) = e^{\lambda t}\big(\tfrac{1}{2}\,t^2\,\mathbf{w}_1 + t\,\mathbf{w}_2 + \mathbf{w}_3\big),$$

$$\textit{and, in general,} \qquad \mathbf{u}_j(t) = e^{\lambda t}\sum_{i=1}^{j}\frac{t^{j-i}}{(j-i)!}\,\mathbf{w}_i, \qquad 1 \le j \le k. \tag{9.16}$$

The proof is by direct substitution of the formulae into the differential equation, using the defining relations (8.45) of the Jordan chain; details are left to the reader. If λ is a complex eigenvalue, then the Jordan chain solutions (9.16) will involve complex exponentials. As usual, they can be split into their real an imaginary parts which, provided A is a real matrix, are independent real solutions.

Example 9.11. The coefficient matrix of the system

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}\mathbf{u}$$

is incomplete; it has only 2 linearly independent eigenvectors associated with the eigenvalues 1 and -2. Using the Jordan basis computed in Example 8.44, we produce the following

5 linearly independent solutions:

$$\mathbf{u}_1(t) = e^t\,\mathbf{v}_1, \qquad \mathbf{u}_2(t) = e^t\,(t\,\mathbf{v}_1 + \mathbf{v}_2), \qquad \mathbf{u}_3(t) = e^t\,(\tfrac{1}{2}\,t^2\,\mathbf{v}_1 + t\,\mathbf{v}_2 + \mathbf{v}_3),$$
$$\mathbf{u}_4(t) = e^{-2\,t}\,\mathbf{v}_4, \qquad \mathbf{u}_5(t) = e^{-2\,t}\,(t\,\mathbf{v}_4 + \mathbf{v}_5),$$

or, explicitly,

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ -e^t \\ e^t \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -e^t \\ 0 \\ -t\,e^t \\ (1+t)\,e^t \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -t\,e^t \\ 0 \\ \left(1 - \frac{1}{2}\,t^2\right) e^t \\ \left(t + \frac{1}{2}\,t^2\right) e^t \end{pmatrix}, \quad \begin{pmatrix} -e^{-2\,t} \\ e^{-2\,t} \\ e^{-2\,t} \\ -2\,e^{-2\,t} \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -(1+t)\,e^{-2\,t} \\ t\,e^{-2\,t} \\ t\,e^{-2\,t} \\ -2\,(1+t)\,e^{-2\,t} \\ e^{-2\,t} \end{pmatrix}.$$

The first three are associated with the $\lambda_1 = 1$ Jordan chain, the last two with the $\lambda_2 = -2$ chain; the eigenvector solutions are the pure exponentials $\mathbf{u}_1(t), \mathbf{u}_4(t)$. The general solution is an arbitrary linear combination of these five basis solutions.

Theorem 9.12. *Let A be an $n \times n$ matrix. Then the solutions* (9.16) *constructed from the Jordan chains in a Jordan basis of A form a basis for the n-dimensional solution space for the corresponding linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$.*

While the full computational details can be quite messy, in practical situations one can glean a significant amount of information about the solutions to the system without much fuss. The following result outlines a general characterization of the solutions of homogeneous linear systems of ordinary differential equations. The result is direct consequence of the general solution formulae in (9.16).

Theorem 9.13. *Let A be a real, square matrix. The general real solution to any constant coefficient homogeneous linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$ is a linear combination of n linearly independent solutions of the following types*:

(a) *If λ is a complete real eigenvalue of multiplicity m, then there exist m linearly independent solutions of the form*

$$\mathbf{u}_k(t) = e^{\lambda\,t}\,\mathbf{v}_k, \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{v}_1, \dots, \mathbf{v}_m$ are linearly independent eigenvectors.

(b) *If $\mu \pm \mathrm{i}\,\nu$ form a pair of complete complex conjugate eigenvalues of multiplicity m, then there exist $2\,m$ linearly independent real solutions of the forms*

$$\begin{aligned} \mathbf{u}_k(t) &= e^{\mu\,t}\big[\,\cos(\nu\,t)\,\mathbf{w}_k - \sin(\nu\,t)\,\mathbf{z}_k\,\big], \\ \widehat{\mathbf{u}}_k(t) &= e^{\mu\,t}\big[\,\sin(\nu\,t)\,\mathbf{w}_k + \cos(\nu\,t)\,\mathbf{z}_k\,\big], \end{aligned} \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{v}_k = \mathbf{w}_k \pm \mathrm{i}\,\mathbf{z}_k$ are the associated complex conjugate eigenvectors.

(c) *If λ is an incomplete real eigenvalue of multiplicity m and $r = \dim V_\lambda$, then there exist m linearly independent solutions of the form*

$$\mathbf{u}_k(t) = e^{\lambda\,t}\,\mathbf{p}_k(t), \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{p}_k(t)$ is a vector of polynomials of degree $\le m - r$.

(d) *If $\mu \pm \mathrm{i}\,\nu$ form a pair of incomplete complex conjugate eigenvalues of multiplicity m and $r = \dim V_\lambda$, then there exist $2m$ linearly independent real solutions*

$$\begin{aligned} \mathbf{u}_k(t) &= e^{\mu t}\,\big[\,\cos(\nu t)\,\mathbf{p}_k(t) - \sin(\nu t)\,\mathbf{q}_k(t)\,\big], \\ \widehat{\mathbf{u}}_k(t) &= e^{\mu t}\,\big[\,\sin(\nu t)\,\mathbf{p}_k(t) + \cos(\nu t)\,\mathbf{q}_k(t)\,\big], \end{aligned} \qquad k = 1, \dots, m,$$

where $\mathbf{p}_k(t), \mathbf{q}_k(t)$ are vectors of polynomials of degree $\leq m - r$.

Corollary 9.14. *Every real solution to a homogeneous linear system of ordinary differential equations is a vector-valued function whose entries are linear combinations of functions of the particular form $t^k\, e^{\mu t} \cos \nu t$ and $t^k\, e^{\mu t} \sin \nu t$, i.e., sums of products of exponentials, trigonometric functions and polynomials. The exponents μ are the real parts of the eigenvalues of the coefficient matrix; the trigonometric frequencies ν are the imaginary parts of the eigenvalues; nonconstant polynomials appear only if the matrix is incomplete.*

Example 9.15. The incomplete cases should remind the reader of the solution to a single scalar ordinary differential equation in the case of a repeated root to the characteristic equation. For example, to solve the second order equation

$$\frac{d^2u}{dt^2} - 2\,\frac{du}{dt} + u = 0,$$

we substitute the exponential ansatz $u = e^{\lambda t}$, leading to the characteristic equation

$$\lambda^2 - 2\,\lambda + 1 = 0.$$

There is only one double root, $\lambda = 1$, and hence, up to scalar multiple, only one exponential solution $u_1(t) = e^t$. In the scalar case, the second "missing" solution is obtained by just multiplying by t, so that $u_2(t) = t\,e^t$. The general solution is

$$u(t) = c_1\,u_1(t) + c_2\,u_2(t) = c_1\,e^t + c_2\,t\,e^t.$$

The equivalent phase plane system (9.9) is

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix} \mathbf{u}, \qquad \text{where} \qquad \mathbf{u}(t) = \begin{pmatrix} u(t) \\ \dot{u}(t) \end{pmatrix}.$$

Note that the coefficient matrix is incomplete — it has $\lambda = 1$ as a double eigenvalue, but only one independent eigenvector, namely $\mathbf{v} = (\,1, 1\,)^T$. The two linearly independent solutions to the phase plane system can be constructed from the solutions $u_1(t) = e^t, u_2(t) = t\,e^t$ to the original equation, and so

$$\mathbf{u}_1(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}, \qquad \mathbf{u}_2(t) = \begin{pmatrix} t\,e^t \\ t\,e^t + e^t \end{pmatrix}. \tag{9.17}$$

Note the appearance of the polynomial factor t in the solution formula. The general solution is obtained as a linear combination of these two basic solutions. *Warning*: In (9.17), the second vector solution $\mathbf{u}_2$ is *not* obtained from the first by merely multiplying by t. Incomplete systems are not that easy to handle!

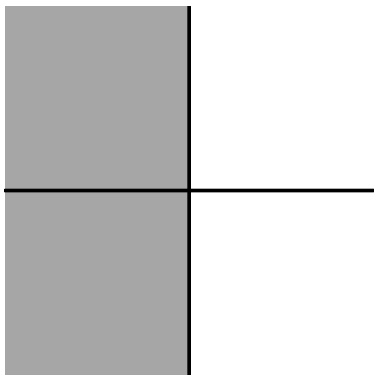


Figure 9.2. The Left Half Plane.

9.2. Stability of Linear Systems.

With the solution formulae in hand, we are now ready to study the qualitative features of first order linear dynamical systems. Our primary focus will be on stability properties of the equilibrium solution(s). The starting point is a simple calculus lemma, whose proof is left to the reader.

Lemma 9.16. *Let μ, ν be real and $k \geq 0$ an integer. A function of the form*

$$f(t) = t^k \, e^{\mu t} \cos \nu t \qquad \text{or} \qquad t^k \, e^{\mu t} \sin \nu t \tag{9.18}$$

will decay to zero for large t, so $\lim_{t \to \infty} f(t) = 0$, if and only if $\mu < 0$. The function remains bounded, so $| f(t) | \leq C$ for all $t \geq 0$, if and only if either $\mu < 0$, or $\mu = 0$ and $k = 0$.

In other words, exponential decay, where $\mu < 0$, will always cancel out polynomial growth, while trigonometric functions remain bounded. Now, in the solution to our ordinary differential equation, the functions (9.18) come from the eigenvalues $\lambda = \mu + \mathrm{i}\, \nu$ of the coefficient matrix. The lemma implies that the asymptotic behavior of the solutions, and hence their stability, depends on the sign of $\mu = \operatorname{Re} \lambda$. If $\mu < 0$, then the solutions decay to zero at an exponential rate as $t \to \infty$. If $\mu > 0$, then the solutions become unbounded as $t \to \infty$. In the borderline case $\mu = 0$, the solutions remain bounded provided they don't involve any powers of t.

Asymptotic stability of the equilibrium zero solution requires that all other solutions tend to $\mathbf{0}$ as $t \to \infty$, and hence all the eigenvalues must satisfy $\mu = \operatorname{Re} \lambda < 0$. Or, stated another way, all eigenvalues must lie in the *left half plane* — the subset of $\mathbb{C}$ to the left of the imaginary axis, as in Figure 9.2. In this manner, we have demonstrated the fundamental asymptotic stability criterion for linear systems.

Theorem 9.17. *A first order linear, homogeneous, constant-coefficient system of ordinary differential equations $\dot{\mathbf{u}} = A\,\mathbf{u}$ has asymptotically stable zero solution if and only if all the eigenvalues of the coefficient matrix A lie in the left half plane*: $\operatorname{Re} \lambda < 0$. *On the other hand, if A has one or more eigenvalues with positive real part, $\operatorname{Re} \lambda > 0$, then the zero solution is unstable.*

Example 9.18. Consider the system

$$\frac{du}{dt} = 2\,u - 6\,v + w, \qquad \frac{dv}{dt} = 3\,u - 3\,v - w, \qquad \frac{dw}{dt} = 3\,u - v - 3\,w.$$

The coefficient matrix $A = \begin{pmatrix} 2 & -6 & 1 \\ 3 & -3 & -1 \\ 3 & -1 & -3 \end{pmatrix}$ is found to have eigenvalues $\lambda_1 = -2$, $\lambda_2 = -1 + \mathrm{i}\,\sqrt{6}\,$, $\lambda_3 = -1 - \mathrm{i}\,\sqrt{6}\,$, with respective real parts $-2, -1, -1$. Therefore, according to Theorem 9.17, the zero solution $u \equiv v \equiv w \equiv 0$ is asymptotically stable. Indeed, the solutions involve linear combinations of the functions $e^{-2\,t}$, $e^{-t}\cos\sqrt{6}\,t$, and $e^{-t}\sin\sqrt{6}\,t$, all of which decay to 0 at an exponential rate. The latter two have the slowest decay rate, and so most solutions to the linear system go to $\mathbf{0}$ like a multiple of e^{-t}, i.e., at an exponential rate determined by the least negative real part.

A particularly important class of systems are the linear *gradient flows*

$$\frac{d\mathbf{u}}{dt} = -\,K\,\mathbf{u}, \tag{9.19}$$

in which $K > 0$ is a symmetric, positive definite matrix. According to Theorem 8.23, all the eigenvalues of K are real and positive. Therefore, the eigenvalues of the negative definite coefficient matrix $-K$ of the gradient flow system (9.19) are real and negative. Applying Theorem 9.17, we conclude that the zero solution to any gradient flow system (9.19) with negative definite coefficient matrix $-K$ is asymptotically stable.

Example 9.19. Using the methods of Chapter 3, the matrix $K = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}$ is found to be positive definite. The associated gradient flow is

$$\frac{du}{dt} = -u - v, \qquad \frac{dv}{dt} = -u - 5\,v. \tag{9.20}$$

The eigenvalues and eigenvectors of $-K = \begin{pmatrix} -1 & -1 \\ -1 & -5 \end{pmatrix}$ are

$$\lambda_1 = -3 + \sqrt{5}\,, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 - \sqrt{5} \end{pmatrix}, \qquad \lambda_2 = -3 - \sqrt{5}\,, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 + \sqrt{5} \end{pmatrix}.$$

Therefore, the general solution to the system is

$$\mathbf{u}(t) = c_1\, e^{(-3+\sqrt{5}\,)t} \begin{pmatrix} 1 \\ 2 - \sqrt{5} \end{pmatrix} + c_2\, e^{(-3-\sqrt{5}\,)t} \begin{pmatrix} 1 \\ 2 + \sqrt{5} \end{pmatrix},$$

or, in components,

$$\begin{aligned} u(t) &= c_1\, e^{(-3+\sqrt{5}\,)t} + c_2\, e^{(-3-\sqrt{5}\,)t}, \\ v(t) &= c_1\,(2 - \sqrt{5})\, e^{(-3+\sqrt{5}\,)t} + c_2\,(2 + \sqrt{5})\, e^{(-3-\sqrt{5}\,)t}. \end{aligned}$$

The solutions clearly tend to zero as $t \to \infty$ at the exponential rate prescribed by the least negative eigenvalue: $-3 + \sqrt{5} = -0.7639\ldots$. This confirms the asymptotic stability of the gradient flow.

The reason for the term "gradient flow" is that the vector field $-K\,\mathbf{u}$ appearing on the right hand side of (9.19) is, in fact, the negative of the gradient of the quadratic function

$$q(\mathbf{u}) = \tfrac{1}{2}\,\mathbf{u}^T K\,\mathbf{u} = \frac{1}{2}\sum_{i,j=1}^{n} k_{ij}\,u_i\,u_j, \tag{9.21}$$

namely $K\,\mathbf{u} = \nabla q(\mathbf{u})$. Thus, we can write (9.19) as

$$\frac{d\mathbf{u}}{dt} = -\,\nabla q(\mathbf{u}). \tag{9.22}$$

For the particular system (9.20),

$$q(u,v) = \tfrac{1}{2}\begin{pmatrix} u & v \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \tfrac{1}{2}\,u^2 + u\,v + \tfrac{5}{2}\,v^2,$$

and so the gradient flow is given by

$$\frac{du}{dt} = -\,\frac{\partial q}{\partial u} = -u - v, \qquad \frac{dv}{dt} = -\,\frac{\partial q}{\partial v} = -u - 5\,v.$$

Remark: The negative gradient $-\nabla q$ of a function points in the direction of its steepest decrease, [**9**]. Thus, the solutions to the gradient flow system (9.22) will decrease $q(\mathbf{u})$ as rapidly as possible, ending up at the minimum $\mathbf{u}^\star = \mathbf{0}$. For instance, if $q(u,v)$ represents the height of a hill at position (u,v), then the solutions to (9.22) are the paths of steepest descent followed by, say, water flowing down the hill. In physical applications, the quadratic function (9.21) often represents the energy in the system, and the gradient flow models the natural behavior of systems that seek to minimize their energy.

Example 9.20. Let us solve the first order system

$$\frac{du}{dt} = -8\,u - w, \qquad \frac{dv}{dt} = -8\,v - w, \qquad \frac{dw}{dt} = -u - v - 7\,w,$$

subject to initial conditions

$$u(0) = 1, \qquad v(0) = -3, \qquad w(0) = 2.$$

The coefficient matrix for the system is

$$\begin{pmatrix} -8 & 0 & -1 \\ 0 & -8 & -1 \\ -1 & -1 & -7 \end{pmatrix} = -\begin{pmatrix} 8 & 0 & 1 \\ 0 & 8 & 1 \\ 1 & 1 & 7 \end{pmatrix} = -K,$$

which is minus the positive definite matrix analyzed in Example 8.24. Using the computed eigenvalues and eigenvectors, we conclude that the general solution has the form

$$\mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \\ w(t) \end{pmatrix} = c_1\,e^{-6\,t}\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} + c_2\,e^{-8\,t}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_3\,e^{-9\,t}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

The coefficients are prescribed by the initial conditions, which read

$$\mathbf{u}(0) = \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix} = c_1 \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} + c_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3.$$

Rather than solve this linear system directly, we make use of the fact that the matrix is symmetric, and hence its eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ form an orthogonal basis. Thus, we can apply the orthogonal basis formula (5.8) to compute the coefficients

$$c_1 = \frac{\langle\, \mathbf{u}(0)\,;\mathbf{v}_1\,\rangle}{\|\,\mathbf{v}_1\,\|^2} = \frac{6}{6} = 1, \qquad c_2 = \frac{\langle\, \mathbf{u}(0)\,;\mathbf{v}_2\,\rangle}{\|\,\mathbf{v}_2\,\|^2} = \frac{-4}{2} = -2, \qquad c_3 = \frac{\langle\, \mathbf{u}(0)\,;\mathbf{v}_3\,\rangle}{\|\,\mathbf{v}_3\,\|^2} = 0.$$

We conclude that the solution to the initial value problem is

$$\mathbf{u}(t) = \begin{pmatrix} -e^{-6t} + 2e^{-8t} \\ -e^{-6t} - 2e^{-8t} \\ 2e^{-6t} \end{pmatrix}.$$

In particular, the exponential decay rate is 6 — as indicated by the largest eigenvalue of K — since e^{-6t} is the slowest decaying exponential in the solution.

Extension of the asymptotic stability criterion of Theorem 9.17 to stable equilibria is not difficult.

Theorem 9.21. *A first order linear, homogeneous, constant-coefficient system of ordinary differential equations* (9.1) *has stable zero solution if and only if all the eigenvalues satisfy* $\operatorname{Re}\lambda \le 0$, *and, moreover, any eigenvalue lying on the imaginary axis,* $\operatorname{Re}\lambda = 0$, *is complete, meaning that it has as many independent eigenvectors as its multiplicity.*

Proof: The proof is the same as above, using Corollary 9.14 and the decay properties in Lemma 9.16. All the eigenvalues with negative real part lead to exponentially decaying solutions — even if they are incomplete. If a purely imaginary eigenvalue is complete, then the associated solutions only involve trigonometric functions, and hence remain bounded. This suffices to maintain stability. On the other hand, solutions associated with incomplete purely imaginary eigenvalues contain powers of t multiplying sines and cosines, and hence cannot remain bounded as $t \to \infty$. *Q.E.D.*

Example 9.22. A *Hamiltonian system* in the plane takes the form

$$\frac{du}{dt} = \frac{\partial H}{\partial v}, \qquad \frac{dv}{dt} = -\frac{\partial H}{\partial u}, \tag{9.23}$$

where $H(u, v)$ is known as the *Hamiltonian function*. If

$$H(u, v) = \tfrac{1}{2}\, a\, u^2 + b\, u\, v + \tfrac{1}{2}\, c\, v^2 \tag{9.24}$$

is a quadratic form, then the Hamiltonian system is

$$\dot{u} = b\, u + c\, v, \qquad \dot{v} = -a\, u - b\, v, \tag{9.25}$$

homogeneous, linear with coefficient matrix $A = \begin{pmatrix} b & c \\ -a & -b \end{pmatrix}$. The characteristic equation is

$$\det(A - \lambda \, \mathrm{I}\,) = \lambda^2 + (a\,c - b^2) = 0.$$

If H is positive or negative definite, then $a\,c - b^2 > 0$, and so the roots of the characteristic equation are purely imaginary: $\lambda = \pm \, \mathrm{i} \, \sqrt{a\,c - b^2}$. Since the eigenvalues are simple, the stability criterion of Theorem 9.21 holds and we conclude that planar Hamiltonian systems with definite Hamiltonian function are stable.

Remark: The basic equations of classical mechanics, such as motion of masses under gravitational attraction, can be formulated as Hamiltonian systems. The Hamiltonian function represents the energy. The Hamiltonian formulation is a crucial first step in the physical process of quantizing the classical equations to determine the quantum mechanical equations of motion, [**100**, **104**].

9.3. Two-Dimensional Systems.

The two-dimensional case is particularly instructive, since many of the most important phenomena are already made manifest there. Moreover, the solutions can be easily pictured by their phase portraits. In this section, we will present a complete classification of the possible qualitative behaviors of real, planar linear dynamical systems.

Setting $\mathbf{u}(t) = (u(t), v(t))^T$, such a system $\dot{\mathbf{u}} = A\,\mathbf{u}$ has the explicit form

$$\frac{du}{dt} = a\,u + b\,v, \qquad \frac{dv}{dt} = c\,u + d\,v, \tag{9.26}$$

where $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is the coefficient matrix. As in Section 9.1, we will refer to the (u, v)–plane as the *phase plane*. In particular, phase plane equivalents (9.9) of second order scalar equations form a special class.

According to (8.21), the characteristic equation for the given 2×2 matrix is

$$\det(A - \lambda \, \mathrm{I}\,) = \lambda^2 - \tau \, \lambda + \delta = 0, \tag{9.27}$$

where

$$\tau = \operatorname{tr} A = a + d, \qquad \delta = \det A = a\,d - b\,c, \tag{9.28}$$

are, respectively, the trace and the determinant of A. The nature of the eigenvalues, and hence the solutions, is therefore almost entirely determined by these two quantities. The sign of the *discriminant*

$$\Delta = \tau^2 - 4\,\delta = (\operatorname{tr} A)^2 - 4 \det A = (a - d)^2 - 4\,b\,c \tag{9.29}$$

determines whether the roots or eigenvalues

$$\lambda = \frac{\tau \pm \sqrt{\Delta}}{2} \tag{9.30}$$

are real or complex, and thereby plays a key role in the classification.

Let us summarize the different possibilities as classified by their qualitative behavior. Each situation will be illustrated by a representative phase portrait, which plots several typical solution trajectories in the phase plane. The complete portrait gallery of planar systems can be found in Figure 9.3 below.

Distinct Real Eigenvalues

The coefficient matrix A has two real, distinct eigenvalues $\lambda_1 < \lambda_2$ if and only if the discriminant (9.29) of the quadratic equation (9.27) is positive: $\Delta > 0$. In this case, the solutions take the exponential form

$$\mathbf{u}(t) = c_1\, e^{\lambda_1 t}\, \mathbf{v}_1 + c_2\, e^{\lambda_2 t}\, \mathbf{v}_2, \tag{9.31}$$

where $\mathbf{v}_1, \mathbf{v}_2$ are the eigenvectors and c_1, c_2 are arbitrary constants, to be determined by the initial conditions. The asymptotic behavior of the solutions is governed by the size of the eigenvalues. Let $V_k = \{c\,\mathbf{v}_k\}$, $k = 1, 2$, denote the "eigenlines", i.e., the one-dimensional eigenspaces associated with each eigenvalue λ_k.

There are five qualitatively different cases, depending upon the signs of the two eigenvalues. These are listed by their descriptive name, followed by the required conditions on the discriminant, trace and determinant of the coefficient matrix.

Ia. Stable Node: $\Delta > 0, \qquad \operatorname{tr} A < 0, \qquad \det A > 0.$

If $\lambda_1 < \lambda_2 < 0$ are both negative, then $\mathbf{0}$ is an asymptotically *stable node*. The solutions all tend to $\mathbf{0}$ as $t \to \infty$. Since the first exponential $e^{\lambda_1 t}$ decreases much faster than the second $e^{\lambda_2 t}$, the first term in the solution (9.31) will soon become negligible, and hence $\mathbf{u}(t) \approx c_2\, e^{\lambda_2 t}\, \mathbf{v}_2$ when t is large. Therefore, all solutions with $c_2 \neq 0$ will arrive at the origin along a curve that is tangent to the eigenline V_2. The solutions with $c_2 = 0$ come in to the origin directly along the eigenline V_1, and at a faster rate. Conversely, as $t \to -\infty$, all solutions become unbounded: $\| \mathbf{u}(t) \| \to \infty$. In this case, the first exponential grows faster than the second, and so the solutions $\mathbf{u}(t) \approx c_1\, e^{\lambda_1 t}\, \mathbf{v}_1$ for $t \ll 0$. Thus, as they escape to ∞, the solution trajectories become more and more parallel to the eigenline V_1 — except for those with $c_1 = 0$ that remain in the eigenline V_2.

Ib. Saddle Point: $\Delta > 0, \qquad \det A < 0.$

If $\lambda_1 < 0 < \lambda_2$, then $\mathbf{0}$ is an unstable *saddle point*. Solutions (9.31) with $c_2 = 0$ start out on the eigenline V_1 and go in to $\mathbf{0}$ as $t \to \infty$, while solutions with $c_1 = 0$ start on V_2 and go to $\mathbf{0}$ as $t \to -\infty$. All other solutions become unbounded at both large positive and large negative times. As $t \to +\infty$, the solutions approach the unstable eigenline V_2, while as $t \to -\infty$, they asymptote to the stable eigenline V_1. The eigenline V_1 is called the *stable manifold*, indicating that solutions that start on it eventually go to the equilibrium point $\mathbf{0}$, while V_2 is the *unstable manifold*, meaning that solutions on it go to equilibrium as $t \to -\infty$.

Ic. Unstable Node: $\Delta > 0, \qquad \operatorname{tr} A > 0, \qquad \det A > 0.$

If the eigenvalues $0 < \lambda_1 < \lambda_2$ are both positive, then $\mathbf{0}$ is an *unstable node*. The phase portrait is the same as that of a stable node, but the solution trajectories are traversed in

the opposite direction. Time reversal $t \to -t$ will convert an unstable node into a stable node and vice versa; see Exercise ■. Thus, in the unstable case, the solutions all tend to $\mathbf{0}$ as $t \to -\infty$ and off to ∞ as $t \to \infty$. Except for the solutions on the eigenlines, they asymptote to V_1 as $t \to -\infty$, and become parallel to V_2 as $t \to \infty$.

Id. Stable Line: $\Delta > 0, \qquad \operatorname{tr} A < 0, \qquad \det A = 0.$

If $\lambda_1 < \lambda_2 = 0$, then every point on the eigenline V_2 associated with the zero eigenvalue is an equilibrium point. Every other solution moves along a straight line parallel to V_1 and tends to one of the equilibria on V_2 as $t \to \infty$.

Ie. Unstable Line: $\Delta > 0, \qquad \operatorname{tr} A > 0, \qquad \det A = 0.$

This is merely the time reversal of a stable line. If $0 = \lambda_1 < \lambda_2$, then every point on the eigenline V_1 is an equilibrium. Every other solution moves off to ∞ along a straight line parallel to V_2 as $t \to \infty$, and tends to one of the equilibria on V_1 as $t \to -\infty$.

Complex Conjugate Eigenvalues

The coefficient matrix A has two complex conjugate eigenvalues

$$\lambda = \mu \pm \mathrm{i}\,\nu, \qquad \text{where} \qquad \mu = \tfrac{1}{2}\,\tau = \tfrac{1}{2}\,\operatorname{tr} A, \qquad \nu = \sqrt{-\Delta},$$

if and only if its discriminant is negative: $\Delta < 0$. In this case, the real solutions can be written in the phase–amplitude form (2.7):

$$\mathbf{u}(t) = r\, e^{\mu\, t}\, [\, \cos(\nu t - \sigma)\, \mathbf{w} + \sin(\nu t - \sigma)\, \mathbf{z}\,], \tag{9.32}$$

where $\mathbf{w} \pm \mathrm{i}\,\mathbf{z}$ are the complex eigenvectors. As noted above, the two vectors $\mathbf{w}, \mathbf{z}$ are always linearly independent. The amplitude r and phase shift σ are uniquely prescribed by the initial conditions. There are three subcases, depending upon the sign of the real part μ, or, equivalently, the sign of the trace of A.

IIa. Stable Focus: $\Delta < 0, \qquad \operatorname{tr} A < 0.$

If $\mu < 0$, then $\mathbf{0}$ is an asymptotically *stable focus*. As $t \to \infty$, the solutions all spiral in to $\mathbf{0}$ with "frequency" ν — meaning it takes time $2\pi/\nu$ for the solution to go once around the origin. As $t \to -\infty$, the solutions spiral off to ∞ with the same frequency.

IIb. Center: $\Delta < 0, \qquad \operatorname{tr} A = 0.$

If $\mu = 0$, then $\mathbf{0}$ is a *center*. The solutions all move periodically around elliptical orbits, with common frequency ν and period $2\pi/\nu$. In particular, solutions that start out near $\mathbf{0}$ stay nearby, and hence a center is a stable, but not asymptotically stable, equilibrium.

IIc. Unstable Focus: $\Delta < 0, \qquad \operatorname{tr} A > 0.$

If $\mu > 0$, then $\mathbf{0}$ is an *unstable focus*. The phase portrait is the time reversal, $t \mapsto -t$, of a stable focus, with solutions spiraling off to ∞ as $t \to \infty$ and in to the origin as $t \to -\infty$, again with a common "frequency" ν.

Incomplete Double Real Eigenvalue

The matrix will have a double real eigenvalue $\lambda = \frac{1}{2}\tau = \frac{1}{2}\operatorname{tr} A$ if and only if the discriminant vanishes: $\Delta = 0$. The formula for the solutions depends on whether the eigenvalue λ is complete or not. If λ is an incomplete eigenvalue, admitting only one independent eigenvector $\mathbf{v}$, then the solutions are no longer given by simple exponentials. The general formula is

$$\mathbf{u}(t) = (c_1 + c_2\, t)e^{\lambda t}\,\mathbf{v} + c_2\, e^{\lambda t}\,\mathbf{w}, \tag{9.33}$$

where $(A - \lambda\,\mathrm{I}\,)\mathbf{w} = \mathbf{v}$, and so $\mathbf{v}, \mathbf{w}$ form a Jordan chain for the coefficient matrix. We let $V = \{c\,\mathbf{v}\}$ denote the eigenline associated with the genuine eigenvector $\mathbf{v}$.

IIIa. Stable Improper Node: $\Delta = 0, \qquad \operatorname{tr} A < 0, \qquad A \neq \lambda\,\mathrm{I}\,.$

If $\lambda < 0$ then $\mathbf{0}$ is an asymptotically *stable improper node*. Since $t\,e^{\lambda t}$ is larger than $e^{\lambda t}$ for $t > 1$, the solutions $\mathbf{u}(t) \approx c_2\, t\, e^{\lambda t}$ tend to $\mathbf{0}$ as $t \to \infty$ along a curve that is tangent to the eigenline V. Similarly, as $t \to -\infty$, the solutions go off to ∞, becoming more and more parallel to the eigenline, but moving away in the opposite direction from their approach.

IIIb. Linear Motion: $\Delta = 0, \qquad \operatorname{tr} A = 0, \qquad A \neq \lambda\,\mathrm{I}\,.$

If $\lambda = 0$, then, as in case Id, every point on the eigenline V is an equilibrium point. Every other solution is a linear, or, more correctly, affine function of T, and so moves along a straight line parallel to V, going off to ∞ in either direction. The origin is an unstable equilibrium point.

IIIc. Unstable Improper Node: $\Delta = 0, \qquad \operatorname{tr} A > 0, \qquad A \neq \lambda\,\mathrm{I}\,.$

If $\lambda > 0$, then $\mathbf{0}$ is an *unstable improper node*. The phase portrait is the time reversal of the stable improper node.

Complete Double Real Eigenvalue

In this case, *every* vector in $\mathbb{R}^2$ is an eigenvector, and so the real solutions take the form $\mathbf{u}(t) = e^{\lambda t}\,\mathbf{v}$, where $\mathbf{v}$ is an *arbitrary* constant vector. In fact, this case occurs if and only if $A = \lambda\,\mathrm{I}$ is a multiple of the identity matrix.

IVa. Stable Star: $A = \lambda\,\mathrm{I}\,, \qquad \lambda < 0.$

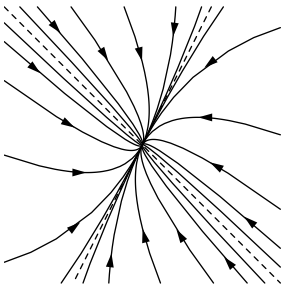
If $\lambda < 0$ then $\mathbf{0}$ is an asymptotically stable star. The solution trajectories are the rays emanating from the origin, and the solutions go to $\mathbf{0}$ at an exponential rate as $t \to \infty$.

IVb. Trivial: $A = \mathrm{O}.$

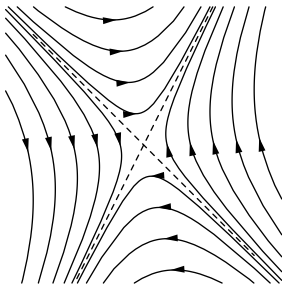
If $\lambda = 0$ then the only possibility is $A = \mathrm{O}$. Now every solution is constant and every point is a (stable) equilibrium point. Nothing happens! This is the only case not pictured in Figure 9.3.

IVc. Unstable Star: $A = \lambda\,\mathrm{I}\,, \qquad \lambda < 0.$

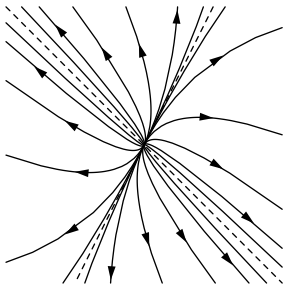
If $\lambda > 0$ then $\mathbf{0}$ is unstable. The phase portrait is the time reversal of the stable star, and so the solutions move along rays, and tend to $\mathbf{0}$ as $t \to -\infty$.



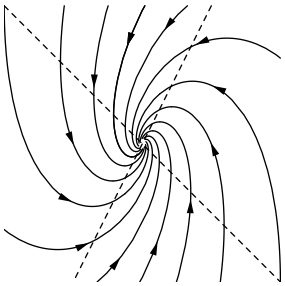
Ia. Stable Node



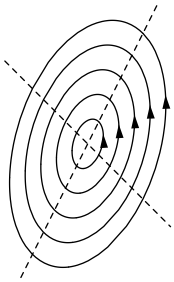
Ib. Saddle Point



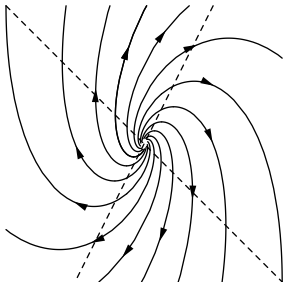
Ic. Unstable Node



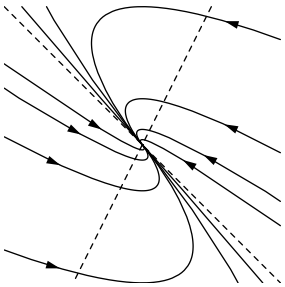
IIa. Stable Focus



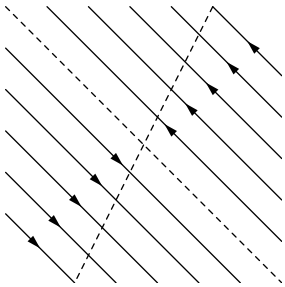
IIb. Center



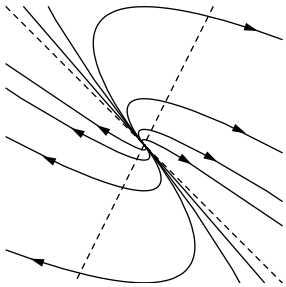
IIc. Unstable Focus



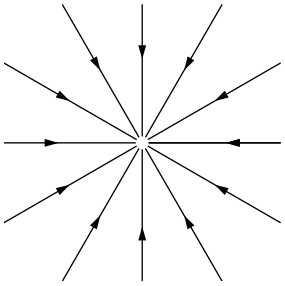
IIIa. Stable Improper Node



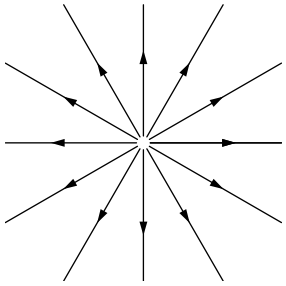
IIIb. Linear Motion



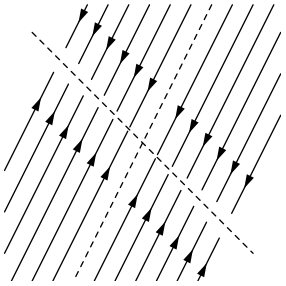
IIIc. Unstable Improper Node



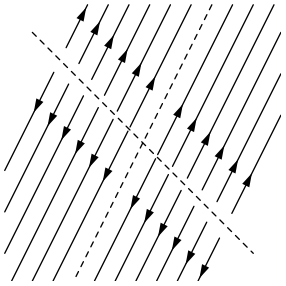
IVa. Stable Star



IVc. Unstable Star



Id. Stable Line



Ie. Unstable Line

Figure 9.3. Phase Portraits.

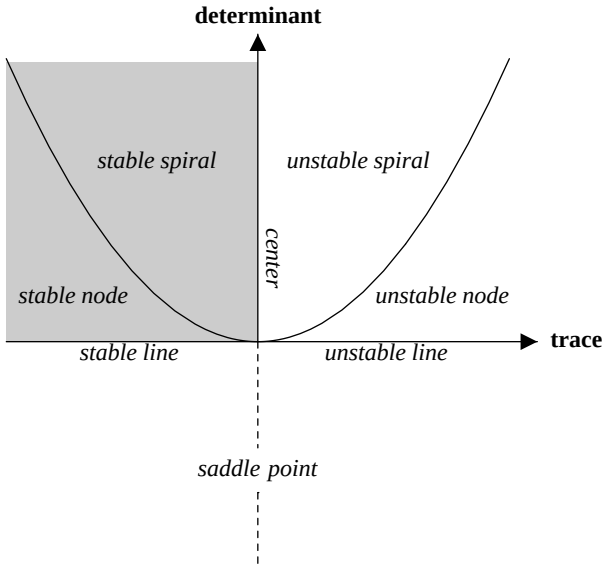


Figure 9.4. Stability Regions for Two–Dimensional Linear Systems.

Figure 9.4 indicates where the different possibilities lie, as prescribed by the trace and determinant of the coefficient matrix. The horizontal axis indicates the value of $\tau = \operatorname{tr} A$, while the vertical axis refers to $\delta = \det A$. Points on the parabola $\tau^2 = 4\,\delta$ represent the cases with vanishing discriminant $\Delta = 0$, and correspond to either stars or improper nodes — except for the origin which is either linear motion or trivial. All the asymptotically stable cases lie in the shaded upper left quadrant where $\operatorname{tr} A < 0$ and $\det A > 0$. The borderline points on the coordinate axes are either stable centers, when $\operatorname{tr} A = 0$, $\det A > 0$, or stable lines, when $\operatorname{tr} A < 0$, $\det A = 0$, or the origin, which may or may not be stable depending upon whether A is the zero matrix or not. All other values for the trace and determinant result in unstable equilibria.

Remark: Time reversal $t \to -t$ changes the sign of the coefficient matrix $A \to -A$, and hence the sign of its trace, $\tau \to -\tau$, while the determinant $\delta = \det A = \det(-A)$ is unchanged. Thus, the effect is to reflect the plot in Figure 9.4 through the vertical axis, interchanging the stable nodes and spirals with their unstable counterparts, while leaving saddle points in the same qualitative form.

In physical applications, the coefficient matrix A is usually not known exactly, and so the physical system may, in fact, be a slight perturbation of the mathematical model. Thus, it is important to know which systems are *structurally stable*, meaning the basic qualitative features are preserved under sufficiently small changes in the coefficients.

Now, a small perturbation will alter the entries of the coefficient matrix slightly, and hence move the trace and determinant by a comparably small amount. The net effect is to slightly perturb its eigenvalues. Therefore, the question of structural stability reduces to whether the eigenvalues have moved sufficiently far to send the system into a different stability regime. Asymptotically stable systems remain stable under small enough perturbations, since the property that the eigenvalues have negative real parts is preserved under small perturbation. For a similar reason, unstable systems remain unstable under small perturbations. On the other hand, a borderline stable system — either a center or the trivial system — could become either asymptotically stable or unstable under an adverse perturbation.

Structural stability requires more, since the overall phase portrait should not significantly change. A system in any of the open regions in the Stability Figure 9.4, e.g., a stable spiral, unstable node, saddle point, etc., is structurally stable, whereas a system that lies on the parabola $\tau^2 = 4\,\delta$, or the horizontal axis, or positive vertical axis, e.g., an improper node, a stable line, etc., is not, since a small perturbation could send it into either of the adjoining regions. In other words, structural stability requires that the eigenvalues be distinct and have non-zero real part: Re $\lambda \neq 0$. This final result also applies to systems in higher dimensions, [**80**].

9.4. Matrix Exponentials.

So far, we have focussed all our attention on vector-valued solutions to linear systems of ordinary differential equations

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}. \tag{9.34}$$

An evident, and, in fact, useful generalization is to look for *matrix solutions*. Specifically, we mean a matrix-valued function $U(t)$ that satisfies the corresponding matrix differential equation

$$\frac{dU}{dt} = A\,U(t). \tag{9.35}$$

As with vectors, the individual entries of $U(t)$ are functions of t, and we differentiate entry-wise. If A is an $n \times n$ matrix, compatibility of matrix multiplication requires that $U(t)$ be of size $n \times k$ for some k.

Since matrix multiplication acts column-wise, the individual columns of the matrix solution must solve the original system, and hence $U(t) = (\,\mathbf{u}_1(t)\,\ldots\,\mathbf{u}_k(t)\,)$ where each column $\mathbf{u}_i(t)$ is a solution to (9.34). Thus, a matrix solution is merely a convenient way of collecting together several different vector solutions to the system.

Example 9.23. According to Example 9.8, the vector-valued functions

$$\mathbf{u}_1(t) = \begin{pmatrix} e^{-4\,t} \\ -2\,e^{-4\,t} \end{pmatrix}, \qquad \mathbf{u}_2(t) = \begin{pmatrix} e^{-t} \\ e^{-t} \end{pmatrix},$$

are both solutions to the linear system

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} \mathbf{u}.$$

They can be combined to form the matrix solution

$$U(t) = \begin{pmatrix} e^{-4\,t} & e^{-t} \\ -2\,e^{-4\,t} & e^{-t} \end{pmatrix} \qquad \text{satisfying} \qquad \frac{dU}{dt} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} U.$$

Indeed, by direct calculation

$$\frac{dU}{dt} = \begin{pmatrix} -4\,e^{-4\,t} & -e^{-t} \\ 8\,e^{-4\,t} & -e^{-t} \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} \begin{pmatrix} e^{-4\,t} & e^{-t} \\ -2\,e^{-4\,t} & e^{-t} \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} U.$$

The existence and uniqueness theorems immediately apply to matrix solutions, and show that there is a unique matrix solution to the system having initial conditions

$$U(t_0) = B, \tag{9.36}$$

where B is a given $n \times k$ matrix. The j^{th} column $\mathbf{u}_j(t)$ of the matrix solution satisfies the initial value problem

$$\frac{d\mathbf{u}_j}{dt} = A\,\mathbf{u}_j, \qquad \mathbf{u}_j(0) = \mathbf{b}_j,$$

where $\mathbf{b}_j$ denotes the j^{th} column of B.

The most important case — beyond vector-valued solutions — is when the matrix solution is square, of size $n \times n$ and so consists of n distinct solutions to the system. In the scalar case, when $n = 1$, the solution to the particular initial value problem

$$\frac{du}{dt} = a\,u, \qquad u(0) = 1,$$

is the ordinary exponential function $u(t) = e^{t\,a}$. With the exponential in hand, the solution to the more general initial value problem $u(t_0) = b$ can then be written as $u(t) = b\,e^{(t-t_0)\,a}$.

Let us introduce a similar initial value problem in the case of linear systems. Now, when dealing with matrices, the role of the number 1 is played by the identity matrix I. Therefore, working by analogy, we are led to define the matrix exponential solution to an n-dimensional linear system.

Definition 9.24. Let A be a square matrix. The *matrix exponential* $U(t) = e^{t\,A}$ is the unique solution to the matrix initial value problem

$$\frac{dU}{dt} = A\,U, \qquad U(0) = \mathrm{I}\,. \tag{9.37}$$

We can rewrite the defining properties (9.37) in the more suggestive form

$$\frac{d}{dt}\,e^{t\,A} = A\,e^{t\,A}, \qquad e^{0\,A} = \mathrm{I}\,. \tag{9.38}$$

Once we know the matrix exponential, we are in a position to solve the general initial value problem. The solution formulae are in direct analogy with the scalar case.

Lemma 9.25. *The solution to the initial value problem*

$$\frac{dU}{dt} = AU, \qquad U(t_0) = B, \qquad \textit{is} \qquad U(t) = e^{(t-t_0)\,A}\,B. \tag{9.39}$$

Proof: Since B is a constant matrix,

$$\frac{dU}{dt} = \frac{d}{dt}\,e^{(t-t_0)\,A}\,B = A\,e^{(t-t_0)\,A}\,B = A\,U,$$

where we used the first defining property in (9.38) along with the chain rule. Thus, $U(t)$ is a matrix solution to the system. Moreover, by the second property in (9.38),

$$U(0) = e^{0\,A}\,B = \mathrm{I}\,B = B$$

has the correct initial conditions. *Q.E.D.*

Remark: The computation used in the proof is a particular instance of the general Leibniz rule

$$\frac{d}{dt}\big[\, M(t)\, N(t)\,\big] = \frac{dM(t)}{dt}\; N(t) + M(t)\; \frac{dN(t)}{dt} \tag{9.40}$$

for the derivative of the product of matrix-valued functions $M(t)$ and $N(t)$ of compatible sizes. The reader is asked to prove this formula in Exercise ■.

In particular, the solution to the vector initial value problem

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}, \qquad \mathbf{u}(t_0) = \mathbf{b},$$

can be written in terms of the matrix exponential:

$$\mathbf{u}(t) = e^{(t-t_0)\,A}\,\mathbf{b}. \tag{9.41}$$

The solution formula (9.39) also gives us a means of computing the matrix exponential. Suppose $U(t)$ is any $n \times n$ matrix solution to the system. Then, by uniqueness, $U(t) = e^{t\,A}\,U(0)$, and hence, provided $U(0)$ is a nonsingular matrix,

$$e^{t\,A} = U(t)\,U(0)^{-1}. \tag{9.42}$$

The condition that $U(0)$ be nonsingular means that the columns of $U(t)$ are linearly independent solutions.

Thus, to construct the exponential of an $n \times n$ matrix A, you first need to find a basis of n linearly independent solutions $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$ to the linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$ using the eigenvalues and eigenvectors, or, in the incomplete case, Jordan chains. The resulting $n \times n$ matrix solution $U(t) = \big(\, \mathbf{u}_1(t) \;\; \dots \;\; \mathbf{u}_n(t) \,\big)$ is, by linear independence, nonsingular at each time t, and hence the formula for $e^{t\,A}$ follows from (9.42).

Example 9.26. For the matrix $A = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix}$ in Example 9.23, we already constructed the nonsingular matrix solution $U(t) = \begin{pmatrix} e^{-4t} & e^{-t} \\ -2e^{-4t} & e^{-t} \end{pmatrix}$ Therefore, by (9.42), the matrix exponential is

$$\begin{aligned} e^{t\,A} &= U(t)\,U(0)^{-1} \\ &= \begin{pmatrix} e^{-4t} & e^{-t} \\ -2e^{-4t} & e^{-t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{3}\,e^{-4t} + \frac{2}{3}\,e^{-t} & -\frac{1}{3}\,e^{-4t} + \frac{1}{3}\,e^{-t} \\ -\frac{2}{3}\,e^{-4t} + \frac{2}{3}\,e^{-t} & \frac{2}{3}\,e^{-4t} + \frac{1}{3}\,e^{-t} \end{pmatrix}. \end{aligned}$$

In particular, we obtain e^A by setting $t = 1$ in this formula:

$$\exp\begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3}\,e^{-4} + \frac{2}{3}\,e^{-1} & -\frac{1}{3}\,e^{-4} + \frac{1}{3}\,e^{-1} \\ -\frac{2}{3}\,e^{-4} + \frac{2}{3}\,e^{-1} & \frac{2}{3}\,e^{-4} + \frac{1}{3}\,e^{-1} \end{pmatrix}.$$

Note that the matrix exponential is *not* obtained by exponentiating the individual matrix entries. Let us use the matrix exponential constructed in Example 9.26 to solve the initial value problem

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix} \mathbf{u}, \qquad \mathbf{u}(0) = \mathbf{b} = \begin{pmatrix} 3 \\ 0 \end{pmatrix}.$$

Employing formula (9.39),

$$\mathbf{u}(t) = e^{t\,A}\,\mathbf{b} = \begin{pmatrix} \frac{1}{3}\,e^{-4\,t} + \frac{2}{3}\,e^{-t} & -\frac{1}{3}\,e^{-4\,t} + \frac{1}{3}\,e^{-t} \\ -\frac{2}{3}\,e^{-4\,t} + \frac{2}{3}\,e^{-t} & \frac{2}{3}\,e^{-4\,t} + \frac{1}{3}\,e^{-t} \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{-4\,t} + 2\,e^{-t} \\ -\,2\,e^{-4\,t} + 2\,e^{-t} \end{pmatrix},$$

reproducing our earlier solution (9.15).

Example 9.27. Suppose $A = \begin{pmatrix} -1 & -2 \\ 2 & -1 \end{pmatrix}$. The characteristic equation $\det(A - \lambda\,\mathrm{I}\,) = \lambda^2 + 2\,\lambda + 5 = 0$ has roots $\lambda = -1 \pm 2\,\mathrm{i}$, which are thus the complex conjugate eigenvalues of A. The corresponding eigenvectors are $\mathbf{v} = (\,\pm\,\mathrm{i}\,, 1\,)^T$, leading to the complex conjugate solutions

$$\mathbf{u}_1(t) = \begin{pmatrix} \mathrm{i}\,e^{(-1+2\,\mathrm{i}\,)\,t} \\ e^{(-1+2\,\mathrm{i}\,)\,t} \end{pmatrix}, \qquad \mathbf{u}_2(t) = \begin{pmatrix} -\,\mathrm{i}\,e^{(-1-2\,\mathrm{i}\,)\,t} \\ e^{(-1-2\,\mathrm{i}\,)\,t} \end{pmatrix}.$$

We assemble these to form the (complex) matrix solution

$$U(t) = \begin{pmatrix} \mathrm{i}\,e^{(-1+2\,\mathrm{i}\,)\,t} & -\,\mathrm{i}\,e^{(-1-2\,\mathrm{i}\,)\,t} \\ e^{(-1+2\,\mathrm{i}\,)\,t} & e^{(-1-2\,\mathrm{i}\,)\,t} \end{pmatrix}.$$

The corresponding matrix exponential is, therefore,

$$e^{t\,A} = U(t)\,U(0)^{-1} = \begin{pmatrix} \mathrm{i}\,e^{(-1+2\,\mathrm{i}\,)\,t} & -\,\mathrm{i}\,e^{(-1-2\,\mathrm{i}\,)\,t} \\ e^{(-1+2\,\mathrm{i}\,)\,t} & e^{(-1-2\,\mathrm{i}\,)\,t} \end{pmatrix} \begin{pmatrix} \mathrm{i} & -\,\mathrm{i} \\ 1 & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} \dfrac{e^{(-1+2\,\mathrm{i}\,)\,t} + e^{(-1-2\,\mathrm{i}\,)\,t}}{2} & \dfrac{-\,e^{(-1+2\,\mathrm{i}\,)\,t} + e^{(-1-2\,\mathrm{i}\,)\,t}}{2\,\mathrm{i}} \\ \dfrac{e^{(-1+2\,\mathrm{i}\,)\,t} - e^{(-1-2\,\mathrm{i}\,)\,t}}{2\,\mathrm{i}} & \dfrac{e^{(-1+2\,\mathrm{i}\,)\,t} + e^{(-1-2\,\mathrm{i}\,)\,t}}{2} \end{pmatrix} = \begin{pmatrix} e^{-t}\cos 2\,t & -\,e^{-t}\sin 2\,t \\ e^{-t}\sin 2\,t & e^{-t}\cos 2\,t \end{pmatrix}.$$

Note that the final expression for the matrix exponential is real, as it must be since A is a real matrix. (See Exercise ∎.) Also note that we didn't need to find the real solutions to construct the matrix exponential — although this would have also worked and given the same result. Indeed, the two columns of $e^{t\,A}$ form a basis for the space of (real) solutions to the linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$.

The matrix exponential turns out to enjoy all the properties you might expect from its scalar counterpart. First, w Let us finish by listing some further important properties of the matrix exponential, all of which are direct analogues of the usual scalar exponential function. First, the multiplicative property survives:

$$e^{(t+s)\,A} = e^{t\,A}\,e^{s\,A}, \qquad \text{for any} \qquad s, t \in \mathbb{R}. \tag{9.43}$$

Indeed, if we differentiate both sides of the equation with respect to t, we find that they both define matrix solutions to the system $\dot{\mathbf{u}} = A\,\mathbf{u}$. Moreover, at $t = 0$ they both have the same initial conditions, and hence, again by uniqueness, they must be the same. In particular, if we set $s = -t$, the left hand side of (9.43) reduces to the identity matrix, and hence

$$e^{-\,t\,A} = \big(\,e^{t\,A}\,\big)^{-1}. \tag{9.44}$$

In particular, this implies that, for any A and any $t \in \mathbb{R}$, the exponential matrix $e^{t\,A}$ is nonsingular.

Warning: On the other hand, in general, for matrices A, B of the same size,

$$e^{t\,(A+B)} \neq e^{t\,A}\, e^{t\,B}. \tag{9.45}$$

Indeed, as we show in Exercise ∎, the left and right hand sides of (9.45) are equal for all t if and only if $A\,B = B\,A$ are commuting matrices.

Finally, we note that the standard exponential series is also valid for the matrix exponential:

$$e^{t\,A} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, A^n = \mathrm{I} + t\,A + \frac{t^2}{2}\, A^2 + \frac{t^3}{6}\, A^3 + \cdots. \tag{9.46}$$

The proof of convergence of the series will be deferred until we have had a chance to discuss matrix norms in Chapter 10. Assuming convergence, the proof that it satisfies the defining initial value problem (9.39) is straightforward:

$$\frac{d}{dt} \sum_{n=0}^{\infty} \frac{t^n}{n!}\, A^n = \sum_{n=1}^{\infty} \frac{t^{n-1}}{(n-1)!}\, A^n = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, A^{n+1} = A \sum_{n=0}^{\infty} \frac{t^n}{n!}\, A^n,$$

while at $t = 0$ the sum collapses to the identity matrix.

Inhomogeneous Linear Systems

We now direct our attention to general inhomogeneous linear systems of ordinary differential equations. For simplicity, we consider only first order† systems of the form

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u} + \mathbf{f}(t), \tag{9.47}$$

where A is a constant $n \times n$ matrix and $\mathbf{f}(t)$ is a vector of functions that represents external forcing to the system. According to our general Theorem 7.37, the solution to the inhomogeneous system will have the form

$$\mathbf{u}(t) = \mathbf{u}^\star(t) + \mathbf{z}(t)$$

where $\mathbf{u}^\star(t)$ is a particular solution and $\mathbf{z}(t)$ is a general solution to the homogeneous system (8.9). Physically, one interprets the solution as a combination of a particular response to the external forcing coupled with the system's own internal motion.

Since we already know how to find the solution $\mathbf{z}(t)$ to the homogeneous system, the only task is to determine one particular solution to the inhomogeneous system. The method used to construct the solution is known as *variation of parameters*, and will work even when the matrix A depends on t. The student may have encountered the scalar version of this method in a first course on ordinary differential equations, and the same basic idea applies to systems.

† Higher order systems can, as remarked earlier, always be converted into first order systems involving additional variables.

Recall that, in the scalar case, to solve the inhomogeneous equation

$$\frac{du}{dt} = a\,u + f(t), \qquad \text{we set} \qquad u(t) = e^{t\,a}\,v(t).$$

Thus, by the product rule,

$$\frac{du}{dt} = a\,e^{t\,a}\,v(t) + e^{t\,a}\,\frac{dv}{dt} = a\,u + e^{t\,a}\,\frac{dv}{dt}\,,$$

and so $v(t)$ solves the differential equation

$$\frac{dv}{dt} = e^{-\,t\,a}\,f(t).$$

The latter equation can be solved by direct integration.

A entirely analogous method works in the vector case. We replace the scalar exponential by the exponential of the coefficient matrix, setting

$$\mathbf{u}(t) = e^{t\,A}\,\mathbf{v}(t).$$

Using the product rule for matrix multiplication and the fourth property in (9.38), we find

$$\frac{d\mathbf{u}}{dt} = A\,e^{t\,A}\,\mathbf{v}(t) + e^{t\,A}\,\frac{d\mathbf{v}}{dt} = A\,\mathbf{u} + e^{t\,A}\,\frac{d\mathbf{v}}{dt}\,.$$

We conclude that

$$\frac{d\mathbf{v}}{dt} = e^{-\,t\,A}\,\mathbf{f}(t).$$

The resulting differential equation can then be solved by direct integration:

$$\mathbf{v}(t) = \mathbf{v}(t_0) + \int_{t_0}^{t} e^{-\,s\,A}\,\mathbf{f}(s)\,ds, \tag{9.48}$$

Substituting the formula for $\mathbf{u}(t)$, we have therefore established the following general formula for the solution to an first order, inhomogeneous linear system with constant coefficient matrix.

Theorem 9.28. *The solution to the initial value problem* $\dfrac{d\mathbf{u}}{dt} = A\,\mathbf{u} + \mathbf{f}(t),\ \mathbf{u}(t_0) = \mathbf{b},$ *is*

$$\mathbf{u}(t) = e^{(t-t_0)\,A}\,\mathbf{b} + \int_{t_0}^{t} e^{(t-s)\,A}\,\mathbf{f}(s)\,ds. \tag{9.49}$$

Example 9.29. Our goal is to solve the initial value problem

$$\begin{aligned} \dot{u}_1 &= 2u_1 - u_2, & u_1(0) &= 1, \\ \dot{u}_2 &= 4u_1 - 3u_2 + e^t & u_2(0) &= 0. \end{aligned} \tag{9.50}$$

The eigenvalues and eigenvectors of the coefficient matrix $A = \begin{pmatrix} 2 & -1 \\ 4 & -3 \end{pmatrix}$ are

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = -2, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

We use these to form the nonsingular matrix solution $U(t) = \begin{pmatrix} e^t & e^{-2t} \\ e^t & 4e^{-2t} \end{pmatrix}$, whence

$$e^{tA} = U(t)\,U(0)^{-1} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{4}{3}e^{-2t} \end{pmatrix}.$$

We can compute the solution directly from formula (9.49):

$$\begin{aligned} \int_0^t e^{(t-s)\,A}\,\mathbf{f}(s)\,ds &= \int_0^t \begin{pmatrix} \frac{4}{3}e^{t-s} - \frac{1}{3}e^{-2(t-s)} & -\frac{1}{3}e^{t-s} + \frac{1}{3}e^{-2(t-s)} \\ \frac{4}{3}e^{t-s} - \frac{4}{3}e^{-2(t-s)} & -\frac{1}{3}e^{t-s} + \frac{4}{3}e^{-2(t-s)} \end{pmatrix} \begin{pmatrix} 0 \\ e^s \end{pmatrix} ds \\ &= \begin{pmatrix} \displaystyle\int_0^t -\tfrac{1}{3}e^t + \tfrac{1}{3}e^{-2t+3s}\,ds \\ \displaystyle\int_0^t -\tfrac{1}{3}e^t + \tfrac{4}{3}e^{-2t+3s)}\,ds \end{pmatrix} = \begin{pmatrix} -\frac{1}{3}t\,e^t + \frac{1}{9}(e^t - 1) \\ -\frac{1}{3}t\,e^t + \frac{4}{9}(e^t - 1) \end{pmatrix}. \end{aligned}$$

This is the particular solution for the homogeneous initial conditions $\mathbf{u}(0) = \mathbf{0}$. To obtain the solution that satisfies the given initial conditions, we compute the first term in (9.49)

$$e^{tA}\,\mathbf{b} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{4}{3}e^{-2t} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} \end{pmatrix},$$

which is the solution to the homogeneous system for the given nonzero initial conditions. We sum these two to finally obtain the solution to our initial value problem:

$$\mathbf{u}(t) = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} - \frac{1}{3}t\,e^t + \frac{1}{9}(e^t - 1) \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} - \frac{1}{3}t\,e^t + \frac{4}{9}(e^t - 1) \end{pmatrix}.$$

Applications in Geometry

The connection between a matrix and its exponential plays an important role in geometry, group theory and, eventually, [**32**, **117**], the symmetry analysis of differential equations.

Let A be an $n \times n$ matrix. For each $t \in \mathbb{R}$, the corresponding matrix exponential e^{tA} is itself an $n \times n$ matrix and thus defines a linear transformation on the vector space $\mathbb{R}^n$:

$$L_t[\mathbf{x}] = e^{tA}\,\mathbf{x} \qquad \text{for} \qquad \mathbf{x} \in \mathbb{R}^n.$$

The resulting family of linear transformations, parametrized by $t \in \mathbb{R}$, obeys the following properties

$$L_t \circ L_s = L_{t+s}, \qquad L_0 = \mathrm{I}\,, \qquad L_{-t} = L_t^{-1}, \tag{9.51}$$

which are merely restatements of three of the basic matrix exponential properties listed in (9.38), (9.44). In geometric terms, the transformations $L_t = e^{tA}$ are said to form *one-parameter group*, [**117**]. The matrix A is often referred to as the *infinitesimal generator* of the one-parameter group. Indeed, by the series formula (9.38) for the matrix exponential,

$$L_t[\mathbf{x}] = e^{tA}\,\mathbf{x} = \left(\,\mathrm{I} + t\,A + \tfrac{1}{2}t^2\,A^2 + \,\cdots\,\right)\mathbf{x} = \mathbf{x} + t\,A\,\mathbf{x} + \,\cdots\,,$$

and so the linear approximation to the group transformations is to move straight in the direction $A\,\mathbf{x}$. This line is the tangent line to the nonlinear group motion.

Most of the interesting linear transformations of importance in geometry, computer graphics and animation arise in this fashion. Let us consider a few basic examples.

(*a*) When $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ then $e^{t\,A} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$ is a rotation matrix. Thus, the infinitesimal generator of the one-parameter group of plane rotations is the simplest skew-symmetric matrix.

(*b*) When $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ then $e^{t\,A} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ represents a shearing transformation.

(*c*) When $A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ then $e^{t\,A} = \begin{pmatrix} e^t & 0 \\ 0 & e^t \end{pmatrix}$ represents a scaling or stretching transformation.

(*d*) When $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ then $e^{t\,A} = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$ represents a stretch in the x direction and a contraction in the y direction.

Remark: Remember that in chaptereq, when we discussed the motions of planar and space structures, we were unable to handle the true geometric motions of the ends of a bar (rotations, mechanisms, etc.) in our linear systems framework, and ended up using a linear approximation. What we were doing was, in fact, replacing the one-parameter group transformations by their infinitesimal approximations.

More generally, rotations in three and higher dimensions are also generated by skew-symmetric matrices.

Lemma 9.30. *If $A^T = -A$ is a skew-symmetric matrix, then $Q(t) = e^{t\,A}$ is a proper orthogonal matrix.*

Proof: According to equation (9.44) and Exercise ∎,

$$Q(t)^{-1} = e^{-t\,A} = e^{t\,A^T} = \left(e^{t\,A}\right)^T = Q(t)^T,$$

which proves orthogonality. Properness follow from Lemma 9.31. *Q.E.D.*

Thus, the $\frac{1}{2}\,n(n-1)$ dimensional vector space of $n \times n$ skew symmetric matrices generates the group of rotations in n-dimensional Euclidean space. In the three-dimensional case, every skew-symmetric 3×3 matrix has zero determinant: $\det A = 0$, and hence admits a null eigenvector $\mathbf{v} \in \ker A$. We claim that $e^{t\,A}$ represents the group of rotations around the eigenvector axis $\mathbf{v}$. For instance, if $A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$, then $\mathbf{v} = \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and $e^{t\,A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}$ represents the rotation around the $\mathbf{e}_1$ or x axis. The set of all skew-symmetric matrices forms a three-dimensional vector space, with basis

$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, corresponding to rotations around the x, y and z axes. Any other skew-symmetric matrix can be written as a linear combination of these three, indicating that the full group of rotations is built up from these three basic types. This quantifies our earlier observations about the umber of independent infinitesimal rigid motions in 2 and three-dimensional space.

Another important result gives a simple formula for the determinant of a matrix exponential in terms of the trace of the generating matrix.

Lemma 9.31. *For any square matrix,* $\det e^{t\,A} = e^{t\,\mathrm{tr}\,A} > 0$.

Proof: According to Exercise ■, if A has eigenvalues $\lambda_1, \dots, \lambda_n$, then $e^{t\,A}$ has eigenvalues[†] $e^{t\,\lambda_1}, \dots, e^{t\,\lambda_n}$ are the corresponding eigenvalues of A. Moreover, using (8.26) the determinant of $\det e^{t\,A}$ is the produuct of its eigenvalues, so

$$\det e^{t\,A} = e^{t\,\lambda_1}\, e^{t\,\lambda_2} \,\cdots\, e^{t\,\lambda_n} = e^{t\,(\lambda_1 + \lambda_2 + \,\cdots\, + \lambda_n)} = e^{t\,\mathrm{tr}\,A},$$

where, by (8.25), we identify the sum of its eigenvalues as the trace of A. *Q.E.D.*

In particular, if $\mathrm{tr}\,A = 0$, its exponential has unit determinant, $\det e^{t\,A} = 1$, and so represents a family of area- or volume-preserving linear transformations. In the two-dimensional examples listed above, the rotations, shears and combined contraction/stretch are all area preserving, while the pure stretches expand areas by a uniform factor $e^{2\,t}$. Thus, if we start with a unit square, under a rotation is remains a unit square. Under a shear it becomes a parallelogram with unit area. Under a stretch it becomes a sqaure with side lengths e^t and hence total area $e^{2\,t}$, while under the combined contraction/stretch it becomes a unit rectangle. Similar remarks hold for a unit circular disk; in the last case it is transformed into an ellipse of area 1.

Noncommutativity of linear transformations is refelcted in noncommutativity of their infinitesimal generators. Recall, (commutator■), that the *commutator* of two $n \times n$ matrices A, B is

$$[\,A, B\,] = A\,B - B\,A.$$

Thus, A and B commute if and only if $[\,A, B\,] = \mathrm{O}$. We use the exponential series (9.46) to evaluate the commutator of the matrix exponentials $e^{t\,A}$ and $e^{t\,B}$:

$$\begin{aligned} \big[\, e^{t\,A}, e^{t\,B} \,\big] &= e^{t\,A}\, e^{t\,B} - e^{t\,B}\, e^{t\,A} \\ &= \big(\, \mathrm{I} + t\,A + \,\cdots\, \big)\big(\, \mathrm{I} + t\,B + \,\cdots\, \big) - \big(\, \mathrm{I} + t\,B + \,\cdots\, \big)\big(\, \mathrm{I} + t\,A + \,\cdots\, \big) \\ &= t\,(A\,B - B\,A) + \,\cdots\, = t\,[\,A, B\,] + \,\cdots\,. \end{aligned}$$

Therefore, for small values of t, the commutator between the two one-parameter groups is governed by the commutator of their generators. In particular, if the groups commute, then $[\,A, B\,] = \mathrm{O}$; the converse is also true, as follows easily from the previous computation.

[†] In Exercise ■ you prove that repeated eigenvalues have the same multiplicities.

Proposition 9.32. *The matrix exponentials $e^{t\,A}$ and $e^{t\,B}$ commute if and only if $A\,B = B\,A$.*

For instance, the non-commutativity of three-dimensional rotations follows from the non-commutativity of their infinitesimal skew-symmetric generators. For instance, $X = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$ generates (righthanded) rotations around the x axis, while $Y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}$ generates (righthanded) rotations around the y axis. Their commutator

$$\begin{aligned}\big[\,X,Y\,\big] &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &= Z\end{aligned}$$

is the generator of rotations around the z axis. Hence, in a linear approximation, the difference between x and y rotations is, interestingly, a z rotation.

9.5. Dynamics of Structures.

Chapter 6 was concerned with the equilibrium configurations of mass-spring chains and, more generally, structures made out of elastic bars. We are now able to analyze the dynamical motions of such structures. Consider first a linear mass/spring chain consisting of n masses $m_1, \dots, m_n$ connected together and, possibly, to the top and bottom supports by springs. Let $u_i(t)$ denote the displacement[†] from equilibrium of the i^{th} mass, and $e_j(t)$ the elongation of the j^{th} spring. Since we are now interested in dynamics, both of these are allowed to depend on time, t.

The motion of each mass is governed by Newton's Second Law,

$$\text{Force} = \text{Mass} \times \text{Acceleration}. \tag{9.52}$$

The acceleration of the i^{th} mass is the second derivative $\ddot{u}_i = d^2u_i/dt^2$ of its displacement. The right hand sides of Newton's Law are thus $m_i\,\ddot{u}_i$, which we collect together in vector form $M\,\ddot{\mathbf{u}}$ by multiplying the second derivative of the displacement vector $\mathbf{u}(t) = (\,u_1(t), \dots, u_n(t)\,)^T$ by the diagonal, positive definite mass matrix $M = \operatorname{diag}(m_1, \dots, m_n)$. Incidentally, the masses of the springs are assumed to be negligible in this approximation.

If, to begin with, we assume no external forcing of the mass/spring system and no frictional effects, then the only force exerted on each mass is due to the elongations of its two connecting springs, which is measured by the components of the internal force vector

$$\mathbf{f} = -\,K\,\mathbf{u} = -\,A^T C\,A\,\mathbf{u}. \tag{9.53}$$

Here $K = A^T C\,A$ the stiffness matrix for the chain, which is constructed from the (reduced) incidence matrix A and the diagonal matrix of spring constants C, as in (6.11).

[†] As in Section 6.1, the masses are only allowed to move in the direction of the chain, that is, we restrict our attention to one-dimensional motion.

Substituting the internal force formula (9.53) into Newton's Law (9.52) leads immediately to the fundamental dynamical equations

$$M \frac{d^2\mathbf{u}}{dt^2} = -K\mathbf{u} \tag{9.54}$$

governing the free, frictionless motions of the system. The goal is to understand the solutions of this system of second order ordinary differential equations, and then, rather straightforwardly, generalize the methods to cover structures in two and three dimensions as well as electrical circuits containing inductors, resistors and capacitors, all of which are governed by the same basic second order system (9.54) based on the appropriate stiffness or resistivity matrix K.

Example 9.33. The simplest case is that of a single mass connected to a fixed support by a spring. The dynamical system (9.54) reduces to a scalar equation

$$m \frac{d^2u}{dt^2} + k\,u = 0. \tag{9.55}$$

Here $m > 0$ is the mass, while $k > 0$ is the spring's stiffness. The general solution to this elementary homogeneous, second order linear ordinary differential equation is

$$u(t) = c_1 \cos\omega\, t + c_2 \sin\omega\, t = r\cos(\omega\, t - \delta), \qquad \text{where} \qquad \omega = \sqrt{\frac{k}{m}} \tag{9.56}$$

is the natural frequency of vibration. We have used the phase-amplitude equation (2.7) to rewrite the solution as a single cosine with an amplitude $r = \sqrt{c_1^2 + c_2^2}$, and phase lag $\delta = \tan^{-1} c_2/c_1$. The motion is periodic, with period $P = 2\,\pi/\omega$. The frequency formula $\omega = \sqrt{k/m}$ tells us that stiffer the spring or the lighter the mass, the faster the vibrations. Take note of the square root; it tells us that, for instance, quadrupling the mass only slows down the vibrations by a factor of two.

The constants c_1, c_2 — or their phase-amplitude counterparts r, δ — are determined by the initial conditions. Physically, we need to specify both an initial position and an initial velocity in order to uniquely prescribe the subsequent motion of the system:

$$u(t_0) = a, \qquad \dot{u}(t_0) = b. \tag{9.57}$$

The resulting solution is most conveniently written in the form

$$u(t) = a\cos\omega\,(t - t_0) + \frac{b}{\omega}\,\sin\omega\,(t - t_0) = r\cos\big[\,\omega\,(t - t_0) - \delta\,\big] \tag{9.58}$$

which has amplitude and phase given by

$$r = \sqrt{a^2 + \frac{\beta^2}{\omega^2}}\,, \qquad \delta = \tan^{-1}\frac{b}{a\,\omega}\,. \tag{9.59}$$

A typical solution is plotted in Figure 9.5.

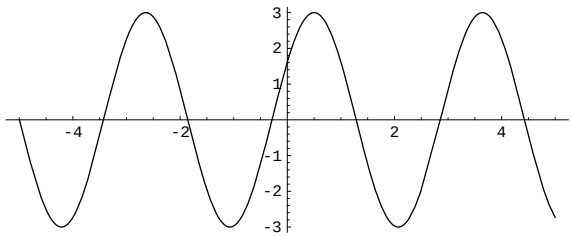


Figure 9.5. Vibration of a Mass.

Let us turn to a more general mass-spring chain or structure. Just as exponentials form the basic building blocks for the solution of systems of first order ordinary differential equations, trigonometric functions form the basic building blocks for solutions to undamped mechanical (and electrical) vibrations governed by second order systems. For simplicity, let us first assume that the masses are all the same and equal to 1 (in some appropriate units), so that (9.54) reduces to

$$\frac{d^2\mathbf{u}}{dt^2} = -K\,\mathbf{u}. \tag{9.60}$$

Mimicking our success in the first order case, let us try substituting the trigonometric ansatz

$$\mathbf{u}(t) = \cos(\omega\, t)\,\mathbf{v}, \tag{9.61}$$

with $\mathbf{v} \neq \mathbf{0}$ denoting a constant vector, into the system (9.60). Differentiating (9.61) directly, we find

$$\frac{d\mathbf{u}}{dt} = -\,\omega\,\sin(\omega\, t)\,\mathbf{v}, \qquad \frac{d^2\mathbf{u}}{dt^2} = -\,\omega^2\,\cos(\omega\, t)\,\mathbf{v}.$$

Therefore, our ansatz (9.61) will solve (9.60) if and only if

$$K\,\mathbf{v} = \omega^2\,\mathbf{v},$$

which means that $\mathbf{v}$ is an eigenvector of K with eigenvalue

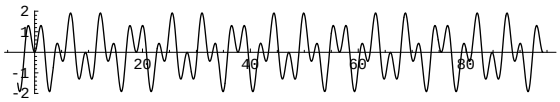
$$\lambda = \omega^2. \tag{9.62}$$

Now, there is nothing special about the cosine function — the same computation also applies to the sine function, and tells us that $\mathbf{u}(t) = \sin(\omega\, t)\,\mathbf{v}$ is also a solution whenever $\mathbf{v}$ is an eigenvector with eigenvalue $\lambda = \omega^2$. Summarizing:

Lemma 9.34. *If $\mathbf{v}$ is an eigenvector of the matrix K with eigenvalue $\lambda = \omega^2$, then the trigonometric vector functions $\mathbf{u}(t) = \cos(\omega\, t)\,\mathbf{v}$ and $\mathbf{u}(t) = \sin(\omega\, t)\,\mathbf{v}$ are solutions to the second order system $\ddot{\mathbf{u}} = -\,K\,\mathbf{u}$.*

Remark: Alternatively, one can utilize the complex exponential solutions $e^{\,\mathrm{i}\,\omega\, t}\,\mathbf{v}$ and $e^{-\,\mathrm{i}\,\omega\, t}\,\mathbf{v}$, which are related to the trigonometric solutions via Euler's formula (3.76). This is common practice in electrical circuit analysis — although electrical engineers tend to use j instead of i to denote the square root of -1.

$\cos t + \cos \sqrt{5}\, t$



$\cos t + \cos \frac{7}{3}\, t$

Figure 9.6. Quasi–Periodic and Periodic Functions.

Stable Structures

Let us next analyze the motion of a stable structure, of the type introduced in Section 6.3. According to Theorem 6.8, stability requires that the reduced stiffness matrix be positive definite: $K > 0$. Theorem 8.25 says that all the eigenvalues of K are strictly positive, $\lambda_i > 0$, which is good, since it implies that the eigenvalue/frequency relation (9.62) yields real frequencies $\omega_i = \sqrt{\lambda_i}$. Moreover, all positive definite matrices are complete, and so, even when there are fewer than n different eigenvalues, there always exist a complete system of n linearly independent real eigenvectors that form an orthogonal basis for $\mathbb{R}^n$.

Since (9.60) is a second order system of homogeneous linear equations in n unknowns, we require $2n$ linearly independent solutions. Lemma 9.34 produces 2 independent solutions for each positive eigenvalue (counted with multiplicity), and hence, assuming positive definiteness, there are indeed $2n$ linearly independent solutions,

$$\begin{aligned} \mathbf{u}_i(t) &= \cos(\omega_i t)\, \mathbf{v}_i = \cos\big(\sqrt{\lambda_i}\, t\big)\, \mathbf{v}_i, \\ \widetilde{\mathbf{u}}_i(t) &= \sin(\omega_i t)\, \mathbf{v}_i = \sin\big(\sqrt{\lambda_i}\, t\big)\, \mathbf{v}_i, \end{aligned} \qquad i = 1, \dots, n, \tag{9.63}$$

governed by the n mutually orthogonal (or even orthonormal) eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ of K. The general solution to (9.60) is an arbitrary linear combination,

$$\mathbf{u}(t) = \sum_{i=1}^{n} \big[\, c_i \cos \omega_i t + d_i \sin \omega_i t \,\big]\, \mathbf{v}_i = \sum_{i=1}^{n} r_i \cos(\omega_i t - \delta_i)\, \mathbf{v}_i, \tag{9.64}$$

of these $2n$ basic solutions. The $2n$ coefficients c_i, d_i — or their phase–amplitude counterparts $r_i > 0$, and $0 \le \delta_i < 2\pi$ — are uniquely determined by the initial conditions. As in (9.57), we need to specify both the initial positions and initial velocities of all the masses; this requires a total of $2n$ initial conditions

$$\mathbf{u}(t_0) = \mathbf{a}, \qquad \dot{\mathbf{u}}(t_0) = \mathbf{b}. \tag{9.65}$$

The individual solutions (9.63) are known as the *normal modes of vibration* of our system, and the $\omega_i = \sqrt{\lambda_i}$ the *normal frequencies*, which are the *square roots of the eigenvalues of the stiffness matrix*. Each is a periodic, vector-valued function of period $P_i = 2\pi/\omega_i$. Linear combinations of such periodic functions are, in general, called *quasi-periodic*. Unless the ratios ω_i/ω_j between the frequencies are all rational numbers, such a quasi-periodic function will never precisely repeat itself, and so can appear to be chaotic, even though it is built up from a few very simple periodic constituents. The reader will find it very instructive to graph some simple quasiperiodic functions, say

$$f(t) = c_1 \cos t + c_2 \cos \sqrt{5}\, t$$

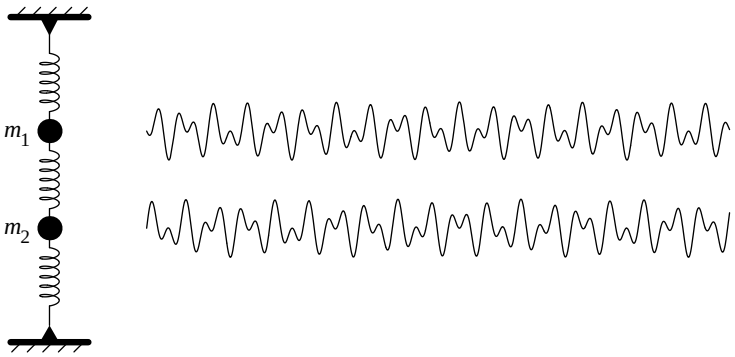


Figure 9.7. Motion of a Double Mass/Spring Chain with Fixed Supports.

for various values of c_1, c_2. Comparison with a case where the frequencies are all rational, say

$$f(t) = c_1 \cos t + c_2 \cos \tfrac{7}{3}\, t$$

is also instructive. The former is truly quasiperiodic, while the the latter is, in fact, periodic with period 6π. Most structures and circuits exhibit quasi-periodic vibrational motions. Let us analyze a couple of simple examples.

Example 9.35. Consider a chain consisting of two equal unit masses connected in a row to supports by three springs, as in Figure 9.7. If the spring constants are c_1, c_2, c_3 (from top to bottom), then the stiffness matrix is

$$K = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & c_3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 + c_3 \end{pmatrix}$$

The eigenvalues and eigenvectors of K will prescribe the normal modes of vibration and natural frequencies of our two–mass chain.

. Let us look in detail at the case of identical springs, and choose our units so that $c_1 = c_2 = c_3 = 1$. Then $K = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ has eigenvalues and eigenvectors

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = 3, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The general solution to the system is then

$$\mathbf{u}(t) = r_1 \cos(t - \delta_1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + r_2 \cos(\sqrt{3}\, t - \delta_2) \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The first summand is the normal mode vibrating at the relatively slow frequency $\omega_1 = 1$, with the two masses moving in tandem. The second summand is the normal mode that vibrates faster, with frequency $\omega_2 = \sqrt{3}$, in which the two masses move in opposing directions. The general motion is a linear combination of these two normal modes. Since the frequency ratio $\omega_2/\omega_1 = \sqrt{2}$ is irrational, the motion is quasi-periodic. The system

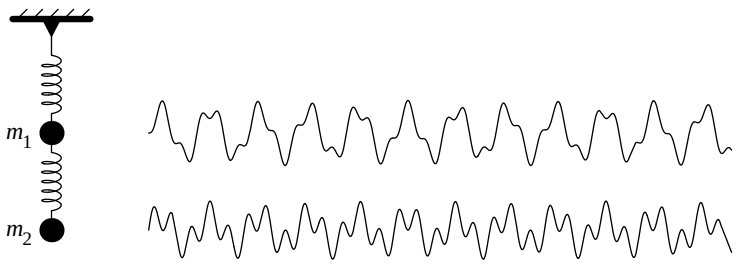


Figure 9.8. Motion of a Double Mass/Spring Chain with One Free End.

never quite returns to its initial configuration — unless it happens to be vibrating in only one of the normal modes. A graph of some typical displacements of the masses is plotted in Figure 9.7.

If we eliminate the bottom spring, so the masses are just hanging from the top support as in Figure 9.8, then the reduced incidence matrix $A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ loses its last row. Assuming that the springs have unit stiffnesses $c_1 = c_2 = 1$, the corresponding stiffness matrix is

$$K = A^T A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors are

$$\lambda_1 = \frac{3-\sqrt{5}}{2}, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ \frac{1+\sqrt{5}}{2} \end{pmatrix}, \qquad \lambda_2 = \frac{3+\sqrt{5}}{2}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ \frac{1-\sqrt{5}}{2} \end{pmatrix}.$$

The general solution to the system is then

$$\mathbf{u}(t) = r_1 \cos\left(\sqrt{\tfrac{3-\sqrt{5}}{2}}\, t - \delta_1\right) \begin{pmatrix} 1 \\ \frac{1+\sqrt{5}}{2} \end{pmatrix} + r_2 \cos\left(\sqrt{\tfrac{3+\sqrt{5}}{2}}\, t - \delta_2\right) \begin{pmatrix} 1 \\ \frac{1-\sqrt{5}}{2} \end{pmatrix}.$$

The slower normal mode, with frequency $\omega_1 = \sqrt{\frac{3-\sqrt{5}}{2}}$, has the masses moving in tandem, with the bottom mass moving proportionally $\dfrac{1+\sqrt{5}}{2}$ farther. The faster normal mode, with frequency $\omega_2 = \sqrt{\frac{3+\sqrt{5}}{2}}$, has the masses moving in opposite directions, with the top mass experiencing the larger displacement. Moreover, both modes vibrate slower than when there is a bottom support. A typical solution is plotted in Figure 9.8.

Example 9.36. Consider a three mass/spring chain, with unit springs and masses, and both ends attached to fixed supports. The stiffness matrix $K = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$

has eigenvalues and eigenvectors

$$\lambda_1 = 2 - \sqrt{2}\,, \qquad \lambda_2 = 2, \qquad \lambda_3 = 2 + \sqrt{2}\,,$$
$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}.$$

The three normal modes, from slowest to fastest, have frequencies

(a) $\omega_1 = \sqrt{2 - \sqrt{2}}$: all three masses move in tandem, with the middle one moving $\sqrt{2}$ times as far.

(b) $\omega_2 = \sqrt{2}$: the two outer masses move in opposing directions, while the middle mass does not move.

(c) $\omega_3 = \sqrt{2 + \sqrt{2}}$: the two outer masses move in tandem, while the inner mass moves $\sqrt{2}$ times as far in the opposite direction.

The general motion is a quasi-periodic combination of these three normal modes. As such, to the naked eye it can look very complicated. Our mathematical analysis unmasks the innate simplicity, where the complex dynamics are, in fact, entirely governed by just three fundamental modes of vibration.

Unstable Structures

So far, we have just dealt with the stable case, when the reduced incidence matrix has trivial kernel, $\ker A = \{\mathbf{0}\}$, and so the stiffness matrix $K = A^T C A$ is positive definite. Unstable configurations, which can admit rigid motions and/or mechanisms, will provide additional complications. The simplest version is a single mass that is not attached to any spring. The mass experiences no restraining force, and has motion governed by the elementary second order ordinary differential equation

$$m\,\frac{d^2u}{dt^2} = 0. \tag{9.66}$$

The general solution

$$u(t) = c\,t + d \tag{9.67}$$

has the mass either sitting still at a specified position or moving in a straight line with constant velocity $c \neq 0$.

More generally, suppose that the stiffness matrix $K = A^T C\,A$ for our structure is only positive semi-definite. Each vector $\mathbf{0} \neq \mathbf{v} \in \ker A = \ker K$ represents a mode of instability of the system. Since $K\mathbf{v} = \mathbf{0}$, we can interpret $\mathbf{v}$ as a *null eigenvector* of K, with eigenvalue $\lambda = 0$. Lemma 9.34 gives us two solutions to the dynamical equations (9.60) with associated "frequency" $\omega = \sqrt{\lambda} = 0$. The first, $\mathbf{u}(t) = \cos(\omega\,t)\,\mathbf{v} = \mathbf{v}$ is a constant solution, i.e., an equilibrium configuration of the system. Thus, an unstable system does not have a unique equilibrium configuration, since every null eigenvector $\mathbf{v} \in \ker K$ gives a constant solution. On the other hand, the second solution, $\mathbf{u}(t) = \sin(\omega\,t)\,\mathbf{v} = \mathbf{0}$, is trivial, and of no help for constructing the general solution. But, to obtain the general solution to the system, we

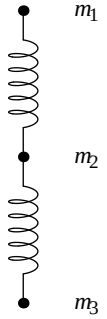


Figure 9.9. A Triatomic Molecule.

still need a second independent solution coming from the null eigenvalue. In analogy with the scalar case (9.67), let us try the solution ansatz $\mathbf{u}(t) = t\,\mathbf{v}$, which works, since

$$\frac{d\mathbf{u}}{dt} = \mathbf{v}, \qquad \mathbf{0} = \frac{d^2\mathbf{u}}{dt^2} = K\,\mathbf{u} = t\,K\,\mathbf{v}.$$

Therefore, to each element of the kernel of the stiffness matrix — i.e., each rigid motion and mechanism — there is a two-dimensional family of solutions

$$\mathbf{u}(t) = (c\,t + d)\,\mathbf{v}. \tag{9.68}$$

When $c = 0$, it reduces to a constant equilibrium solution; when $c \neq 0$, the solution is moving with constant velocity in the null direction $\mathbf{v}$ representing an unstable mode in the system. The general solution will be a linear superposition of the vibrational modes corresponding to the positive eigenvalues and these unstable linear motions corresponding to the zero eigenvalues.

Remark: If the null direction $\mathbf{v}$ represents a rigid translation, then the entire structure will move in that direction. If $\mathbf{v}$ represents an infinitesimal rotation, then, owing to our linear approximation to the true nonlinear bar motions, the individual masses will move in straight lines, which are the tangent approximations to the circular motion that occurs in the true physical, nonlinear regime. We refer to the earlier discussion in Chapter 6 for details. Finally, if we excite a mechanism, then the masses will again follow straight lines, moving in different directions, whereas in the full nonlinear physical regime the masses may move along much more complicated curved trajectories.

Example 9.37. Consider a system of three unit masses connected in a line by two unit springs, but not attached to any fixed supports, as illustrated in Figure 9.9. This structure could be viewed as a simplified model of a triatomic molecule that is only allowed to move the vertical direction. The incidence matrix is $A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$ and, since we are dealing with unit springs, the stiffness matrix is

$$K = A^T A = \begin{pmatrix} -1 & 0 \\ -1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors of K are easily found:

$$\lambda_1 = 0, \qquad \lambda_2 = 1, \qquad \lambda_3 = 3\,,$$
$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Each positive eigenvalue provides two trigonometric solutions, while the zero eigenvalue leads to solutions that depend linearly on t. This yields the required six basis solutions:

$$\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{u}_3(t) = \begin{pmatrix} \cos t \\ 0 \\ -\cos t \end{pmatrix}, \qquad \mathbf{u}_5(t) = \begin{pmatrix} \cos \sqrt{3}\, t \\ -2\cos \sqrt{3}\, t \\ \cos \sqrt{3}\, t \end{pmatrix},$$
$$\mathbf{u}_2(t) = \begin{pmatrix} t \\ t \\ t \end{pmatrix}, \qquad \mathbf{u}_4(t) = \begin{pmatrix} \sin t \\ 0 \\ -\sin t \end{pmatrix}, \qquad \mathbf{u}_6(t) = \begin{pmatrix} \cos \sqrt{3}\, t \\ -2\cos \sqrt{3}\, t \\ \cos \sqrt{3}\, t \end{pmatrix}.$$

The first solution $\mathbf{u}_1(t)$ is a constant, equilibrium mode, where the masses rest at a fixed common distance from their reference positions. The second $\mathbf{u}_2(t)$ is the unstable mode, corresponding to a uniform vertical translational motion of the masses without any stretch of the interconnecting springs. The final four solutions represent vibrational modes. In the first pair $\mathbf{u}_3(t), \mathbf{u}_4(t)$, the two outer masses move in opposing directions, while the middle mass remains fixed, while the final pair $\mathbf{u}_5(t), \mathbf{u}_6(t)$ has the two outer masses moving in tandem, while the inner mass moves twice as far in the opposing direction. The general solution is a linear combination of the six normal modes,

$$\mathbf{u}(t) = c_1\, \mathbf{u}_1(t) + \, \cdots \, + c_6\, \mathbf{u}_6(t), \tag{9.69}$$

and corresponds to the molecule moving along its axis at a fixed speed while the individual masses perform a quasi-periodic vibration.

Let us see if we can predict the motion of the molecule from its initial conditions

$$\mathbf{u}(0) = \mathbf{a}, \qquad \dot{\mathbf{u}}(0) = \boldsymbol{\alpha},$$

where $\mathbf{a} = (\, a, b, c\,)^T$ is the initial displacements of the three atoms, while $\boldsymbol{\alpha} = (\, \alpha, \beta, \gamma\,)^T$ is their initial velocities. Substituting the solution formula (9.69) leads to the linear systems

$$c_1\, \mathbf{v}_1 + c_3\, \mathbf{v}_2 + c_5\, \mathbf{v}_3 = \mathbf{a}, \qquad c_2\, \mathbf{v}_1 + c_4\, \mathbf{v}_2 + \sqrt{3}\, c_6\, \mathbf{v}_3 = \boldsymbol{\alpha},$$

for the coefficients $c_1, \ldots, c_6$. Since the eigenvectors of the symmetric matrix K are mutually orthogonal, we can use our orthogonality formula to immediately compute the coefficients:

$$c_1 = \frac{\mathbf{a} \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2} = \frac{a+b+c}{3}\,, \qquad c_3 = \frac{\mathbf{a} \cdot \mathbf{v}_2}{\| \mathbf{v}_2 \|^2} = \frac{a-c}{2}\,, \qquad c_5 = \frac{\mathbf{a} \cdot \mathbf{v}_3}{\| \mathbf{v}_3 \|^2} = \frac{a-2\,b+c}{6}\,,$$
$$c_2 = \frac{\boldsymbol{\alpha} \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2} = \frac{\alpha+\beta+\gamma}{3}\,, \qquad c_4 = \frac{\boldsymbol{\alpha} \cdot \mathbf{v}_2}{\| \mathbf{v}_2 \|^2} = \frac{\alpha-\gamma}{2}\,, \qquad c_6 = \frac{\boldsymbol{\alpha} \cdot \mathbf{v}_3}{\sqrt{3}\, \| \mathbf{v}_3 \|^2} = \frac{\alpha-2\,\beta+\gamma}{6\sqrt{3}}\,.$$

In particular, the unstable translational mode is excited if and only if its coefficient $c_2 \neq 0$ is non-zero, and this occurs if and only if there is a nonzero net initial velocity of the molecule: $\alpha + \beta + \gamma \neq 0$. In this case the vibrating molecule will move off to ∞ at a uniform velocity $c = c_2 = \frac{1}{3}(\alpha + \beta + \gamma)$ equal to the average of the individual initial velocities. On the other hand, if $\alpha + \beta + \gamma = 0$, then the unstable mode will not be excited and the molecule will vibrate quasiperiodically, with frequencies 1 and $\sqrt{3}$, while sitting at a fixed location.

The observations established in this example hold, in fact, in complete generality. Let us state the result, leaving the details of the proof as an exercise for the reader.

Theorem 9.38. *The solution to unstable second order linear system with positive semi-definite coefficient matrix $K = A^T C A$ is a combination of a quasi-periodic vibration and a uniform motion at a fixed velocity in the direction of a null eigenvector $\mathbf{v} \in \ker A$. In particular, the system does not experience any unstable motion, and so will just vibrate around a fixed position, if and only if the initial velocity $\dot{\mathbf{u}}(t_0) \in (\ker K)^\perp = \operatorname{rng} K$ is orthogonal to the subspace $\ker A = \ker K$ of all unstable directions.*

As usual, the unstable modes correspond to either translations or rotations, or to mechanisms of the structure. To prevent a structure from exhibiting an unstable motion, one has to ensure that the initial velocity is orthogonal to all of the unstable directions. The result is in direct analogy with Theorem 6.8 that requires a force to be orthogonal to all such unstable modes in order to maintain equilibrium in the structure.

Systems with Different Masses

When a structure has differing masses at the nodes, the Newtonian equations of motion take the more general form

$$M\,\ddot{\mathbf{u}} = -K\,\mathbf{u}, \qquad \text{or, equivalently,} \qquad \ddot{\mathbf{u}} = -M^{-1}K\,\mathbf{u} = -P\,\mathbf{u}. \tag{9.70}$$

The mass matrix $M > 0$ is positive definite (and, usually, diagonal, although the general theory does not require this latter restriction), while the stiffness matrix $K = A^T C A$ is either positive definite or, in the unstable situation when $\ker A \neq \{\mathbf{0}\}$, positive semi-definite. The coefficient matrix

$$P = M^{-1}K = M^{-1}A^T C A \tag{9.71}$$

is *not* in general symmetric, and so we cannot directly apply the preceding constructions. However, P does have the more general self-adjoint form (7.68) based on the weighted inner products

$$\langle\, \mathbf{u}\,;\widetilde{\mathbf{u}}\,\rangle = \mathbf{u}^T M\,\widetilde{\mathbf{u}}, \qquad\qquad \langle\!\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle\!\rangle = \mathbf{v}^T C\,\widetilde{\mathbf{v}}, \tag{9.72}$$

on, respectively, the domain and target spaces for A.

If $\ker A = \{\mathbf{0}\}$, then $P > 0$ is positive definite in the generalized sense of Definition 7.58. In this case, substituting our standard trigonometric solution ansatz $\mathbf{u}(t) = \cos(\omega\, t)\,\mathbf{v}$ into the system results in a *generalized matrix eigenvalue problem*

$$K\,\mathbf{v} = \lambda\, M\,\mathbf{v}, \qquad \text{or, equivalently,} \qquad P\,\mathbf{v} = \lambda\,\mathbf{v}, \qquad \text{with} \qquad \lambda = \omega^2. \tag{9.73}$$

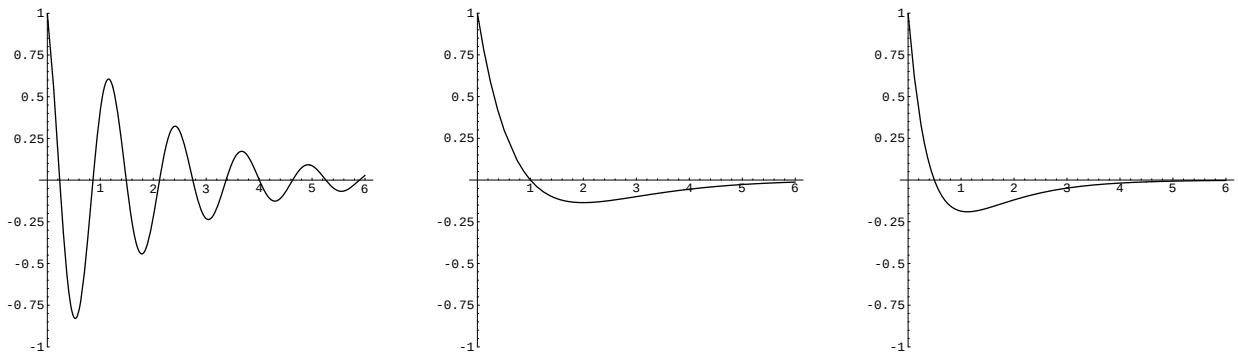


Figure 9.10. Damped Vibrations.

The matrix M plays the role of the identity matrix I in the standard eigenvalue equation (8.13). The proofs for the standard eigenvalue problem are easily modified to handle this situation, and demonstrate that all the eigenvalues are real and non-negative. Moreover the eigenvectors are orthogonal, but now with respect to the weighted inner product $\langle\, \mathbf{u}\,;\widetilde{\mathbf{u}}\,\rangle$ governed by the mass matrix M. Details are relegated to the exercises.

Friction and Damping

So far, we have not allowed frictional forces to affect the motion of our dynamical equations. In many physical systems, friction exerts a force on a mass in motion which is proportional to its velocity. In the simplest case of a single mass attached to a spring, one amends the balance of forces in the undamped Newton equation (9.55) to obtain

$$m\,\frac{d^2u}{dt^2} + \beta\,\frac{du}{dt} + k\,u = 0. \tag{9.74}$$

As before, $m > 0$ is the mass, and $k > 0$ the spring stiffness, while $\beta > 0$ measures the effect of a velocity-dependent frictional force — the larger β the greater the frictional damping of the motion.

The solution of this more general second order homogeneous linear ordinary differential equation is found by substituting the usual exponential ansatz $u(t) = e^{\lambda t}$ into the equation, leading to the quadratic characteristic equation

$$m\,\lambda^2 + \beta\,\lambda + k = 0. \tag{9.75}$$

There are three possible cases, illustrated in Figure 9.10:

Underdamped: If $0 < \beta^2 < 4\,m\,k$, then (9.75) has two complex-conjugate roots

$$\lambda \;=\; -\,\frac{\beta}{2\,m} \;\pm\; \mathrm{i}\;\frac{\sqrt{4\,m\,k - \beta^2}}{2\,m} \;\;=\;\; -\,\mu \pm \mathrm{i}\,\nu\,. \tag{9.76}$$

The general solution to the differential equation is

$$u(t) = e^{-\,\mu\,t}\big(\,c_1\cos\nu\,t + c_2\sin\nu\,t\,\big) = r\,e^{-\,\mu\,t}\cos(\nu\,t - \delta), \tag{9.77}$$

which represents a damped periodic motion. The time-dependent amplitude of vibration $a(t) = r\, e^{-\mu t}$ decays to zero at an exponential rate as $t \to \infty$. The formula for the rate of decay, $\mu = \beta/(2m)$, tells us that more friction or less mass will cause the system to return to equilibrium faster. (Of course, mathematically, it never quite gets there, but in a real physical system after a sufficiently long time the difference is not noticeable.) On the other hand, the frequency of vibration,

$$\nu \;=\; \frac{\sqrt{4mk - \beta^2}}{2m} \;=\; \sqrt{\frac{k}{m} - \frac{\beta^2}{4m^2}}\,, \tag{9.78}$$

remains fixed throughout the motion. The frictionally modified vibrational frequency ν is strictly smaller than the undamped frequency $\omega = \sqrt{k/m}$, and hence friction has the effect of slowing down vibrations while progressively diminishing their amplitudes. As the friction approaches a critical threshold, $\beta \nearrow 2\sqrt{m\,k}$, the vibrational frequency goes to zero, $\nu \to 0$, and so the period of vibration $P = 2\pi/\nu$ goes to ∞.

Overdamped: If $\beta^2 > 4mk$, then the characteristic equation (9.75) has two negative real roots

$$\lambda_1 \;=\; -\,\frac{\beta + \sqrt{\beta^2 - 4mk}}{2m}\,, \qquad\qquad \lambda_2 \;=\; -\,\frac{\beta - \sqrt{\beta^2 - 4mk}}{2m}\,,$$

with $\lambda_1 < \lambda_2 < 0$. The solution

$$u(t) = c_1\, e^{\lambda_1 t} + c_2\, e^{\lambda_2 t} \tag{9.79}$$

is a linear combination of two decaying exponentials. An overdamped system models the motion of a mass in a vat of molasses. Its "vibration" is so slow that it can pass at most once through its equilibrium position $u = 0$, and then only when its initial velocity is quite large. In the long term, since $\lambda_1 < \lambda_2$, the first exponential $e^{\lambda_1 t}$ will decay to zero faster, and hence the overall decay rate of the solution is (unless $c_2 = 0$) governed by the less negative eigenvalue λ_2.

Critically Damped: The borderline case occurs when $\beta^2 = 4mk$, which means that the characteristic equation (9.75) has only a single negative real root:

$$\lambda_1 = -\,\frac{\beta}{2m}\,.$$

In this case, our ansatz only supplies one exponential solution $e^{\lambda_1 t} = e^{-\beta t/2m}$. The second linearly independent solution is obtained by multiplication by t, leading to the general solution

$$u(t) = (c_1\, t + c_2) e^{-\beta t/2m}. \tag{9.80}$$

Even though the formula looks quite different, its qualitative behavior is very similar to the overdamped case. The factor of t plays an unimportant role, since the asymptotics of this solution are almost entirely governed by the decaying exponential function. This represents the nonvibrating solution that has the slowest possible decay rate — reducing the frictional coefficient any further will permit a damped periodic vibration to appear.

In all three cases, provided the frictional coefficient is positive, $\beta > 0$, the zero solution is globally asymptotically stable. Physically, since there is no external forcing, all solutions eventually return to equilibrium as the friction gradually overwhelms any initial motion.

Remark: You may, if you prefer, convert the second order equation (9.74) into a first order system by adopting the phase plane variables u and $v = \dot{u}$. The coefficient matrix of the equivalent phase plane system $\dot{\mathbf{u}} = A\,\mathbf{u}$ is $A = \begin{pmatrix} 0 & 1 \\ -c/m & -b/m \end{pmatrix}$. In terms of our classification of two-dimensional systems, the undamped case corresponds to a center, the underdamped case to a stable focus, the critically damped case to a stable improper node, and the overdamped case to a stable node. The reader should verify that the relevant conditions are met in each case and correlate the phase portraits with the time plots in Figure 9.10.

This concludes our discussion of the scalar case. Similar considerations apply to mass/spring chains, and two and three-dimensional structures. The frictionally damped system has the general form

$$M\,\frac{d^2\mathbf{u}}{dt^2} + B\,\frac{d\mathbf{u}}{dt} + K\,\mathbf{u} = \mathbf{0}, \tag{9.81}$$

where the mass matrix $M > 0$ and the matrix of frictional coefficients $B > 0$ are both diagonal, positive definite, while the stiffness matrix $K = A^T C\,A \geq 0$ is a positive semi-definite Gram matrix constructed from the reduced incidence matrix A. The mathematical details in this case are sufficiently complicated that we shall leave their analysis as an advanced project for the motivated student.

9.6. Forcing and Resonance.

So far, we have allowed our structure to vibrate on its own. It is now time to start applying external forces — to see what happens when we shake it. In this section, we will investigate the effects of periodic forcing on both undamped and damped systems. More general types of forcing can be handled by the variation of parameters method, cf. [**24**].

The simplest case is that of a single mass connected to a spring without any frictional damping. We append an external forcing function $f(t)$ to the homogeneous (unforced) equation (9.55), leading to the inhomogeneous ordinary differential equation

$$m\,\frac{d^2u}{dt^2} + k\,u = f(t), \tag{9.82}$$

in which $m > 0$ is the mass and $k > 0$ the spring stiffness. We are particularly interested in the case of periodic forcing

$$f(t) = \alpha \cos \eta\, t \tag{9.83}$$

of frequency $\eta > 0$ and amplitude α. To find a particular solution to (9.82), (9.83), we use the method of undetermined coefficients[†] which tells us to guess a solution ansatz of the form

$$u_\star(t) = a\,\cos \eta\, t + b\,\sin \eta\, t, \tag{9.84}$$

[†] One can also use variation of parameters, although the intervening calculations are slightly more complicated.

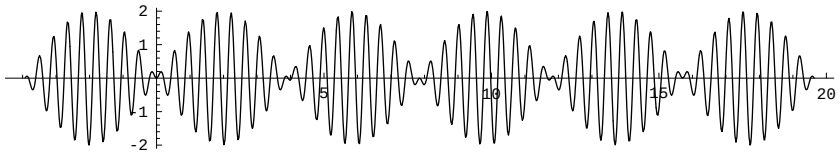


Figure 9.11. Beats in a Periodically Forced Vibration.

where a, b are constant. Substituting this ansatz into the differential equation, we find

$$m\,\frac{d^2u_\star}{dt^2} + k\,u_\star = a\,(k - m\,\eta^2)\cos\eta\,t + b\,(k - m\,\eta^2)\sin\eta\,t = \alpha\cos\eta\,t.$$

We can solve for

$$a = \frac{\alpha}{k - m\,\eta^2} = \frac{\alpha}{m(\omega^2 - \eta^2)}\,, \qquad\qquad b = 0,$$

provided the denominator is nonzero:

$$k - m\,\eta^2 = m(\omega^2 - \eta^2) \neq 0. \tag{9.85}$$

Here

$$\omega = \sqrt{\frac{k}{m}} \tag{9.86}$$

refers to the natural, unforced vibrational frequency of the system, while η is the forcing frequency. Therefore, provided the forcing frequency is *not* equal to the system's natural frequency, $\eta \neq \omega$, there exists a particular solution

$$u_\star(t) = a\,\cos\eta\,t = \frac{\alpha}{m(\omega^2 - \eta^2)}\,\cos\eta\,t \tag{9.87}$$

that vibrates at the same frequency as the forcing function.

The general solution to the inhomogeneous system (9.82) is found, as usual, by adding in an arbitrary solution to the homogeneous equation, (9.56), yielding

$$u(t) = r\,\cos(\omega\,t - \delta) + a\,\cos\eta\,t, \qquad \text{where} \qquad a = \frac{\alpha}{m(\omega^2 - \eta^2)}\,, \tag{9.88}$$

and where r and δ are determined by the initial conditions. The solution is therefore a quasiperiodic combination of two periodic motions — the first, vibrating with frequency ω, represents the internal or natural vibrations of the system, while the second, with frequency η, represents the response of the system to the periodic forcing. Due to the factor $\omega^2 - \eta^2$ in the denominator of (9.88), the closer the forcing frequency is to the natural frequency, the larger the overall amplitude of the response, and the more likely the spring breaks. displays the graph of

Suppose we start the mass initially at equilibrium, so the initial conditions are

$$u(0) = 0, \qquad\qquad \dot u(0) = 0. \tag{9.89}$$

Substituting the solution formula (9.88) and solving for r, δ, we find that

$$r = -\,a, \qquad\qquad \delta = 0.$$

Thus, the solution to the initial value problem can be written in the form

$$u(t) = a\big(\cos\eta\, t - \cos\omega\, t\big) = 2\,a\,\sin\left(\frac{\omega-\eta}{2}\,t\right)\,\sin\left(\frac{\omega+\eta}{2}\,t\right), \tag{9.90}$$

using a standard trigonometric identity, cf. Exercise ■. The factor $\sin\frac{1}{2}(\omega+\eta)t$ represents a periodic motion whose frequency is the average of the natural and the forcing frequencies. If the forcing frequency η is close to the natural frequency ω, then the initial factor $2\,a\,\sin\frac{1}{2}(\omega-\eta)t$ can be viewed as a periodically varying amplitude, whose vibrational frequency $\frac{1}{2}(\omega-\eta)$ is much slower. This factor is responsible for the phenomenon of *beats*, heard, for example, when two tuning forks of close but not exactly equal pitch vibrate near each other. The resulting sound periodically waxes and wanes in intensity. Figure 9.11 displays the graph of the particular function

$$\cos 14\,t - \cos 15.6\,t = 2\,\sin .8\,t\;\sin 14.8\,t.$$

The slowly varying amplitude $2\,\sin .8\,t$ is clearly visible as the envelope of the relatively rapid vibrations of frequency 14.8.

If we force the system at exactly the natural frequency $\eta = \omega$, then the trigonometric ansatz (9.84) does not work. This is because both terms are now solutions to the homogeneous equation, and so cannot be combined to form a solution to the inhomogeneous version. In this situation, there is a simple modification to the ansatz, namely multiplication by t, that does the trick. Substituting

$$u_\star(t) = a\,t\,\cos\omega\, t + b\,t\,\sin\omega\, t \tag{9.91}$$

into the differential equation (9.82), we find

$$m\,\frac{d^2u_\star}{dt^2} + k\,u_\star = -\,2\,a\,m\,\omega\,\sin\omega\, t + 2\,b\,m\,\omega\,\cos\omega\, t = \alpha\cos\omega\, t,$$

and so

$$a = 0, \qquad\qquad b = \frac{\alpha}{2\,m\,\omega}\,.$$

Combining the resulting particular solution with the solution to the homogeneous equation leads to the general solution

$$u(t) = r\cos(\omega\, t - \delta) + \frac{\alpha}{2\,m\,\omega}\;t\,\sin\omega\, t. \tag{9.92}$$

Both terms vibrate with frequency ω, but the second has a linearly growing amplitude that gets larger and larger as $t \to \infty$; see Figure 9.12. The mass will oscillate more and more wildly until the spring eventually breaks. In this situation, the system is said to be in *resonance*, and the increasingly wild oscillations are provoked by forcing it at the resonant frequency ω.

If we are very close to resonance, the oscillations induced by the particular solution (9.90) will have extremely large, although not unbounded, amplitude a. The lesson is, never force a system at or close to its natural frequency (or frequencies) of vibration. The classic example is the 1940 Tacoma Narrows Bridge disaster, when the vibration in the

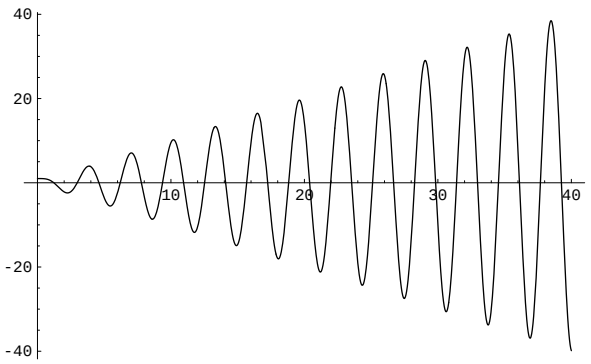


Figure 9.12. Resonance.

bridge caused by a strong wind was close enough to the bridge's natural frequency to cause it to oscillate wildly and collapse! A movie taken at the time is particularly impressive. A second example is the old practice of British (and subsequently, U.S.) infantry who, learning from experience, do not march in unison across a bridge so as not to set off a resonant frequency and cause it to collapse.

If we include frictional effects, then we can partially mollify the wild behavior near the resonant frequency. The frictionally damped vibrations of a mass on a spring, when subject to periodic forcing, can be described by the inhomogeneous version

$$m\,\frac{d^2u}{dt^2} + \beta\,\frac{du}{dt} + k\,u = \alpha\cos\eta\,t \tag{9.93}$$

of equation (9.74). Let us assume that the friction is sufficiently small as to keep us in the underdamped regime $\beta^2 < 4\,m\,k$. Since neither summand solves the homogeneous system, we can use the trigonometric solution ansatz (9.84) to construct the particular solution

$$u_\star(t) = a\cos(\eta\,t - \varepsilon) \qquad \text{where} \qquad a = \frac{\alpha}{\sqrt{m^2(\omega^2-\eta^2)^2 + \beta^2\,\eta^2}} \tag{9.94}$$

represents the amplitude of the response to the periodic forcing, with $\omega = \sqrt{k/m}$ continuing to denote the undamped resonant frequency (9.86), while

$$\varepsilon = \tan^{-1}\frac{\beta\,\omega}{m(\omega^2-\eta^2)} \tag{9.95}$$

represents a *phase lag* in the response of the system that is due to the friction.

The general solution is

$$u(t) = r\,e^{-\mu\,t}\cos(\nu\,t - \delta) + a\cos(\eta\,t - \varepsilon), \tag{9.96}$$

where r, δ are determined by the initial conditions, while $\lambda = \mu + \mathrm{i}\,\nu$ are the roots of the characteristic equation, cf. (9.76). The first term — the solution to the homogeneous equation — is called the *transient* since it decays exponentially fast to zero. Thus, at large times, the internal motion of the system that might have been excited by the initial conditions dies out, and only the particular solution (9.94) incited by the forcing persists. The amplitude of the persistent response (9.94) is at a maximum at the resonant frequency

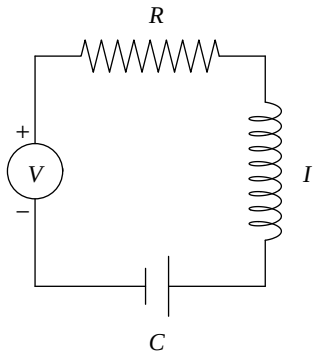


Figure 9.13. The Basic RLC Circuit.

$\eta = \omega$, where it takes the value $a_{max} = \alpha/(\beta\,\omega)$. Thus, the smaller the frictional coefficient β (or the slower the resonant frequency ω) the more likely the breakdown of the system due to an overly large response.

Friction also induces a phase shift ε in the response of the system to the external forcing. Speeding up the forcing frequency η increases the overall phase shift, which has the value of $\frac{1}{2}\,\pi$ at the resonant frequency $\eta = \omega$, so the system lags a quarter period behind the forcing, and reaches a maximum $\varepsilon = \pi$ as $\eta \to \infty$. Thus, the response of the system to a high frequency forcing is almost exactly out of phase — the mass is moving downwards when the force is pulling it upwards, and vice versa!

Electrical Circuits

The Electrical–Mechanical Correspondence will continue to operate in the dynamical universe. As we learned in Chapter 6, the equations governing the equilibria of simple electrical circuits and the mechanical systems such as mass/spring chains and structures are modeled by the same basic mathematical structure. In a similar manner, circuits with time-varying currents can also be modeled by linear dynamical systems of ordinary differential equations.

In this section, we analyze the simplest situation of an RLC *circuit* consisting of a *resistor* R, an *inductor* L and a *capacitor* C connected together in a loop as illustrated in Figure 9.13. Let $u(t)$ denote the current in the circuit at time t. As the current passes through each circuit element, it induces a corresponding voltage, which we denote by v_R, v_L and v_C. The voltages are prescribed by the basic laws of electrical circuit design.

(a) First, as we know from Section 6.2, the resistance $R \geq 0$ in the circuit is the proportionality factor between voltage and current, so $v_R = R\,u$.

(b) The voltage passing through an inductor is proportional to the rate of change in the current. Thus, $v_L = L\,\dot{u}$, where $L > 0$ is the *inductance*, and the dot indicates time derivative.

(c) On the other hand, the current passing through a capacitor is proportional to the rate of change in the voltage, and so $u = C\,\dot{v}_C$, where $C > 0$ denotes the *capacitance*. We integrate this relation to produce the capacitor voltage $v_C = \displaystyle\int \frac{u(t)}{C}\, dt$.

The combination of all the induced voltages must equal the externally applied voltage from, say, a battery. The precise rules governing these voltages are:

The voltage balance law tells us that the total of these individual voltages must equal any externally applied voltage coming from, say, a battery or generator. Therefore,

$$v_R + v_L + v_C = v_B,$$

where $v_B = f(t)$ denotes the applied voltage due to a time-varying source. Substituting the preceding formulae, we deduce that the current $u(t)$ in our circuit satisfies following linear integro-differential equation

$$L\,\frac{du}{dt} + R\,u + \int \frac{u}{C}\,dt = f(t). \tag{9.97}$$

We can convert this into a differential equation by differentiating both sides with respect to t. Assuming, for simplicity, that L, R and C are constant, the result is the linear second order ordinary differential equation

$$L\,\frac{d^2u}{dt^2} + R\,\frac{du}{dt} + \frac{1}{C}\,u = f'(t). \tag{9.98}$$

In particular, the homogeneous version, with $f'(t) \equiv 0$, governs the current in an RLC circuit with a constant applied voltage source.

Comparing (9.98) with the equation (9.74) for a mechanically vibrating mass, we see that the analogy between electrical circuits and mechanical structures developed in Chapter 6 continues to hold in the dynamical regime. The current corresponds to the displacement. The inductance plays the role of mass, the resistance corresponds to friction, while the reciprocal $1/C$ of capacitance is analogous to the spring stiffness. Thus, all of our analytical conclusions regarding stability of equilibria, qualitative behavior and formulae for solutions, etc., that we established in the mechanical context can, suitably re-interpreted, be immediately applied to electrical circuit theory.

In particular, an RLC circuit is *underdamped* if $R^2 < 4\,L/C$, and the current $u(t)$ oscillates with frequency

$$\nu = \sqrt{\frac{1}{C\,L} - \frac{R^2}{4\,L^2}}\,, \tag{9.99}$$

while slowly dying off to zero. In the overdamped and critically damped cases $R^2 \geq 4\,L/C$, where the resistance of the circuit is large, the current merely decays to zero exponentially fast and there is no longer any oscillatory behavior in the circuit. Attaching an alternating current source $f(t) = f_0 + a\,\sin \eta\, t$ to the circuit will induce resonance in the case of no resistance if the forcing frequency is equal to the circuits natural internal frequency. Details are relegated to the exercises.

Forcing and Resonance in Systems

Let us very briefly discuss the effect of periodic forcing on a more complicated system. For undamped mass/spring chains, structures and more complicated resistanceless LC circuits, we are led to consider a periodically forced second order system

$$M\,\ddot{\mathbf{u}} + K\,\mathbf{u} = \cos(\omega\, t)\,\mathbf{a}, \tag{9.100}$$

where $\mathbf{a}$ is a constant vector representing both a magnitude and a "direction" of the forcing. Here $M > 0$ is the diagonal mass matrix (or inductance matrix in a circuit), while $K = A^T C A$ the (semi-)definite stiffness (or conductance) matrix for the system. We are ignoring friction (resistance) for simplicity. More general periodic and quasiperiodic forcing terms can be built up, via the general inhomogeneous superposition principle of Theorem 7.42, as a linear combination of such simple solutions.

To find a particular solution to the system, let us try the trigonometric ansatz

$$\mathbf{u}^\star(t) = \cos(\omega\, t)\, \mathbf{w} \tag{9.101}$$

where $\mathbf{w}$ is a constant vector. Substituting into (9.100) leads to a linear algebraic system

$$(K - \lambda M)\, \mathbf{w} = \mathbf{a}, \qquad \text{where} \qquad \lambda = \omega^2. \tag{9.102}$$

If equation (9.102) has a solution, then our ansatz (9.101) is valid, and we have produced a particular vibration of the system that has the same frequency as the forcing vibration. The general solution, then, will be a quasi-periodic combination of this particular solution coupled with the vibrations at the system's natural, unforced frequencies. In particular, if $\lambda = \omega^2$ is *not* a generalized eigenvalue[†] of the matrix pair K, M, as in (9.73), then the coefficient matrix $K - \lambda M$ is nonsingular, and so (9.102) can be solved for any right hand side $\mathbf{a}$.

The more interesting case is when $K - \lambda M$ is singular, its kernel being equal to the generalized eigenspace V_λ. In this case, (9.102) will have a solution $\mathbf{w}$ if and only if $\mathbf{a}$ lies in the range of $K - \lambda M$. According to the Fredholm Alternative Theorem 5.51, the range is the orthogonal complement of the cokernel, which, since the coefficient matrix is symmetric, is the same as the kernel. Therefore, (9.102) will have a solution if and only if $\mathbf{a}$ is orthogonal to V_λ, i.e., $\mathbf{a} \cdot \mathbf{v} = \mathbf{0}$ for every eigenvector $\mathbf{v}$ for the eigenvalue λ. Thus, one can force a system at a natural frequency without inciting resonance provided the "direction" of forcing, as governed by the vector $\mathbf{a}$, is orthogonal to the natural directions of motion of the system, as governed by the eigenvectors for that particular frequency.

If this orthogonality constraint is not satisfied, then the periodic solution ansatz (9.101) does not apply, and we are in a truly resonant situation. Inspired by the scalar solution, let us try the *resonant ansatz*

$$\mathbf{u}^\star(t) = t \sin(\omega t)\, \mathbf{y} + \cos(\omega t)\, \mathbf{w}. \tag{9.103}$$

Since

$$\frac{d^2\mathbf{u}^\star}{dt^2} = -\omega^2\, t \sin(\omega\, t)\, \mathbf{y} + \cos(\omega\, t)\, (2\omega\, \mathbf{y} - \omega^2\, \mathbf{w}),$$

the function (9.103) will be a solution to the differential equation (9.100) provided

$$(K - \lambda M)\mathbf{y} = \mathbf{0}, \qquad (K - \lambda M)\mathbf{w} = \mathbf{a} - 2\omega\, \mathbf{y}. \tag{9.104}$$

The first equation requires that $\mathbf{y} \in V_\lambda$ be a generalized eigenvector of the matrix pair K, M. The Fredholm alternative Theorem 5.51 implies that, since the coefficient matrix

[†] When $M = \mathrm{I}$ the system reduces to the standard eigenvalue equation for K.

$K - \lambda M$ is symmetric, the second equation will be solvable for $\mathbf{w}$ if and only if $\mathbf{a} - 2\omega\mathbf{y}$ is orthogonal to the eigenspace $V_\lambda = \operatorname{coker}(K - \lambda M) = \ker(K - \lambda M)$. Thus, $2\omega\mathbf{y}$ must be the orthogonal projection of $\mathbf{a}$ onto V_λ. With this choice of $\mathbf{y}$ and $\mathbf{w}$, formula (9.103) produces a resonant solution to the system.

Summarizing, we have shown that, generically, forcing a system at a resonant frequency induces resonance.

Theorem 9.39. *An undamped vibrational system will be periodically forced into resonance if and only if the forcing* $\mathbf{f} = \cos(\omega\, t)\, \mathbf{a}$ *is at a natural frequency of the system and the direction of forcing* $\mathbf{a}$ *is not orthogonal to the natural direction(s) of motion of the system for that frequency.*

Example 9.40. Consider the peridoically forced system

$$\frac{d^2\mathbf{u}}{dt^2} + \begin{pmatrix} 3 & -2 \\ -2 & 3 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \cos t \\ 0 \end{pmatrix}.$$

The eigenvalues of the coefficient matrix are $\lambda_1 = 5, \lambda_2 = 1$, with corresponding orthogonal eigenvectors $\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. The resonant frequencies are $\omega_1 = \sqrt{\lambda_1} = \sqrt{5}, \omega_2 = \sqrt{\lambda_2} = 1$, and hence we are forcing at a resonant frequency. To obtain the resonant solution (9.103), we first note that $\mathbf{a} = (\,1, 0\,)^T$ has orthogonal projection $\mathbf{p} = \left(\,\frac{1}{2}, \frac{1}{2}\,\right)^T$ onto the eigenline spanned by $\mathbf{v}_2$, and hence $\mathbf{y} = \mathbf{p}/(2\omega) = \left(\,\frac{1}{4}, \frac{1}{4}\,\right)^T$. We can then solve

$$(K - \mathrm{I}\,)\mathbf{w} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \mathbf{w} = \mathbf{a} - \mathbf{p} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \qquad \text{for} \qquad \mathbf{w} = \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix}.$$

(We can safely ignore the arbitrary multiple of the eigenvector that can be added to $\mathbf{w}$ as we only need find one particular solution; these will reappear anyway once we assemble the general solution to the system.) Therefore, the particular resonant solution is

$$\mathbf{u}^\star(t) = t \sin t\, \mathbf{y} + \cos t\, \mathbf{w} = \begin{pmatrix} \frac{1}{4}\, t \sin t + \frac{1}{4} \cos t \\ \frac{1}{4}\, t \sin t \end{pmatrix}.$$

The general solution to the system is

$$\mathbf{u}(t) = \begin{pmatrix} \frac{1}{4}\, t \sin t + \frac{1}{4} \cos t \\ \frac{1}{4}\, t \sin t \end{pmatrix} + r_1 \cos\left(\,\sqrt{5}\, t - \delta_1\,\right) \begin{pmatrix} -1 \\ 1 \end{pmatrix} + r_2 \cos(t - \delta_2) \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where the amplitudes r_1, r_2 and phase shifts δ_1, δ_2, are fixed by the initial conditions. Eventually the resonant term $\frac{1}{4}\, t \sin t\, (\,1, 1\,)^T$ dominates the solution, inducing progressively larger and larger oscillations.

Chapter 8

Eigenvalues and Dynamics

So far, we have concentrated on statics: unchanging equilibrium configurations of mass/spring chains, circuits, and structures. It is now time to introduce motion into our universe. In general, a *dynamical system* refers to the (differential) equations governing the temporal behavior of some physical system. In a discrete system, the dynamical behavior of, say, a mass–spring chain, a simple electrical circuit, or the vibrations of a structure, is governed by a system of ordinary differential equations. Dynamics of continuous media — fluids, solids and gases — are governed by partial differential equations, and will form the focus of the later chapters.

The goal of this chapter is to understand the behavior of the simplest class of dynamical systems — constant coefficient linear systems of ordinary differential equations. We begin with a very quick review of the scalar case, whose solutions are exponential functions. Applying a similar exponential ansatz to the vector version leads us naturally to the absolutely fundamental notions of eigenvalue and eigenvector for a square matrix.

The next five sections are devoted to the basic properties of eigenvalues and eigenvectors. Most square matrices are complete, meaning that their (complex) eigenvectors form a basis of the underlying vector space. Complete matrices are diagonalizable, meaning that they are similar to a diagonal matrix when written in the eigenvector basis; the net result is that computations become significantly simpler when performed in the eigenvector basis. The most important class are the symmetric matrices, whose eigenvectors form an orthogonal basis of $\mathbb{R}^n$; in fact, this is by far the most common way for orthogonal bases to appear. In Section 8.4 we discuss incomplete matrices, which do not have eigenvector bases. The general Jordan canonical form, named after the nineteenth century French mathematician Camille Jordan (and not the Jordan of Gauss–Jordan fame), replaces the diagonal form in the incomplete cases. In Section 8.6 we discuss the singular values, which can be viewed as a generalization of eigenvalues to more general rectangular matrices. The *singular value decomposition*, abbreviated as SVD, of a matrix governs its condition number and hence the degree of difficulty of solving associated linear systems. Singular values have been appearing in an increasingly wide range of significant applications, ranging over image processing, data clustering, object recognition, semantics and language processing, and learning theory.

For a first order system of differential equations, the "eigenstates" describe the basic modes of exponential growth, decay, or periodic behavior. The stability of the equilibrium solution is almost entirely determined by the eigenvalues of the associated matrix, which explains their ubiquitous role in physical phenomena. Most of the important phenomena already appear in the two-dimensional systems, and we devote Section 8.9 to a complete

description of the possible behaviors.

In a mechanical system without damping or frictional effects, the eigenstates are the "normal modes" of the system, each periodically vibrating with its associated fundamental frequency. Linearity then allows us to describe the general motion as a linear superposition of the individual pure periodic normal modes of vibration. Such a linear combination will, in general, no longer be periodic, and so the motion can appear to be quite erratic. Nevertheless, it is merely the superposition of very simple periodic motions — called "quasi-periodic", and, unlike a chaotic, nonlinear system, is eminently predictable. When the system is forced, the result is a superposition of the free quasi-periodic motion and a particular reaction of the system to the forcing. In particular, periodic forcing will typically lead to quasiperiodic motion, unless we try to force the system at one of the fundamental frequencies; this will lead to the phenomenon of resonance, where the vibrations become larger and larger and the system breaks apart.

Many of the observations in this chapter are fundamental to general dynamical systems, and, as we shall see, apply equally well to the continuous case, where the physical system is governed by a linear partial differential equation. For example, the orthogonal bases of functions appearing in Fourier analysis and solution of partial differential equations arise as the eigenvectors, or, rather, eigenfunctions of "symmetric" boundary value problems for linear differential operators. However, before making this leap in abstraction, we need to properly understand the finite-dimensional matrix version first. Finally, even when the physics forces us to consider nonlinear systems, the tools from the linear regime will be essential for navigating these far more treacherous waters.

8.1. First Order Linear Systems of Ordinary Differential Equations.

The simplest dynamical system consists of n linear ordinary differential equations

$$\begin{aligned}
\frac{du_1}{dt} &= a_{11}\,u_1 + a_{12}\,u_2 + \,\cdots\, + a_{1n}\,u_n,\\
\frac{du_2}{dt} &= a_{21}\,u_1 + a_{22}\,u_2 + \,\cdots\, + a_{2n}\,u_n,\\
&\;\;\vdots \qquad\qquad\qquad\quad \vdots\\
\frac{du_n}{dt} &= a_{n1}\,u_1 + a_{n2}\,u_2 + \,\cdots\, + a_{nn}\,u_n,
\end{aligned} \tag{8.1}$$

involving n unknown functions $u_1(t), u_2(t), \ldots, u_n(t)$ depending on a scalar variable $t \in \mathbb{R}$, which we usually view as time. In the constant coefficient case, which is the only one to be treated in depth, the coefficients a_{ij} are assumed to be (real) constants. Such systems can be written in the compact matrix form

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}, \tag{8.2}$$

where A is a constant $n \times n$ matrix, and $\mathbf{u}(t) = (\,u_1(t), \ldots, u_n(t)\,)^T$ a vector-valued function. We seek not only to develop basic solution techniques for such dynamical systems, but to also understand their behavior from both a qualitative and quantitative standpoint.

The Scalar Case

We begin by analyzing the elementary scalar ordinary differential equation

$$\frac{du}{dt} = a\,u. \tag{8.3}$$

in detail. Here $a \in \mathbb{R}$ is a real constant, while the unknown $u(t)$ is a scalar function.

As you learned in calculus[†], the general solution to (8.3) is an exponential function

$$u(t) = c\,e^{a\,t}. \tag{8.4}$$

The integration constant c is uniquely determined by a single initial condition

$$u(t_0) = u_0 \tag{8.5}$$

imposed at an initial time t_0. Substituting $t = t_0$ into the solution formula (8.4), we find

$$u(t_0) = c\,e^{a\,t_0} = u_0, \qquad \text{and so} \qquad c = u_0\,e^{-a\,t_0}.$$

We conclude that there is a unique solution to the scalar initial value problem (8.3), (8.5), namely

$$u(t) = u_0\,e^{a(t-t_0)}. \tag{8.6}$$

Example 8.1. The radioactive decay of an isotope, say Uranium 238, is governed by the differential equation

$$\frac{du}{dt} = -\,\gamma\,u. \tag{8.7}$$

Here $u(t)$ denotes the amount of the isotope remaining at time t, and the coefficient $\gamma > 0$ governs the decay rate. The solution is given by an exponentially decaying function $u(t) = c\,e^{-\gamma\,t}$, where $c = u(0)$ is the initial amount of radioactive material.

The *half-life* $t^\star$ is the time it takes for half of a sample to decay, that is when $u(t^\star) = \frac{1}{2}\,u(0)$. To determine $t^\star$, we solve the algebraic equation

$$e^{-\gamma\,t^\star} = \tfrac{1}{2}, \qquad \text{so that} \qquad t^\star = \frac{\log 2}{\gamma}\,. \tag{8.8}$$

At each integer multiple $n\,t^\star$ of the half-life, exactly half of the isotope has decayed, i.e., $u(n\,t^\star) = 2^{-n}\,u(0)$.

Let us make some elementary, but pertinent observations about this simple linear dynamical system. First of all, since the equation is homogeneous, the zero function $u(t) \equiv 0$ (corresponding to $c = 0$) is a constant solution, known as an *equilibrium solution* or *fixed point*, since it does not depend on t. If the coefficient $a > 0$ is positive, then the solutions (8.4) are exponentially growing (in absolute value) as $t \to +\infty$. This implies that the zero equilibrium solution is *unstable*. The initial condition $u(t_0) = 0$ produces the zero solution, but if we make a tiny error (either physical, numerical, or mathematical) in the

[†] See also Section 19.1.

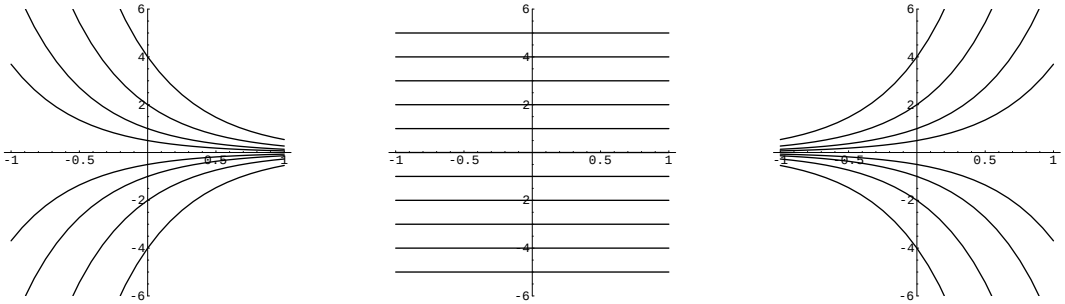


Figure 8.1. Solutions to $\dot{u} = a\,u$.

initial data, say $u(t_0) = \varepsilon$, then the solution $u(t) = \varepsilon\, e^{a(t-t_0)}$ will eventually get very far away from equilibrium. More generally, any two solutions with very close, but not equal, initial data, will eventually become arbitrarily far apart: $|\,u_1(t) - u_2(t)\,| \to \infty$ as $t \to \infty$. One consequence is the inherent difficulty in accurately computing the long time behavior of the solution, since small numerical errors will eventually have very large effects.

On the other hand, if $a < 0$, the solutions are exponentially decaying in time. In this case, the zero solution is *stable*, since a small error in the initial data will have a negligible effect on the solution. In fact, the zero solution is *globally asymptotically stable*. The phrase "asymptotically stable" implies that solutions that start out near zero eventually return; more specifically, if $u(t_0) = \varepsilon$ is small, then $u(t) \to 0$ as $t \to \infty$. The adjective "globally" implies that this happens no matter how large the initial data is. In fact, for a linear system, the stability (or instability) of an equilibrium solution is always a global phenomenon.

The borderline case is when $a = 0$. Then all the solutions to (8.3) are constant. In this case, the zero solution is *stable* — indeed, globally stable — but not asymptotically stable. The solution to the initial value problem $u(t_0) = \varepsilon$ is $u(t) \equiv \varepsilon$. Therefore, a solution that starts out near equilibrium will remain near, but will not asymptotically return. The three qualitatively different possibilities are illustrated in Figure 8.1.

Analogous stability results hold for linear systems (8.2) in several unknowns, but to properly formulate them, we must acquire some familiarity with the basic solution techniques.

The Phase Plane

Many fundamental physical phenomena are modeled by second order ordinary differential equations. The simplest scalar version is a linear, homogeneous equation

$$\frac{d^2u}{dt^2} + \alpha\,\frac{du}{dt} + \beta\,u = 0, \tag{8.9}$$

in which α, β are prescribed constants. In your first course on ordinary differential equations, you learned how to solve such equations; the basic method is reviewed in Example 7.31 and in the following example.

Example 8.2. Consider the second order equation

$$\frac{d^2u}{dt^2} + \frac{du}{dt} - 6\,u = 0. \tag{8.10}$$

To solve the equation, we substitute an exponential formula or ansatz[†] $u(t) = e^{\lambda t}$ into the equation. The result is the *characteristic equation* for the unspecified exponent λ:

$$\lambda^2 + \lambda - 6 = 0, \qquad \text{with solutions} \qquad \lambda_1 = 2, \qquad \lambda_2 = -3.$$

In view of Theorem 7.33, we conclude that e^{2t} and e^{-3t} form a basis for the two-dimensional solution space to (8.10). The general solution can be written as a linear combination

$$u(t) = c_1\, e^{2t} + c_2\, e^{-3t},$$

where c_1, c_2 are arbitrary constants.

There is a standard trick to convert any second order scalar equation, e.g., (8.9) into a first order system. One introduces the variables[‡]

$$u_1 = u, \qquad\qquad u_2 = \dot u = \frac{du}{dt}\,. \tag{8.11}$$

In view of (8.9), these variables satisfy

$$\frac{du_1}{dt} = \frac{du}{dt} = u_2, \qquad\qquad \frac{du_2}{dt} = \frac{d^2u}{dt^2} = -\beta\,u - \alpha\,\frac{du}{dt} = -\beta\,u_1 - \alpha\,u_2.$$

In this manner, the second order equation is converted into the equivalent first order system

$$\dot{\mathbf{u}} = A\,\mathbf{u}, \qquad \text{where} \qquad \mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 1 \\ -\beta & -\alpha \end{pmatrix}. \tag{8.12}$$

The $(u_1, u_2) = (u, \dot u)$ plane is referred to as the *phase plane*. The solutions $\mathbf{u}(t)$ to (8.12) parametrize curves in the phase plane — the solution *trajectories* or *orbits*. In particular, the equilibrium solution $\mathbf{u}(t) \equiv \mathbf{0}$ remains fixed at the origin, and so its trajectory is a single point All other solutions describe genuine curves. The collection of all possible solution trajectories is called the *phase portrait* of the system. An important fact is that, for a (constant coefficient) first order system, *the phase plane trajectories never cross*. This striking property, which is also valid for nonlinear systems, is a consequence of the uniqueness properties of solutions, and will be discussed in detail in Section 19.2. Thus, the phase portrait consists of a family of non-intersecting curves and equilibrium points that fill out the entire phase plane. The direction of motion along the trajectory is indicated by a small arrow. The one feature that is not so easily pictured in the phase portrait is the speed at which the solution moves along the phase curves — this would require a more complicated three-dimensional plot in which the third axis indicates time.

[†] See the footnote on p. 344 for an explanation of this term.

[‡] We will often use dots as a shorthand notation for time derivatives.

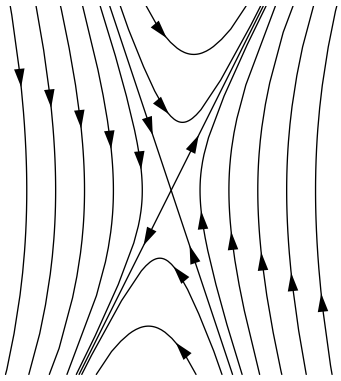


Figure 8.2. Phase Plane Trajectories for $\dot{u}_1 = u_2$, $\dot{u}_2 = 6\,u_1 - u_2$.

It is not hard to verify that every solution $u(t)$ to the second order equation yields a solution $\mathbf{u}(t) = \big(\, u(t), \dot{u}(t) \,\big)^T$ to the phase plane system (8.12). Vice versa, if $\mathbf{u}(t) = \big(\, u_1(t), u_2(t) \,\big)^T$ is any solution to the system (8.12), then its first component $u(t) = u_1(t)$ defines a solution to the original scalar equation (8.9). We conclude that the scalar equation and its phase plane version are completely equivalent; solving one will immediately lead to a solution of the other.

Example 8.3. For the second order equation (8.10), the equivalent phase plane system is

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 1 \\ 6 & -1 \end{pmatrix} \mathbf{u}, \qquad \text{or, in full detail,} \qquad \begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= 6\,u_1 - u_2. \end{aligned} \tag{8.13}$$

Our identification (8.11) of the phase plane variables tells us that the solution to the system (8.13) is given by

$$\begin{aligned} u_1(t) &= u(t) = c_1\, e^{2\,t} + c_2\, e^{-3\,t}, \\ u_2(t) &= \frac{du}{dt} = 2\,c_1\, e^{2\,t} - 3\,c_2\, e^{-3\,t}, \end{aligned}$$

and hence

$$\mathbf{u}(t) = \begin{pmatrix} c_1\, e^{2\,t} + c_2\, e^{-3\,t} \\ 2\,c_1\, e^{2\,t} - 3\,c_2\, e^{-3\,t} \end{pmatrix} = c_1 \begin{pmatrix} e^{2\,t} \\ 2\,e^{2\,t} \end{pmatrix} + c_2 \begin{pmatrix} e^{-3\,t} \\ -3\,e^{-3\,t} \end{pmatrix}.$$

A plot of the phase plane trajectories $\mathbf{u}(t)$ for various choices of the constants c_1, c_2 appears in Figure 8.2. The horizontal axis represents the solution $u_1 = u(t)$ whereas the vertical axis represents is derivative $u_2 = \dot{u}(t)$. With some practice, one learns to understand the temporal behavior of the solution from studying its phase plane trajectory. Many more examples will appear in Section 8.9 below.

8.2. Eigenvalues and Eigenvectors.

Let us now focus on our primary mission — solving a first order linear system of ordinary differential equations

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}, \tag{8.14}$$

with constant coefficient matrix A. Drawing our inspiration from the solution (8.4) in the scalar case, let us investigate whether the vector system has any solutions of a similar exponential form

$$\mathbf{u}(t) = e^{\lambda t}\,\mathbf{v}, \tag{8.15}$$

in which λ is a constant scalar, so $e^{\lambda t}$ is a scalar function of t, while $\mathbf{v} \in \mathbb{R}^n$ is a constant vector. In other words, the components $u_i(t) = v_i\, e^{\lambda t}$ of our desired solution are assumed to be constant multiples of the *same* exponential function. Since $\mathbf{v}$ is assumed to be constant, the derivative of $\mathbf{u}(t)$ is easily found:

$$\frac{d\mathbf{u}}{dt} = \frac{d}{dt}\left(e^{\lambda t}\,\mathbf{v}\right) = \lambda\, e^{\lambda t}\,\mathbf{v}.$$

On the other hand, since $e^{\lambda t}$ is a scalar, it commutes with matrix multiplication, and so

$$A\,\mathbf{u} = A\, e^{\lambda t}\,\mathbf{v} = e^{\lambda t} A\,\mathbf{v}.$$

Therefore, $\mathbf{u}(t)$ will solve the system (8.14) if and only if

$$\lambda\, e^{\lambda t}\,\mathbf{v} = e^{\lambda t} A\,\mathbf{v},$$

or, canceling the common scalar factor $e^{\lambda t}$,

$$\lambda\,\mathbf{v} = A\,\mathbf{v}.$$

The result is a system of algebraic equations relating the vector $\mathbf{v}$ and the scalar λ.

The preceding analysis motivates the following absolutely fundamental definition.

Definition 8.4. Let A be an $n \times n$ matrix. A scalar λ is called an *eigenvalue* of A if there is a *non-zero* vector $\mathbf{v} \neq \mathbf{0}$, called an *eigenvector*, such that

$$A\,\mathbf{v} = \lambda\,\mathbf{v}. \tag{8.16}$$

Remark: The odd-looking terms "eigenvalue" and "eigenvector" are hybrid German–English words. In the original German, they are *Eigenwert* and *Eigenvektor*, which can be fully translated as "proper value" and "proper vector". For some reason, the half-translated terms have acquired a certain charm, and are now standard. The alternative English terms *characteristic value* and *characteristic vector* can be found in some (mostly older) texts. Oddly, the term *characteristic equation*, to be defined below, is still used.

The requirement that the eigenvector $\mathbf{v}$ be nonzero is important, since $\mathbf{v} = \mathbf{0}$ is a trivial solution to the eigenvalue equation (8.16) for *any* scalar λ. Moreover, as far as solving linear ordinary differential equations goes, the zero vector $\mathbf{v} = \mathbf{0}$ only gives the trivial zero solution $\mathbf{u}(t) \equiv \mathbf{0}$.

The eigenvalue equation (8.16) is a system of linear equations for the entries of the eigenvector $\mathbf{v}$ — provided the eigenvalue λ is specified in advance — but is "mildly" nonlinear as a combined system for λ and $\mathbf{v}$. Gaussian elimination per se will not solve the problem, and we are in need of a new idea. Let us begin by rewriting the equation in the form

$$(A - \lambda\,\mathrm{I}\,)\mathbf{v} = \mathbf{0}, \tag{8.17}$$

where I is the identity matrix of the correct size[†]. Now, for given λ, equation (8.17) is a homogeneous linear system for $\mathbf{v}$, and always has the trivial zero solution $\mathbf{v} = \mathbf{0}$. But we are specifically seeking a nonzero solution! According to Theorem 1.45, a homogeneous linear system has a nonzero solution $\mathbf{v} \neq \mathbf{0}$ if and only if its coefficient matrix, which in this case is $A - \lambda \, \mathrm{I}$, is singular. This observation is the key to resolving the eigenvector equation.

Theorem 8.5. *A scalar λ is an eigenvalue of the $n \times n$ matrix A if and only if the matrix $A - \lambda \, \mathrm{I}$ is singular, i.e., of rank $< n$. The corresponding eigenvectors are the nonzero solutions to the eigenvalue equation $(A - \lambda \, \mathrm{I})\mathbf{v} = \mathbf{0}$.*

We know a number of ways to characterize singular matrices, including the determinantal criterion given in Theorem 1.50. Therefore, the following result is an immediate corollary of Theorem 8.5.

Proposition 8.6. *A scalar λ is an eigenvalue of the matrix A if and only if λ is a solution to the* characteristic equation

$$\det(A - \lambda \, \mathrm{I}) = 0. \tag{8.18}$$

In practice, when finding eigenvalues and eigenvectors by hand, one first solves the characteristic equation (8.18). Then, for each eigenvalue λ one uses standard linear algebra methods, i.e., Gaussian elimination, to solve the corresponding linear system (8.17) for the eigenvector $\mathbf{v}$.

Example 8.7. Consider the 2×2 matrix

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

We compute the determinant in the characteristic equation using (1.34):

$$\det(A - \lambda \, \mathrm{I}) = \det \begin{pmatrix} 3 - \lambda & 1 \\ 1 & 3 - \lambda \end{pmatrix} = (3 - \lambda)^2 - 1 = \lambda^2 - 6\lambda + 8.$$

The characteristic equation is a quadratic polynomial equation, and can be solved by factorization:

$$\lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2) = 0.$$

We conclude that A has two eigenvalues: $\lambda_1 = 4$ and $\lambda_2 = 2$.

For each eigenvalue, the corresponding eigenvectors are found by solving the associated homogeneous linear system (8.17). For the first eigenvalue, the corresponding eigenvector equation is

$$(A - 4\,\mathrm{I})\,\mathbf{v} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \text{or} \qquad \begin{aligned} -x + y &= 0, \\ x - y &= 0. \end{aligned}$$

[†] Note that it is not legal to write (8.17) in the form $(A - \lambda)\mathbf{v} = \mathbf{0}$ since we do not know how to subtract a scalar λ from a matrix A. Worse, if you type $A - \lambda$ in MATLAB, it will subtract λ from *all* the entries of A, which is *not* what we are after!

The general solution is

$$x = y = a, \qquad \text{so} \qquad \mathbf{v} = \begin{pmatrix} a \\ a \end{pmatrix} = a \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where a is an arbitrary scalar. Only the nonzero solutions[†] count as eigenvectors, and so the eigenvectors for the eigenvalue $\lambda_1 = 4$ must have $a \neq 0$, i.e., they are all nonzero scalar multiples of the basic eigenvector $\mathbf{v}_1 = (\,1, 1\,)^T$.

Remark: In general, if $\mathbf{v}$ is an eigenvector of A for the eigenvalue λ, then so is any nonzero scalar multiple of $\mathbf{v}$. In practice, we only distinguish linearly independent eigenvectors. Thus, in this example, we shall say "$\mathbf{v}_1 = (\,1, 1\,)^T$ is *the* eigenvector corresponding to the eigenvalue $\lambda_1 = 4$", when we really mean that the eigenvectors for $\lambda_1 = 4$ consist of all nonzero scalar multiples of $\mathbf{v}_1$.

Similarly, for the second eigenvalue $\lambda_2 = 2$, the eigenvector equation is

$$(A - 2\,\mathrm{I}\,)\,\mathbf{v} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The solution $(\,-a, a\,)^T = a\,(\,-1, 1\,)^T$ is the set of scalar multiples of the eigenvector $\mathbf{v}_2 = (\,-1, 1\,)^T$. Therefore, the complete list of eigenvalues and eigenvectors (up to scalar multiple) is

$$\lambda_1 = 4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = 2, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Example 8.8. Consider the 3×3 matrix

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

Using the formula (1.82) for a 3×3 determinant, we compute the characteristic equation

$$\begin{aligned} 0 = \det(A - \lambda\,\mathrm{I}\,) &= \det \begin{pmatrix} -\lambda & -1 & -1 \\ 1 & 2 - \lambda & 1 \\ 1 & 1 & 2 - \lambda \end{pmatrix} \\ &= (-\lambda)(2 - \lambda)^2 + (-1) \cdot 1 \cdot 1 + (-1) \cdot 1 \cdot 1 - \\ &\qquad - 1 \cdot (2 - \lambda)(-1) - 1 \cdot 1 \cdot (-\lambda) - (2 - \lambda) \cdot 1 \cdot (-1) \\ &= -\lambda^3 + 4\,\lambda^2 - 5\,\lambda + 2. \end{aligned}$$

[†] If, at this stage, you end up with a linear system with only the trivial zero solution, you've done something wrong! Either you don't have an correct eigenvalue — maybe you made a mistake setting up and/or solving the characteristic equation — or you've made an error solving the homogeneous eigenvector system.

The resulting cubic polynomial can be factorized:

$$-\lambda^3 + 4\lambda^2 - 5\lambda + 2 = -(\lambda - 1)^2(\lambda - 2) = 0.$$

Most 3×3 matrices have three different eigenvalues, but this particular one has only two: $\lambda_1 = 1$, which is called a double eigenvalue since it is a double root of the characteristic equation, along with a simple eigenvalue $\lambda_2 = 2$.

The eigenvector equation (8.17) for the double eigenvalue $\lambda_1 = 1$ is

$$(A - \mathrm{I})\mathbf{v} = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The general solution to this homogeneous linear system

$$\mathbf{v} = \begin{pmatrix} -a - b \\ a \\ b \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + b \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

depends upon two free variables, $y = a$, $z = b$. Any nonzero solution forms a valid eigenvector for the eigenvalue $\lambda_1 = 1$, and so the general eigenvector is any non-zero linear combination of the two "basis eigenvectors" $\mathbf{v}_1 = (-1, 1, 0)^T$, $\widehat{\mathbf{v}}_1 = (-1, 0, 1)^T$.

On the other hand, the eigenvector equation for the simple eigenvalue $\lambda_2 = 2$ is

$$(A - 2\,\mathrm{I})\mathbf{v} = \begin{pmatrix} -2 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The general solution

$$\mathbf{v} = \begin{pmatrix} -a \\ a \\ a \end{pmatrix} = a \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$$

consists of all scalar multiple of the eigenvector $\mathbf{v}_2 = (-1, 1, 1)^T$.

In summary, the eigenvalues and (basis) eigenvectors for this matrix are

$$\begin{aligned} &\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \widehat{\mathbf{v}}_1 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \\ &\lambda_2 = 2, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}. \end{aligned} \tag{8.19}$$

In general, given an eigenvalue λ, the corresponding *eigenspace* $V_\lambda \subset \mathbb{R}^n$ is the subspace spanned by all its eigenvectors. Equivalently, the eigenspace is the kernel

$$V_\lambda = \ker(A - \lambda\,\mathrm{I}). \tag{8.20}$$

In particular, λ is an eigenvalue if and only if $V_\lambda \neq \{\mathbf{0}\}$ is a nontrivial subspace, and then every nonzero element of V_λ is a corresponding eigenvector. The most economical way to indicate each eigenspace is by writing out a basis, as in (8.19).

Example 8.9. The characteristic equation of the matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$ is

$$0 = \det(A - \lambda\, \mathrm{I}\,) = -\lambda^3 + \lambda^2 + 5\lambda + 3 = -(\lambda + 1)^2 (\lambda - 3).$$

Again, there is a double eigenvalue $\lambda_1 = -1$ and a simple eigenvalue $\lambda_2 = 3$. However, in this case the matrix

$$A - \lambda_1 \mathrm{I} = A + \mathrm{I} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 0 & 1 \\ 2 & 0 & 2 \end{pmatrix}$$

has only a one-dimensional kernel, spanned by $(2, -1, -2)^T$. Thus, even though λ_1 is a double eigenvalue, it only admits a one-dimensional eigenspace. The list of eigenvalues and eigenvectors is, in a sense, incomplete:

$$\lambda_1 = -1, \qquad \mathbf{v}_1 = \begin{pmatrix} 2 \\ -1 \\ -2 \end{pmatrix}, \qquad\qquad \lambda_2 = 3, \qquad \mathbf{v}_2 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}.$$

Example 8.10. Finally, consider the matrix $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & -2 \\ 2 & 2 & -1 \end{pmatrix}$. The characteristic equation is

$$0 = \det(A - \lambda\, \mathrm{I}\,) = -\lambda^3 + \lambda^2 - 3\lambda - 5 = -(\lambda + 1)(\lambda^2 - 2\lambda + 5).$$

The linear factor yields the eigenvalue -1. The quadratic factor leads to two complex roots, $1 + 2\,\mathrm{i}$ and $1 - 2\,\mathrm{i}$, which can be obtained via the quadratic formula. Hence A has one real and two complex eigenvalues:

$$\lambda_1 = -1, \qquad\qquad \lambda_2 = 1 + 2\,\mathrm{i}\,, \qquad\qquad \lambda_3 = 1 - 2\,\mathrm{i}\,.$$

Complex eigenvalues are as important as real eigenvalues, and we need to be able to handle them too. To find the corresponding eigenvectors, which will also be complex, we need to solve the usual eigenvalue equation (8.17), which is now a complex homogeneous linear system. For example, the eigenvector(s) for $\lambda_2 = 1 + 2\,\mathrm{i}$ are found by solving

$$(A - (1 + 2\,\mathrm{i}\,)\,\mathrm{I}\,)\mathbf{v} = \begin{pmatrix} -2\,\mathrm{i} & 2 & 0 \\ 0 & -2\,\mathrm{i} & -2 \\ 2 & 2 & -2 - 2\,\mathrm{i} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This linear system can be solved by Gaussian elimination (with complex pivots). A simpler approach is to work directly: the first equation $-2\,\mathrm{i}\,x + 2y = 0$ tells us that $y = \mathrm{i}\,x$, while the second equation $-2\,\mathrm{i}\,y - 2z = 0$ says $z = -\mathrm{i}\,y = x$. If we trust our calculations so far, we do not need to solve the final equation $2x + 2y + (-2 - 2\,\mathrm{i}\,)z = 0$, since we know that the coefficient matrix is singular and hence it must be a consequence of the first two equations. (However, it does serve as a useful check on our work.) So, the general

solution $\mathbf{v} = (\,x, \mathrm{i}\,x, x\,)^T$ is an arbitrary constant multiple of the complex eigenvector $\mathbf{v}_2 = (\,1, \mathrm{i}\,, 1\,)^T$.

Summarizing, the matrix under consideration has three complex eigenvalues and three corresponding eigenvectors, each unique up to (complex) scalar multiple:

$$\begin{aligned} &\lambda_1 = -1, && \lambda_2 = 1 + 2\,\mathrm{i}\,, && \lambda_3 = 1 - 2\,\mathrm{i}\,, \\ &\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, && \mathbf{v}_2 = \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix}, && \mathbf{v}_3 = \begin{pmatrix} 1 \\ -\mathrm{i} \\ 1 \end{pmatrix}. \end{aligned}$$

Note that the third complex eigenvalue is the complex conjugate of the second, and the eigenvectors are similarly related. This is indicative of a general fact for real matrices:

Proposition 8.11. *If A is a real matrix with a complex eigenvalue $\lambda = \mu + \mathrm{i}\,\nu$ and corresponding complex eigenvector $\mathbf{v} = \mathbf{x} + \mathrm{i}\,\mathbf{y}$, then the complex conjugate $\overline{\lambda} = \mu - \mathrm{i}\,\nu$ is also an eigenvalue with complex conjugate eigenvector $\overline{\mathbf{v}} = \mathbf{x} - \mathrm{i}\,\mathbf{y}$.*

Proof: First take complex conjugates of the eigenvalue equation (8.16)

$$\overline{A}\,\overline{\mathbf{v}} = \overline{A\,\mathbf{v}} = \overline{\lambda\,\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{v}}.$$

Using the fact that a real matrix is unaffected by conjugation, so $\overline{A} = A$, we conclude

$$A\,\overline{\mathbf{v}} = \overline{\lambda}\,\overline{\mathbf{v}}, \tag{8.21}$$

which is the eigenvalue equation for the eigenvalue $\overline{\lambda}$ and eigenvector $\overline{\mathbf{v}}$. *Q.E.D.*

As a consequence, when dealing with real matrices, one only needs to compute the eigenvectors for *one* of each complex conjugate pair of eigenvalues. This observation effectively halves the amount of work in the unfortunate event that we are confronted with complex eigenvalues.

Remark: The reader may recall that we said one should never use determinants in practical computations. So why have we reverted to using determinants to find eigenvalues? The truthful answer is that the practical computation of eigenvalues and eigenvectors *never* resorts to the characteristic equation! There are just too many numerical pitfalls and inefficiencies in (a) computing the determinant, then (b) solving the resulting polynomial equation, and finally (c) solving each of the resulting linear eigenvector systems. Nevertheless, the characteristic equation does give us important theoretical insight into the structure of the eigenvalues of a matrix, and can be used on small, e.g., 2×2 and 3×3, matrices. Numerical algorithms for computing eigenvalues and eigenvectors are based on completely different ideas, and will be discussed in Section 9.6.

Basic Properties of Eigenvalues

If A is an $n \times n$ matrix, then its *characteristic polynomial* is

$$p_A(\lambda) = \det(A - \lambda\,\mathrm{I}\,) = c_n\,\lambda^n + c_{n-1}\,\lambda^{n-1} + \,\cdots\, + c_1\,\lambda + c_0. \tag{8.22}$$

The fact that $p_A(\lambda)$ is a polynomial of degree n is a consequence of the general determinantal formula (1.81). Indeed, every term is plus or minus a product of matrix entries containing one from each row and one from each column. The term corresponding to the identity permutation is obtained by multiplying the the diagonal entries together, which, in this case, is

$$(a_{11}-\lambda)\,(a_{22}-\lambda)\,\cdots\,(a_{nn}-\lambda) = (-1)^n\lambda^n + (-1)^{n-1}\big(\,a_{11}+a_{22}+\,\cdots\,+a_{nn}\,\big)\,\lambda^{n-1} + \cdots\,, \tag{8.23}$$

All of the other terms have at most $n-2$ diagonal factors $a_{ii}-\lambda$, and so are polynomials of degree $\le n-2$ in λ. Thus, (8.23) is the only summand containing the monomials λ^n and λ^{n-1}, and so their respective coefficients are

$$c_n = (-1)^n, \qquad c_{n-1} = (-1)^{n-1}(a_{11}+a_{22}+\,\cdots\,+a_{nn}) = (-1)^{n-1}\,\mathrm{tr}\,A, \tag{8.24}$$

where $\mathrm{tr}\,A$, the sum of its diagonal entries, is called the *trace* of the matrix A. The other coefficients $c_{n-2},\dots,c_1$ in (8.22) are more complicated combinations of the entries of A. However, setting $\lambda = 0$ implies $p_A(0) = \det A = c_0$, and hence the constant term equals the determinant of the matrix. In particular, if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a 2×2 matrix, its characteristic polynomial has the form

$$\begin{aligned} p_A(\lambda) = \det(A - \lambda\,\mathrm{I}\,) &= \det\begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} \\ &= \lambda^2 - (a+d)\lambda + (a\,d - b\,c) = \lambda^2 - (\mathrm{tr}\,A)\lambda + (\det A). \end{aligned} \tag{8.25}$$

As a result of these considerations, the characteristic equation of an $n\times n$ matrix A is a polynomial equation of degree n, namely $p_A(\lambda) = 0$. According to the Fundamental Theorem of Algebra (see Corollary 15.63) every (complex) polynomial of degree n can be completely factored:

$$p_A(\lambda) = (-1)^n(\lambda-\lambda_1)(\lambda-\lambda_2)\,\cdots\,(\lambda-\lambda_n). \tag{8.26}$$

The complex numbers $\lambda_1,\dots,\lambda_n$, some of which may be repeated, are the *roots* of the characteristic equation $p_A(\lambda) = 0$, and hence the eigenvalues of the matrix A. Therefore, we immediately conclude:

Theorem 8.12. *An $n\times n$ matrix A has at least one and at most n distinct complex eigenvalues.*

Most $n\times n$ matrices — meaning those for which the characteristic polynomial factors into n *distinct* factors — have *exactly* n complex eigenvalues. More generally, an eigenvalue λ_j is said to have *multiplicity* m if the factor $(\lambda-\lambda_j)$ appears exactly m times in the factorization (8.26) of the characteristic polynomial. An eigenvalue is *simple* if it has multiplicity 1. In particular, A has n distinct eigenvalues if and only if all its eigenvalues are simple. In all cases, when the eigenvalues are counted in accordance with their multiplicity, every $n\times n$ matrix has a total of n possibly repeated eigenvalues.

An example of a matrix with just one eigenvalue, of multiplicity n, is the $n\times n$ identity matrix I, whose only eigenvalue is $\lambda = 1$. In this case, *every* nonzero vector in $\mathbb{R}^n$ is an

eigenvector of the identity matrix, and so the eigenspace is all of $\mathbb{R}^n$. At the other extreme, the "bidiagonal" *Jordan block matrix*

$$J_\lambda = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \tag{8.27}$$

also has only one eigenvalue, λ, again of multiplicity n. But in this case, J_λ has only one eigenvector (up to scalar multiple), which is the standard basis vector $\mathbf{e}_n$, and so its eigenspace is one-dimensional.

Remark: If λ is a complex eigenvalue of multiplicity k for the real matrix A, then its complex conjugate $\overline{\lambda}$ also has multiplicity k. This is because complex conjugate roots of a real polynomial necessarily appear with identical multiplicities.

Remark: If $n \leq 4$, then one can, in fact, write down an explicit formula for the solution to a polynomial equation of degree n, and hence explicit (but not particularly helpful) formulae for the eigenvalues of general $2 \times 2, 3 \times 3$ and 4×4 matrices. As soon as $n \geq 5$, there is no explicit formula (at least in terms of radicals), and so one must usually resort to numerical approximations. This remarkable and deep algebraic result was proved by the young Norwegian mathematician Nils Hendrik Abel in the early part of the nineteenth century, [**53**].

If we explicitly multiply out the factored product (8.26) and equate the result to the characteristic polynomial (8.22), we find that its coefficients $c_0, c_1, \dots c_{n-1}$ can be written as certain polynomials of the roots, known as the *elementary symmetric polynomials*. The first and last are of particular importance:

$$c_0 = \lambda_1 \lambda_2 \cdots \lambda_n, \qquad c_{n-1} = (-1)^{n-1} (\lambda_1 + \lambda_2 + \cdots + \lambda_n). \tag{8.28}$$

Comparison with our previous formulae for the coefficients c_0 and c_{n-1} leads us to the following useful result.

Proposition 8.13. *The sum of the eigenvalues of a matrix equals its trace*:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{tr} A = a_{11} + a_{22} + \cdots + a_{nn}. \tag{8.29}$$

The product of the eigenvalues equals its determinant:

$$\lambda_1 \lambda_2 \cdots \lambda_n = \det A. \tag{8.30}$$

Remark: For repeated eigenvalues, one must add or multiply them in the formulae (8.29), (8.30) according to their multiplicity.

Example 8.14. The matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ 2 & 0 & 1 \end{pmatrix}$ considered in Example 8.9 has trace and determinant

$$\operatorname{tr} A = 1, \qquad \det A = 3.$$

These fix, respectively, the coefficient of λ^2 and the constant term in the characteristic equation. This matrix has two distinct eigenvalues, -1, which is a double eigenvalue, and 3, which is simple. For this particular matrix, formulae (8.29), (8.30) become

$$1 = \operatorname{tr} A = (-1) + (-1) + 3, \qquad 3 = \det A = (-1)(-1)\,3.$$

8.3. Eigenvector Bases and Diagonalization.

Most of the vector space bases that play a distinguished role in applications consist of eigenvectors of a particular matrix. In this section, we show that the eigenvectors for any "complete" matrix automatically form a basis for $\mathbb{R}^n$ or, in the complex case, $\mathbb{C}^n$. In the following subsection, we use the eigenvector basis to rewrite the linear transformation determined by the matrix in a simple diagonal form.

The first task is to show that eigenvectors corresponding to distinct eigenvalues are automatically linearly independent.

Lemma 8.15. *If $\lambda_1, \dots, \lambda_k$ are* distinct *eigenvalues of the same matrix A, then the corresponding eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ are linearly independent.*

Proof: We use induction on the number of eigenvalues. The case $k = 1$ is immediate since an eigenvector cannot be zero. Assume that we know the result for $k-1$ eigenvalues. Suppose we have a linear combination

$$c_1\,\mathbf{v}_1 + \,\cdots\, + c_{k-1}\,\mathbf{v}_{k-1} + c_k\,\mathbf{v}_k = \mathbf{0} \tag{8.31}$$

which vanishes. Let us multiply this equation by the matrix A:

$$\begin{aligned} A\big(\,c_1\,\mathbf{v}_1 + \,\cdots\, + c_{k-1}\,\mathbf{v}_{k-1} + c_k\,\mathbf{v}_k\,\big) &= c_1\,A\,\mathbf{v}_1 + \,\cdots\, + c_{k-1}\,A\,\mathbf{v}_{k-1} + c_k\,A\,\mathbf{v}_k \\ &= c_1\,\lambda_1\,\mathbf{v}_1 + \,\cdots\, + c_{k-1}\,\lambda_{k-1}\,\mathbf{v}_{k-1} + c_k\,\lambda_k\,\mathbf{v}_k = \mathbf{0}. \end{aligned}$$

On the other hand, if we just multiply the original equation by λ_k, we also have

$$c_1\,\lambda_k\,\mathbf{v}_1 + \,\cdots\, + c_{k-1}\,\lambda_k\,\mathbf{v}_{k-1} + c_k\,\lambda_k\,\mathbf{v}_k = \mathbf{0}.$$

Subtracting this from the previous equation, the final terms cancel and we are left with the equation

$$c_1(\lambda_1 - \lambda_k)\mathbf{v}_1 + \,\cdots\, + c_{k-1}(\lambda_{k-1} - \lambda_k)\mathbf{v}_{k-1} = \mathbf{0}.$$

This is a vanishing linear combination of the first $k-1$ eigenvectors, and so, by our induction hypothesis, can only happen if all the coefficients are zero:

$$c_1(\lambda_1 - \lambda_k) = 0, \qquad \dots \qquad c_{k-1}(\lambda_{k-1} - \lambda_k) = 0.$$

The eigenvalues were assumed to be distinct, so $\lambda_j \neq \lambda_k$ when $j \neq k$; consequently, $c_1 = \cdots = c_{k-1} = 0$. Substituting these values back into (8.31), we find $c_k \mathbf{v}_k = \mathbf{0}$, and so $c_k = 0$ also, since the eigenvector $\mathbf{v}_k \neq \mathbf{0}$. Thus we have proved that (8.31) holds if and only if $c_1 = \cdots = c_k = 0$, which implies the linear independence of the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$. This completes the induction step. *Q.E.D.*

The most important consequence of this result is stated in the corollary.

Theorem 8.16. *If the $n \times n$ real matrix A has n distinct real eigenvalues $\lambda_1, \ldots, \lambda_n$, then the corresponding real eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis of $\mathbb{R}^n$. If A (which may now be either a real or a complex matrix) has n distinct complex eigenvalues, then its eigenvectors form a basis of $\mathbb{C}^n$.*

If a matrix has multiple eigenvalues, then there may or may not be an eigenvector basis of $\mathbb{R}^n$ (or $\mathbb{C}^n$). The matrix in Example 8.8 has an eigenvector basis, whereas the matrix in Example 8.9 does not. In general, it can be proved that the dimension of the eigenspace is less than or equal to the multiplicity of the eigenvalue. In particular, a simple eigenvalue has a one-dimensional eigenspace, and hence, up to scalar multiple, only one associated eigenvector.

Definition 8.17. An eigenvalue λ of a matrix A is called *complete* if its eigenspace $V_\lambda = \ker(A - \lambda \, \mathrm{I})$ has the same dimension as its multiplicity. The matrix A is *complete* if all its eigenvalues are.

Remark: The multiplicity of an eigenvalue λ_i is sometimes referred to as its *algebraic multiplicity*. The dimension of the eigenspace V_λ is called the *geometric multiplicity*, and so completeness requires that the two multiplicities are equal.

Note that a simple eigenvalue is automatically complete, and so only multiple eigenvalues can cause the incompleteness of a matrix.

Theorem 8.18. *An $n \times n$ real or complex matrix A is complete if and only if its eigenvectors span $\mathbb{C}^n$. In particular, any $n \times n$ matrix that has n distinct eigenvalues is complete.*

A $n \times n$ matrix is *incomplete* if it does not have n linearly independent complex eigenvectors. Most matrices, including those with all simple eigenvalues, are complete. Incomplete matrices are more tricky to deal with, and we relegate most of the messy details to Section 8.4.

Remark: We already noted that complex eigenvectors of a real matrix always appear in conjugate pairs: $\mathbf{v} = \mathbf{x} \pm \mathrm{i}\,\mathbf{y}$. It can be shown that the real and imaginary parts of these vectors form a real basis for $\mathbb{R}^n$. (See Exercise ■ for the underlying principle.) For instance, in Example 8.10, the complex eigenvectors are $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \pm \mathrm{i} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$. The vectors

$\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, consisting of the real eigenvector and the real and imaginary parts of the complex eigenvectors, form a basis for $\mathbb{R}^3$.

Diagonalization

Every $n \times n$ matrix A represents a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^n$, namely the function $L[\mathbf{u}] = A\,\mathbf{u}$ given by matrix multiplication. As we learned in Section 7.2, the matrix representing a linear transformation depends upon the choice basis of $\mathbb{R}^n$. Some bases give a particular simple matrix representation.

For example, the linear transformation $L\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x - y \\ 2x - 4y \end{pmatrix}$ studied in Example 7.18 is represented by the matrix $A = \begin{pmatrix} 1 & -1 \\ 2 & 4 \end{pmatrix}$ — when expressed in terms of the standard basis of $\mathbb{R}^2$. In terms of the alternative basis $\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$, the linear transformation is represented by the diagonal matrix $\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. This follows from the action of the linear transformation on the new basis: $A\,\mathbf{v}_1 = 2\mathbf{v}_1$ and $A\,\mathbf{v}_2 = 3\mathbf{v}_2$. Now we can understand the reason for this simplification. *The new basis consists of the two eigenvectors of the matrix A.* This observation is indicative of a general fact: representing a linear transformation in terms of an eigenvector basis has the effect of replacing its matrix representative by a simple diagonal form. The effect is to *diagonalize* the original coefficient matrix.

According to (7.22), if $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a basis of $\mathbb{R}^n$, then the corresponding matrix representative of the linear transformation $L[\mathbf{v}] = A\,\mathbf{v}$ is given by the similar matrix $B = S^{-1}A\,S$, where $S = (\,\mathbf{v}_1\ \mathbf{v}_2\ \dots\ \mathbf{v}_n\,)^T$ is the matrix whose columns are the basis vectors. In the preceding example, $S = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$, and we find that $S^{-1}A\,S = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$ is a diagonal matrix.

Definition 8.19. A square matrix A is called *diagonalizable* if there exists a nonsingular matrix S and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ such that

$$S^{-1}A\,S = \Lambda. \tag{8.32}$$

A diagonal matrix represents a linear transformation that simultaneously stretches[†] in the direction of the basis vectors. Thus, every diagonalizable matrix represents a elementary combination of (complex) stretching transformations.

To understand the diagonalization equation (8.32), we rewrite it in the equivalent form

$$A\,S = S\,\Lambda. \tag{8.33}$$

[†] A negative diagonal entry represents the combination of a reflection and stretch. Complex entries correspond to a complex stretching transformation. See Section 7.2 for details.

Using the basic property (1.11) of matrix multiplication, one easily sees that the k^{th} column of this $n \times n$ matrix equation is given by

$$A\,\mathbf{v}_k = \lambda_k \mathbf{v}_k.$$

Therefore, the columns of S are necessarily eigenvectors, and the entries of the diagonal matrix Λ are the corresponding eigenvalues! And, as a result, a diagonalizable matrix A must have n linearly independent eigenvectors, i.e., an eigenvector basis, to form the columns of the diagonalizing matrix S. Since the diagonal form Λ contains the eigenvalues along its diagonal, it is uniquely determined up to a permutation of its entries.

Now, as we know, not every matrix has an eigenvector basis. Moreover, even when it exists, the eigenvector basis may be complex, in which case S is a complex matrix, and the entries of the diagonal matrix Λ are the complex eigenvalues. Thus, we should distinguish between matrices that are diagonalizable over the complex numbers and the more restrictive class of matrices which can be diagonalized by a real matrix S.

We have now proved the following important result.

Theorem 8.20. *A matrix is complex diagonalizable if and only if it is complete. A matrix is real diagonalizable if and only if it is complete and has all real eigenvalues.*

Remark: Many authors use the term "diagonalizable" for what we have called complete matrices.

Example 8.21. The 3×3 matrix $A = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ considered in Example 8.7 has eigenvector basis

$$\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

We assemble these to form the eigenvector matrix

$$S = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \qquad \text{and so} \qquad S^{-1} = \begin{pmatrix} -1 & 0 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

The diagonalization equation (8.32) becomes

$$S^{-1} A\, S = \begin{pmatrix} -1 & 0 & -1 \\ -1 & -1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \Lambda,$$

with the eigenvalues of A appearing on the diagonal of Λ, in the same order as the eigenvectors.

Remark: If a matrix is not complete, then it cannot be diagonalized. Incomplete matrices represent generalized shearing transformations, and will be the subject of the following subsection. A simple example is a matrix of the form $\begin{pmatrix} 1 & c \\ 0 & 1 \end{pmatrix}$ for $c \neq 0$, which represents a shear in the direction of the x axis.

8.4. Incomplete Matrices and the Jordan Canonical Form.

Unfortunately, not all matrices are complete. Matrices without any eigenvector basis are considerably less convenient to deal with. However, as they occasionally appear in applications, it is worth learning how to handle them. The key is to supplement the eigenvectors in order to obtain a basis in which the matrix assumes a simple, but now non-diagonal form.

Throughout this section, A will be an $n \times n$ matrix, with either real or complex entries. We let $\lambda_1, \dots, \lambda_k$ denote the distinct eigenvalues of A. We recall that Theorem 8.12 guarantees that every matrix has at least one (complex) eigenvalue, so $k \geq 1$.

Definition 8.22. A *Jordan chain* of length j is a sequence of non-zero vectors $w_1, \dots, w_j \in \mathbb{C}^m$ that satisfies

$$A\,\mathbf{w}_1 = \lambda\,\mathbf{w}_1, \qquad A\,\mathbf{w}_i = \lambda\,\mathbf{w}_i + \mathbf{w}_{i-1}, \qquad i = 2, \dots, j, \tag{8.34}$$

where λ is an eigenvalue of A.

Note that the initial vector $\mathbf{w}_1$ in a Jordan chain is a genuine eigenvector. The others, $\mathbf{w}_2, \dots, \mathbf{w}_j$, are *generalized eigenvectors*, in accordance with the following definition.

Definition 8.23. A nonzero vector $\mathbf{w} \neq \mathbf{0}$ that satisfies

$$(A - \lambda\,\mathrm{I}\,)^k\,\mathbf{w} = \mathbf{0} \tag{8.35}$$

for some $k > 0$ and $\lambda \in \mathbb{C}$ is called a *generalized eigenvector* of the matrix A.

Note that every ordinary eigenvector is automatically a generalized eigenvector, since we can just take $k = 1$ in (8.35), but the converse is not necessarily valid. We shall call the minimal value of k for which (8.35) holds the *index* of the generalized eigenvector. Thus, an ordinary eigenvector is a generalized eigenvector of index 1. Since $A - \lambda\,\mathrm{I}$ is nonsingular whenever λ is not an eigenvalue of A, its k^{th} power $(A - \lambda\,\mathrm{I}\,)^k$ is also nonsingular. Therefore, generalized eigenvectors can only exist when λ is an ordinary eigenvalue of A — there are no "generalized eigenvalues".

Lemma 8.24. *The i^{th} vector $\mathbf{w}_i$ in a Jordan chain* (8.34) *is a generalized eigenvector of index i.*

Proof: By definition, $(A - \lambda\,\mathrm{I}\,)\,\mathbf{w}_1 = \mathbf{0}$, and so $\mathbf{w}_1$ is an eigenvector. Next, we have $(A - \lambda\,\mathrm{I}\,)\,\mathbf{w}_2 = \mathbf{w}_1$, and so $(A - \lambda\,\mathrm{I}\,)^2\,\mathbf{w}_2 = (A - \lambda\,\mathrm{I}\,)\,\mathbf{w}_1 = \mathbf{0}$. Thus, $\mathbf{w}_2$ a generalized eigenvector of index 2. A simple induction proves that $(A - \lambda\,\mathrm{I}\,)^i\,\mathbf{w}_i = \mathbf{0}$. *Q.E.D.*

Example 8.25. Consider the 3×3 Jordan block $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$. The only eigenvalue is $\lambda = 2$, and $A - 2\,\mathrm{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$. We claim that the standard basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ form a Jordan chain. Indeed, $A\mathbf{e}_1 = 2\mathbf{e}_1$, and hence $\mathbf{e}_1 \in \ker(A - 2\,\mathrm{I})$ is a genuine eigenvector. Furthermore, $A\mathbf{e}_2 = 2\mathbf{e}_2 + \mathbf{e}_1$, and $A\mathbf{e}_3 = 2\mathbf{e}_3 + \mathbf{e}_2$, as you can easily check. Thus, $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ satisfy the Jordan chain equations for the eigenvalue $\lambda = 2$. Note that $\mathbf{e}_2$ lies in the kernel of $(A - 2\,\mathrm{I})^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$, and so is a generalized eigenvector of index 2. Indeed, every vector of the form $\mathbf{w} = a\mathbf{e}_1 + b\mathbf{e}_2$ with $b \neq 0$ is a generalized eigenvector of index 2. (When $b = 0$, $a \neq 0$, the vector $\mathbf{w} = a\mathbf{e}_1$ is an ordinary eigenvector of index 1.) Finally, $(A - 2\,\mathrm{I})^3 = \mathrm{O}$, and so every vector $\mathbf{v} \in \mathbb{R}^3$, including $\mathbf{e}_3$, is a generalized eigenvector of index 3 (or less).

Given a matrix A, a basis of $\mathbb{R}^n$ or $\mathbb{C}^n$ is called a *Jordan basis* if it consists of one or more nonoverlapping Jordan chains. Thus, for the Jordan matrix in Example 8.25, the standard basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is, in fact, a Jordan basis. An eigenvector basis qualifies as a Jordan basis, since each eigenvector belongs to a Jordan chain of length 1. Jordan bases are the generalization of eigenvector bases for incomplete matrices that we are after.

Theorem 8.26. *Every $n \times n$ matrix admits a Jordan basis of $\mathbb{C}^n$. The first elements of the Jordan chains form a maximal system of linearly independent eigenvectors. Moreover, the number of generalized eigenvectors in the Jordan basis associated with an eigenvalue λ is the same as its multiplicity.*

Example 8.27. Consider the matrix $A = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}$. With some work, its characteristic equation is found to be

$$p_A(\lambda) = \det(A - \lambda\,\mathrm{I}) = \lambda^5 + \lambda^4 - 5\lambda^3 - \lambda^2 + 8\lambda - 4 = (\lambda - 1)^3(\lambda + 2)^2 = 0,$$

and hence A has two eigenvalues: 1, which is a triple eigenvalue, and -2, which is double. Solving the associated homogeneous system $(A - \lambda\,\mathrm{I})\mathbf{v} = \mathbf{0}$, we discover that, up to constant multiple, there are only two eigenvectors: $\mathbf{v}_1 = (\,0, 0, 0, -1, 1\,)^T$ for $\lambda_1 = 1$ and, anticipating our final numbering, $\mathbf{v}_4 = (\,-1, 1, 1, -2, 0\,)^T$ for $\lambda_2 = -2$. Thus, A is far from complete.

To construct a Jordan basis, we first note that since A has 2 linearly independent eigenvectors, the Jordan basis will contain two Jordan chains; the one associated with the triple eigenvalue $\lambda_1 = 1$ has length 3, while $\lambda_2 = -2$ admits a length 2 Jordan chain. To construct the former, we need to first solve the system $(A - \mathrm{I})\mathbf{w} = \mathbf{v}_1$. Note that the coefficient matrix is singular — it must be since 1 is an eigenvalue — and the general

solution is $\mathbf{w} = \mathbf{v}_2 + t\,\mathbf{v}_1$ where $\mathbf{v}_2 = (\,0,1,0,0,-1\,)^T$, and t is the free variable. The appearance of an arbitrary multiple of the eigenvector $\mathbf{v}_1$ in the solution is not unexpected; indeed, the kernel of $A - \mathrm{I}$ is the eigenspace for $\lambda_1 = 1$. We can choose any solution, e.g., $\mathbf{v}_2$ as the second element in the Jordan chain. To find the last element of the chain, we solve $(A - \mathrm{I}\,)\mathbf{w} = \mathbf{v}_2$ to find $\mathbf{w} = \mathbf{v}_3 + t\,\mathbf{v}_1$ where $\mathbf{v}_3 = (\,0,0,0,1,0\,)^T$ can be used as the Jordan chain element. Similarly, to construct the Jordan chain for the second eigenvalue, we solve $(A + 2\,\mathrm{I}\,)\mathbf{w} = \mathbf{v}_4$ and find $\mathbf{w} = \mathbf{v}_5 + t\,\mathbf{v}_4$ where $\mathbf{v}_5 = (\,-1,0,0,-2,1\,)^T$. Thus, the desired Jordan basis is

$$\mathbf{v}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -2 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_5 = \begin{pmatrix} -1 \\ 0 \\ 0 \\ -2 \\ 1 \end{pmatrix},$$

with $A\,\mathbf{v}_1 = \mathbf{v}_1$, $A\,\mathbf{v}_2 = \mathbf{v}_1 + \mathbf{v}_2$, $A\,\mathbf{v}_3 = \mathbf{v}_2 + \mathbf{v}_2$, $A\,\mathbf{v}_4 = -2\,\mathbf{v}_4$, $A\,\mathbf{v}_5 = \mathbf{v}_4 - 2\,\mathbf{v}_5$.

To prove Theorem 8.26, we begin with a simple lemma.

Lemma 8.28. *If $\mathbf{v}_1, \dots, \mathbf{v}_n$ forms a Jordan basis for the matrix A, it also forms a Jordan basis for $B = A - c\,\mathrm{I}$, for any scalar c.*

Proof: We note that the eigenvalues of B are of the form $\lambda - c$, where λ is an eigenvalue of A. Moreover, given a Jordan chain $\mathbf{w}_1, \dots, \mathbf{w}_j$ of A, we have

$$B\,\mathbf{w}_1 = (\lambda - c)\,\mathbf{w}_1, \qquad B\,\mathbf{w}_i = (\lambda - c)\,\mathbf{w}_i + \mathbf{w}_{i-1}, \qquad i = 2, \dots, j,$$

so $\mathbf{w}_1, \dots, \mathbf{w}_j$ is also a Jordan chain for B corresponding to the eigenvalue $\lambda - c$. *Q.E.D.*

The proof of Theorem 8.26 will be done by induction on the size n of the matrix. The case $n = 1$ is trivial, since any basis of $\mathbb{C}$ is a Jordan basis for a 1×1 matrix $A = (a)$. To perform the induction step, we assume that the result is valid for all matrices of size $\leq n - 1$. Let A be an $n \times n$ matrix. According to Theorem 8.12, A has at least one complex eigenvalue λ. Let $B = A - \lambda\,\mathrm{I}$. Since λ is an eigenvalue of A, we know that 0 is an eigenvalue of B. This means that $\ker B \neq \{\mathbf{0}\}$, and so $r = \operatorname{rank} B < n$. Moreover, by Lemma 8.28, any Jordan basis of B is also a Jordan basis for A, and so we can concentrate all our attention on the singular matrix B from now on.

According to Exercise ■, $W = \operatorname{rng} B \subset \mathbb{C}^n$ is an invariant subspace, i.e., $B\,\mathbf{w} \in W$ whenever $\mathbf{w} \in W$. Moreover, since B is singular, its range has dimension $\dim W = r = \operatorname{rank} B < n$, cf. Theorem 2.47. Thus, by fixing a basis of W, we can realize the restriction $B\colon W \to W$ as multiplication by an $r \times r$ matrix. The fact that $r < n$ allows us to invoke the induction hypothesis, and deduce the existence of a Jordan basis $\mathbf{w}_1, \dots, \mathbf{w}_r \in W \subset \mathbb{C}^n$ for B restricted to W. Our goal is to complete this collection of vectors to a full Jordan basis for B on $\mathbb{C}^n$.

To this end, we append two additional kinds of vectors. Suppose that the Jordan basis of W contains k null Jordan chains associated with the zero eigenvalue. Each null Jordan chain consists of vectors $\mathbf{w}_1, \dots, \mathbf{w}_j \in W$ satisfying

$$B\,\mathbf{w}_1 = \mathbf{0}, \qquad B\,\mathbf{w}_2 = \mathbf{w}_1, \qquad \dots \qquad B\,\mathbf{w}_j = \mathbf{w}_{j-1}. \tag{8.36}$$

The number of null Jordan chains is equal to the number of linearly independent null eigenvectors of B in $W = \operatorname{rng} B$, that is $k = \dim(\ker B \cap \operatorname{rng} B)$. To each null Jordan chain $\mathbf{w}_1, \dots, \mathbf{w}_j \in W$, we append a vector $\mathbf{w}_{j+1} \in \mathbb{C}^n$ such that

$$B\,\mathbf{w}_{j+1} = \mathbf{w}_j; \tag{8.37}$$

the existence of $\mathbf{w}_{j+1}$ comes from our condition that $\mathbf{w}_j \in \operatorname{rng} B$. Appending (8.37) to (8.36), we deduce that $\mathbf{w}_1, \dots, \mathbf{w}_{j+1} \in \mathbb{C}^n$ forms a null Jordan chain, of length $j+1$, for B. The resulting collection contains $r + k$ vectors in $\mathbb{C}^n$ arranged in nonoverlapping Jordan chains. To complete to a basis, we include $n - r - k$ additional linearly independent null vectors $\mathbf{z}_1, \dots, \mathbf{z}_{n-r-k} \in \ker B \setminus \operatorname{rng} B$ that lie outside its range. Since $B\,\mathbf{z}_j = \mathbf{0}$, each $\mathbf{z}_j$ forms a null Jordan chain of length 1. We claim that the complete collection consisting of the non-null Jordan chains, the k extended null chains, and the additional null vectors $\mathbf{z}_1, \dots, \mathbf{z}_{n-r-k}$ forms the desired Jordan basis. By construction, it consists of nonoverlapping Jordan chains. The only remaining technical issue is proving that the vectors are linear independent, which is left as a challenge for the reader in Exercise ∎.

Just as an eigenvector basis diagonalizes a complete matrix, a Jordan basis provides a particularly simple form for an incomplete matrix, known as the *Jordan canonical form*.

Definition 8.29. A $n \times n$ matrix of the form[†]

$$J_{\lambda,n} = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \tag{8.38}$$

in which λ is a real or complex number, is known as a *Jordan block*.

In particular, a 1×1 Jordan block is merely a scalar $J_{\lambda,1} = \lambda$. Since every matrix has at least one (complex) eigenvector the Jordan block matrices have the least possible number of eigenvectors.

Lemma 8.30. *The $n \times n$ Jordan block matrix $J_{\lambda,n}$ has a single eigenvalue, λ, and a single independent eigenvector, $\mathbf{e}_1$. The standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ form a Jordan chain for $J_{\lambda,n}$.*

Definition 8.31. A *Jordan matrix* is a square matrix of block diagonal form

$$J = \operatorname{diag}(J_{\lambda_1,n_1}, J_{\lambda_2,n_2}, \dots, J_{\lambda_k,n_k}) = \begin{pmatrix} J_{\lambda_1,n_1} & & & \\ & J_{\lambda_2,n_2} & & \\ & & \ddots & \\ & & & J_{\lambda_k,n_k} \end{pmatrix}, \tag{8.39}$$

[†] All non-displayed entries are zero.

in which one or more Jordan blocks, not necessarily of the same size, lie along the diagonal, while all off-diagonal blocks are zero.

Note that the only non-zero entries in a Jordan matrix are those on the diagonal, which can have any complex value, and those on the superdiagonal, which are either 1 or 0. The positions of the superdiagonal 1's uniquely prescribes the Jordan blocks.

For example, the 6×6 matrices

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

are all Jordan matrices; the first is a diagonal matrix, consisting of 6 distinct 1×1 Jordan blocks; the second has a 4×4 Jordan block followed by a 2×2 block that happen to have the same diagonal entries; the last has three 2×2 Jordan blocks.

As a simple corollary of Lemma 8.30 combined with the block structure, as outlined in Exercise ∎, we obtain a complete classification of the eigenvectors and eigenvalues of a Jordan matrix.

Lemma 8.32. *The Jordan matrix* (8.39) *has eigenvalues* $\lambda_1, \dots, \lambda_k$. *The standard basis vectors* $\mathbf{e}_1, \dots, \mathbf{e}_n$ *form a Jordan basis; the Jordan chains are labeled by the Jordan blocks.*

Thus, in the preceding examples of Jordan matrices, the first has three double eigenvalues, $1, 2, 3$, and corresponding linearly independent eigenvectors $\mathbf{e}_1, \mathbf{e}_6; \mathbf{e}_2, \mathbf{e}_5; \mathbf{e}_3, \mathbf{e}_4$, each of which belongs to a Jordan chain of length 1. The second matrix has only one eigenvalue, -1, but two Jordan chains, namely $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ and $\mathbf{e}_5, \mathbf{e}_6$. The last has eigenvalues $0, 1, 2$ and three Jordan chains, namely $\mathbf{e}_1, \mathbf{e}_2$, and $\mathbf{e}_3, \mathbf{e}_4$, and $\mathbf{e}_5, \mathbf{e}_6$. In particular, the only complete Jordan matrices are the diagonal matrices, all of whose Jordan blocks are of size 1×1.

Theorem 8.33. *Let* A *be an* $n \times n$ *real or complex matrix. Let* $S = (\, \mathbf{w}_1 \ \mathbf{w}_2 \ \dots \ \mathbf{w}_n \,)^T$ *be the matrix whose columns are a Jordan basis of* A. *Then* S *places* A *in* Jordan canonical form

$$S^{-1} A S = J = \operatorname{diag}(J_{\lambda_1, n_1}, J_{\lambda_2, n_2}, \dots, J_{\lambda_k, n_k}). \tag{8.40}$$

The diagonal entries of the resulting Jordan matrix J *are the eigenvalues of* A. *The Jordan canonical form of* A *is uniquely determined up to a permutation of the diagonal Jordan blocks. In particular,* A *is complete* (*diagonalizable*) *if and only if every Jordan block is of size* 1×1 *or, equivalently, all Jordan chains are of length* 1.

For instance, the matrix $A = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}$ considered in Example 8.27

has the following Jordan basis matrix and Jordan canonical form

$$S = \begin{pmatrix} 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 1 & -2 & -2 \\ 1 & -1 & 0 & 0 & 1 \end{pmatrix}, \qquad J = S^{-1}AS = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 0 & -2 \end{pmatrix}.$$

8.5. Eigenvalues of Symmetric Matrices.

Fortunately, the matrices that arise in most applications possess some additional structure that ameliorates the calculation of their eigenvalues and eigenvectors. The most prevalent are the real diagonalizable matrices, i.e., complete matrices that have only real eigenvalues and hence a real eigenvector basis. The most important class of matrices with this property are the symmetric, including the positive definite, matrices. In fact, not only are the eigenvalues of a symmetric matrix necessarily real, the eigenvectors always form an *orthogonal basis*. In such situations, we can tap into the dramatic simplification and power of orthogonal bases as developed in Chapter 5. In fact, this is by far the most common way for orthogonal bases to appear — as the eigenvector bases of a symmetric matrices.

Theorem 8.34. *If $A = A^T$ be a real symmetric $n \times n$ matrix, Then*

(a) *All the eigenvalues of A are real.*

(b) *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

(c) *There is an orthonormal basis of $\mathbb{R}^n$ consisting of n eigenvectors of A.*

In particular, all symmetric matrices are complete.

Remark: Orthogonality is with respect to the standard dot product on $\mathbb{R}^n$. As we noted in Section 7.5, the transpose operation is intimately connected with the dot product. Introducing a more general inner product on $\mathbb{R}^n$ leads to the concept of a self-adjoint linear transformation, and an analogous result holds in this more general context; see Exercise ■.

Example 8.35. The 2×2 matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ considered in Example 8.7 is symmetric, and so has real eigenvalues $\lambda_1 = 4$ and $\lambda_2 = 2$. You can easily check that the corresponding eigenvectors $\mathbf{v}_1 = (\,1 \;\; 1\,)^T$, $\mathbf{v}_2 = (\,-1 \;\; 1\,)^T$ are orthogonal: $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, and hence form on orthogonal basis of $\mathbb{R}^2$. An orthonormal basis is provided by the unit eigenvectors

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \tag{8.41}$$

obtained by dividing each eigenvector by its length: $\mathbf{u}_k = \mathbf{v}_k / \| \mathbf{v}_k \|$.

Proof of Theorem 8.34: First note that if $A = A^T$ is real, symmetric, then

$$(A\,\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\,\mathbf{w}) \qquad \text{for all} \qquad \mathbf{v}, \mathbf{w} \in \mathbb{C}^n, \tag{8.42}$$

where we use the Euclidean dot product for real vectors and, more generally, the Hermitian dot product $\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^T \overline{\mathbf{w}}$ when they are complex. (See Exercise ■.)

To prove property (a), suppose λ is a complex eigenvalue with complex eigenvector $\mathbf{v} \in \mathbb{C}^n$. Consider the Hermitian dot product of the complex vectors $A\mathbf{v}$ and $\mathbf{v}$:

$$(A\mathbf{v}) \cdot \mathbf{v} = (\lambda \mathbf{v}) \cdot \mathbf{v} = \lambda \, \| \mathbf{v} \|^2.$$

On the other hand, by (8.42), since $A^T = A$ is a real matrix, (8.21) implies

$$(A\mathbf{v}) \cdot \mathbf{v} = \mathbf{v} \cdot (A\mathbf{v}) = \mathbf{v} \cdot (\lambda \mathbf{v}) = \mathbf{v}^T \, \overline{\lambda \mathbf{v}} = \overline{\lambda} \, \| \mathbf{v} \|^2.$$

Equating these two expressions, we deduce

$$\overline{\lambda} \, \| \mathbf{v} \|^2 = \lambda \, \| \mathbf{v} \|^2.$$

Since $\mathbf{v}$ is an eigenvector, it is nonzero, $\mathbf{v} \neq \mathbf{0}$, and so $\overline{\lambda} = \lambda$. This proves that the eigenvalue λ is real.

To prove (b), suppose

$$A\mathbf{v} = \lambda \mathbf{v}, \qquad\qquad A\mathbf{w} = \mu \mathbf{w},$$

where $\lambda \neq \mu$ are distinct real eigenvalues. Then, again by (8.42),

$$\lambda \mathbf{v} \cdot \mathbf{w} = (A\mathbf{v}) \cdot \mathbf{w} = \mathbf{v} \cdot (A\mathbf{w}) = \mathbf{v} \cdot (\mu \mathbf{w}) = \mu \mathbf{v} \cdot \mathbf{w},$$

and hence

$$(\lambda - \mu) \mathbf{v} \cdot \mathbf{w} = 0.$$

Since $\lambda \neq \mu$, this implies that $\mathbf{v} \cdot \mathbf{w} = 0$ and hence the eigenvectors $\mathbf{v}, \mathbf{w}$ are orthogonal.

Finally, the proof of (c) is easy if all the eigenvalues of A are distinct. Theorem 8.16 implies that the eigenvectors form a basis of $\mathbb{R}^n$, and part (b) proves they are orthogonal. (An alternative proof starts with orthogonality, and then applies Proposition 5.4 to prove that the eigenvectors form a basis.) To obtain an orthonormal basis, we merely divide the eigenvectors by their lengths: $\mathbf{u}_k = \mathbf{v}_k / \| \mathbf{v}_k \|$, as in Lemma 5.2. A general proof can be found in [**125**]. *Q.E.D.*

Example 8.36. Consider the symmetric matrix $A = \begin{pmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \\ 2 & 2 & -1 \end{pmatrix}$. A straightforward computation produces its eigenvalues and eigenvectors:

$$\lambda_1 = 9, \qquad \lambda_2 = 3, \qquad \lambda_3 = -3,$$
$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}.$$

As the reader can check, the eigenvectors form an orthogonal basis of $\mathbb{R}^3$. The orthonormal eigenvector basis promised by Theorem 8.34 is obtained by dividing each eigenvector by

its norm:

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \end{pmatrix}.$$

Finally, we can characterize positive definite matrices by their eigenvalues.

Theorem 8.37. *A symmetric matrix $K = K^T$ is positive definite if and only if all of its eigenvalues are strictly positive.*

Proof: First, if $K > 0$, then, by definition, $\mathbf{x}^T K \mathbf{x} > 0$ for all nonzero vectors $\mathbf{x} \in \mathbb{R}^n$. In particular, if $\mathbf{x} = \mathbf{v}$ is an eigenvector with (necessarily real) eigenvalue λ, then

$$0 < \mathbf{v}^T K \mathbf{v} = \mathbf{v}^T (\lambda \mathbf{v}) = \lambda \, \| \mathbf{v} \|^2, \tag{8.43}$$

which immediately proves that $\lambda > 0$. Conversely, suppose K has all positive eigenvalues. Let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the orthonormal eigenvector basis of $\mathbb{R}^n$ guaranteed by Theorem 8.34, with $K \mathbf{u}_j = \lambda_j \mathbf{u}_j$. Then, writing

$$\mathbf{x} = c_1 \mathbf{u}_1 + \, \cdots \, + c_n \mathbf{u}_n, \qquad \text{we have} \qquad K \mathbf{x} = c_1 \lambda_1 \mathbf{u}_1 + \, \cdots \, + c_n \lambda_n \mathbf{u}_n.$$

Therefore, using the orthonormality of the eigenvectors,

$$\mathbf{x}^T K \mathbf{x} = (c_1 \mathbf{u}_1 + \, \cdots \, + c_n \mathbf{u}_n) \cdot (c_1 \lambda_1 \mathbf{u}_1 + \, \cdots \, + c_n \lambda_n \mathbf{u}_n) = \lambda_1 c_1^2 + \, \cdots \, + \lambda_n c_n^2 \geq 0,$$

Moreover, the result is strictly positive for $\mathbf{x} \neq \mathbf{0}$ since not all the coefficients $c_1, \dots, c_n$ can be zero. This proves that that K is positive definite. *Q.E.D.*

Remark: The same proof shows that K is positive semi-definite if and only if all its eigenvalues satisfy $\lambda \geq 0$. A positive semi-definite matrix that is not positive definite admits a zero eigenvalue and one or more *null eigenvectors*, i.e., solutions to $K \mathbf{v} = \mathbf{0}$. Every nonzero element $\mathbf{0} \neq \mathbf{v} \in \ker K$ of the kernel is a null eigenvector.

In elasticity, the stress tensor is represented by a positive definite matrix. Its eigenvalues are known as the principal stretches and its eigenvectors the principal directions of stretch for the body.

Example 8.38. Consider the symmetric matrix $K = \begin{pmatrix} 8 & 0 & 1 \\ 0 & 8 & 1 \\ 1 & 1 & 7 \end{pmatrix}$. Its characteristic equation is

$$\det(K - \lambda I) = -\lambda^3 + 23\lambda^2 - 174\lambda + 432 = -(\lambda - 9)(\lambda - 8)(\lambda - 6),$$

and so its eigenvalues are $9, 8$ and 6. Since they are all positive, we conclude that K is a positive definite matrix. The associated eigenvectors are

$$\lambda_1 = 9, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \lambda_2 = 8, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \qquad \lambda_3 = 6, \qquad \mathbf{v}_3 = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

Note that the eigenvectors form an orthogonal basis of $\mathbb{R}^3$, as guaranteed by Theorem 8.34. We can construct an orthonormal eigenvector basis

$$\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix}, \qquad \mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \qquad \mathbf{u}_3 = \begin{pmatrix} -\frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{pmatrix},$$

by dividing each eigenvector by its norm.

The Spectral Theorem

Since a real, symmetric matrix admits an eigenvector basis, it is diagonalizable. Moreover, since we can arrange that the eigenvectors form an orthonormal basis, the diagonalizing matrix takes a particularly simple form. Recall that an $n \times n$ matrix Q is called *orthogonal* if and only if its columns form an orthonormal basis of $\mathbb{R}^n$. Alternatively, one characterizes orthogonal matrices by the condition $Q^{-1} = Q^T$, as per Definition 5.18.

The *Spectral Theorem* governs the diagonalization of real, symmetric matrices.

Theorem 8.39. *Let A be a real, symmetric matrix. Then there exists an orthogonal matrix Q such that*

$$A = Q\,\Lambda\,Q^{-1} = Q\,\Lambda\,Q^T, \tag{8.44}$$

where Λ is a real diagonal matrix. The eigenvalues of A appear on the diagonal of Λ, while the eigenvectors are the corresponding columns of Q.

Proof: The proof is an immediate consequence of the diagonalization Theorem 8.20 coupled with Theorem 8.34. One merely replaces the general eigenvector matrix S by the orthogonal matrix Q whose columns consist of our orthonormal eigenvector basis. *Q.E.D.*

Remark: The term "spectrum" refers to the eigenvalues of a matrix or, more generally, a linear operator. The terminology comes from physics. The spectral energy lines of atoms, molecules and nuclei are characterized as the eigenvalues of the governing quantum mechanical linear operators!

Warning: The spectral diagonalization $A = Q\,\Lambda\,Q^T$ and the Gaussian diagonalization $A = L\,D\,L^T$ of a regular symmetric matrix, cf. (1.52), are completely different. In particular, the eigenvalues are *not* the pivots: $\Lambda \neq D$.

The spectral factorization (8.44) provides us with an alternative means of diagonalizing the associated quadratic form $q(\mathbf{x}) = \mathbf{x}^T A\,\mathbf{x}$, i.e., of completing the square. We write

$$q(\mathbf{x}) = \mathbf{x}^T A\,\mathbf{x} = \mathbf{x}^T\,Q\,\Lambda\,Q^T\,\mathbf{x} = \mathbf{y}^T\Lambda\,\mathbf{y} = \sum_{i=1}^{n} \lambda_i\,y_i^2, \tag{8.45}$$

where $\mathbf{y} = Q^T\mathbf{x} = Q^{-1}\mathbf{x}$ are the coordinates of $\mathbf{x}$ with respect to the orthonormal eigenvalue basis of A, cf. (7.21). In particular, $q(\mathbf{x}) > 0$ for all nonzero $\mathbf{x} = \mathbf{0}$ — which means A is positive definite — if and only if each eigenvalue $\lambda_i > 0$ is strictly positive. This provides an alternative proof of Theorem 8.37.

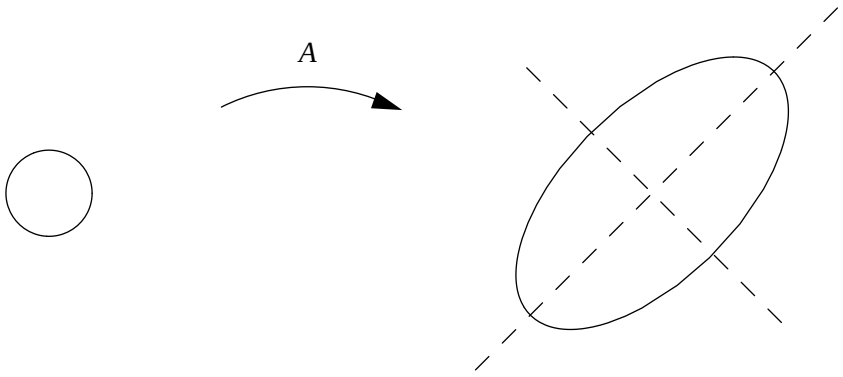


Figure 8.3. Stretching a Circle into an Ellipse.

Example 8.40. For the 2×2 matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ considered in Example 8.35, the orthonormal eigenvectors (8.41) produce the diagonalizing orthogonal rotation matrix $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$. The reader can check the spectral factorization

$$\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} = A = Q\,\Lambda\,Q^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

According to (8.45), the associated quadratic form is diagonalized as

$$q(\mathbf{x}) = 3\,x_1^2 + 2\,x_1\,x_2 + 3\,x_2^2 = 4\,y_1^2 + 2\,y_2^2,$$

where $\mathbf{y} = Q^T\mathbf{x}$, i.e., $y_1 = \dfrac{x_1 + x_2}{\sqrt{2}}$, $y_2 = \dfrac{-x_1 + x_2}{\sqrt{2}}$.

We note that you can choose Q to be a proper orthogonal matrix, so $\det Q = 1$, since an improper orthogonal matrix can be made proper by multiplying one of its columns by -1, which does not affect its status as an eigenvector matrix. Since a proper orthogonal matrix Q represents a rigid rotation of $\mathbb{R}^n$, the diagonalization of a symmetric matrix can be interpreted as a rotation of the coordinate system in which the orthogonal eigenvectors line up along the coordinate axes. Therefore, a linear transformation $L(\mathbf{x}) = A\,\mathbf{x}$ represented by a positive definite matrix $A > 0$ can be regarded as a combination of stretches along a mutually orthogonal set of directions. In elasticity, the stress tensor is represented by a positive definite matrix. Its eigenvalues are known as the *principal stretches* and its eigenvectors the principal directions of stretch.

. A good way to visualize this is to consider the effect of the linear transformation on the unit (Euclidean) sphere $S_1 = \{\, \|\,\mathbf{x}\,\| = 1 \,\}$. Stretching the sphere in orthogonal directions will map it into an ellipsoid E whose axes are aligned with the directions of stretch. For example, the matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$ considered in the preceding example

represents the linear transformation

$$\widetilde{x} = 3\,x + y, \qquad \widetilde{y} = x + 3\,y.$$

Solving for x, y, we find that the unit circle $x^2 + y^2 = 1$ will be mapped to the ellipse

$$\left(\frac{3\,\widetilde{x} - \widetilde{y}}{8}\right)^2 + \left(\frac{-\,\widetilde{x} + 3\,\widetilde{y}}{8}\right)^2 = \tfrac{5}{32}\,\widetilde{x}^2 - \tfrac{3}{16}\,\widetilde{x}\,\widetilde{y} + \tfrac{5}{32}\,\widetilde{y}^2 = 1,$$

whose principal axes lie in the directions of the eigenvectors $\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$, $\mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$. Their lengths, $4, 2$, are the ellipse's *semi-axes*, which are equal to the corresponding eigenvalues. The effect of this linear transformation is illustrated in Figure 8.3.

Optimization Principles for Eigenvalues

As we learned in Chapter 4, the solution to a linear system with positive definite coefficient matrix can be characterized by a minimization principle. Thus, it should come as no surprise that eigenvalues of positive definite, and even more general symmetric matrices, can also be characterized by some sort of optimization procedure. A number of basic numerical algorithms for computing eigenvalues, of both matrices and, later on, differential operators are based on such optimization principles.

First consider the case of a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. We assume that the diagonal entries, which are the *same* as the eigenvalues, appear in decreasing order,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \tag{8.46}$$

so λ_1 is the largest eigenvalue, while λ_n is the smallest. The effect of Λ on a vector $\mathbf{y} \in \mathbb{R}^n$ is to multiply its entries by the diagonal eigenvalues: $\Lambda\,\mathbf{y} = (\,\lambda_1\,y_1, \lambda_2\,y_2, \dots, \lambda_n\,y_n\,)^T$. In other words, the linear transformation represented by the coefficient matrix Λ has the effect of stretching[†] in the i^{th} coordinate direction by the factor λ_i. In particular, the maximal stretch occurs in the first direction, with factor λ_1, while the minimal stretch occurs in the last direction, with factor λ_n. The germ of the optimization principles for characterizing the extreme eigenvalues is contained in this geometrical observation.

Let us turn our attention to the associated quadratic form

$$q(\mathbf{y}) = \mathbf{y}^T \Lambda\,\mathbf{y} = \lambda_1\,y_1^2 + \lambda_2\,y_2^2 + \cdots + \lambda_n\,y_n^2. \tag{8.47}$$

In the positive definite case, when all the $\lambda_i \geq 0$, the minimal value of $q(\mathbf{y})$ is 0, obtained when $\mathbf{y} = \mathbf{0}$; if $\lambda_n < 0$, then the minimal value is $-\infty$, since $q(c\,\mathbf{e}_n) = \lambda_n\,c^2$. Thus, in either case, a strict minimization of $q(\mathbf{y})$ is not of much help.

Suppose, however, that we try to minimize $q(\mathbf{y})$ when $\mathbf{y}$ is restricted to be a unit vector (in the Euclidean norm): $\|\,\mathbf{y}\,\| = 1$. First note that $q(\mathbf{e}_i) = \lambda_i$, where $\mathbf{e}_i$ denotes the i^{th} standard basis vector. Moreover, in view of (8.46) and the positivity of each y_i^2,

$$q(\mathbf{y}) = \lambda_1\,y_1^2 + \lambda_2\,y_2^2 + \cdots + \lambda_n\,y_n^2 \leq \lambda_1\,y_1^2 + \lambda_1\,y_2^2 + \cdots + \lambda_1\,y_n^2 = \lambda_1\big(\,y_1^2 + \cdots + y_n^2\,\big) = \lambda_1,$$

[†] If $\lambda_i < 0$, then the effect is to stretch and reflect.

whenever $\| \mathbf{y} \|^2 = y_1^2 + \cdots + y_n^2 = 1$. Since $q(\mathbf{e}_1) = \lambda_1$, the maximal value of $q(\mathbf{y})$ over all unit vectors is the largest eigenvalue of Λ:

$$\lambda_1 = \max \{ \, q(\mathbf{y}) \mid \| \mathbf{y} \| = 1 \, \}.$$

By the same reasoning, $q(\mathbf{y})$ also has a minimal value

$$\lambda_n = \min \{ \, q(\mathbf{y}) \mid \| \mathbf{y} \| = 1 \, \}$$

equal to the smallest eigenvalue. Thus, we can represent the two extreme eigenvalues by optimization principles, albeit of a slightly different character than we treated in Chapter 4.

Now suppose A is any symmetric matrix. We use the Spectral Theorem 8.39 to write

$$A = Q \, \Lambda \, Q^T,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix containing the eigenvalues of A along its diagonal, written in increasing order, while Q is an orthogonal matrix whose columns are the orthonormal eigenvector basis. We use the spectral factorization to diagonalize the associated quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T A \, \mathbf{x} = \mathbf{x}^T \, Q \, \Lambda \, Q^T \, \mathbf{x} = \mathbf{y}^T \Lambda \, \mathbf{y}, \qquad \text{where} \qquad \mathbf{y} = Q^T \mathbf{x} = Q^{-1} \mathbf{x},$$

as in (8.45). According to the preceding discussion, the minimum of $\mathbf{y}^T \Lambda \, \mathbf{y}$ over all unit vectors $\| \mathbf{y} \| = 1$ is the smallest eigenvalue λ_1 of Λ, which is the *same* as the smallest eigenvalue of A. Moreover, since Q is an orthogonal matrix, Proposition 7.23 tell us that it maps unit vectors to unit vectors:

$$1 = \| \mathbf{y} \| = \| \, Q^T \mathbf{x} \, \| = \| \, \mathbf{x} \, \|.$$

Thus, minimizing $q(\mathbf{x})$ over all unit vectors $\mathbf{y} = Q^T \mathbf{x}$ is the same as minimizing over all unit vectors $\mathbf{x}$. Similar reasoning applies to the largest eigenvalue. In this fashion, we have established the basic optimization principles for the largest and smallest eigenvalues of a symmetric matrix.

Theorem 8.41. *If A is a symmetric matrix, then*

$$\lambda_1 = \max \left\{ \, \mathbf{x}^T A \, \mathbf{x} \, \middle| \, \| \mathbf{x} \| = 1 \, \right\}, \qquad \lambda_n = \min \left\{ \, \mathbf{x}^T A \, \mathbf{x} \, \middle| \, \| \mathbf{x} \| = 1 \, \right\}, \tag{8.48}$$

are, respectively its largest and smallest eigenvalues.

The maximal value is achieved when we set $\mathbf{x} = \pm \mathbf{u}_1$ equal to either unit eigenvector corresponding to the largest eigenvalue; similarly, the minimal value is at $\mathbf{x} = \pm \mathbf{u}_n$.

Remark: You may have learned about Lagrange multipliers for constrained minimization problems. In the present situation, the Lagrange multiplier is equal to the eigenvalue.

Example 8.42. The problem is to maximize the value of the quadratic form

$$q(x, y) = 3 \, x^2 + 2 \, x \, y + 3 \, y^2$$

for all x, y lying on the unit circle $x^2 + y^2 = 1$. This maximization problem is precisely of form (8.48). The coefficient matrix is $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$, whose eigenvalues are, according to Example 8.7, $\lambda_1 = 4$ and $\lambda_2 = 2$. Theorem 8.41 implies that the maximal value for the quadratic form on the unit circle is the largest eigenvalue, and hence equal to 4, while its minimal value is the smallest eigenvalue, and hence equal to 2. Indeed, if we evaluate $q(x, y)$ on the unit eigenvectors, we obtain $q\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = 2$, $q\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = 4$, while $2 \leq q(x, y) \leq 4$ for all x, y such that $x^2 + y^2 = 1$.

Restricting the quadratic form to unit vectors may not be particularly convenient. One can, however, easily rephrase the optimization principles to apply to more general vectors. If $\mathbf{v} \neq \mathbf{0}$ is any nonzero vector, then $\mathbf{x} = \mathbf{v}/\| \mathbf{v} \|$ is a unit vector. Substituting this expression for $\mathbf{x}$ in the quadratic form leads to the following optimization principles for the extreme eigenvalues of a symmetric matrix:

$$\lambda_1 = \max \left\{ \frac{\mathbf{v}^T A \mathbf{v}}{\| \mathbf{v} \|^2} \;\middle|\; \mathbf{v} \neq \mathbf{0} \right\}, \qquad \lambda_n = \min \left\{ \frac{\mathbf{v}^T A \mathbf{v}}{\| \mathbf{v} \|^2} \;\middle|\; \mathbf{v} \neq \mathbf{0} \right\}. \tag{8.49}$$

Thus, we replace optimization of a quadratic polynomial over the unit sphere by optimization of a rational function over all of $\mathbb{R}^n \setminus \{\mathbf{0}\}$. For example, the maximum value of

$$r(x, y) = \frac{3x^2 + 2xy + 3y^2}{x^2 + y^2} \qquad \text{for all} \qquad \begin{pmatrix} x \\ y \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

is equal to 4, the same maximal eigenvalue of the corresponding coefficient matrix.

What about if we are interested in the intermediate eigenvalues? Then we need to be a little more sophisticated in designing the minimization or maximization principle. To motivate the construction, look first at the diagonal case. If we restrict the quadratic form (8.47) to vectors $\widetilde{\mathbf{y}} = (\, 0, y_2, \ldots, y_n \,)^T$ whose first component is zero, we obtain

$$q(\widetilde{\mathbf{y}}) = q(0, y_2, \ldots, y_n) = \lambda_2 \, y_2^2 + \, \cdots \, + \lambda_n \, y_n^2.$$

The maximum value of $q(\widetilde{\mathbf{y}})$ over all such $\widetilde{\mathbf{y}}$ of norm 1 is, by the same reasoning, the second largest eigenvalue λ_2. Moreover, we can characterize such vectors geometrically by noting that they are orthogonal to the first standard basis vector, $\widetilde{\mathbf{y}} \cdot \mathbf{e}_1 = 0$, which also happens to be the eigenvector of Λ corresponding to the eigenvalue λ_1. Similarly, if we want to find the j^{th} largest eigenvalue λ_j, we maximize $q(\mathbf{y})$ over all unit vectors $\widehat{\mathbf{y}}$ whose first $j - 1$ components vanish, $y_1 = \cdots = y_{j-1} = 0$, or, stated geometrically, over all $\widetilde{\mathbf{y}}$ such that $\| \widehat{\mathbf{y}} \| = 1$ and $\widehat{\mathbf{y}} \cdot \mathbf{e}_1 = \cdots = \widehat{\mathbf{y}} \cdot \mathbf{e}_{j-1} = 0$, i.e., over all vectors orthogonal to the first $j - 1$ eigenvectors of Λ.

A similar reasoning based on the Spectral Theorem 8.39 and the orthogonality of eigenvectors of symmetric matrices, leads to the general result.

Theorem 8.43. *Let $A > 0$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and corresponding orthogonal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Then the maximal value of the quadratic form $\mathbf{x}^T A \mathbf{x}$ over all unit vectors which are orthogonal to the first $j - 1$*

eigenvectors is the j^{th} eigenvalue:

$$\lambda_j = \max \left\{ \mathbf{x}^T A \mathbf{x} \;\middle|\; \| \mathbf{x} \| = 1, \;\; \mathbf{x} \cdot \mathbf{v}_1 = \cdots = \mathbf{x} \cdot \mathbf{v}_{j-1} = 0 \right\}. \tag{8.50}$$

8.6. Singular Values.

We have already indicated the absolutely fundamental importance of the eigenvalues of a square matrix in a wide range of applications. Alas, rectangular matrices have no eigenvalues, and so, at first glance, they do not appear to possess any quantities of comparable significance. However, our analysis of least squares minimization problems as well as the equilibrium equations for strucutres and circuits relied on the symmetric, positive semi-definite *square* Gram matrix $K = A^T A$, which can be constructed from any matrix A. It turns out that the eigenvalues of K play a comparable role. Since they are not easily related to the eigenvalues of A — which, in the truly rectangular case, don't even exist — we shall endow them with a new name.

Definition 8.44. The *singular values* $\sigma_1, \dots, \sigma_n$ of an $m \times n$ matrix A are the positive square roots, $\sigma_i = \sqrt{\lambda_i} > 0$, of the nonzero eigenvalues of the Gram matrix $K = A^T A$. The corresponding eigenvectors of K are known as the *singular vectors* of A.

Since K is positive semi-definite, its eigenvalues are always non-negative, $\lambda_i \geq 0$, and hence the singular values of A are also all positive, $\sigma_i > 0$ — no matter whether A itself has positive, negative, or even complex eigenvalues, or is rectangular and has no eigenvalues at all. However, for symmetric matrices, there is a direct connection between the two:

Proposition 8.45. *If $A = A^T$ is a symmetric matrix, its singular values are the absolute values of its nonzero eigenvalues: $\sigma_i = | \lambda_i |$, and its singular vectors coincide with the corresponding eigenvectors.*

Proof: In this case $K = A^T A = A^2$. If $A\mathbf{v} = \lambda \mathbf{v}$, then $K\mathbf{v} = A^2 \mathbf{v} = \lambda^2 \mathbf{v}$. Thus, every eigenvector $\mathbf{v}$ of A is also an eigenvector of K with eigenvalue λ^2. To demonstrate that K has no other eigenvalues, we use the spectral factorization $A = Q \Lambda Q^T$, with Λ the diagonal eigenvalue matrix and Q the orthogonal matrix whose columns form the orthonormal eigenvector basis of A. Then, by orthogonality, $K = A^2 = (Q \Lambda Q^T)^2 = Q \Lambda^2 Q^T$ is the spectral factorization of K. Thus, the eigenvalues of K are the diagonal entries of Λ^2, i.e., the squares of the eigenvalues of A, while its eigenvectors are the same columns of Q. *Q.E.D.*

The standard convention is to number the singular values in *decreasing* order, so that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Thus, σ_1 will always denote the largest or *dominant* singular value. If $A^T A$ has repeated eigenvalues, then the singular values of A are repeated with the same multiplicities.

Example 8.46. Let $A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}$. The associated Gram matrix is $K = A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$, with eigenvalues $\lambda_1 = 40$ and $\lambda_2 = 10$. Thus, the singular values of A are $\sigma_1 = \sqrt{40} \approx 6.3246\ldots$ and $\sigma_2 = \sqrt{10} \approx 3.1623\ldots$. Note that these are not the same as the eigenvalues of A, namely $\lambda_1 = \frac{1}{2}(3 + \sqrt{89}) \approx 6.2170\ldots$, $\lambda_2 = \frac{1}{2}(3 + \sqrt{89}) \approx -3.2170$.

A rectangular matrix M will be called *diagonal* if its only nonzero entries are on the main diagonal starting in the upper left hand corner, and so $m_{ij} = 0$ for $i \neq j$. An example is the matrix

$$M = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

whose only nonzero entries are in the diagonal $(1,1)$ and $(2,2)$ positions. (The last diagonal entry happens to be zero.)

The generalization of the spectral factorization to non-symmetric matrices is known as the *singular value decomposition*, commonly abbreviated as SVD. The decomposition applies to arbitrary real rectangular matrices.

Theorem 8.47. *Any real $m \times n$ matrix A can be factorized*

$$A = P \Sigma Q^T \tag{8.51}$$

into the product of an $m \times m$ orthogonal matrix P, the $m \times n$ diagonal matrix Σ that has the singular values of A as its nonzero diagonal entries, followed by an $n \times n$ orthogonal matrix Q^T.

Proof: Writing the factorization (8.51) as $AQ = P\Sigma$, and looking at the columns of the resulting matrix equation, we find

$$A \mathbf{u}_i = \sigma_i \mathbf{v}_i, \qquad i = 1, \dots, n, \tag{8.52}$$

where $\mathbf{u}_1, \dots, \mathbf{u}_n$ are the columns of $Q = (\, \mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_n \,)$, $\mathbf{v}_1, \dots, \mathbf{v}_m$ the columns of $P = (\, \mathbf{v}_1 \ \mathbf{v}_2 \ \dots \ \mathbf{v}_m \,)$. The scalars σ_i are the diagonal entries of Σ or, if $m < i \leq n$, equal to zero. The fact that P and Q are both orthogonal matrices means that their column vectors form orthonormal bases for, respectively, $\mathbb{R}^m$ and $\mathbb{R}^n$ under the Euclidean dot product. In this manner, the singular values of the linear transformation represented by the matrix A indicate how far it stretches a distinguished set of orthonormal basis vectors.

To construct the required bases, we let $\mathbf{u}_1, \dots, \mathbf{u}_n$ be the orthonormal eigenvector basis of the Gram matrix $K = A^T A$, and so

$$A^T A \mathbf{u}_i = K \mathbf{u}_i = \lambda_i \mathbf{u}_i = \sigma_i^2 \mathbf{u}_i.$$

We claim that the image vectors $\mathbf{w}_i = A \mathbf{u}_i$ are automatically orthogonal. Indeed,

$$\mathbf{w}_i \cdot \mathbf{w}_j = \mathbf{w}_i^T \mathbf{w}_j = (A \mathbf{u}_i)^T A \mathbf{u}_j = \mathbf{u}_i^T A^T A \mathbf{u}_j = \lambda_j \mathbf{u}_i^T \mathbf{u}_j = \lambda_j \mathbf{u}_i \cdot \mathbf{u}_j. \tag{8.53}$$

Thus, by orthogonality of the eigenvector basis, $\mathbf{w}_i \cdot \mathbf{w}_j = 0$ for $i \neq j$, while setting $i = j$ in (8.53) gives

$$\| \mathbf{w}_i \|^2 = \lambda_i \mathbf{u}_i^T \mathbf{u}_i = \lambda_i = \sigma_i^2, \qquad \text{and so} \qquad \| \mathbf{w}_i \| = \sigma_i.$$

Since $\mathbf{u}_1, \dots, \mathbf{u}_n$ form a basis of $\mathbb{R}^n$, their images $\mathbf{w}_1 = A \mathbf{u}_1, \dots, \mathbf{w}_n = A \mathbf{u}_n$ span $\operatorname{rng} A$. Suppose that A has r non-zero singular values, so $\sigma_i > 0$ for $i \leq r$. Then the corresponding image vectors $\mathbf{w}_1, \dots, \mathbf{w}_r$ are non-zero, mutually orthogonal vectors, and

hence form an orthogonal basis for $\operatorname{rng} A$. Since the dimension of $\operatorname{rng} A$ is equal to its rank, this implies that the number of singular values is $r = \operatorname{rank} A$. The corresponding unit vectors

$$\mathbf{v}_i = \frac{\mathbf{w}_i}{\sigma_i} = \frac{A\,\mathbf{u}_i}{\sigma_i}\,, \qquad\qquad i = 1, \dots, r, \tag{8.54}$$

are an orthonormal basis for $\operatorname{rng} A$. Let us choose an orthonormal basis $\mathbf{v}_{r+1}, \dots, \mathbf{v}_m$ for its orthogonal complement $\operatorname{coker} A = (\operatorname{rng} A)^{\perp}$. The combined set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_m$ clearly forms an orthonormal basis of $\mathbb{R}^m$. The orthonormal bases $\mathbf{u}_1, \dots, \mathbf{u}_n$ and $\mathbf{v}_1, \dots, \mathbf{v}_m$ have been designed to satisfy (8.52), and hence form the respective columns of the orthogonal matrices in the singular value decomposition (8.51). *Q.E.D.*

Example 8.48. For the matrix $A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix}$ considered in Example 8.46, the orthonormal eigenvector basis of $K = A^T A = \begin{pmatrix} 25 & 15 \\ 15 & 25 \end{pmatrix}$ is given by $\mathbf{u}_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ and $\mathbf{u}_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$. Thus, $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$. On the other hand, according to (8.54),

$$\mathbf{v}_1 = \frac{A\,\mathbf{u}_1}{\sigma_1} = \frac{1}{\sqrt{40}} \begin{pmatrix} 4\sqrt{2} \\ 2\sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{pmatrix}, \qquad \mathbf{v}_2 = \frac{A\,\mathbf{u}_2}{\sigma_2} = \frac{1}{\sqrt{10}} \begin{pmatrix} \sqrt{2} \\ -2\sqrt{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ -\frac{2}{\sqrt{5}} \end{pmatrix}.$$

Therefore, $P = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix}$. You may wish to validate the resulting singular value decomposition

$$A = \begin{pmatrix} 3 & 5 \\ 4 & 0 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \sqrt{40} & 0 \\ 0 & \sqrt{10} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = P \Sigma Q^T.$$

As their name suggests, the singular values can be used to detect singular matrices. Indeed, the singular value decomposition tells us some important new geometrical information about the action of the matrix, filling in further details in the discussion begun in Section 2.5 and continued in Section 5.6. The result follows directly from the proof of Theorem 8.47.

Theorem 8.49. *Let $\sigma_1, \dots, \sigma_r$ be the singular values of the $m \times n$ matrix A. Let $\mathbf{v}_1, \dots, \mathbf{v}_m$ and $\mathbf{u}_1, \dots, \mathbf{u}_n$ the orthonormal bases of, respectively, $\mathbb{R}^m$ and $\mathbb{R}^n$ provided by the columns of P and Q in its singular value decomposition $A = P \Sigma Q^T$. Then*

(*i*) $r = \operatorname{rank} A$,
(*ii*) $\mathbf{u}_1, \dots, \mathbf{u}_r$ *form an orthonormal basis for* $\operatorname{corng} A$,
(*iii*) $\mathbf{u}_{r+1}, \dots, \mathbf{u}_n$ *form an orthonormal basis for* $\ker A$,
(*iv*) $\mathbf{v}_1, \dots, \mathbf{v}_r$ *form an orthonormal basis for* $\operatorname{rng} A$,
(*v*) $\mathbf{v}_{r+1}, \dots, \mathbf{v}_m$ *form an orthonormal basis for* $\operatorname{coker} A$.

We already noted in Section 5.6 that the linear transformation $L: \mathbb{R}^n \to \mathbb{R}^m$ defined by matrix multiplication by A can be interpreted as a projection from $\mathbb{R}^n$ to corng A followed by an invertible map from corng A to rng A. The singular value decompostion tells us that not only is the latter map invertible, it is simply a combination of r stretches in mutually orthogonal directions, whose magnitudes equal the nonzero singular values. In this way, we have at last reached a complete understanding of the subtle geometry underlying the simple operation of matrix multiplication.

An alternative useful interpretation is to view the two orthogonal matrices in (8.51) as defining rigid rotations or reflections. Therefore, in all cases, a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ is composed of three constituents:

(*i*) A rotation/reflection of the domain space $\mathbb{R}^n$, followed by

(*ii*) a simple stretching map of the coordinate vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ of domain space, mapping $\mathbf{e}_i$ to $\sigma_i \mathbf{e}_i$ in the target space $\mathbb{R}^m$, followed by

(*iii*) a rotation/reflection of the target space.

In most cases,we can choose both orthogonal matrices to represent proper rotations; see Exercise ■.

The singular values not only provide a nice geometric interpretation of the action of the matrix, they also play a key role in modern computational algorithms. The relative magnitudes of the singular values can be used to distinguish well-behaved linear systems from ill-conditioned systems which are much trickier to solve accurately. A sqaure matrix with a zero singular value is singular; a matrix with one or more very small singular values is considered to be close to singular, and hence ill-conditioned in the sense that it is hard to invert numerically. Such ill-conditioning is traditionally quantified as follows.

Definition 8.50. The *condition number* of an $n \times n$ matrix is the ratio between the largest and smallest singular value: $\kappa(A) = \sigma_1/\sigma_n$.

A matrix with a very large condition number is said to be *ill-conditioned*; in practice, this occurs when the condition number is larger than the reciprocal of the machine's precision, e.g., 10^6 for single precision aritmetic. As the name implies, it is much harder to solve a linear system $A\mathbf{x} = \mathbf{b}$ when its coefficient matrix is ill-conditioned. In the extreme case when A has one or more zero singular values, its condition number is infinite, and the linear system is singular, with either no solution or infinitely many solutions.

With the singular value decomposition in hand, we are able to introduce a generalization of the inverse of a matrix that applies even in cases when the matrix in question is singular or rectangular. We begin with the diagonal case. Let Σ be an $m \times n$ diagonal matrix with r nonzero diagonal entries $\sigma_1, \dots, \sigma_r$. We define the *pseudoinverse* of Σ to be the $n \times m$ diagonal matrix Σ^+ whose nonzero diagonal entries are the reciprocals $1/\sigma_1, \dots, 1/\sigma_r$. For example, if

$$\Sigma = \begin{pmatrix} 5 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \text{then} \qquad \Sigma^+ = \begin{pmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In particular, if Σ is a nonsingular diagonal matrix (and necessarily square) then its pseudoinverse $\Sigma^+ = \Sigma^{-1}$ is the same as its ordinary inverse.

Definition 8.51. The *pseudoinverse* of an $m \times n$ matrix A with singular value decomposition $A = P\,\Sigma\,Q^T$ is the $n \times m$ matrix $A^+ = Q\,\Sigma^+\,P^T$.

If A is a non-singular square matrix, then its pseudoinverse agrees with its ordinary inverse, since

$$A^{-1} = (P\,\Sigma\,Q^T)^{-1} = Q^{-1}\,\Sigma^{-1}\,P^{-T} = Q\,\Sigma^+\,P^T = A^+,$$

where we used the fact that the inverse of an orthogonal matrix is equal to its transpose. The pseudoinverse provides us with a direct route to the least squares solution to a system of linear equations.

Theorem 8.52. *Consider the linear system $A\,\mathbf{x} = \mathbf{b}$. Let $\mathbf{y} = A^+\,\mathbf{b}$, where A^+ is the pseudoinverse of A. If $\ker A = \{\mathbf{0}\}$, then $\mathbf{y}$ is the least squares solution to the system. If, more generally, $\ker A \neq \{\mathbf{0}\}$, then $\mathbf{y}$ is the vector of minimal Euclidean norm among all vectors that minimize the least squares error $\| A\,\mathbf{x} - \mathbf{b} \|$.*

Proof: To show that $\mathbf{y} = A^+\,\mathbf{b}$ is the least squares solution to the system, we must check that it satisfies the normal equations $A^T A\,\mathbf{y} = A^T\mathbf{b}$. Using the definition of the pseudoinverse and the singular value decomposition (8.51), we find

$$A^T A\,\mathbf{y} = (P\,\Sigma\,Q^T)^T(P\,\Sigma\,Q^T)(Q\,\Sigma^+\,P^T)\mathbf{b} = Q\Sigma^T\Sigma\,\Sigma^+P^T\mathbf{b} = Q\Sigma^T P^T\mathbf{b} = A^T\mathbf{b},$$

where the next to last equality is left as Exercise ∎ for the reader. This proves that $\mathbf{y}$ solves the normal equations, and hence minimizes the least squares error[†]. Moreover, since the last $n - r$ rows of Σ^+ are zero, the last $n - r$ entries of the vector $\mathbf{c} = \Sigma^+P^T\mathbf{b} = (\,c_1, \dots c_r, 0, \dots, 0\,)^T$. Therefore,

$$\mathbf{y} = A^+\,\mathbf{b} = Q\,\Sigma^+\,P^T\mathbf{b} = Q\,\mathbf{c} = c_1\,\mathbf{u}_1 + \,\cdots\, + c_r\,\mathbf{u}_r$$

is a linear combination of the r singular vectors, and hence $\mathbf{y} \in \operatorname{corng} A$ is the least squares solution that is orthogonal to the kernel of A. The general least squares solution has the form $\mathbf{x} = \mathbf{y} + \mathbf{z}$ where $\mathbf{z} \in \ker A$, and the fact that $\| \mathbf{y} \|$ is minimized follows as in Theorem 5.55. *Q.E.D.*

We immediately see that very small singular values lead to very large entries in Σ^+, which will cause numerical problems when computing the solution $\mathbf{x} = A^{-1}\mathbf{b}$ to the linear system. A common and effective computational strategy to avoid the effects of small singular values is to replace the corresponding diagonal entries of Σ^{-1} by 0, and thereby convert the system $A\,\mathbf{x} = \mathbf{b}$ into a least squares linear system.

[†] In Chapter 4, this was proved under the assumption that $\ker A = \{\mathbf{0}\}$. The general case is left as Exercise ∎ for the reader.

8.7. Linear Dynamical Systems.

Now we have accumulated enough experience in the theory and computation of eigenvalues to be able to analyze dynamical systems governed by linear, homogeneous, constant coefficient ordinary differential equations. Our primary focus will be on systems

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u} \tag{8.55}$$

consisting of n first order linear ordinary differential equations in n unknowns $\mathbf{u}(t) = (\,u_1(t), \dots, u_n(t)\,)^T \in \mathbb{R}^n$. The coefficient matrix A, of size $n \times n$, is assumed to be a constant real matrix — although extensions to complex systems are not difficult.

As we saw at the beginning of Section 8.2, the vector-valued exponential function $\mathbf{u}(t) = e^{\lambda t}\,\mathbf{v}$ is a (non-zero) solution if and only if λ is an eigenvalue of A and $\mathbf{v}$ is the corresponding eigenvector. If the coefficient matrix A is complete, then it admits n linearly independent eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$, which, along with their associated eigenvalues $\lambda_1, \dots, \lambda_n$ will produce n distinct exponential solutions

$$\mathbf{u}_1(t) = e^{\lambda_1 t}\,\mathbf{v}_1, \qquad \dots \qquad \mathbf{u}_n(t) = e^{\lambda_n t}\,\mathbf{v}_n. \tag{8.56}$$

Linear superposition, based on the general principle in Theorem 7.29 (although this is easy to prove directly), implies that, for any choice of constants $c_1, \dots, c_n$, the linear combination

$$\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \,\cdots\, + c_n\mathbf{u}_n(t) = c_1\,e^{\lambda_1 t}\,\mathbf{v}_1 + \,\cdots\, + c_n\,e^{\lambda_n t}\,\mathbf{v}_n \tag{8.57}$$

is also a solution to the linear system.

Are there any other solutions? The answer is no — in fact (8.57) represents the most general solution to the system. This result is a consequence of the basic existence and uniqueness theorem for linear systems of ordinary differential equations, which we discuss next.

Example 8.53. Consider the linear system

$$\frac{du}{dt} = 3\,u + v, \qquad \frac{dv}{dt} = u + 3\,v. \tag{8.58}$$

We first write the system in matrix form

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u}, \quad \text{with unknown } \mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} \text{ and coefficient matrix } A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

According to Example 8.7, the eigenvalues and eigenvectors of A are

$$\lambda_1 = 4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = 2, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We use the eigenvalues and eigenvectors to construct the two particular exponential solutions

$$\mathbf{u}_1(t) = e^{4t}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} e^{4t} \\ e^{4t} \end{pmatrix}, \qquad \mathbf{u}_2(t) = e^{2t}\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -e^{2t} \\ e^{2t} \end{pmatrix}.$$

According to the preceding remark, to be justified below, the general solution to (8.58) is then given as a linear combination

$$\mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} = c_1\, e^{4\,t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\, e^{2\,t} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1\, e^{4\,t} - c_2\, e^{2\,t} \\ c_1\, e^{4t} + c_2\, e^{2t} \end{pmatrix},$$

where c_1, c_2 are arbitrary constants.

Existence and Uniqueness

Before proceeding further, it will help to briefly summarize the existence and uniqueness properties for solutions to linear systems of ordinary differential equations. These are direct consequences of the general existence and uniqueness theorem for nonlinear systems of ordinary differential equations, to be discussed in Section 19.1. Even though we will only study the constant coefficient case in detail in this text, the results are equally applicable to homogeneous linear systems with variable coefficients. So, in this subsection we allow the coefficient matrix to depend continuously on t.

The key fact is that a system of n first order ordinary differential equations requires n initial conditions — one for each variable — in order to specify its solution uniquely. More specifically:

Theorem 8.54. *Let $A(t)$ be an $n \times n$ matrix of continuous functions on the interval $a < t < b$. Given an initial time $a < t_0 < b$ and an initial vector $\mathbf{b} \in \mathbb{R}^n$, the* initial value problem

$$\frac{d\mathbf{u}}{dt} = A(t)\,\mathbf{u}, \qquad\qquad \mathbf{u}(t_0) = \mathbf{b}, \tag{8.59}$$

admits a unique solution $\mathbf{u}(t)$ which is defined for all $a < t < b$.

In particular, an initial value problem for a constant coefficient system (8.55) admits a unique solution $\mathbf{u}(t)$ that is defined for all $-\infty < t < \infty$. Uniqueness of solutions implies that, for such homogeneous systems, the solution with zero initial conditions $\mathbf{u}(t_0) = \mathbf{0}$ is the trivial zero solution $\mathbf{u}(t) \equiv \mathbf{0}$. Uniqueness has the important consequence that linear independence needs only be checked at a single point.

Lemma 8.55. *The solutions $\mathbf{u}_1(t), \dots, \mathbf{u}_k(t)$ to a first order homogeneous linear system $\dot{\mathbf{u}} = A(t)\,\mathbf{u}$ are linearly independent functions if and only if their initial values $\mathbf{u}_1(t_0), \dots, \mathbf{u}_k(t_0)$ are linearly independent vectors in $\mathbb{R}^n$.*

Proof: If the solutions are linearly dependent, then

$$\mathbf{u}(t) = c_1\,\mathbf{u}_1(t) + \,\cdots\, + c_k\,\mathbf{u}_k(t) \equiv \mathbf{0} \tag{8.60}$$

for some constant scalars $c_1, \dots, c_k$ not all zero. The equation holds, in particular, at $t = t_0$,

$$\mathbf{u}(t_0) = c_1\,\mathbf{u}_1(t_0) + \,\cdots\, + c_k\,\mathbf{u}_k(t_0) = \mathbf{0}, \tag{8.61}$$

proving linear dependence of the initial vectors. Conversely, if the initial values are linearly dependent, then (8.61) hold for some $c_1, \dots, c_k$ not all zero. Linear superposition implies that the corresponding linear combination $\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \,\cdots\, + c_k\mathbf{u}_k(t)$ is a solution to the system, with zero initial condition. By uniqueness, $\mathbf{u}(t) \equiv \mathbf{0}$ for all t, and so (8.60) holds, proving linear dependence of the solutions. *Q.E.D.*

Warning: This result is *not* true if the functions are not solutions to a *first order* system! For example, $\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$, $\mathbf{u}_2(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$, are linearly independent vector-valued functions, but, at time $t = 0$, $\mathbf{u}_1(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \mathbf{u}_2(0)$ are linearly dependent vectors. Even worse, $\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ t \end{pmatrix}$, $\mathbf{u}_2(t) = \begin{pmatrix} t \\ t^2 \end{pmatrix}$, define linearly dependent vectors at every fixed value of t, but as vector-valued functions they are, nonetheless, linearly independent. In view of Lemma 8.55, neither pair of functions can be solutions to a common linear ordinary differential equation.

The next result tells us how many different solutions we need in order to construct the general solution by linear superposition.

Theorem 8.56. *Let $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$ be n linearly independent solutions to the homogeneous system of n first order linear ordinary differential equations $\dot{\mathbf{u}} = A(t)\,\mathbf{u}$. then the general solution is a linear combination $\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \cdots + c_n\mathbf{u}_n(t)$ depending on n arbitrary constants $c_1, \dots, c_n$.*

Proof: If we have n linearly independent solutions, then Lemma 8.55 implies that, at the initial time t_0, the vectors $\mathbf{u}_1(t_0), \dots, \mathbf{u}_n(t_0)$ are linearly independent, and hence form a basis for $\mathbb{R}^n$. This means that we can express any initial condition

$$\mathbf{u}(t_0) = \mathbf{b} = c_1\mathbf{u}_1(t_0) + \cdots + c_n\mathbf{u}_n(t_0)$$

as a linear combination of the initial vectors. Superposition and uniqueness of solutions implies that the corresponding solution to the initial value problem (8.59) is given by the same linear combination

$$\mathbf{u}(t) = \mathbf{b} = c_1\mathbf{u}_1(t) + \cdots + c_n\mathbf{u}_n(t).$$

We conclude that every solution to the ordinary differential equation can be written in the prescribed form. *Q.E.D.*

Complete Systems

Now we have assembled the basic ingredients that will enable us to construct the complete solution to most first order homogeneous, linear, constant coefficient systems of ordinary differential equations. For a system of n equations, the goal is to find n linearly independent solutions. Each eigenvalue and eigenvector leads to an exponential solution of the form $e^{\lambda t}\,\mathbf{v}$. The solutions will be linearly independent if and only if the eigenvectors are — this will follow easily from Lemma 8.55. Thus, if the $n \times n$ matrix admits an eigenvector basis, i.e., it is complete, then we have the requisite number of solutions, and hence have solved the differential equation.

Theorem 8.57. *If the $n \times n$ matrix A is complete, then the general (complex) solution to the constant coefficient linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$ is given by*

$$\mathbf{u}(t) = c_1\,e^{\lambda_1 t}\,\mathbf{v}_1 + \cdots + c_n\,e^{\lambda_n t}\,\mathbf{v}_n, \tag{8.62}$$

where $\mathbf{v}_1, \dots, \mathbf{v}_n$ *are the linearly independent eigenvectors,* $\lambda_1, \dots, \lambda_n$ *the corresponding eigenvalues, and* $c_1, \dots, c_n$ *arbitrary constants, which are uniquely specified by the initial conditions* $\mathbf{u}(t_0) = \mathbf{b}$.

Proof: Since the eigenvectors are linearly independent, the solutions define linearly independent vectors $\mathbf{u}_k(0) = \mathbf{v}_k$ time $t = 0$. Thus, Lemma 8.55 implies that the functions $\mathbf{u}_k(t)$ are, indeed, linearly independent. Thus, the result is an immediate consequence of Theorem 8.56. *Q.E.D.*

Example 8.58. Let us solve the initial value problem

$$\begin{aligned} \dot{u}_1 &= -2\,u_1 + u_2, & u_1(0) &= 3, \\ \dot{u}_2 &= 2\,u_1 - 3\,u_2 & u_2(0) &= 0. \end{aligned}$$

The coefficient matrix of the system is $A = \begin{pmatrix} -2 & 1 \\ 2 & -3 \end{pmatrix}$. A straightforward computation produces the following eigenvalues and eigenvectors of A:

$$\lambda_1 = -4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \qquad\qquad \lambda_2 = -1, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The corresponding exponential soltions $\mathbf{u}_1(t) = e^{-4\,t} \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \mathbf{u}_2(t) = e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ form a basis for the two-dimensional solution space. The general solution is an arbitrary linear combination

$$\mathbf{u}(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} = c_1\, e^{-4\,t} \begin{pmatrix} 1 \\ -2 \end{pmatrix} + c_2\, e^{-t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1\, e^{-4\,t} + c_2\, e^{-t} \\ -2\,c_1\, e^{-4\,t} + c_2\, e^{-t} \end{pmatrix},$$

where c_1, c_2 are constant scalars. Once we have the general solution in hand, the final step is to determine the values of c_1, c_2 so as to satisfy the initial conditions. Evaluating the solution at $t = 0$, we find we need to solve the linear system

$$c_1 + c_2 = 3, \qquad\qquad -2\,c_1 + c_2 = 0,$$

for $c_1 = 1$, $c_2 = 2$. Thus, the (unique) solution to the initial value problem is

$$u_1(t) = e^{-4\,t} + 2\,e^{-t}, \qquad\qquad u_2(t) = -2\,e^{-4\,t} + 2\,e^{-t}.$$

Note that both components of the solution decay exponentially fast to 0 as $t \to \infty$.

Example 8.59. Consider the linear system

$$\dot{u}_1 = u_1 + 2\,u_2, \qquad \dot{u}_2 = u_2 - 2\,u_3, \qquad \dot{u}_3 = 2\,u_1 + 2\,u_2 - u_3.$$

The coefficient matrix is

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

In Example 8.10 we computed the eigenvalues and eigenvectors:

$$\lambda_1 = -1, \qquad \lambda_2 = 1 + 2\,\mathrm{i}\,, \qquad \lambda_3 = 1 - 2\,\mathrm{i}\,,$$
$$\mathbf{v}_1 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -\mathrm{i} \\ 1 \end{pmatrix}.$$

The first leads to a real solution, but the second and third lead to complex solutions to our real system of equations, e.g., $\widehat{\mathbf{u}}_2(t) = e^{(1+2\,\mathrm{i}\,)\,t}\,(\,1,\,\mathrm{i}\,,1\,)^T$. While this is a perfectly valid complex solution, it is not so convenient to work with if, as in most applications, we require real-valued functions. Since the underlying linear system is real, the general reality principle of Theorem 7.47 says that any complex solution can be broken up into its real and imaginary parts, each of which is a *real* solution. Applying Euler's formula (3.76) to the complex exponential, we find

$$\widehat{\mathbf{u}}_2(t) = e^{(1+2i)t}\begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix} = \big(\,e^t\cos 2t + \,\mathrm{i}\,e^t \sin 2t\,\big)\begin{pmatrix} 1 \\ \mathrm{i} \\ 1 \end{pmatrix} = \begin{pmatrix} e^t\cos 2t \\ -e^t\sin 2t \\ e^t\cos 2t \end{pmatrix} + \mathrm{i}\,\begin{pmatrix} e^t\sin 2t \\ e^t\cos 2t \\ e^t\sin 2t \end{pmatrix},$$

which yields two real vector-valued solutions to the system, as you can readily check. In this manner, we have produced three linearly independent real solutions to our system:

$$\mathbf{u}_1(t) = \begin{pmatrix} -e^{-t} \\ e^{-t} \\ e^{-t} \end{pmatrix}, \qquad \mathbf{u}_2(t) = \begin{pmatrix} e^t\cos 2t \\ -e^t\sin 2t \\ e^t\cos 2t \end{pmatrix}, \qquad \mathbf{u}_3(t) = \begin{pmatrix} e^t\sin 2t \\ e^t\cos 2t \\ e^t\sin 2t \end{pmatrix}.$$

Theorem 8.56 tells us that the general solution is a linear combination of the 3 independent solutions:

$$\mathbf{u}(t) = c_1\,\mathbf{u}_1(t) + c_2\,\mathbf{u}_2(t) + c_3\,\mathbf{u}_3(t) = \begin{pmatrix} -c_1\,e^{-t} + c_2\,e^t\cos 2t + c_3\,e^t\sin 2t \\ c_1\,e^{-t} - c_2\,e^t\sin 2t + c_3\,e^t\cos 2t \\ c_1\,e^{-t} + c_2\,e^t\cos 2t + c_3\,e^t\sin 2t \end{pmatrix}.$$

The constants c_1, c_2, c_3 are uniquely prescribed by imposing initial conditions; for example, the solution with $\mathbf{u}(0) = (\,2,-1,-2\,)^T$ requires $c_1 = -2, c_2 = 0, c_3 = 1$, and so the solution's components are $u_1(t) = 2\,e^{-t} + e^t\sin 2t$, $u_2(t) = -2\,e^{-t} + e^t\cos 2t$, $u_3(t) = -2\,e^{-t} + e^t\sin 2t$.

Incidentally, the third complex solution also produces two real solutions, but these reproduce the ones we have already listed. In fact, since $\lambda_3 = \overline{\lambda_2}$ is the complex conjugate of the eigenvalue λ_2, its eigenvector $\mathbf{v}_3 = \overline{\mathbf{v}_2}$ is also the complex conjugate of the eigenvector $\mathbf{v}_2$, and, finally, the solutions are also related by complex conjugation:

$$\widehat{\mathbf{u}}_3(t) = e^{(1-2i)t}\begin{pmatrix} 1 \\ -\mathrm{i} \\ 1 \end{pmatrix} = \begin{pmatrix} e^t\cos 2t \\ -e^t\sin 2t \\ e^t\cos 2t \end{pmatrix} - \mathrm{i}\,\begin{pmatrix} e^t\sin 2t \\ e^t\cos 2t \\ e^t\sin 2t \end{pmatrix} = \overline{\widehat{\mathbf{u}}_2(t)}\,.$$

In general, when dealing with complex eigenvalues of real systems, you only need to look at one eigenvalue from each complex conjugate pair to find a complete system of real solutions.

The General Case

If the matrix A is not complete, then the formulae for the solutions are a little more intricate, and involve polynomials as well as (complex) exponentials. When dealing with an incomplete matrix, we do not have sufficient eigenvectors to construct all the solutions, and so make use of its Jordan basis. Let us first describe the solutions associated with a Jordan chain.

Lemma 8.60. *Suppose $\mathbf{w}_1, \dots, \mathbf{w}_k$ form a Jordan chain of length k for the eigenvalue λ of the matrix A. Then there are k linearly independent solutions to the corresponding first order system $\dot{\mathbf{u}} = A\,\mathbf{u}$ having the form*

$$\mathbf{u}_1(t) = e^{\lambda t}\mathbf{w}_1, \qquad \mathbf{u}_2(t) = e^{\lambda t}(t\,\mathbf{w}_1 + \mathbf{w}_2), \qquad \mathbf{u}_3(t) = e^{\lambda t}\big(\tfrac{1}{2}t^2\,\mathbf{w}_1 + t\,\mathbf{w}_2 + \mathbf{w}_3\big),$$

$$\text{and, in general,} \qquad \mathbf{u}_j(t) = e^{\lambda t}\sum_{i=1}^{j}\frac{t^{j-i}}{(j-i)!}\,\mathbf{w}_i, \qquad 1 \le j \le k. \tag{8.63}$$

The proof is by direct substitution of the formulae into the differential equation, using the defining relations (8.34) of the Jordan chain; details are left to the reader. If λ is a complex eigenvalue, then the Jordan chain solutions (8.63) will involve complex exponentials. As usual, they can be split into their real an imaginary parts which, provided A is a real matrix, are independent real solutions.

Example 8.61. The coefficient matrix of the system

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ -2 & 2 & -4 & 1 & 1 \\ -1 & 0 & -3 & 0 & 0 \\ -4 & -1 & 3 & 1 & 0 \\ 4 & 0 & 2 & -1 & 0 \end{pmatrix}\mathbf{u}$$

is incomplete; it has only 2 linearly independent eigenvectors associated with the eigenvalues 1 and -2. Using the Jordan basis computed in Example 8.27, we produce the following 5 linearly independent solutions:

$$\mathbf{u}_1(t) = e^{t}\,\mathbf{v}_1, \qquad \mathbf{u}_2(t) = e^{t}\,(t\,\mathbf{v}_1 + \mathbf{v}_2), \qquad \mathbf{u}_3(t) = e^{t}\,(\tfrac{1}{2}t^2\,\mathbf{v}_1 + t\,\mathbf{v}_2 + \mathbf{v}_3),$$
$$\mathbf{u}_4(t) = e^{-2\,t}\,\mathbf{v}_4, \qquad \mathbf{u}_5(t) = e^{-2\,t}\,(t\,\mathbf{v}_4 + \mathbf{v}_5),$$

or, explicitly,

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ -e^{t} \\ e^{t} \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -e^{t} \\ 0 \\ -t\,e^{t} \\ (1+t)\,e^{t} \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -t\,e^{t} \\ 0 \\ \big(1-\frac{1}{2}t^2\big)\,e^{t} \\ \big(t+\frac{1}{2}t^2\big)\,e^{t} \end{pmatrix}, \quad \begin{pmatrix} -e^{-2\,t} \\ e^{-2\,t} \\ e^{-2\,t} \\ -2\,e^{-2\,t} \\ 0 \end{pmatrix}, \quad \begin{pmatrix} -(1+t)\,e^{-2\,t} \\ t\,e^{-2\,t} \\ t\,e^{-2\,t} \\ -2\,(1+t)\,e^{-2\,t} \\ e^{-2\,t} \end{pmatrix}.$$

The first three are associated with the $\lambda_1 = 1$ Jordan chain, the last two with the $\lambda_2 = -2$ chain; the eigenvector solutions are the pure exponentials $\mathbf{u}_1(t), \mathbf{u}_4(t)$. The general solution is an arbitrary linear combination of these five basis solutions.

Theorem 8.62. *Let A be an $n \times n$ matrix. Then the solutions* (8.63) *constructed from the Jordan chains in a Jordan basis of A form a basis for the n-dimensional solution space for the corresponding linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$.*

While the full computational details can be quite messy, in practical situations one can glean a significant amount of information about the solutions to the system without much fuss. The following result outlines a general characterization of the solutions of homogeneous linear systems of ordinary differential equations. The result is direct consequence of the general solution formulae in (8.63).

Theorem 8.63. *Let A be a real, square matrix. The general real solution to any constant coefficient homogeneous linear system $\dot{\mathbf{u}} = A\,\mathbf{u}$ is a linear combination of n linearly independent solutions of the following types*:

(a) *If λ is a complete real eigenvalue of multiplicity m, then there exist m linearly independent solutions of the form*

$$\mathbf{u}_k(t) = e^{\lambda t}\,\mathbf{v}_k, \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{v}_1, \dots, \mathbf{v}_m$ are linearly independent eigenvectors.

(b) *If $\mu \pm \mathrm{i}\,\nu$ form a pair of complete complex conjugate eigenvalues of multiplicity m, then there exist $2m$ linearly independent real solutions of the forms*

$$\begin{aligned} \mathbf{u}_k(t) &= e^{\mu t}\big[\,\cos \nu t\,\mathbf{w}_k - \sin \nu t\,\mathbf{z}_k\,\big], \\ \widehat{\mathbf{u}}_k(t) &= e^{\mu t}\big[\,\sin \nu t\,\mathbf{w}_k + \cos \nu t\,\mathbf{z}_k\,\big], \end{aligned} \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{v}_k = \mathbf{w}_k \pm \mathrm{i}\,\mathbf{z}_k$ are the associated complex conjugate eigenvectors.

(c) *If λ is an incomplete real eigenvalue of multiplicity m and $r = \dim V_\lambda$, then there exist m linearly independent solutions of the form*

$$\mathbf{u}_k(t) = e^{\lambda t}\,\mathbf{p}_k(t), \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{p}_k(t)$ is a vector of polynomials of degree $\le m - r$.

(d) *If $\mu \pm \mathrm{i}\,\nu$ form a pair of incomplete complex conjugate eigenvalues of multiplicity m and $r = \dim V_\lambda$, then there exist $2m$ linearly independent real solutions*

$$\begin{aligned} \mathbf{u}_k(t) &= e^{\mu t}\,\big[\,\cos \nu t\,\mathbf{p}_k(t) - \sin \nu t\,\mathbf{q}_k(t)\,\big], \\ \widehat{\mathbf{u}}_k(t) &= e^{\mu t}\,\big[\,\sin \nu t\,\mathbf{p}_k(t) + \cos \nu t\,\mathbf{q}_k(t)\,\big], \end{aligned} \qquad\qquad k = 1, \dots, m,$$

where $\mathbf{p}_k(t), \mathbf{q}_k(t)$ are vectors of polynomials of degree $\le m - r$.

Corollary 8.64. *Every real solution to a homogeneous linear system of ordinary differential equations is a vector-valued function whose entries are linear combinations of functions of the particular form $t^k\,e^{\mu t}\cos \nu t$ and $t^k\,e^{\mu t}\sin \nu t$, i.e., sums of products of exponentials, trigonometric functions and polynomials. The exponents μ are the real parts of the eigenvalues of the coefficient matrix; the trigonometric frequencies ν are the imaginary parts of the eigenvalues; nonconstant polynomials appear only if the matrix is incomplete.*

Example 8.65. The incomplete cases should remind the reader of the solution to a single scalar ordinary differential equation in the case of a repeated root to the characteristic equation. For example, to solve the second order equation

$$\frac{d^2u}{dt^2} - 2\,\frac{du}{dt} + u = 0,$$

we substitute the exponential ansatz $u = e^{\lambda t}$, leading to the characteristic equation

$$\lambda^2 - 2\,\lambda + 1 = 0.$$

There is only one double root, $\lambda = 1$, and hence, up to scalar multiple, only one exponential solution $u_1(t) = e^t$. In the scalar case, the second "missing" solution is obtained by just multiplying by t, so that $u_2(t) = t\,e^t$. The general solution is

$$u(t) = c_1\,u_1(t) + c_2\,u_2(t) = c_1\,e^t + c_2\,t\,e^t.$$

The equivalent phase plane system (8.12) is

$$\frac{d\mathbf{u}}{dt} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}\mathbf{u}, \qquad \text{where} \qquad \mathbf{u}(t) = \begin{pmatrix} u(t) \\ \dot{u}(t) \end{pmatrix}.$$

Note that the coefficient matrix is incomplete — it has $\lambda = 1$ as a double eigenvalue, but only one independent eigenvector, namely $\mathbf{v} = (\,1, 1\,)^T$. The two linearly independent solutions to the phase plane system can be constructed from the solutions $u_1(t) = e^t, u_2(t) = t\,e^t$ to the original equation, and so

$$\mathbf{u}_1(t) = \begin{pmatrix} e^t \\ e^t \end{pmatrix}, \qquad \mathbf{u}_2(t) = \begin{pmatrix} t\,e^t \\ t\,e^t + e^t \end{pmatrix}. \tag{8.64}$$

Note the appearance of the polynomial factor t in the solution formula. The general solution is obtained as a linear combination of these two basic solutions. *Warning*: In (8.64), the second vector solution $\mathbf{u}_2$ is *not* obtained from the first by merely multiplying by t. Incomplete systems are not that easy to handle!

8.8. Stability of Linear Systems.

With the solution formulae in hand, we are now ready to study the qualitative features of first order linear dynamical systems. Our primary focus will be on stability properties of the equilibrium solution(s). The starting point is a simple calculus lemma, whose proof is left to the reader.

Lemma 8.66. *Let μ, ν be real and $k \geq 0$ an integer. A function of the form*

$$f(t) = t^k\,e^{\mu\,t}\cos\nu\,t \qquad \text{or} \qquad t^k\,e^{\mu\,t}\sin\nu\,t \tag{8.65}$$

will decay to zero for large t, so $\lim_{t\to\infty} f(t) = 0$, if and only if $\mu < 0$. The function remains bounded, so $|\,f(t)\,| \leq C$ for all $t \geq 0$, if and only if either $\mu < 0$, or $\mu = 0$ and $k = 0$.

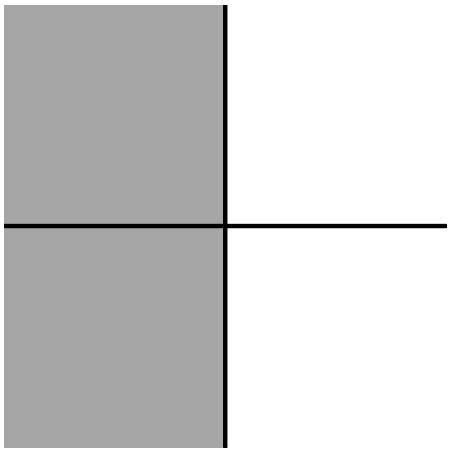


Figure 8.4. The Left Half Plane.

In other words, exponential decay, where $\mu < 0$, will always cancel out polynomial growth, while trigonometric functions remain bounded. Now, in the solution to our ordinary differential equation, the functions (8.65) come from the eigenvalues $\lambda = \mu + \mathrm{i}\,\nu$ of the coefficient matrix. The lemma implies that the asymptotic behavior of the solutions, and hence their stability, depends on the sign of $\mu = \operatorname{Re}\lambda$. If $\mu < 0$, then the solutions decay to zero at an exponential rate as $t \to \infty$. If $\mu > 0$, then the solutions become unbounded as $t \to \infty$. In the borderline case $\mu = 0$, the solutions remain bounded provided they don't involve any powers of t.

Asymptotic stability of the equilibrium zero solution requires that all other solutions tend to $\mathbf{0}$ as $t \to \infty$, and hence all the eigenvalues must satisfy $\mu = \operatorname{Re}\lambda < 0$. Or, stated another way, all eigenvalues must lie in the *left half plane* — the subset of $\mathbb{C}$ to the left of the imaginary axis, as in Figure 8.4. In this manner, we have demonstrated the fundamental asymptotic stability criterion for linear systems.

Theorem 8.67. *A first order linear, homogeneous, constant-coefficient system of ordinary differential equations $\dot{\mathbf{u}} = A\,\mathbf{u}$ has asymptotically stable zero solution if and only if all the eigenvalues of the coefficient matrix A lie in the left half plane*: $\operatorname{Re}\lambda < 0$. *On the other hand, if A has one or more eigenvalues with positive real part,* $\operatorname{Re}\lambda > 0$, *then the zero solution is unstable.*

Example 8.68. Consider the system

$$\frac{du}{dt} = 2\,u - 6\,v + w, \qquad \frac{dv}{dt} = 3\,u - 3\,v - w, \qquad \frac{dw}{dt} = 3\,u - v - 3\,w.$$

The coefficient matrix $A = \begin{pmatrix} 2 & -6 & 1 \\ 3 & -3 & -1 \\ 3 & -1 & -3 \end{pmatrix}$ is found to have eigenvalues $\lambda_1 = -2$, $\lambda_2 = -1 + \mathrm{i}\,\sqrt{6}$, $\lambda_3 = -1 - \mathrm{i}\,\sqrt{6}$, with respective real parts $-2, -1, -1$. Therefore, according to Theorem 8.67, the zero solution $u \equiv v \equiv w \equiv 0$ is asymptotically stable. Indeed, the solutions involve linear combinations of the functions $e^{-2\,t}$, $e^{-t}\cos\sqrt{6}\,t$, and $e^{-t}\sin\sqrt{6}\,t$, all of which decay to 0 at an exponential rate. The latter two have the slowest decay

rate, and so most solutions to the linear system go to $\mathbf{0}$ like a multiple of e^{-t}, i.e., at an exponential rate determined by the least negative real part.

A particularly important class of systems are the linear *gradient flows*

$$\frac{d\mathbf{u}}{dt} = -K\,\mathbf{u}, \tag{8.66}$$

in which $K > 0$ is a symmetric, positive definite matrix. According to Theorem 8.37, all the eigenvalues of K are real and positive. Therefore, the eigenvalues of the negative definite coefficient matrix $-K$ of the gradient flow system (8.66) are real and negative. Applying Theorem 8.67, we conclude that the zero solution to any gradient flow system (8.66) with negative definite coefficient matrix $-K$ is asymptotically stable.

Example 8.69. Using the methods of Chapter 3, the matrix $K = \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}$ is found to be positive definite. The associated gradient flow is

$$\frac{du}{dt} = -u - v, \qquad \frac{dv}{dt} = -u - 5\,v. \tag{8.67}$$

The eigenvalues and eigenvectors of $-K = \begin{pmatrix} -1 & -1 \\ -1 & -5 \end{pmatrix}$ are

$$\lambda_1 = -3 + \sqrt{5}\,, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 - \sqrt{5} \end{pmatrix}, \qquad \lambda_2 = -3 - \sqrt{5}\,, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 + \sqrt{5} \end{pmatrix}.$$

Therefore, the general solution to the system is

$$\mathbf{u}(t) = c_1\, e^{(-3+\sqrt{5}\,)t} \begin{pmatrix} 1 \\ 2 - \sqrt{5} \end{pmatrix} + c_2\, e^{(-3-\sqrt{5}\,)t} \begin{pmatrix} 1 \\ 2 + \sqrt{5} \end{pmatrix},$$

or, in components,

$$\begin{aligned} u(t) &= c_1\, e^{(-3+\sqrt{5}\,)t} + c_2\, e^{(-3-\sqrt{5}\,)t}, \\ v(t) &= c_1\,(2 - \sqrt{5})\, e^{(-3+\sqrt{5}\,)t} + c_2\,(2 + \sqrt{5})\, e^{(-3-\sqrt{5}\,)t}. \end{aligned}$$

The solutions clearly tend to zero as $t \to \infty$ at the exponential rate prescribed by the least negative eigenvalue: $-3 + \sqrt{5} = -0.7639\ldots$. This confirms the asymptotic stability of the gradient flow.

The reason for the term "gradient flow" is that the vector field $-K\,\mathbf{u}$ appearing on the right hand side of (8.66) is, in fact, the negative of the gradient of the quadratic function

$$q(\mathbf{u}) = \tfrac{1}{2}\,\mathbf{u}^T K\,\mathbf{u} = \frac{1}{2} \sum_{i,j=1}^{n} k_{ij}\, u_i\, u_j, \tag{8.68}$$

namely $K\,\mathbf{u} = \nabla q(\mathbf{u})$. Thus, we can write (8.66) as

$$\frac{d\mathbf{u}}{dt} = -\nabla q(\mathbf{u}). \tag{8.69}$$

For the particular system (8.67),

$$q(u,v) = \frac{1}{2}\begin{pmatrix} u & v \end{pmatrix}^T \begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \tfrac{1}{2}\,u^2 + u\,v + \tfrac{5}{2}\,v^2,$$

and so the gradient flow is given by

$$\frac{du}{dt} = -\frac{\partial q}{\partial u} = -u - v, \qquad \frac{dv}{dt} = -\frac{\partial q}{\partial v} = -u - 5\,v.$$

In applications, the quadratic function (8.68) often represents the energy in the system. Its negative gradient $-\nabla q$ points in the direction of steepest decrease of q. A good analogy is water flowing down the side of the hill. If $q(u,v)$ denotes the height of the hill at position (u,v) then $\nabla q = (\partial q/\partial u, \partial q/\partial v)$ points directly uphill while $-\nabla q$ points downhill. The water will flow in the direction of steepest decrease, and so satisfy the gradient flow equations (8.69). Since $q(\mathbf{u})$ is positive definite, the zero solution is the minimum of q, and so, by asymptotic stability, solutions to the gradient flow equations will end up at the bottom of the graph of q. The general features of nonlinear gradient flows will be more fully explored in Chapter 19.

Example 8.70. Let us solve the first order system

$$\frac{du}{dt} = -8\,u - w, \qquad \frac{dv}{dt} = -8\,v - w, \qquad \frac{dw}{dt} = -u - v - 7\,w,$$

subject to initial conditions

$$u(0) = 1, \qquad v(0) = -3, \qquad w(0) = 2.$$

The coefficient matrix for the system is

$$\begin{pmatrix} -8 & 0 & -1 \\ 0 & -8 & -1 \\ -1 & -1 & -7 \end{pmatrix} = -\begin{pmatrix} 8 & 0 & 1 \\ 0 & 8 & 1 \\ 1 & 1 & 7 \end{pmatrix} = -K,$$

which is minus the positive definite matrix analyzed in Example 8.38. Using the computed eigenvalues and eigenvectors, we conclude that the general solution has the form

$$\mathbf{u}(t) = \begin{pmatrix} u(t) \\ v(t) \\ w(t) \end{pmatrix} = c_1\, e^{-6\,t} \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} + c_2\, e^{-8\,t} \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_3\, e^{-9\,t} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

The coefficients are prescribed by the initial conditions, which read

$$\mathbf{u}(0) = \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix} = c_1 \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix} + c_2 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = c_1\,\mathbf{v}_1 + c_2\,\mathbf{v}_2 + c_3\,\mathbf{v}_3.$$

Rather than solve this linear system directly, we make use of the fact that the matrix is symmetric, and hence its eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ form an orthogonal basis. Thus, we can apply the orthogonal basis formula (5.8) to compute the coefficients

$$c_1 = \frac{\langle\, \mathbf{u}(0)\,;\mathbf{v}_1\,\rangle}{\|\,\mathbf{v}_1\,\|^2} = \frac{6}{6} = 1, \qquad c_2 = \frac{\langle\, \mathbf{u}(0)\,;\mathbf{v}_2\,\rangle}{\|\,\mathbf{v}_2\,\|^2} = \frac{-4}{2} = -2, \qquad c_3 = \frac{\langle\, \mathbf{u}(0)\,;\mathbf{v}_3\,\rangle}{\|\,\mathbf{v}_3\,\|^2} = 0.$$

We conclude that the solution to the initial value problem is

$$\mathbf{u}(t) = \begin{pmatrix} -e^{-6t} + 2e^{-8t} \\ -e^{-6t} - 2e^{-8t} \\ 2e^{-6t} \end{pmatrix}.$$

In particular, the exponential decay rate is 6 — as indicated by the largest eigenvalue of K — since e^{-6t} is the slowest decaying exponential in the solution.

Extension of the asymptotic stability criterion of Theorem 8.67 to stable equilibria is not difficult.

Theorem 8.71. *A first order linear, homogeneous, constant-coefficient system of ordinary differential equations* (8.55) *has stable zero solution if and only if all the eigenvalues satisfy* $\operatorname{Re}\lambda \leq 0$, *and, moreover, any eigenvalue lying on the imaginary axis,* $\operatorname{Re}\lambda = 0$, *is complete, meaning that it has as many independent eigenvectors as its multiplicity.*

Proof: The proof is the same as above, using Corollary 8.64 and the decay properties in Lemma 8.66. All the eigenvalues with negative real part lead to exponentially decaying solutions — even if they are incomplete. If a purely imaginary eigenvalue is complete, then the associated solutions only involve trigonometric functions, and hence remain bounded. This suffices to maintain stability. On the other hand, solutions associated with incomplete purely imaginary eigenvalues contain powers of t multiplying sines and cosines, and hence cannot remain bounded as $t \to \infty$. *Q.E.D.*

Example 8.72. A *Hamiltonian system* in the plane takes the form

$$\frac{du}{dt} = \frac{\partial H}{\partial v}, \qquad \frac{dv}{dt} = -\frac{\partial H}{\partial u}, \tag{8.70}$$

where $H(u, v)$ is known as the *Hamiltonian function*. If

$$H(u, v) = \tfrac{1}{2}\, a\, u^2 + b\, u\, v + \tfrac{1}{2}\, c\, v^2 \tag{8.71}$$

is a quadratic form, then the Hamiltonian system is

$$\dot{u} = b\, u + c\, v, \qquad \dot{v} = -a\, u - b\, v, \tag{8.72}$$

homogeneous, linear with coefficient matrix $A = \begin{pmatrix} b & c \\ -a & -b \end{pmatrix}$. The characteristic equation is

$$\det(A - \lambda\, \mathrm{I}\,) = \lambda^2 + (a\, c - b^2) = 0.$$

If H is positive or negative definite, then $a\, c - b^2 > 0$, and so the roots of the characteristic equation are purely imaginary: $\lambda = \pm\, \mathrm{i}\, \sqrt{a\, c - b^2}$. Since the eigenvalues are simple, the stability criterion of Theorem 8.71 holds and we conclude that planar Hamiltonian systems with definite Hamiltonian function are stable.

8.9. Two-Dimensional Systems.

The two-dimensional case is particularly instructive, since many of the most important phenomena are already made manifest there. Moreover, the solutions can be easily pictured by their phase portraits. In this section, we will present a complete classification of the possible qualitative behaviors of real, planar linear dynamical systems.

Setting $\mathbf{u}(t) = (u(t), v(t))^T$, such a system $\dot{\mathbf{u}} = A\,\mathbf{u}$ has the explicit form

$$\frac{du}{dt} = a\,u + b\,v, \qquad \frac{dv}{dt} = c\,u + d\,v, \tag{8.73}$$

where $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is the coefficient matrix. As in Section 8.1, we will refer to the (u, v)–plane as the *phase plane*. In particular, phase plane equivalents (8.12) of second order scalar equations form a special class.

According to (8.25), the characteristic equation for the given 2×2 matrix is

$$\det(A - \lambda\,\mathrm{I}\,) = \lambda^2 - \tau\,\lambda + \delta = 0, \tag{8.74}$$

where

$$\tau = \operatorname{tr} A = a + d, \qquad \delta = \det A = a\,d - b\,c, \tag{8.75}$$

are, respectively, the trace and the determinant of A. The nature of the eigenvalues, and hence the solutions, is therefore almost entirely determined by these two quantities. The sign of the *discriminant*

$$\Delta = \tau^2 - 4\,\delta = (\operatorname{tr} A)^2 - 4 \det A = (a - d)^2 - 4\,b\,c \tag{8.76}$$

determines whether the roots or eigenvalues

$$\lambda = \frac{\tau \pm \sqrt{\Delta}}{2} \tag{8.77}$$

are real or complex, and thereby plays a key role in the classification.

Let us summarize the different possibilities as classified by their qualitative behavior. Each situation will be illustrated by a representative phase portrait, which plots a representative sample of the solution trajectories in the phase plane. The complete taxonomy appears in Figure 8.5 below.

Distinct Real Eigenvalues

The coefficient matrix A has two real, distinct eigenvalues $\lambda_1 < \lambda_2$ if and only if the discriminant (8.76) of the quadratic equation (8.74) is positive: $\Delta > 0$. In this case, the solutions take the exponential form

$$\mathbf{u}(t) = c_1\,e^{\lambda_1 t}\,\mathbf{v}_1 + c_2\,e^{\lambda_2 t}\,\mathbf{v}_2, \tag{8.78}$$

where $\mathbf{v}_1, \mathbf{v}_2$ are the eigenvectors and c_1, c_2 are arbitrary constants, to be determined by the initial conditions. The asymptotic behavior of the solutions is governed by the size of the

eigenvalues. Let $V_k = \{c\,\mathbf{v}_k\}$, $k = 1, 2$, denote the "eigenlines", i.e., the one-dimensional eigenspaces associated with each eigenvalue λ_k.

There are five qualitatively different cases, depending upon the signs of the two eigenvalues. These are listed by their descriptive name, followed by the required conditions on the discriminant, trace and determinant of the coefficient matrix.

Ia. Stable Node: $\Delta > 0, \qquad \operatorname{tr} A < 0, \qquad \det A > 0.$

If $\lambda_1 < \lambda_2 < 0$ are both negative, then $\mathbf{0}$ is an asymptotically *stable node*. The solutions all tend to $\mathbf{0}$ as $t \to \infty$. Since the first exponential $e^{\lambda_1 t}$ decreases much faster than the second $e^{\lambda_2 t}$, the first term in the solution (8.78) will soon become negligible, and hence $\mathbf{u}(t) \approx c_2\, e^{\lambda_2 t}\, \mathbf{v}_2$ when t is large. Therefore, all solutions with $c_2 \neq 0$ will arrive at the origin along a curve that is tangent to the eigenline V_2. The solutions with $c_2 = 0$ come in to the origin directly along the eigenline V_1, and at a faster rate. Conversely, as $t \to -\infty$, all solutions become unbounded: $\| \mathbf{u}(t) \| \to \infty$. In this case, the first exponential grows faster than the second, and so the solutions $\mathbf{u}(t) \approx c_1\, e^{\lambda_1 t}\, \mathbf{v}_1$ for $t \ll 0$. Thus, as they escape to ∞, the solution trajectories become more and more parallel to the eigenline V_1 — except for those with $c_1 = 0$ that remain in the eigenline V_2.

Ib. Saddle Point: $\Delta > 0, \qquad \det A < 0.$

If $\lambda_1 < 0 < \lambda_2$, then $\mathbf{0}$ is an unstable *saddle point*. Solutions (8.78) with $c_2 = 0$ start out on the eigenline V_1 and go in to $\mathbf{0}$ as $t \to \infty$, while solutions with $c_1 = 0$ start on V_2 and go to $\mathbf{0}$ as $t \to -\infty$. All other solutions become unbounded at both large positive and large negative times. As $t \to +\infty$, the solutions approach the unstable eigenline V_2, while as $t \to -\infty$, they asymptote to the stable eigenline V_1. The eigenline V_1 is called the *stable manifold*, indicating that solutions that start on it eventually go to the equilibrium point $\mathbf{0}$, while V_2 is the *unstable manifold*, meaning that solutions on it go to equilibrium as $t \to -\infty$.

Ic. Unstable Node: $\Delta > 0, \qquad \operatorname{tr} A > 0, \qquad \det A > 0.$

If the eigenvalues $0 < \lambda_1 < \lambda_2$ are both positive, then $\mathbf{0}$ is an *unstable node*. The phase portrait is the same as that of a stable node, but the solution trajectories are traversed in the opposite direction. Time reversal $t \to -t$ will convert an unstable node into a stable node and vice versa; see Exercise ■. Thus, in the unstable case, the solutions all tend to $\mathbf{0}$ as $t \to -\infty$ and off to ∞ as $t \to \infty$. Except for the solutions on the eigenlines, they asymptote to V_1 as $t \to -\infty$, and become parallel to V_2 as $t \to \infty$.

Id. Stable Line: $\Delta > 0, \qquad \operatorname{tr} A < 0, \qquad \det A = 0.$

If $\lambda_1 < \lambda_2 = 0$, then every point on the eigenline V_2 associated with the zero eigenvalue is an equilibrium point. Every other solution moves along a straight line parallel to V_1 and tends to one of the equilibria on V_2 as $t \to \infty$.

Ie. Unstable Line: $\Delta > 0, \qquad \operatorname{tr} A > 0, \qquad \det A = 0.$

This is merely the time reversal of a stable line. If $0 = \lambda_1 < \lambda_2$, then every point on the eigenline V_1 is an equilibrium. Every other solution moves off to ∞ along a straight line parallel to V_2 as $t \to \infty$, and tends to one of the equilibria on V_1 as $t \to -\infty$.

Complex Conjugate Eigenvalues

The coefficient matrix A has two complex conjugate eigenvalues

$$\lambda = \mu \pm \mathrm{i}\,\nu, \qquad \text{where} \qquad \mu = \tfrac{1}{2}\,\tau = \tfrac{1}{2}\,\operatorname{tr} A, \qquad \nu = \sqrt{-\Delta},$$

if and only if its discriminant is negative: $\Delta < 0$. In this case, the real solutions can be written in the phase–amplitude form (2.7):

$$\mathbf{u}(t) = r\, e^{\mu\, t}\, [\, \cos(\nu\, t - \sigma)\, \mathbf{w} + \sin(\nu\, t - \sigma)\, \mathbf{z}\,], \tag{8.79}$$

where $\mathbf{v}_{\pm} = \mathbf{w} \pm \mathrm{i}\,\mathbf{z}$ are the complex eigenvectors. As noted above, the two vectors $\mathbf{w}, \mathbf{z}$ are always linearly independent. The amplitude r and phase shift σ are uniquely prescribed by the initial conditions. There are three subcases, depending upon the sign of the real part μ, or, equivalently, the sign of the trace of A.

IIa. Stable Focus: $\Delta < 0, \qquad \operatorname{tr} A < 0.$

If $\mu < 0$, then $\mathbf{0}$ is an asymptotically *stable focus*. As $t \to \infty$, the solutions all spiral in to $\mathbf{0}$ with "frequency" ν — meaning it takes time $2\pi/\nu$ for the solution to go once around the origin. As $t \to -\infty$, the solutions spiral off to ∞ with the same frequency.

IIb. Center: $\Delta < 0, \qquad \operatorname{tr} A = 0.$

If $\mu = 0$, then $\mathbf{0}$ is a *center*. The solutions all move periodically around elliptical orbits, with common frequency ν and period $2\pi/\nu$. In particular, solutions that start out near $\mathbf{0}$ stay nearby, and hence a center is a stable, but not asymptotically stable, equilibrium.

IIc. Unstable Focus: $\Delta < 0, \qquad \operatorname{tr} A > 0.$

If $\mu > 0$, then $\mathbf{0}$ is an *unstable focus*. The phase portrait is the time reversal, $t \mapsto -t$, of a stable focus, with solutions spiraling off to ∞ as $t \to \infty$ and in to the origin as $t \to -\infty$, again with a common "frequency" ν.

Incomplete Double Real Eigenvalue

The matrix will have a double real eigenvalue $\lambda = \frac{1}{2}\,\tau = \frac{1}{2}\,\operatorname{tr} A$ if and only if the discriminant vanishes: $\Delta = 0$. The formula for the solutions depends on whether the eigenvalue λ is complete or not. If λ is an incomplete eigenvalue, admitting only one independent eigenvector $\mathbf{v}$, then the solutions are no longer given by simple exponentials. The general formula is

$$\mathbf{u}(t) = (c_1 + c_2\, t) e^{\lambda\, t}\, \mathbf{v} + c_2\, e^{\lambda\, t}\, \mathbf{w}, \tag{8.80}$$

where $(A - \lambda\, \mathrm{I}\,)\mathbf{w} = \mathbf{v}$, and so $\mathbf{v}, \mathbf{w}$ form a Jordan chain for the coefficient matrix. We let $V = \{c\,\mathbf{v}\}$ denote the eigenline associated with the genuine eigenvector $\mathbf{v}$.

IIIa. Stable Improper Node: $\Delta = 0, \qquad \operatorname{tr} A < 0, \qquad A \neq \lambda\, \mathrm{I}\,.$

If $\lambda < 0$ then $\mathbf{0}$ is an asymptotically *stable improper node*. Since $t\, e^{\lambda\, t}$ is larger than $e^{\lambda\, t}$ for $t > 1$, the solutions $\mathbf{u}(t) \approx c_2\, t\, e^{\lambda\, t}$ tend to $\mathbf{0}$ as $t \to \infty$ along a curve that is tangent to the eigenline V. Similarly, as $t \to -\infty$, the solutions go off to ∞, becoming

more and more parallel to the eigenline, but moving away in the opposite direction from their approach.

IIIb. Linear Motion: $\Delta = 0, \qquad \operatorname{tr} A = 0, \qquad A \neq \lambda \, \mathrm{I}\,.$

If $\lambda = 0$, then, as in case Id, every point on the eigenline V is an equilibrium point. Every other solution is a linear, or, more correctly, affine function of T, and so moves along a straight line parallel to V, going off to ∞ in either direction. The origin is an unstable equilibrium point.

IIIc. Unstable Improper Node: $\Delta = 0, \qquad \operatorname{tr} A > 0, \qquad A \neq \lambda \, \mathrm{I}\,.$

If $\lambda > 0$, then $\mathbf{0}$ is an *unstable improper node*. The phase portrait is the time reversal of the stable improper node.

Complete Double Real Eigenvalue

In this case, *every* vector in $\mathbb{R}^2$ is an eigenvector, and so the real solutions take the form $\mathbf{u}(t) = e^{\lambda t}\,\mathbf{v}$, where $\mathbf{v}$ is an *arbitrary* constant vector. In fact, this case occurs if and only if $A = \lambda \, \mathrm{I}$ is a multiple of the identity matrix.

IVa. Stable Star: $A = \lambda \, \mathrm{I}\,, \qquad \lambda < 0.$

If $\lambda < 0$ then $\mathbf{0}$ is an asymptotically stable star. The solution trajectories are the rays emanating from the origin, and the solutions go to $\mathbf{0}$ at an exponential rate as $t \to \infty$.

IVb. Trivial: $A = \mathrm{O}.$

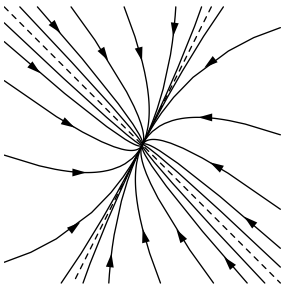
If $\lambda = 0$ then the only possibility is $A = \mathrm{O}$. Now every solution is constant and every point is a (stable) equilibrium point. Nothing happens! This is the only case not pictured in Figure 8.5.

IVc. Unstable Star: $A = \lambda \, \mathrm{I}\,, \qquad \lambda < 0.$

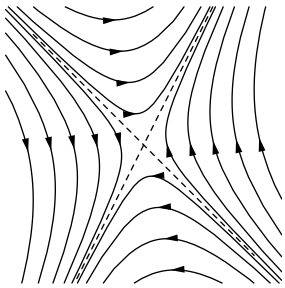
If $\lambda > 0$ then $\mathbf{0}$ is unstable. The phase portrait is the time reversal of the stable star, and so the solutions move along rays, and tend to $\mathbf{0}$ as $t \to -\infty$.

Figure 8.6 indicates where the different possibilities lie, as prescribed by the trace and determinant of the coefficient matrix. The horizontal axis indicates the value of $\tau = \operatorname{tr} A$, while the vertical axis refers to $\delta = \det A$. Points on the parabola $\tau^2 = 4\,\delta$ represent the cases with vanishing discriminant $\Delta = 0$, and correspond to either stars or improper nodes — except for the origin which is either linear motion or trivial. All the asymptotically stable cases lie in the shaded upper left quadrant where $\operatorname{tr} A < 0$ and $\det A > 0$. The borderline points on the coordinate axes are either stable centers, when $\operatorname{tr} A = 0$, $\det A > 0$, or stable lines, when $\operatorname{tr} A < 0$, $\det A = 0$, or the origin, which may or may not be stable depending upon whether A is the zero matrix or not. All other values for the trace and determinant result in unstable equilibria.

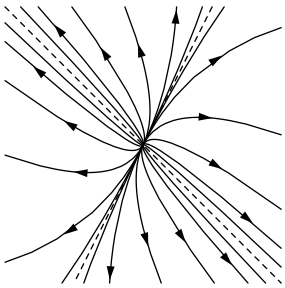
Remark: Time reversal $t \to -t$ changes the sign of the coefficient matrix $A \to -A$, and hence the sign of its trace, $\tau \to -\tau$, while the determinant $\delta = \det A = \det(-A)$ is unchanged. Thus, the effect is to reflect the plot in Figure 8.6 through the vertical axis, interchanging the stable nodes and spirals with their unstable counterparts, while leaving saddle points in the same qualitative form.



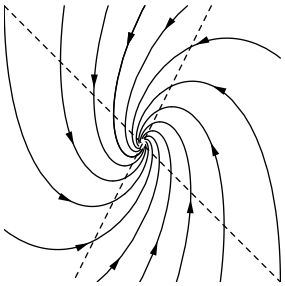
Ia. Stable Node



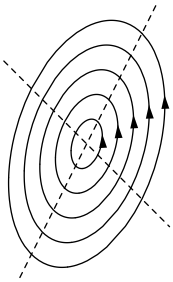
Ib. Saddle Point



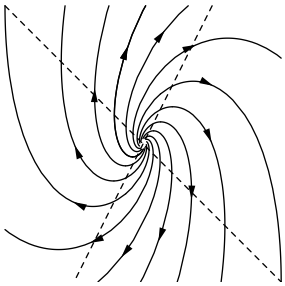
Ic. Unstable Node



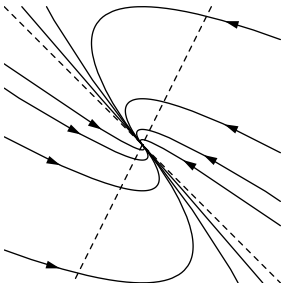
IIa. Stable Focus



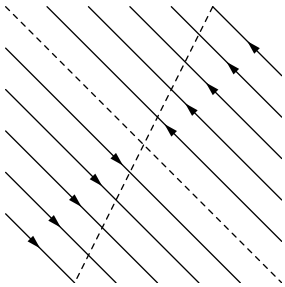
IIb. Center



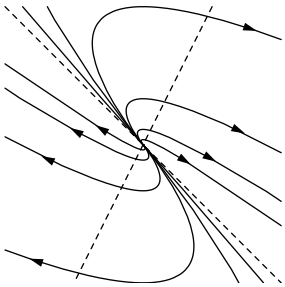
IIc. Unstable Focus



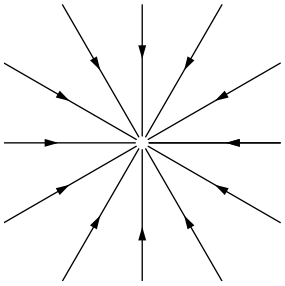
IIIa. Stable Improper Node



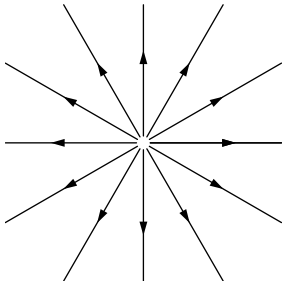
IIIb. Linear Motion



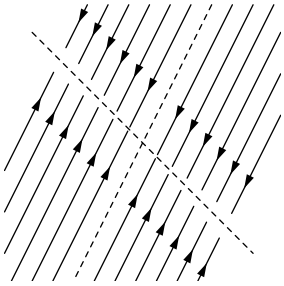
IIIc. Unstable Improper Node



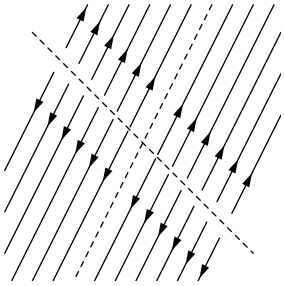
IVa. Stable Star



IVc. Unstable Star



Id. Stable Line



Ie. Unstable Line

Figure 8.5. Phase Portraits.

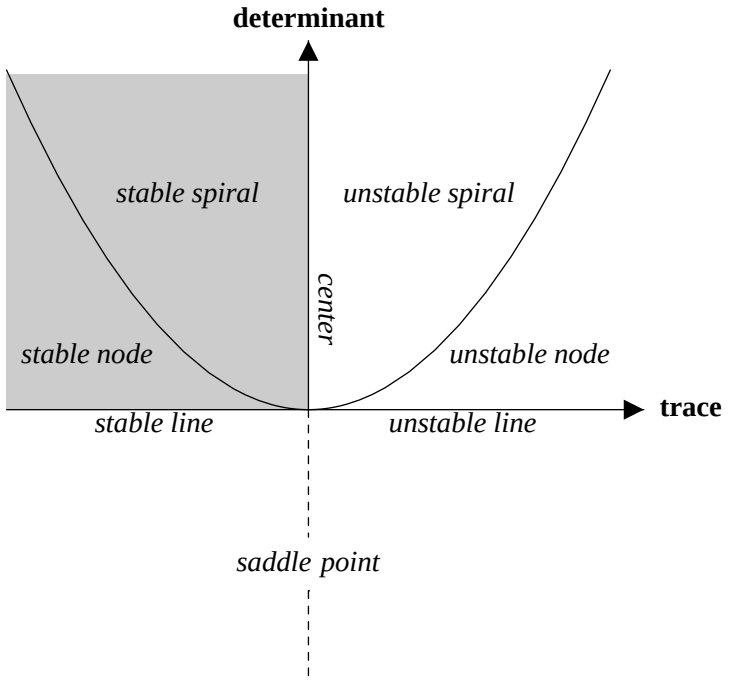


Figure 8.6. Stability Regions for Two–Dimensional Linear Systems.

In physical applications, the coefficient matrix A is usually not known exactly, and so the physical system may, in fact, be a slight perturbation of the mathematical model. Thus, it is important to know which systems are *structurally stable*, meaning the basic qualitative features are preserved under sufficiently small changes in the coefficients.

Now, a small perturbation will alter the entries of the coefficient matrix slightly, and hence move the trace and determinant by a comparably small amount. The net effect is to slightly perturb its eigenvalues. Therefore, the question of structural stability reduces to whether the eigenvalues have moved sufficiently far to send the system into a different stability regime. Asymptotically stable systems remain stable under small enough perturbations, since the property that the eigenvalues have negative real parts is preserved under small perturbation. For a similar reason, unstable systems remain unstable under small perturbations. On the other hand, a borderline stable system — either a center or the trivial system — could become either asymptotically stable or unstable under an adverse perturbation.

Structural stability requires more, since the overall phase portrait should not significantly change. A system in any of the open regions in the Stability Figure 8.6, e.g., a stable spiral, unstable node, saddle point, etc., is structurally stable, whereas a system that lies on the parabola $\tau^2 = 4\,\delta$, or the horizontal axis, or positive vertical axis, e.g., an improper node, a stable line, etc., is not, since a small perturbation could send it into either of the adjoining regions. In other words, structural stability requires that the eigenvalues be distinct and have non-zero real part: $\operatorname{Re} \lambda \neq 0$. This final result also applies to systems in higher dimensions, [**73**].

8.10. Dynamics of Structures.

Chapter 6 was concerned with the equilibrium configurations of mass-spring chains and, more generally, structures made out of elastic bars. We are now able to analyze the dynamical motions of such structures. Consider first a linear mass/spring chain consisting of n masses $m_1, \dots, m_n$ connected together and, possibly, to the top and bottom supports by springs. Let $u_i(t)$ denote the displacement† from equilibrium of the i^{th} mass, and $e_j(t)$ the elongation of the j^{th} spring. Since we are now interested in dynamics, both of these are allowed to depend on time, t.

The motion of each mass is governed by Newton's Second Law,

$$\text{Force} = \text{Mass} \times \text{Acceleration}. \tag{8.81}$$

The acceleration of the i^{th} mass is the second derivative $\ddot{u}_i = d^2u_i/dt^2$ of its displacement. The right hand sides of Newton's Law are thus $m_i\, \ddot{u}_i$, which we collect together in vector form $M\, \ddot{\mathbf{u}}$ by multiplying the second derivative of the displacement vector $\mathbf{u}(t) = (\, u_1(t), \dots, u_n(t)\,)^T$ by the diagonal, positive definite mass matrix $M = \operatorname{diag}(m_1, \dots, m_n)$. Incidentally, the masses of the springs are assumed to be negligible in this approximation.

If, to begin with, we assume no external forcing of the mass/spring system and no frictional effects, then the only force exerted on each mass is due to the elongations of its two connecting springs, which is measured by the components of the internal force vector

$$\mathbf{f} = -K\mathbf{u} = -A^TCA\mathbf{u}. \tag{8.82}$$

Here $K = A^TCA$ the stiffness matrix for the chain, which is constructed from the (reduced) incidence matrix A and the diagonal matrix of spring constants C, as in (6.11).

Substituting the internal force formula (8.82) into Newton's Law (8.81) leads immediately to the fundamental dynamical equations

$$M\,\frac{d^2\mathbf{u}}{dt^2} = -K\mathbf{u} \tag{8.83}$$

governing the free, frictionless motions of the system. The goal is to understand the solutions of this system of second order ordinary differential equations, and then, rather straightforwardly, generalize the methods to cover structures in two and three dimensions as well as electrical circuits containing inductors, resistors and capacitors, all of which are governed by the same basic second order system (8.83) based on the appropriate stiffness or resistivity matrix K.

Example 8.73. The simplest case is that of a single mass connected to a fixed support by a spring. The dynamical system (8.83) reduces to a scalar equation

$$m\,\frac{d^2u}{dt^2} + k\,u = 0. \tag{8.84}$$

† As in Section 6.1, the masses are only allowed to move in the direction of the chain, that is, we restrict our attention to one-dimensional motion.

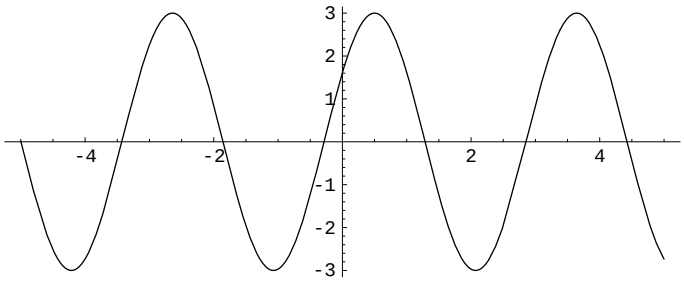


Figure 8.7. Vibration of a Mass.

Here $m > 0$ is the mass, while $k > 0$ is the spring's stiffness. The general solution to this elementary homogeneous, second order linear ordinary differential equation is

$$u(t) = c_1 \cos \omega\, t + c_2 \sin \omega\, t = r \cos(\omega\, t - \delta), \qquad \text{where} \qquad \omega = \sqrt{\frac{k}{m}} \tag{8.85}$$

is the natural frequency of vibration. We have used the phase-amplitude equation (2.7) to rewrite the solution as a single cosine with an amplitude $r = \sqrt{c_1^2 + c_2^2}$, and phase lag $\delta = \tan^{-1} c_2/c_1$. The motion is periodic, with period $P = 2\,\pi/\omega$. The frequency formula $\omega = \sqrt{k/m}$ tells us that stiffer the spring or the lighter the mass, the faster the vibrations. Take note of the square root; it tells us that, for instance, quadrupling the mass only slows down the vibrations by a factor of two.

The constants c_1, c_2 — or their phase-amplitude counterparts r, δ — are determined by the initial conditions. Physically, we need to specify both an initial position and an initial velocity in order to uniquely prescribe the subsequent motion of the system:

$$u(t_0) = a, \qquad \dot{u}(t_0) = b. \tag{8.86}$$

The resulting solution is most conveniently written in the form

$$u(t) = a \cos \omega\,(t - t_0) + \frac{b}{\omega} \sin \omega\,(t - t_0) = r \cos\big[\,\omega\,(t - t_0) - \delta\,\big] \tag{8.87}$$

which has amplitude and phase given by

$$r = \sqrt{a^2 + \frac{\beta^2}{\omega^2}}\,, \qquad \delta = \tan^{-1} \frac{b}{a\,\omega}\,. \tag{8.88}$$

A typical solution is plotted in Figure 8.7.

Let us turn to a more general mass-spring chain or structure. Just as exponentials form the basic building blocks for the solution of systems of first order ordinary differential equations, trigonometric functions form the basic building blocks for solutions to undamped mechanical (and electrical) vibrations governed by second order systems. For simplicity, let us first assume that the masses are all the same and equal to 1 (in some appropriate units), so that (8.83) reduces to

$$\frac{d^2\mathbf{u}}{dt^2} = -\,K\,\mathbf{u}. \tag{8.89}$$

Mimicking our success in the first order case, let us try substituting the trigonometric ansatz

$$\mathbf{u}(t) = \cos(\omega\, t)\, \mathbf{v}, \tag{8.90}$$

with $\mathbf{v} \neq \mathbf{0}$ denoting a constant vector, into the system (8.89). Differentiating (8.90) directly, we find

$$\frac{d\mathbf{u}}{dt} = -\,\omega\, \sin(\omega\, t)\, \mathbf{v}, \qquad\qquad \frac{d^2\mathbf{u}}{dt^2} = -\,\omega^2\, \cos(\omega\, t)\, \mathbf{v}.$$

Therefore, our ansatz (8.90) will solve (8.89) if and only if

$$K\, \mathbf{v} = \omega^2\, \mathbf{v},$$

which means that $\mathbf{v}$ is an eigenvector of K with eigenvalue

$$\lambda = \omega^2. \tag{8.91}$$

Now, there is nothing special about the cosine function — the same computation also applies to the sine function, and tells us that $\mathbf{u}(t) = \sin(\omega\, t)\, \mathbf{v}$ is also a solution whenever $\mathbf{v}$ is an eigenvector with eigenvalue $\lambda = \omega^2$. Summarizing:

Lemma 8.74. *If $\mathbf{v}$ is an eigenvector of the matrix K with eigenvalue $\lambda = \omega^2$, then the trigonometric vector functions $\mathbf{u}(t) = \cos(\omega\, t)\, \mathbf{v}$ and $\mathbf{u}(t) = \sin(\omega\, t)\, \mathbf{v}$ are solutions to the second order system $\ddot{\mathbf{u}} = -\,K\, \mathbf{u}$.*

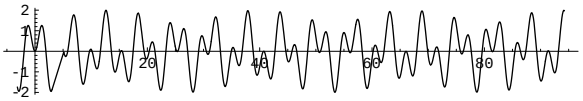
Remark: Alternatively, one can utilize the complex exponential solutions $e^{\,\mathrm{i}\,\omega\, t}\, \mathbf{v}$ and $e^{-\,\mathrm{i}\,\omega\, t}\, \mathbf{v}$, which are related to the trigonometric solutions via Euler's formula (3.76). This is common practice in electrical circuit analysis — although electrical engineers tend to use j instead of i to denote the square root of -1.

Stable Structures

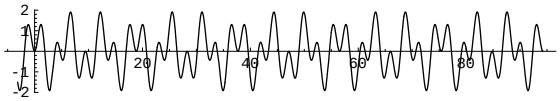
Let us next analyze the motion of a stable structure, of the type introduced in Section 6.3. According to Theorem 6.8, stability requires that the reduced stiffness matrix be positive definite: $K > 0$. Theorem 8.39 says that all the eigenvalues of K are strictly positive, $\lambda_i > 0$, which is good, since it implies that the eigenvalue/frequency relation (8.91) yields real frequencies $\omega_i = \sqrt{\lambda_i}$. Moreover, all positive definite matrices are complete, and so, even when there are fewer than n different eigenvalues, there always exist a complete system of n linearly independent real eigenvectors that form an orthogonal basis for $\mathbb{R}^n$.

Since (8.89) is a second order system of homogeneous linear equations in n unknowns, we require $2\,n$ linearly independent solutions. Lemma 8.74 produces 2 independent solutions for each positive eigenvalue (counted with multiplicity), and hence, assuming positive definiteness, there are indeed $2\,n$ linearly independent solutions,

$$\begin{aligned} \mathbf{u}_i(t) &= \cos(\omega_i\, t)\, \mathbf{v}_i = \cos\big(\sqrt{\lambda_i}\, t\,\big)\, \mathbf{v}_i, \\ \widetilde{\mathbf{u}}_i(t) &= \sin(\omega_i\, t)\, \mathbf{v}_i = \sin\big(\sqrt{\lambda_i}\, t\,\big)\, \mathbf{v}_i, \end{aligned} \qquad\qquad i = 1, \dots, n, \tag{8.92}$$



$\cos t + \cos \sqrt{5}\, t$



$\cos t + \cos \frac{7}{3}\, t$

Figure 8.8. Quasi–Periodic and Periodic Functions.

governed by the n mutually orthogonal (or even orthonormal) eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ of K. The general solution to (8.89) is an arbitrary linear combination,

$$\mathbf{u}(t) = \sum_{i=1}^{n} \big[\, c_i \cos \omega_i t + d_i \sin \omega_i t \,\big]\, \mathbf{v}_i = \sum_{i=1}^{n} r_i \cos(\omega_i t - \delta_i)\, \mathbf{v}_i, \tag{8.93}$$

of these $2n$ basic solutions. The $2n$ coefficients c_i, d_i — or their phase–amplitude counterparts $r_i > 0$, and $0 \le \delta_i < 2\pi$ — are uniquely determined by the initial conditions. As in (8.86), we need to specify both the initial positions and initial velocities of all the masses; this requires a total of $2n$ initial conditions

$$\mathbf{u}(t_0) = \mathbf{a}, \qquad \dot{\mathbf{u}}(t_0) = \mathbf{b}. \tag{8.94}$$

The individual solutions (8.92) are known as the *normal modes of vibration* of our system, and the $\omega_i = \sqrt{\lambda_i}$ the *normal frequencies*, which are the *square roots of the eigenvalues of the stiffness matrix*. Each is a periodic, vector-valued function of period $P_i = 2\pi/\omega_i$. Linear combinations of such periodic functions are, in general, called *quasi-periodic*. Unless the ratios ω_i/ω_j between the frequencies are all rational numbers, such a quasi-periodic function will never precisely repeat itself, and so can appear to be chaotic, even though it is built up from a few very simple periodic constituents. The reader will find it very instructive to graph some simple quasiperiodic functions, say

$$f(t) = c_1 \cos t + c_2 \cos \sqrt{5}\, t$$

for various values of c_1, c_2. Comparison with a case where the frequencies are all rational, say

$$f(t) = c_1 \cos t + c_2 \cos \tfrac{7}{3}\, t$$

is also instructive. The former is truly quasiperiodic, while the the latter is, in fact, periodic with period 6π. Most structures and circuits exhibit quasi-periodic vibrational motions. Let us analyze a couple of simple examples.

Example 8.75. Consider a chain consisting of two equal unit masses connected in a row to supports by three springs, as in Figure 8.9. If the spring constants are c_1, c_2, c_3 (from top to bottom), then the stiffness matrix is

$$K = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} c_1 & 0 & 0 \\ 0 & c_2 & 0 \\ 0 & 0 & c_3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 + c_3 \end{pmatrix}$$

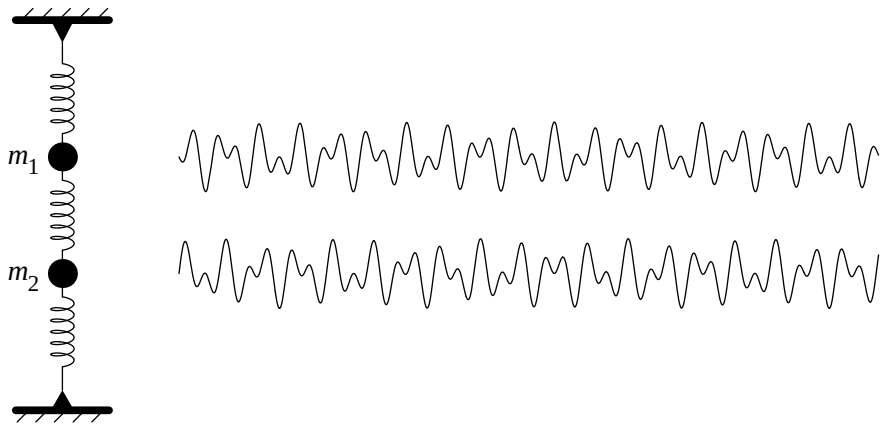


Figure 8.9. Motion of a Double Mass/Spring Chain with Fixed Supports.

The eigenvalues and eigenvectors of K will prescribe the normal modes of vibration and natural frequencies of our two–mass chain.

. Let us look in detail at the case of identical springs, and choose our units so that $c_1 = c_2 = c_3 = 1$. Then $K = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ has eigenvalues and eigenvectors

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = 3, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The general solution to the system is then

$$\mathbf{u}(t) = r_1 \cos(t - \delta_1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + r_2 \cos(\sqrt{3}\, t - \delta_2) \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The first summand is the normal mode vibrating at the relatively slow frequency $\omega_1 = 1$, with the two masses moving in tandem. The second summand is the normal mode that vibrates faster, with frequency $\omega_2 = \sqrt{3}$, in which the two masses move in opposing directions. The general motion is a linear combination of these two normal modes. Since the frequency ratio $\omega_2/\omega_1 = \sqrt{2}$ is irrational, the motion is quasi-periodic. The system never quite returns to its initial configuration — unless it happens to be vibrating in only one of the normal modes. Figure 8.9. A typical graph of the displacements of the masses is plotted in

If we eliminate the bottom spring, so the masses are just hanging from the top support as in Figure 8.10, then the reduced incidence matrix $A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$ loses its last row. Assuming that the springs have unit stiffnesses $c_1 = c_2 = 1$, the corresponding stiffness matrix is

$$K = A^T A = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors are

$$\lambda_1 = \frac{3 - \sqrt{5}}{2}, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ \frac{1+\sqrt{5}}{2} \end{pmatrix}, \qquad\qquad \lambda_2 = \frac{3 + \sqrt{5}}{2}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ \frac{1-\sqrt{5}}{2} \end{pmatrix}.$$

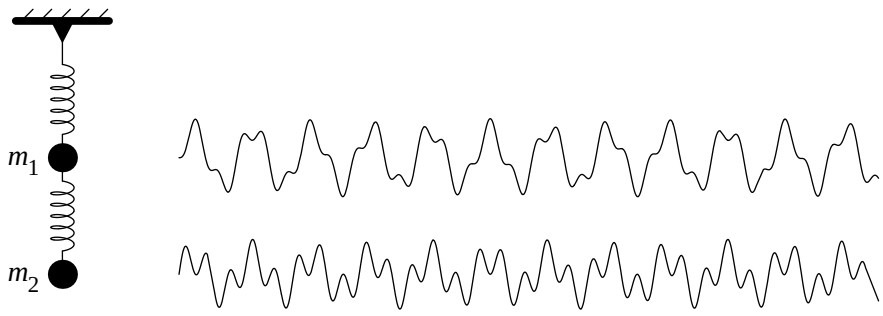


Figure 8.10. Motion of a Double Mass/Spring Chain with One Free End.

The general solution to the system is then

$$\mathbf{u}(t) = r_1 \cos\left(\sqrt{\tfrac{3-\sqrt{5}}{2}}\, t - \delta_1\right) \begin{pmatrix} 1 \\ \frac{1+\sqrt{5}}{2} \end{pmatrix} + r_2 \cos\left(\sqrt{\tfrac{3+\sqrt{5}}{2}}\, t - \delta_2\right) \begin{pmatrix} 1 \\ \frac{1-\sqrt{5}}{2} \end{pmatrix}.$$

The slower normal mode, with frequency $\omega_1 = \sqrt{\frac{3-\sqrt{5}}{2}}$, has the masses moving in tandem, with the bottom mass moving proportionally $\dfrac{1+\sqrt{5}}{2}$ farther. The faster normal mode, with frequency $\omega_2 = \sqrt{\frac{3+\sqrt{5}}{2}}$, has the masses moving in opposite directions, with the top mass experiencing the larger displacement. Moreover, both modes vibrate slower than when there is a bottom support. A typical solution is plotted in Figure 8.10.

Example 8.76. Consider a three mass/spring chain, with unit springs and masses, and both ends attached to fixed supports. The stiffness matrix $K = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ has eigenvalues and eigenvectors

$$\lambda_1 = 2 - \sqrt{2}\,, \qquad \lambda_2 = 2, \qquad \lambda_3 = 2 + \sqrt{2}\,,$$
$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}.$$

The three normal modes, from slowest to fastest, have frequencies

(*a*) $\omega_1 = \sqrt{2-\sqrt{2}}$: all three masses move in tandem, with the middle one moving $\sqrt{2}$ times as far.

(*b*) $\omega_2 = \sqrt{2}$: the two outer masses move in opposing directions, while the middle mass does not move.

(*c*) $\omega_3 = \sqrt{2+\sqrt{2}}$: the two outer masses move in tandem, while the inner mass moves $\sqrt{2}$ times as far in the opposite direction.

The general motion is a quasi-periodic combination of these three normal modes. As such, to the naked eye it can look very complicated. Our mathematical analysis unmasks the innate simplicity, where the complex dynamics are, in fact, entirely governed by just three fundamental modes of vibration.

Unstable Structures

So far, we have just dealt with the stable case, when the reduced incidence matrix has trivial kernel, $\ker A = \{\mathbf{0}\}$, and so the stiffness matrix $K = A^T C A$ is positive definite. Unstable configurations, which can admit rigid motions and/or mechanisms, will provide additional complications. The simplest version is a single mass that is not attached to any spring. The mass experiences no restraining force, and has motion governed by the elementary second order ordinary differential equation

$$m\,\frac{d^2u}{dt^2} = 0. \tag{8.95}$$

The general solution

$$u(t) = c\,t + d \tag{8.96}$$

has the mass either sitting still at a specified position or moving in a straight line with constant velocity $c \neq 0$.

More generally, suppose that the stiffness matrix $K = A^T C\,A$ for our structure is only positive semi-definite. Each vector $\mathbf{0} \neq \mathbf{v} \in \ker A = \ker K$ represents a mode of instability of the system. Since $K\mathbf{v} = \mathbf{0}$, we can interpret $\mathbf{v}$ as a *null eigenvector* of K, with eigenvalue $\lambda = 0$. Lemma 8.74 gives us two solutions to the dynamical equations (8.89) with associated "frequency" $\omega = \sqrt{\lambda} = 0$. The first, $\mathbf{u}(t) = \cos(\omega\,t)\,\mathbf{v} = \mathbf{v}$ is a constant solution, i.e., an equilibrium configuration of the system. Thus, an unstable system does not have a unique equilibrium configuration, since every null eigenvector $\mathbf{v} \in \ker K$ gives a constant solution. On the other hand, the second solution, $\mathbf{u}(t) = \sin(\omega\,t)\,\mathbf{v} = \mathbf{0}$, is trivial, and of no help for constructing the general solution. But, to obtain the general solution to the system, we still need a second independent solution coming from the null eigenvalue. In analogy with the scalar case (8.96), let us try the solution ansatz $\mathbf{u}(t) = t\,\mathbf{v}$, which works, since

$$\frac{d\mathbf{u}}{dt} = \mathbf{v}, \qquad\qquad \mathbf{0} = \frac{d^2\mathbf{u}}{dt^2} = K\,\mathbf{u} = t\,K\,\mathbf{v}.$$

Therefore, to each element of the kernel of the stiffness matrix — i.e., each rigid motion and mechanism — there is a two-dimensional family of solutions

$$\mathbf{u}(t) = (c\,t + d)\,\mathbf{v}. \tag{8.97}$$

When $c = 0$, it reduces to a constant equilibrium solution; when $c \neq 0$, the solution is moving with constant velocity in the null direction $\mathbf{v}$ representing an unstable mode in the system. The general solution will be a linear superposition of the vibrational modes corresponding to the positive eigenvalues and these unstable linear motions corresponding to the zero eigenvalues.

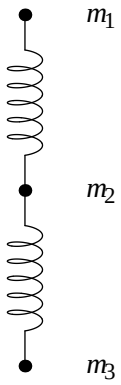


Figure 8.11. A Triatomic Molecule.

Remark: If the null direction $\mathbf{v}$ represents a rigid translation, then the entire structure will move in that direction. If $\mathbf{v}$ represents an infinitesimal rotation, then, owing to our linear approximation to the true nonlinear bar motions, the individual masses will move in straight lines, which are the tangent approximations to the circular motion that occurs in the true physical, nonlinear regime. We refer to the earlier discussion in Chapter 6 for details. Finally, if we excite a mechanism, then the masses will again follow straight lines, moving in different directions, whereas in the full nonlinear physical regime the masses may move along much more complicated curved trajectories.

Example 8.77. Consider a system of three unit masses connected in a line by two unit springs, but not attached to any fixed supports, as illustrated in Figure 8.11. This structure could be viewed as a simplified model of a triatomic molecule that is only allowed to move the vertical direction. The incidence matrix is $A = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$ and, since we are dealing with unit springs, the stiffness matrix is

$$K = A^T A = \begin{pmatrix} -1 & 0 \\ -1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors of K are easily found:

$$\lambda_1 = 0, \quad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \lambda_2 = 1, \quad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \qquad \lambda_3 = 3\,, \quad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

Each positive eigenvalue provides two trigonometric solutions, while the zero eigenvalue

leads to solutions that depend linearly on t. This yields the required six basis solutions:

$$\mathbf{u}_1(t) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{u}_3(t) = \begin{pmatrix} \cos t \\ 0 \\ -\cos t \end{pmatrix}, \qquad \mathbf{u}_5(t) = \begin{pmatrix} \cos \sqrt{3}\, t \\ -2\cos \sqrt{3}\, t \\ \cos \sqrt{3}\, t \end{pmatrix},$$
$$\mathbf{u}_2(t) = \begin{pmatrix} t \\ t \\ t \end{pmatrix}, \qquad \mathbf{u}_4(t) = \begin{pmatrix} \sin t \\ 0 \\ -\sin t \end{pmatrix}, \qquad \mathbf{u}_6(t) = \begin{pmatrix} \cos \sqrt{3}\, t \\ -2\cos \sqrt{3}\, t \\ \cos \sqrt{3}\, t \end{pmatrix}.$$

The first solution $\mathbf{u}_1(t)$ is a constant, equilibrium mode, where the masses rest at a fixed common distance from their reference positions. The second $\mathbf{u}_2(t)$ is the unstable mode, corresponding to a uniform vertical translational motion of the masses without any stretch of the interconnecting springs. The final four solutions represent vibrational modes. In the first pair $\mathbf{u}_3(t), \mathbf{u}_4(t)$, the two outer masses move in opposing directions, while the middle mass remains fixed, while the final pair $\mathbf{u}_5(t), \mathbf{u}_6(t)$ has the two outer masses moving in tandem, while the inner mass moves twice as far in the opposing direction. The general solution is a linear combination of the six normal modes,

$$\mathbf{u}(t) = c_1\, \mathbf{u}_1(t) + \, \cdots \, + c_6\, \mathbf{u}_6(t), \tag{8.98}$$

and corresponds to the molecule moving alon its axis at a fixed speed while the individual masses perform a quasi-periodic vibration.

Let us see if we can predict the motion of the molecule from its initial conditions

$$\mathbf{u}(0) = \mathbf{a}, \qquad \dot{\mathbf{u}}(0) = \boldsymbol{\alpha},$$

where $\mathbf{a} = (\, a, b, c\,)^T$ is the initial displacements of the three atoms, while $\boldsymbol{\alpha} = (\, \alpha, \beta, \gamma\,)^T$ is their initial velocities. Substituting the solution formula (8.98) leads to the linear systems

$$c_1\, \mathbf{v}_1 + c_3\, \mathbf{v}_2 + c_5\, \mathbf{v}_3 = \mathbf{a}, \qquad c_2\, \mathbf{v}_1 + c_4\, \mathbf{v}_2 + \sqrt{3}\, c_6\, \mathbf{v}_3 = \boldsymbol{\alpha},$$

for the coefficients $c_1, \dots, c_6$. Since the eigenvectors of the symmetric matrix K are mutually orthogonal, we can use our orthogonality formula to immediately compute the coefficients:

$$c_1 = \frac{\mathbf{a} \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2} = \frac{a+b+c}{3}, \qquad c_3 = \frac{\mathbf{a} \cdot \mathbf{v}_2}{\| \mathbf{v}_2 \|^2} = \frac{a-c}{2}, \qquad c_5 = \frac{\mathbf{a} \cdot \mathbf{v}_3}{\| \mathbf{v}_3 \|^2} = \frac{a-2b+c}{6},$$
$$c_2 = \frac{\boldsymbol{\alpha} \cdot \mathbf{v}_1}{\| \mathbf{v}_1 \|^2} = \frac{\alpha+\beta+\gamma}{3}, \qquad c_4 = \frac{\boldsymbol{\alpha} \cdot \mathbf{v}_2}{\| \mathbf{v}_2 \|^2} = \frac{\alpha-\gamma}{2}, \qquad c_6 = \frac{\boldsymbol{\alpha} \cdot \mathbf{v}_3}{\sqrt{3}\, \| \mathbf{v}_3 \|^2} = \frac{\alpha-2\beta+\gamma}{6\sqrt{3}}.$$

In particular, the unstable translational mode is excited if and only if its coefficient $c_2 \neq 0$ is non-zero, and this occurs if and only if there is a nonzero net initial velocity of the molecule: $\alpha + \beta + \gamma \neq 0$. In this case the vibrating molecule will move off to ∞ at a uniform velocity $c = c_2 = \frac{1}{3}(\alpha + \beta + \gamma)$ equal to the average of the individual initial velocities. On the other hand, if $\alpha + \beta + \gamma = 0$, then the unstable mode will not be excited and the molecule will vibrate quasiperiodically, with frequencies 1 and $\sqrt{3}$, while sitting at a fixed location.

The observations established in this example hold, in fact, in complete generality. Let us state the result, leaving the details of the proof as an exercise for the reader.

Theorem 8.78. *The solution to unstable second order linear system with positive semi-definite coefficient matrix $K = A^T C\, A$ is a combination of a quasi-periodic vibration and a uniform motion at a fixed velocity in the direction of a null eigenvector $\mathbf{v} \in \ker A$. In particular, the system does not experience any unstable motion, and so will just vibrate around a fixed position, if and only if the initial velocity $\dot{\mathbf{u}}(t_0) \in (\ker K)^\perp = \operatorname{rng} K$ is orthogonal to the subspace $\ker A = \ker K$ of all unstable directions.*

As usual, the unstable modes correspond to either translations or rotations, or to mechanisms of the structure. To prevent a structure from exhibiting an unstable motion, one has to ensure that the initial velocity is orthogonal to all of the unstable directions. The result is in direct analogy with Theorem 6.8 that requires a force to be orthogonal to all such unstable modes in order to maintain equilibrium in the structure.

Systems with Different Masses

When a structure has differing masses at the nodes, the Newtonian equations of motion take the more general form

$$M\,\ddot{\mathbf{u}} = -K\,\mathbf{u}, \qquad \text{or, equivalently,} \qquad \ddot{\mathbf{u}} = -\,M^{-1}K\,\mathbf{u} = -\,P\,\mathbf{u}. \tag{8.99}$$

The mass matrix $M > 0$ is positive definite (and, usually, diagonal, although the general theory does not require this latter restriction), while the stiffness matrix $K = A^T C\, A$ is either positive definite or, in the unstable situation when $\ker A \neq \{\mathbf{0}\}$, positive semi-definite. The coefficient matrix

$$P = M^{-1}K = M^{-1}A^T C\, A \tag{8.100}$$

is *not* in general symmetric, and so we cannot directly apply the preceding constructions. However, P does have the more general self-adjoint form (7.68) based on the weighted inner products

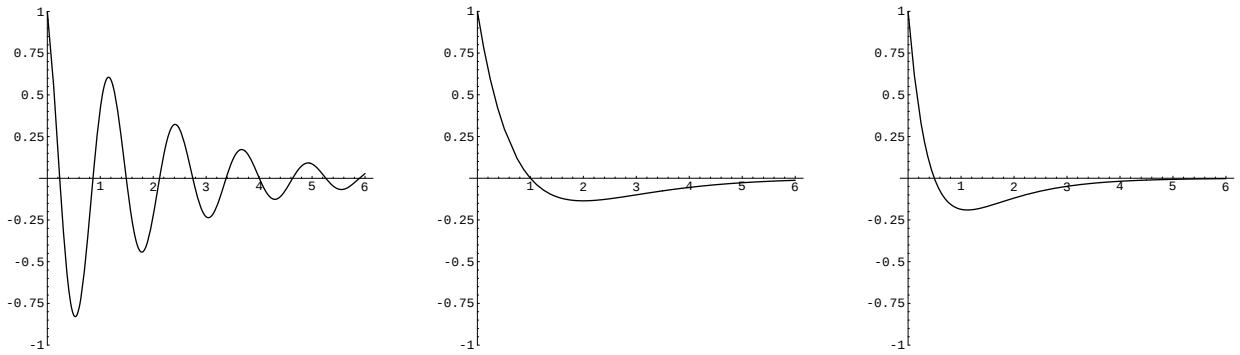
$$\langle\, \mathbf{u}\,;\widetilde{\mathbf{u}}\,\rangle = \mathbf{u}^T M\,\widetilde{\mathbf{u}}, \qquad\qquad \langle\!\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle\!\rangle = \mathbf{v}^T C\,\widetilde{\mathbf{v}}, \tag{8.101}$$

on, respectively, the domain and target spaces for A.

If $\ker A = \{\mathbf{0}\}$, then $P > 0$ is positive definite in the generalized sense of Definition 7.58. In this case, substituting our standard trigonometric solution ansatz $\mathbf{u}(t) = \cos(\omega\, t)\,\mathbf{v}$ into the system results in a *generalized matrix eigenvalue problem*

$$K\,\mathbf{v} = \lambda\, M\,\mathbf{v}, \qquad \text{or, equivalently,} \qquad P\,\mathbf{v} = \lambda\,\mathbf{v}, \qquad \text{with} \qquad \lambda = \omega^2. \tag{8.102}$$

The matrix M plays the role of the identity matrix I in the standard eigenvalue equation (8.17). The proofs for the standard eigenvalue problem are easily modified to handle this situation, and demonstrate that all the eigenvalues are real and non-negative. Moreover the eigenvectors are orthogonal, but now with respect to the weighted inner product $\langle\, \mathbf{u}\,;\widetilde{\mathbf{u}}\,\rangle$ governed by the mass matrix M. Details are relegated to the exercises.



Underdamped Critically Damped Overdamped

Figure 8.12. Damped Vibrations.

Friction and Damping

So far, we have not allowed frictional forces to affect the motion of our dynamical equations. In many physical systems, friction exerts a force on a mass in motion which is proportional to its velocity. In the simplest case of a single mass attached to a spring, one amends the balance of forces in the undamped Newton equation (8.84) to obtain

$$m\,\frac{d^2u}{dt^2} + \beta\,\frac{du}{dt} + k\,u = 0. \tag{8.103}$$

As before, $m > 0$ is the mass, and $k > 0$ the spring system, while $\beta > 0$ measures the effect of a velocity-dependent frictional force — the larger β the greater the frictional damping of the motion.

The solution of this more general second order homogeneous linear ordinary differential equation is found by substituting the usual exponential ansatz $u(t) = e^{\lambda t}$ into the equation, leading to the quadratic characteristic equation

$$m\,\lambda^2 + \beta\,\lambda + k = 0. \tag{8.104}$$

There are three possible cases, illustrated in Figure 8.12:

Underdamped: If $0 < \beta^2 < 4\,m\,k$, then (8.104) has two complex-conjugate roots

$$\lambda \;=\; -\,\frac{\beta}{2\,m} \;\pm\; \mathrm{i}\;\frac{\sqrt{4\,m\,k - \beta^2}}{2\,m} \;\;=\;\; -\,\mu \pm \mathrm{i}\,\nu\,. \tag{8.105}$$

The general solution to the differential equation is

$$u(t) = e^{-\mu\,t}\big(\,c_1\cos\nu\,t + c_2\sin\nu\,t\,\big) = r\,e^{-\mu\,t}\cos(\nu\,t - \delta), \tag{8.106}$$

which represents a damped periodic motion. The time-dependent amplitude of vibration $a(t) = r\,e^{-\mu\,t}$ decays to zero at an exponential rate as $t \to \infty$. The formula for the rate of decay, $\mu = \beta/(2\,m)$, tells us that more friction or less mass will cause the system to return to equilibrium faster. (Of course, mathematically, it never quite gets there, but in

a real physical system after a sufficiently long time the difference is not noticeable.) On the other hand, the frequency of vibration,

$$\nu \;=\; \frac{\sqrt{4\,m\,k - \beta^2}}{2\,m} \;=\; \sqrt{\frac{k}{m} - \frac{\beta^2}{4\,m^2}}\,, \tag{8.107}$$

remains fixed throughout the motion. The frictionally modified vibrational frequency ν is strictly smaller than the undamped frequency $\omega = \sqrt{k/m}$, and hence friction has the effect of slowing down vibrations while progressively diminishing their amplitudes. As the friction approaches a critical threshold, $\beta \nearrow 2\sqrt{m\,k}$, the vibrational frequency goes to zero, $\nu \to 0$, and so the period of vibration $P = 2\pi/\nu$ goes to ∞.

Overdamped: If $\beta^2 > 4\,m\,k$, then the characteristic equation (8.104) has two negative real roots

$$\lambda_1 \;=\; -\,\frac{\beta + \sqrt{\beta^2 - 4\,m\,k}}{2\,m}\,, \qquad\qquad \lambda_2 \;=\; -\,\frac{\beta - \sqrt{\beta^2 - 4\,m\,k}}{2\,m}\,,$$

with $\lambda_1 < \lambda_2 < 0$. The solution

$$u(t) = c_1\, e^{\lambda_1\, t} + c_2\, e^{\lambda_2\, t} \tag{8.108}$$

is a linear combination of two decaying exponentials. An overdamped system models the motion of a mass in a vat of molasses. Its "vibration" is so slow that it can pass at most once through its equilibrium position $u = 0$, and then only when its initial velocity is quite large. In the long term, since $\lambda_1 < \lambda_2$, the first exponential $e^{\lambda_1\, t}$ will decay to zero faster, and hence the overall decay rate of the solution is (unless $c_2 = 0$) governed by the less negative eigenvalue λ_2.

Critically Damped: The borderline case occurs when $\beta^2 = 4\,m\,k$, which means that the characteristic equation (8.104) has only a single negative real root:

$$\lambda_1 = -\,\frac{\beta}{2\,m}\,.$$

In this case, our ansatz only supplies one exponential solution $e^{\lambda_1\, t} = e^{-\beta\, t/2\, m}$. The second linearly independent solution is obtained by multiplication by t, leading to the general solution

$$u(t) = (c_1\, t + c_2) e^{-\beta\, t/2\, m}. \tag{8.109}$$

Even though the formula looks quite different, its qualitative behavior is very similar to the overdamped case. The factor of t plays an unimportant role, since the asymptotics of this solution are almost entirely governed by the decaying exponential function. This represents the nonvibrating solution that has the slowest possible decay rate — reducing the frictional coefficient any further will permit a damped periodic vibration to appear.

In all three cases, provided the frictional coefficient is positive, $\beta > 0$, the zero solution is globally asymptotically stable. Physically, since there is no external forcing, all solutions eventually return to equilibrium as the friction gradually overwhelms any initial motion.

Remark: You may, if you prefer, convert the second order equation (8.103) into a first order system by adopting the phase plane variables u and $v = \dot{u}$. The coefficient matrix of the equivalent phase plane system $\dot{\mathbf{u}} = A\,\mathbf{u}$ is $A = \begin{pmatrix} 0 & 1 \\ -c/m & -b/m \end{pmatrix}$. In terms of our classification of two-dimensional systems, the undamped case corresponds to a center, the underdamped case to a stable focus, the critically damped case to a stable improper node, and the overdamped case to a stable node. The reader should verify that the relevant conditions are met in each case and correlate the phase portraits with the time plots in Figure 8.12.

This concludes our discussion of the scalar case. Similar considerations apply to mass/spring chains, and two and three-dimensional structures. The frictionally damped system has the general form

$$M\,\frac{d^2\mathbf{u}}{dt^2} + B\,\frac{d\mathbf{u}}{dt} + K\,\mathbf{u} = \mathbf{0}, \tag{8.110}$$

where the mass matrix $M > 0$ and the matrix of frictional coefficients $B > 0$ are both diagonal, positive definite, while the stiffness matrix $K = A^T C\,A \geq 0$ is a positive semi-definite Gram matrix constructed from the reduced incidence matrix A. The mathematical details in this case are sufficiently complicated that we shall leave their analysis as an advanced project for the motivated student.

8.11. Forcing and Resonance.

So far, we have allowed our structure to vibrate on its own. It is now time to start applying external forces — to see what happens when we shake it. In this section, we will investigate the effects of periodic forcing on both undamped and damped systems. More general types of forcing can be handled by the variation of parameters method, cf. [**23**].

The simplest case is that of a single mass connected to a spring without any frictional damping. We append an external forcing function $f(t)$ to the homogeneous (unforced) equation (8.84), leading to the inhomogeneous ordinary differential equation

$$m\,\frac{d^2u}{dt^2} + k\,u = f(t), \tag{8.111}$$

in which $m > 0$ is the mass and $k > 0$ the spring stiffness. We are particularly interested in the case of periodic forcing

$$f(t) = \alpha \cos \eta\, t \tag{8.112}$$

of frequency $\eta > 0$ and amplitude α. To find a particular solution to (8.111), (8.112), we use the method of undetermined coefficients[†] which tells us to guess a solution ansatz of the form

$$u_\star(t) = a\,\cos \eta\, t + b\,\sin \eta\, t, \tag{8.113}$$

[†] One can also use variation of parameters, although the intervening calculations are slightly more complicated.

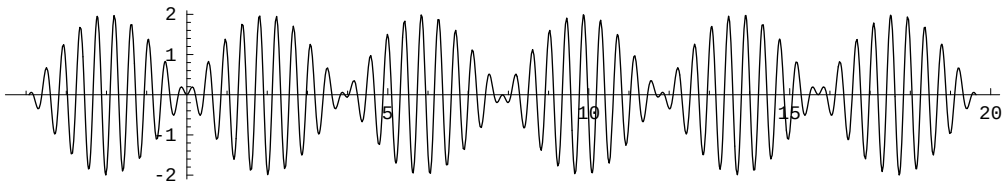


Figure 8.13. Beats in a Periodically Forced Vibration.

where a, b are constant. Substituting this ansatz into the differential equation, we find

$$m\,\frac{d^2u_\star}{dt^2} + k\,u_\star = a\,(k - m\eta^2)\cos\eta\,t + b\,(k - m\eta^2)\sin\eta\,t = \alpha\cos\eta\,t.$$

We can solve for

$$a = \frac{\alpha}{k - m\eta^2} = \frac{\alpha}{m(\omega^2 - \eta^2)}\,, \qquad\qquad b = 0,$$

provided the denominator is nonzero:

$$k - m\eta^2 = m(\omega^2 - \eta^2) \neq 0. \tag{8.114}$$

Here

$$\omega = \sqrt{\frac{k}{m}} \tag{8.115}$$

refers to the natural, unforced vibrational frequency of the system, while η is the forcing frequency. Therefore, provided the forcing frequency is *not* equal to the system's natural frequency, $\eta \neq \omega$, there exists a particular solution

$$u_\star(t) = a\,\cos\eta\,t = \frac{\alpha}{m(\omega^2 - \eta^2)}\,\cos\eta\,t \tag{8.116}$$

that vibrates at the same frequency as the forcing function.

The general solution to the inhomogeneous system (8.111) is found, as usual, by adding in an arbitrary solution to the homogeneous equation, (8.85), yielding

$$u(t) = r\,\cos(\omega\,t - \delta) + a\,\cos\eta\,t, \qquad \text{where} \qquad a = \frac{\alpha}{m(\omega^2 - \eta^2)}\,, \tag{8.117}$$

and where r and δ are determined by the initial conditions. The solution is therefore a quasiperiodic combination of two periodic motions — the first, vibrating with frequency ω, represents the internal or natural vibrations of the system, while the second, with frequency η, represents the response of the system to the periodic forcing. Due to the factor $\omega^2 - \eta^2$ in the denominator of (8.117), the closer the forcing frequency is to the natural frequency, the larger the overall amplitude of the response, and the more likely the spring breaks. displays the graph of

Suppose we start the mass initially at equilibrium, so the initial conditions are

$$u(0) = 0, \qquad\qquad \dot u(0) = 0. \tag{8.118}$$

Substituting the solution formula (8.117) and solving for r, δ, we find that

$$r = -\,a, \qquad\qquad \delta = 0.$$

Thus, the solution to the initial value problem can be written in the form

$$u(t) = a\big(\cos \eta\, t - \cos \omega\, t\big) = 2a \sin\left(\frac{\omega - \eta}{2\, t}\right) \sin\left(\frac{\omega + \eta}{2\, t}\right), \tag{8.119}$$

using a standard trigonometric identity, cf. Exercise ■. The factor $\sin \frac{1}{2}(\omega + \eta)t$ represents a periodic motion whose frequency is the average of the natural and the forcing frequencies. If the forcing frequency η is close to the natural frequency ω, then the initial factor $2a \sin \frac{1}{2}(\omega - \eta)t$ can be viewed as a periodically varying amplitude, whose vibrational frequency $\frac{1}{2}(\omega - \eta)$ is much slower. This factor is responsible for the phenomenon of *beats*, heard, for example, when two tuning forks of close but not exactly equal pitch vibrate near each other. The resulting sound periodically waxes and wanes in intensity. Figure 8.13 displays the graph of the particular function

$$\cos 14\, t - \cos 15.6\, t = 2\, \sin .8\, t\ \sin 14.8\, t.$$

The slowly varying amplitude $2 \sin .8\, t$ is clearly visible as the envelope of the relatively rapid vibrations of frequency 14.8.

If we force the system at exactly the natural frequency $\eta = \omega$, then the trigonometric ansatz (8.113) does not work. This is because both terms are now solutions to the homogeneous equation, and so cannot be combined to form a solution to the inhomogeneous version. In this situation, there is a simple modification to the ansatz, namely multiplication by t, that does the trick. Substituting

$$u_\star(t) = a\, t\, \cos \omega\, t + b\, t\, \sin \omega\, t \tag{8.120}$$

into the differential equation (8.111), we find

$$m\, \frac{d^2 u_\star}{dt^2} + k\, u_\star = -\, 2\, a\, m\, \omega\, \sin \omega\, t + 2\, b\, m\, \omega\, \cos \omega\, t = \alpha \cos \omega\, t,$$

and so

$$a = 0, \qquad\qquad b = \frac{\alpha}{2\, m\, \omega}\,.$$

Combining the resulting particular solution with the solution to the homogeneous equation leads to the general solution

$$u(t) = r \cos(\omega\, t - \delta) + \frac{\alpha}{2\, m\, \omega}\, t\, \sin \omega\, t. \tag{8.121}$$

Both terms vibrate with frequency ω, but the second has a linearly growing amplitude that gets larger and larger as $t \to \infty$; see Figure 8.14. The mass will oscillate more and more wildly until the spring eventually breaks. In this situation, the system is said to be in *resonance*, and the increasingly wild oscillations are provoked by forcing it at the resonant frequency ω.

If we are very close to resonance, the oscillations induced by the particular solution (8.119) will have extremely large, although not unbounded, amplitude a. The lesson is, never force a system at or close to its natural frequency (or frequencies) of vibration. The classic example is the 1940 Tacoma Narrows Bridge disaster, when the vibration in the

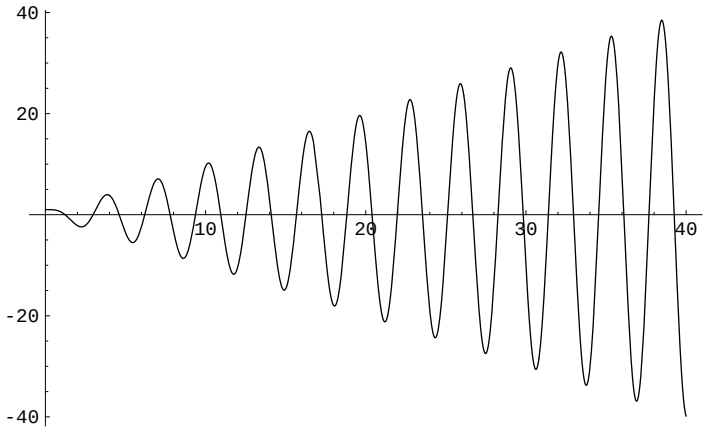


Figure 8.14. Resonance.

bridge caused by a strong wind was close enough to the bridge's natural frequency to cause it to oscillate wildly and collapse! A movie taken at the time is particularly impressive. A second example is the old practice of British (and subsequently, U.S.) infantry who, learning from experience, do not march in unison across a bridge so as not to set off a resonant frequency and cause it to collapse.

If we include frictional effects, then we can partially mollify the wild behavior near the resonant frequency. The frictionally damped vibrations of a mass on a spring, when subject to periodic forcing, can be described by the inhomogeneous version

$$m\,\frac{d^2u}{dt^2} + \beta\,\frac{du}{dt} + k\,u = \alpha\cos\eta\,t \tag{8.122}$$

of equation (8.103). Let us assume that the friction is sufficiently small as to keep us in the underdamped regime $\beta^2 < 4\,m\,k$. Since neither summand solves the homogeneous system, we can use the trigonometric solution ansatz (8.113) to construct the particular solution

$$u_\star(t) = a\cos(\eta\,t - \varepsilon) \qquad \text{where} \qquad a = \frac{\alpha}{\sqrt{m^2(\omega^2-\eta^2)^2 + \beta^2\,\eta^2}} \tag{8.123}$$

represents the amplitude of the response to the periodic forcing, with $\omega = \sqrt{k/m}$ continuing to denote the undamped resonant frequency (8.115), while

$$\varepsilon = \tan^{-1}\frac{\beta\,\omega}{m(\omega^2-\eta^2)} \tag{8.124}$$

represents a *phase lag* in the response of the system that is due to the friction.

The general solution is

$$u(t) = r\,e^{-\mu\,t}\cos(\nu\,t - \delta) + a\cos(\eta\,t - \varepsilon), \tag{8.125}$$

where r, δ are determined by the initial conditions, while $\lambda = \mu + \mathrm{i}\,\nu$ are the roots of the characteristic equation, cf. (8.105). The first term — the solution to the homogeneous equation — is called the *transient* since it decays exponentially fast to zero. Thus, at large times, the internal motion of the system that might have been excited by the initial

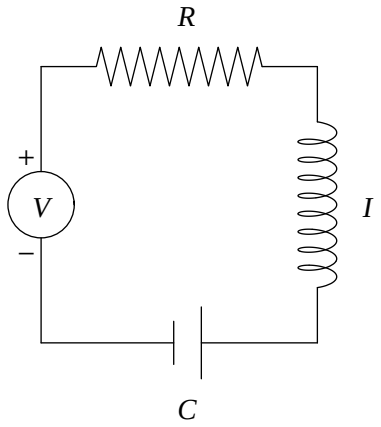


Figure 8.15. The Basic RLC Circuit.

conditions dies out, and only the particular solution (8.123) incited by the forcing persists. The amplitude of the persistent response (8.123) is at a maximum at the resonant frequency $\eta = \omega$, where it takes the value $a_{max} = \alpha/(\beta\,\omega)$. Thus, the smaller the frictional coefficient β (or the slower the resonant frequency ω) the more likely the breakdown of the system due to an overly large response.

Friction also induces a phase shift ε in the response of the system to the external forcing. Speeding up the forcing frequency η increases the overall phase shift, which has the value of $\frac{1}{2}\,\pi$ at the resonant frequency $\eta = \omega$, so the system lags a quarter period behind the forcing, and reaches a maximum $\varepsilon = \pi$ as $\eta \to \infty$. Thus, the response of the system to a high frequency forcing is almost exactly out of phase — the mass is moving downwards when the force is pulling it upwards, and vice versa!

Electrical Circuits

The Electrical–Mechanical Correspondence will continue to operate in the dynamical universe. As we learned in Chapter 6, the equations governing the equilibria of simple electrical circuits and the mechanical systems such as mass/spring chains and structures are modeled by the same basic mathematical structure. In a similar manner, circuits with time-varying currents can also be modeled by linear dynamical systems of ordinary differential equations.

In this section, we analyze the simplest situation of an RLC *circuit* consisting of a *resistor* R, an *inductor* L and a *capacitor* C connected together in a loop as illustrated in Figure 8.15. Let $u(t)$ denote the current in the circuit at time t. As the current passes through each circuit element, it induces a corresponding voltage, which we denote by v_R, v_L and v_C. The voltages are prescribed by the basic laws of electrical circuit design.

(a) First, as we know from Section 6.2, the resistance $R \geq 0$ in the circuit is the proportionality factor between voltage and current, so $v_R = R\,u$.

(b) The voltage passing through an inductor is proportional to the rate of change in the current. Thus, $v_L = L\,\dot{u}$, where $L > 0$ is the *inductance*, and the dot indicates time derivative.

(c) On the other hand, the current passing through a capacitor is proportional to the rate of change in the voltage, and so $u = C\,\dot{v}_C$, where $C > 0$ denotes the *capacitance*. We integrate this relation to produce the capacitor voltage $v_C = \displaystyle\int \frac{u(t)}{C}\, dt$.

The combination of all the induced voltages must equal the externally applied voltage from, say, a battery. The precise rules governing these voltages are:

The voltage balance law tells us that the total of these individual voltages must equal any externally applied voltage coming from, say, a battery or generator. Therefore,

$$v_R + v_L + v_C = v_B,$$

where $v_B = f(t)$ denotes the applied voltage due to a time-varying source. Substituting the preceding formulae, we deduce that the current $u(t)$ in our circuit satisfies following linear integro-differential equation

$$L\,\frac{du}{dt} + R\,u + \int \frac{u}{C}\, dt = f(t). \tag{8.126}$$

We can convert this into a differential equation by differentiating both sides with respect to t. Assuming, for simplicity, that L, R and C are constant, the result is the linear second order ordinary differential equation

$$L\,\frac{d^2u}{dt^2} + R\,\frac{du}{dt} + \frac{1}{C}\,u = f'(t). \tag{8.127}$$

In particular, the homogeneous version, with $f'(t) \equiv 0$, governs the current in an RLC circuit with a constant applied voltage source.

Comparing (8.127) with the equation (8.103) for a mechanically vibrating mass, we see that the analogy bewtween electrical circuits and mechanical structures developed in Chapter 6 continues to hold in the dynamical regime. The current corresponds to the displacement. The inductance plays the role of mass, the resistance corresponds to friction, while the reciprocal $1/C$ of capacitance is analogous to the spring stiffness. Thus, all of our analytical conclusions regarding stability of equilibria, qualitative behavior and formulae for solutions, etc., that we established in the mechanical context can, suitably re-interpreted, be immediately applied to electrical circuit theory.

In particular, an RLC circuit is *underdamped* if $R^2 < 4\,L/C$, and the current $u(t)$ oscillates with frequency

$$\nu = \sqrt{\frac{1}{C\,L} - \frac{R^2}{4\,L^2}}\,, \tag{8.128}$$

while slowly dying off to zero. In the overdamped and critically damped cases $R^2 \geq 4\,L/C$, where the resistance of the circuit is large, the current merely decays to zero exponentially fast and there is no longer any oscillatory behavior in the circuit. Attaching an alternating current source $f(t) = f_0 + a\,\sin \eta\, t$ to the circuit will induce resonance in the case of no resistance if the forcing frequency is equal to the circuits natural internal frequency. Details are relegated to the exercises.

Forcing and Resonance in Systems

Let us very briefly discuss the effect of periodic forcing on a more complicated system. For undamped mass/spring chains, structures and more complicated resistanceless LC circuits, we are led to consider a periodically forced second order system

$$M\,\ddot{\mathbf{u}} + K\,\mathbf{u} = \cos(\omega\,t)\,\mathbf{a}, \tag{8.129}$$

where $\mathbf{a}$ is a constant vector representing both a magnitude and a "direction" of the forcing. Here $M > 0$ is the diagonal mass matrix (or inductance matrix in a circuit), while $K = A^T C\,A$ the (semi-)definite stiffness (or conductance) matrix for the system. We are ignoring friction (resistance) for simplicity. More general periodic and quasiperiodic forcing terms can be built up, via the general inhomogeneous superposition principle of Theorem 7.42, as a linear combination of such simple solutions.

To find a particular solution to the system, let us try the trigonometric ansatz

$$\mathbf{u}^\star(t) = \cos(\omega\,t)\,\mathbf{w} \tag{8.130}$$

where $\mathbf{w}$ is a constant vector. Substituting into (8.129) leads to a linear algebraic system

$$(K - \lambda\,M)\,\mathbf{w} = \mathbf{a}, \qquad \text{where} \qquad \lambda = \omega^2. \tag{8.131}$$

If equation (8.131) has a solution, then our ansatz (8.130) is valid, and we have produced a particular vibration of the system that has the same frequency as the forcing vibration. The general solution, then, will be a quasi-periodic combination of this particular solution coupled with the vibrations at the system's natural, unforced frequencies. In particular, if $\lambda = \omega^2$ is *not* a generalized eigenvalue[†] of the matrix pair K, M, as in (8.102), then the coefficient matrix $K - \lambda\,M$ is nonsingular, and so (8.131) can be solved for any right hand side $\mathbf{a}$.

The more interesting case is when $K - \lambda\,M$ is singular, its kernel being equal to the generalized eigenspace V_λ. In this case, (8.131) will have a solution $\mathbf{w}$ if and only if $\mathbf{a}$ lies in the range of $K - \lambda\,M$. According to the Fredholm Alternative Theorem 5.51, the range is the orthogonal complement of the cokernel, which, since the coefficient matrix is symmetric, is the same as the kernel. Therefore, (8.131) will have a solution if and only if $\mathbf{a}$ is orthogonal to V_λ, i.e., $\mathbf{a} \cdot \mathbf{v} = \mathbf{0}$ for every eigenvector $\mathbf{v}$ for the eigenvalue λ. Thus, one can force a system at a natural frequency without inciting resonance provided the "direction" of forcing, as governed by the vector $\mathbf{a}$, is orthogonal to the natural directions of motion of the system, as governed by the eigenvectors for that particular frequency.

If this orthogonality constraint is not satisfied, then the periodic solution ansatz (8.130) does not apply, and we are in a truly resonant situation. Inspired by the scalar solution, let us try the *resonant ansatz*

$$\mathbf{u}^\star(t) = t\,\sin(\omega\,t)\,\mathbf{y} + \cos(\omega\,t)\,\mathbf{w}. \tag{8.132}$$

[†] When $M = \mathrm{I}$ the system reduces to the standard eigenvalue equation for K.

We compute

$$\frac{d^2\mathbf{u}^\star}{dt^2} = -\,\omega^2\, t\, \sin(\omega t)\, \mathbf{y} + \cos(\omega\, t)\, (2\omega\, \mathbf{y} - \omega^2\, \mathbf{w}).$$

Therefore (8.132) will be a solution to the differential equation (8.129) provided

$$(K - \lambda\, M)\mathbf{y} = \mathbf{0}, \qquad\qquad (K - \lambda\, M)\mathbf{w} = \mathbf{a} - 2\,\omega\, \mathbf{y}.$$

The first equation requires that $\mathbf{y} \in V_\lambda$ be a generalized eigenvector of the matrix pair K, M. The second will be solvable for $\mathbf{w}$ if and only if $\mathbf{a} - 2\,\omega\, \mathbf{y}$ is orthogonal to the eigenspace V_λ, which requires $2\,\omega\, \mathbf{y}$ to be the orthogonal projection of $\mathbf{a}$ onto V_λ. With this choice of $\mathbf{y}$ and $\mathbf{w}$, the basic resonant solution ansatz produces a resonant solution to the system. Summarizing, we find that a generic forcing at a resonant frequency induces resonance in the system.

Theorem 8.79. *An undamped vibrational system will be periodically forced into resonance if and only if the forcing* $\mathbf{f} = \cos(\omega\, t)\, \mathbf{a}$ *is at a natural frequency of the system and the direction of forcing* $\mathbf{a}$ *is not orthogonal to the natural direction(s) of motion of the system for that frequency.*

Chapter 10

Iteration of Linear Systems

Iteration, or repeated application of a function, appears in a surprisingly wide range of applications. Discrete dynamical systems, in which the continuous time variable has been "quantized" in individual units (seconds, days, years, etc.) are modeled by iterative systems. Most numerical solution methods, for both linear and nonlinear systems, are based on an iterative procedure. Starting with an initial guess, the successive iterates lead to closer and closer approximations to the true solution. For linear systems of equations, iterative solution methods can be used as an attractive alternative to Gaussian elimination, and are particularly effective for solving the very large, sparse systems arising in the numerical solution to both ordinary and partial differential equations. In probability theory, population dynamics and other applications, iterative models known as Markov processes govern basic probabilistic processes. All practical methods for computing eigenvalues and eigenvectors are based on a form of iteration.

In this chapter, we concentrate on the iteration of linear systems. As always, proper understanding of the linear situation is an essential prerequisite for tackling the more challenging nonlinear systems, which will be deferred until Chapter 19.Linear iteration coincides with multiplication by successive powers of a matrix. The convergence of the iterates depends on the magnitude of the eigenvalues of the coefficient matrix. The largest eigenvalue (in modulus) is known as the "spectral radius" of the matrix, and convergence requires a spectral radius smaller than one. While accurate computation of the eigenvalues is not an easy task, the simple but effective Gerschgorin Circle Theorem yields useful estimates, that can, in favorable situations, readily ensure convergence. Matrix norms are another practical alternative, since iterative methods with coefficient matrices of norm less than one are guaranteed to converge.

We will then turn our attention to the three most important iterative schemes used to accurately approximate the solutions to linear systems of algebraic equations. The classical Jacobi method is the simplest, while an evident modification leads to the popular Gauss–Seidel method. Completely general conditions ensuring convergence of these schemes to the solution of the original system are hard to formulate, although convergence is assured for the important class of diagonally dominant matrices that arise in many applications. A simple modification of the Gauss–Seidel scheme known as Successive Over-Relaxation (SOR) can dramatically speed up the convergence rate, and is the method of choice in many modern applications.

In the final section we discuss the computation of eigenvalues and eigenvectors of matrices. Needless to say, we completely avoid trying to solve (or even write down) the characteristic polynomial equation. The simple power method and its variants, all based

on linear iteration, provide an effective means of approximating selected eigenvalues. For constructing a complete system of eigenvalues and eigenvectors, the remarkable $Q\,R$ algorithm, which is based of the Gram–Schmidt orthogonalization procedure, is the method of choice, and we shall close with a new proof of its convergence.

10.1. Linear Iterative Systems.

We begin with the basic definition of an iterative system of linear equations.

Definition 10.1. A *linear iterative system* takes the form

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}, \qquad \mathbf{u}^{(0)} = \mathbf{a}. \tag{10.1}$$

The *coefficient matrix* T has size $n \times n$. We will consider both real and complex systems, and so the *iterates* $\mathbf{u}^{(k)}$ are vectors either in $\mathbb{R}^n$ (which assumes that the coefficient matrix T is also real) or in $\mathbb{C}^n$. A linear iterative system can be viewed as a discretized version of a first order system of linear ordinary differential equations, as in (8.9), in which the state of system, as represented by the vector $\mathbf{u}^{(k)}$, changes at discrete time intervals, labeled by the index k. For $k = 1, 2, 3, \ldots$, the solution $\mathbf{u}^{(k)}$ is uniquely determined by the *initial conditions* $\mathbf{u}^{(0)} = \mathbf{a}$.

Scalar Systems

As usual, one begins with an analysis of the scalar version. Consider the iterative equation

$$u^{(k+1)} = \lambda\,u^{(k)}, \qquad u^{(0)} = a. \tag{10.2}$$

The general solution to (10.2) is easily found:

$$u^{(1)} = \lambda\,u^{(0)} = \lambda\,a, \qquad u^{(2)} = \lambda\,u^{(1)} = \lambda^2\,a, \qquad u^{(3)} = \lambda\,u^{(2)} = \lambda^3\,a,$$

and, in general,

$$u^{(k)} = \lambda^k\,a. \tag{10.3}$$

If the initial condition is $a = 0$, then the solution $u^{(k)} \equiv 0$ is constant. Therefore, 0 is a *fixed point* or *equilibrium solution* for the iterative system.

Example 10.2. Banks add interest to a savings account at discrete time intervals. For example, if the bank offers 5% interest compounded yearly, this means that the account balance will increase by 5% each year. Thus, assuming no deposits or withdrawals, the balance $u^{(k)}$ after k years will satisfy the iterative equation (10.2) with $\lambda = 1 + r$ where r is the interest rate, and the 1 indicates that the money in the account remains there. For example, if your initial deposit is $u^{(0)} = a = \$1,000$, after 1 year your account has $u^{(1)} = \$1,050$, after 10 years $u^{(10)} = \$1,628.89$, after 50 years $u^{(50)} = \$11,467.40$, and after 200 years $u^{(200)} = \$17,292,580.82$.

When the compounding is done monthly, the interest rate is still quoted on a yearly basis, and so you receive $\frac{1}{12}$ of the interest compounded each month. If $\widehat{u}^{(k)}$ denotes the balance after k months, then, after n years, the account balance is $\widehat{u}^{(12\,n)} = \left(1 + \frac{1}{12}\,r\right)^{12\,n} a$. Thus, when the interest rate of 5% is compounded monthly, your account balance is

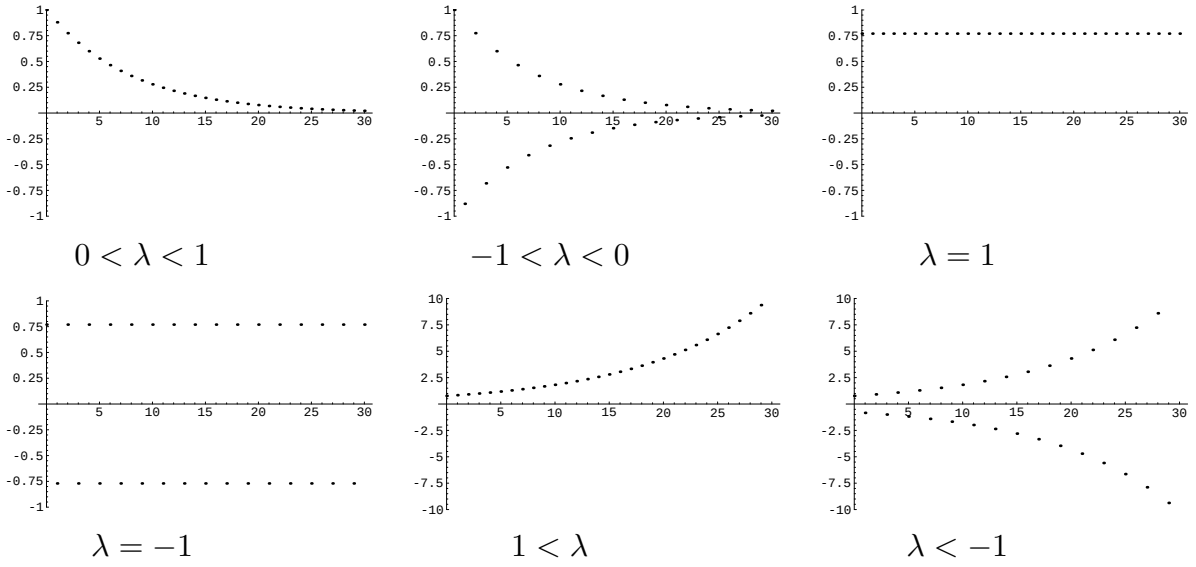


Figure 10.1. One Dimensional Real Linear Iterative Systems..

$\widehat{u}^{(12)} = \$1,051.16$ after 1 year, $\widehat{u}^{(120)} = \$1,647.01$ after 10 years, $\widehat{u}^{(600)} = \$12,119.38$ after 50 years, and $\widehat{u}^{(2400)} = \$21,573,572.66$ dollars after 200 years. So, if you wait sufficiently long, compounding has a dramatic effect. Daily compounding replaces 12 by 365.25, the number of days in a year.

Let us analyze the solutions of iterative equations when $\lambda \in \mathbb{R}$ is a real constant. Apart from the equilibrium solution, the iterates exhibit five qualitatively different behaviors, depending on the size of the coefficient λ.

(*a*) If $\lambda = 0$, the solution immediately becomes zero, and stays there, so $u^{(k)} = 0$ for all $k \geq 1$.

(*b*) If $0 < \lambda < 1$, then the solution is of one sign, and tends monotonically to zero, so $u^{(k)} \to 0$ as $k \to \infty$.

(*c*) If $-1 < \lambda < 0$, then the solution tends to zero, $u^{(k)} \to 0$ as $k \to \infty$. Successive iterates have alternating signs.

(*d*) If $\lambda = 1$, the solution is constant, $u^{(k)} = a$, for all $k \geq 0$.

(*e*) If $\lambda = -1$, the solution switches back and forth between two values; $u^{(k)} = (-1)^k\, a$.

(*f*) If $1 < \lambda < \infty$, then the iterates $u^{(k)}$ become unbounded. If $a > 0$, they go monotonically to $+\infty$; if $a < 0$, to $-\infty$.

(*g*) If $-\infty < \lambda < -1$, then the iterates $u^{(k)}$ also become unbounded. Successive iterates have alternating signs.

In Figure 10.1 we exhibit representative *scatter plots* for the nontrivial cases (b – g). The horizontal axis is the index k and the vertical axis the solution value u.

To describe the different scenarios, we adopt a terminology that already appeared in the continuous realm. In the first three cases, the fixed point $u = 0$ is said to be *globally asymptotically stable* since all solutions tend to 0 as $k \to \infty$. In cases (d) and (e), the zero solution is *stable*, since solutions with nearby initial data, $|\, a \,| \ll 1$, remain nearby. In the final two cases, the zero solution is *unstable*; any nonzero initial data $a \neq 0$ — no

matter how small — will give rise to a solution that eventually goes arbitrarily far away from equilibrium.

Let us next consider the case of a complex scalar iterative system. The coefficient λ and the initial data a in (10.2) are allowed to be complex numbers. The solution is the same, (10.3), but now we need to know what happens when we raise a complex number λ to a high power. The secret is to write $\lambda = r\, e^{\,\mathrm{i}\,\theta}$ in polar form (3.77), where $r = |\,\lambda\,|$ is its modulus and $\theta = \mathrm{ph}\,\lambda$ its angle or phase. Then $\lambda^k = r^k\, e^{\,\mathrm{i}\,k\,\theta}$. Since $|\, e^{\,\mathrm{i}\,k\,\theta}\,| = 1$, we have $|\,\lambda^k\,| = |\,\lambda\,|^k$, and so the solutions (10.3) have modulus $|\,u^{(k)}\,| = |\,\lambda^k\, a\,| = |\,\lambda\,|^k\,|\,a\,|$. As a result, $u^{(k)}$ will remain bounded if and only if $|\,\lambda\,| \le 1$, and will tend to zero as $k \to \infty$ if and only if $|\,\lambda\,| < 1$.

We have thus established the basic stability criteria for scalar, linear systems.

Theorem 10.3. *The zero solution to a (real or complex) scalar iterative system* $u^{(k+1)} = \lambda\, u^{(k)}$ *is*

(*a*) asymptotically stable *if and only if* $|\,\lambda\,| < 1$,

(*b*) stable *if and only if* $|\,\lambda\,| \le 1$,

(*c*) unstable *if and only if* $|\,\lambda\,| > 1$.

Powers of Matrices

The solution to the general linear matrix iterative system

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}, \qquad \mathbf{u}^{(0)} = \mathbf{a}, \tag{10.4}$$

is also, at least at first glance, immediate. Clearly,

$$\mathbf{u}^{(1)} = T\,\mathbf{u}^{(0)} = T\,\mathbf{a}, \qquad \mathbf{u}^{(2)} = T\,\mathbf{u}^{(1)} = T^2\mathbf{a}, \qquad \mathbf{u}^{(3)} = T\,\mathbf{u}^{(2)} = T^3\mathbf{a},$$

and, in general,

$$\mathbf{u}^{(k)} = T^k\mathbf{a}. \tag{10.5}$$

Thus, the iterates are simply determined by multiplying the initial vector $\mathbf{a}$ by the successive powers of the coefficient matrix T.

However, unlike real or complex scalars, the general formulae and qualitative behavior of the powers of a square matrix are not nearly so immediately apparent. (Before continuing, the reader is urged to experiment with simple 2×2 matrices, and try to detect patterns.) To resolve this dilemma, recall that we managed to solve linear systems of differential equations by suitably adapting the known exponential solution from the scalar version. In the discrete case, we no longer have exponentials, but rather powers, in our scalar solution formula (10.3). This motivates us to try the power ansatz

$$\mathbf{u}^{(k)} = \lambda^k\,\mathbf{v}, \tag{10.6}$$

where λ is a scalar and $\mathbf{v}$ is a fixed vector, as a possible solution. We find

$$\mathbf{u}^{(k+1)} = \lambda^{k+1}\,\mathbf{v}, \qquad \text{while} \qquad T\,\mathbf{u}^{(k)} = T(\lambda^k\,\mathbf{v}) = \lambda^k\,T\,\mathbf{v}.$$

These two expressions will be equal if and only if

$$T\,\mathbf{v} = \lambda\,\mathbf{v}.$$

Therefore, (10.6) is a nontrivial solution to (10.4) if and only if λ is an *eigenvalue* of T and $\mathbf{v}$ an associated *eigenvector*.

Thus, to each eigenvector and eigenvalue of the coefficient matrix, we can construct a solution to the iterative system. We can then use linear superposition, as in Theorem 7.29, to combine the basic power solutions to form more general solutions. In particular, if the coefficient matrix is complete, then this method will, as in the case of linear ordinary differential equations, produce the general solution.

Theorem 10.4. *If the coefficient matrix T is complete, then the general solution to the linear iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$ is given by*

$$\mathbf{u}^{(k)} = c_1\,\lambda_1^k\,\mathbf{v}_1 + c_2\,\lambda_2^k\,\mathbf{v}_2 + \,\cdots\, + c_n\,\lambda_n^k\,\mathbf{v}_n, \tag{10.7}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the linearly independent eigenvectors and $\lambda_1, \ldots, \lambda_n$ the corresponding eigenvalues of T. The coefficients $c_1, \ldots, c_n$ are arbitrary scalars, and are uniquely prescribed by the initial conditions $\mathbf{u}^{(0)} = \mathbf{a}$.

Proof: Since we already know that (10.7) is a solution to the system for arbitrary $c_1, \ldots, c_n$, it suffices to show that we can match any prescribed initial conditions. We need to solve the linear system

$$\mathbf{u}^{(0)} = c_1\,\mathbf{v}_1 + \,\cdots\, + c_n\,\mathbf{v}_n = \mathbf{a}. \tag{10.8}$$

Completeness of T implies that its eigenvectors form a basis of $\mathbb{C}^n$, and hence (10.8) always admits a solution. In matrix form, we can rewrite (10.8) as

$$S\,\mathbf{c} = \mathbf{a}, \qquad \text{so that} \qquad \mathbf{c} = S^{-1}\mathbf{a},$$

where $S = (\,\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_n\,)$ is the (nonsingular) matrix whose columns are the eigenvectors. *Q.E.D.*

Remark: Incomplete cases rely on the Jordan canonical form of Section 8.6; see Exercise ∎ for details.

Example 10.5. Consider the iterative system

$$x^{(k+1)} = \tfrac{3}{10}\,x^{(k)} + \tfrac{1}{10}\,y^{(k)}, \qquad y^{(k+1)} = \tfrac{1}{10}\,x^{(k)} + \tfrac{3}{10}\,y^{(k)}, \tag{10.9}$$

with initial conditions

$$x^{(0)} = a, \qquad y^{(0)} = b. \tag{10.10}$$

The system can be rewritten in our matrix form (10.4) with

$$T = \begin{pmatrix} .3 & .1 \\ .1 & .3 \end{pmatrix}, \qquad \mathbf{u}^{(k)} = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}, \qquad \mathbf{a} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Solving the characteristic equation

$$\det(T - \lambda\, \mathrm{I}\,) = \lambda^2 - .6\lambda - .08 = 0$$

produces the eigenvalues $\lambda_1 = .4, \lambda_2 = .2$. We then solve the associated linear systems $(T - \lambda_j\, \mathrm{I}\,)\mathbf{v}_j = \mathbf{0}$ for the corresponding eigenvectors:

$$\lambda_1 = .4, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = .2, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Therefore, the basic power solutions are

$$\mathbf{u}_1^{(k)} = (.4)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \mathbf{u}_2^{(k)} = (.2)^k \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Theorem 10.4 tells us that the general solution is given as a linear combination,

$$\mathbf{u}^{(k)} = c_1\, \mathbf{u}_1^{(k)} + c_2\, \mathbf{u}_2^{(k)} = c_1\, (.4)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\, (.2)^k \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1\, (.4)^k - c_2\, (.2)^k \\ c_1\, (.4)^k + c_2\, (.2)^k \end{pmatrix},$$

where c_1, c_2 are arbitrary scalars, whose values are determined by the initial conditions:

$$\mathbf{u}^{(0)} = \begin{pmatrix} c_1 - c_2 \\ c_1 + c_2 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \text{and hence} \qquad c_1 = \frac{a+b}{2}\,, \qquad c_2 = \frac{b-a}{2}\,.$$

Therefore, the explicit formula for the solution to (10.9), (10.10) is

$$x^{(k)} = (.4)^k\, \frac{a+b}{2} - (.2)^k\, \frac{b-a}{2}, \qquad\qquad y^{(k)} = (.4)^k\, \frac{a+b}{2} + (.2)^k\, \frac{b-a}{2}.$$

In particular, as $k \to \infty$, the iterates $\mathbf{u}^{(k)} \to \mathbf{0}$ converge to zero at a rate governed by the larger eigenvalue $\lambda_1 = .4$. Thus, (10.9) defines a stable iterative system.

Example 10.6. The *Fibonacci numbers* are defined by the second order iterative scheme

$$u^{(k+2)} = u^{(k+1)} + u^{(k)}, \tag{10.11}$$

with initial conditions

$$u^{(0)} = a, \qquad u^{(1)} = b. \tag{10.12}$$

The classical Fibonacci integers follow from $a = 0$, $b = 1$. Thus, to obtain the next Fibonacci number, we add the previous two; the first few Fibonacci integers are

$$u^{(0)} = 0,\ \ u^{(1)} = 1,\ \ u^{(2)} = 1,\ \ u^{(3)} = 2,\ \ u^{(4)} = 3,\ \ u^{(5)} = 5,\ \ u^{(6)} = 8,\ \ u^{(7)} = 13,\ \ \ldots.$$

The Fibonacci integers occur in a surprising range of natural objects, including leaves, flowers, and fruit, [**11**]. They were originally introduced by the Renaissance mathematician Fibonacci (Leonardo of Pisa) as a crude model of the growth of a population of rabbits. In Fibonacci's model, the k^{th} Fibonacci number $u^{(k)}$ measures the total number of pairs of rabbits at year k. We start the process with a single juvenile pair† at year 0. Once a

† We ignore important details like the sex of the offspring.

year, each pair of rabbits produces a new pair of offspring, but it takes a year for a rabbit pair to mature enough to produce offspring of their own.

Just as every higher order ordinary differential equation can be replaced by an equivalent first order system, so every higher order iterative system can be replaced by a first order iterative system. In this particular case, we define the vector

$$\mathbf{u}^{(k)} = \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix} \in \mathbb{R}^2,$$

and note that (10.11) is equivalent to the matrix system

$$\begin{pmatrix} u^{(k+1)} \\ u^{(k+2)} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix}, \quad \text{or} \quad \mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}, \quad \text{where} \quad T = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

To find an explicit formula for the Fibonacci numbers, we need to determine the eigenvalues and eigenvectors of the coefficient matrix T. A straightforward computation produces

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \;=\; 1.618034\ldots, \qquad \lambda_2 = \frac{1-\sqrt{5}}{2} \;=\; -.618034\ldots,$$
$$\mathbf{v}_1 = \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix}.$$

Therefore, according to (10.7), the general solution to the Fibonacci system is

$$\mathbf{u}^{(k)} = \begin{pmatrix} u^{(k+1)} \\ u^{(k)} \end{pmatrix} = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^k \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix} + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^k \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix}. \tag{10.13}$$

The initial data

$$\mathbf{u}^{(0)} = c_1 \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

uniquely specifies the coefficients

$$c_1 = \frac{2\,a + (1+\sqrt{5})\,b}{2\sqrt{5}}, \qquad c_2 = -\,\frac{2\,a + (1-\sqrt{5})\,b}{2\sqrt{5}}.$$

The first entry of the solution vector (10.13) produces the formula

$$u^{(k)} \;=\; \frac{(-1+\sqrt{5})\,a + 2\,b}{2\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^k \;+\; \frac{(1+\sqrt{5})\,a - 2\,b}{2\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^k \tag{10.14}$$

for the k^{th} Fibonacci number. For the particular initial conditions $a = 0$, $b = 1$, formula (10.14) reduces to the classical *Binet formula*

$$u^{(k)} = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k \right] \tag{10.15}$$

for the k^{th} Fibonacci integer. It is a remarkable fact that, for every value of k, all the $\sqrt{5}$'s cancel out, and the Binet formula does indeed produce the Fibonacci integers tabulated above. Another useful observation is that since

$$0 < |\,\lambda_2\,| = \frac{\sqrt{5}-1}{2} < 1 < \lambda_1 = \frac{1+\sqrt{5}}{2}\,,$$

the terms involving λ_1^k go to ∞ (and so the zero solution to this iterative system is unstable) while the terms involving λ_2^k go to zero. Therefore, even for k moderately large, the first term in (10.14) is an excellent approximation (and one that gets more and more accurate with increasing k) to the k^{th} Fibonacci number.

The dominant eigenvalue $\lambda_1 = \frac{1}{2}\big(\,1+\sqrt{5}\,\big) = 1.618034\ldots \equiv \phi$ is known as the *golden ratio* and plays an important role in spiral growth in nature, as well as in art, architecture and design, [**11**]. It describes the overall growth rate of the Fibonacci integers, and, in fact, any sequence of Fibonacci numbers with initial conditions $b \neq \frac{1}{2}\big(\,1-\sqrt{5}\,\big)\,a$.

Example 10.7. Let $T = \begin{pmatrix} -3 & 1 & 6 \\ 1 & -1 & -2 \\ -1 & -1 & 0 \end{pmatrix}$ be the coefficient matrix for a three-dimensional iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$. The eigenvalues and corresponding eigenvectors are

$$\lambda_1 = -2, \qquad \mathbf{v}_1 = \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix}, \qquad \lambda_2 = -1+\mathrm{i}\,, \qquad \mathbf{v}_2 = \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix}, \qquad \lambda_3 = -1-\mathrm{i}\,, \qquad \mathbf{v}_3 = \begin{pmatrix} 2+\mathrm{i} \\ -1 \\ 1 \end{pmatrix}.$$

Therefore, according to (10.7), the general complex solution to the iterative system is

$$\mathbf{u}^{(k)} = b_1\,(-2)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + b_2\,(-1+\mathrm{i}\,)^k \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix} + b_3\,(-1-\mathrm{i}\,)^k \begin{pmatrix} 2+\mathrm{i} \\ -1 \\ 1 \end{pmatrix},$$

where b_1, b_2, b_3 are arbitrary complex scalars.

If we are only interested in real solutions, we can, as in the case of systems of differential equations, break up any complex solution into its real and imaginary parts, each of which constitutes a real solution. (This is another manifestation of the general Reality Theorem 7.47, but is not hard to prove directly.) We begin by writing $\lambda_2 = -1+\mathrm{i} = \sqrt{2}\,e^{3\pi\,\mathrm{i}\,/4}$, and hence

$$(-1+\mathrm{i}\,)^k = 2^{k/2}\,e^{3k\pi\,\mathrm{i}\,/4} = 2^{k/2}\,\big(\cos\tfrac{3}{4}\,k\pi + \mathrm{i}\,\sin\tfrac{3}{4}\,k\pi\,\big)\,.$$

Therefore, the complex solution

$$(-1+\mathrm{i}\,)^k \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix} = 2^{k/2} \begin{pmatrix} 2\cos\frac{3}{4}\,k\pi + \sin\frac{3}{4}\,k\pi \\ -\cos\frac{3}{4}\,k\pi \\ \cos\frac{3}{4}\,k\pi \end{pmatrix} + \mathrm{i}\;2^{k/2} \begin{pmatrix} 2\sin\frac{3}{4}\,k\pi - \cos\frac{3}{4}\,k\pi \\ -\sin\frac{3}{4}\,k\pi \\ \sin\frac{3}{4}\,k\pi \end{pmatrix}$$

is a complex combination of two independent real solutions. The complex conjugate eigenvalue $\lambda_3 = -1 - \mathrm{i}$ leads, as before, to the complex conjugate solution — and the same two real solutions. The general real solution $\mathbf{u}^{(k)}$ to the system can be written as a linear combination of the three independent real solutions:

$$c_1\,(-2)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + c_2\,2^{k/2} \begin{pmatrix} 2\cos\frac{3}{4}\,k\,\pi + \sin\frac{3}{4}\,k\,\pi \\ -\cos\frac{3}{4}\,k\,\pi \\ \cos\frac{3}{4}\,k\,\pi \end{pmatrix} + c_3\,2^{k/2} \begin{pmatrix} 2\sin\frac{3}{4}\,k\,\pi - \cos\frac{3}{4}\,k\,\pi \\ -\sin\frac{3}{4}\,k\,\pi \\ \sin\frac{3}{4}\,k\,\pi \end{pmatrix}, \tag{10.16}$$

where c_1, c_2, c_3 are arbitrary real scalars, uniquely prescribed by the initial conditions.

10.2. Stability.

With the solution formula (10.7) in hand, we are now in a position to understand the qualitative behavior of solutions to (complete) linear iterative systems. The most important case for applications is when all the iterates converge to $\mathbf{0}$.

Definition 10.8. The equilibrium solution $\mathbf{u}^\star = \mathbf{0}$ to a linear iterative system (10.1) is called *asymptotically stable* if and only if all solutions $\mathbf{u}^{(k)} \to \mathbf{0}$ as $k \to \infty$.

Stability of the solutions to an iterative system relies on the following property of the coefficient matrix.

Definition 10.9. A matrix T is called *convergent* if its powers $T^k \to \mathrm{O}$ converge to the zero matrix as $k \to \infty$.

We note that convergence of a sequence of matrices or vectors is equivalent to convergence of their individual entries. The equivalence of the convergence condition and stability of the iterative system follows immediately from the solution formula (10.5).

Proposition 10.10. *The linear iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$ has asymptotically stable zero solution if and only if T is a convergent matrix.*

Indeed, since $\mathbf{u}^{(k)} = T^k\,\mathbf{a}$ and the initial condition $\mathbf{a}$ is arbitrary, the only way that all solutions tend to zero as $k \to \infty$ is if the coefficient matrices $T^k \to \mathrm{O}$.

For the analysis of convergence, we shall adopt a norm $\|\cdot\|$ on our underlying vector space, $\mathbb{R}^n$ or $\mathbb{C}^n$. The reader may be inclined to choose the Euclidean (or Hermitian) norm, but, in practice, the L^∞ norm

$$\|\,\mathbf{u}\,\|_\infty = \max\big\{\,|\,u_1\,|,\ \ldots\ ,|\,u_n\,|\,\big\} \tag{10.17}$$

prescribed by the vector's maximal entry (in modulus) is usually much easier to work with. Convergence of the iterates is equivalent to convergence of their norms:

$$\mathbf{u}^{(k)} \to \mathbf{0} \qquad \text{if and only if} \qquad \|\,\mathbf{u}^{(k)}\,\| \to \mathbf{0} \qquad \text{as} \qquad k \to \infty.$$

(See also Section 12.5 for additional details on convergence in finite-dimensional vector spaces.)

The fundamental stability criterion relies on the magnitude of the eigenvalues of the coefficient matrix.

Theorem 10.11. *A linear iterative system* (10.1) *has asymptotically stable zero solution if and only if all its* (*complex*) *eigenvalues have modulus strictly less than one*: $|\lambda_j| < 1$.

Proof: Let us prove this result assuming that the coefficient matrix T is complete. (The proof in the incomplete case relies on the Jordan canonical form, and is outlined in the exercises.) If λ_j is an eigenvalue such that $|\lambda_j| < 1$, then the corresponding basis solution $\mathbf{u}_j^{(k)} = \lambda_j^k \mathbf{v}_j$ tends to zero as $k \to \infty$; indeed,

$$\| \mathbf{u}_j^{(k)} \| = \| \lambda_j^k \mathbf{v}_j \| = |\lambda_j|^k \| \mathbf{v}_j \| \longrightarrow 0 \qquad \text{since} \qquad |\lambda_j| < 1.$$

Therefore, if all eigenvalues are less than 1 in modulus, all terms in the solution formula (10.7) tend to zero, which proves asymptotic stability: $\mathbf{u}^{(k)} \to \mathbf{0}$. *Q.E.D.*

Consequently, the necessary and sufficient condition for asymptotic stability of a linear iterative system is that all the eigenvalues of the coefficient matrix lie strictly inside the unit circle in the complex plane[†]: $|\lambda_j| < 1$. Let us formalize this key result.

Definition 10.12. The *spectral radius* of a matrix T is defined as the maximal modulus of all of its real and complex eigenvalues: $\rho(T) = \max\{ |\lambda_1|, \ldots, |\lambda_k| \}$.

We can restate the Stability Theorem 10.11 as follows.

Theorem 10.13. *The matrix T is convergent if and only if its spectral radius is strictly less than one: $\rho(T) < 1$.*

If T is complete, then we can apply the triangle inequality to (10.7) to estimate

$$\begin{aligned} \| \mathbf{u}^{(k)} \| &= \| c_1 \lambda_1^k \mathbf{v}_1 + \cdots + c_n \lambda_n^k \mathbf{v}_n \| \\ &\le |\lambda_1|^k \| c_1 \mathbf{v}_1 \| + \cdots + |\lambda_n|^k \| c_n \mathbf{v}_n \| \\ &\le \rho(T)^k \big(|c_1| \| \mathbf{v}_1 \| + \cdots + |c_n| \| \mathbf{v}_n \| \big) = C\, \rho(T)^k, \end{aligned} \tag{10.18}$$

for some constant $C > 0$ that depends only upon the initial conditions. In particular, if $\rho(T) < 1$, then

$$\| \mathbf{u}^{(k)} \| \le C\, \rho(T)^k \longrightarrow 0 \qquad \text{as} \qquad k \to \infty, \tag{10.19}$$

in accordance with Theorem 10.13. Thus, the spectral radius prescribes the rate of convergence of the solutions to equilibrium. The smaller the spectral radius, the faster the solutions converge to $\mathbf{0}$.

If T has only one largest (simple) eigenvalue, so $|\lambda_1| > |\lambda_j|$ for all $j > 1$, then the first term in the solution formula (10.7) will eventually dominate all the others: $\| \lambda_1^k \mathbf{v}_1 \| \gg \| \lambda_j^k \mathbf{v}_j \|$ for $j > 1$ and $k \gg 0$ large. Therefore, provided $c_1 \neq 0$, the solution (10.7) has the asymptotic formula

$$\mathbf{u}^{(k)} \approx c_1 \lambda_1^k \mathbf{v}_1, \tag{10.20}$$

[†] Note that this is *not* the same as the stability criterion for ordinary differential equations, which requires the eigenvalues of the coefficient matrix to lie in the left half plane.

and so most solutions end up parallel to the dominant eigenvector $\mathbf{v}_1$. In particular, if $|\lambda_1| = \rho(T) < 1$, such a solution approaches $\mathbf{0}$ along the direction of the dominant eigenvector $\mathbf{v}_1$ at a rate governed by the modulus of the dominant eigenvalue.

The exceptional solutions, with $c_1 = 0$, tend to $\mathbf{0}$ at a faster rate, along one of the other eigendirections. However, in practical computations, one rarely observes the exceptional solutions. Indeed, even if one begins with initial conditions for which there is no dominant eigenvector component, round off error will almost inevitably introduce a small component in the direction of $\mathbf{v}_1$, which will, if you wait long enough, eventually dominate the computation.

Remark: The inequality (10.18) only applies to complete matrices. In the general case, one can prove that the solution satisfies the slightly weaker inequality

$$\| \mathbf{u}^{(k)} \| \leq C\, \sigma^k \qquad \text{for all} \qquad k \geq 0, \qquad \text{where} \qquad \sigma > \rho(T) \tag{10.21}$$

is any number larger than the spectral radius, while $C > 0$ is a positive constant (that may depend on how close σ is to ρ).

Example 10.14. According to Example 10.7, the matrix

$$T = \begin{pmatrix} -3 & 1 & 6 \\ 1 & -1 & -2 \\ -1 & -1 & 0 \end{pmatrix} \qquad \text{has eigenvalues} \qquad \begin{aligned} \lambda_1 &= -2, \\ \lambda_2 &= -1 + \mathrm{i}\,, \\ \lambda_3 &= -1 - \mathrm{i}\,. \end{aligned}$$

Since $|\lambda_1| = 2 > |\lambda_2| = |\lambda_3| = \sqrt{2}$, the spectral radius is $\rho(T) = |\lambda_1| = 2$. We conclude that T is not a convergent matrix. As the reader can check, either directly, or from the solution formula (10.16), the vectors $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)}$ obtained by repeatedly multiplying any nonzero initial vector $\mathbf{u}^{(0)}$ by T rapidly go off to ∞, at a rate roughly equal to $\rho(T)^k = 2^k$.

On the other hand, the matrix

$$\widetilde{T} = -\tfrac{1}{3}\, T = \begin{pmatrix} 1 & -\frac{1}{3} & -2 \\ -\frac{1}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix} \qquad \text{with eigenvalues} \qquad \begin{aligned} \lambda_1 &= \tfrac{2}{3}, \\ \lambda_2 &= \tfrac{1}{3}(1 - \mathrm{i}\,), \\ \lambda_3 &= \tfrac{1}{3}(1 + \mathrm{i}\,), \end{aligned}$$

has spectral radius $\rho(\widetilde{T}) = \frac{2}{3}$, and hence is a convergent matrix. According to (10.20), if we write the initial data $\mathbf{u}^{(0)} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3$ as a linear combination of the eigenvectors, then, provided $c_1 \neq 0$, the iterates have the asymptotic form $\mathbf{u}^{(k)} \approx c_1 \left(-\frac{2}{3}\right)^k \mathbf{v}_1$, where $\mathbf{v}_1 = (\,4, -2, 1\,)^T$ is the eigenvector corresponding to the dominant eigenvalue $\lambda_1 = -\frac{2}{3}$. Thus, for most initial vectors, the iterates end up decreasing in length by a factor of almost exactly $\frac{2}{3}$ and become eventually parallel to the dominant eigenvector. This is borne out by a sample computation; starting with $\mathbf{u}^{(0)} = (\,1, 1, 1\,)^T$, the first ten iterates are

$$\begin{pmatrix} -0.0936 \\ 0.0462 \\ -0.0231 \end{pmatrix}, \quad \begin{pmatrix} -0.0627 \\ 0.0312 \\ -0.0158 \end{pmatrix}, \quad \begin{pmatrix} -0.0416 \\ 0.0208 \\ -0.0105 \end{pmatrix}, \quad \begin{pmatrix} -0.0275 \\ 0.0138 \\ -0.0069 \end{pmatrix}, \quad \begin{pmatrix} -0.0182 \\ 0.0091 \\ -0.0046 \end{pmatrix},$$

$$\begin{pmatrix} -0.0121 \\ 0.0061 \\ -0.0030 \end{pmatrix}, \quad \begin{pmatrix} -0.0081 \\ 0.0040 \\ -0.0020 \end{pmatrix}, \quad \begin{pmatrix} -0.0054 \\ 0.0027 \\ -0.0013 \end{pmatrix}, \quad \begin{pmatrix} -0.0036 \\ 0.0018 \\ -0.0009 \end{pmatrix}, \quad \begin{pmatrix} -0.0024 \\ 0.0012 \\ -0.0006 \end{pmatrix},$$

Fixed Points

The zero vector $\mathbf{0}$ is always a fixed point for a linear iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$. Are there any others? The answer is immediate: $\mathbf{u}^\star$ is a fixed point if and only if $\mathbf{u}^\star = T\,\mathbf{u}^\star$, and hence any nonzero $\mathbf{u}^\star$ must be an eigenvector of T with eigenvalue 1. Thus, the system has a nonzero fixed point if and only if the coefficient matrix T has 1 as an eigenvalue. Since any scalar multiple of the eigenvector $\mathbf{u}^\star$ is also an eigenvector, in such cases the system admits infinitely many fixed points.

The stability of such fixed points, at least if the coefficient matrix is complete, is governed by the same solution formula (10.7). If the eigenvalue $\lambda_1 = 1$ is simple, and all other eigenvalues are less than one in modulus, so $|\,\lambda_2\,|, \dots, |\,\lambda_n\,| < 1$, then the solution takes the asymptotic form

$$\mathbf{u}^{(k)} = c_1\,\mathbf{v}_1 + c_2\,\lambda_2^k\,\mathbf{v}_2 + \,\cdots\, + c_n\,\lambda_1^k\,\mathbf{v}_n \;\longrightarrow\; c_1\,\mathbf{v}_1, \qquad \text{as} \qquad k \longrightarrow \infty, \tag{10.22}$$

converging to one of the fixed points, i.e., a multiple of the eigenvector $\mathbf{v}_1$. The actual multiple c_1 is determined by the initial conditions, as in (10.8). The rate of convergence is governed by the modulus $|\,\lambda_2\,|$ of the *subdominant eigenvalue*.

The general convergence result governing the stability of fixed points for general coefficient matrices follows.

Theorem 10.15. *Suppose that T has a simple (or, more generally, complete) eigenvalue $\lambda_1 = 1$, and, moreover, all other eigenvalues satisfy $|\,\lambda_j\,| < 1$, for $j \geq 2$. Then all solutions to the linear iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$ converge to a vector $\mathbf{v} \in V_1$ in the eigenspace for the eigenvalue $\lambda_1 = 1$.*

Remark: If $\lambda = 1$ is an incomplete eigenvalue, then the solutions do not, in general, converge.

Example 10.16. For the matrix $T = \begin{pmatrix} \frac{3}{2} & -\frac{1}{2} & -3 \\ -\frac{1}{2} & \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$, the eigenvalues and corresponding eigenvectors are

$$\lambda_1 = 1, \qquad \lambda_2 = \frac{1+\mathrm{i}}{2}, \qquad \lambda_3 = \frac{1-\mathrm{i}}{2},$$
$$\mathbf{v}_1 = \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 2+\mathrm{i} \\ -1 \\ 1 \end{pmatrix}.$$

Since $\lambda_1 = 1$, any multiple of the eigenvector $\mathbf{v}_1$ is a fixed point. The fixed points are stable since the remaining eigenvalues have modulus $|\,\lambda_2\,| = |\,\lambda_3\,| = \frac{1}{2}\sqrt{2} \approx 0.7071 < 1$. Thus, the iterates $\mathbf{u}^{(k)} = T^k\mathbf{a} \to c_1\,\mathbf{v}_1$ will eventually converge, at a rate of about .7, to a

multiple of the first eigenvector. For example, starting with $\mathbf{u}^{(0)} = (\,1, 1, 1\,)^T$, leads to the iterates[†]

$$\mathbf{u}^{(5)} = \begin{pmatrix} -9.5 \\ 4.75 \\ -2.75 \end{pmatrix}, \qquad \mathbf{u}^{(10)} = \begin{pmatrix} -7.9062 \\ 3.9062 \\ -1.9062 \end{pmatrix}, \qquad \mathbf{u}^{(15)} = \begin{pmatrix} -7.9766 \\ 4.0 \\ -2.0 \end{pmatrix},$$

$$\mathbf{u}^{(20)} = \begin{pmatrix} -8.0088 \\ 4.0029 \\ -2.0029 \end{pmatrix}, \qquad \mathbf{u}^{(25)} = \begin{pmatrix} -7.9985 \\ 3.9993 \\ -1.9993 \end{pmatrix}, \qquad \mathbf{u}^{(30)} = \begin{pmatrix} -8.0001 \\ 4.0001 \\ -2.0001 \end{pmatrix},$$

which are slowly converging to the particular eigenvector $(\,-8, 4, -2\,)^T = -2\,\mathbf{v}_1$. This can be predicted in advance by decomposing the initial condition into a linear combination of eigenvectors:

$$\mathbf{u}^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -2 \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + \frac{3 + 3\,\mathrm{i}}{2} \begin{pmatrix} 2 - \mathrm{i} \\ -1 \\ 1 \end{pmatrix} + \frac{3 - 3\,\mathrm{i}}{2} \begin{pmatrix} 2 + \mathrm{i} \\ -1 \\ 1 \end{pmatrix},$$

whence

$$\mathbf{u}^{(k)} = \begin{pmatrix} -8 \\ 4 \\ -2 \end{pmatrix} + \frac{3 + 3\,\mathrm{i}}{2} \left(\frac{1 + \mathrm{i}}{2} \right)^k \begin{pmatrix} 2 - \mathrm{i} \\ -1 \\ 1 \end{pmatrix} + \frac{3 - 3\,\mathrm{i}}{2} \left(\frac{1 - \mathrm{i}}{2} \right)^k \begin{pmatrix} 2 + \mathrm{i} \\ -1 \\ 1 \end{pmatrix},$$

and so $\mathbf{u}^{(k)} \to (\,-8, 4, -2\,)^T$ as $k \to \infty$,

10.3. Matrix Norms.

The convergence of a linear iterative system is governed by the spectral radius of the coefficient matrix, and hence knowledge of its eigenvalues is essential. Unfortunately, a priori information on the eigenvalues is not so easy to come by. Indeed, computing accurate approximations to the eigenvalues of a general matrix is a difficult computational problem, and completely satisfactory general numerical algorithms are not known. Indeed, the simplest way to determine the spectral radius is, in fact, to explicitly iterate the matrix and observe how fast the resulting vectors grow or decay. But this defeats its purpose!

An alternative, more practical approach to convergence is based on the concept of a matrix norm. Matrix norms are a natural class of norms on the vector space of $n \times n$ matrices. They often provide comparable convergence information for linear iterative systems, and are simpler to compute.

We work exclusively with real $n \times n$ matrices in this section, although the results straightforwardly extend to complex $n \times n$ matrices. Let us fix a norm $\| \cdot \|$ on $\mathbb{R}^n$. The norm may or may not come from an inner product — this is irrelevant as far as the construction goes. Roughly speaking, the matrix norm tells us how far the matrix stretches vectors relative to the given norm.

[†] Since the convergence is slow, we only display every fifth one.

Theorem 10.17. *If $\| \cdot \|$ is any norm on $\mathbb{R}^n$, then the quantity*

$$\| A \| = \max \{ \| A\,\mathbf{u} \| \mid \| \mathbf{u} \| = 1 \} \tag{10.23}$$

defines a norm on the vector space $\mathcal{M}_{n\times n}$ of all $n \times n$ matrices, called the natural matrix norm *associated with the given norm $\| \cdot \|$ on $\mathbb{R}^n$.*

Proof: First note that $\| A \| < \infty$ since the maximum is taken on a closed and bounded subset, namely the unit sphere $S_1 = \{ \| \mathbf{u} \| = 1 \}$ of the given norm. To show that (10.23) defines a norm, we need to verify the three basic axioms of Definition 3.12. Non-negativity, $\| A \| \geq 0$, is immediate. Suppose $\| A \| = 0$. This means that, for every unit vector, $\| A\,\mathbf{u} \| = 0$, and hence $A\,\mathbf{u} = \mathbf{0}$ whenever $\| \mathbf{u} \| = 1$. If $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$ is any nonzero vector, then $\mathbf{u} = \mathbf{v}/r$, where $r = \| \mathbf{v} \|$, is a unit vector, and so

$$A\,\mathbf{v} = A(r\,\mathbf{u}) = r\,A\,\mathbf{u} = \mathbf{0}. \tag{10.24}$$

Therefore, $A\,\mathbf{v} = \mathbf{0}$ for every $\mathbf{v} \in \mathbb{R}^n$, which implies $A = \mathrm{O}$ is the zero matrix. This serves to prove the positivity property. As for homogeneity, if $c \in \mathbb{R}$ is any scalar,

$$\| c\,A \| = \max \{ \| c\,A\,\mathbf{u} \| \} = \max \{ | c | \, \| A\,\mathbf{u} \| \} = | c | \, \| A \|.$$

Finally, to prove the triangle inequality, we use the fact that the maximum of the sum of quantities is bounded by the sum of their individual maxima. Therefore, since the norm on $\mathbb{R}^n$ satisfies the triangle inequality,

$$\begin{aligned} \| A + B \| = \max \{ \| A\,\mathbf{u} + B\,\mathbf{u} \| \} &\leq \max \{ \| A\,\mathbf{u} \| + \| B\,\mathbf{u} \| \} \\ &\leq \max \{ \| A\,\mathbf{u} \| \} + \max \{ \| B\,\mathbf{u} \| \} = \| A \| + \| B \|. \end{aligned}$$

This completes the proof that the matrix norm satisfies the three basic axioms. *Q.E.D.*

The property that distinguishes a matrix norm over a generic norm on the space of matrices is the fact that it obeys a very useful *product inequality*.

Theorem 10.18. *A natural matrix norm satisfies*

$$\| A\,\mathbf{v} \| \leq \| A \| \, \| \mathbf{v} \|, \qquad \text{for all} \qquad A \in \mathcal{M}_{n\times n}, \quad \mathbf{v} \in \mathbb{R}^n. \tag{10.25}$$

Furthermore,

$$\| A\,B \| \leq \| A \| \, \| B \|, \qquad \text{for all} \qquad A, B \in \mathcal{M}_{n\times n}. \tag{10.26}$$

Proof: Note first that, by definition $\| A\,\mathbf{u} \| \leq \| A \|$ for all unit vectors $\| \mathbf{u} \| = 1$. Then, letting $\mathbf{v} = r\,\mathbf{u}$ where $\mathbf{u}$ is a unit vector and $r = \| \mathbf{v} \|$, we have

$$\| A\,\mathbf{v} \| = \| A(r\,\mathbf{u}) \| = r \, \| A\,\mathbf{u} \| \leq r \, \| A \| = \| \mathbf{v} \| \, \| A \|,$$

proving the first inequality. To prove the second, we apply the first to compute

$$\begin{aligned} \| A\,B \| = \max \{ \| A\,B\,\mathbf{u} \| \} &= \max \{ \| A\,(B\,\mathbf{u}) \| \} \\ &\leq \max \{ \| A \| \, \| B\,\mathbf{u} \| \} = \| A \| \, \max \{ \| B\,\mathbf{u} \| \} = \| A \| \, \| B \|. \end{aligned}$$

This completes the proof. *Q.E.D.*

Remark: A norm on the vector space of $n \times n$ matrices is called a *matrix norm* if it also satisfies the multiplicative inequality (10.26). Most, but not all, matrix norms used in applications come from norms on the underlying vector space.

The multiplicative inequality (10.26) implies, in particular, that $\| A^2 \| \leq \| A \|^2$; equality is not necessarily true. More generally,

Lemma 10.19. *If A is a square matrix, then $\| A^k \| \leq \| A \|^k$. In particular, if $\| A \| < 1$, then $\| A^k \| \to 0$ as $k \to \infty$, and hence A is a convergent matrix: $A^k \to \mathrm{O}$.*

The converse is not quite true; a convergent matrix does not necessarily have matrix norm less than 1, or even ≤ 1 — see Example 10.24 below for an explicit example. An alternative proof of Lemma 10.19 can be based on the following useful estimate.

Theorem 10.20. *The spectral radius of a matrix is bounded by its matrix norm*:

$$\rho(A) \leq \| A \|. \tag{10.27}$$

Proof: If λ is a real eigenvalue, and $\mathbf{u}$ a corresponding unit eigenvector, so that $A\,\mathbf{u} = \lambda\,\mathbf{u}$ with $\| \mathbf{u} \| = 1$, then

$$\| A\,\mathbf{u} \| = \| \lambda\,\mathbf{u} \| = | \lambda |\, \| \mathbf{u} \| = | \lambda |. \tag{10.28}$$

Since $\| A \|$ is the maximum of $\| A\,\mathbf{u} \|$ over all possible unit vectors, this implies that

$$| \lambda | \leq \| A \|. \tag{10.29}$$

If all the eigenvalues of A are real, then the spectral radius is the maximum of their absolute values, and so it too is bounded by $\| A \|$, proving (10.27).

If A has complex eigenvalues, then we need to work a little harder. Let $\lambda = r\, e^{\,\mathrm{i}\,\theta}$ be a complex eigenvalue with complex eigenvector $\mathbf{z} = \mathbf{x} + \mathrm{i}\,\mathbf{y}$. Define

$$m = \min \left\{ \, \| \operatorname{Re} e^{\,\mathrm{i}\,\varphi}\, \mathbf{z} \| = \| (\cos\varphi)\,\mathbf{x} - (\sin\varphi)\,\mathbf{y} \| \;\big|\; 0 \leq \varphi \leq 2\pi \, \right\}. \tag{10.30}$$

Since the indicated subset is a closed curve that does not go through the origin[†], $m > 0$. Let φ_0 denote the value of the angle that produces the minimum, so

$$m = \| (\cos\varphi_0)\,\mathbf{x} - (\sin\varphi_0)\,\mathbf{y} \| = \| \operatorname{Re} \left(e^{\,\mathrm{i}\,\varphi_0}\, \mathbf{z} \right) \|.$$

Define the real unit vector

$$\mathbf{u} = \frac{\operatorname{Re} \left(e^{\,\mathrm{i}\,\varphi_0}\, \mathbf{z} \right)}{m} = \frac{(\cos\varphi_0)\,\mathbf{x} - (\sin\varphi_0)\,\mathbf{y}}{m}, \qquad \text{so that} \qquad \| \mathbf{u} \| = 1.$$

Then

$$A\,\mathbf{u} = \frac{1}{m} \operatorname{Re} \left(e^{\,\mathrm{i}\,\varphi_0}\, A\,\mathbf{z} \right) = \frac{1}{m} \operatorname{Re} \left(r\, e^{\,\mathrm{i}\,\varphi_0}\, e^{\,\mathrm{i}\,\theta)}\, \mathbf{z} \right) = \frac{r}{m} \operatorname{Re} \left(e^{\,\mathrm{i}\,(\varphi_0 + \theta)}\, \mathbf{z} \right).$$

Therefore, using the fact that m is the minimal value in (10.30),

$$\| A \| \geq \| A\,\mathbf{u} \| = \frac{r}{m} \| \operatorname{Re} \left(e^{\,\mathrm{i}\,(\varphi_0 + \theta)}\, \mathbf{z} \right) \| \geq r = | \lambda |, \tag{10.31}$$

and so (10.29) also holds for complex eigenvalues. *Q.E.D.*

[†] This relies on the fact that $\mathbf{x}, \mathbf{y}$ are linearly independent, which was shown in Exercise ■.

Explicit Formulae

Let us now determine the explicit formulae for the matrix norms corresponding to our most important vector norms, introduced in Example 3.13. Let us begin with the ∞ matrix norm.

Definition 10.21. The i^{th} *absolute row sum* of a matrix A is the sum of the absolute values (moduli) of the entries in the i^{th} row:

$$s_i = |\, a_{i1} \,| + \cdots + |\, a_{in} \,| = \sum_{j=1}^{n} |\, a_{ij} \,|. \tag{10.32}$$

Proposition 10.22. *The* ∞ matrix norm *of a matrix A is equal to the maximal absolute row sum:*

$$\| A \|_\infty = \max\{s_1, \dots, s_n\} = \max \left\{ \sum_{j=1}^{n} |\, a_{ij} \,| \;\middle|\; 1 \le i \le n \right\}. \tag{10.33}$$

Proof: Let $s = \max\{s_1, \dots, s_n\}$ denote the right hand side of (10.33). Given any $\mathbf{v} \in \mathbb{R}^n$, we compute

$$\begin{aligned} \| A\mathbf{v} \|_\infty &= \max \left\{ \left| \sum_{j=1}^{n} a_{ij} v_j \right| \right\} \le \max \left\{ \sum_{j=1}^{n} |\, a_{ij} v_j \,| \right\} \\ &\le \max \left\{ \sum_{j=1}^{n} |\, a_{ij} \,| \right\} \max \{\, |\, v_j \,| \,\} = s \, \| \mathbf{v} \|_\infty. \end{aligned}$$

In particular, by specializing to $\| \mathbf{v} \|_\infty = 1$, we deduce that $\| A \|_\infty \le s$.

On the other hand, suppose the maximal absolute row sum occurs at row i, so

$$s_i = \sum_{j=1}^{n} |\, a_{ij} \,| = s. \tag{10.34}$$

Let $\mathbf{u}$ be defined so that $u_j = +1$ if $a_{ij} > 0$, while $u_j = -1$ if $a_{ij} < 0$. Then $\| \mathbf{u} \|_\infty = 1$. Moreover, the i^{th} entry of $A\mathbf{u}$ is equal to the i^{th} row sum (10.34). This implies that

$$\| A \|_\infty \ge \| A\mathbf{u} \|_\infty \ge s. \qquad \textit{Q.E.D.}$$

Corollary 10.23. *If A has maximal absolute row sum strictly less than* 1, *then* $\| A \|_\infty < 1$ *and hence A is a convergent matrix.*

This is an immediate consequence of Lemma 10.19.

Example 10.24. Consider the symmetric matrix $A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{4} \end{pmatrix}$. Its two absolute row sums are $\left|\frac{1}{2}\right| + \left|-\frac{1}{3}\right| = \frac{5}{6}$, $\left|-\frac{1}{3}\right| + \left|\frac{1}{4}\right| = \frac{7}{12}$ and so

$$\| A \|_\infty = \max\left\{ \tfrac{5}{6}, \tfrac{7}{12} \right\} = \tfrac{5}{6} \approx .83333\ldots .$$

Since the norm is less than 1, A is a convergent matrix. Indeed, its eigenvalues are

$$\lambda_1 = \frac{9+\sqrt{73}}{24} \approx .7310\ldots , \qquad \lambda_2 = \frac{9-\sqrt{73}}{24} \approx .0190\ldots ,$$

and hence the spectral radius is

$$\rho(A) = \frac{9+\sqrt{73}}{24} \approx .7310\ldots ,$$

which is slightly smaller than its ∞ norm.

The row sum test for convergence is not always conclusive. For example, the matrix

$$A = \begin{pmatrix} \frac{1}{2} & -\frac{3}{5} \\ \frac{3}{5} & \frac{1}{4} \end{pmatrix} \qquad \text{has matrix norm} \qquad \| A \|_\infty = \tfrac{11}{10} > 1.$$

On the other hand, its eigenvalues are $(15 \pm \sqrt{601}\,)/40$, and hence its spectral radius is

$$\rho(A) = \frac{15+\sqrt{601}}{40} \approx .98788\ldots ,$$

which implies that A is (just barely) convergent, even though its maximal row sum is larger than 1.

The Euclidean matrix norm relies on the singular value decomposition of Theorem 8.33.

Proposition 10.25. *The matrix norm corresponding to the Euclidean norm is its maximal singular value*

$$\| A \|_2 = \max\{\sigma_1, \ldots, \sigma_n\}. \tag{10.35}$$

Proof: We use the singular value decomposition (8.41) to write

$$A = Q_1 \, \Sigma \, Q_2^T$$

where Q_1 and Q_2 are orthogonal matrices, while $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ is the diagonal matrix containing the singular values of A. Using the Euclidean norm-preserving property (7.32) of orthogonal matrices, we have

$$\| A\,\mathbf{u} \|_2 = \| \, Q_1 \, \Sigma \, Q_2^T \mathbf{u} \, \|_2 = \| \, \Sigma \, Q_2^T \mathbf{u} \, \|_2 .$$

Now, if $\mathbf{u}$ is a unit vector, $\| \, \mathbf{u} \, \|_2 = 1$, then so is $\widetilde{\mathbf{u}} = Q_2^T \mathbf{u}$. Therefore,

$$\begin{aligned} \| A \|_2 &= \max \{ \, \| A\,\mathbf{u} \|_2 \; | \;\; \| \, \mathbf{u} \, \|_2 = 1 \, \} \\ &= \max \left\{ \, \| \, \Sigma \, Q_2^T \mathbf{u} \, \|_2 \;\middle|\;\; \| \, \mathbf{u} \, \|_2 = 1 \, \right\} = \max \{ \, \| \, \Sigma \, \widetilde{\mathbf{u}} \, \|_2 \; | \;\; \| \, \widetilde{\mathbf{u}} \, \|_2 = 1 \, \} . \end{aligned}$$

If we order the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, then

$$\| \Sigma \widetilde{\mathbf{u}} \|_2 = \sqrt{\sigma_1^2 \widetilde{u}_1^2 + \cdots + \sigma_n^2 \widetilde{u}_n^2} \leq \sigma_1 \sqrt{\widetilde{u}_1^2 + \cdots + \widetilde{u}_n^2} = \sigma_1 \qquad \text{since} \qquad \| \widetilde{\mathbf{u}} \|_2 = 1.$$

On the other hand, if we set $\widetilde{\mathbf{u}} = \mathbf{e}_1$ to be the first standard basis vector, then $\| \Sigma \mathbf{e}_1 \|_2 = \| \sigma_1 \mathbf{e}_1 \|_2 = \sigma_1$. When put together, these imply that $\| A \|_2 = \sigma_1$, which proves the result. *Q.E.D.*

Corollary 10.26. *If A is symmetric, its Euclidean matrix norm is equal to its spectral radius.*

Proof: This follows directly from the fact, proved in Proposition 8.31, that the singular values of a symmetric matrix are just the absolute values of its eigenvalues. *Q.E.D.*

Example 10.27. Consider the matrix $A = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}$. The Gram matrix

$$A^T A = \begin{pmatrix} 0.2225 & 0.0800 & 0.1250 \\ 0.0800 & 0.1511 & -0.1111 \\ 0.1250 & -0.1111 & 0.3611 \end{pmatrix},$$

has eigenvalues $\lambda_1 = 0.4472, \lambda_2 = 0.2665, \lambda_3 = 0.0210$, and hence the singular values of A are the square roots: $\sigma_1 = 0.6687, \sigma_2 = 0.5163, \sigma_3 = 0.1448$. The Euclidean matrix norm of A is the largest singular value, and so $\| A \|_2 = 0.6687$, proving that A is a convergent matrix. Note that, as always, the matrix norm overestimates the spectral radius $\rho(A) = .5$.

Unfortunately, as we discovered in Example 10.24, matrix norms are not a foolproof test of convergence There exist convergent matrices such that $\rho(A) < 1$ and yet have matrix norm $\| A \| \geq 1$. In such cases, we will not be able to predict the convergence of the iterative system based on the matrix, although we would expect the convergence to be quite slow. Although such pathology might show up in one particular matrix norm, it turns out that one can always find some matrix norm which is less than 1. A proof of this result can be found in [**119**].

Theorem 10.28. *Let A have spectral radius $\rho(A)$. If $\varepsilon > 0$ is any positive number, then there exists a matrix norm $\| \cdot \|$ such that*

$$\rho(A) \leq \| A \| < \rho(A) + \varepsilon. \tag{10.36}$$

Corollary 10.29. *If A is a convergent matrix, then there exists a matrix norm such that $\| A \| < 1$.*

Proof: By definition, A is convergent if and only if $\rho(A) < 1$. Choose $\varepsilon > 0$ such that $\rho(A) + \varepsilon < 1$. Any norm that satisfies (10.36) has the desired property. *Q.E.D.*

Remark: Based on the accumulated evidence, one might be tempted to speculate that the spectral radius itself defines a matrix norm. Unfortunately, this is not the case. For example, the nonzero matrix $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has zero spectral radius, $\rho(A) = 0$, violating a basic norm axiom.

The Gerschgorin Circle Theorem

In general, precisely computing the eigenvalues, and hence the spectral radius of a matrix is not easy, and, in most cases, must be done through a numerical eigenvalue routine. In many applications, though, one does not need their exact numerical values, but only their approximate locations. The *Gerschgorin Circle Theorem* serves to restrict the eigenvalues to a certain well-defined region in the complex plane. In favorable situations, this information, which is relatively easy to obtain, is sufficient to demonstrate the convergence of the matrix.

Definition 10.30. Let A be an $n \times n$ matrix, either real or complex. For each $1 \le i \le n$, define the *Gerschgorin disk*

$$D_i = \{\, | z - a_{ii} | \le r_i \mid \ z \in \mathbb{C} \,\}, \qquad \text{where} \qquad r_i = \sum_{\substack{i=1 \\ i \neq j}}^{n} | a_{ij} |. \tag{10.37}$$

The *Gerschgorin domain* $D = \bigcup_{i=1}^{n} \subset \mathbb{C}$ is the union of the Gerschgorin disks.

Thus, the i^{th} Gerschgorin disk D_i is centered at the i^{th} diagonal entry a_{ii}, and has radius r_i equal to the sum of the absolute values of the off-diagonal entries that are in the i^{th} row of A.

Theorem 10.31. *All real and complex eigenvalues of the matrix A lie in its Gerschgorin domain D.*

Example 10.32. The matrix $A = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 4 & -1 \\ -1 & -1 & -3 \end{pmatrix}$ has Gerschgorin disks

$$D_1 = \{\, | z - 2 | \le 1 \,\}, \qquad D_2 = \{\, | z - 4 | \le 2 \,\}, \qquad D_3 = \{\, | z + 3 | \le 2 \,\},$$

which are plotted in Figure 10.2. The eigenvalues of A are

$$\lambda_1 = 3, \qquad \lambda_2 = 3.1623\ldots, \qquad \lambda_3 = -3.1623\ldots.$$

Observe that λ_1 belongs to both D_1 and D_2, while λ_2 lies in D_2, and λ_3 in D_3. We thus confirm that all three eigenvalues are in the Gerschgorin domain $D = D_1 \cup D_2 \cup D_3$.

Proof of Theorem 10.31: Let $\mathbf{v}$ be an eigenvector of A with eigenvalue λ. Let $\mathbf{u} = \mathbf{v}/\| \mathbf{v} \|_\infty$ be the corresponding unit eigenvector with respect to the ∞ norm, so

$$\| \mathbf{u} \|_\infty = \max\{\, | u_1 |, \ \ldots \ , | u_n | \,\} = 1.$$

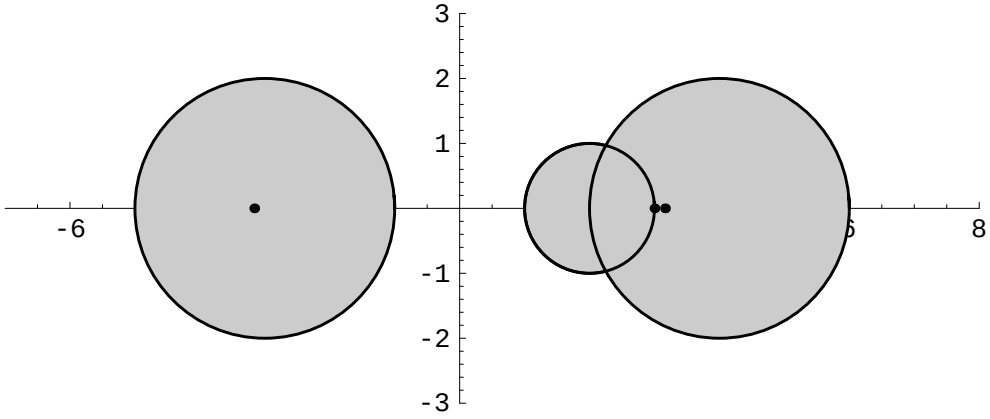


Figure 10.2. Gerschgorin Disks and Eigenvalues.

Let u_i be an entry of $\mathbf{u}$ that achieves the maximum: $|\, u_i \,| = 1$. Writing out the eigenvalue equation $A\,\mathbf{u} = \lambda\,\mathbf{u}$ in components, we find

$$\sum_{j=1}^{n} a_{ij}\, u_j = \lambda\, u_i, \qquad \text{which we rewrite as} \qquad \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\, u_j = (\lambda - a_{ii})\, u_i.$$

Therefore, since all $|\, u_i \,| \leq 1$,

$$|\, \lambda - a_{ii} \,|\, |\, u_i \,| \;\leq\; \left| \sum_{j\neq i} a_{ij}\, u_j \right| \;\leq\; \sum_{j\neq i} |\, a_{ij} \,|\, |\, u_j \,| \;\leq\; \sum_{j\neq i} |\, a_{ij} \,| = r_i.$$

Since we chose u_i so that $|\, u_i \,| = 1$, we conclude that λ satisfies

$$|\, \lambda - a_{ii} \,| \;\leq\; r_i,$$

and hence $\lambda \in D_i \subset D$ belongs to the i^{th} Gerschgorin disk. *Q.E.D.*

The Gerschgorin Theorem 10.31 can be used to give a direct proof of Corollary 10.23. If A is any matrix, then the modulus of all points $z \in D_i$ contained in its i^{th} Gerschgorin disk is bounded by the i^{th} absolute row sum,

$$|\, z \,| \leq |\, z - a_{ii} \,| + |\, a_{ii} \,| \leq r_i + |\, a_{ii} \,| = s_i,$$

where the final equality follows by comparison of (10.37) and (10.32). Thus, every point $z \in D$ in the Gerschgorin set has modulus

$$|\, z \,| \leq \max\{s_1, \ldots, s_n\} = \|\, A \,\|_\infty,$$

bounded by the maximal row sum. Since all eigenvalues λ_j of A are contained in D, they too satisfy

$$|\, \lambda_j \,| \leq \|\, A \,\|_\infty, \qquad \text{and hence} \qquad \rho(A) \leq \|\, A \,\|_\infty. \tag{10.38}$$

By hypothesis, $1 > \|\, A \,\|_\infty \geq \rho(A)$, and hence A is a convergent matrix.

As a second application, we give a simple direct test that guarantees invertibility of a matrix without requiring Gaussian elimination or computing determinants. Recall that a matrix is nonsingular if and only if it does not have a zero eigenvalue. Thus, if its Gerschgorin domain does not contain $0 \not\in \mathcal{D}$, then the matrix cannot have 0 as an eigenvalue, and hence is necessarily invertible. This condition requires that the matrix have large diagonal entries, as quantified by the following definition.

Definition 10.33. A square matrix A is called *strictly diagonally dominant* if

$$| a_{ii} | \; > \; \sum_{\substack{i=1 \\ i \neq j}}^{n} | a_{ij} |, \qquad \text{for all} \qquad i = 1, \dots, n. \tag{10.39}$$

In other words, for A to be diagonally dominant, its diagonal entry must be larger, in absolute value, than the sum of *all* the other entries in its row. For example, the matrix $A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}$ is strictly diagonally dominant since

$$| 3 | > | 1 | + | -1 |, \qquad | -4 | > | 1 | + | 2 |, \qquad | 5 | > | -2 | + | -1 |.$$

Diagonally dominant matrices arise in many applications, particularly in finite difference and finite element methods for numerically solving boundary value problems. As we shall see, they are the most common class of matrices to which iterative solution methods can be successfully applied.

Proposition 10.34. *A strictly diagonally dominant matrix is nonsingular.*

Proof: The diagonal dominance inequalities (10.39) imply that the radius of the i^{th} Gerschgorin disk is strictly less than the modulus of its center: $r_i < | a_{ii} |$. Thus, the disk cannot contain 0; indeed, if $z \in D_i$, then, by the triangle inequality

$$r_i > | z - a_{ii} | \geq | a_{ii} | - | z | > r_i - | z |, \qquad \text{and hence} \qquad | z | > 0.$$

Thus, $0 \not\in \mathcal{D}$ does not lie in the Gerschgorin domain and hence cannot be an eigenvalue. *Q.E.D.*

Warning: The converse is obviously not true. There are plenty of nonsingular matrices that are not diagonally dominant.

10.4. Markov Processes.

A discrete process in which the probability of a system being in a particular state during a given time period depends only its state in the immediately preceding time period is known as a *Markov chain*, in honor of the pioneering studies of the Russian mathematician Andrei Markov. Markov chains are the beginning of the theory of stochastic processes. They are described by linear iterative systems whose coefficient matrices have a special form, and hence can be analyzed by our eigenvalue methods.

To take a very simple example, suppose you are interested in predicting whether the weather in your city on a particular day will be either sunny or cloudy. Consulting weather records over the past decade, you determine that

(i) If today is sunny, there is a 70% chance that tomorrow will also be sunny,

(ii) But, if today is cloudy, the chances are 80% that tomorrow is also cloudy.

Question: given that today is sunny, what is the probability that next Saturday's weather will also be sunny?

To mathematically formulate this process, we let $s^{(k)}$ denote the probability that day k is sunny and $c^{(k)}$ the probability that it is cloudy. If we assume that these are the only possibilities, then the individual probabilities must sum to 1, so

$$s^{(k)} + c^{(k)} = 1.$$

According to our data, the probability that the next day is sunny or cloudy is expressed by the equations

$$s^{(k+1)} = .7\, s^{(k)} + .2\, c^{(k)}, \qquad c^{(k+1)} = .3\, s^{(k)} + .8\, c^{(k)}. \tag{10.40}$$

Indeed, day $k+1$ could be sunny either if day k was, with a 70% chance, or, if day k was cloudy, there is still a 20% chance of day $k+1$ being sunny. We rewrite (10.40) in a more convenient matrix form:

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}, \qquad \text{where} \qquad T = \begin{pmatrix} .7 & .2 \\ .3 & .8 \end{pmatrix}, \qquad \mathbf{u}^{(k)} = \begin{pmatrix} s^{(k)} \\ c^{(k)} \end{pmatrix}. \tag{10.41}$$

In a Markov process, the vector of probabilities $\mathbf{u}^{(k)}$ is known as the k^{th} *state vector* and the matrix T is known as the *transition matrix*, whose entries fix the transition probabilities between the states.

By assumption, our initial state vector is $\mathbf{u}^{(0)} = (\,1, 0\,)^T$, since we know for certain that today is sunny. Rounding off to three decimal places, the subsequent state vectors are

$$\mathbf{u}^{(1)} = \begin{pmatrix} .7 \\ .3 \end{pmatrix}, \quad \mathbf{u}^{(2)} = \begin{pmatrix} 0.55 \\ 0.45 \end{pmatrix}, \quad \mathbf{u}^{(3)} = \begin{pmatrix} 0.475 \\ 0.525 \end{pmatrix}, \quad \mathbf{u}^{(4)} = \begin{pmatrix} 0.438 \\ 0.563 \end{pmatrix},$$
$$\mathbf{u}^{(5)} = \begin{pmatrix} 0.419 \\ 0.581 \end{pmatrix}, \quad \mathbf{u}^{(6)} = \begin{pmatrix} 0.410 \\ 0.591 \end{pmatrix}, \quad \mathbf{u}^{(7)} = \begin{pmatrix} 0.405 \\ 0.595 \end{pmatrix}, \quad \mathbf{u}^{(8)} = \begin{pmatrix} 0.402 \\ 0.598 \end{pmatrix}.$$

The iterates converge fairly rapidly to $(\,.4, .6\,)^T$, which is a fixed point for the iterative system (10.41). Thus, in the long run, 40% of the days will be sunny and 60% will be cloudy. Let us explain why this happens.

Definition 10.35. A vector $\mathbf{u} = (\,u_1, u_2, \ldots, u_n\,)^T \in \mathbb{R}^n$ is called a *probability vector* if all its individual entries lie between 0 and 1, so $0 \le u_i \le 1$, and, moreover, the sum of its entries is unity: $u_1 + \cdots + u_n = 1$.

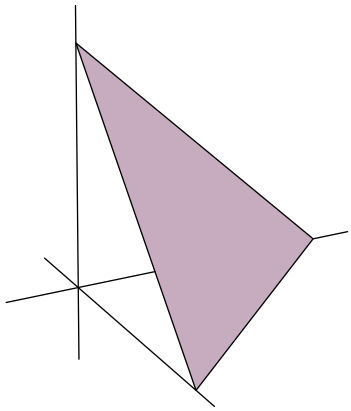


Figure 10.3. Probability Vectors in $\mathbb{R}^3$.

For example, the possible probability vectors $\mathbf{u} \in \mathbb{R}^3$ fill the equilateral triangle plotted in Figure 10.3. We interpret the entry u_i of a probability vector as the probability the system is in state number i. The fact that the entries add up to 1 means that they represent a complete list of probabilities for the possible states of the system.

Remark: Any nonzero vector $\mathbf{0} \neq \mathbf{v} = (\, v_1, v_2, \dots, v_n \,)^T$ with all non-negative entries: $v_i \geq 0$ for $i = 1, \dots, n$, can be converted into a parallel probability vector by dividing by the sum of its entries:

$$\mathbf{u} = \frac{\mathbf{v}}{v_1 + \cdots + v_n}\,. \tag{10.42}$$

For example, if $\mathbf{v} = (\, 3, 2, 0, 1 \,)^T$, then $\mathbf{u} = \left(\, \frac{1}{2}, \frac{1}{3}, 0, \frac{1}{6} \,\right)^T$ is the corresponding probability vector.

In general, a *Markov chain* is represented by a first order linear iterative system

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}. \tag{10.43}$$

The *transition matrix*

$$T = \big(\, t_{ij} \,\big), \qquad 0 \leq t_{ij} \leq 1, \qquad t_{1j} + \cdots + t_{nj} = 1, \tag{10.44}$$

contains all the transitional probabilities. The entry t_{ij} represents the probability that the system will switch from state j to state i. (Note the reversal of indices.) Since this covers all possible transitions, the *column sums* of the transition matrix are all equal to 1, and hence each column of T is a probability vector. An easy Exercise ∎ shows that if $\mathbf{u}^{(k)}$ is a probability vector, so is $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$. Thus, the solution $\mathbf{u}^{(k)} = T^k\,\mathbf{u}^{(0)}$ to the Markov process represents a sequence or "chain" of probability vectors.

It can be proved, [**94**] that every transition matrix T is complete, and hence admits an eigenvector basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ with associated eigenvalues $\lambda_1, \dots, \lambda_n$ (some of which may be repeated). Therefore, by Theorem 10.4, the solution to the Markov process (10.43) is

$$\mathbf{u}^{(k)} = T^k\,\mathbf{u}^{(0)} = c_1\,\lambda_1^k\,\mathbf{v}_1 + \;\cdots\; + c_n\,\lambda_n^k\,\mathbf{v}_n, \tag{10.45}$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues and $\mathbf{v}_1, \dots, \mathbf{v}_n$ the corresponding eigenvectors.

Let us now investigate the convergence of the Markov chain. This will not happen in general, but requires some additional mild restrictions on the transition matrix.

Definition 10.36. A transition matrix (10.44) is *regular* if some power T^k contains no zero entries. In particular, if T itself has no transition probabilities equal to 0, then it is regular.

Warning: The term "regular transition matrix" is *not* the same as our earlier term "regular matrix", which was used to describe matrices with an LU factorization.

The asymptotic behavior of a regular Markov chain is governed by the following key result.

Theorem 10.37. *If T is a regular transition matrix, then it admits a unique* probability eigenvector $\mathbf{u}^\star$ *with eigenvalue $\lambda_1 = 1$. Moreover, any Markov chain with coefficient matrix T will converge to the distinguished probability vector*: $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ *as $k \to \infty$.*

The proof of this result appears at the end of this section.

Example 10.38. For the weather transition matrix (10.41), the eigenvalues and eigenvectors are

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} \frac{2}{3} \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = .5, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

The first eigenvector is then converted into a probability vector via (10.42):

$$\mathbf{u}^\star = \mathbf{u}_1 = \frac{1}{1 + \frac{2}{3}} \begin{pmatrix} \frac{2}{3} \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2}{5} \\ \frac{3}{5} \end{pmatrix}.$$

This distinguished probability eigenvector represents the final asymptotic state of the system after many iterations, *no matter what the initial state*. Thus, our earlier observation that about 40% of the days will be sunny and 60% will be cloudy holds *no matter what the initial weather is*.

Example 10.39. A taxi company in Minnesota serves the cities of Minneapolis and St. Paul, as well as the nearby suburbs. Records indicate that, on average, 10% of the customers taking a taxi in Minneapolis go to St. Paul and 30% go to the suburbs. Customers boarding in St. Paul have a 30% chance of going to Minneapolis and 30% chance of going to the suburbs, while suburban customers choose Minneapolis 40% of the time and St. Paul 30% of the time. The owner of the taxi company is interested in knowing where the taxis will end up, on average. We write this as a Markov process. The entries of the state vector $\mathbf{u}^{(k)} = (u_1^{(k)}, u_2^{(k)}, u_3^{(k)})^T$ tell what proportion of the taxi fleet is, respectively, in Minneapolis, St. Paul and the suburbs. Using the data, we construct the relevant transition matrix

$$T = \begin{pmatrix} .6 & .3 & .4 \\ .1 & .4 & .3 \\ .3 & .3 & .3 \end{pmatrix}.$$

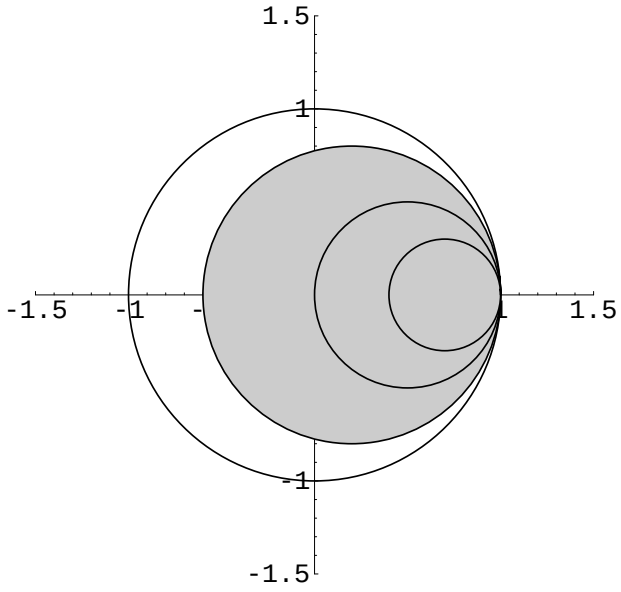


Figure 10.4. Gerschgorin Disks for a Transition Matrix.

Note that T regular since it has no zero entries. The probability eigenvector

$$\mathbf{u}^\star = (\, 0.471429\ldots \quad 0.228571\ldots \quad 0.3 \,)^T$$

corresponding to the unit eigenvalue $\lambda_1 = 1$ is found by first solving the linear system $(T - \mathrm{I})\mathbf{v} = 0$ and then converting the solution[†] $\mathbf{v}$ into a valid probability vector by use of formula (10.42). According to Theorem 10.37, no matter how the taxis are initially distributed, ultimately about 47% of the taxis will be in Minneapolis, 23% in St. Paul, and 30% in the suburbs. This can be confirmed by running numerical experiments on the system.

Remark: The convergence rate of the Markov chain to its steady state is governed by the size of the subdominant or second largest eigenvalue λ_2. The smaller $|\,\lambda_2\,|$ is to 0, the faster the process converges. In the taxi example, $\lambda_2 = .3$ (and $\lambda_3 = 0$) and so the convergence to steady state is fairly rapid.

Proof of Theorem 10.37: We begin the proof by replacing T by its transpose $M = T^T$, keeping in mind that every eigenvalue of T is also an eigenvalue of M, cf. Exercise ■. The conditions (10.44) tell us that the matrix M has entries $0 \le m_{ij} = t_{ji} \le 1$, and, moreover, the (absolute) *row sums* $s_i = \sum_{i=1}^{n} m_{ij} = 1$ of M, being the same as the corresponding column sums of T, are all equal to 1. Since $M^k = (T^k)^T$, regularity of T implies that some power M^k has all positive entries.

[†] Theorem 10.37 guarantees an eigenvector $\mathbf{v}$ with all non-negative entries.

According to Exercise ■, if $\mathbf{z} = (1, \dots, 1)^T$ is the column vector all of whose entries are equal to 1, then the entries of $M\,\mathbf{z}$ are the row sums of M. Therefore, $M\,\mathbf{z} = \mathbf{z}$, which implies that $\mathbf{z}$ is an eigenvector of M with eigenvalue $\lambda_1 = 1$. As a consequence, T also has 1 as an eigenvalue — although it is associated with a different eigenvector, not necessarily a multiple of $\mathbf{z}$.

Let us next prove that $\lambda_1 = 1$ is a simple eigenvalue. Since T is complete, this is equivalent to the statement that the only vectors satisfying $M\,\mathbf{v} = \mathbf{v}$ are those with all equal entries $v_1 = \cdots = v_n = a$, and hence $\mathbf{v} = a\,\mathbf{z}$ is a scalar multiple of the particular eigenvector $\mathbf{z}$. Let us first prove this assuming all of the entries of M are positive, and so $0 < m_{ij} = t_{ji} < 1$ for all i, j. Suppose $\mathbf{v}$ is an eigenvector with not all equal entries. Let v_k be the minimal entry of $\mathbf{v}$, so $v_k \le v_i$ for all $i \neq k$ and at least one inequality is strict, say $v_k < v_j$. Then the k^{th} entry of the eigenvector equation $\mathbf{v} = M\,\mathbf{v}$ is

$$v_k = \sum_{j=1}^{n} m_{kj}\, v_j < \sum_{j=1}^{n} m_{kj}\, v_k = v_k,$$

where the strict inequality follows from the positivity of the entries of M, and the final equality follows from the fact that M has unit row sums. Thus, we are led to a contradiction, and the claim follows. If M has one or more 0 entries, but M^k has all positive entries, then we apply the previous argument to the equation $M^k\,\mathbf{v} = \mathbf{v}$ which follows from $M\,\mathbf{v} = \mathbf{v}$.

Finally, let us prove that all the other eigenvalues of M are less than 1 in modulus. For this we appeal to the Gerschgorin Circle Theorem 10.31. The Gerschgorin disk D_i is centered at m_{ii} and has radius $r_i = s_i - m_{ii} = 1 - m_{ii}$. Thus the disk lies strictly inside the open unit disk $|\,z\,| < 1$ *except* for a single boundary point at $z = 1$; see Figure 10.4. The Circle Theorem 10.31 implies that all eigenvalues except the unit eigenvalue $\lambda_1 = 1$ must lie strictly inside the unit disk, and so $|\,\lambda_j\,| < 1$ for $j \ge 2$.

Therefore, the matrix M, and, hence, also T satisfy the hypotheses of Theorem 10.15. We conclude that the iterates $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)} \to \mathbf{u}^\star$ converge to a multiple of the unit eigenvector of T. If the initial condition $\mathbf{u}^{(0)}$ is a probability vector, then so is every subsequent state vector $\mathbf{u}^{(k)}$, and so their limit $\mathbf{u}^\star$ must also be a probability vector. This completes the proof of the theorem. *Q.E.D.*

10.5. Iterative Solution of Linear Systems.

In this section, we introduce several basic iterative methods that are used to approximate the solution of certain classes of linear systems

$$A\,\mathbf{u} = \mathbf{b}, \tag{10.46}$$

consisting of n equations in n unknowns. The resulting algorithms will provide an attractive alternative to Gaussian elimination, particularly when dealing with the large, sparse systems that arise in the numerical solution to differential equations. One major advantage of an iterative technique is that it produces progressively more and more accurate approximations to the solution, and hence, by prolonging the iterations, one can, in principle, compute the solution to any desired order of accuracy — although, in practice, the

round-off errors due to the finite precision of the computer will eventually be an issue. Moreover, even performing just a few iterations may produce a reasonable approximation to the true solution — in stark contrast to Gaussian elimination, where one must continue the algorithm through to the bitter end before any useful information can be extracted. A partially completed Gaussian elimination is of scant use! On the other hand, specific iterative schemes are not universally applicable to all linear systems, and their design relies upon the detailed structure of the coefficient matrix.

We shall be attempting to solving (10.46) by an iterative system of the form

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{c}, \qquad \mathbf{u}^{(0)} = \mathbf{u}_0, \tag{10.47}$$

where T is a fixed $n \times n$ matrix and $\mathbf{c}$ a fixed vector. This is a slight generalization of the linear iterative system (10.1), in that the right hand side is now an affine function of $\mathbf{u}^{(k)}$. If the solutions to the affine iterative system converge, $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ as $k \to \infty$, then $\mathbf{u}^\star$ solves the fixed-point equation

$$\mathbf{u}^\star = T\,\mathbf{u}^\star + \mathbf{c}. \tag{10.48}$$

Indeed, both $\mathbf{u}^{(k)}$ and $\mathbf{u}^{(k+1)}$ in (10.47) converge to the same $\mathbf{u}^\star$, and so the system converges to the limiting fixed point equation (10.48). Thus we need to design our system so that

(*a*) The solution to the fixed-point system (10.48) coincides with the solution to the original system (10.46), and

(*b*) The iterates defined by (10.47) converge to the solution $\mathbf{u}^\star$.

Before exploring these issues in depth, let us look at a simple example.

Example 10.40. Consider the linear system

$$3\,x + y - z = 3, \qquad x - 4\,y + 2\,z = -1, \qquad -2\,x - y + 5\,z = 2, \tag{10.49}$$

which we rewrite in matrix form $A\,\mathbf{u} = \mathbf{b}$, with

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}.$$

One easy way to rewrite the system in fixed-point form (10.48) is to set

$$T = \mathrm{I} - A = \begin{pmatrix} -2 & -1 & 1 \\ -1 & 5 & -2 \\ 2 & 1 & -4 \end{pmatrix}, \qquad \mathbf{c} = \mathbf{b} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}. \tag{10.50}$$

Clearly, $A\,\mathbf{u} = \mathbf{b}$ if and only if $T\,\mathbf{u} + \mathbf{b} = (\mathrm{I} - A)\mathbf{u} + \mathbf{b} = \mathbf{u}$, and hence the fixed point coincides with the solution to the original system. The resulting iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{c}$ has the explicit form

$$\begin{aligned} x^{(k+1)} &= -2\,x^{(k)} - y^{(k)} + z^{(k)} + 3, \\ y^{(k+1)} &= -x^{(k)} + 5\,y^{(k)} - 2\,z^{(k)} - 1, \\ z^{(k+1)} &= 2\,x^{(k)} + y^{(k)} - 4\,z^{(k)} + 2. \end{aligned}$$

Another possibility is to solve the first equation in (10.49) for x, the second for y and the third for z, so that

$$x = -\tfrac{1}{3}\,y + \tfrac{1}{3}\,z + 1, \qquad y = \tfrac{1}{4}\,x + \tfrac{1}{2}\,z + \tfrac{1}{4}\,, \qquad z = \tfrac{2}{5}\,x + \tfrac{1}{5}\,y + \tfrac{2}{5}\,.$$

The solution to this fixed point system also coincide with that of the original linear system. The corresponding iteration takes the form

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{3}\,y^{(k)} + \tfrac{1}{3}\,z^{(k)} + 1, \\ y^{(k+1)} &= \tfrac{1}{4}\,x^{(k)} + \tfrac{1}{2}\,z^{(k)} + \tfrac{1}{4}, \\ z^{(k+1)} &= \tfrac{2}{5}\,x^{(k)} + \tfrac{1}{5}\,y^{(k)} + \tfrac{2}{5}. \end{aligned} \tag{10.51}$$

In matrix notation, this becomes

$$\mathbf{u}^{(k+1)} = \widehat{T}\,\mathbf{u}^{(k)} + \widehat{\mathbf{c}}, \qquad \text{where} \qquad \widehat{T} = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}, \qquad \widehat{\mathbf{c}} = \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{2}{5} \end{pmatrix}. \tag{10.52}$$

Do the resulting iterative schemes (10.47) converge to the solution $x = y = z = 1$? The results, starting with initial guess $\mathbf{u}^{(0)} = (0, 0, 0)$, appear in the following table.

k	$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{b}$			$\mathbf{u}^{(k+1)} = \widehat{T}\,\mathbf{u}^{(k)} + \widehat{\mathbf{c}}$		
0	0	0	0	0	0	0
1	3	−1	2	1	0.25	0.4
2	0	−13	−1	1.05	0.7	0.85
3	15	−64	−7	1.05	0.9375	0.96
4	30	−322	−4	1.0075	0.9925	1.0075
5	261	−1633	−244	1.005	1.00562	1.0015
6	870	−7939	−133	0.9986	1.002	1.0031
7	6069	−40300	−5665	1.0004	1.0012	0.9999
8	22500	−196240	−5500	0.9995	1.0000	1.0004
9	145743	−992701	−129238	1.0001	1.0001	0.9998
10	571980	−4850773	−184261	0.9999	0.9999	1.0001
11	3522555	−24457324	−2969767	1.0000	1.0000	1.0000

For the first scheme, the answer is no — the iterations become successively wilder and wilder. Indeed, this occurs no matter how close the initial guess is to the actual solution — unless it happens to be exactly equal: $\mathbf{u}^{(0)} = \mathbf{u}^\star$. (And even then, numerical errors could creep in and send the iterations off to ∞.) In the second case, the convergence is quite good, and it does not take too long, even starting from a bad initial guess, to obtain an accurate approximation to the solution.

Of course, in such a simple example, it would be silly to use iteration, when Gaussian elimination can be done by hand and produces the solution almost immediately. However, we use the small examples for illustrative purposes, reserving the full-fledged application of the iterative schemes to the large linear systems arising in applications.

The convergence of solutions to (10.47) to the fixed point $\mathbf{u}^\star$ is based on the behavior of the *error vectors*

$$\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^\star, \tag{10.53}$$

which measure how close the iterates are to the actual solution. Let us find out how the successive error vectors are related. We compute

$$\mathbf{e}^{(k+1)} = \mathbf{u}^{(k+1)} - \mathbf{u}^\star = (T\,\mathbf{u}^{(k)} + \mathbf{a}) - (T\,\mathbf{u}^\star + \mathbf{a}) = T(\mathbf{u}^{(k)} - \mathbf{u}^\star) = T\,\mathbf{e}^{(k)}.$$

Therefore, the error vectors satisfy a linear iterative system

$$\mathbf{e}^{(k+1)} = T\,\mathbf{e}^{(k)}, \tag{10.54}$$

with the *same* coefficient matrix T. Therefore, the errors are given by the explicit formula $\mathbf{e}^{(k)} = T^k\,\mathbf{e}^{(0)}$. Now, the solutions to (10.47) converge to the fixed point, $\mathbf{u}^{(k)} \to \mathbf{u}^\star$, if and only if the error vectors $\mathbf{e}^{(k)} \to \mathbf{0}$ as $k \to \infty$. Consequently, our convergence results for linear iterative systems, as summarized in Proposition 10.10, imply the following basic result.

Proposition 10.41. *The iterative system* (10.47) *will converge to the solution to the fixed point equation* (10.48) *if and only if T is a convergent matrix: $\rho(T) < 1$.*

For example, in the two iterative schemes presented in Example 10.40, the spectral radii of the coefficient matrices are found to be

$$\rho(T) = 4.9675\ldots\,, \qquad\qquad \rho(\,\widehat{T}\,) = 0.5.$$

Therefore, T is not a convergent matrix, which explains the behavior of its iterates, whereas $\widehat{T}$ is convergent, and one expects the error to roughly decrease by a factor of $\frac{1}{2}$ with each new iterate.

The spectral radius $\rho(T)$ of the coefficient matrix will govern the speed of convergence. Therefore, the main goal is to construct an iterative scheme whose coefficient matrix has as small a spectral radius as possible. At the very least, the spectral radius must be less than 1.

The Jacobi Method

The first general iterative scheme for solving linear systems is based on the same simple idea used in our illustrative Example 10.40. Namely, we solve the i^{th} equation in the system $A\,\mathbf{u} = \mathbf{b}$, which is

$$\sum_{j=1}^{n} a_{ij}\,u_j = b_i,$$

for the i^{th} variable. To do this, we need to assume that all the diagonal entries of A are nonzero: $a_{ii} \neq 0$. The result is

$$u_i = -\frac{1}{a_{ii}} \sum_{i \neq j = 1}^{n} a_{ij}\, u_j + \frac{b_i}{a_{ii}} = \sum_{j=1}^{n} t_{ij}\, u_j + c_i, \tag{10.55}$$

where

$$t_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}}\,, & i \neq j, \\ 0, & i = j, \end{cases} \qquad \text{and} \qquad c_i = \frac{b_i}{a_{ii}}\,. \tag{10.56}$$

Equation (10.55) can be rewritten in fixed point form $\mathbf{u} = T\,\mathbf{u} + \mathbf{c}$, and forms the basis of the *Jacobi method*

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{c}, \qquad \mathbf{u}^{(0)} = \mathbf{u}_0, \tag{10.57}$$

named after the influential nineteenth century German analyst Carl Jacobi. The explicit form of the Jacobi iterative scheme is

$$u_i^{(k+1)} = -\frac{1}{a_{ii}} \sum_{i \neq j = 1}^{n} a_{ij}\, u_j^{(k)} + \frac{b_i}{a_{ii}}\,. \tag{10.58}$$

Let us rederive the Jacobi method in a direct matrix form. We begin by decomposing the coefficient matrix

$$A = L + D + U, \tag{10.59}$$

into the sum of a strictly lower triangular matrix L, a diagonal matrix D, and a strictly upper triangular matrix U, each of which is uniquely specified. For example, in the case of the coefficient matrix

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \tag{10.60}$$

the decomposition (10.59) yields

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -2 & -1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Warning: The L, D, U in the elementary additive decomposition (10.59) have nothing to do with the L, D, U in factorizations arising from Gaussian elimination. The latter play no role in the iterative solution methods considered here.

We then rewrite the system

$$A\,\mathbf{u} = (L + D + U)\,\mathbf{u} = \mathbf{b} \quad \text{ in the alternative form } \quad D\,\mathbf{u} = -\,(L + U)\,\mathbf{u} + \mathbf{b}.$$

The Jacobi fixed point equation amounts to solving for

$$\mathbf{u} = T\,\mathbf{u} + \mathbf{c}, \qquad \text{where} \qquad T = -\,D^{-1}(L + U), \qquad \mathbf{c} = D^{-1}\mathbf{b}. \tag{10.61}$$

For the example (10.60), we recover the Jacobi iteration matrix by

$$T = -D^{-1}(L+U) = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}.$$

Deciding whether or not the Jacobi method converges for a specific matrix is not an easy task. However, it can be shown that Jacobi iteration will always converge for matrices that have large diagonal terms: the diagonally dominant matrices of Definition 10.33.

Theorem 10.42. *If A is strictly diagonally dominant, then the associated Jacobi iteration scheme converges.*

Proof: We shall prove that $\| T \|_\infty < 1$, and so Corollary 10.23 implies that T is a convergent matrix. The row sums of the Jacobi matrix $T = -D^{-1}(L+U)$ are, according to (10.56),

$$s_i = \sum_{j=1}^{n} | t_{ij} | = \frac{1}{| a_{ii} |} \sum_{i \neq j = 1}^{n} | a_{ij} | < 1 \tag{10.62}$$

because A is strictly diagonally dominant. Thus, $\| T \|_\infty = \max\{s_1, \dots, s_n\} < 1$, and the result follows. *Q.E.D.*

Example 10.43. Consider the linear system

$$\begin{aligned} 4x + y + w &= 1, \\ x + 4y + z + v &= 2, \\ y + 4z + w &= -1, \\ x + z + 4w + v &= 2, \\ y + w + 4v &= 1. \end{aligned}$$

The Jacobi method solves the respective equations for x, y, z, w, v, leading to the iterative scheme

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{4} y^{(k)} - \tfrac{1}{4} w^{(k)} + 1, \\ y^{(k+1)} &= -\tfrac{1}{4} x^{(k)} - \tfrac{1}{4} z^{(k)} - \tfrac{1}{4} v^{(k)} + 2, \\ z^{(k+1)} &= -\tfrac{1}{4} y^{(k)} - \tfrac{1}{4} w^{(k)} - 1, \\ w^{(k+1)} &= -\tfrac{1}{4} x^{(k)} - \tfrac{1}{4} z^{(k)} - \tfrac{1}{4} v^{(k)} + 2, \\ v^{(k+1)} &= -\tfrac{1}{4} y^{(k)} - \tfrac{1}{4} w^{(k)} + 1. \end{aligned}$$

The coefficient matrix of the original system

$$A = \begin{pmatrix} 4 & 1 & 0 & 1 & 0 \\ 1 & 4 & 1 & 0 & 1 \\ 0 & 1 & 4 & 1 & 0 \\ 1 & 0 & 1 & 4 & 1 \\ 0 & 1 & 0 & 1 & 4 \end{pmatrix},$$

is diagonally dominant, and so we are guaranteed that the Jacobi iterations will eventually converge to the solution. Indeed, the Jacobi scheme takes the iterative form (10.61), with

$$T = \begin{pmatrix} 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ -\frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix}.$$

Note that $\| T \|_\infty = \frac{3}{4} < 1$, and hence the convergence rate of the iterates to the solution is at least .75, which slightly overestimates the true convergence rate, as determined by the spectral radius $\rho(T) = .6124$. To obtain four decimal place accuracy in the solution, we anticipate[†] about $\log(.5 \times 10^{-4})/\log .6124 \approx 20$ iterations. Indeed, starting with the initial guess $x^{(0)} = y^{(0)} = z^{(0)} = w^{(0)} = v^{(0)} = 0$, the Jacobi iterates converge to the exact solution $x = -.1$, $y = .7$, $z = -.6$, $w = .7$, $v = -.1$, to four decimal places in exactly 20 iterations.

The Gauss–Seidel Method

The Gauss–Seidel method relies on a slightly more sophisticated implementation of the Jacobi process. To understand how it works, it will help to write out the Jacobi iteration scheme (10.57) in full detail:

$$\begin{aligned}
u_1^{(k+1)} &= \qquad\quad t_{12}\, u_2^{(k)} + t_{13}\, u_3^{(k)} + \;\cdots\; + t_{1,n-1}\, u_{n-1}^{(k)} + t_{1n}\, u_n^{(k)} + c_1, \\
u_2^{(k+1)} &= t_{21}\, u_1^{(k)} + \qquad\quad t_{23}\, u_3^{(k)} + \;\cdots\; + t_{2,n-1}\, u_{n-1}^{(k)} + t_{2n}\, u_n^{(k)} + c_2, \\
u_3^{(k+1)} &= t_{31}\, u_1^{(k)} + t_{32}\, u_2^{(k)} + \qquad\quad \cdots\; + t_{3,n-1}\, u_{n-1}^{(k)} + t_{3n}\, u_n^{(k)} + c_3, \\
&\;\;\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \ddots \qquad\qquad\qquad\qquad \ddots \qquad\quad \vdots \\
u_n^{(k+1)} &= t_{n1}\, u_1^{(k)} + t_{n2}\, u_2^{(k)} + t_{n3}\, u_3^{(k)} + \;\cdots\; + t_{n,n-1}\, u_{n-1}^{(k)} \qquad\quad + c_n,
\end{aligned} \tag{10.63}$$

where we are explicitly noting the fact that the diagonal entries of T vanish. Observe that we are using the entries of $\mathbf{u}^{(k)}$ to compute *all* of the updated values of $\mathbf{u}^{(k+1)}$. Presumably, if the iterates $\mathbf{u}^{(k)}$ are converging to the solution $\mathbf{u}^\star$, then their individual entries are also converging, and so each $u_j^{(k+1)}$ should be a better approximation to $u_j^\star$ than $u_j^{(k)}$ is. Therefore, if we begin the k^{th} Jacobi iteration by computing $u_1^{(k+1)}$ using the first equation, then we are tempted to use this new value instead of the previous, less accurate value $u_1^{(k)}$ in each of the subsequent equations. In particular, we employ the modified equation

$$u_2^{(k+1)} = t_{21}\, u_1^{(k+1)} + t_{23}\, u_3^{(k)} + \;\cdots\; + t_{1n}\, u_n^{(k)} + c_2$$

[†] If we were to use the matrix norm instead of the spectral radius, we would overestimate the proposed number of iterates to be $\log(.5 \times 10^{-4})/\log .75 \approx 34$.

to update the second component of our iterate. This more accurate value should then be used to update $u_3^{(k+1)}$, and so on.

The upshot of these considerations is the *Gauss–Seidel* iteration scheme

$$u_i^{(k+1)} = t_{i1}\, u_1^{(k+1)} + \cdots + t_{i,i-1}\, u_{i-1}^{(k+1)} + t_{i,i+1}\, u_{i+1}^{(k)} + \cdots + t_{in}\, u_n^{(k)} + c_i, \qquad (10.64)$$

named after Gauss (as usual!) and the German astronomer/mathematician Philipp von Seidel. At the k^{th} stage of the iteration, we use (10.64) to compute the updated entries $u_1^{(k+1)}$, $u_2^{(k+1)}, \dots, u_n^{(k+1)}$ in their numerical order. Once an entry has been updated, the new value is immediately used in all subsequent computations.

Example 10.44. For the linear system

$$3x + y - z = 3, \qquad x - 4y + 2z = -1, \qquad -2x - y + 5z = 2,$$

the Jacobi iteration method was given in (10.51). To obtain the corresponding Gauss–Seidel scheme we use updated values of x, y and z as they become available. Explicitly,

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{3}\, y^{(k)} + \tfrac{1}{3}\, z^{(k)} + 1, \\ y^{(k+1)} &= \tfrac{1}{4}\, x^{(k+1)} + \tfrac{1}{2}\, z^{(k)} + \tfrac{1}{4}, \\ z^{(k+1)} &= \tfrac{2}{5}\, x^{(k+1)} + \tfrac{1}{5}\, y^{(k+1)} + \tfrac{2}{5}. \end{aligned} \qquad (10.65)$$

The resulting iterates starting with $\mathbf{u}^{(0)} = \mathbf{0}$ are

$$\mathbf{u}^{(1)} = \begin{pmatrix} 1.0000 \\ 0.5000 \\ 0.9000 \end{pmatrix}, \quad \mathbf{u}^{(2)} = \begin{pmatrix} 1.1333 \\ 0.9833 \\ 1.0500 \end{pmatrix}, \quad \mathbf{u}^{(3)} = \begin{pmatrix} 1.0222 \\ 1.0306 \\ 1.0150 \end{pmatrix}, \quad \mathbf{u}^{(4)} = \begin{pmatrix} 0.9948 \\ 1.0062 \\ 0.9992 \end{pmatrix},$$

$$\mathbf{u}^{(5)} = \begin{pmatrix} 0.9977 \\ 0.9990 \\ 0.9989 \end{pmatrix}, \quad \mathbf{u}^{(6)} = \begin{pmatrix} 1.0000 \\ 0.9994 \\ 0.9999 \end{pmatrix}, \quad \mathbf{u}^{(7)} = \begin{pmatrix} 1.0001 \\ 1.0000 \\ 1.0001 \end{pmatrix}, \quad \mathbf{u}^{(8)} = \begin{pmatrix} 1.0000 \\ 1.0000 \\ 1.0000 \end{pmatrix}.$$

The iterations have converged to the solution, to 4 decimal places, after only 8 iterations — as opposed to the 11 iterations required by the Jacobi method. In this example, the Gauss–Seidel method is converging roughly 50% faster.

The Gauss–Seidel iteration scheme is particularly suited to implementation on a serial computer, since one can immediately replace each component $u_i^{(k)}$ by its updated value $u_i^{(k+1)}$, thereby also saving on storage in the computer's memory. In contrast, the Jacobi scheme requires us to retain all the old values $\mathbf{u}^{(k)}$ until all of the new values in $\mathbf{u}^{(k+1)}$ have been computed. Moreover, Gauss–Seidel typically (although not always) converges faster than Jacobi, making it the iterative algorithm of choice for serial processors. On the other hand, with the advent of parallel processing machines, variants of the Jacobi scheme have been making a comeback. Whereas Gauss–Seidel necessitates performing only one entry update at a time, the Jacobi method can be more easily parallelized.

What is Gauss–Seidel really up to? Let us rewrite the basic iterative equation (10.64) by multiplying by a_{ii} and moving the terms involving $\mathbf{u}^{(k+1)}$ to the left hand side. In view

of the formula (10.56) for the entries of T, the resulting equation is

$$a_{i1}\,u_1^{(k+1)} + \,\cdots\, + a_{i,i-1}\,u_{i-1}^{(k+1)} + a_{ii}\,u_i^{(k+1)} = -\,a_{i,i+1}\,u_{i+1}^{(k)} - \,\cdots\, - a_{in}\,u_n^{(k)} + b_i.$$

In matrix form, taking (10.59) into account, this reads

$$(L+D)\mathbf{u}^{(k+1)} = -\,U\,\mathbf{u}^{(k)} + \mathbf{b}, \tag{10.66}$$

and so can be viewed as a linear system of equations for $\mathbf{u}^{(k+1)}$ with lower triangular coefficient matrix $L+D$. Note that the fixed point of (10.66), namely $(L+D)\,\mathbf{u} = -\,U\,\mathbf{u}+\mathbf{b}$, coincides with the solution to the original system $A\,\mathbf{u} = (L+D+U)\,\mathbf{u} = \mathbf{b}$. The Gauss–Seidel procedure is merely implementing Forward Substitution to solve the lower triangular system (10.66) for the next iterate:

$$\mathbf{u}^{(k+1)} = -\,(L+D)^{-1}U\,\mathbf{u}^{(k)} + (L+D)^{-1}\,\mathbf{b}.$$

The latter is in our more usual iterative form

$$\mathbf{u}^{(k+1)} = \widetilde{T}\,\mathbf{u}^{(k)} + \widetilde{\mathbf{c}}, \qquad \text{where} \qquad \widetilde{T} = -\,(L+D)^{-1}U, \qquad \widetilde{\mathbf{c}} = (L+D)^{-1}\,\mathbf{b}.$$

Consequently, the convergence of the Gauss–Seidel iterates is governed by the spectral radius of its coefficient matrix $\widetilde{T}$.

For example, in the case of the coefficient matrix in Example 10.44, we have

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \qquad L+D = \begin{pmatrix} 3 & 0 & 0 \\ 1 & -4 & 0 \\ -2 & -1 & 5 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, the Gauss–Seidel coefficient matrix is

$$\widetilde{T} = -\,(L+D)^{-1}U = \begin{pmatrix} 0 & -0.3333 & 0.3333 \\ 0 & -0.0833 & 0.5833 \\ 0 & -0.1500 & 0.2500 \end{pmatrix}.$$

The matrix $\widetilde{T}$ has eigenvalues 0 and $0.0833 \pm 0.2444\,\mathrm{i}$, and hence its spectral radius is $\rho(\widetilde{T}) \approx 0.2582$. This is roughly the square of the Jacobi spectral radius of .5, and tell us that the Gauss-Seidel iterations will converge about twice as fast to the solution, in accordance with our earlier observation. Indeed, although examples exist where the Jacobi method converges faster, in many practical situation, the Gauss–Seidel scheme tends to converge roughly twice as fast.

General conditions guaranteeing the convergence of the Gauss–Seidel method are hard to establish. But, like the Jacobi method, diagonally dominant matrices are still handled well.

Theorem 10.45. *If A is strictly diagonally dominant, then the Gauss–Seidel iteration scheme converges.*

Proof: Let $\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^\star$ denote the k^{th} Gauss–Seidel error vector. As in (10.54), the error vectors satisfy the homogeneous iteration $\mathbf{e}^{(k+1)} = \widetilde{T}\,\mathbf{e}^{(k)}$. We write out this equation in components:

$$e_i^{(k+1)} = t_{i1}\, e_1^{(k+1)} + \,\cdots\, t_{i,i-1}\, e_{i-1}^{(k+1)} + t_{i,i+1}\, e_{i+1}^{(k)} + \,\cdots\, + t_{in}\, e_n^{(k)}. \tag{10.67}$$

Let

$$m^{(k)} = \| \mathbf{e}^{(k)} \|_\infty = \max\{\, |\, e_1^{(k)}\, |,\, \dots\, , |\, e_n^{(k)}\, | \,\} \tag{10.68}$$

denote the ∞ norm of the error vector. We claim that diagonal dominance of A implies that

$$m^{(k+1)} \le s\, m^{(k)}, \qquad \text{where} \qquad s = \| T \|_\infty < 1 \tag{10.69}$$

denotes the ∞ matrix norm of the Jacobi matrix, which, by (10.62), is less than 1. We infer that $m^{(k)} \le s^k\, m^{(0)} \to 0$ as $k \to \infty$. Thus, $\mathbf{e}^{(k)} \to \mathbf{0}$, demonstrating the theorem.

To prove (10.69), we use induction on $i = 1, \dots, n$. Our induction hypothesis is

$$|\, e_j^{(k+1)}\, | \le s\, m^{(k)} \le m^{(k)} \qquad \text{for} \qquad j = 1, \dots, i-1.$$

Moreover, by (10.68),

$$|\, e_j^{(k)}\, | \le m^{(k)} \qquad \text{for all} \qquad j = 1, \dots, n.$$

We use these two inequalities to estimate $|\, e_i^{(k+1)}\, |$ from (10.67):

$$\begin{aligned} |\, e_i^{(k+1)}\, | &\le |\, t_{i1}\, |\, |\, e_1^{(k+1)}\, | + \,\cdots\, + |\, t_{i,i-1}\, |\, |\, e_{i-1}^{(k+1)}\, | + |\, t_{i,i+1}\, |\, |\, e_{i+1}^{(k)}\, | + \,\cdots\, + |\, t_{in}\, |\, |\, e_n^{(k)}\, | \\ &\le \big(\, |\, t_{i1}\, | + \,\cdots\, + |\, t_{in}\, |\, \big)\, m^{(k)} = s_i\, m^{(k)} \le s\, m^{(k)}, \end{aligned}$$

which completes the induction step. As a result, the maximum

$$m^{(k+1)} = \max\{\, |\, e_1^{(k+1)}\, |,\, \dots\, , |\, e_n^{(k+1)}\, | \,\} \le s\, m^{(k)}$$

also satisfies the same bound, and hence (10.69) follows. *Q.E.D.*

Example 10.46. For the linear system considered in Example 10.43, the Gauss–Seidel iterations take the form

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{4}\, y^{(k)} - \tfrac{1}{4}\, w^{(k)} + 1, \\ y^{(k+1)} &= -\tfrac{1}{4}\, x^{(k+1)} - \tfrac{1}{4}\, z^{(k)} - \tfrac{1}{4}\, v^{(k)} + 2, \\ z^{(k+1)} &= -\tfrac{1}{4}\, y^{(k+1)} - \tfrac{1}{4}\, w^{(k)} - 1, \\ w^{(k+1)} &= -\tfrac{1}{4}\, x^{(k+1)} - \tfrac{1}{4}\, z^{(k+1)} - \tfrac{1}{4}\, v^{(k)} + 2, \\ v^{(k+1)} &= -\tfrac{1}{4}\, y^{(k+1)} - \tfrac{1}{4}\, w^{(k+1)} + 1. \end{aligned}$$

Starting with $x^{(0)} = y^{(0)} = z^{(0)} = w^{(0)} = v^{(0)} = 0$, the Gauss–Seidel iterates converge to the solution $x = -.1, y = .7, z = -.6, w = .7, v = -.1$, to four decimal places in 11 iterations, again roughly twice as fast as the Jacobi scheme.

Indeed, the convergence rate is governed by the corresponding Gauss-Seidel matrix $\widetilde{T}$, which is

$$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 \\ 1 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -0.2500 & 0 & -0.2500 & 0 \\ 0 & 0.0625 & -0.2500 & 0.0625 & -0.2500 \\ 0 & -0.0156 & 0.0625 & -0.2656 & 0.0625 \\ 0 & 0.0664 & -0.0156 & 0.1289 & -0.2656 \\ 0 & -0.0322 & 0.0664 & -0.0479 & 0.1289 \end{pmatrix}.$$

Its spectral radius is $\rho(\widetilde{T}) = .3936$, which is, as in the previous example, approximately the square of the spectral radius of the Jacobi coefficient matrix. This explains its doubly fast rate of convergence.

Successive Over–Relaxation (*SOR*)

As we know, the smaller the spectral radius (or matrix norm) of the coefficient matrix, the faster the convergence of the iterative method. The goal of researchers in numerical linear algebra is to design new methods for accelerating the convergence. In his 1950 thesis, the American mathematician David Young discovered a simple modification of the Jacobi and Gauss–Seidel methods that can, in many common applications, lead to a dramatic speed up in the rate of convergence. The method, known as *successive over-relaxation*, and often abbreviated as SOR, has become the iterative method of choice in many modern applications. In this subsection, we give a brief overview of the SOR iterative scheme.

In practice, designing the optimal iterative algorithm to solve a given linear system is as hard a solving the system itself. Therefore, one relies on a few tried and true techniques for building a good iterative scheme that works in a range of examples. Every decomposition

$$A = M - N \tag{10.70}$$

of the coefficient matrix of the system $A\,\mathbf{u} = \mathbf{b}$ into the difference of two matrices leads to an equivalent system of the form

$$M\,\mathbf{u} = N\,\mathbf{u} + \mathbf{b}. \tag{10.71}$$

Provided we take M to be invertible, we can rewrite the system in the fixed point form

$$\mathbf{u} = M^{-1}N\,\mathbf{u} + M^{-1}\mathbf{b} = T\,\mathbf{u} + \mathbf{c}, \qquad \text{where} \qquad T = M^{-1}N, \quad \mathbf{c} = M^{-1}\mathbf{b}.$$

Now, we are free to choose any such M, which then specifies $N = A - M$ uniquely. However, for the resulting iterative scheme $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{c}$ to be practical we must arrange that

(a) $T = M^{-1}N$ is a convergent matrix, and

(b) M can be easily inverted.

The second requirement ensures that the iterative equations

$$M\,\mathbf{u}^{(k+1)} = N\,\mathbf{u}^{(k)} + \mathbf{b} \tag{10.72}$$

can be solved for $\mathbf{u}^{(k+1)}$ with minimal computational effort. Typically, this requires that M be either a diagonal matrix, in which case the inversion is immediate, or upper or lower triangular, in which case one employs back or forward substitution to solve for $\mathbf{u}^{(k+1)}$.

With this in mind, we now introduce the SOR method. It relies on a slight generalization of the Gauss–Seidel decomposition (10.66) of the matrix into lower plus diagonal and upper triangular parts. The starting point is to write

$$A = L + D + U = \big[\, L + \alpha\, D \,\big] - \big[\, (\alpha - 1)\, D - U \,\big], \tag{10.73}$$

where $0 \neq \alpha$ is an adjustable scalar parameter. We decompose the system $A\,\mathbf{u} = \mathbf{b}$ as

$$(L + \alpha\, D)\mathbf{u} = \big[\, (\alpha - 1)\, D - U \,\big]\mathbf{u} + \mathbf{b}. \tag{10.74}$$

It turns out to be slightly more convenient to divide (10.74) through by α, and write the resulting iterative system in the form

$$(\omega\, L + D)\mathbf{u}^{(k+1)} = \big[\, (1 - \omega)\, D - \omega\, U \,\big]\mathbf{u}^{(k)} + \omega\, \mathbf{b}, \tag{10.75}$$

where $\omega = 1/\alpha$ is called the *relaxation parameter*. Assuming, as usual, that all diagonal entries of A are nonzero, the matrix $\omega\, L + D$ is an invertible lower triangular matrix, and so we can use forward substitution to solve the iterative system (10.75) to recover $\mathbf{u}^{(k+1)}$. The explicit formula for its i^{th} entry is

$$\begin{aligned} u_i^{(k+1)} = \omega\, t_{i1}\, u_1^{(k+1)} + \,\cdots\, + \omega\, t_{i,i-1}\, u_{i-1}^{(k+1)} + (1 - \omega)\, u_i^{(k)} + \\ + \omega\, t_{i,i+1}\, u_{i+1}^{(k)} + \,\cdots\, + \omega\, t_{in}\, u_n^{(k)} + \omega\, c_i, \end{aligned} \tag{10.76}$$

where t_{ij} and c_i denote the original Jacobi values (10.56). As in the Gauss–Seidel approach, we update the entries $u_i^{(k+1)}$ in numerical order $i = 1, \dots, n$. Thus, to obtain the SOR scheme (10.76), we merely multiply the right hand side of the Gauss–Seidel scheme (10.64) by the adjustable relaxation parameter ω and append the diagonal term $(1 - \omega)\, u_i^{(k)}$. In particular, if we set $\omega = 1$, then the SOR method reduces to the Gauss–Seidel method. Choosing $\omega < 1$ leads to an *under-relaxed* method, while $\omega > 1$, known as *over-relaxation*, is the choice that works in most practical instances.

To analyze the convergence rate of the SOR scheme (10.75), we rewrite it in the fixed point form

$$\mathbf{u}^{(k+1)} = T_\omega\, \mathbf{u}^{(k)} + \mathbf{c}_\omega, \tag{10.77}$$

where

$$T_\omega = (\omega\, L + D)^{-1} \big[\, (1 - \omega)\, D - \omega\, U \,\big], \qquad \mathbf{c}_\omega = (\omega\, L + D)^{-1}\, \omega\, \mathbf{b}. \tag{10.78}$$

The rate of convergence of the SOR method is governed by the spectral radius of its coefficient matrix T_ω. The goal is to choose the relaxation parameter ω so as to make the spectral radius of T_ω as small as possible. As we will see, a clever choice of ω will result in a dramatic speed up in the convergence of the iterative method. Before stating some general facts (albeit without proof) let us analyze a simple example.

Example 10.47. Consider the matrix $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, which we write as $L+D+U$, where

$$L = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

Jacobi iteration uses the coefficient matrix $T = -D^{-1}(L+U) = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$. The Jacobi spectral radius is $\rho(T) = .5$, and hence it takes, on average, roughly $3.3 \approx -1/\log_{10} .5$ iterations to produce each new decimal place of accuracy in the solution.

The SOR scheme (10.75) takes the explicit form

$$\begin{pmatrix} 2 & 0 \\ -\omega & 2 \end{pmatrix} \mathbf{u}^{(k+1)} = \begin{pmatrix} 2(1-\omega) & \omega \\ 0 & 2\,(1-\omega) \end{pmatrix} \mathbf{u}^{(k)} + \omega\, \mathbf{b},$$

where Gauss–Seidel is the particular case $\omega = 1$. The SOR coefficient matrix is

$$T_\omega = \begin{pmatrix} 2 & 0 \\ -\omega & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2\,(1-\omega) & \omega \\ 0 & 2\,(1-\omega) \end{pmatrix} = \begin{pmatrix} 1-\omega & \frac{1}{2}\,\omega \\ \frac{1}{2}\,\omega(1-\omega) & \frac{1}{4}\,(2-\omega)^2 \end{pmatrix}.$$

To compute the eigenvalues of T_ω, we form its characteristic equation

$$\begin{aligned} 0 = \det(T_\omega - \lambda\, \mathrm{I}\,) &= \lambda^2 - \big(\,2 - 2\omega + \tfrac{1}{4}\,\omega^2\,\big)\lambda + (1-\omega)^2 \\ &= (\lambda + \omega - 1)^2 - \tfrac{1}{4}\,\lambda\omega^2. \end{aligned} \tag{10.79}$$

Our goal is to choose ω so that

(a) Both eigenvalues are less than 1 in modulus, so $|\,\lambda_1\,|, |\,\lambda_2\,| < 1$. This is the minimal requirement for convergence of the method.

(b) The largest eigenvalue (in modulus) is as small as possible. This will give the smallest spectral radius for T_ω and hence the fastest convergence rate.

By (8.26), the product of the two eigenvalues is the determinant,

$$\lambda_1\,\lambda_2 = \det T_\omega = (1-\omega)^2.$$

If $\omega \le 0$ or $\omega \ge 2$, then $\det A \ge 1$, and hence least one of the eigenvalues would have modulus larger than 1. Thus, in order to ensure convergence, we must require $0 < \omega < 2$. For Gauss–Seidel, at $\omega = 1$, the eigenvalues are $\lambda_1 = 0$, $\lambda_2 = \frac{1}{4}$, and the spectral radius is $\rho(T_1) = .25$. This is exactly the square of the Jacobi spectral radius, and hence the Gauss–Seidel iterates converge twice as fast — it only takes, on average, about 1.65 Gauss–Seidel iterations to produce a new decimal place of accuracy. It can be shown (Exercise ■) that as ω increases above 1, the two eigenvalues move together, the larger one decreasing in size. They are equal when

$$\omega = \omega_\star = 8 - 4\sqrt{3} \;\approx\; 1.07.$$

At that point, $\lambda_1 = \lambda_2 = .07 = \rho(T_\omega)$, which is the convergence rate of the optimal[†] SOR scheme. Each iteration produces slightly more than one new decimal place in the solution, which represents a significant improvement over the Gauss–Seidel convergence rate of .25. It takes about twice as many Gauss–Seidel iterations (and four times as many Jacobi iterations) to produce the same accuracy as this optimal SOR method.

[†] In Exercise ■, the reader is asked to complete the proof of optimality.

Of course, in such a simple 2×2 example, it is not so surprising that we can construct the optimal relaxation parameter by hand. In his 1950 thesis, cf. [**154**], Young found the optimal value of the relaxation parameter for a broad class of matrices that includes most of those arising in the finite difference and finite element numerical solutions to ordinary and partial differential equations. For the matrices in Young's class, the Jacobi eigenvalues occur in signed pairs. If $\pm\mu$ are a pair of eigenvalues for the Jacobi method, then the corresponding eigenvalues of the SOR iteration matrix satisfy the quadratic equation

$$(\lambda + \omega - 1)^2 = \lambda\, \omega^2\, \mu^2. \tag{10.80}$$

If $\omega = 1$, so we have standard Gauss–Seidel, then $\lambda^2 = \lambda\, \mu^2$, and so the corresponding Gauss–Seidel eigenvalues are $\lambda = 0$, $\lambda = \mu^2$. The Gauss–Seidel spectral radius is therefore the square of the Jacobi spectral radius, and so (at least for matrices in the Young class) its iterates converge twice as fast. The quadratic equation (10.80) has the same properties as in the 2×2 version (10.79) (which corresponds to the case $\mu = \frac{1}{2}$), and hence the optimal value of ω will be the one at which the two roots are equal,

$$\lambda_1 = \lambda_2 = \omega - 1, \qquad \text{which occurs when} \qquad \omega = \frac{2 - 2\sqrt{1 - \mu^2}}{\mu^2} = \frac{2}{1 + \sqrt{1 - \mu^2}}\,.$$

Therefore, if $\rho_J = \max |\, \mu\, |$ denotes the spectral radius of the Jacobi method, then the Gauss–Seidel has spectral radius $\rho_{GS} = \rho_J^2$, while the SOR method with optimal relaxation parameter

$$\omega_\star = \frac{2}{1 + \sqrt{1 - \rho_J^2}}, \qquad \text{has spectral radius} \qquad \rho_\star = \omega_\star - 1. \tag{10.81}$$

For example, if $\rho_J = .99$, which is quite slow convergence (but common for iterative solutions of partial differential equations), then $\rho_{GS} = 0.9801$, which is twice as fast, but still quite slow, while SOR with $\omega_\star = 1.7527$ has $\rho_\star = 0.7527$, which is dramatically faster. Indeed, since $\rho_\star \approx (\rho_{GS})^{14} \approx (\rho_J)^{28}$, it takes about 14 Gauss–Seidel (and 28 Jacobi) iterations to produce the same accuracy as one SOR step. The fact that such a simple idea can have such a dramatic effect on the convergence rate is amazing.

Conjugate Gradients

So far, we have established two broad classes of algorithms for solving linear systems. The first, the *direct methods*, based on some version of Gaussian elimination or matrix factorization, eventually[†] obtain the solution, but must be carried through to completion before any useful information is obtained. The alternative, *iterative methods* discussed in the present chapter, lead to closer and closer approximations of the solution, but never reach the actual value exactly. One might ask whether there are algorithms that combine the best of both: *semi-direct methods* that give closer and closer approximations to the solution, but are guaranteed to eventually terminate with the exact solution in hand.

[†] This assumes that we are dealing with a fully accurate implementation, i.e., without round-off or other numerical error. For this discussion, numerical instability will be left aside as a separate, albeit ultimately important, issue.

For instance, one might ask for an algorithm that successively computes each entry of the solution vector $\mathbf{u}_\star$. This seems unlikely, but if we recall that the entries of the solution are merely its coordinates with respect to the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_n$, then one might try instead to compute the coordinates $t_1, \dots, t_n$ of $\mathbf{u}_\star = t_1 \mathbf{v}_1 + \cdots + t_n \mathbf{v}_n$ with respect to some basis that is specially adapted to the linear system. Ideally, $\mathbf{v}_1, \dots, \mathbf{v}_n$ should be an orthogonal basis — but orthogonality with respect to the standard Euclidean dot product is not typically relevant. A better idea is to arrange that the basis be orthogonal with respect to an inner product that is adapted to the system under consideration. In particular, if the linear system to be solved takes the form

$$K\,\mathbf{u} = \mathbf{f}, \tag{10.82}$$

in which the coefficient matrix is *positive definite*, as occurs in many applications, then orthogonality with respect to the induced inner product

$$\langle\!\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle\!\rangle = \mathbf{v}^T K\,\mathbf{w} = \mathbf{v} \cdot K\,\mathbf{w} \tag{10.83}$$

is very natural. Vectors that are orthogonal with respect to the inner product induced by the coefficient matrix K are known as *conjugate vectors*, which explain half the name of the conjugate gradient algorithm, first introduced by Hestenes and Stiefel, [**78**].

The term "gradient" stems from the minimization principle. According to Theorem 4.1, the solution $\mathbf{u}_\star$ to the positive definite linear system (10.82) is the unique minimizer of the quadratic function[†]

$$p(\mathbf{u}) = \tfrac{1}{2}\,\mathbf{u}^T K\,\mathbf{u} - \mathbf{u}^T \mathbf{f}. \tag{10.84}$$

Thus, one way to solve the system is to minimize $p(\mathbf{u})$. Suppose we find ourselves at a point $\mathbf{u}$ which is not the minimizer. In which direction should we travel to find $\mathbf{u}_\star$? The key result from multivariable calculus is that the gradient vector $\nabla p(\mathbf{u})$ of a (nonlinear) function points in the direction of its steepest increase at the point, while its negative $-\nabla p(\mathbf{u})$ points in the direction of steepest decrease. Our discussion of gradient flow systems (9.19) used the same idea; full details appear in Section 19.3. For the particular quadratic function (10.84), its negative gradient is easily found:

$$-\nabla p(\mathbf{u}) = \mathbf{f} - K\,\mathbf{u} = \mathbf{r},$$

where $\mathbf{r}$ is known as the *residual vector* for the point $\mathbf{u}$. Note that $\mathbf{r} = \mathbf{0}$ if and only if $\mathbf{u} = \mathbf{u}_\star$ is the solution, and so the size of $\mathbf{r}$ measures, in a certain sense, how accurately $\mathbf{u}$ comes to solving the system. Moreover, the residual vector indicates the direction of steepest decrease in the quadratic function, and is thus a good choice of direction to head off in search of the true minimizer.

The initial result is the *gradient descent algorithm*, in which each successive approximation $\mathbf{u}_k$ to the solution is obtained by going a certain distance in the residual direction:

$$\mathbf{u}_{k+1} = \mathbf{u}_k + t_k\,\mathbf{r}_k, \qquad \text{where} \qquad \mathbf{r}_k = \mathbf{f} - K\,\mathbf{u}_k. \tag{10.85}$$

[†] Here, we include an irrelevant factor of $\frac{1}{2}$ for later convenience.

The scalar factor t_k can be specified by the requirement that $p(\mathbf{u}_{k+1})$ is as small as possible; in Exercise ■ you are asked to find this value. A second option is to make the residual vector at $\mathbf{u}_{k+1}$ as small as possible. The initial guess $\mathbf{u}_0$ for the solution can be chosen as desired, with $\mathbf{u}_0 = \mathbf{0}$ the default choice. Gradient descent is a reasonable algorithm, and will lead to the solution in favorable situations. It is also used to minimize more general nonlinear functions. However, in many circumstances, the iterative method based on gradient descent can take an exceedingly long time to converge to an accurate approximation to the solution, and so is often not a competitive algorithm.

However, if we supplement the gradient descent idea by the use of conjugate vectors, we are led to a very powerful semi-direct solution algorithm. We shall construct the solution $\mathbf{u}_\star$ by successive approximation, with the k^{th} iterate having the form

$$\mathbf{u}_k = t_1\,\mathbf{v}_1 + \,\cdots\, + t_k\,\mathbf{v}_k, \qquad \text{so that} \qquad \mathbf{u}_{k+1} = \mathbf{u}_k + t_{k+1}\,\mathbf{v}_{k+1},$$

where, as advertised, the conjugate vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a K–orthogonal basis. The secret is not to try to specify the conjugate basis vectors in advance, but rather to successively construct them during the course of the algorithm. We begin, merely for convenience, with an initial guess $\mathbf{u}_0 = \mathbf{0}$ for the solution. The residual vector $\mathbf{r}_0 = \mathbf{f} - K\,\mathbf{u}_0 = \mathbf{f}$ indicates the direction of steepest decrease of $p(\mathbf{u})$ at $\mathbf{u}_0$, and we update our original guess by moving in this direction, taking $\mathbf{v}_1 = \mathbf{r}_0 = \mathbf{f}$ as our first conjugate direction. The next iterate is $\mathbf{u}_1 = \mathbf{u}_0 + t_1\,\mathbf{v}_1 = t_1\,\mathbf{v}_1$, and we choose the parameter t_1 so that the corresponding residual vector

$$\mathbf{r}_1 = \mathbf{f} - K\,\mathbf{u}_1 = \mathbf{r}_0 - t_1\,K\,\mathbf{v}_1 \tag{10.86}$$

is as close to $\mathbf{0}$ (in the Euclidean norm) as possible. This occurs when $\mathbf{r}_1$ is orthogonal to $\mathbf{r}_0$ (why?), and so we require

$$0 = \mathbf{r}_0 \cdot \mathbf{r}_1 = \| \mathbf{r}_0 \|^2 - t_1\,\mathbf{r}_0 \cdot K\,\mathbf{v}_1 = \| \mathbf{r}_0 \|^2 - t_1 \langle\!\langle\, \mathbf{r}_0 \,;\, \mathbf{v}_1 \,\rangle\!\rangle = \| \mathbf{r}_0 \|^2 - t_1 \langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle. \tag{10.87}$$

Therefore we set

$$t_1 = \frac{\| \mathbf{r}_0 \|^2}{\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle} \qquad \text{and so} \qquad \mathbf{u}_1 = \mathbf{u}_0 + \frac{\| \mathbf{r}_0 \|^2}{\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle}\,\mathbf{v}_1 \tag{10.88}$$

is the new approximation to the solution.

Note: We will consistently use $\| \mathbf{v} \|$ to denote the standard Euclidean norm, and $\langle\!\langle\, \mathbf{v} \,;\, \mathbf{w} \,\rangle\!\rangle$ the adapted inner product (10.83), which has its own norm $\sqrt{\langle\!\langle\, \mathbf{v} \,;\, \mathbf{v} \,\rangle\!\rangle}$.

The gradient descent algorithm would tell us to update $\mathbf{u}_1$ by moving in the residual direction $\mathbf{r}_1$. But in the conjugate gradient algorithm, we choose a direction $\mathbf{v}_2$ which is conjugate, meaning K–orthogonal to the first direction $\mathbf{v}_1 = \mathbf{r}_0$. Thus, we slightly modify the residual direction by setting $\mathbf{v}_2 = \mathbf{r}_1 + s_1\,\mathbf{v}_1$, where the scalar factor s_1 is determined by the orthogonality requirement

$$0 = \langle\!\langle\, \mathbf{v}_2 \,;\, \mathbf{v}_1 \,\rangle\!\rangle = \langle\!\langle\, \mathbf{v}_2 \,;\, \mathbf{r}_1 + s_1\,\mathbf{v}_1 \,\rangle\!\rangle = \langle\!\langle\, \mathbf{r}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle + s_1 \langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle, \quad \text{so} \quad s_1 = -\frac{\langle\!\langle\, \mathbf{r}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle}{\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle}\,.$$

Now, using (10.87) twice,

$$\langle\!\langle\, \mathbf{r}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle = \mathbf{r}_1 \cdot K\mathbf{v}_1 = \frac{1}{t_1}\,\mathbf{r}_1 \cdot (\mathbf{r}_0 - \mathbf{r}_1) = -\,\frac{1}{t_1}\,\|\,\mathbf{r}_1\,\|^2,$$
$$\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle = \mathbf{v}_1 \cdot K\mathbf{v}_1 = \frac{1}{t_1}\,\|\,\mathbf{r}_0\,\|^2,$$

and therefore the second conjugate direction is

$$\mathbf{v}_2 = \mathbf{r}_1 + s_1\,\mathbf{v}_1, \qquad \text{where} \qquad s_1 = \frac{\|\,\mathbf{r}_1\,\|^2}{\|\,\mathbf{r}_0\,\|^2}\,.$$

We then update

$$\mathbf{u}_2 = \mathbf{u}_1 + t_2\,\mathbf{v}_2 = t_1\,\mathbf{v}_1 + t_2\,\mathbf{v}_2$$

so as to make the corresponding residual vector

$$\mathbf{r}_2 = \mathbf{f} - K\mathbf{u}_2 = \mathbf{r}_1 - t_2\,K\mathbf{v}_2$$

as small as possible, which is accomplished by requiring it to be orthogonal to $\mathbf{r}_1$. Thus, using the K–orthogonality of $\mathbf{v}_1$ and $\mathbf{v}_2$,

$$0 = \mathbf{r}_1 \cdot \mathbf{r}_2 = \|\,\mathbf{r}_1\,\|^2 - t_2\,\mathbf{r}_1 \cdot K\mathbf{v}_2 = \|\,\mathbf{r}_1\,\|^2 - t_2\,\langle\!\langle\, \mathbf{v}_2 \,;\, \mathbf{v}_2 \,\rangle\!\rangle, \quad \text{and so} \quad t_2 = \frac{\|\,\mathbf{r}_1\,\|^2}{\langle\!\langle\, \mathbf{v}_2 \,;\, \mathbf{v}_2 \,\rangle\!\rangle}\,.$$

Continuing in this manner, at the k^{th} stage, we have already constructed the conjugate vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$, and the solution approximation $\mathbf{u}_k$ as a suitable linear combination of them. The next conjugate direction is given by

$$\mathbf{v}_{k+1} = \mathbf{r}_k + s_k\,\mathbf{v}_k, \qquad \text{where} \qquad s_k = \frac{\|\,\mathbf{r}_k\,\|^2}{\|\,\mathbf{r}_{k-1}\,\|^2} \tag{10.89}$$

is a result of the K–orthogonality requirement $\langle\!\langle\, \mathbf{v}_i \,;\, \mathbf{v}_k \,\rangle\!\rangle = 0$ for $i < k$. The updated solution approximation

$$\mathbf{u}_{k+1} = \mathbf{u}_k + t_{k+1}\,\mathbf{v}_{k+1} \qquad \text{where} \qquad t_{k+1} = \frac{\|\,\mathbf{r}_k\,\|^2}{\langle\!\langle\, \mathbf{v}_{k+1} \,;\, \mathbf{v}_{k+1} \,\rangle\!\rangle}\,, \tag{10.90}$$

is then fixed so as to make the corresponding residual $\mathbf{r}_{k+1} = \mathbf{f} - K\mathbf{u}_{k+1} = \mathbf{r}_k - t_{k+1}\,K\mathbf{v}_{k+1}$ as small as possible, which requires that it be orthogonal to $\mathbf{r}_k$. Starting with an initial guess $\mathbf{u}_0$, the iterative equations (10.89), (10.90) implement the complete conjugate gradient algorithm. Note that the only matrix operation required is a multiplication $K\mathbf{v}_k$ in the computation of t_k; all the other operations are fast Euclidean dot products. Unlike Gaussian elimination, the method produces a sequence of successive approximations $\mathbf{u}_1, \mathbf{u}_2, \dots$ to the solution $\mathbf{u}_\star$, and so the method can be stopped when a desired solution accuracy is reached. On the other hand, unlike purely iterative methods, the algorithm does eventually terminate at the exact solution, because, as remarked at the outset, there are at most n conjugate directions, forming an K orthogonal basis of $\mathbb{R}^n$. Therefore,

Conjugate Gradient Method for Solving $K\,\mathbf{u} = \mathbf{f}$

start

  choose $\mathbf{u}_0$, e.g. $\mathbf{u}_0 = \mathbf{0}$

  for $k = 1$ to m

    set $\mathbf{r}_{k-1} = \mathbf{f} - K\,\mathbf{u}_{k-1}$

    if $k = 1$ set $\mathbf{v}_1 = \mathbf{r}_0$

      else set $\mathbf{v}_k = \mathbf{r}_{k-1} + \dfrac{\| \mathbf{r}_{k-1} \|^2}{\| \mathbf{r}_{k-2} \|^2}\, \mathbf{v}_{k-1}$

    set $\mathbf{u}_k = \mathbf{u}_{k-1} + \dfrac{\| \mathbf{r}_{k-1} \|^2}{\mathbf{v}_k \cdot K\,\mathbf{v}_k}\, \mathbf{v}_k$

  next k

end

$\mathbf{u}_n = t_1\,\mathbf{v}_1 + \cdots + t_n\,\mathbf{v}_n = \mathbf{u}_\star$ must be the solution since its residual $\mathbf{r}_n = \mathbf{f} - K\,\mathbf{u}_n$ is orthogonal to all the $\mathbf{v}_i$, and hence equal to $\mathbf{0}$.

A pseudocode program is attached; at each stage $\mathbf{u}_k$ represents the updated approximation to the solution. The initial guess $\mathbf{u}_0$ can be chosen by the user, with $\mathbf{u}_0 = \mathbf{0}$ the default. The iteration number $m \leq n$ can be chosen by the user in advance; alternatively, one can impose a stopping criterion based on the size of the residual vector, $\| \mathbf{r}_{k-1} \|$, or, alternatively, the distance between successive iterates, $\| \mathbf{u}_k - \mathbf{u}_{k-1} \|$. If the process is carried on to the bitter end, i.e., for $m = n$, then, in the absence of round-off errors, the result is the exact solution to the system.

Example 10.48. Consider the linear system $K\,\mathbf{u} = \mathbf{f}$ with

$$K = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \qquad \mathbf{f} = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}.$$

The exact solution is $\mathbf{u}_\star = (\,2, 5, -6\,)^T$. In order to implement the method of conjugate gradients, we start with the initial guess $\mathbf{u}_0 = (\,0, 0, 0\,)^T$. The corresponding residual vector is merely $\mathbf{r}_0 = \mathbf{f} - K\,\mathbf{u}_0 = (\,1, 2, -1\,)^T$. The first conjugate direction is $\mathbf{v}_1 = \mathbf{r}_0 = \mathbf{f} = (\,1, 2, -1\,)^T$, and we use (10.88) to obtain the updated approximation to the solution

$$\mathbf{u}_1 = \mathbf{u}_0 + \frac{\| \mathbf{r}_0 \|^2}{\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle}\, \mathbf{v}_1 = \frac{6}{4} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \end{pmatrix},$$

noting that $\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_1 \,\rangle\!\rangle = \mathbf{v}_1 \cdot K\,\mathbf{v}_1 = 4$. In the next stage of the algorithm, we compute the

corresponding residual $\mathbf{r}_1 = \mathbf{f} - K\,\mathbf{u}_1 = \left(-\frac{1}{2}, -1, -\frac{5}{2}\right)^T$. The conjugate direction is

$$\mathbf{v}_2 = \mathbf{r}_1 + \frac{\| \mathbf{r}_1 \|^2}{\| \mathbf{r}_0 \|^2}\, \mathbf{v}_1 = \begin{pmatrix} -\frac{1}{2} \\ -1 \\ -\frac{5}{2} \end{pmatrix} + \frac{\frac{15}{2}}{6} \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ -\frac{15}{4} \end{pmatrix}.$$

We note that, as designed, it satisfies the conjugacy condition $\langle\!\langle\, \mathbf{v}_1 \,;\, \mathbf{v}_2 \,\rangle\!\rangle = \mathbf{v}_1 \cdot K\,\mathbf{v}_2 = 0$. Each entry of the new approximation

$$\mathbf{u}_2 = \mathbf{u}_1 + \frac{\| \mathbf{r}_1 \|^2}{\langle\!\langle\, \mathbf{v}_2 \,;\, \mathbf{v}_2 \,\rangle\!\rangle}\, \mathbf{v}_2 = \begin{pmatrix} \frac{3}{2} \\ 3 \\ -\frac{3}{2} \end{pmatrix} + \frac{\frac{15}{2}}{\frac{27}{4}} \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ -\frac{15}{4} \end{pmatrix} = \begin{pmatrix} \frac{7}{3} \\ \frac{14}{3} \\ -\frac{17}{3} \end{pmatrix} = \begin{pmatrix} 2.333\ldots \\ 4.666\ldots \\ -5.666\ldots \end{pmatrix}$$

is within a $\frac{1}{3}$ of the exact solution $\mathbf{u}_\star$.

Since we are dealing with a 3×3 system, we will recover the exact solution by one more iteration of the algorithm. The new residual is $\mathbf{r}_2 = \mathbf{f} - K\,\mathbf{u}_2 = \left(-\frac{4}{3}, \frac{2}{3}, 0\right)^T$. The last conjugate direction is

$$\mathbf{v}_3 = \mathbf{r}_2 + \frac{\| \mathbf{r}_2 \|^2}{\| \mathbf{r}_1 \|^2}\, \mathbf{v}_2 = \begin{pmatrix} -\frac{4}{3} \\ \frac{2}{3} \\ 0 \end{pmatrix} + \frac{\frac{20}{9}}{\frac{15}{2}} \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ -\frac{15}{4} \end{pmatrix} = \begin{pmatrix} -\frac{10}{9} \\ \frac{10}{9} \\ -\frac{10}{9} \end{pmatrix},$$

which, as you can check, is conjugate to both $\mathbf{v}_1$ and $\mathbf{v}_2$. The solution is obtained from

$$\mathbf{u}_3 = \mathbf{u}_2 + \frac{\| \mathbf{r}_2 \|^2}{\langle\!\langle\, \mathbf{v}_3 \,;\, \mathbf{v}_3 \,\rangle\!\rangle}\, \mathbf{v}_3 = \begin{pmatrix} \frac{7}{3} \\ \frac{14}{3} \\ -\frac{17}{3} \end{pmatrix} + \frac{\frac{20}{9}}{\frac{200}{27}} \begin{pmatrix} -\frac{10}{9} \\ \frac{10}{9} \\ -\frac{10}{9} \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ -6 \end{pmatrix}.$$

Of course, in larger examples, one would not carry through the algorithm to the bitter end — indeed reverting to ordinary Gaussian elimination is probably a better strategy in that case — since a decent approximation to the solution is typically obtained with only a few iterations. The result is a substantial saving in computational time and effort to produce a reasonable approximation to the solution.

10.6. Numerical Computation of Eigenvalues.

The importance of the eigenvalues of a square matrix for both continuous and discrete dynamical systems has been amply demonstrated in this chapter and its predecessor. However, finding the eigenvalues and associated eigenvectors is not an easy task. The classical method of constructing the characteristic equation of the matrix through the determinantal formula, then solving the resulting polynomial equation for the eigenvalues, and finally producing the eigenvectors by solving the associated homogeneous linear system, is hopelessly inefficient, fraught with difficulty and numerical dangers. We are in need of a completely new approach if we have any hopes of designing efficient, numerical approximation schemes for computing eigenvalues and eigenvectors.

In this section, we present a few of the most basic numerical schemes for accurately computing eigenvalues and eigenvectors of matrices. The most direct are based on the connections between the eigenvalues and the high powers of a matrix. A more sophisticated technique is based on the $Q\,R$ factorization that we learned in Section 5.3, and will be presented at the end of the section.

The Power Method

We have already noted the role played by the eigenvalues and eigenvectors in the solution to linear iterative systems. Now we are going to turn the tables, and use the iterative system as a mechanism for approximating the eigenvalues, or, more correctly, selected eigenvalues of the coefficient matrix. The resulting computational procedure is known as the *power method*.

We assume, for simplicity, that A is a complete[†] $n \times n$ matrix. Let $\mathbf{v}_1, \dots, \mathbf{v}_n$ denote the eigenvector basis, and $\lambda_1, \dots, \lambda_n$ the corresponding eigenvalues. As we have learned, the solution to the linear iterative system

$$\mathbf{v}^{(k+1)} = A\,\mathbf{v}^{(k)}, \qquad \mathbf{v}^{(0)} = \mathbf{v}, \tag{10.91}$$

is obtained by multiplying the initial vector $\mathbf{v}$ by the successive powers of the coefficient matrix: $\mathbf{v}^{(k)} = A^k\,\mathbf{v}$. If we write the initial vector in terms of the eigenvector basis

$$\mathbf{v} = c_1\,\mathbf{v}_1 + \,\cdots\, + c_n\,\mathbf{v}_n, \tag{10.92}$$

then the solution takes the explicit form given in Theorem 10.4, namely

$$\mathbf{v}^{(k)} = A^k\,\mathbf{v} = c_1\,\lambda_1^k\,\mathbf{v}_1 + \,\cdots\, + c_n\,\lambda_n^k\,\mathbf{v}_n. \tag{10.93}$$

Suppose further that A has a single dominant *real* eigenvalue, λ_1, that is larger than any other in magnitude, so

$$|\,\lambda_1\,| > |\,\lambda_j\,| \qquad \text{for all} \qquad j > 1. \tag{10.94}$$

The largest eigenvalue will completely dominate the iteration (10.93). Indeed, since

$$|\,\lambda_1\,|^k \gg |\,\lambda_j\,|^k \qquad \text{for all} \;\; j > 1 \;\; \text{and all} \;\; k \gg 0,$$

the first term in the iterative formula (10.93) will eventually be much larger than all the rest, and so, provided $c_1 \neq 0$,

$$\mathbf{v}^{(k)} \;\approx\; c_1\,\lambda_1^k\,\mathbf{v}_1 \qquad \text{for} \qquad k \gg 0.$$

Therefore, under the assumption (10.94), the solution to the iterative system (10.91) will, almost always, end up being a multiple of the first eigenvector of the coefficient matrix.

[†] This is not a very severe restriction. Most matrices are complete. Moreover, perturbations caused by numerical inaccuracies will almost always make an incomplete matrix complete.

k	$\mathbf{v}^{(k)}$			λ
0	1	0	0	
1	−1	−1	−3	−1.
2	−7	11	−27	7.
3	−25	17	−69	3.5714
4	−79	95	−255	3.1600
5	−241	209	−693	3.0506
6	−727	791	−2247	3.0166
7	−2185	2057	−6429	3.0055
8	−6559	6815	−19935	3.0018
9	−19681	19169	−58533	3.0006
10	−59047	60071	−178167	3.0002
11	−177145	175097	−529389	3.0001
12	−531439	535535	−1598415	3.0000

Furthermore, the entries of $\mathbf{v}^{(k)}$ are given by $v_i^{(k)} = \lambda_1^k v_i$. Hence, for any nonzero eigenvector component $v_i \neq 0$, we can recover the eigenvalue λ_1 itself by taking a ratio between the i^{th} components of successive iterates:

$$\lambda_1 \;\approx\; \frac{v_i^{(k+1)}}{v_i^{(k)}}, \qquad \text{provided} \qquad v_i^{(k)} \neq 0. \tag{10.95}$$

Example 10.49. Consider the matrix $A = \begin{pmatrix} -1 & 2 & 2 \\ -1 & -4 & -2 \\ -3 & 9 & 7 \end{pmatrix}$. As the reader can check, its eigenvalues and eigenvectors are

$$\lambda_1 = 3, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}, \qquad \lambda_2 = -2, \qquad \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \qquad \lambda_3 = 1, \qquad \mathbf{v}_3 = \begin{pmatrix} -1 \\ 1 \\ -2 \end{pmatrix}.$$

Repeatedly multiplying the particular initial vector $\mathbf{v} = (\,1, 0, 0\,)^T$ by A results in the vectors $\mathbf{v}^{(k)} = A^k \mathbf{v}$ listed in the accompanying table. The last column indicates the ratio $v_1^{(k)}$ and $v_1^{(k-1)}$ between the first components of successive iterates. (One could equally well use the second or third components.) These ratios are converging to the third and largest eigenvalue $\lambda_3 = 3$, while $\mathbf{v}^{(k)}$ is converging to a very large multiple of the corresponding eigenvector $\mathbf{v}_3$.

The success of the power method requires that A have a unique dominant eigenvalue of maximal modulus, which, by definition, equals its spectral radius: $|\,\lambda_1\,| = \rho(A)$. The rate of convergence of the method is governed by the ratio $|\,\lambda_2/\lambda_1\,|$ between the subdominant

and dominant eigenvalues. Thus, the further the dominant eigenvalue lies away from the rest, the faster the power method converges.

Since complex eigenvalues of real matrices come in complex conjugate pairs, of the same modulus, matrices whose dominant eigenvalue is complex are not covered by the method. Indeed, one could hardly expect to compute a complex eigenvalue as a ratio of real vectors! However, a slightly more sophisticated version of the method can handle the cases when there is a single complex-conjugate pair of dominant eigenvalues; see Exercise ■. We also assumed that the initial vector $\mathbf{v}^{(0)}$ includes a nonzero multiple of the dominant eigenvector, i.e., $c_1 \neq 0$. As we do not know the eigenvectors, it is not so easy to guarantee this in advance, although one must be quite unlucky to make such a poor choice of initial vector. (Of course, the stupid choice $\mathbf{v}^{(0)} = \mathbf{0}$ is not counted.) Moreover, even if c_1 happens to be 0 initially, numerical round-off error will typically come to one's rescue, since it will almost inevitably introduce a tiny component of the eigenvector $\mathbf{v}_1$ into some iterate, and this component will eventually dominate the computation. The trick is to wait long enough for it to show up!

Since the iterates of A are, typically, getting either very large — when $\rho(A) > 1$ — or very small — when $\rho(A) < 1$ — the iterated vectors will be increasingly subject to round-off error, if not numerical over– or under–flow. One way to avoid this problem is to work with unit vectors $\mathbf{u}^{(k)} = \| \mathbf{v}^{(k)} \|^{-1} \mathbf{v}^{(k)}$, whose entries cannot get too large, and so are less likely to cause numerical errors in the computations. Here $\| \cdot \|$ is any convenient norm — the 1 and ∞ norms being slightly easier to compute than the Euclidean norm. The unit vectors $\mathbf{u}^{(k)}$ can be computed directly by the iterative scheme

$$\mathbf{u}^{(0)} = \frac{\mathbf{v}^{(0)}}{\| \mathbf{v}^{(0)} \|}, \qquad \text{and} \qquad \mathbf{u}^{(k+1)} = \frac{A\,\mathbf{u}^{(k)}}{\| A\,\mathbf{u}^{(k)} \|}. \tag{10.96}$$

If the largest eigenvalue $\lambda_1 > 0$ is positive, then $\mathbf{u}^{(k)} \to \mathbf{u}_1$ will converge to one of the two unit eigenvectors (the other is $-\mathbf{u}_1$) corresponding to the eigenvalue λ_1. If $\lambda_1 < 0$, then the iterates will switch back and forth between the two eigenvectors $\mathbf{u}^{(k)} \approx (-1)^k \mathbf{u}_1$. In either case, the eigenvalue λ_1 is obtained as a limiting ratio between nonzero entries of $\mathbf{u}^{(k)}$ and $A\,\mathbf{u}^{(k)}$. If some other behavior is observed, it means that one of our assumptions is not valid; either A has more than one dominant eigenvalue of maximum modulus, or it is not complete.

Example 10.50. For the matrix considered in Example 10.49, if we multiply the initial vector $\mathbf{u}^{(k)} = (\,1, 0, 0\,)^T$ by A, the resulting unit vectors $\mathbf{u}^{(k)} = A\mathbf{u}^{(k-1)}/\| A\mathbf{u}^{(k-1)} \|$ for the Euclidean norm are given in the table. The last column, being the ratio between the first components of $\mathbf{u}^{(k)}$ and $A\mathbf{u}^{(k)}$, converges to the dominant eigenvalue $\lambda_1 = 3$.

Variants of the power method for computing the other eigenvalues of the matrix are explored in the exercises.

The Q R Algorithm

As stated, the power method only produces the largest eigenvalue of a matrix A. The inverse power method of Exercise ■ can be used to find the smallest eigenvalue. Additional eigenvalues can be found by using the shifted inverse power method of Exercise ■, or the

k	$\mathbf{u}^{(k)}$			λ
1	-0.3015	-0.3015	-0.9045	-1.0000
2	-0.2335	0.3669	-0.9005	7.0000
3	-0.3319	0.2257	-0.9159	3.5714
4	-0.2788	0.3353	-0.8999	3.1600
5	-0.3159	0.2740	-0.9084	3.0506
6	-0.2919	0.3176	-0.9022	3.0166
7	-0.3080	0.2899	-0.9061	3.0055
8	-0.2973	0.3089	-0.9035	3.0018
9	-0.3044	0.2965	-0.9052	3.0006
10	-0.2996	0.3048	-0.9041	3.0002
11	-0.3028	0.2993	-0.9048	3.0001
12	-0.3007	0.3030	-0.9043	3.0000

deflation method of Exercise ■. However, if we need to know *all* the eigenvalues, these methods are too time-consuming and impractical.

The most important scheme for simultaneously approximating all the eigenvalues of a matrix A is the remarkable QR algorithm, first proposed in 1961, by Francis, [**61**], and Kublanovskaya, [**99**]. The underlying idea is simple but surprising. The first step is to factor the matrix

$$A = A_0 = Q_0 R_0$$

into a product of an orthogonal matrix Q_0 and a positive upper triangular matrix (i.e., with all positive entries along the diagonal) R_0 using the Gram–Schmidt orthogonalization procedure of Theorem 5.24. Next, multiply the two factors together *in the wrong order*! The result is the new matrix

$$A_1 = R_0 Q_0.$$

We then repeat these two steps. Thus, we next factor

$$A_1 = Q_1 R_1$$

using the Gram–Schmidt process, and then multiply the factors in the reverse order to produce $A_2 = R_2 Q_2$. The general algorithm is

$$A = Q_0 R_0, \qquad A_{k+1} = R_k Q_k = Q_{k+1} R_{k+1}, \qquad k = 0, 1, 2, \dots , \tag{10.97}$$

where Q_k, R_k come from the previous step, and the subsequent orthogonal matrix Q_{k+1} and positive upper triangular matrix R_{k+1} are computed using the numerically stable form of the Gram–Schmidt algorithm.

The astonishing fact is that, for many matrices A, the iterates $A_k \longrightarrow V$ converge to a upper triangular matrix V whose diagonal entries are the eigenvalues of A. Thus, after

a sufficient number of iterations, say k, the matrix A_k will have very small entries below the diagonal, and one can read off a complete system of (approximate) eigenvalues along its diagonal. For each eigenvalue, the computation of the corresponding eigenvector can be done by solving the appropriate homogeneous linear system, or by applying the shifted inverse power method of Exercise ■.

Example 10.51. Consider the matrix $A = \begin{pmatrix} 2 & 1 \\ 2 & 3 \end{pmatrix}$. The initial Gram–Schmidt factorization $A = Q_0 R_0$ yields

$$Q_0 = \begin{pmatrix} .7071 & .7071 \\ -.7071 & .7071 \end{pmatrix}, \qquad R_0 = \begin{pmatrix} 2.8284 & 2.8284 \\ 0 & 1.4142 \end{pmatrix}.$$

These are multiplied in the reverse order to give $A_1 = R_0 Q_0 = \begin{pmatrix} 4 & 0 \\ 1 & 1 \end{pmatrix}$. We refactor $A_1 = Q_1 R_1$ via Gram–Schmidt, and then reverse multiply to produce

$$Q_1 = \begin{pmatrix} .9701 & -.2425 \\ .2425 & .9701 \end{pmatrix}, \qquad R_1 = \begin{pmatrix} 4.1231 & .2425 \\ 0 & .9701 \end{pmatrix},$$
$$A_2 = R_1 Q_1 = \begin{pmatrix} 4.0588 & -.7647 \\ .2353 & .9412 \end{pmatrix}.$$

The next iteration yields

$$Q_2 = \begin{pmatrix} .9983 & -.0579 \\ .0579 & .9983 \end{pmatrix}, \qquad R_2 = \begin{pmatrix} 4.0656 & -.7090 \\ 0 & .9839 \end{pmatrix},$$
$$A_3 = R_2 Q_2 = \begin{pmatrix} 4.0178 & -.9431 \\ .0569 & .9822 \end{pmatrix}.$$

Continuing in this manner, after 9 iterations we find, to four decimal places

$$Q_9 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad R_9 = \begin{pmatrix} 4 & -1 \\ 0 & 1 \end{pmatrix}, \qquad A_{10} = R_9 Q_9 = \begin{pmatrix} 4 & -1 \\ 0 & 1 \end{pmatrix}.$$

The eigenvalues of A, namely 4 and 1, appear along the diagonal of A_{10}. Additional iterations produce very little further change, although they can be used for increasing the accuracy of the computed eigenvalues.

If the original matrix A happens to be symmetric and positive definite, then the limiting matrix $A_k \to V = \Lambda$ is, in fact, the diagonal matrix containing the eigenvalues of A. Moreover, if, in this case, we recursively define

$$S_k = S_{k-1} Q_k = Q_0 Q_1 \cdots Q_{k-1} Q_k, \tag{10.98}$$

then $S_k \to S$ have, as their limit, the orthogonal matrix appearing in the Spectral Theorem 8.25, whose columns are the orthonormal eigenvector basis of A.

Example 10.52. Consider the symmetric matrix $A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & -1 \\ 0 & -1 & 6 \end{pmatrix}$. The initial $A = Q_0 R_0$ factorization produces

$$S_0 = Q_0 = \begin{pmatrix} .8944 & -.4082 & -.1826 \\ .4472 & .8165 & .3651 \\ 0 & -.4082 & .9129 \end{pmatrix}, \qquad R_0 = \begin{pmatrix} 2.2361 & 2.2361 & -.4472 \\ 0 & 2.4495 & -3.2660 \\ 0 & 0 & 5.1121 \end{pmatrix},$$

and so

$$A_1 = R_0 Q_0 = \begin{pmatrix} 3 & 1.0954 & 0 \\ 1.0954 & 3.3333 & -2.0870 \\ 0 & -2.0870 & 4.6667 \end{pmatrix}.$$

We refactor $A_1 = Q_1 R_1$ and reverse multiply to produce

$$Q_1 = \begin{pmatrix} .9393 & -.2734 & -.2071 \\ .3430 & .7488 & .5672 \\ 0 & -.6038 & .7972 \end{pmatrix}, \qquad R_1 = \begin{pmatrix} 3.1937 & 2.1723 & -.7158 \\ 0 & 3.4565 & -4.3804 \\ 0 & 0 & 2.5364 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 3.7451 & 1.1856 & 0 \\ 1.1856 & 5.2330 & -1.5314 \\ 0 & -1.5314 & 2.0219 \end{pmatrix}, \qquad S_1 = \begin{pmatrix} .7001 & -.4400 & -.5623 \\ .7001 & .2686 & .6615 \\ -.1400 & -.8569 & .4962 \end{pmatrix},$$

where $S_1 = S_0 Q_1 = Q_0 Q_1$. Continuing in this manner, after 10 iterations we find

$$Q_{10} = \begin{pmatrix} 1.0000 & -.0067 & 0 \\ .0067 & 1.0000 & 0.0001 \\ 0 & -.0001 & 1.0000 \end{pmatrix}, \qquad R_{10} = \begin{pmatrix} 6.3229 & .0647 & 0 \\ 0 & 3.3582 & -.0006 \\ 0 & 0 & 1.3187 \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} 6.3232 & 0.0224 & 0 \\ .0224 & 3.3581 & -.0002 \\ 0 & -.0002 & 1.3187 \end{pmatrix}, \qquad S_{10} = \begin{pmatrix} .0753 & -.5667 & -.8205 \\ .3128 & -.7679 & .5591 \\ -.9468 & -.2987 & .1194 \end{pmatrix}.$$

After 20 iterations, the process has completely settled down, and

$$Q_{20} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad R_{20} = \begin{pmatrix} 6.3234 & .0001 & 0 \\ 0 & 3.3579 & 0 \\ 0 & 0 & 1.3187 \end{pmatrix},$$

$$A_{21} = \begin{pmatrix} 6.3234 & 0 & 0 \\ 0 & 3.3579 & 0 \\ 0 & 0 & 1.3187 \end{pmatrix}, \qquad S_{20} = \begin{pmatrix} .0710 & -.5672 & -.8205 \\ .3069 & -.7702 & .5590 \\ -.9491 & -.2915 & .1194 \end{pmatrix}.$$

The eigenvalues of A appear along the diagonal of A_{21}, while the columns of S_{20} are the corresponding orthonormal eigenvector basis, both correct to 4 decimal places.

We will devote the remainder of this section to a justification of the $Q\,R$ algorithm. The secret is that the method is, in fact, intimately connected with the more primitive power method. To keep the exposition under control, let us make the simplifying assumption

that the matrix $A > 0$ is symmetric and positive definite with distinct positive eigenvalues that we label in decreasing order:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n. \tag{10.99}$$

The corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ can be chosen to form an orthonormal basis of $\mathbb{R}^n$. While not the most general case to which the QR algorithm applies, the positive definite matrices are a the most important subclass, and the ones for which the basic algorithm applies as stated.

If one were to implement a version of the power method to capture all the eigenvectors and eigenvalues of A, one might think of iterating a complete basis $\mathbf{u}_1^{(0)}, \ldots, \mathbf{u}_n^{(0)}$ of $\mathbb{R}^n$ instead of just one individual vector. The problem is that, for almost all vectors, the matrix power iterates $\mathbf{u}_j^{(k)} = A^k \mathbf{u}_j$ all tend to a multiple of the eigenvector $\mathbf{u}_1$ corresponding to the dominant eigenvalue. Normalizing the vectors at each step of the algorithm, as in (10.96), is not any better, since then they merely converge to one of the two dominant unit eigenvectors $\pm \mathbf{u}_1$. However, if, inspired by the form of the basis, we *orthonormalize* the vectors at each step, then we effectively keep them separate and so prevent them from all accumulating at the dominant unit eigenvector, and so, with some luck, they may decide to converge to the other eigenvectors. This inspired hope is the heart of the QR algorithm! After all, orthonormalizing a basis is equivalent to the QR matrix factorization.

Thus, we start with any orthonormal basis, which, for simplicity, we take to be the standard basis vectors of $\mathbb{R}^n$, and so $\mathbf{u}_j^{(0)} = \mathbf{e}_j$, $j = 1, \ldots, n$. At the k^{th} stage of the algorithm, we set $\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_n^{(k)}$ to be the orthonormal vectors that result from applying the Gram–Schmidt algorithm to the power vectors $\mathbf{v}_j^{(k)} = A^k \mathbf{e}_j$. In matrix language, the vectors $\mathbf{v}_1^{(k)}, \ldots, \mathbf{v}_n^{(k)}$ are merely the columns of A^k, and the orthonormal basis $\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_n^{(k)}$ are the columns of the orthogonal matrix S_k in the QR decomposition of the k^{th} power of A, which we denote by

$$A^k = S_k P_k, \tag{10.100}$$

where P_k is positive upper triangular. Note that, in view of (10.97)

$$\begin{aligned} A &= Q_0 R_0, & A^2 &= Q_0 R_0 Q_0 R_0 = Q_0 Q_1 R_1 R_0, \\ A^3 &= Q_0 R_0 Q_0 R_0 Q_0 R_0 = Q_0 Q_1 R_1 Q_1 R_1 R_0 = Q_0 Q_1 Q_2 R_2 R_1 R_0, \end{aligned}$$

and, in general,

$$A^k = \big(Q_0 Q_1 \cdots Q_{k-1} Q_k \big) \big(R_k R_{k-1} \cdots R_1 R_0 \big). \tag{10.101}$$

The product of orthogonal matrices is also orthogonal, and the product of positive upper triangular matrices is also positive upper triangular. Comparing (10.100), (10.101), we conclude that

$$S_k = Q_0 Q_1 \cdots Q_{k-1} Q_k == S_{k-1} Q_k, \qquad P_k = R_k R_{k-1} \cdots R_1 R_0 = R_k P_{k-1}, \tag{10.102}$$

since the QR factorization is unique once one requires that all the diagonal entries of R be positive.

Let $S = (\,\mathbf{u}_1\ \mathbf{u}_2\ \dots\ \mathbf{u}_n\,)$ denote the orthogonal eigenvector matrix. The Spectral Theorem 8.25 tells us that

$$A = S\,\Lambda\,S^T, \qquad \text{where} \qquad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

is the diagonal eigenvalue matrix. Substituting the spectral factorization into (10.100), we find

$$A^k = S\,\Lambda^k\,S^T = S_k\,P_k.$$

We now make one additional assumption on the eigenvectors by requiring that S^T be a regular matrix, and so, by Gaussian elimination, admits a factorization $S^T = LU$ into a product of special lower and upper triangular matrices. This holds generically, and is the analog of the condition that our original vector in the basic power method includes a nonzero component of the dominant eigenvector. (The reader can trace through the argument in the general case that S^T requires a permuted LU factorization.) Under this assumption,

$$A^k = S\,\Lambda^k\,LU = S_k\,P_k, \qquad \text{and hence} \qquad S\,\Lambda^k\,L = S_k\,P_k\,U^{-1}.$$

Multiplying on the right by Λ^{-k} we obtain

$$S\,\Lambda^k\,L\,\Lambda^{-k} = S_k\,T_k, \qquad \text{where} \qquad T_k = P_k\,U^{-1}\,\Lambda^{-k} \tag{10.103}$$

is also a positive upper triangular matrix.

Now consider what happens as $k \to \infty$. Since L is lower triangular, the entries of the matrix $\Lambda^k\,L\,\Lambda^{-k}$ below the diagonal are $l_{ij}\,(\lambda_j/\lambda_i)^k \longrightarrow 0$, since $i > j$, and, by (10.99), $0 < \lambda_j/\lambda_i < 1$. Its diagonal entries are all equal to 1, and therefore, in the limit,

$$\Lambda^k\,L\,\Lambda^{-k} \quad \longrightarrow \quad \mathrm{I}\,,$$

with convergence rate governed by the ratio λ_2/λ_1 between the subdominant and the dominant eigenvalues. As a consequence, in the limit as $k \to \infty$, the left hand side of (10.103) tends to the orthogonal eigenvector matrix S:

$$S_k\,T_k \quad \longrightarrow \quad S. \tag{10.104}$$

We now make use of the following result, whose proof will be given after we finish the justification of the QR algorithm.

Lemma 10.53. *The products of orthogonal and positive upper triangular matrices have an orthogonal limit, as in* (10.104)*, if and only if the individual matrices have limits*

$$S_k \longrightarrow S, \qquad T_k \longrightarrow \mathrm{I}\,. \tag{10.105}$$

Therefore, as claimed, the orthogonal matrices S_k do converge to the orthogonal eigenvector matrix. Moreover, by (10.102), (10.103),

$$R_k = P_k\,P_{k-1}^{-1} = \big(\,T_k\,\Lambda^k\,U^{-1}\,\big)\,\big(\,T_{k-1}\,\Lambda^{k-1}\,U^{-1}\,\big)^{-1} = T_k\,\Lambda\,T_{k-1}^{-1}.$$

Since both T_k and T_{k-1} converge to the identity matrix, in the limit $R_k \to \Lambda$ converges to the diagonal eigenvalue matrix, as claimed. The eigenvalues appear in decreasing order along the diagonal — this follows from our regularity assumption on the transposed eigenvector matrix S^T.

Theorem 10.54. *If A is symmetric, satisfies* (10.99), *and S^T is a regular matrix, then the matrices $S_k \to S$ and $R_k \to \Lambda$ in the QR algorithm converge to, respectively the eigenvector matrix and the diagonal eigenvalue matrix. The rate of convergence is governed by the ratio between the subdominant and dominant eigenvalues.*

The last remaining detail is a proof of Lemma 10.53. We write $S = (\,\mathbf{u}_1\ \mathbf{u}_2\ \dots\ \mathbf{u}_n\,)$, $S_k = (\mathbf{u}_1^{(k)}, \dots, \mathbf{u}_n^{(k)})$ in columnar form, and let $t_{ij}^{(k)}$ denote the matrix entries of the positive upper triangular matrix T_k. The last column of the limiting equation (10.104) reads $t_{nn}^{(k)}\,\mathbf{u}_n^{(k)} \to \mathbf{u}_n$. Since both $\mathbf{u}_n^{(k)}$ and $\mathbf{u}_n$ are unit vectors, and $t_{nn}^{(k)} > 0$, the norm of the limit implies $t_{nn}^{(k)} \to 1$ and hence the last column $\mathbf{u}_n^{(k)} \to \mathbf{u}_n$. The next to last column of (10.104) reads

$$t_{n-1,n-1}^{(k)}\,\mathbf{u}_{n-1}^{(k)} + t_{n-1,n}^{(k)}\,\mathbf{u}_n^{(k)} \ \longrightarrow\ \mathbf{u}_{n-1}.$$

Taking the inner product with $\mathbf{u}_n^{(k)} \to \mathbf{u}_n$ and using orthonormality, we deduce $t_{n-1,n}^{(k)} \to 0$, and so $t_{n-1,n-1}^{(k)}\,\mathbf{u}_{n-1}^{(k)} \to \mathbf{u}_{n-1}$, which, by the previous reasoning, implies $t_{n-1,n-1}^{(k)} \to 1$ and $\mathbf{u}_{n-1}^{(k)} \to \mathbf{u}_{n-1}$. The proof is completed through a reverse induction on the columns, and the remaining details are left to the reader.

Tridiagonalization

In practical implementations, the direct QR algorithm is not very efficient, and takes too long to provide reasonable approximations to the eigenvalues of large matrices. Fortunately, the algorithm can be made efficient by a simple preprocessing step. The key observation is that the QR algorithm preserves the class of symmetric tridiagonal matrices, cf. Exercise ∎, and, moreover, like Gaussian elimination, is much faster when applied to this class of matrices. Alston Householder devised a simple method that converts any symmetric matrix into tridiagonal form while preserving all the eigenvalues. Thus, by first applying the Householder tridiagonalization algorithm, and then applying the QR method to the resulting tridiagonal matrix, the result is an efficient and practical algorithm for computing eigenvalues of symmetric matrices.

Consider the *Householder* or *elementary reflection matrix*

$$H = \mathrm{I} - 2\,\mathbf{u}\,\mathbf{u}^T \tag{10.106}$$

in which $\mathbf{u}$ is a unit vector (in the Euclidean norm). The matrix H represents a reflection of vectors through the plane perpendicular to $\mathbf{u}$, as illustrated in Figure 10.5. According to Exercise ∎, H is a symmetric orthogonal matrix, and so

$$H^T = H, \qquad H^2 = \mathrm{I}, \qquad H^{-1} = H.$$

The proof is straightforward: symmetry is immediate, while

$$H\,H^T = H^2 = (\,\mathrm{I} - 2\,\mathbf{u}\,\mathbf{u}^T)\,(\,\mathrm{I} - 2\,\mathbf{u}\,\mathbf{u}^T) = \mathrm{I} - 4\,\mathbf{u}\,\mathbf{u}^T + 4\,\mathbf{u}\,(\mathbf{u}^T\mathbf{u})\,\mathbf{u}^T = \mathrm{I}$$

since, by assumption, $\mathbf{u}^T\mathbf{u} = \|\,\mathbf{u}\,\|^2 = 1$. By suitably prescribing the unit vector $\mathbf{u}$, we can construct an elementary reflection matrix that interchanges any two vectors of the same length.

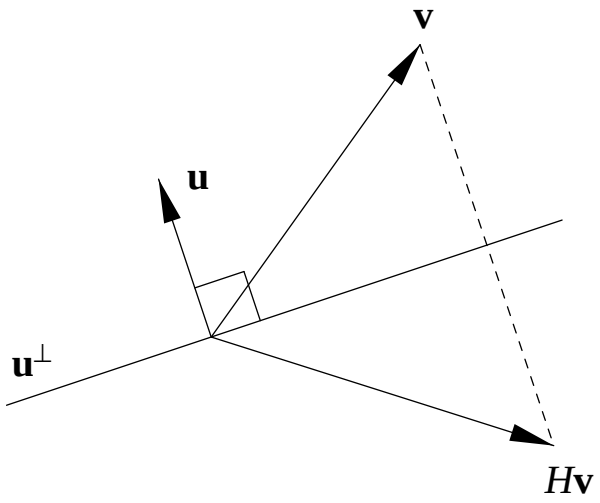


Figure 10.5. Elementary Reflection Matrix.

Lemma 10.55. *Let* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ *with* $\| \mathbf{x} \| = \| \mathbf{y} \|$. *Set* $\mathbf{u} = \dfrac{\mathbf{x} - \mathbf{y}}{\| \mathbf{x} - \mathbf{y} \|}$ *and let* $H = \mathrm{I} - 2\,\mathbf{u}\,\mathbf{u}^T$ *be the corresponding elementary reflection matrix. Then* $H\,\mathbf{x} = \mathbf{y}$ *and* $H\,\mathbf{y} = \mathbf{x}$.

Proof: Keeping in mind that $\mathbf{x}$ and $\mathbf{y}$ have the same Euclidean norm, we compute

$$\begin{aligned} H\,\mathbf{x} = (\,\mathrm{I} - 2\,\mathbf{u}\,\mathbf{u}^T)\,\mathbf{x} &= \mathbf{x} - 2\,\frac{(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T\mathbf{x}}{\| \mathbf{x} - \mathbf{y} \|^2} \\ &= \mathbf{x} - 2\,\frac{(\mathbf{x}-\mathbf{y})\,\big(\,\| \mathbf{x} \|^2 - \mathbf{y}\cdot\mathbf{x}\,\big)}{2\,\| \mathbf{x} \|^2 - 2\,\mathbf{x}\cdot\mathbf{y}} = \mathbf{x} - (\mathbf{x}-\mathbf{y}) = \mathbf{y}. \end{aligned}$$

The proof of the second equation is similar. *Q.E.D.*

Given a symmetric $n \times n$ matrix A, our goal is to devise a similar tridiagonal matrix by applying Householder reflections. We begin by setting

$$\mathbf{x}_1 = \begin{pmatrix} 0 \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{pmatrix}, \qquad \mathbf{y}_1 = \begin{pmatrix} 0 \\ \pm r_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \text{where} \qquad r_1 = \| \mathbf{x}_1 \| = \| \mathbf{y}_1 \|,$$

so that $\mathbf{x}_1$ consists of all off-diagonal entries of the first column of A. Let

$$H_1 = \mathrm{I} - 2\,\mathbf{u}_1\,\mathbf{u}_1^T, \qquad \text{where} \qquad \mathbf{u}_1 = \frac{\mathbf{x}_1 - \mathbf{y}_1}{\| \mathbf{x}_1 - \mathbf{y}_1 \|}$$

be the corresponding elementary reflection matrix that maps $\mathbf{x}_1$ to $\mathbf{y}_1$. Either $\pm$ sign in the formula for $\mathbf{y}_1$ works in the algorithm; the practical choice is to set it to be the opposite of the sign of the entry a_{21}, which minimizes the possible effects of round–off error when

computing the unit vector $\mathbf{u}_1$. A direct computation, based on Lemma 10.55 and the fact that the first entry of $\mathbf{u}_1$ is zero, proves that

$$A_2 = H_1\, A\, H_1 = \begin{pmatrix} a_{11} & r_1 & 0 & \dots & 0 \\ r_1 & \widetilde{a}_{22} & \widetilde{a}_{23} & \dots & \widetilde{a}_{2n} \\ 0 & \widetilde{a}_{32} & \widetilde{a}_{33} & \dots & \widetilde{a}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \widetilde{a}_{n2} & \widetilde{a}_{n} & \dots & \widetilde{a}_{nn} \end{pmatrix} \tag{10.107}$$

for some $\widetilde{a}_{ij}$. Thus, by a single Householder transformation, we can arrange that the first row and column of A are of tridiagonal form. At the next stage, we work on the second row and column of the new matrix A_2. We set

$$\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ \widetilde{a}_{32} \\ \widetilde{a}_{42} \\ \dots \\ \widetilde{a}_{n2} \end{pmatrix}, \qquad \mathbf{y}_1 = \begin{pmatrix} 0 \\ 0 \\ \pm r_2 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \qquad \text{where} \qquad r_2 = \| \mathbf{x}_2 \| = \| \mathbf{y}_2 \|,$$

and the sign is chosen to be the same as that of $\widetilde{a}_{32}$. We set $H_2 = \mathrm{I} - 2\,\mathbf{u}_2\,\mathbf{u}_2^T$, leading to

$$A_3 = H_2\, A_2\, H_2 = \begin{pmatrix} a_{11} & r_1 & 0 & 0 & \dots & 0 \\ r_1 & \widetilde{a}_{22} & r_2 & 0 & \dots & 0 \\ 0 & r_2 & \widehat{a}_{33} & \widehat{a}_{34} & \dots & \widehat{a}_{3n} \\ 0 & 0 & \widehat{a}_{43} & \widehat{a}_{44} & \dots & \widehat{a}_{4n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \widehat{a}_{n3} & \widehat{a}_{n4} & \dots & \widehat{a}_{nn} \end{pmatrix},$$

whose first two rows and columns are in tridiagonal form. The remaining steps in the algorithm should now be clear; once they are reached, the final two columns need not be modified since the resulting matrix will be in tridiagonal form. Let us illustrate the method by an example.

Example 10.56. To tridiagonalize matrix $A = \begin{pmatrix} 4 & 1 & -1 & 2 \\ 1 & 4 & 1 & -1 \\ -1 & 1 & 4 & 1 \\ 2 & -1 & 1 & 4 \end{pmatrix}$, we begin with its first column. We set $\mathbf{x}_1 = \begin{pmatrix} 0 \\ 1 \\ -1 \\ 2 \end{pmatrix}$, and so $\mathbf{y}_1 = \begin{pmatrix} 0 \\ \sqrt{6} \\ 0 \\ 0 \end{pmatrix} \approx \begin{pmatrix} 0 \\ 2.4495 \\ 0 \\ 0 \end{pmatrix}$. Therefore,

the unit vector is $\mathbf{u}_1 = \dfrac{\mathbf{x}_1 - \mathbf{y}_1}{\| \mathbf{x}_1 - \mathbf{y}_1 \|} = \begin{pmatrix} 0 \\ .8391 \\ -.2433 \\ .4865 \end{pmatrix}$, with corresponding Householder matrix

$$H_1 = \mathrm{I} - 2\,\mathbf{u}_1\,\mathbf{u}_1^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -.4082 & .4082 & -.8165 \\ 0 & .4082 & .8816 & .2367 \\ 0 & -.8165 & .2367 & .5266 \end{pmatrix}$$

and so

$$A_2 = H_1\,A\,H_1 = \begin{pmatrix} 4.0000 & -2.4495 & 0 & 0 \\ -2.4495 & 2.3333 & -.3865 & -.8599 \\ 0 & -.3865 & 4.9440 & -.1246 \\ 0 & -.8599 & -.1246 & 4.7227 \end{pmatrix}$$

In the next phase, $\mathbf{x}_2 = \begin{pmatrix} 0 \\ 0 \\ -.3865 \\ -.8599 \end{pmatrix}$, $\mathbf{y}_2 = \begin{pmatrix} 0 \\ 0 \\ -.9428 \\ 0 \end{pmatrix}$, so $\mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \\ -.8396 \\ -.5431 \end{pmatrix}$, and

$$H_2 = \mathrm{I} - 2\,\mathbf{u}_2\,\mathbf{u}_2^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -.4100 & -.9121 \\ 0 & 0 & -.9121 & .4100 \end{pmatrix}.$$

The resulting matrix

$$T = A_3 = H_2\,A_2\,H_2 = \begin{pmatrix} 4.0000 & -2.4495 & 0 & 0 \\ -2.4495 & 2.3333 & .9428 & 0 \\ 0 & .9428 & 4.6667 & 0 \\ 0 & 0 & 0 & 5.0000 \end{pmatrix}$$

is now in tridiagonal form.

Now, the key point is that, because the Householder matrices are their own inverses, the resulting matrices

$$A = A_1, \qquad A_2 = H_1\,A\,H_1^{-1}, \qquad A_3 = H_2\,A_2\,H_2^{-1} = (H_1\,H_2)\,A\,(H_1\,H_2)^{-1}, \qquad \dots$$

are all similar and hence have the same eigenvalues. Thus the final result is a tridiagonal matrix $T = A_n$ that has the *same eigenvalues* as the original symmetric matrix A.

We may then apply the $Q\,R$ algorithm to the final tridiagonal matrix T to approximate its eigenvalues, and hence the eigenvalues of A. According to Exercise ∎, in the resulting Gram–Schmidt procedure, the iterates A_k are all tridiagonal. Moreover, the number of arithmetic operations is relatively small; in Exercise ∎ you are asked to quantify this. For instance, in the preceding example, after we apply 20 iterations of the $Q\,R$ algorithm

directly to T, the upper triangular factor has become

$$R_{20} = \begin{pmatrix} 6.0000 & -0.0065 & 0 & 0 \\ 0 & 4.5616 & 0 & 0 \\ 0 & 0 & 5.0000 & 0 \\ 0 & 0 & 0 & .4384 \end{pmatrix},$$

The eigenvalues of T and hence also A are along the diagonal, and are correct to 4 decimal places.

Finally, if A is not symmetric, then the Householder transformations proceed as before, but the result is no longer tridiagonal, but rather an *upper Hessenberg matrix*, which means that all its entries below the subdiagonal are zero. Thus, a 5×5 upper Hessenberg matrix looks like

$$\begin{pmatrix} \star & \star & \star & \star & \star \\ \star & \star & \star & \star & \star \\ 0 & \star & \star & \star & \star \\ 0 & 0 & \star & \star & \star \\ 0 & 0 & 0 & \star & \star \end{pmatrix}$$

where the starred entries can be anything. The QR algorithm maintains the upper Hessenberg form, and, while not nearly as efficient as the tridiagonal case, still produces a significant savings in computational effort required to find the eigenvalues.

Chapter 11

Boundary Value Problems in One Dimension

In this chapter, we begin our analysis of continuous mechanical systems. The equilibrium equations of one-dimensional continuum mechanics — bars, beams, strings and the like — are formulated as boundary value problems for ordinary differential equations. The basic framework introduced for discrete mechanical systems in Chapter 6 will carry over, in essence, to the infinite-dimensional setting appropriate to such problems. The underlying Euclidean vector space $\mathbb{R}^n$ becomes a function space. Vectors change into functions, while matrices turn into linear differential operators. We shall characterize the underlying linear boundary value problems as self-adjoint and positive (semi-)definite with respect to a suitable inner product on the function space. Stable configurations lead to positive definite boundary value problems whose equilibrium solutions can then be characterized by a general minimization principle based on a quadratic functional representing the total energy in the system. As always, Nature continues seeking to minimize energy.

Many of the basic linear algebra techniques that we developed in the preceding chapters can be systematically translated into this new context. Finite-dimensional linear algebra not only provides us with important insights into the underlying mathematical structure of the problems and their solution spaces, but also motivates basic analytical, and, ultimately, numerical solution schemes. In the infinite-dimensional function space framework underlying these boundary value problems, the general superposition principle becomes reformulated in terms of the response of the system to a unit impulse force concentrated at a single point. However, constructing a function that represents a concentrated impulse turns out to be a highly non-trivial mathematical issue; ordinary functions do not suffice, and we are forced to develop a new theory and calculus of so-called generalized functions or distributions, of inestimable importance to further developments in the subject and its applications. The most essential generalized function is the delta function representing a concentrated impulse. The response of the system to a unit impulse force is known as the Green's function of the boundary value problem, in honor of the self-taught English mathematician (and miller) George Green. With the Green's function in hand, the general solution to the inhomogeneous system can be reconstructed by superimposing the effects of suitably scaled impulse responses on the entire domain. Understanding this construction will become increasingly important as we progress on to partial differential equations, where direct analytical solution techniques are far harder to come by.

In simple situations, we are able to produce explicit formulae for the solution. One should never underestimate the value of such formulae for providing insight into the underlying physical processes, understanding the behavior of more general systems, as well as checking the accuracy of numerical integrators. However, more complicated systems can-

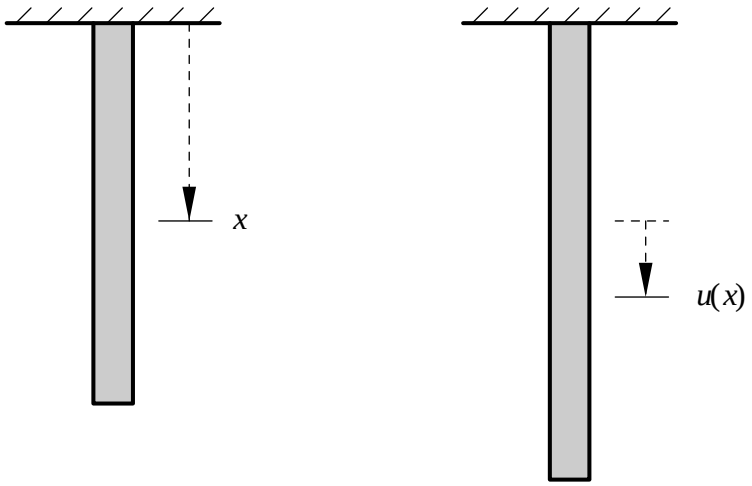


Figure 11.1. Bar with One Fixed Support.

not be solved in closed form, and one must rely on suitable numerical algorithms, which is the focus of the final section of this chapter. Numerical solutions to positive definite boundary value problems will be based on the finite element method, which relies on the characterization of solutions through a minimization principle. The differential equations are converted into a system of linear algebraic equations by minimizing the restriction of the energy functional to a suitably chosen finite-dimensional subspace of the full function space. The implementation of this seminal idea converts the differential equation into a finite-dimensional linear system, which can then be solved by one of the direct or iterative methods that we already learned. An alternative formulation of the finite element solution, that can be applied even in situations where there is no minimum principle available, is based on the idea of a weak solution to the boundary value problem, where one relaxes the classical differentiability requirements.

11.1. Elastic Bars.

A *bar* is a mathematical idealization of a one-dimensional linearly elastic continuum that can be stretched or contracted in the longitudinal direction, but is not allowed to bend in a transverse direction. (Materials that can bend are called beams, and will be analyzed in Section 11.4.) We will view the bar as the continuum limit of a one-dimensional chain of masses and springs — a system that we already analyzed in Section 6.1. Intuitively, the continuous bar consists of an infinite number of masses connected by infinitely short springs. The individual masses can be thought of as the atoms in the bar, although one should not try to read too much into the physics of this interpretation.

We shall derive the basic equilibrium equations for the bar from first principles. Recall the three basic steps we already used to establish the corresponding equilibrium equations for discrete mechanical systems (mass–spring chains and structures):

(i) First, use geometry to relate the displacement of the masses to the elongation in

the connecting springs.

(*ii*) Second, use the constitutive assumptions such as Hooke's Law to relate the strain to the stress or internal force in the system.

(*iii*) Finally, impose a force balance between external and internal forces.

The remarkable fact, which will, when suitably formulated, carry over to the continuum, is that the force balance law is directly related to the geometrical displacement law by a transpose or adjoint operation.

Consider a bar of length ℓ hanging from a fixed support, with the bottom end left free, as illustrated in Figure 11.1. We use $0 \le x \le \ell$ to refer to the reference or unstressed configuration of the bar, so x measures the distance along the bar from the fixed end $x = 0$ to the free end $x = \ell$. Note that we are adopting the convention that the positive x axis points *down*. Let $u(x)$ denote the *displacement* of the bar from its reference configuration. This means that the "atom" that started at position x has moved to position $x + u(x)$. With our convention, $u(x) > 0$ means that the atom has moved down, while if $u(x) < 0$ the atom has moved up. In particular,

$$u(0) = 0 \tag{11.1}$$

because we are assuming that the top end is fixed and cannot move.

The *strain* in the bar measures the relative amount of stretching or elongation. Two nearby atoms, at respective positions x and $x + \Delta x$, are moved to positions $x + u(x)$ and $x + \Delta x + u(x + \Delta x)$. The original, unstressed length of this small section of bar was Δx, while in the new configuration the same section has length

$$\big[\, x + \Delta x + u(x + \Delta x) \,\big] - \big[\, x + u(x) \,\big] = \Delta x + \big[\, u(x + \Delta x) - u(x) \,\big].$$

Therefore, this segment of the bar has been elongated by an amount $u(x + \Delta x) - u(x)$. The dimensionless strain measures the relative elongation, and so is obtained by dividing by the reference length: $[u(x + \Delta x) - u(x)]/\Delta x$. We now take the continuum limit by letting the two atoms become infinitesimally close. Mathematically, we set $y = x + \Delta x$ and let the interatomic spacing $\Delta x \to 0$. The result is the strain function

$$v(x) = \lim_{\Delta x \to 0} \frac{u(x + \Delta x) - u(x)}{\Delta x} = \frac{du}{dx} \tag{11.2}$$

that measures the local stretch in the bar at position x.

We may approximate the bar by a chain of n masses connected by n springs, and letting the bottom mass hang free. The mass/spring chain will have total length ℓ, and so the individual springs have reference length

$$\Delta x = \frac{\ell}{n}.$$

The bar can be viewed as the *continuum limit* of such a mass/spring chain, where the number of masses $n \to \infty$ and the spring lengths $\Delta x \to 0$. The k^{th} mass starts out at position

$$x_k = k\,\Delta x = \frac{k\,\ell}{n}\,,$$

and, under forcing, experiences a displacement u_k. The strain or relative elongation of the k^{th} spring is

$$v_k = \frac{e_k}{\Delta x} = \frac{u_{k+1} - u_k}{\Delta x}. \tag{11.3}$$

In particular, since the fixed end cannot move, the first value $u_0 = 0$ is omitted from the subsequent equations.

Remark: We will find it helpful to label the springs from $k = 0$ to $k = n-1$ here. This will facilitate comparisons with the bar, which, by convention, starts at position $x_0 = 0$.

The relation (11.3) between displacement and strain takes the familiar matrix form

$$\mathbf{v} = A\,\mathbf{u}, \qquad \mathbf{v} = \big(\, v_0, v_1, \dots, v_{n-1} \,\big)^T, \qquad \mathbf{u} = \big(\, u_1, u_2, \dots, u_n \,\big)^T,$$

where

$$A = \frac{1}{\Delta x} \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \end{pmatrix} \approx \frac{d}{dx} \tag{11.4}$$

is the *scaled incidence matrix* of the mass/spring chain. The derivative operator d/dx that relates displacement to strain in the bar equation (11.2) can be viewed as the continuum limit, as the number of masses $n \to \infty$ and the spring lengths $\Delta x \to 0$, of the scaled incidence matrix (11.4). Vice versa, the incidence matrix can be viewed as a discrete, numerical approximation to the derivative operator. Indeed, if we regard the discrete displacements and strains as approximations to the sample values of their continuous counterparts, so

$$u_k \approx u(x_k), \qquad \varepsilon_k \approx \varepsilon(x_k),$$

then (11.3) takes the form

$$v(x_k) \;=\; \frac{u(x_{k+1}) - u(x_k)}{\Delta x} \;=\; \frac{u(x_k + \Delta x) - u(x_k)}{\Delta x} \;\approx\; \frac{du}{dx}\,(x_k).$$

justifying the identification (11.4). The passage back and forth between the discrete and the continuous forms the foundation of continuum mechanics — solids, fluids, gases. Discrete models both motivate and provide numerical approximations to continuum systems, which in turn simplify and provide insight into the discrete domain.

The next part of the framework is to use the constitutive relations of the bar to relate the strain to the *stress*, or internal force experienced by the bar. To keep matters simple, we shall only consider bars that are modeled by a linear relation between stress and strain. For physical bars, this is a pretty good assumption as long as the bar is not stretched beyond its elastic limits. Let $w(x)$ denote the stress on the part of the bar that was at reference position x. Hooke's Law implies that

$$w(x) = c(x)\, v(x), \tag{11.5}$$

where $c(x)$ measures the stiffness of the bar at position x. For a homogeneous bar, made out of a uniform material, $c(x) \equiv c$ is a constant function. The constitutive function $c(x)$ is the continuum limit of the diagonal matrix

$$C = \begin{pmatrix} c_0 & & & \\ & c_1 & & \\ & & \ddots & \\ & & & c_{n-1} \end{pmatrix}$$

of individual spring constants c_k appearing in the discrete constitutive law

$$w_k = c_k\, v_k, \qquad \text{or} \qquad \mathbf{w} = C\,\mathbf{v}, \tag{11.6}$$

Indeed, writing (11.6) as $w(x_k) = c(x_k)\, v(x_k)$ makes the identification immediate.

Finally, we need to impose a force balance at each point of the bar. Suppose $f(x)$ is an external force at position x on the bar, where $f(x) > 0$ means the force is acting downwards. Physical examples include mechanical, gravitational, or electrostatic forces acting solely in the vertical direction. In equilibrium[†], the bar will deform so as to balance the external force with its own internal force resulting from stretching. Now, the internal force per unit length on the section of the bar lying between nearby positions x and $x+\Delta x$ is the difference in stress at the two ends, $[w(x+\Delta x) - w(x)]/\Delta x$. The force balance law requires that, in the limit,

$$0 = f(x) + \lim_{\Delta x \to 0} \frac{w(x+\Delta x) - w(x)}{\Delta x} = f(x) + \frac{dw}{dx}\,,$$

or

$$f = -\,\frac{dw}{dx}\,. \tag{11.7}$$

The force balance law is the continuum limit of the mass–spring chain version,

$$f_k = \frac{w_{k-1} - w_k}{\Delta x}, \qquad\qquad w_n = 0, \tag{11.8}$$

where the final condition implies the correct formula for the force on the free-hanging bottom mass. (Remember that the springs are numbered from 0 to $n-1$.) This indicates that we should also impose an analogous boundary condition

$$w(\ell) = 0 \tag{11.9}$$

at the bottom end of the bar $x_n = \ell$ which is hanging freely and so is unable to support any stress. The matrix form of the discrete system (11.8) is

$$\mathbf{f} = A^T \mathbf{w},$$

[†] The dynamical processes leading to equilibrium will be discussed in Chapter 14.

where the transposed scaled incidence matrix

$$A^T = \frac{1}{\Delta x}\begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & -1 & & \\ & & & 1 & -1 & \\ & & & & \ddots & \ddots \\ & & & & & \end{pmatrix} \approx \; - \frac{d}{dx}\,, \tag{11.10}$$

should approximate the differential operator $-d/dx$ that appears in the continuum force balance law (11.7). Thus, we should somehow interpret the differential operator $-d/dx$ as the "transpose" or "adjoint" of the differential operator d/dx. This important point will be developed properly in Section 11.3. But before trying to go any further in the theory, let us analyze the mathematical equations governing some simple configurations.

But first, let us summarize our progress so far. The three basic equilibrium equations (11.2), (11.5), (11.7) are

$$v(x) = \frac{du}{dx}\,, \qquad\qquad w(x) = c(x)\,v(x), \qquad\qquad f(x) = -\,\frac{dw}{dx}\,. \tag{11.11}$$

Substituting the first equation into the second, and then the resulting formula into the last equation, leads to the equilibrium equation

$$K[u] = -\,\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad\qquad 0 < x < \ell. \tag{11.12}$$

Thus, the displacement $u(x)$ of the bar is obtained as the solution to a second order ordinary differential equation. As such, it will depend on two arbitrary constants, which will be uniquely determined by the boundary conditions[†] (11.1), (11.9) at the two ends:

$$u(0) = 0, \qquad\qquad w(\ell) = c(\ell)\,u'(\ell) = 0. \tag{11.13}$$

Usually $c(\ell) > 0$, in which case it can be omitted from the second boundary condition, which simply becomes $u'(\ell) = 0$. the resulting boundary value problem is viewed as the continuum limit of the linear system

$$K\,\mathbf{u} = A^T C\,A\,\mathbf{u} = \mathbf{f} \tag{11.14}$$

governing a mass-spring chain with one free end. The individual constituents of the stiffness matrix become

$$A \;\longrightarrow\; \frac{d}{dx}\,, \qquad C \;\longrightarrow\; c(x), \qquad A^T \;\longrightarrow\; -\,\frac{d}{dx}\,, \qquad \mathbf{u} \;\longrightarrow\; u(x), \qquad \mathbf{f} \;\longrightarrow\; f(x),$$

and so, in the same continuum limit, the matrix system (11.14) turns into the second order boundary value problem (11.12), (11.13). And, as we will see, further features of the finite-dimensional problem also have, when suitably interpreted, direct continuum counterparts.

[†] We will sometimes use primes, as in $u' = du/dx$, to denote derivatives with respect to x.

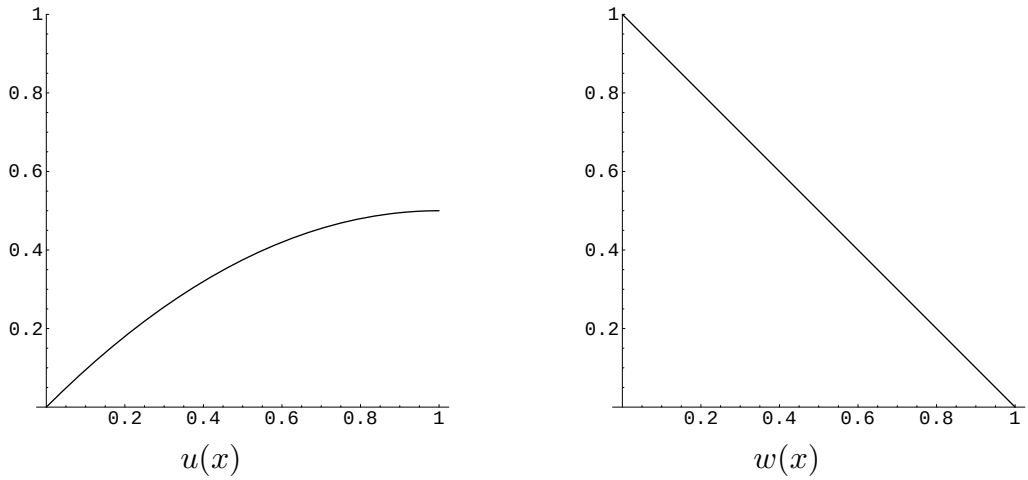


Figure 11.2. Displacement and Stress of Bar with One Fixed End.

Example 11.1. Consider the simplest case of a uniform bar of unit length $\ell = 1$ subjected to a uniform force, e.g., gravity. The equilibrium equation (11.12) is

$$-c\,\frac{d^2u}{dx^2} = f, \tag{11.15}$$

where we are assuming that the force f is constant. This elementary second order ordinary differential equation can be immediately integrated,

$$u(x) = -\tfrac{1}{2}\,\alpha\,x^2 + a\,x + b, \qquad \text{where} \qquad \alpha = \frac{f}{c} \tag{11.16}$$

is the ratio of the force to the stiffness of the bar. The values of the integration constants a and b are fixed by the boundary conditions (11.13), so

$$u(0) = b = 0, \qquad u'(1) = -\alpha + a = 0.$$

Therefore, there is a unique solution to the boundary value problem, yielding the displacement

$$u(x) = \alpha\,x - \tfrac{1}{2}\,\alpha\,x^2, \tag{11.17}$$

which is graphed in Figure 11.2. Note that the displacement reaches its maximum, α, at the free end of the bar, the point which moves downwards the farthest. The stronger the force, or the weaker the bar, the farther the overall displacement. Also note the parabolic shape of the displacement graph, with zero derivative, indicating no strain, at the free end.

Remark: This example illustrates the simplest way to solve boundary value problems. It is modeled on the usual method for solving initial value problems. First, solve the differential equation by standard methods (if possible). For a second order equation, the general solution will involve two arbitrary constants. The values of the constants are found by substituting the general solution into the two boundary conditions. Unlike initial value problems, the existence and/or uniqueness of the solution to a general boundary value problem is not always guaranteed, and so one may encounter situations where one cannot complete the solution; see, for instance, Example 7.41. A more sophisticated approach, based on the Green's function, will be discussed in the following section.

As in the discrete situation, this particular mechanical configuration is *statically determinate*, meaning that we can solve directly for the stress $w(x)$ in terms of the external force $f(x)$ without having to compute the displacement $u(x)$ first. In this particular example, we need to solve the first order boundary value problem

$$-\frac{dw}{dx} = f, \qquad\qquad w(1) = 0,$$

arising from the force balance law (11.7), which yields

$$w(x) = f\,(1-x), \qquad \text{and} \qquad v(x) = \frac{w(x)}{c} = \alpha\,(1-x).$$

Note that the boundary condition determines the integration constant uniquely. We can then find the displacement $u(x)$ by solving another boundary value problem

$$\frac{du}{dx} = v(x) = \alpha\,(1-x), \qquad\qquad u(0) = 0,$$

resulting from (11.2), which again leads to (11.17). As before, the appearance of one boundary condition implies that we can find a unique solution to the differential equation.

Remark: We motivated the boundary value problem for the bar by taking the continuum limit of the mass–spring chain. Let us see to what extent this limiting procedure can be justified. To compare the solutions, we keep the reference length of the chain fixed at $\ell = 1$ and its total mass fixed at m. So, if we have n identical masses, each spring has length $\Delta x = 1/n$. The k^{th} mass will start out at reference position $x_k = k/n$ and has mass $m_k = m/n$. Using static determinacy, we can solve the system (11.8), which reads

$$w_k = w_{k+1} + \frac{f}{n}\,, \qquad\qquad w_n = 0,$$

directly for the stresses:

$$w_k = f\,\left(1 - \frac{k}{n}\right) = f\,(1 - x_k)\,.$$

Thus, in this particular case, the continuous bar and the discrete chain have equal stresses at the sample points: $w(x_k) = w_k$. The strains also are in agreement:

$$v_k = \frac{1}{c}\,w_k = \alpha\left(1 - \frac{k}{n}\right) = \alpha\,(1 - x_k) = v(x_k)\,,$$

where $\alpha = f/c$ as above. We then obtain the displacements by solving

$$u_{k+1} = u_k + \frac{v_k}{n} = u_k + \frac{\alpha}{n}\left(1 - \frac{k}{n}\right).$$

Since $u_0 = 0$, the solution is

$$u_k = \frac{\alpha}{n}\sum_{i=1}^{k} 1 - \frac{\alpha}{n^2}\sum_{i=1}^{k} k = \alpha\left(\frac{k}{n} - \frac{k(k+1)}{2\,n^2}\right) = \alpha\,(x_k - x_k^2) - \frac{\alpha\,x_k}{2\,n} = u(x_k) - \frac{\alpha\,x_k}{2\,n}\,. \tag{11.18}$$

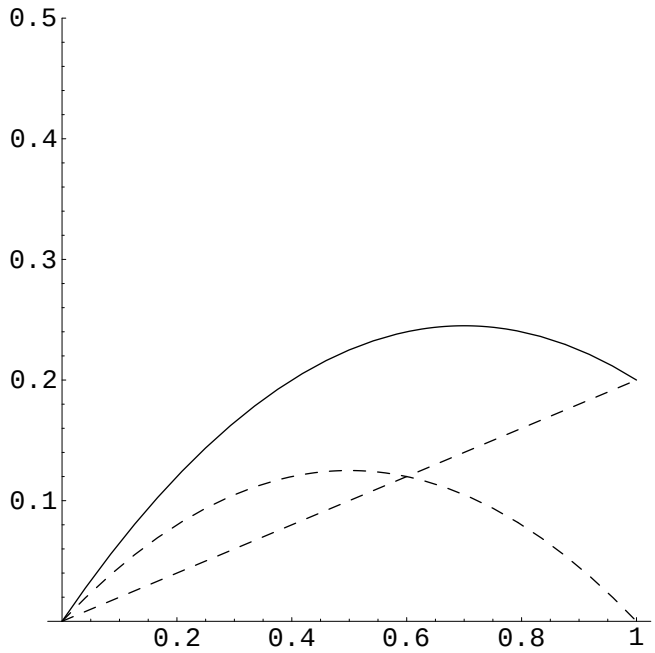


Figure 11.3. Displacements of a Bar with Two Fixed Ends.

Now the sampled displacement $u(x_k)$ is not exactly equal to u_k, but their difference tends to zero as the number of masses $n \to \infty$. In this way, we have completely justified our approximation scheme.

Example 11.2. Consider the same uniform, unit length bar as in the previous example, again subject to a uniform constant force, but now with two fixed ends. We impose the inhomogeneous boundary conditions

$$u(0) = 0, \qquad u(1) = d, \tag{11.19}$$

where the top end is fixed, while the bottom end is displaced an amount d. (Note that $d > 0$ means the bar is stretched, while $d < 0$ means it is compressed.) The general solution to the equilibrium equation (11.15) is, as before, given by (11.16). The values of the arbitrary constants a, b are again determined by plugging into the boundary conditions (11.19), so

$$u(0) = b = 0, \qquad u(1) = -\tfrac{1}{2}\,\alpha + d = 0.$$

Again, there is a unique solution to the boundary value problem,

$$u(x) = \tfrac{1}{2}\,\alpha\,(x - x^2) + d\,x. \tag{11.20}$$

The displacement is a superposition of two functions; the first constituent is due to the external force f, while the second is a uniform stretch due to the boundary condition. (As in Example 7.44, linearity of the boundary value problem allows us to combine the responses to different inhomogeneities.) In Figure 11.3, the dotted curves represent the two simple responses, and the solid graph is their sum, the actual displacement.

Unlike a bar with a free end, this configuration is *statically indeterminate*. There is no boundary condition on the force balance equation

$$-\frac{dw}{dx} = f,$$

and so the integration constant a in the stress $w(x) = a - f\,x$ *cannot* be determined without first figuring out the displacement (11.20):

$$w(x) = c\,\frac{du}{dx} = f\,\left(\tfrac{1}{2} - x\right) + d.$$

Remark: The particular boundary value problems that govern the mechanical equilibria of a simple bar arise in many other physical systems. For example, the equation for the thermal equilibrium of a bar under an external heat source is modeled by the same boundary value problem (11.12); in this case, $u(x)$ represents the temperature of the bar, $c(x)$ represents the *diffusivity* or *thermal conductivity* of the material at position x, while $f(x)$ represents an external heat source. A fixed boundary condition $u(\ell) = a$ corresponds to an end that is held at a fixed temperature a, while a free boundary condition $u'(\ell) = 0$ represents an insulated end that does not allow heat energy to enter or leave the bar. Details of the physical derivation can be found in Section 14.1.

Example 11.3. Finally, consider the case when both ends of the bar are left free. The boundary value problem

$$-u'' = f(x), \qquad u'(0) = 0, \qquad u'(\ell) = 0, \tag{11.21}$$

represents the continuum limit of a mass–spring chain with two free ends, and represents to a bar floating in outer space, subject to a nonconstant external force. Based on our finite-dimensional experience, we expect the solution to manifest an underlying instability of the physical problem. Solving the differential equation, we find

$$u(x) = a\,x + b - \int_0^x \left(\int_0^y f(z)\,dz \right) dy,$$

where the constants a, b are to be determined by the boundary conditions. Since

$$u'(x) = a - \int_0^x f(z)\,dz,$$

the first boundary condition $u'(0) = 0$ requires $a = 0$. The second boundary condition requires

$$u'(\ell) = \int_0^\ell f(x)\,dx = 0, \tag{11.22}$$

which is not automatically valid! The integral represents the total force per unit length exerted on the bar. As in the case of a mass-spring chain with two free ends, if there is a non-zero net force, the bar cannot remain in equilibrium, but will move off in space and the equilibrium boundary value problem (11.21) has no solution. On the other hand, if the

forcing satisfies the constraint (11.22), then the resulting solution of the boundary value problem has the form

$$u(x) = b - \int_0^x \left(\int_0^y f(z)\, dz \right) dy, \tag{11.23}$$

where the constant b is arbitrary. Thus, when it exists, the solution to the boundary value problem is not unique. The constant b solves the corresponding homogeneous problem, and represents a rigid translation of the entire bar by a distance b.

Physically, the free boundary value problem corresponds to an unstable structure with an instability. If both ends of the bar are left free, then there is a translational instability, where the bar moves rigidly in the longitudinal direction. Only those forces with mean zero will not excite this instability. Furthermore, when it does exist, the equilibrium solution is not unique since there is nothing to tie the bar down to any particular spatial position.

This should remind you of our earlier study of linear matrix systems. Indeed, according to Theorem 7.37, the solution to a linear inhomogeneous system is not unique if and only if the corresponding homogeneous system, with trivial forcing, $f(x) \equiv 0$, has a non-trivial solution. In the current situation, any constant function $u(x) \equiv b$ satisfies the homogeneous boundary value problem. We identify the solutions to the homogeneous boundary value problem with the kernel of the linear operator K defining the boundary value problem. The boundary value problem has a unique solution if and only if $\ker K = \{0\}$ is trivial, whereas in the present situation $\ker K$ is a one-dimensional subspace consisting of all constant displacements $u(x) \equiv b$.

The constraint (11.22) on the forcing function is, in fact, a manifestation of the Fredholm alternative (5.71), that requires it to be orthogonal to all of the functions in the cokernel of K. In Section 11.3, we will show that, like its finite-dimensional matrix counterpart, the differential operator K is self-adjoint, and hence its kernel and cokernel coincide: $\operatorname{coker} K = \ker K$, with basis provided by the constant function 1. Thus, orthogonality of the forcing function to the kernel of K will be ensured by the vanishing of the L^2 inner product,

$$\langle\, f\,;1\,\rangle = \int_0^\ell f(x)\, dx = 0,$$

which is precisely the condition (11.22) required for a solution to exist.

11.2. Generalized Functions and the Green's Function.

The general superposition principle for inhomogeneous linear systems, as summarized in Theorem 7.42, inspires a important, alternative approach to the solution of boundary value problems. This method relies on the solution to a particular set of inhomogeneities, namely concentrated unit impulses. The resulting family of fundamental solutions are collectively known as the Green's function for the system. The Green's function has the important property that the solution induced by all other inhomogeneities can be built up as a continuous superposition of these fundamental solutions.

To motivate the construction, let us return briefly to the case of a mass–spring chain. Given the equilibrium equations

$$K\,\mathbf{u} = \mathbf{f}, \tag{11.24}$$

let us decompose the external forcing $\mathbf{f} = (\, f_1, f_2, \ldots, f_n \,)^T \in \mathbb{R}^n$ into a linear combination

$$\mathbf{f} = f_1\, \mathbf{e}_1 + f_2\, \mathbf{e}_2 + \, \cdots \, + f_n\, \mathbf{e}_n \tag{11.25}$$

of the standard basis vectors of $\mathbb{R}^n$. Each $\mathbf{e}_i$ represents a unit force which is applied solely to the i^{th} mass in the chain. Suppose we know how to solve each of the individual systems

$$K\,\mathbf{u}_i = \mathbf{e}_i, \qquad\qquad i = 1, \ldots, n. \tag{11.26}$$

The solution $\mathbf{u}_i$ represents the response of the chain to a single unit force concentrated on the i^{th} mass. Formula (11.25) shows how to decompose any other force vector as a superposition of impulse forces, with f_i representing the strength of the impulse applied to the i^{th} mass. The general superposition principle for linear systems says that we can then write the solution to the inhomogeneous system (11.24) as the same linear combination,

$$\mathbf{u} = f_1\, \mathbf{u}_1 + \, \cdots \, + f_n\, \mathbf{u}_n \tag{11.27}$$

of the individual responses.

Remark: The alert reader will recognize that we are, in fact, reconstructing the solution to the linear system (11.24) by inverting the matrix K. Thus, this observation does not lead to an efficient solution technique for discrete systems. In contrast, in the case of continuous boundary value problems, this idea leads to one of the most important solution paradigms, in both practical and theoretical developments.

The Delta Function

Our aim is to extend this basic superposition principle to the boundary value problem for an elastic bar. Therefore, the key question is how to characterize a force or impulse that is concentrated on a single atom[†] of the bar. A *unit impulse* at position $x = y$ will be described by something called the *delta function*, and denoted by $\delta_y(x)$. Since the impulse is supposed to be concentrated solely at $x = y$, we should have

$$\delta_y(x) = 0 \qquad \text{for} \qquad x \neq y. \tag{11.28}$$

Moreover, since it is a *unit* impulse, we want the total amount of force exerted on the bar to be equal to one. The total force is the sum of the individual forces, which, in the continuum limit, is represented by an integral of the force function $f(x)$ over the length of the bar. Thus, to represent a unit impulse, we must also require that the delta function satisfy

$$\int_0^\ell \delta_y(x)\, dx = 1, \qquad\qquad \text{provided} \qquad\qquad 0 < y < \ell. \tag{11.29}$$

Alas, there is no function that enjoys both of the required properties! At least, not one that behaves like a function in the usual mathematical sense. Indeed, according to the basic facts of Riemann (or even Lebesgue) integration, two functions which are the same

[†] Here, as before, "atom" is used in a figurative sense.

everywhere except at one single point have exactly the same integral, [**125**]. Thus, since δ_y is zero except at one point, its integral should be 0, not 1. The mathematical conclusion is that the two requirements, (11.28), (11.29) are incompatible for ordinary functions!

This unfortunate fact stopped mathematicians dead in their tracks. It took the imagination of a British engineer, Oliver Heaviside, who was not deterred by the lack of rigorous justification, to start utilizing delta functions in practical applications — with remarkable effect. Despite his success, Heaviside was ridiculed by the pure mathematicians of his day, and eventually succumbed to mental illness. But, some thirty years later, the great theoretical physicist Paul Dirac resurrected the delta function for quantum mechanical applications, and this finally made theoreticians sit up and take notice. (Indeed, the term "Dirac delta function" is quite common.) In 1944, the French mathematician Laurent Schwartz finally established a rigorous theory of *distributions* that incorporated such useful, but rather unusual generalized functions, [**130**]. It is beyond the scope of this introductory text to develop a fully rigorous theory of distributions. Rather, in the spirit of Heaviside, we shall concentrate on learning, through practice with applications and computations, how to domesticate these wild mathematical beasts.

There are two distinct ways to introduce the delta function. Both are important and worth knowing.

Method #1. Limits: The first approach is to regard the delta function $\delta_y(x)$ as a limit, as $n \to \infty$, of a sequence of ordinary smooth functions[†] $g_n(x)$. These functions will represent more and more concentrated unit forces, which, in the limit, converge to the desired unit impulse concentrated at a single point, $x = y$. Thus, we require

$$\lim_{n \to \infty} g_n(x) = 0, \qquad x \neq y, \tag{11.30}$$

while the total amount of force remains fixed at

$$\int_0^\ell g_n(x)\, dx = 1. \tag{11.31}$$

On a formal level, the limit "function"

$$\delta_y(x) = \lim_{n \to \infty} g_n(x)$$

should satisfy the key properties (11.28), (11.29).

A simple explicit example of such a sequence is provided by the rational functions

$$g_n(x) = \frac{n}{\pi\big(1 + n^2 x^2\big)} \tag{11.32}$$

that are graphed in Figure 11.4. These functions satisfy

$$\lim_{n \to \infty} g_n(x) = \begin{cases} 0, & x \neq 0, \\ \infty, & x = 0, \end{cases} \tag{11.33}$$

[†] We suppress the dependence of the functions g_n on the point y where the limiting delta function is concentrated.

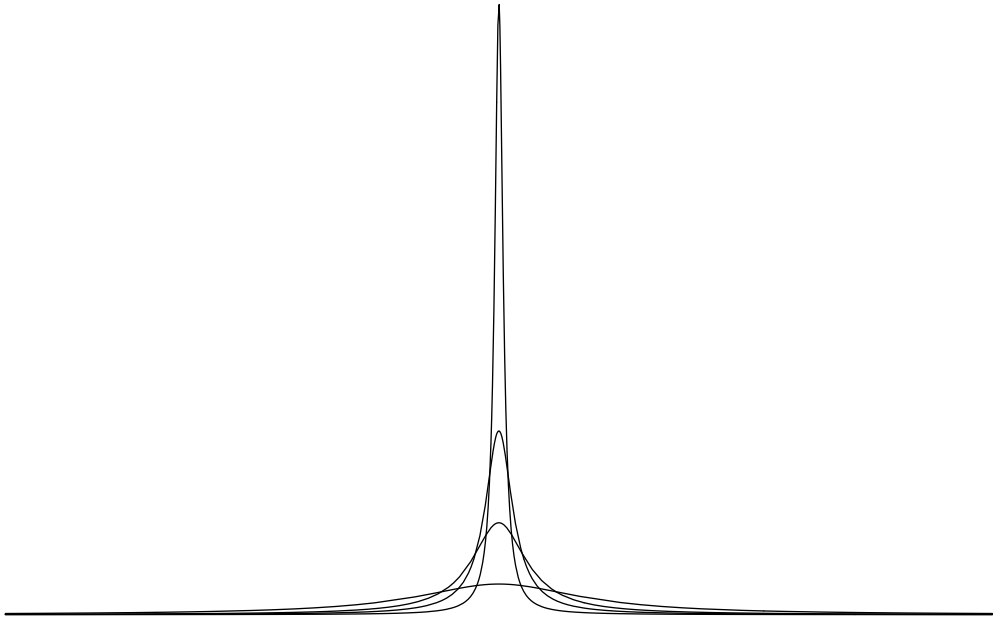


Figure 11.4. Delta Function as Limit.

while[‡]

$$\int_{-\infty}^{\infty} g_n(x)\,dx = \frac{1}{\pi}\,\tan^{-1} n\,x \,\Big|_{x=-\infty}^{\infty} = 1. \tag{11.34}$$

Therefore, formally, we identify the limiting function

$$\lim_{n\to\infty}\, g_n(x) = \delta(x) = \delta_0(x),$$

with the unit impulse delta function concentrated at $x = 0$. As n gets larger and larger, each function $g_n(x)$ is a closer and closer approximation to the delta function, and forms a more and more concentrated spike, while maintaining a unit total area under its graph. The limiting delta function "looks like" an infinitely tall spike of zero width, entirely concentrated at the origin.

Remark: This construction of the delta function highlights the perils of interchanging limits and integrals without proper justification. In Riemann's or Lebesgue's integration theories, the limit of the functions g_n would be indistinguishable from the zero function and so the limit of their integrals (11.34) would *not* equal the integral of their limit:

$$1 = \lim_{n\to\infty} \int_{-\infty}^{\infty} g_n(x)\,dx \;\neq\; \int_{-\infty}^{\infty} \lim_{n\to\infty}\, g_n(x)\,dx = 0.$$

The delta function is, in a sense, a means of sidestepping this analytic inconvenience. The full ramifications and theoretical constructions underlying such limits and generalized functions must, however, be deferred to a rigorous course in real analysis, [**125**].

[‡] It is slightly simpler here to consider the entire real line — corresponding to a bar of infinite length. See Exercise ■ for modifications on a finite interval.

Remark: There are many other possible choices for the limiting functions $g_n(x)$. See Exercise ∎ for another important example.

Once we have found the delta function $\delta(x) = \delta_0(x)$ concentrated at the origin, we can obtain the delta function concentrated at any other position y by a simple translation:

$$\delta_y(x) = \delta(x - y). \tag{11.35}$$

Thus, $\delta_y(x)$ can be realized as the limit of the translated functions

$$\widehat{g}_n(x) = g_n(x - y) = \frac{n}{\pi\left(1 + n^2(x-y)^2\right)}. \tag{11.36}$$

Method #2. Duality: The second approach is a bit more abstract, but much closer to the proper rigorous formulation. Here, we view a generalized function like the delta function as a real-valued linear operator $L\colon \mathrm{C}^0[0,\ell] \to \mathbb{R}$ on a suitable function space — in this case the vector space of continuous functions on the interval $[0,\ell]$. As in the general (7.3), linearity requires that $L[c\,f + d\,g] = c\,L[f] + d\,L[g]$ for all functions f, g and all scalars (constants) $c, d \in \mathbb{R}$.

The key observation is that if $u(x)$ is any continuous function, then

$$\int_0^\ell \delta_y(x)\,u(x)\,dx = u(y), \qquad \text{for} \qquad 0 < y < \ell. \tag{11.37}$$

Indeed, since $\delta_y(x) = 0$ for $x \neq y$, the integrand only depends on the value of $u(x)$ at the point $x = y$, and so

$$\int_0^\ell \delta_y(x)\,u(x)\,dx = \int_0^\ell \delta_y(x)\,u(y)\,dx = u(y)\int_0^\ell \delta_y(x)\,dx = u(y).$$

Equation (11.37) serves to define a linear operator[†] $L_y\colon \mathrm{C}^0[0,\ell] \to \mathbb{R}$ that maps a continuous function $u \in \mathrm{C}^0[0,\ell]$ to its value

$$L_y[u] = u(y) \in \mathbb{R}$$

at the point $x = y$. In the dual approach to generalized functions, the delta function is, in fact, *defined* as this particular linear operator. The function $u(x)$ is sometimes referred to as a *test function* since it serves to test the actual form of the linear operator L.

Remark: If the impulse point y lies outside the integration domain, then

$$\int_0^\ell \delta_y(x)\,u(x)\,dx = 0, \qquad\qquad y < 0 \qquad \text{or} \qquad y > \ell, \tag{11.38}$$

because the integrand is identically zero on the entire interval. For technical reasons, we will not attempt to define the integral (11.38) if the impulse point $y = 0$ or $y = \ell$ lies on the boundary of the interval of integration.

[†] Linearity was demonstrated in Example 7.7.

The interpretation of the linear operator L_y as a kind of function $\delta_y(x)$ is based on the identification between vectors and real-valued linear functions. According to Theorem 7.10, every linear function $L\colon V \to \mathbb{R}$ on a finite-dimensional inner product space is given by an inner product $L[\mathbf{u}] = \langle\, \mathbf{a}\,;\mathbf{u}\,\rangle$ with a fixed element $\mathbf{a} \in V$. Similarly, on the infinite-dimensional function space $\mathrm{C}^0[0,\ell]$, the L^2 inner product

$$L_g[u] = \langle\, g\,;u\,\rangle = \int_0^\ell g(x)\,u(x)\,dx \tag{11.39}$$

with a fixed function $g \in \mathrm{C}^0[0,\ell]$ defines a real-valued linear function $L_g\colon \mathrm{C}^0[0,\ell] \to \mathbb{R}$. However, unlike the finite-dimensional situation, *not* every real-valued linear function on function space has this form! In particular, there is no continuous (ore even integrable) function $\delta_y(x)$ such that the inner product identity

$$\langle\, \delta_y\,;u\,\rangle = \int_0^\ell \delta_y(x)\,u(x)\,dx = u(y) \tag{11.40}$$

holds for every continuous function $u(x)$. This fact highlights yet another of the profound differences between finite- and infinite-dimensional vector spaces!

The dual interpretation of generalized functions acts as if this were true. *Generalized functions are real-valued linear operators on function space, which, formally, are identified as functions via the inner product.* One can, with a little care, manipulate generalized functions as if they were actual functions, but always keeping in mind that a rigorous justification of such computations must ultimately rely on their formal characterization as linear operators.

The two approaches — limits and duality — are completely compatible. Indeed, with a little extra work, one can justify the dual formula (11.37) as the limit

$$u(y) = \lim_{n\to\infty} \int_0^\ell g_n(x)\,u(x)\,dx = \int_0^\ell \delta_y(x)\,u(x)\,dx \tag{11.41}$$

of the inner products of the function u with the approximating concentrated impulse functions $g_n(x)$. In this manner, the linear operator $L[u] = u(y)$ represented by the delta function is the limit, $L_y = \lim_{n\to\infty} L_n$, of the approximating linear operators

$$L_n[u] = \int_0^\ell g_n(x)\,u(x)\,dx.$$

Thus, the choice of interpretation of the generalized delta function is, in some ways, a matter of taste. For the student, the limit interpretation of the delta function is perhaps the easier to digest at first, although the dual, linear operator interpretation has stronger connections with the rigorous theory and, even in applications, offers some significant advantages.

Although on the surface, the delta function might look a little bizarre, its utility in modern applied mathematics and mathematical physics more than justifies including it in your analytical toolbox. Even if you are not yet comfortable with either definition, you are

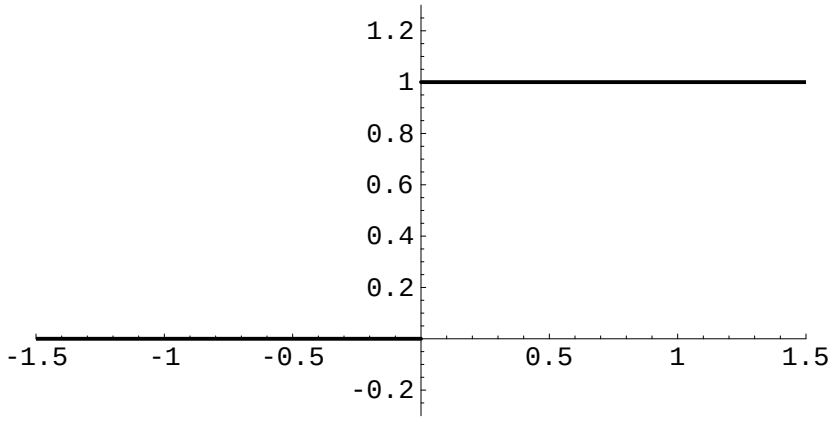


Figure 11.5. The Step Function.

advised to press on to gain a good working relationship with the delta function through its basic properties. You usually won't go far wrong by treating it as if it were a genuine function. After you gain experience in working with it as a practical tool, you can, if desired, return to contemplate just exactly what kind of object the delta function really is.

Calculus of Generalized Functions

Since we are going to use the delta function to solve differential equations, we need to find out how it behaves under the basic operations of calculus — differentiation and integration. The integral of the delta function is known as a *step function*. More specifically, the basic formulae (11.37), (11.38) imply that

$$\int_a^x \delta_y(t)\,dt = \sigma_y(x) = \sigma(x-y) = \begin{cases} 0, & x < y, \\ 1, & x > y. \end{cases} \qquad \text{provided} \qquad a < y. \tag{11.42}$$

Figure 11.5 shows the graph of $\sigma(x) = \sigma_0(x)$. Unlike the delta function, the step function $\sigma_y(x)$ is an ordinary function. It is continuous — indeed constant — except at $x = y$. The value of the step function at the discontinuity $x = y$ is left unspecified, although a popular choice, motivated by Fourier theory, is to set $\sigma_y(y) = \frac{1}{2}$, the average of its left and right hand limits.

We observe that the integration formula (11.42) is compatible with our characterization of the delta function as the limit of highly concentrated forces. If we integrate the approximating functions (11.32), we obtain

$$f_n(x) = \int_{-\infty}^x g_n(t)\,dt = \frac{1}{\pi}\,\tan^{-1} n\,x + \frac{1}{2}\,.$$

Since

$$\lim_{y \to \infty} \tan^{-1} y = \tfrac{1}{2}\,\pi, \qquad \text{while} \qquad \lim_{y \to -\infty} \tan^{-1} y = -\tfrac{1}{2}\,\pi,$$

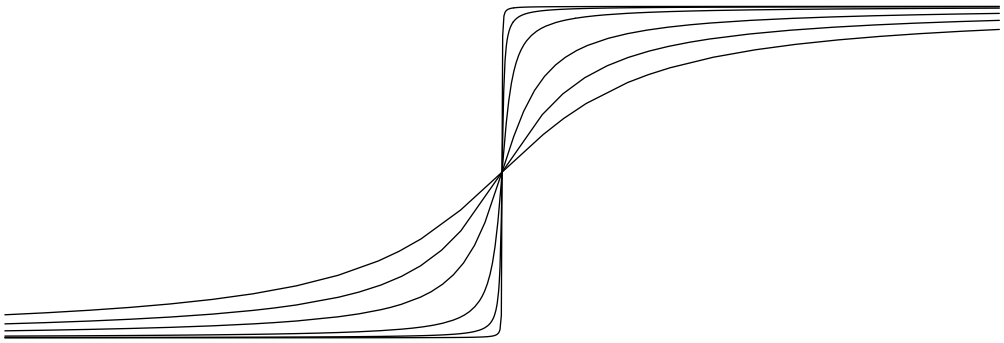


Figure 11.6. Step Function as Limit.

these functions converge to the step function:

$$\lim_{n \to \infty} f_n(x) = \sigma(x) = \begin{cases} 1, & x > 0, \\ \frac{1}{2}, & x = 0, \\ -1, & x < 0. \end{cases} \tag{11.43}$$

A graphical illustration of this limiting procedure is sketched in Figure 11.6.

Motivated by the Fundamental Theorem of Calculus, we shall use (11.42) to identify the derivative of the step function with the delta function

$$\frac{d\sigma}{dx} = \delta\,. \tag{11.44}$$

This fact is highly significant. In basic calculus, one is not allowed to differentiate a discontinuous function. Here, we discover that the derivative is defined, not as an ordinary function, but rather as a generalized delta function.

This observation is a particular instance of a general result. We use

$$f(y^-) = \lim_{x \to y^-} f(x), \qquad\qquad f(y^+) = \lim_{x \to y^+} f(x), \tag{11.45}$$

to denote, respectively, the left and right sided limits of a function at a point y. The function $f(x)$ is *continuous* at the point y if and only if its one-sided limits exist and are equal to its value: $f(y) = f(y^-) = f(y^+)$. If the one-sided limits are the same, but not equal to its value $f(y)$, then the function is said to have a *removable discontinuity*, since by redefining $f(y) = f(y^-) = f(y^+)$ makes it continuous at the point in question. An example is the function $f(x)$ that is equal to 0 for all $x \neq 0$, but has[†] $f(0) = 1$. Removing the discontinuity by setting $f(0) = 0$ makes $f(x) \equiv 0$ equal to the continuous constant 0 function. Removable discontinuities play no role in our theory or applications, and will always be removed if they appear.

[†] This function is *not* a version of the delta function — its integral is 0, not 1. Also, removable discontinuities only occur in ordinary functions; 0 is *not* a removable discontinuity for the delta function.

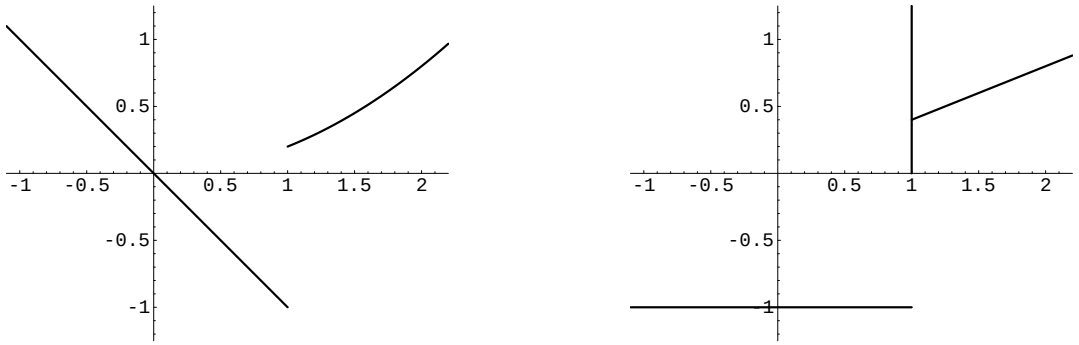


Figure 11.7. The Derivative of a Discontinuous Function.

Finally, if both the left and right limits exist, but are not equal, then f is said to have a *jump discontinuity* at the point y. The *magnitude* of the jump is the difference

$$\beta = f(y^+) - f(y^-) = \lim_{x \to y^+} f(x) - \lim_{x \to y^-} f(x), \tag{11.46}$$

between the right and left limits. Note the value of the function at the point, $f(y)$, which may not even be defined, does not play a role in the specification of the jump. The magnitude of the jump is positive if the function jumps up, when moving from left to right, and negative for a downwards jump. For example, the step function $\sigma(x)$ has a unit, i.e., magnitude 1, jump discontinuity at the origin:

$$\sigma(0^+) - \sigma(0^-) = 1 - 0 = 1,$$

and is continuous everywhere else.

In general, the derivative of a function with jump discontinuities is a generalized function that includes delta functions concentrated at each discontinuity. More explicitly, suppose that $f(x)$ is differentiable, in the usual calculus sense, everywhere except at the point y where it has a jump discontinuity of magnitude β. We can re-express the function in the convenient form

$$f(x) = g(x) + \beta\, \sigma(x - y), \tag{11.47}$$

where $g(x)$ is continuous everywhere, and differentiable except possibly at the jump. Differentiating (11.47), we find

$$f'(x) = g'(x) + \beta\, \delta(x - y), \tag{11.48}$$

has a delta spike of magnitude β at the discontinuity. Thus, the derivatives of f and g coincide everywhere except at the discontinuity.

Example 11.4. Consider the function

$$f(x) = \begin{cases} -x, & x < 1, \\ \frac{1}{5}\, x^2, & x > 1, \end{cases} \tag{11.49}$$

which we graph in Figure 11.7. We note that f has a single jump discontinuity of magnitude $\frac{6}{5}$ at $x = 1$. This means that

$$f(x) = g(x) + \tfrac{6}{5}\, \sigma(x - 1), \qquad \text{where} \qquad g(x) = \begin{cases} -x, & x < 1, \\ \frac{1}{5}\, x^2 - \frac{6}{5}, & x > 1, \end{cases}$$

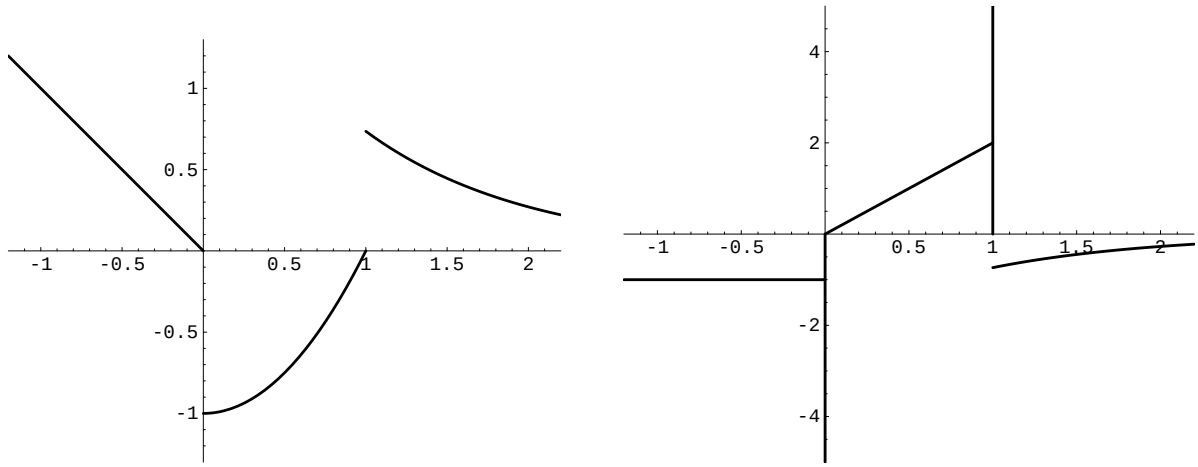


Figure 11.8. The Derivative of a Discontinuous Function.

is continuous everywhere, since its right and left hand limits at the original discontinuity are equal: $g(1^+) = g(1^-) = -1$. Therefore,

$$f'(x) = g'(x) + \tfrac{6}{5}\,\delta(x-1), \qquad \text{where} \qquad g'(x) = \begin{cases} -1, & x < 0, \\ \frac{2}{5}\,x, & x > 1. \end{cases}$$

In Figure 11.7, the delta spike in the derivative of f is symbolized by a vertical line — although this pictorial device fails to indicate its magnitude of $\frac{6}{5}$. Note that $g'(x)$ can be found by directly differentiating the formula (11.49) for $f(x)$. This implies that, once we determine the magnitude and location of the jump discontinuities of $f(x)$, we can compute its derivative directly without introducing to the auxiliary continuous function $g(x)$.

Example 11.5. As a second example, consider the function

$$f(x) = \begin{cases} -x, & x < 0, \\ x^2 - 1, & 0 < x < 1, \\ 2\,e^{-x}, & x > 1, \end{cases}$$

which is plotted in Figure 11.8. This function has jump discontinuities of magnitude -1 at $x = 0$, and of magnitude $2/e$ at $x = 1$. Therefore, in light of the preceding remark,

$$f'(x) = -\,\delta(x) + \frac{2}{e}\,\delta(x-1) + \begin{cases} -1, & x < 0, \\ 2\,x, & 0 < x < 1, \\ -2\,e^{-x}, & x > 1, \end{cases}$$

where the final terms are obtained by directly differentiating $f(x)$.

The integral of the discontinuous step function (11.42) is the continuous ramp function,

$$\int_a^x \sigma_y(z)\,dz = \rho_y(x) = \rho(x-y) = \begin{cases} 0, & a < x < y, \\ x - y, & x > y > a, \end{cases} \tag{11.50}$$

which is graphed in Figure 11.9. Note that $\rho(x-y)$ has a corner at $x = y$, and so is not differentiable there; indeed, its derivative $\dfrac{d\rho}{dx} = \sigma$ has a jump discontinuity, and its second

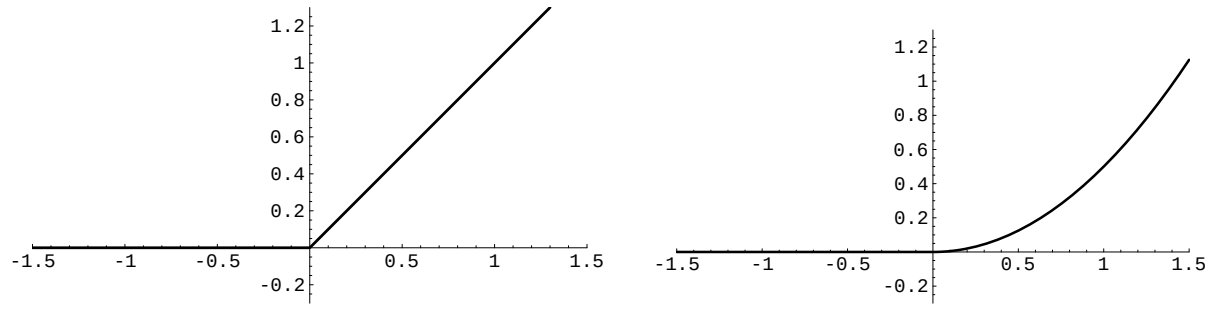


Figure 11.9. First and Second Order Ramp Functions.

derivative $\dfrac{d^2\rho}{dx^2} = \delta$ is no longer an ordinary function. We can continue to integrate; the n^{th} integral of the delta function is the n^{th} order ramp function

$$\rho_n(x-y) = \begin{cases} \dfrac{(x-y)^n}{n!}, & x > y, \\ 0, & x < y. \end{cases} \tag{11.51}$$

Example 11.6. The derivative of the absolute value function

$$a(x) = |\, x \,| = \begin{cases} x, & x > 0, \\ -x, & x < 0, \end{cases}$$

is the *sign function*

$$s(x) = a'(x) = \begin{cases} +1, & x > 0, \\ -1, & x < 0. \end{cases} \tag{11.52}$$

Note that there is no delta function in $a'(x)$ because $a(x)$ is continuous everywhere. Since $s(x)$ has a jump of magnitude 2 at the origin and is otherwise constant, its derivative $s'(x) = a''(x) = 2\,\delta(x)$ is twice the delta function.

Conversely, we can also differentiate the delta function. Its first derivative

$$\delta_y'(x) = \delta\,'(x-y)$$

can be interpreted in two ways. First, we may view $\delta\,'(x)$ as the limit

$$\frac{d\delta}{dx} = \lim_{n\to\infty} \frac{dg_n}{dx} = \lim_{n\to\infty} \frac{-2\,n^3\,x}{\pi(1+n^2\,x^2)^2} \tag{11.53}$$

of the derivatives of the approximating functions (11.32). The graphs of these rational functions take the form of more and more concentrated spiked "doublets", as illustrated in Figure 11.10. To determine the effect of the derivative on a test function $u(x)$, we compute the limiting integral

$$\begin{aligned} \langle\, \delta\,' ; u \,\rangle &= \int_{-\infty}^{\infty} \delta\,'(x)\, u(x)\, dx = \lim_{n\to\infty} \int_{-\infty}^{\infty} g_n'(x)\, u(x)\, dx \\ &= -\lim_{n\to\infty} \int_{-\infty}^{\infty} g_n(x)\, u'(x)\, dx = -\int_{-\infty}^{\infty} \delta(x)\, u'(x)\, dx = -\,u'(0). \end{aligned} \tag{11.54}$$

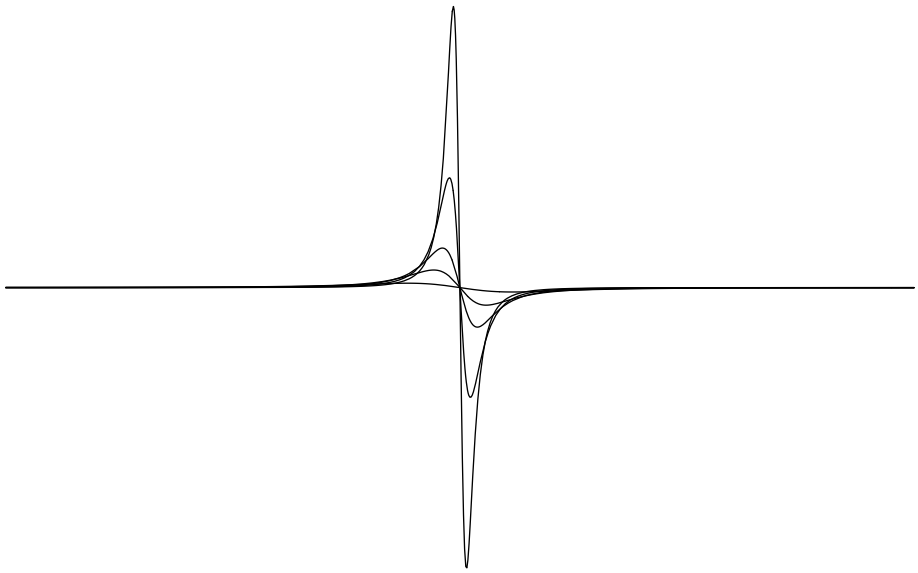


Figure 11.10. Derivative of Delta Function as Limit of Doublets.

In the middle step, we used an integration by parts; the boundary terms at $\pm\infty$ vanish provided $u(x)$ is continuously differentiable and bounded as $|\,x\,| \to \infty$. Pay attention to the minus sign in the final answer.

In the dual interpretation, the generalized function $\delta_y'(x)$ corresponds to the linear operator

$$L_y'[u] = -\,u'(y) = \langle\, \delta_y' \,;\, u \,\rangle = \int_0^\ell \delta_y'(x)\, u(x)\, dx, \qquad \text{where} \qquad 0 < y < \ell, \tag{11.55}$$

that maps a continuously differentiable function $u(x)$ to *minus* its derivative at the point y. We note that (11.55) is compatible with a formal integration by parts

$$\int_0^\ell \delta\,'(x-y)\, u(x)\, dx = \delta(x-y)\, u(x)\,\Big|_{x=0}^{\ell} - \int_0^\ell \delta(x-y)\, u'(x)\, dx = -\,u'(y).$$

The boundary terms at $x = 0$ and $x = \ell$ automatically vanish since $\delta(x-y) = 0$ for $x \neq y$.

Warning: The functions $\widetilde{g}_n(x) = g_n(x) + g_n'(x)$ satisfy $\lim_{n\to\infty} \widetilde{g}_n(x) = 0$ for all $x \neq y$, while $\displaystyle\int_{-\infty}^{\infty} \widetilde{g}_n(x)\, dx = 1$. However, $\lim_{n\to\infty} \widetilde{g}_n = \lim_{n\to\infty} g_n + \lim_{n\to\infty} g_n' = \delta + \delta\,'$. Thus, our original conditions (11.30), (11.31) are *not* in fact sufficient to characterize whether a sequence of functions has the delta function as a limit. To be absolutely sure, one must, in fact, verify the more comprehensive limiting definition (11.41).

The Green's Function

To further cement our new-found friendship, we now discuss how the delta function is used to solve inhomogeneous boundary value problems. Consider a bar of length ℓ subject

to a unit impulse force $\delta_y(x) = \delta(x-y)$ concentrated at position $0 < y < \ell$ along the bar. The underlying differential equation (11.12) takes the form

$$-\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = \delta(x-y), \qquad 0 < x < \ell. \tag{11.56}$$

Coupled with the appropriate boundary conditions, this represents the continuum analog of the discrete unit impulse equilibrium system (11.26). The solution to the boundary value problem associated with (11.56) is known as the *Green's function*, and will be denoted by $G_y(x) = G(x,y)$.

Example 11.7. Let us look at the simple case of a homogeneous bar with uniform stiffness $c(x) \equiv 1$, of unit length $\ell = 1$, and fixed at both ends. The boundary value problem for the Green's function $G(x,y)$ takes the form

$$-u'' = \delta(x-y), \qquad u(0) = 0 = u(1), \tag{11.57}$$

where $0 < y < 1$ indicates the point at which we apply the impulse force. The solution to the differential equation is obtained directly by two integrations. First, by (11.42),

$$u'(x) = -\sigma(x-y) + a,$$

where a is a constant of integration. A second integration leads to

$$u(x) = -\rho(x-y) + a\,x + b, \tag{11.58}$$

where ρ is the ramp function (11.50). The integration constants a, b are fixed by the boundary conditions; since $0 < y < 1$, we have

$$u(0) = b = 0, \qquad u(1) = -(1-y) + a + b = 0, \qquad \text{and so} \qquad a = 1 - y.$$

Therefore, the Green's function for the problem is

$$G(x,y) = -\rho(x-y) + (1-y)x = \begin{cases} x(1-y), & x \le y, \\ y(1-x), & x \ge y, \end{cases} \tag{11.59}$$

See Figure 11.11 for a graph of $G(x,y)$. Note that, for each fixed y, it is a continuous and piecewise affine function of x — meaning that its graph consists of connected straight line segments, with a corner where the unit impulse force is being applied.

We observe the following fundamental properties that serve to uniquely characterize the Green's function (11.59). First, since the delta forcing vanishes except at the point $x = y$, the Green's function satisfies the homogeneous differential equation[†]

$$\frac{\partial^2 G}{\partial x^2}\,(x,y) = 0 \qquad \text{for all} \qquad x \neq y. \tag{11.60}$$

[†] Since $G(x,y)$ is a function of two variables, we switch to partial derivative notation to indicate its derivatives.

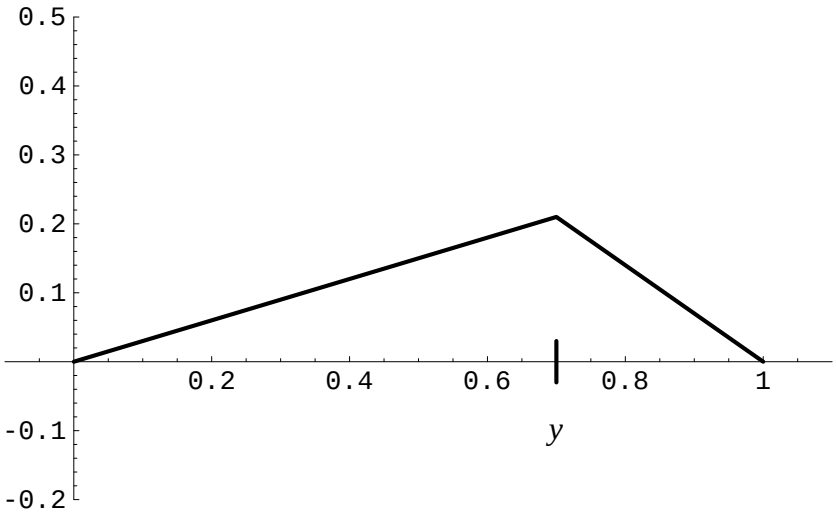


Figure 11.11. Green's function for a Bar with Fixed Ends.

Secondly, by construction, it must satisfy the boundary conditions,

$$G(0, y) = 0 = G(1, y).$$

Thirdly, G is continuous, but has a 90° corner at $x = y$, which implies that its derivative $\partial G/\partial x$ has a jump discontinuity of magnitude -1 there. The second derivative $\partial^2 G/\partial x^2$ has a delta function discontinuity at $x = y$, and thereby solves the original boundary value problem (11.57). Finally, we observe that the Green's function is symmetric in x and y:

$$G(x, y) = G(y, x). \tag{11.61}$$

This symmetry property is a consequence of the underlying symmetry or "self-adjointness" of the boundary value problem; this aspect will be discussed in more depth in the following section. Symmetry has the interesting physical consequence that the response of the bar at position x due to an impulse force concentrated at position y is exactly the same as the response of the bar at position y due to an impulse being applied at position x. This turns out to be a rather general, although perhaps unanticipated phenomenon. analogous results for the discrete cases of mass-spring chains, circuits, and structures can be found in Exercises ■, ■ and ■.

Once we have determined the Green's function for the system, we can solve the general forced boundary value problem

$$-u'' = f(x), \qquad u(0) = 0 = u(1), \tag{11.62}$$

by linear superposition — in direct analogy with the superposition solution (11.27) of the discrete problem. In the continuum case, we need to express the forcing function $f(x)$ as a linear combination of impulses that are concentrated at each point along the bar. Since there is a continuum of possible positions y at which the impulse forces may be applied, we need to replace the finite sum by an integral, writing the external force as

$$f(x) = \int_0^1 f(y)\, \delta(x - y)\, dy. \tag{11.63}$$

We will interpret (11.63) as the continuous superposition of an infinite collection of impulses $f(y)\,\delta(x-y)$, of respective magnitudes $f(y)$ and concentrated at position y.

The general linear superposition principle states that linear combinations of inhomogeneities produce linear combinations of solutions. Again, we adapt this principle to the continuum by replacing the sums by integrals. (Indeed, the original definition of the Riemann integral is as a limit of Riemann sums, [**9**].) Thus, we write the differential equation (11.62) as

$$-u'' = \int_0^1 f(y)\,\delta(x-y)\,dy,$$

and write the solution as the same continuous superposition

$$u(x) = \int_0^1 f(y)\,G(x,y)\,dy \tag{11.64}$$

of the Green's function solutions to the individual unit impulse problems.

For the particular boundary value problem (11.62), plugging (11.59) into (11.64), and breaking the integral up into two parts, for $y < x$ and $y > x$, we arrive at an explicit formula

$$u(x) = \int_0^x (1-x)\,y\,f(y)\,dy + \int_x^1 x\,(1-y)\,f(y)\,dy \tag{11.65}$$

for its solution. For example, under a constant unit force $f(x) \equiv 1$, the solution (11.65) is

$$u(x) = \int_0^x (1-x)\,y\,dy + \int_x^1 x\,(1-y)\,dy = \tfrac{1}{2}\,(1-x)\,x^2 + \tfrac{1}{2}\,x\,(1-x)^2 = \tfrac{1}{2}\,x - \tfrac{1}{2}\,x^2,$$

in agreement with (11.20) when $\alpha = 1$ and $d = 0$. Although this relatively simple problem was perhaps easier to solve directly, the use of the Green's function has the advantage of providing a unified framework that fits all of the special solution techniques for inhomogeneous boundary value problems, and really comes into its own in higher dimensional situations.

Let us, finally, convince ourselves that the superposition formula (11.65) does indeed give the correct answer. First,

$$\begin{aligned}\frac{du}{dx} &= (1-x)\,x\,f(x) + \int_0^x [-y\,f(y)]\,dy - x\,(1-x)\,f(x) + \int_x^1 (1-y)\,f(y)\,dy\\ &= -\int_0^1 y\,f(y)\,dy + \int_x^1 f(y)\,dy.\end{aligned}$$

Differentiating again, we conclude that

$$\frac{d^2u}{dx^2} = -\,f(x),$$

as desired. As with all limiting processes, one must always be careful interchanging the order of differentiation and integration. In all the examples considered here, the integrand will always be sufficiently nice to allow this to be done.

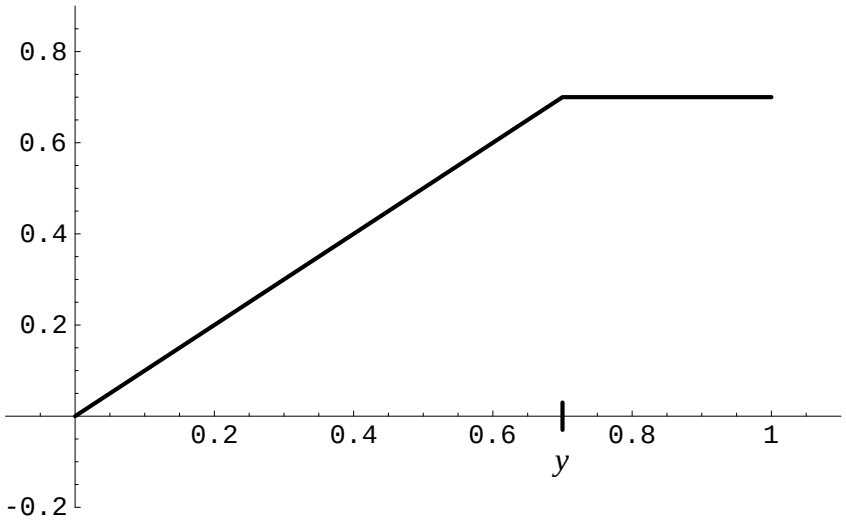


Figure 11.12. Green's Function for Bar with One Fixed and One Free End.

Remark: In computing the derivatives of u, we made use of the calculus formula

$$\frac{d}{dx}\int_{\alpha(x)}^{\beta(x)} F(x,y)\,dy = F(x,\beta(x))\,\frac{d\beta}{dx} - F(x,\alpha(x))\,\frac{d\alpha}{dx} + \int_{\alpha(x)}^{\beta(x)} \frac{\partial F}{\partial x}\,(x,y)\,dy \tag{11.66}$$

for the derivative of an integral with variable limits, which is a straightforward consequence of the Fundamental Theorem of Calculus and the Chain Rule, [**9**].

Example 11.8. Consider next a uniform bar of length $\ell = 1$ with one fixed and one free end. To determine the Green's function, we must solve the boundary value problem

$$-c\,u'' = \delta(x-y), \qquad u(0) = 0, \qquad u'(1) = 0, \tag{11.67}$$

where c is the elastic constant of the bar. Integrating twice, we find the general solution to the differential equation can be written in terms of the ramp function

$$u(x) = -\frac{1}{c}\,\rho(x-y) + a\,x + b.$$

The integration constants a, b are fixed by the boundary conditions

$$u(0) = b = 0, \qquad u'(1) = -\frac{1}{c} + a = 0.$$

Therefore, the Green's function for this problem is

$$G(x,y) = \begin{cases} x/c, & x \le y, \\ y/c, & x \ge y, \end{cases} \tag{11.68}$$

and is graphed in Figure 11.12.

As in the previous example, the Green's function is piecewise affine, and so solves the homogeneous differential equation $-u'' = 0$, except at the impulse point $x = y$, where it has a corner. Its first derivative $\partial G/\partial x$ has a jump discontinuity of magnitude $-1/c$, as

required for its second derivative $\partial^2 G/\partial x^2$ to produce the correct delta function singularity. Moreover, it satisfies both boundary conditions. Note that $G(x,y)$ is constant for $x > y$ since the unit impulse at $x = y$ will only stretch the section of the bar that lies above it, while the section below the impulse hangs freely. And, as in the preceding example, the Green's function is a symmetric function: $G(x,y) = G(y,x)$, which admits a similar physical interpretation.

Once we have the Green's function in hand, the solution to the general forced boundary value problem

$$-cu'' = f(x), \qquad u(0) = 0, \qquad u'(1) = 0, \tag{11.69}$$

can be expressed using a superposition formula (11.64), namely

$$u(x) = \int_0^1 G(x,y) f(y)\, dy \;=\; \frac{1}{c}\int_0^x x\, f(y)\, dy + \frac{1}{c}\int_x^1 y\, f(y)\, dy. \tag{11.70}$$

The reader may wish to verify this directly, as we did in the previous example.

Let us conclude this section by summarizing the fundamental properties that characterize the Green's function of a boundary value problem.

Basic Properties of the Green's Function

(a) Solves the homogeneous differential equation:

$$-\frac{\partial}{\partial x}\left(c(x)\,\frac{\partial}{\partial x}\, G(x,y)\right) = 0, \qquad \text{for all} \qquad x \neq y. \tag{11.71}$$

(b) Satisfies the boundary conditions.

(c) Has a jump discontinuity of magnitude $-1/c(y)$ in its derivative $\partial G/\partial x$ at $x = y$.

(d) Is a symmetric function of its arguments: $G(y,x) = G(x,y)$.

(e) Admits a superposition principle for general forcing functions:

$$u(x) = \int_0^\ell G(x,y)\, f(y)\, dy. \tag{11.72}$$

Although derived and stated for the simple case of a one-dimensional boundary value problem governing the equilibrium solution of a bar, these properties, suitably adapted, hold in a very broad range of boundary value problems, including those of higher order and in more dimensions.

Remark: The Green's function represents the continuum limit of the inverse $G = K^{-1}$ of the stiffness matrix of a discrete mass-spring chain, (11.14). The entries G_{ij} of the inverse matrix are approximations to the sampled values $G(x_i, x_j)$ of the Green's function of the limiting bar. Symmetry, $G(x_i, x_j) = G(x_j, x_i)$, of the Green's function corresponds to symmetry, $G_{ij} = G_{ji}$, of the inverse of the symmetric stiffness matrix. In Exercise ∎, you are asked to study this limiting procedure in detail.

11.3. Adjoints and Minimum Principles.

Let us now discuss how the boundary value problems for continuous elastic bars fit into our general equilibrium framework of positive (semi-)definite linear systems. In Chapter 6, we learned that the stable equilibrium configurations of discrete mechanical and electrical systems can be characterized as energy minimizers. This fundamental physical principle has a direct counterpart in the continuum systems that we have begun to consider, and our goal is to understand how to adapt the finite-dimensional constructions to infinite-dimensional function space. In particular, the resulting minimization principles not only lead to an important theoretical characterization of the equilibrium solution, but, through the finite element method, underlies the most important class of numerical approximation algorithms for such boundary value problems.

Adjoints of Differential Operators

In discrete mechanical systems, the crucial observation was that the matrix appearing in the force balance law is the transpose of the incidence matrix relating displacements and elongations or strains. In the continuum limit, the discrete incidence matrix has turned into a differential operator, and so a crucial difficulty is how to take its "transpose". The abstract answer to this quandary can be found in Section 7.5. The transpose of a matrix is a particular instance of the general notion of the *adjoint* of a linear function, which relies on the specification of inner products on its domain and target spaces. In the case of the matrix transpose, the adjoint is prescribed with respect to the standard dot product on Euclidean space. Thus, the correct interpretation of the "transpose" of a differential operator is as the adjoint linear operator with respect to suitable inner products on function space.

For bars and similar continuous one-dimensional media, the role of the incidence matrix is played by the derivative $v = D[u] = du/dx$, which defines a linear operator $D\colon U \to V$ from the vector space of possible displacements $u(x)$, denoted by U, to the vector space of possible strains $v(x)$, denoted by V. In order to compute its adjoint, we need to impose inner products on both the displacement space U and the strain space V. The simplest situation is to adopt the same standard L^2 inner product

$$\langle\, u\,;\widetilde{u}\,\rangle = \int_a^b u(x)\,\widetilde{u}(x)\,dx, \qquad\qquad \langle\!\langle\, v\,;\widetilde{v}\,\rangle\!\rangle = \int_a^b v(x)\,\widetilde{v}(x)\,dx, \tag{11.73}$$

on both vector spaces. These are the continuum analogs of the Euclidean dot product, and, as we shall see, will be appropriate when dealing with homogeneous bars. According to the defining equation (7.64), the adjoint D^* of the derivative operator must satisfy the inner product identity

$$\langle\!\langle\, D[u]\,;v\,\rangle\!\rangle = \langle\, u\,;D^*[v]\,\rangle \qquad \text{for all} \qquad u \in U, \quad v \in V. \tag{11.74}$$

First, we compute the left hand side:

$$\langle\!\langle\, D[u]\,;v\,\rangle\!\rangle = \left\langle\!\!\left\langle\, \frac{du}{dx}\,;\, v\,\right\rangle\!\!\right\rangle = \int_a^b \frac{du}{dx}\; v\; dx. \tag{11.75}$$

On the other hand, the right hand side should equal

$$\langle\, u\,;D^*[\,v\,]\,\rangle = \int_a^b u\, D^*[\,v\,]\,dx. \tag{11.76}$$

Now, in the latter integral, we see u multiplying the result of applying the linear operator D^* to v. To identify this integrand with that in the previous integral (11.75), we need to somehow remove the derivative from u. The secret is integration by parts! It allows us to rewrite the first integral in the form

$$\int_a^b \frac{du}{dx}\, v\,dx = \big[\, u(b)\,v(b) - u(a)\,v(a)\,\big] - \int_a^b u\,\frac{dv}{dx}\,dx. \tag{11.77}$$

If we ignore the boundary terms $u(b)\,v(b) - u(a)\,v(a)$ for a moment, then the remaining integral is equal to an inner product

$$-\int_a^b u\,\frac{dv}{dx}\,dx = \int_a^b u\left(-\frac{dv}{dx}\right)dx = \left\langle\, u\,;-\frac{dv}{dx}\,\right\rangle = \langle\, u\,;-D[\,v\,]\,\rangle. \tag{11.78}$$

Equating (11.76), (11.78), we deduce that

$$\langle\!\langle\, D[\,u\,]\,;v\,\rangle\!\rangle = \left\langle\!\!\left\langle\, \frac{du}{dx}\,;v\,\right\rangle\!\!\right\rangle = \left\langle\, u\,;-\frac{dv}{dx}\,\right\rangle = \langle\, u\,;-D[\,v\,]\,\rangle.$$

Thus, to satisfy, (11.76), we must have

$$\langle\, u\,;D^*[\,v\,]\,\rangle = \langle\, u\,;-D[\,v\,]\,\rangle \qquad \text{for all} \qquad u\in U, \quad v\in V,$$

and so

$$D^* = \left(\frac{d}{dx}\right)^* = -\,\frac{d}{dx} = -\,D. \tag{11.79}$$

The final equation confirms our earlier identification (11.4) of the derivative operator as the continuum limit of the incidence matrix A and its negative as the limit (11.10) of the transposed (or adjoint) incidence matrix $A^T = A^*$.

However, the preceding argument is *only* valid if the boundary terms in (11.77) vanish:

$$u(b)\,v(b) - u(a)\,v(a) = 0, \tag{11.80}$$

which necessitates imposing suitable boundary conditions on the functions u and v. For example, in the case of a bar with both ends fixed, the boundary conditions

$$u(a) = 0, \qquad\qquad u(b) = 0, \tag{11.81}$$

will ensure that (11.80) holds, and therefore validate (11.79). The homogeneous boundary conditions serve to define the vector space

$$U = \big\{\, u(x)\in \mathrm{C}^1[\,a,b\,] \;\big|\; u(a) = u(b) = 0\,\big\}$$

of allowable displacements, consisting of all continuously differentiable functions that vanish on the boundary.

The fixed boundary conditions (11.81) are not the only possible ones that ensure the vanishing of the boundary terms (11.80). An evident alternative is to require that the strain v vanish at both endpoints, $v(a) = v(b) = 0$. This is the case of an unsupported bar with two free ends, where the displacement at the ends is unspecified, but the strain vanishes owing to a lack of support. In this case, the strain space

$$V = \big\{\, v(x) \in \mathrm{C}^1[a,b] \;\big|\; v(a) = v(b) = 0 \,\big\}$$

consists of all functions that vanish on the boundary. Since the derivative $D\colon U \to V$ must map a displacement $u(x)$ to an allowable strain $v(x)$, the vector space of allowable displacements takes the form

$$U = \big\{\, u(x) \in \mathrm{C}^1[a,b] \;\big|\; u'(a) = u'(b) = 0 \,\big\},$$

indicating free boundary conditions at both ends, as in Example 11.3. Again, restricting $D\colon U \to V$ to these particular vector spaces ensures that the boundary terms (11.80) vanish, and so (11.79) holds in this situation too.

Let us list the most important combinations of boundary conditions that will imply the vanishing of the boundary terms (11.80) and ensure that the desired adjoint equation $D^* = -D$ is valid. In all cases, the boundary conditions impose restrictions on the displacement space U and, in cases (b–d), the strain space V too.

Self-Adjoint Boundary Conditions for a Bar

a)	Both ends fixed:	$u(a) = u(b) = 0.$
b)	One free and one fixed end:	$u(a) = 0, \quad u'(b) = 0 \quad \text{or} \quad u'(a) = 0, \quad u(b) = 0.$
c)	Both ends free:	$u'(a) = u'(b) = 0.$
d)	Periodic boundary conditions:	$u(a) = u(b), \quad u'(a) = u'(b).$

A fixed boundary condition $u(a) = 0$ is commonly referred to as a *Dirichlet boundary condition*, in honor the nineteenth century French analyst Lejeune Dirichlet. A free boundary condition $u'(a) = 0$ is known as a *Neumann boundary condition*, in honor of his German contemporary Carl Gottfried Neumann. The Dirichlet boundary value problem has both ends fixed, while the Neumann boundary value problem has both ends free. The intermediate case (b) is known as a *mixed* boundary value problem. The periodic boundary conditions represent a bar that has its ends joined together for form a circular[†] elastic ring. It represents the continuum limit of the periodic mass–spring chain discussed in Exercise ■, and requires $u(x)$ to be a periodic function with period $b - a$.

In the case of a homogeneous bar with stiffness $c(x) \equiv 1$, the connections between strain, displacement and external force take the form

$$v = D[u] = u', \qquad\qquad f = D^*[v] = -v',$$

[†] we are assuming that the circle is sufficiently large that we can ignore any curvature effects.

provided we impose a suitable pair of boundary conditions. The equilibrium equations are then written in the self-adjoint form

$$K[u] = f, \qquad \text{where} \qquad K = D^* \circ D = -D^2. \tag{11.82}$$

Note that

$$K^* = (D^* \circ D)^* = D^* \circ (D^*)^* = D^* \circ D = K, \tag{11.83}$$

which proves self-adjointness of the differential operator; in complete detail,

$$\langle\, K[u]\,;\widetilde{u}\,\rangle = \int_0^\ell \big[\, -u''(x)\,\widetilde{u}(x)\,\big]\,dx = \int_0^\ell \big[\, -u(x)\,\widetilde{u}''(x)\,\big]\,dx = \langle\, u\,;K[\widetilde{u}]\,\rangle \tag{11.84}$$

for all displacements $u, \widetilde{u} \in U$. A direct verification of this formula relies on two integration by parts, employing the Dirichlet boundary conditions to cancel out the ensuing boundary terms.

To deal with inhomogeneous materials in the same framework, we must modify the inner products on the underlying function spaces. To this aim, we retain the ordinary L^2 inner product

$$\langle\, u\,;\widetilde{u}\,\rangle = \int_a^b u(x)\,\widetilde{u}(x)\,dx, \qquad\qquad u, \widetilde{u} \in U, \tag{11.85}$$

on the vector space of possible displacements, but adopt a weighted inner product

$$\langle\!\langle\, v\,;\widetilde{v}\,\rangle\!\rangle = \int_a^b v(x)\,\widetilde{v}(x)\,c(x)\,dx, \qquad\qquad v, \widetilde{v} \in V, \tag{11.86}$$

on the space of strain functions. The weight function $c(x) > 0$ turns out to be the stiffness function for the bar, and so its positivity corroborates the underlying physical hypotheses.

Let us compute the adjoint of the derivative operator $D\colon U \to V$ with respect to these two inner products (11.85), (11.86). Now we need to compare

$$\langle\!\langle\, D[u]\,;v\,\rangle\!\rangle = \int_a^b \frac{du}{dx}\; v(x)\,c(x)\,dx, \qquad \text{with} \qquad \langle\, u\,;D^*[v]\,\rangle = \int_a^b u(x)\,D^*[v]\,dx.$$

Integrating the first expression by parts, we find

$$\int_a^b \frac{du}{dx}\; c\,v\,dx = \big[\, u(b)\,c(b)\,v(b) - u(a)\,c(a)\,v(a)\,\big] - \int_a^b u\,\frac{d(c\,v)}{dx}\,dx = \int_a^b u\left(-\frac{d(c\,v)}{dx}\right)dx, \tag{11.87}$$

provided we choose our boundary conditions so that the boundary terms vanish:

$$u(b)\,c(b)\,v(b) - u(a)\,c(a)\,v(a) = 0,$$

which holds from any of the listed boundary conditions: Dirichlet, Neumann or mixed, as well as the periodic case provided $c(a) = c(b)$. Therefore, in such cases, the weighted adjoint D^* of the derivative operator is

$$\left(\frac{d}{dx}\right)^* v = -\,\frac{d(c\,v)}{dx} = -\,c\,\frac{dv}{dx} - c'\,v. \tag{11.88}$$

The self-adjoint combination $K = D^* \circ D$ is now given by

$$K[u] = -\frac{d}{dx}\left(c(x)\,\frac{du}{dx} \right), \tag{11.89}$$

which agrees with the differential operator (11.12) for a nonuniform bar. In this way, we have formulated a non-uniform bar in the same abstract self-adjoint form.

Positivity and Minimum Principles

In Chapter 6, we learned that the stiffness matrix of a discrete mechanical system is, in all cases, of positive semi-definite Gram matrix form $K = A^T C\, A$. The existence and uniqueness of solutions, and stability of the system under arbitrary external forcing depends upon whether or not its stiffness matrix is positive definite, $K > 0$. Furthermore, the stable equilibrium configurations can be characterized as minimizers of the quadratic energy function. These fundamental principles all have direct counterparts in the continuum problems now under consideration, as we now discuss.

The first step is to understand how a differential operator or associated boundary value problem can be positive definite. According to the abstract Definition 7.58, a linear operator $K\colon U \to U$ on an inner product space U is *positive definite* provided it is (a) self-adjoint, so $K^* = K$, and (b) satisfies the positivity criterion

$$\langle\, K[u]\,;u\,\rangle > 0, \qquad \text{for all} \qquad 0 \neq u \in U. \tag{11.90}$$

Self-adjointness of the product operator $K = D^* \circ D$ was proved in (11.83). Furthermore, K is positive definite if and only if D has trivial kernel: $\ker D = \{0\}$. Indeed, by the definition of the adjoint,

$$\langle\, K[u]\,;u\,\rangle = \langle\, D^*[\,D[u]\,]\,;u\,\rangle = \langle\!\langle\, D[u]\,;D[u]\,\rangle\!\rangle = \|\,D[u]\,\|^2 \geq 0, \tag{11.91}$$

so K is automatically positive semi-definite. Furthermore,

$$\langle\, K[u]\,;u\,\rangle = \|\,D[u]\,\|^2 = 0 \qquad \text{if and only if} \qquad D[u] = 0,$$

and thus the condition $\ker D = \{0\}$ is necessary and sufficient for the positivity criterion (11.90) to hold.

Now, in the absence of constraints, the kernel of the derivative operator D is *not* trivial, but contains all constant functions. However, we are viewing D as a linear operator on the vector space U of allowable displacements, and so the elements of $\ker D$ must also be allowable, meaning that they must satisfy the boundary conditions. Thus, positivity reduces, in the present situation, to the question of whether or not there are any nontrivial constant functions that satisfy the prescribed homogeneous[†] boundary conditions and hence belonging to $\ker D \subset U$.

Clearly, the only constant function that satisfies a homogeneous Dirichlet boundary conditions is the zero function. Therefore, when restricted to the Dirichlet displacement

[†] The inhomogeneous boundary value problem will be discussed later.

space $U = \{u(0) = u(\ell) = 0\}$, the derivative operator has trivial kernel, $\ker D = \{0\}$. As a result, the composition $K = D^* \circ D$ defines a positive definite linear operator on U. A similar argument applies to the the mixed boundary value problems. Again, the only constant function that satisfies the homogeneous boundary conditions is the zero function, which suffices to prove positive definiteness. Indeed, as we saw, both the Dirichlet and mixed boundary value problems are stable, and admit a unique equilibrium solution under arbitrary external forcing.

On the other hand, any constant function satisfies the homogeneous Neumann boundary conditions, and so $\ker D \subset U$ is the one-dimensional subspace of constant functions. Therefore, the free boundary value problem is only positive semi-definite. A similar argument applies to the periodic problem. Indeed, in such unstable configurations, the boundary value problem has either no solution or infinitely many equilibrium solutions, depending on the nature of the external forcing. Thus, the distinction between stable and unstable systems based on the definiteness of the underlying differential operator is in complete agreement with the finite-dimensional story of Chapter 6.

In the positive definite, stable cases, we can characterize the solution to the (homogeneous) boundary value problem $K[u] = f$ as the unique minimizer of the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2} \| D[u] \|^2 - \langle u ; f \rangle = \int_0^\ell \big[\tfrac{1}{2} c(x) u'(x)^2 - f(x) u(x) \big] \, dx. \tag{11.92}$$

Note that the norm in (11.92) refers to the strain space V, and so is associated with the weighted inner product (11.86); indeed, the first term

$$\tfrac{1}{2} \| D[u] \|^2 = \tfrac{1}{2} \| v \|^2 = \int_0^\ell \tfrac{1}{2} c(x) v(x)^2 \, dx = \int_0^\ell \tfrac{1}{2} v(x) w(x) \, dx = \tfrac{1}{2} \langle v ; w \rangle$$

is one half the (unweighted) inner product between stress and strain, and hence represents the internal energy of our bar. The second term represents the potential energy due to the external forcing, and so, as usual, our minimization principle (11.92) represents the total energy for the mechanical configuration.

Example 11.9. Consider the homogeneous Dirichlet boundary value problem

$$-u'' = f, \qquad u(0) = 0, \qquad u(\ell) = 0. \tag{11.93}$$

for a uniform bar with two fixed ends. This is a stable case, and so the underlying differential operator $K = D^* \circ D = -D^2$, when subject to the boundary conditions, is positive definite. Explicitly, positive definiteness requires

$$\langle K[u] ; u \rangle = \int_0^\ell [-u''(x) u(x)] \, dx = \int_0^\ell u'(x)^2 \, dx > 0 \tag{11.94}$$

for all nonzero $u(x)$ / ~~0~~ satisfying the boundary conditions. Note how we employed an integration by parts, using the boundary conditions to eliminate the boundary terms, to

expose the positivity of the integral. The corresponding minimum principle can be written out as

$$\mathcal{P}[u] = \tfrac{1}{2}\| u' \|^2 - \langle u\,; f \rangle = \int_0^\ell \left[\tfrac{1}{2}\, u'(x)^2 - f(x)\, u(x) \right] dx. \tag{11.95}$$

In other words, the solution $u_\star(x)$ to (11.93) is the function for which $\mathcal{P}[u_\star]$ achieves the minimal value over all possible functions $u(x)$ satisfying the boundary conditions.

While the general, abstract proof of the validity of the minimization principle can be found following Theorem 7.60, a direct verification in this special case may be instructive. As in our derivation of the adjoint operator, it relies on an integration by parts. Since $u_\star'' = -f$, we find

$$\begin{aligned} \mathcal{P}[u] &= \int_0^\ell \left[\tfrac{1}{2}\,(u')^2 + u_\star'' \, u \right] dx = u_\star'(b)\, u(b) - u_\star'(a)\, u(a) + \int_0^\ell \left[\tfrac{1}{2}(u')^2 - u_\star'\, u' \right] dx \\ &= \int_0^\ell \tfrac{1}{2}(u' - u_\star')^2 \, dx + \int_0^\ell \tfrac{1}{2}(u_\star')^2 \, dx. \end{aligned} \tag{11.96}$$

The first integral is always ≥ 0, and is actually equal to 0 if and only if $u' = u_\star'$. Since u and $u_\star$ are both assumed to satisfy the boundary conditions, $\mathcal{P}[u]$ will assume its minimum value when $u = u_\star$.

Inhomogeneous Boundary Conditions

So far, we have restricted our attention to homogeneous boundary value problems. Inhomogeneous boundary conditions a little trickier, since the spaces of allowable displacements and allowable strains are no longer vector spaces, and so the abstract theory, as developed in Chapter 7, will not directly apply.

One way to circumvent this difficulty is to appeal to linear superposition in order to modify the displacement function so as to incorporate the boundary conditions and thereby revert to the homogeneous situation. Consider, for example, the inhomogeneous Dirichlet boundary value problem

$$K[u] = -\frac{d}{dx}\left(c(x)\, \frac{du}{dx} \right) = f(x), \qquad u(0) = \alpha, \qquad u(\ell) = \beta. \tag{11.97}$$

Choose a function $h(x)$ that satisfies the boundary conditions:

$$h(0) = \alpha, \qquad h(\ell) = \beta.$$

Note that we are *not* requiring h to satisfy the differential equation, and so one, but by no means the only, possible choice is the affine function

$$h(x) = \alpha + \frac{\beta - \alpha}{\ell}\, x. \tag{11.98}$$

Since u and h have the same boundary values, their difference

$$\widetilde{u}(x) = u(x) - h(x) \tag{11.99}$$

satisfies the homogeneous Dirichlet boundary conditions

$$\widetilde{u}(0) = \widetilde{u}(\ell) = 0. \tag{11.100}$$

Moreover, by linearity, $\widetilde{u}$ satisfies the modified equation

$$K[\widetilde{u}] = K[u - h] = K[u] - K[h] = f - K[h] \equiv \widetilde{f},$$

or, explicitly,

$$-\frac{d}{dx}\left(c(x)\,\frac{d\widetilde{u}}{dx}\right) = \widetilde{f}(x), \qquad \text{where} \qquad \widetilde{f} = f + \frac{d}{dx}\left(c(x)\,\frac{dh}{dx}\right). \tag{11.101}$$

For the particular choice (11.98),

$$\widetilde{f}(x) = f(x) + \frac{\beta - \alpha}{\ell}\,c'(x).$$

Thus, we have managed to convert the inhomogeneous problem into a homogeneous boundary value problem given in (11.100), (11.101). Once we have solved the latter, the solution to the original inhomogeneous problem is then reconstructed from the formula

$$u(x) = \widetilde{u}(x) + h(x). \tag{11.102}$$

We know that the homogeneous Dirichlet boundary value problem is positive definite, and so we can characterize its solution by a minimum principle, namely as the minimizer of the quadratic energy functional

$$\mathcal{P}[\widetilde{u}] = \tfrac{1}{2}\|\,\widetilde{u}'\,\|^2 - \langle\,\widetilde{u}\,;\widetilde{f}\,\rangle = \int_0^\ell \big[\,\tfrac{1}{2}\,c(x)\,\widetilde{u}'(x)^2 - \widetilde{f}(x)\,\widetilde{u}(x)\,\big]\,dx. \tag{11.103}$$

Let us rewrite the minimization principle in terms of the original displacement function $u(x)$. We replace $\widetilde{u}$ and $\widetilde{f}$ by their formulae (11.99), (11.101); the result is

$$\begin{aligned}\mathcal{P}[\widetilde{u}] &= \tfrac{1}{2}\|\,u' - h'\,\|^2 - \langle\,u - h\,;f - K[h]\,\rangle\\ &= \big(\,\tfrac{1}{2}\|\,u'\,\|^2 - \langle\,u\,;f\,\rangle\,\big) - \big(\,\langle\!\langle\,u'\,;h'\,\rangle\!\rangle - \langle\,u\,;K[h]\,\rangle\,\big) + \big(\,\tfrac{1}{2}\|\,h'\,\|^2 + \langle\,h\,;f - K[h]\,\rangle\,\big)\\ &= \mathcal{P}[u] - \big(\,\langle\!\langle\,u'\,;h'\,\rangle\!\rangle - \langle\,u\,;K[h]\,\rangle\,\big) + C_0.\end{aligned} \tag{11.104}$$

In the middle formula, the first pair of terms reproduces the quadratic energy functional (11.92) for the actual displacement $u(x)$. The last two terms depend only on the initial choice of $h(x)$, and not on $u(x)$; thus, for the purposes of this argument, they can be regarded as a fixed constant, denoted by C_0. The middle two terms can be explicitly evaluated as follows:

$$\begin{aligned}\langle\!\langle\,u'\,;h'\,\rangle\!\rangle - \langle\,u\,;K[h]\,\rangle &= \int_0^\ell \big[\,c(x)\,h'(x)\,u'(x) + \big(\,c(x)\,h'(x)\,\big)'\,u(x)\,\big]\,dx \qquad (11.105)\\ &= \int_0^\ell \frac{d}{dx}\,\big[\,c(x)\,h'(x)\,u(x)\,\big]\,dx = c(\ell)\,h'(\ell)\,u(\ell) - c(0)\,h'(0)\,u(0).\end{aligned}$$

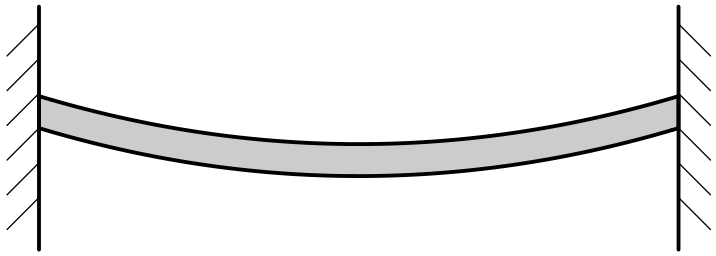


Figure 11.13. Bending of a Beam.

In particular, if $u(x)$ satisfies the inhomogeneous Dirichlet boundary conditions $u(0) = \alpha$, $u(\ell) = \beta$, then these terms

$$\langle\!\langle\, u' \,;\, h' \,\rangle\!\rangle - \langle\, u \,;\, K[\,h\,] \,\rangle = c(\ell)\, h'(\ell)\, \beta - c(0)\, h'(0)\, \alpha \equiv C_1$$

also depend only on the interpolating function h and not on u. Therefore,

$$\mathcal{P}[\,\widetilde{u}\,] = \mathcal{P}[\,u\,] - C_1 + C_0$$

differ by a constant. Consequently, if the function $\widetilde{u}$ minimizes $\mathcal{P}[\,\widetilde{u}\,]$, then $u = \widetilde{u} + h$ necessarily minimizes $\mathcal{P}[\,u\,]$. In this manner, we have characterized the solution to the inhomogeneous Dirichlet boundary value problem by the *same* minimization principle.

Theorem 11.10. *The solution $u_\star(x)$ to the Dirichlet boundary value problem*

$$-\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad u(0) = \alpha, \qquad u(\ell) = \beta,$$

is the unique C^2 *functions that satisfies the indicated boundary conditions and minimizes the energy functional* $\mathcal{P}[\,u\,] = \displaystyle\int_0^\ell \left[\, \tfrac{1}{2}\, c(x)\, u'(x)^2 - f(x)\, u(x) \,\right] dx.$

Warning: When treating the inhomogeneous mixed boundary value problem, we cannot ignore the extra terms (11.105) since they *will* depend upon the choice of function $u(x)$. The details are worked out in Exercise ∎.

11.4. Beams and Splines.

Unlike a bar, which can only stretch in the longitudinal direction, an elastic beam is allowed to bend in a transverse direction. To keep the geometry simple, we consider the case in which the bending of the beam is restricted to the (x, y) plane, as sketched in Figure 11.13. Let $0 \le x \le \ell$ represent the reference position along the beam of length ℓ. In the present, simplified model, we ignore stretching, and assume that the atoms in the beam can only move in the transverse direction. We let $y = u(x)$ represent the transverse displacement of the atom at position x.

The *strain* in a beam measures how much it is bent. Mathematically, bending is equal to the *curvature*[†] of the graph of the displacement function $u(x)$, and is computed by the usual calculus formula

$$\kappa = \frac{u''}{(1 + u'^2)^{3/2}} \,. \tag{11.106}$$

Thus, for beams, the strain is a *nonlinear* function of displacement. Since we are still only willing to deal with linear systems, we shall suppress the nonlinearity by assuming that the beam is not bent too far; more specifically, we assume that the derivative $u'(x) \ll 1$ is small and so the tangent line is nearly horizontal. Under this assumption, the curvature function (11.106) is replaced by its linear approximation

$$\kappa \;\approx\; v = L[u] = \frac{d^2u}{dx^2} \,.$$

From now on, we will identify $v = u''$ as the *strain* in a bending beam. The second derivative operator $L = D^2$ that maps displacement to strain thereby assumes the role of the incidence matrix for the (linearized) beam and describes its underlying geometry.

The next step is to formulate a constitutive relation that relates stress to strain. Physically, the stress $w(x)$ represents the bending moment of the beam, defined as the product of internal force and angular deflection. Our small bending assumption allows us to assume a linear Hooke's law, and so the beam stress function has the form

$$w(x) = c(x)\, v(x) = c(x)\, \frac{d^2u}{dx^2} \,, \tag{11.107}$$

where the proportionality factor $c(x) > 0$ measures the *stiffness* of the beam at the point x. A uniform beam has constant stiffness, $c(x) \equiv c$.

Finally, the differential equation governing the equilibrium configuration of the beam will follow from a balance of the internal and external forces. To compute the internal force, we appeal to our general equilibrium framework, which leads us to apply the adjoint of the incidence operator $L = D^2$ to the stress. Let us compute the adjoint. We use the ordinary L^2 inner product on the space of displacements $u(x)$, and adopt a weighted inner product, based on the stiffness function $c(x)$, between strain functions:

$$\langle\, u\,;\widetilde{u}\,\rangle = \int_a^b u(x)\,\widetilde{u}(x)\,dx, \qquad \langle\!\langle\, v\,;\widetilde{v}\,\rangle\!\rangle = \int_a^b v(x)\,\widetilde{v}(x)\,c(x)\,dx. \tag{11.108}$$

To compute the adjoint $L^* = (D^2)^*$, we need to compare

$$\langle\!\langle\, L[u]\,;v\,\rangle\!\rangle = \int_0^\ell L[u]\,v\,c\,dx \qquad \text{with} \qquad \langle\, u\,;L^*[v]\,\rangle = \int_0^\ell u\,L^*[v]\,dx.$$

[†] By definition, [**9**], the curvature of a curve at a point is equal to the reciprocal, $\kappa = 1/r$ of the radius of the osculating circle; see Exercise ■ for details.

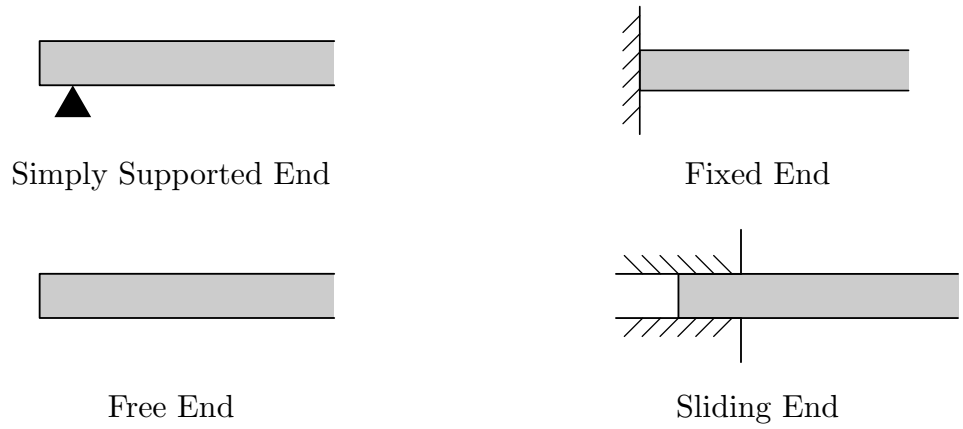


Figure 11.14. Boundary Conditions for a Beam.

As always, the adjoint computation relies on (in this case a double) integration by parts:

$$\begin{aligned} \langle\!\langle\, L[u]\,;v\,\rangle\!\rangle = \int_0^\ell \frac{d^2u}{dx^2}\; c\,v\,dx &= \left[\, \frac{du}{dx}\; c\,v \,\right] \Big|_{x=0}^{\ell} - \int_0^\ell \frac{du}{dx}\; \frac{d(c\,v)}{dx}\,dx \\ &= \left[\, \frac{du}{dx}\; c\,v - u\; \frac{d(c\,v)}{dx} \,\right] \Big|_{x=0}^{\ell} + \int_0^\ell u\; \frac{d^2(c\,v)}{dx^2}\,dx \end{aligned}.$$

Therefore, $L^*[v] = D^2(c\,v)$ provided the boundary terms vanish:

$$\begin{aligned} \left[\, \frac{du}{dx}\; c\,v - u\; \frac{d(c\,v)}{dx} \,\right] \Big|_{x=0}^{\ell} &= \left[\, \frac{du}{dx}\; w - u\; \frac{dw}{dx} \,\right] \Big|_{x=0}^{\ell} \\ &= \big[\, u'(\ell)\, w(\ell) - u(\ell)\, w'(\ell) \,\big] - \big[\, u'(0)\, w(0) - u(0)\, w'(0) \,\big] = 0. \end{aligned} \tag{11.109}$$

Thus, the appropriate force balance equations are

$$L^*[v] = f, \qquad \text{or, explicitly,} \qquad \frac{d^2w}{dx^2} = \frac{d^2(c\,v)}{dx^2} = f(x). \tag{11.110}$$

A justification of (11.110) based on physical principles can be found in [**143**]. Combining (11.107), (11.110), we conclude that the equilibrium configurations of the beam are solutions to the differential equation

$$\frac{d^2}{dx^2}\, \left(c(x)\, \frac{d^2u}{dx^2} \right) = f(x). \tag{11.111}$$

Since we are dealing with a fourth order ordinary differential equation (11.111), we need to impose a total of four boundary conditions — two at each end — so as to make the boundary terms in our integration by parts computation vanish, (11.109). There are a variety of ways in which this can be arranged, and the most important possibilities are illustrated in Figure 11.14:

a) Simply supported end: $u(0) = w(0) = 0.$
b) Fixed (clamped) end: $u(0) = u'(0) = 0.$
c) Free end: $w(0) = w'(0) = 0.$
d) Sliding end: $u'(0) = w'(0) = 0.$

Here $w(x) = c(x)\, v(x) = c(x)\, u''(x)$ is the stress resulting from the displacement $u(x)$.

A second pair of boundary conditions must be imposed at the other end $x = \ell$. One can mix or match these conditions in any combination — for example, a pair of simply supported ends, or one free end and one fixed end, and so on. Inhomogeneous boundary conditions are also allowed, and used to model applied displacements or forces at each end of the beam. Yet another option is to consider a periodic beam, modeling a bendable circular ring, in which one imposes periodic boundary conditions

$$u(0) = u(\ell), \qquad u'(0) = u'(\ell), \qquad w(0) = w(\ell), \qquad w'(0) = w'(\ell).$$

Let us concentrate our efforts on the case of a uniform beam, with $c(x) \equiv 1$, of unit length $\ell = 1$. In the absence of external forcing, the differential equation (11.111) reduces to the homogeneous fourth order ordinary differential equation

$$\frac{d^4 u}{dx^4} = 0. \tag{11.112}$$

The general solution

$$u = a\, x^3 + b\, x^2 + c\, x + d \tag{11.113}$$

is a linear combination of the four basis solutions $1, x, x^2, x^3$, and is easily found by direct integration. Let us use this to completely solve a couple of representative boundary value problems.

First, suppose we fix both ends of the beam, imposing the boundary conditions

$$u(0) = 0, \qquad u'(0) = \beta, \qquad u(1) = 0, \qquad u'(1) = 0, \tag{11.114}$$

so that the left hand end is tilted by a (small) angle $\tan^{-1} \beta$. We substitute the solution formula (11.113) into the boundary conditions (11.114) and solve for

$$a = \beta, \qquad b = -2\,\beta, \qquad c = \beta, \qquad d = 0.$$

The resulting cubic polynomial solution

$$u(x) = \beta\,(x^3 - 2\,x^2 + x) = \beta\, x(1 - x)^2 \tag{11.115}$$

is known as a *Hermite cubic spline*[†], and graphed in Figure 11.15.

[†] We first met Charles Hermite in Section 3.6, and the term "spline" will be explained shortly.

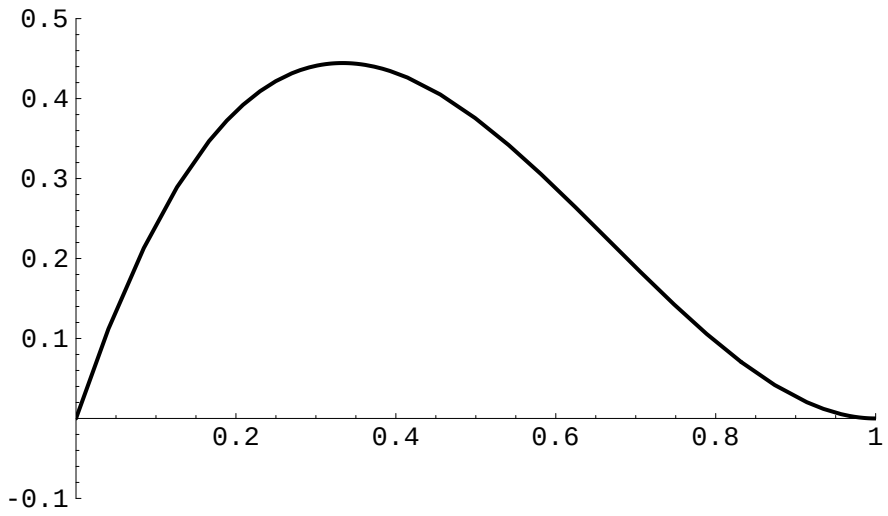


Figure 11.15. Hermite Cubic Spline.

As a second example, suppose that we raise the left hand end of the beam without tilting, which corresponds to the boundary conditions

$$u(0) = \alpha, \qquad u'(0) = 0, \qquad u(1) = 0, \qquad u'(1) = 0. \tag{11.116}$$

Substituting the general formula (11.113) and solving for the coefficients a, b, c, d, we find that the solution is

$$u(x) = \alpha\,(1 - x)^2\,(2\,x + 1). \tag{11.117}$$

If we simultaneously raise and tilt the left hand end, with $u(0) = \alpha$, $u'(0) = \beta$, then we can use superposition to write the solution as the sum of (11.115) and (11.117).

To analyze a forced beam, we can adapt the Green's function method. Let us treat the case when the beam has tow fixed ends, and so subject to the homogeneous boundary conditions

$$u(0) = 0, \qquad u'(0) = 0, \qquad u(1) = 0, \qquad u'(1) = 0. \tag{11.118}$$

To construct the Green's function, we must solve the forced boundary value problem

$$\frac{d^4u}{dx^4} = \delta(x - y) \tag{11.119}$$

corresponding to a concentrated unit impulse applied at position y along the beam. Integrating (11.119) four times, using (11.51) with $n = 4$, we produce the general solution

$$u(x) = a\,x^3 + b\,x^2 + c\,x + d + \begin{cases} \frac{1}{6}\,(x - y)^3, & x > y, \\ 0, & x < y, \end{cases}$$

to the differential equation (11.119). The boundary conditions require

$$\begin{aligned} u(0) &= d = 0, & u(1) &= a + b + \tfrac{1}{6}\,(1 - y)^3 = 0, \\ u'(0) &= c = 0, & u'(1) &= 3\,a + 2\,b + \tfrac{1}{2}\,(1 - y)^2 = 0, \end{aligned}$$

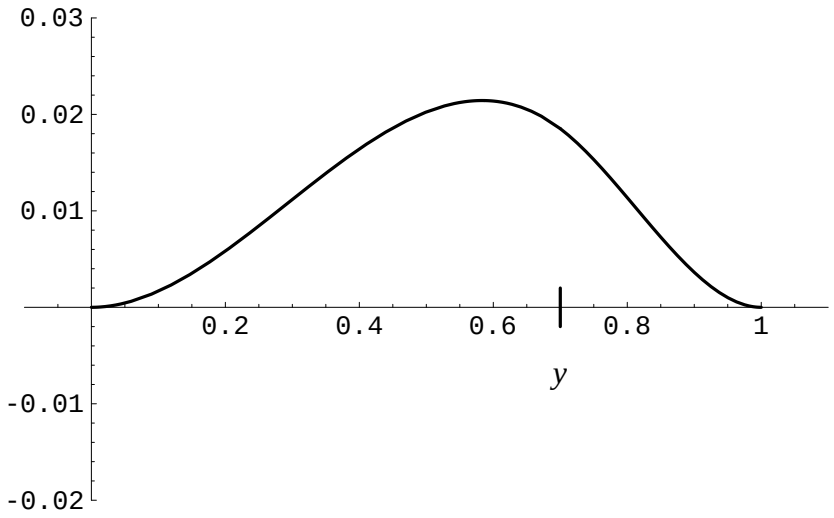


Figure 11.16. Green's Function for a Beam with Two Fixed Ends.

and hence

$$a = \tfrac{1}{3}\,(1-y)^3 - \tfrac{1}{2}\,(1-y)^2, \qquad\qquad b = -\tfrac{1}{2}\,(1-y)^3 + \tfrac{1}{2}\,(1-y)^2.$$

Therefore, the Green's function is

$$G(x,y) = \begin{cases} \tfrac{1}{6}\,x^2\,(1-y)^2\,(3y - x - 2xy), & x < y, \\ \tfrac{1}{6}\,y^2\,(1-x)^2\,(3x - y - 2xy), & x > y. \end{cases} \tag{11.120}$$

As in the second order case of bars, the Green's function is symmetric in x and y, so $G(x,y) = G(y,x)$, which stems from the fact that we are dealing with a self-adjoint system. Physically, symmetry implies that the deflection of the beam at position x due to a concentrated impulse force applied at position y is the same as the deflection at y due to an impulse force of the same magnitude applied at x. Moreover, as a function of x, the Green's function $G(x,y)$ satisfies the homogeneous differential equation for all $x \neq y$. Its first and second derivatives $\partial G/\partial x$, $\partial^2 G/\partial x^2$ are continuous; the third derivative $\partial^3 G/\partial x^3$ has a unit jump discontinuity at $x = y$, which then produces the delta function impulse in its fourth derivative. The Green's function (11.120) is graphed in Figure 11.16, and appears to be quite smooth. Evidently, the human eye cannot easily discern discontinuities in third order derivatives!

The solution to the general forced boundary value problem

$$\frac{d^4u}{dx^4} = f(x), \qquad\qquad u(0) = u'(0) = u(1) = u'(1) = 0, \tag{11.121}$$

is then obtained by invoking the usual superposition principle. We view the forcing function as a linear superposition

$$f(x) = \int_0^\ell f(y)\,\delta(x-y)\,dx$$

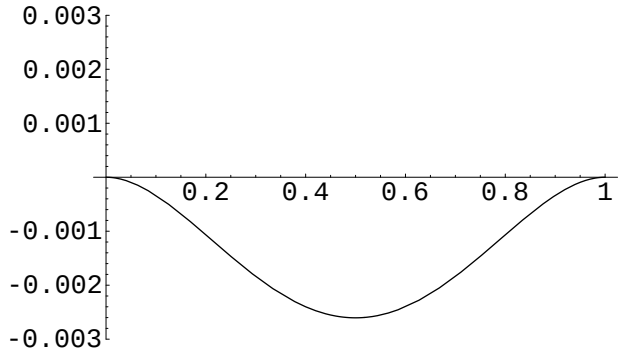


Figure 11.17. Deflection of a Uniform Beam under Gravity.

of impulse delta forces. The response is the self-same linear superposition of Green's function responses:

$$\begin{aligned} u(x) &= \int_0^1 G(x,y) f(y)\, dy \qquad (11.122)\\ &= \frac{1}{6}\int_0^x y^2\,(1-x)^2\,(3x-y-2xy)\, dy + \frac{1}{6}\int_x^1 x^2\,(1-y)^2\,(3y-x-2xy) f(y)\, dy. \end{aligned}$$

For example, under a constant unit downwards force $f(x) \equiv 1$, e.g., gravity, the deflection of the beam is given by

$$u(x) = \tfrac{1}{24}\,x^4 - \tfrac{1}{12}\,x^3 + \tfrac{1}{24}\,x^2 = \tfrac{1}{24}\,x^2\,(1-x)^2,$$

and graphed† in Figure 11.17. Although we could, of course, obtain u by integrating the original differential equation (11.121) directly, writing the solution formula as a single integral is more useful, particularly for numerical computations.

Since the beam operator has the standard self-adjoint form $K = L^* \circ L$, it will be positive definite when subject to the appropriate boundary conditions. As before, the key condition is that $\ker L = \ker D^2 = \{0\}$ on the space of functions satisfying the boundary conditions. Since the second derivative D^2 annihilates all affine functions $u = \alpha + \beta\, x$, the boundary value problem will be positive definite if and only if no non-zero affine function satisfies all four homogeneous boundary conditions. For example, having one fixed end will suffice, while two free ends, or a simply supported plus a free end will not. In the former case, every affine function satisfies the boundary conditions, while in the latter $u(x) = \beta\, x$ satisfies the four boundary conditions $u(0) = u'(0) = 0$, $w(\ell) = w'(\ell) = 0$.

In the positive definite cases, the solution to the beam boundary value problem can

† We have reversed the vertical axis in keeping with our convention that positive deflections go down.

be characterized as the unique minimizer of the quadratic energy functional[‡]

$$\mathcal{P}[u] = \tfrac{1}{2} \| L[u] \|^2 - \langle u ; f \rangle = \int_a^b \left[\tfrac{1}{2} c(x) u''(x)^2 - f(x) u(x) \right] dx. \tag{11.123}$$

Minimizing $\mathcal{P}$ among all functions with homogeneous boundary conditions will lead to the solution to the beam equation (11.111). Inhomogeneous boundary conditions require a little extra work, keeping careful track of the integration by parts required.

Splines

In pre–CAD draftsmanship, a *spline* was a long, thin, flexible strip of wood that was used to draw a smooth curve connecting prescribed points. The points were marked by small pegs, and the spline rested on the pegs. The mathematical theory of splines was first developed in the 1940's by I.J. Schoenberg as an attractive alternative to polynomial interpolation and approximation. It has since become standard in numerical analysis, computer graphics and design, and a broad range of other key applications.

We suppose that the spline coincides with the graph of a function $y = u(x)$. The pegs are fixed at the prescribed data points (x_j, y_j) for $j = 0, \dots, n$, and this requires $u(x)$ to satisfy the interpolation conditions

$$u(x_j) = y_j, \qquad j = 0, \dots, n. \tag{11.124}$$

The *mesh points* $x_0 < x_1 < x_2 < \cdots < x_n$ are distinct, and labeled in increasing order. On the intervals between each successive pair of mesh points, the spline is modeled as an elastic beam, and so satisfies the homogeneous beam equation (11.112). Therefore,

$$u(x) = a_j + b_j (x - x_j) + c_j (x - x_j)^2 + d_j (x - x_j)^3, \qquad \begin{array}{c} x_j \le x \le x_{j+1}, \\ j = 0, \dots, n-1, \end{array} \tag{11.125}$$

is a piecewise cubic function — meaning that between successive mesh points it is a cubic polynomial, but not necessarily the same cubic on each subinterval. The fact that we write the formula (11.125) in terms of $x - x_j$ is merely for computational convenience.

Our problem is to determine the coefficients

$$a_j, \qquad b_j, \qquad c_j, \qquad d_j, \qquad\qquad j = 0, \dots, n-1.$$

Since there are n subintervals between mesh points, there are a total of $4n$ coefficients, and so we require $4n$ equations to prescribe them uniquely. First, we need the spline to satisfy the interpolation conditions (11.124). Since the spline is given by a different formula on each side of the mesh point, this results in a total of $2n$ conditions:

$$\begin{aligned} u(x_j^+) &= a_j = y_j, \\ u(x_{j+1}^-) &= a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 = y_{j+1}, \end{aligned} \qquad j = 0, \dots, n-1, \tag{11.126}$$

[‡] Keep in mind that the norm on the strain functions $v = L[u] = u''$ is based on the weighted inner product $\langle\langle v ; \tilde{v} \rangle\rangle$ in (11.108).

where we abbreviate the length of the j^{th} subinterval by

$$h_j = x_{j+1} - x_j.$$

The next step is to require that the spline be as smooth as possible. The interpolation conditions (11.126) guarantee that $u(x)$ is continuous. The condition $u(x) \in \mathrm{C}^1$ is continuously differentiable requires that $u'(x)$ be continuous at the interior mesh points $x_1, \dots, x_{n-1}$, which imposes the $n-1$ additional conditions

$$b_j + 2\,c_j\,h_j + 3\,d_j\,h_j^2 = u'(x_{j+1}^-) = u'(x_{j+1}^+) = b_{j+1}, \qquad j = 0, \dots, n-2. \tag{11.127}$$

To make $u \in \mathrm{C}^2$, we impose $n-1$ further conditions

$$2\,c_j + 6\,d_j\,h_j = u''(x_{j+1}^-) = u''(x_{j+1}^+) = 2\,c_{j+1}, \qquad j = 0, \dots, n-2, \tag{11.128}$$

to ensure that u'' is continuous at the mesh points. We have now imposed a total of $4n-2$ conditions, namely (11.126), (11.127), and (11.128), on the $4n$ coefficients. The two missing constraints will be imposed as boundary conditions at the two endpoints of the interval, namely x_0 and x_n. There are three common types:

(*i*) *Natural boundary conditions*: $u''(x_0) = u''(x_n) = 0$ so that

$$c_0 = 0, \qquad c_{n-1} + 3\,d_{n-1}\,h_{n-1} = 0. \tag{11.129}$$

Physically, this corresponds to a simply supported spline that rests freely on the first and last pegs.

(*ii*) *Clamped boundary conditions*: $u'(x_0) = \alpha$, $u'(x_n) = \beta$, where α, β are fixed by the user. This requires

$$b_0 = \alpha, \qquad b_{n-1} + 2\,c_{n-1}\,h_{n-1} + 3\,d_{n-1}\,h_{n-1}^2 = \beta. \tag{11.130}$$

Physically, this corresponds to clamping the spline so as to have prescribed slopes at the ends.

(*iii*) *Periodic boundary conditions*: $u'(x_0) = u'(x_n)$, $u''(x_0) = u''(x_n)$, so that

$$b_0 = b_{n-1} + 2\,c_{n-1}\,h_{n-1} + 3\,d_{n-1}\,h_{n-1}^2, \qquad c_0 = c_{n-1} + 3\,d_{n-1}\,h_{n-1}. \tag{11.131}$$

If we also require that the end interpolation values agree,

$$u(x_0) = y_0 = y_n = u(x_n), \tag{11.132}$$

then the resulting spline will be a periodic C^2 function with period $p = x_n - x_0$, so $u(x+p) = u(x)$ for all x. A particularly important application of this case is to computer aided sketching of smooth closed curves.

Theorem 11.11. *Given data points (x_j, y_j) with $a = x_0 < x_1 < \cdots < x_n = b$, there exists a unique piecewise cubic spline function $u(x) \in \mathrm{C}^2[a,b]$ such that $u(x_j) = y_j$ and u satisfies one of the three possible pairs of boundary conditions* (11.129), (11.130), *or* (11.131).

Proof: We first discuss the natural case. The clamped case is left as an exercise for the reader, while the slightly harder periodic case will be done at the end of the section. First, (11.126) says that

$$a_j = y_j, \qquad j = 0, \dots, n-1. \tag{11.133}$$

Second, (11.128), (11.129) imply that

$$d_j = \frac{c_{j+1} - c_j}{3\,h_j}\,. \tag{11.134}$$

This equation holds for $j = n-1$ provided we make the convention that

$$c_n = 0.$$

Substituting (11.134) into (11.127),

$$b_{j+1} = b_j + (c_j + c_{j+1})\,h_j. \tag{11.135}$$

We now substitute (11.133), (11.134) into (11.126), and then solve the resulting equation for

$$b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{(2\,c_j + c_{j+1})\,h_j}{3}\,. \tag{11.136}$$

Substituting this result back into (11.135), and simplifying, we find

$$h_j\,c_j + 2\,(h_j + h_{j+1})\,c_{j+1} + h_{j+1}\,c_{j+2} = 3\left[\frac{y_{j+2} - y_{j+1}}{h_{j+1}} - \frac{y_{j+1} - y_j}{h_j}\right] \equiv z_{j+1}, \tag{11.137}$$

where we use z_{j+1} to denote the right hand side of these equations.

For the natural boundary conditions, we have

$$c_0 = 0, \qquad c_n = 0,$$

and so, setting $\mathbf{c} = \big(c_1, c_2, \dots, c_{n-1}\big)^T$, $\mathbf{z} = \big(z_1, z_2, \dots, z_{n-1}\big)^T$, (11.137) constitutes a tridiagonal linear system

$$A\,\mathbf{c} = \mathbf{z}, \tag{11.138}$$

for the unknown coefficients $c_1, \dots, c_{n-1}$, with coefficient matrix

$$A = \begin{pmatrix} 2\,(h_0 + h_1) & h_1 & & & & \\ h_1 & 2\,(h_1 + h_2) & h_2 & & & \\ & h_2 & 2\,(h_2 + h_3) & h_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & h_{n-3} & 2\,(h_{n-3} + h_{n-2}) & h_{n-2} & \\ & & & h_{n-2} & 2\,(h_{n-2} + h_{n-1}) \end{pmatrix}. \tag{11.139}$$

Once we solve (11.139), we then use (11.133), (11.136), (11.134) to reconstruct the other coefficients a_j, b_j, d_j.

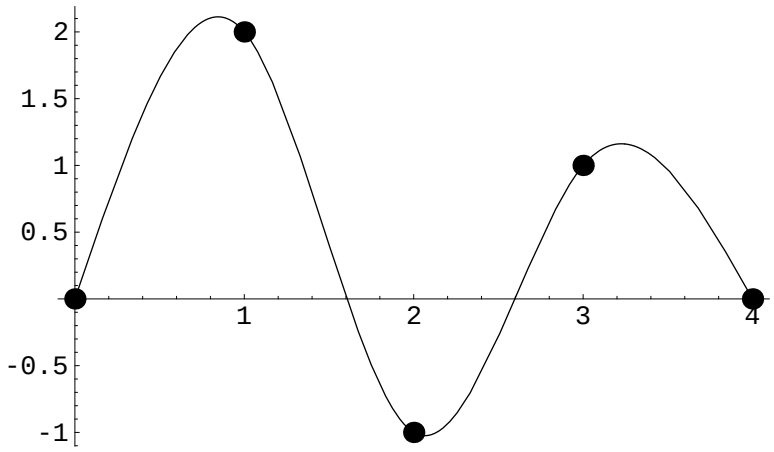


Figure 11.18. A Cubic Spline.

The key observation is that the coefficient matrix A is *strictly diagonally dominant*, as in Definition 10.33, because all the $h_j > 0$, and so

$$2\,(h_{j-1} + h_j) > h_{j-1} + h_j.$$

Proposition 10.34 implies that A is nonsingular, and hence the tridiagonal linear system has a unique solution $\mathbf{c}$. This suffices to prove the theorem in the case of natural boundary conditions. *Q.E.D.*

To actually solve the system and compute the resulting spline function, we can apply our tridiagonal solution algorithm (1.63). Let us consider the most important case, when the mesh points are equally spaced in the interval $[a, b]$, so that

$$x_j = a + j\,h, \qquad \text{where} \qquad h = h_j = \frac{b-a}{n}\,, \qquad j = 0, \dots, n-1.$$

In this case, the coefficient matrix $A = h\,B$ is equal to h times the tridiagonal matrix

$$B = \begin{pmatrix} 4 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & 1 & 4 & 1 & & \\ & & & 1 & 4 & 1 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

that first appeared in Example 1.35. Its LU factorization takes on a particularly simple form, which makes the implementation of the forward and back substitution algorithms particularly easy.

In summary, the construction of the natural spline proceeds as follows: First, determine the coefficients $c_0, \dots, c_n$ by solving the tridiagonal linear system (11.138) to construct $c_1, \dots, c_{n-1}$ and the boundary conditions to determine $c_0 = c_n = 0$. Then use equations (11.133), (11.136), (11.134) to find the other coefficients $a_0, \dots, a_{n-1}$, $b_0, \dots, b_{n-1}$,

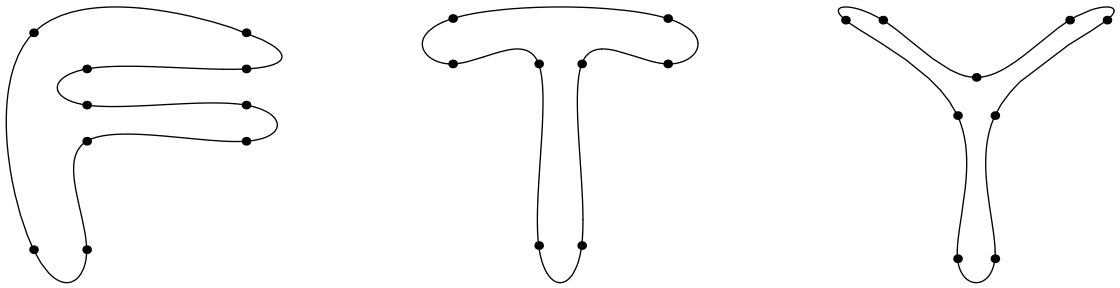


Figure 11.19. Three Sample Spline Letters.

$d_0, \dots, d_{n-1}$. The resulting piecewise cubic spline (11.125) will be the unique natural spline interpolant to the data $u(x_j) = y_j$ for $j = 0, \dots, n$.

Figure 11.18 shows a particular example — a natural spline passing through the data points $(0,0)$, $(1,2)$, $(2,-1)$, $(3,1)$, $(4,0)$. As with the Green's function for the beam, the human eye is unable to discern the discontinuities in its third derivatives, and so the graph appears completely smooth even though it is, in fact, only C^2.

In the periodic case, we set

$$a_n = a_0, \qquad a_{n+1} = a_1, \qquad \text{etc.}$$

and similarly for the other coefficients. The basic systems (11.133), (11.136), (11.134) and (11.137) are the same. Now the coefficient matrix for the linear system

$$A\,\mathbf{c} = \mathbf{z}, \qquad \text{with} \qquad \mathbf{c} = \big(\, c_0, c_1, \dots, c_{n-1} \,\big)^T, \qquad \mathbf{z} = \big(\, z_0, z_1, \dots, z_{n-1} \,\big)^T$$

is no longer tridiagonal, but of "circulant tridiagonal" type:

$$A = \begin{pmatrix} 2\,(h_{n-1}+h_0) & h_0 & & & & & h_{n-1} \\ h_0 & 2\,(h_0+h_1) & h_1 & & & & \\ & h_1 & 2\,(h_1+h_2) & h_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & h_{n-3} & 2\,(h_{n-3}+h_{n-2}) & h_{n-2} & \\ h_{n-1} & & & & h_{n-2} & 2\,(h_{n-2}+h_{n-1}) & \end{pmatrix}. \tag{11.140}$$

Again A is strictly diagonally dominant, and so there is a unique solution, proving Theorem 11.11 in the periodic case. The LU factorization of such "tridiagonal circulant" matrices was discussed in Exercise ■.

One immediate application of splines is curve fitting in computer aided design (CAD) and computer graphics. The basic problem is to draw a smooth curve $\mathbf{x} = \mathbf{u}(t) = \big(\, u(t), v(t) \,\big)^T$ that passes through a set of prescribed data points $\mathbf{x}_k = \big(\, x_k, y_k \,\big)^T$ in the plane. We have the freedom to choose the parameter values $t = t_k$ when the curve passes through the k^{th} point; the simplest and most common choice is to set $t_k = k$. We then construct the functions $x = u(t)$ and $y = v(t)$ as cubic splines interpolating the x and y coordinates of the data points, so $u(k) = x_k$, $v(k) = y_k$. The result is a parametrized curve that interpolates the data points. If the curve is closed, then we require that both

splines be periodic; for curves with ends, either natural or clamped boundary conditions are used. In addition to implementations in most computer graphics packages, this idea also underlies modern font design for laser printing and typography (including the fonts used in this book). The great advantage of spline fonts over their bitmapped counterparts is that they can be readily scaled to arbitrary sizes and arbitrary resolutions. Some sample letter shapes parametrized by periodic splines passing through the indicated data points are plotted in Figure 11.19. Better results can be easily obtained by increasing the number of prescribed data points used to fix the interpolating splines. Various extensions of the basic method to curves, and also to surfaces in three-dimensional computer graphics, design and animation, can be found in [**52**, **129**].

11.5. Sturm–Liouville Boundary Value Problems.

The boundary value problems that govern the equilibrium configurations of bars are particular cases of a very general class of second order boundary value problems that was first systematically investigated by the nineteenth century French mathematicians Jacques Sturm and Joseph Liouville. Sturm–Liouville boundary value problems appear in a very wide range of applications, particularly in the analysis of partial differential equations by the method of separation of variables. A partial list of applications includes

(a) Heat conduction in non-uniform bars,
(b) Vibrations of non-uniform bars and strings,
(c) Quantum mechanics — the one-dimensional Schrödinger equation.
(d) Scattering theory — Hill's equation.
(e) Oscillations of circular membranes (vibrations of drums) — Bessel's equation.
(f) Oscillations of a sphere — Legendre's equation,
(g) Heat flow in cylindrical and spherical bodies.

Details can be found in [**24**, **146**]. In this section, we will show how the class of Sturm–Liouville problems fits into our general equilibrium framework and analyze some elementary examples. However, the most interesting cases will be deferred until needed in our analysis of partial differential equations in Chapters 17 and 18.

The general *Sturm–Liouville boundary value problem* is based on a second order ordinary differential equation of the form

$$-\frac{d}{dx}\left(p(x)\,\frac{du}{dx}\right) + q(x)\,u = f(x), \tag{11.141}$$

coupled with Dirichlet, Neumann, mixed or periodic boundary conditions. To be specific, let us concentrate on the case of homogeneous Dirichlet boundary conditions

$$u(a) = 0, \qquad u(b) = 0. \tag{11.142}$$

To avoid singular points of the differential equation (although we will later discover that most cases of interest in physics have one or more singular points) we assume that $p(x) > 0$ for all $a \le x \le b$. To ensure positive definiteness of the Sturm–Liouville differential operator, we also assume $q(x) \ge 0$. These assumptions suffice to guarantee existence and uniqueness of the solution to the Sturm–Liouville problem. A proof of the following theorem can be found in [**24**].

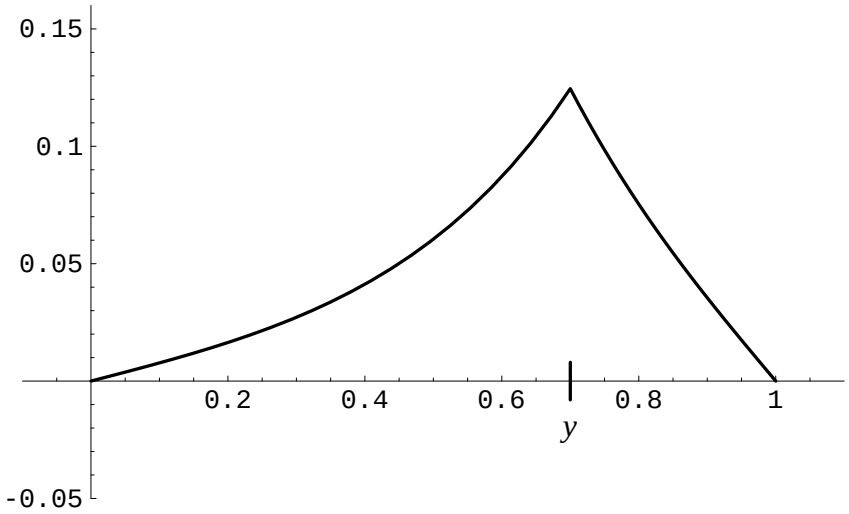


Figure 11.20. Green's Function for the Constant Coefficient Sturm–Liouville Problem.

Theorem 11.12. *Let $p(x) > 0$ and $q(x) \geq 0$ for $a \leq x \leq b$. Then the Sturm–Liouville boundary value problem* (11.141), (11.142) *admits a unique solution.*

Most Sturm–Liouville problems cannot be solved in terms of elementary functions. Indeed, most of the important special functions appearing in mathematical physics, including Bessel functions, Legendre functions, hypergeometric functions, and so on, first arise as solutions to particular Sturm–Liouville equations. But the simplest, constant coefficient case can be solved by standard methods.

Example 11.13. Consider the constant coefficient Sturm–Liouville boundary value problem

$$-u'' + \omega^2 u = f(x), \qquad u(0) = u(1) = 0. \tag{11.143}$$

The functions $p(x) \equiv 1$ and $q(x) \equiv \omega^2 > 0$ are both constant. We will solve this problem by constructing the Green's function. Thus, we first consider the effect of a delta function inhomogeneity

$$-u'' + \omega^2 u = \delta(x - y), \qquad u(0) = u(1) = 0. \tag{11.144}$$

Rather than try to integrate this differential equation directly, let us appeal to the defining properties of the Green's function. The general solution to the homogeneous equation is a linear combination of the two basic exponentials $e^{\omega x}$ and $e^{-\omega x}$, or better, the hyperbolic functions

$$\cosh \omega x = \frac{e^{\omega x} + e^{-\omega x}}{2}, \qquad \sinh \omega x = \frac{e^{\omega x} - e^{-\omega x}}{2}. \tag{11.145}$$

The solutions satisfying the first boundary condition are multiples of $\sinh \omega x$, while the solutions satisfying the second boundary condition are multiples of $\sinh \omega (1-x)$. Therefore, the solution to (11.144) has the form

$$G(x,y) = \begin{cases} a \sinh \omega x, & x < y, \\ b \sinh \omega (1-x), & x > y. \end{cases} \tag{11.146}$$

Continuity of $G(x,y)$ at $x = y$ requires

$$a \sinh \omega y = b \sinh \omega (1-y). \tag{11.147}$$

At $x = y$, the derivative $\partial G/\partial x$ must have a jump discontinuity of magnitude -1 in order that the second derivative term in (11.144) match the delta function. Since

$$\frac{\partial G}{\partial x}(x,y) = \begin{cases} \omega a \cosh \omega x, & x < y, \\ -\omega b \cosh \omega (1-x), & x > y, \end{cases}$$

the jump condition requires

$$\omega a \cosh \omega y - 1 = -\omega b \cosh \omega (1-y). \tag{11.148}$$

If we multiply (11.147) by $\omega \cosh \omega (1-y)$ and (11.148) by $\sinh \omega (1-y)$ and then add the results together, we find

$$\sinh \omega (1-y) = a\omega \big[\sinh \omega y \, \cosh \omega (1-y) + \cosh \omega y \, \sinh \omega (1-y) \big] = a\omega \sinh \omega,$$

where we used the addition formula for the hyperbolic sine:

$$\sinh(\alpha + \beta) = \sinh \alpha \, \cosh \beta + \cosh \alpha \, \sinh \beta, \tag{11.149}$$

cf. Exercise ∎. Therefore,

$$a = \frac{\sinh \omega (1-y)}{\omega \sinh \omega}, \qquad b = \frac{\sinh \omega y}{\omega \sinh \omega},$$

and the Green's function for our boundary value problem is

$$G(x,y) = \begin{cases} \dfrac{\sinh \omega (1-y) \, \sinh \omega x}{\omega \sinh \omega}, & x < y, \\ \dfrac{\sinh \omega (1-x) \sinh \omega y}{\omega \sinh \omega}, & x > y. \end{cases}$$

A graph appears in Figure 11.20; note that the corner, indicating a discontinuity in the first derivative, appears at the point $x = y$ where the impulse force is applied.

The general solution to the inhomogeneous boundary value problem (11.143) is given by the basic superposition formula (11.64), which becomes

$$\begin{aligned} u(x) &= \int_0^1 G(x,y) f(y) \, dx \\ &= \int_0^x \frac{\sinh \omega (1-y) \, \sinh \omega x}{\omega \sinh \omega} f(y) \, dy \; + \; \int_x^1 \frac{\sinh \omega (1-x) \sinh \omega y}{\omega \sinh \omega} f(y) \, dy. \end{aligned}$$

For example, under a constant unit force $f(x) \equiv 1$, the solution is

$$\begin{aligned} u(x) &= \int_0^x \frac{\omega \sinh \omega}{\sinh \omega (1-x) \sinh \omega y} \, dy \; + \; \int_x^1 \frac{\sinh \omega (1-y) \, \sinh \omega x}{\omega \sinh \omega} \, dy \\ &= \frac{\sinh \omega (1-x) \big(\cosh \omega x - 1 \big)}{\omega^2 \sinh \omega} \; + \; \frac{\sinh \omega x \big(\cosh \omega (1-x) - 1 \big)}{\omega^2 \sinh \omega} \\ &= \frac{1}{\omega^2} \; - \; \frac{\sinh \omega x + \sinh \omega (1-x)}{\omega^2 \sinh \omega}. \end{aligned} \tag{11.150}$$

For comparative purposes, the reader may wish to rederive this particular solution by a direct calculation, without appealing to the Green's function.

Finally, to place a Sturm–Liouville boundary value problem in our self-adjoint framework, we proceed as follows. (See Exercise ■ for motivation.) Consider the linear operator

$$L[u] = \begin{pmatrix} u' \\ u \end{pmatrix}$$

that maps $u(x)$ to the vector-valued function whose components are the function and its first derivative. In view of the boundary conditions (11.142), the domain of L will be the vector space

$$U = \{\, u(x) \mid \; u(a) = u(b) = 0 \,\} \subset \mathrm{C}^2[a,b]$$

consisting of all twice continuously differentiable functions that vanish at the endpoints. The target space of L consists of continuously differentiable vector-valued functions $\mathbf{v}(x) = (\, v_1(x), v_2(x) \,)^T$; we denote this vector space as $V = \mathrm{C}^1([a,b], \mathbb{R}^2)$.

We need to compute the adjoint of $L\colon U \to V$. To recover the Sturm–Liouville problem, we use the standard L^2 inner product (11.85) on U, but adopt a weighted inner product

$$\langle\!\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle\!\rangle = \int_a^b \big[\, p(x)\, v_1(x)\, w_1(x) + q(x)\, v_2(x)\, w_2(x)\,\big]\, dx, \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \tag{11.151}$$

on V. The positivity assumptions on the weight functions p, q ensure that this is a *bona fide* inner product. According to the defining equation (7.64), the adjoint $L^*\colon V \to U$ is required to satisfy

$$\langle\!\langle\, L[u]\,;\mathbf{v}\,\rangle\!\rangle = \langle\, u\,; L^*[\mathbf{v}]\,\rangle.$$

As usual, the adjoint computation relies an integration by parts. In this case, we only need to manipulate the first summand:

$$\begin{aligned} \langle\!\langle\, L[u]\,;\mathbf{v}\,\rangle\!\rangle &= \int_a^b \big[\, p\, u'\, v_1 + q\, u\, v_2\,\big]\, dx \\ &= \big[\, p(b)\, u(b)\, v_1(b) - p(a)\, u(a)\, v_1(a)\,\big] + \int_a^b u\, \big[\, -(p\, v_1)' + q\, v_2\,\big]\, dx. \end{aligned}$$

The Dirichlet conditions (11.142) ensure that the boundary terms vanish, and therefore,

$$\langle\!\langle\, L[u]\,;\mathbf{v}\,\rangle\!\rangle = \int_a^b u\, \big[\, -(p\, v_1)' + q\, v_2\,\big]\, dx = \langle\, u\,; L^*[\mathbf{v}]\,\rangle.$$

We conclude that the adjoint operator is given by

$$L^*[\mathbf{v}] = -\,\frac{d(p\, v_1)}{dx} + q\, v_2.$$

The canonical self-adjoint combination

$$K[u] = L^* \circ L[u] = L^* \begin{pmatrix} u' \\ u \end{pmatrix} = -\,\frac{d}{dx}\left(p\, \frac{du}{dx} \right) + q\, u. \tag{11.152}$$

reproduces the Sturm–Liouville differential operator. Moreover, since $\ker L = \{0\}$ is trivial, the boundary value problem is positive definite. As a direct consequence of the general abstract minimization principle established in Theorem 7.61, we deduce that the solution to the Sturm–Liouville boundary value problem (11.141) can be characterized as the minimizer of the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2}\| L[u] \|^2 - \langle u\,;f\rangle = \int_a^b \big[\, \tfrac{1}{2}\,p(x)\,u'(x)^2 + \tfrac{1}{2}\,q(x)\,u(x)^2 - f(x)\,u(x)\,\big]\,dx \quad (11.153)$$

among all C^2 functions satisfying the prescribed boundary conditions. For example, the solution to the constant coefficient Sturm–Liouville problem (11.143) can be characterized as minimizing the quadratic functional

$$\mathcal{P}[u] = \int_0^1 \big[\, \tfrac{1}{2}\,u'(x)^2 + \tfrac{1}{2}\,\omega^2\,u(x)^2 - f(x)\,u(x)\,\big]\,dx$$

among all C^2 functions satisfying $u(0) = u(1) = 0$.

11.6. Finite Elements.

The characterization of the solution to a positive definite boundary value problem via a minimization principle inspires a very powerful and widely applicable numerical solution algorithm, known as the *finite element method*. In this section, we give a brief introduction to the finite element method in the context of one-dimensional boundary value problems involving ordinary differential equations. Extensions to boundary value problems in higher dimensions governed by partial differential equations will appear in Section 15.5.

The underlying idea is strikingly simple. We are trying to find the solution to a boundary value problem by minimizing a quadratic functional $\mathcal{P}[u]$ on an infinite-dimensional vector space U. The solution $u_\star(x) \in U$ to this minimization problem requires the solution to a differential equation with boundary conditions. However, as we learned in Chapter 4, if we were to minimize the functional on a *finite-dimensional* subspace $W \subset U$, then this becomes a problem in linear algebra, and, moreover, one that we have already solved! Of course, restricting the functional $\mathcal{P}[u]$ to the subspace W will not, barring luck, lead to the exact minimizer. Nevertheless, if we choose W to be sufficiently "large" subspace, the resulting minimizer $w_\star \in W$ may very well provide a reasonable approximation to the true minimizer $u_\star \in U$. The analysis of the finite element method, cf. [**139**, **157**], provides a rigorous justification of this process, under appropriate hypotheses. Here, we shall concentrate on trying to understand how to apply the method in practice.

To be a bit more explicit, we consider the basic abstract minimization principle

$$\mathcal{P}[u] = \tfrac{1}{2}\,\| L[u] \|^2 - \langle f\,;u\rangle, \quad (11.154)$$

for the boundary value problem

$$K[u] = f, \qquad \text{where} \qquad K = L^* \circ L.$$

As we have learned, the norm in (11.154) is typically based on some form of weighted inner product $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$ on the space of strains $v = L[u] \in V$, while the inner product term $\langle f\,;u\rangle$

is typically (although not necessarily) unweighted on the space of displacements $u \in U$. The linear operator takes the self-adjoint form $K = L^* \circ L$, and must be positive definite — which requires $\ker L = \{0\}$. Without the positivity assumption, the boundary value problem has either no solutions, or infinitely many; in either event, the basic finite element method will not apply.

Rather than try to minimize $\mathcal{P}[u]$ on the entire function space U, we now seek to minimize it on a suitably chosen finite-dimensional subspace $W \subset U$, the elements of which are required to satisfy the boundary conditions. We begin by choosing a basis[†] $\varphi_1, \dots, \varphi_n$ of our finite-dimensional subspace W. The general element of W is a linear combination

$$w(x) = c_1 \varphi_1(x) + \cdots + c_n \varphi_n(x) \tag{11.155}$$

of the basis functions. Our goal, then, is to determine the coefficients $c_1, \dots, c_n$ such that $w(x)$ minimizes $\mathcal{P}[w]$ among all such functions. Substituting (11.155) into (11.154) and expanding we find

$$\mathcal{P}[w] = \frac{1}{2} \sum_{i,j=1}^{n} m_{ij}\, c_i\, c_j - \sum_{i=1}^{n} b_i\, c_i = \tfrac{1}{2}\, \mathbf{c}^T M \mathbf{c} - \mathbf{c}^T \mathbf{b}, \tag{11.156}$$

where

(a) $\mathbf{c} = (\, c_1, c_2, \dots, c_n \,)^T$ is the vector of unknown coefficients in (11.155),

(b) $M = (m_{ij})$ is the symmetric $n \times n$ matrix with entries

$$m_{ij} = \langle\!\langle\, L[\varphi_i]\,;\, L[\varphi_j]\,\rangle\!\rangle, \qquad i, j = 1, \dots, n, \tag{11.157}$$

(c) $\mathbf{b} = (\, b_1, b_2, \dots, b_n \,)^T$ is the vector with entries

$$b_i = \langle\, f\,;\, \varphi_i\,\rangle, \qquad i = 1, \dots, n. \tag{11.158}$$

Note that once we specify the basis functions φ_i, the coefficients m_{ij} and b_i are all known quantities. Therefore, we have reduced our original problem to a finite-dimensional problem of minimizing a quadratic function (11.156) over all possible vectors $\mathbf{c} \in \mathbb{R}^n$. The coefficient matrix M is, in fact, positive definite, since, by the preceding computation,

$$\mathbf{c}^T M \mathbf{c} = \sum_{i,j=1}^{n} m_{ij}\, c_i\, c_j = \| \, L[c_1 \varphi_1(x) + \cdots + c_n \varphi_n] \, \|^2 = \| \, L[w] \, \|^2 > 0 \tag{11.159}$$

as long as $L[w] \neq 0$. Moreover, our positivity assumption implies that $L[w] = 0$ if and only if $w \equiv 0$, and hence (11.159) is positive for all $\mathbf{c} \neq \mathbf{0}$. We can now invoke the original finite-dimensional minimization Theorem 4.1 to conclude that the unique minimizer to (11.156) is obtained by solving the associated linear system

$$M \mathbf{c} = \mathbf{b}, \tag{11.160}$$

[†] In this case, an orthonormal basis is not of any particular help.

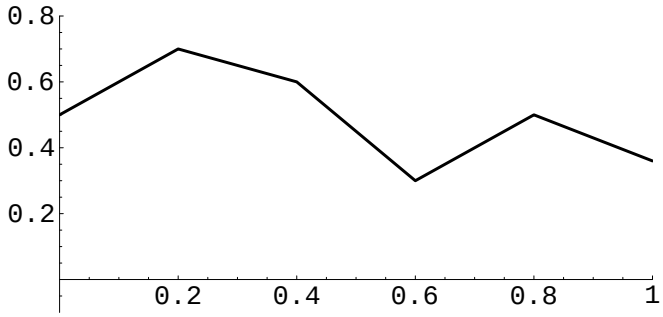


Figure 11.21. A Continuous Piecewise Affine Function.

which can be done by some form of Gaussian elimination, or, alternatively, by an iterative linear system solver, e.g., Gauss–Seidel or SOR.

This constitutes the basic framework for the finite element method. The main issue, then, is how to effectively choose the finite-dimensional subspace W. We already know a few potential candidates. One is the space $\mathcal{P}^{(n)}$ of polynomials of degree $\leq n$. Another is the space $\mathcal{T}^{(n)}$ of trigonometric polynomials of degree $\leq n$, to be the focus of Chapter 12. However, neither of these is particularly suitable in the present situation for a variety of reasons. One criterion is that the functions in W must satisfy the relevant boundary conditions. More importantly, in order to obtain sufficient accuracy, the resulting linear system (11.160) will typically be rather large, and so the coefficient matrix (11.157) should be as sparse as possible, i.e., have lots of zero entries. Otherwise, computing the solution will be too time consuming to be of much practical value. Such considerations will prove to be of critical importance when applying the method to solve boundary value problems for partial differential equations in several variables.

The really innovative contribution of the finite element method is to first (paradoxically) *enlarge* the space of allowable functions upon which to minimize $\mathcal{P}[u]$. The governing differential equation requires its solutions to have a certain degree of smoothness, whereas the associated minimization principles typically requires only half as many derivatives. Thus, for second order boundary value problems, including bars, (11.92), and general Sturm–Liouville problems, (11.153), the quadratic functional only involves first order derivatives. It can be rigorously shown that the functional has the *same* minimizing solution, even if one allows functions that do not have enough derivatives to satisfy the differential equation. Thus, one can try minimizing over subspaces containing fairly "rough" functions. Again, the justification of this method requires some deeper analysis, but this lies beyond the scope of this introductory treatment.

For second order boundary value problems, a popular and effective choice of the finite-dimensional subspace is to use continuous, piecewise affine functions. Recall that a function is affine, $f(x) = a\,x + b$, if and only if its graph is a straight line. The function is *piecewise affine* if its graph consists of a finite number of straight line segments; a typical example is plotted in Figure 11.21. Continuity requires that the individual line segments be connected together at their ends.

Given a boundary value problem on a bounded interval $[a, b]$, let us choose a set of

mesh points

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

The formulas simplify if one uses equally spaced mesh points, but this is not necessary for the method to apply. Let W denote the vector space consisting of all continuous, piecewise affine functions with corners at the nodes that satisfy the boundary conditions. To be specific, let us treat the case of homogeneous Dirichlet (fixed) boundary conditions

$$w(a) = w(b) = 0. \tag{11.161}$$

Thus, on each subinterval

$$w(x) = c_j + b_j(x - x_j), \qquad \text{for} \qquad x_j \le x \le x_{j+1}, \qquad j = 0, \dots, n-1.$$

Continuity of $w(x)$ requires

$$c_j = w(x_j^+) = w(x_j^-) = c_{j-1} + b_{j-1}\, h_{j-1}, \qquad j = 1, \dots, n-1, \tag{11.162}$$

where $h_{j-1} = x_j - x_{j-1}$ denotes the length of the j^{th} subinterval. The boundary conditions (11.161) require

$$w(a) = c_0 = 0, \qquad w(b) = c_{n-1} + h_{n-1}\, b_{n-1} = 0. \tag{11.163}$$

The function $w(x)$ involves a total of $2n$ different coefficients $c_0, \dots, c_{n-1}, b_0, \dots, b_{n-1}$. The continuity conditions (11.162) and the second boundary condition (11.163) uniquely determine the b_j. The first boundary condition specifies c_0, while the remaining $n-1$ coefficients $c_1 = w(x_1), \dots, c_{n-1} = w(x_{n-1})$ can be selected in an arbitrary manner. We conclude that the vector space W has dimension $n-1$, the number of interior mesh points.

Remark: Every function $w(x)$ in our subspace has piecewise constant first derivative $w'(x)$. However, the jump discontinuities in $w'(x)$ imply that the second derivative $w''(x)$ has a delta function impulse at each mesh point, and is therefore far from being a solution to the differential equation. Nevertheless, the finite element minimizer $w_\star(x)$ will, in practice, provide a reasonable approximation to the actual solution $u_\star(x)$.

The most convenient basis for the space W consists of the "hat functions" which are continuous, piecewise affine functions that interpolate the same basis data as the Lagrange polynomials (4.44) and the cardinal splines of Exercise ■, namely

$$\varphi_j(x_k) = \begin{cases} 1, & j = k, \\ 0, & j \ne k, \end{cases} \qquad \text{for} \qquad j = 1, \dots, n-1, \qquad k = 0, \dots, n.$$

The graph of a typical hat function appears in Figure 11.22. The explicit formula is easily established:

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \le x \le x_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \le x \le x_{j+1}, \\ 0, & x \le x_{j-1} \ \text{ or } \ x \ge x_{j+1}, \end{cases} \qquad j = 1, \dots, n-1. \tag{11.164}$$

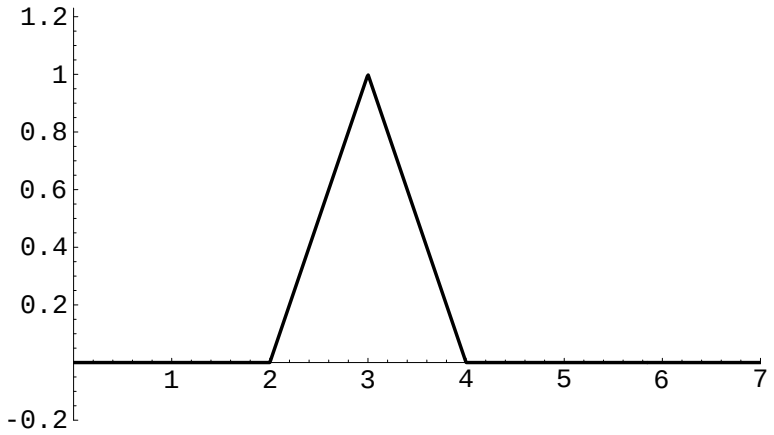


Figure 11.22. A Hat Function.

An advantage of using these basis elements is that the resulting coefficient matrix (11.157) turns out to be tridiagonal. Therefore, the tridiagonal Gaussian elimination algorithm in (1.63), will rapidly produce the solution to the linear system (11.160). Since the accuracy of the finite element solution increases with the number of mesh points, this solution scheme allows us to easily compute very accurate numerical approximations to the solution to the boundary value problem.

Example 11.14. Consider the equilibrium equations (11.12) for a non-uniform bar subject to homogeneous Dirichlet boundary conditions. In order to formulate a finite element approximation scheme, we begin with the minimization principle (11.92) based on the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2} \| u' \|^2 - \langle f \, ; u \rangle = \int_0^\ell \big[\tfrac{1}{2} c(x) u'(x)^2 - f(x) u(x) \big] \, dx. \tag{11.165}$$

We divide the interval $[0, \ell]$ into n equal subintervals, each of length $h = \ell/n$. The resulting uniform mesh consists of

$$x_j = j\, h = \frac{j\, \ell}{n}, \qquad\qquad j = 0, \dots, n. \tag{11.166}$$

The corresponding finite element basis hat functions are explicitly given by

$$\varphi_j(x) = \begin{cases} (x - x_{j-1})/h, & x_{j-1} \le x \le x_j, \\ (x_{j+1} - x)/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad\qquad j = 1, \dots, n-1. \tag{11.167}$$

The associated linear system (11.160) has coefficient matrix entries

$$m_{ij} = \langle\!\langle \, \varphi_i' \, ; \varphi_j' \, \rangle\!\rangle = \int_0^\ell \varphi_i'(x) \, \varphi_j'(x) \, c(x) \, dx, \qquad\qquad i, j = 1, \dots, n-1.$$

Since the function $\varphi_i(x)$ vanishes except on the interval $x_{i-1} < x < x_{i+1}$, while $\varphi_j(x)$ vanishes outside $x_{j-1} < x < x_{j+1}$, the integral will vanish unless $i = j$ or $i = j \pm 1$.

Moreover,

$$\varphi_j'(x) = \begin{cases} 1/h, & x_{j-1} \le x \le x_j, \\ -1/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, \dots, n-1.$$

Therefore, the coefficient matrix has the tridiagonal form

$$M = \frac{1}{h^2} \begin{pmatrix} s_0 + s_1 & -s_1 & & & & \\ -s_1 & s_1 + s_2 & -s_2 & & & \\ & -s_2 & s_2 + s_3 & -s_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -s_{n-3} & s_{n-3} + s_{n-2} & -s_{n-2} \\ & & & & -s_{n-2} & s_{n-2} + s_{n-1} \end{pmatrix}, \tag{11.168}$$

where

$$s_j = \int_{x_j}^{x_{j+1}} c(x)\, dx, \tag{11.169}$$

is the total stiffness on the j^{th} subinterval. The corresponding right hand side has entries

$$\begin{aligned} b_j = \langle\, f\,;\varphi_j\,\rangle &= \int_0^\ell f(x)\,\varphi_j(x)\,dx \\ &= \frac{1}{h}\left[\int_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x)\,dx + \int_{x_j}^{x_{j+1}} (x_{j+1} - x) f(x)\,dx\right], \end{aligned} \tag{11.170}$$

In practice, we do not have to explicitly evaluate the integrals (11.169), (11.170), but may replace them by a suitably close numerical approximation. When $h \ll 1$ is small, then the integrals are taken over small intervals, and we can use the trapezoid rule[†] to approximate them:

$$s_j \;\approx\; \frac{h}{2}\big[\, c(x_j) + c(x_{j+1})\,\big], \qquad\qquad b_j \;\approx\; h\, f(x_j). \tag{11.171}$$

For example, in the homogeneous case $c(x) \equiv 1$, the coefficient matrix (11.168) reduces to the very special form

$$M = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}. \tag{11.172}$$

[†] One might be tempted use more accurate numerical integration procedures, but the improvement in accuracy of the final answer is not very significant, particularly if the step size h is small.

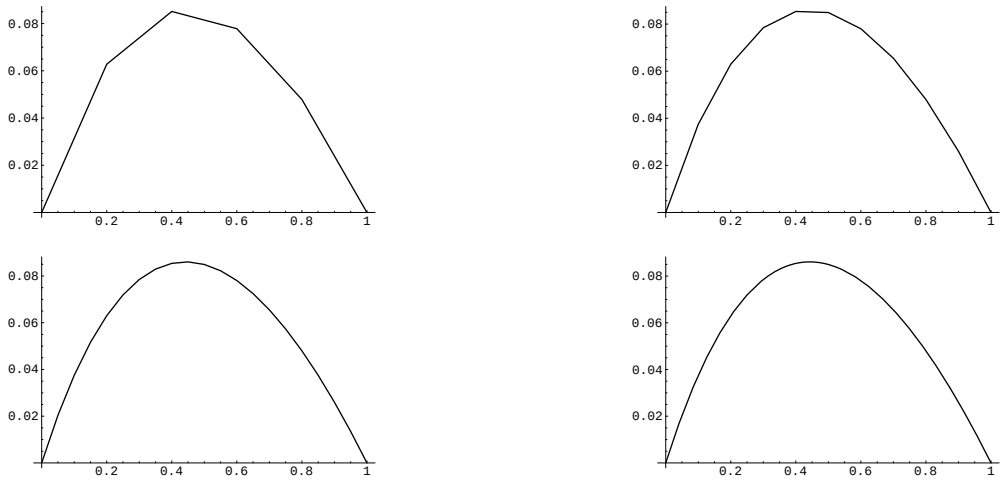


Figure 11.23. Finite Element Solution to (11.174).

The j^{th} entry of the resulting finite element system $M\mathbf{c} = \mathbf{b}$ is, upon dividing by h, given by

$$-\frac{c_{j+1} - 2\,c_j + c_{j-1}}{h^2} \;=\; -\frac{u(x_{j+1}) - 2\,u(x_j) + u(x_{j-1})}{h^2} = -f(x_j). \tag{11.173}$$

Remark: The left hand side of (11.173) is, interestingly, the standard finite difference approximation to minus the second derivative $-u''(x_j)$ of the displacement at the mesh point x_j. (Details concerning finite differences can be found in Section 14.6Therefore, for this particular differential equation, the finite element and finite difference numerical solution methods happen to coincide.

Example 11.15. Consider the boundary value problem

$$-\frac{d}{dx}\,(x+1)\,\frac{du}{dx} = 1, \qquad\qquad u(0) = 0, \qquad\qquad u(1) = 0. \tag{11.174}$$

The explicit solution is easily found by direct integration:

$$u(x) = -\,x + \frac{\log(x+1)}{\log 2}\,. \tag{11.175}$$

It minimizes the associated quadratic functional

$$\mathcal{P}[\,u\,] = \int_0^{\ell} \left[\, \tfrac{1}{2}\,(x+1)\,u'(x)^2 - u(x)\,\right] dx \tag{11.176}$$

over all possible functions $u \in \mathrm{C}^1$ subject to the given boundary conditions. The finite element system (11.160) has coefficient matrix given by (11.168) and right hand side (11.170), where

$$s_j = \int_{x_j}^{x_{j+1}} (1+x)\,dx \;= h\,(1+x_j) + \tfrac{1}{2}\,h^2 = h + h^2\,\left(j + \frac{1}{2}\right), \qquad b_j = \int_{x_j}^{x_{j+1}} 1\,dx \;= h.$$

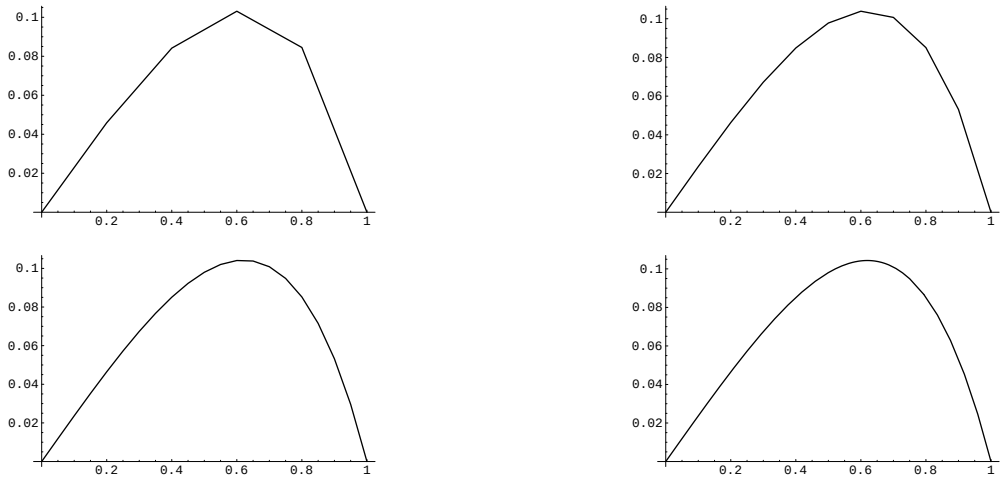


Figure 11.24. Finite Element Solution to (11.177).

The resulting solution is plotted in Figure 11.23. The first three graphs contain, respectively, 5, 10, 20 points in the mesh, so that $h = .2, .1, .05$, while the last plots the exact solution (11.175). Thus, even for rather coarse meshes, the finite element approximation is quite respectable.

Example 11.16. Consider the Sturm–Liouville boundary value problem

$$-u'' + (x+1)u = x\,e^x, \qquad u(0) = 0, \qquad u(1) = 0. \tag{11.177}$$

The solution minimizes the quadratic functional (11.153), which in this particular case is

$$\mathcal{P}[u] = \int_0^1 \left[\tfrac{1}{2}\,u'(x)^2 + \tfrac{1}{2}\,(x+1)\,u(x)^2 - e^x\,u(x) \right] dx, \tag{11.178}$$

over all functions $u(x)$ that satisfy the boundary conditions. We lay out a uniform mesh of step size $h = 1/n$ and the corresponding basis hat functions as in (11.167). Using the trapezoid method to approximate the integrals, the matrix entries are

$$m_{ij} = \int_0^1 \left[\varphi_i'(x)\,\varphi_j'(x) + (x+1)\,\varphi_i(x)\,\varphi_j(x) \right] dx \;\approx\; \begin{cases} \frac{2}{h} + \frac{2}{3}h\,(x_i+1), & i = j, \\ -\frac{1}{h} + \frac{1}{6}\,h\,(x_i+1), & |\,i-j\,| = 1, \\ 0, & \text{otherwise}, \end{cases}$$

while

$$b_i = \langle\, x\,e^x\,;\varphi_i\,\rangle = \int_0^1 x\,e^x\,\varphi_i(x)\,dx \;\approx\; x_i\,e^{x_i}\,h.$$

The resulting solution is plotted in Figure 11.24. As in the previous figure, the first three graphs contain, respectively, 5, 10, 20 points in the mesh, while the last plots the exact solution.

So far, we have only handled homogeneous boundary conditions. An inhomogeneous boundary value problem does not immediately fit into our framework since the set of

functions satisfying the boundary conditions does *not* form a subspace. As discussed at the end of Section 11.3, one way to get around this problem is to replace $u(x)$ by $\widetilde{u}(x) = u(x) - h(x)$, where $h(x)$ is any function that satisfies the boundary conditions. For example, for the inhomogeneous Dirichlet conditions

$$u(a) = \alpha, \qquad\qquad u(b) = \beta,$$

one can use an affine function

$$h(x) = \frac{(\beta - \alpha)x + \alpha\, b - \beta\, a}{b - a}.$$

Linearity implies that the difference $\widetilde{u}(x) = u(x) - h(x)$ will satisfy the modified differential equation

$$K[\widetilde{u}] = \widetilde{f}, \qquad\qquad \text{where} \qquad\qquad \widetilde{f} = f - K[h],$$

with homogeneous boundary conditions. The modified homogeneous boundary value problem can then be solved by the standard finite element method. Another possible choice for the modifier function $h(x)$ is a combination of elements at the endpoints:

$$h(x) = \alpha\, \varphi_0(x) + \beta\, \varphi_n(x),$$

where φ_0, φ_n are again piecewise affine, and equal to 1 at the end nodes $x_0 = a, x_n = b$, respectively, and zero at all other nodes. Details are left to the reader.

Finally, one can use other functions beyond the piecewise affine hat functions (11.164) to span finite element subspace. Another popular choice — essential for higher order boundary value problems such as beams — is to use splines. Thus, once we have chosen our mesh points, we can let $\varphi_j(x)$ be the basis cubic B–splines, as explained in Exercises ■, ■. The one complication is at the endpoints of the interval, where one needs to modify $\varphi_1(x)$ and $\varphi_{n-1}(x)$ to satisfy the boundary conditions. Since $\varphi_j(x) = 0$ for $x \le x_{j-2}$ or $x \ge x_{j+2}$, the coefficient matrix (11.157) is pentadiagonal, which means $m_{ij} = 0$ whenever $|\, i - j\,| > 2$. Pentadiagonal matrices are not quite as nice as their tridiagonal cousins, but they are still quite sparse. Positive definiteness of M implies that an iterative solution technique can be effectively applied to approximate the solution to the linear system, and thereby produce the finite element spline approximation to the boundary value problem.

Weak Solutions

There is an alternative way of introducing the finite element solution method, which also applies when there is no convenient minimization principle available, based on an important analytical extension of the usual notion of what constitutes a solution to a differential equation. One reformulates the differential equation as an integral equation. The resulting "weak solutions", which include non-classical solutions with singularities and discontinuities, are particularly appropriate in the study of discontinuous and nonsmooth physical phenomena, such as shock waves, cracks and dislocations in elastic media, singularities in liquid crystals, and so on; see [**147**] and Section 22.1 for details. The weak solution approach has the advantage that it applies even to equations that do not possess an associated minimization principle. However, the convergence of the induced finite element scheme is harder to justify, and, indeed, not always valid.

The starting point is a trivial observation: the only element of an inner product space which is orthogonal to every other element is zero. More precisely:

Lemma 11.17. *If V is an inner product space, then $\langle\, \mathbf{w}\,;\mathbf{v}\,\rangle = 0$ for all $\mathbf{v} \in V$ if and only if $\mathbf{w} = \mathbf{0}$.*

Proof: Choose $\mathbf{v} = \mathbf{w}$. The orthogonality condition implies $0 = \langle\, \mathbf{w}\,;\mathbf{w}\,\rangle = \|\,\mathbf{w}\,\|^2$, and so $\mathbf{w} = \mathbf{0}$. *Q.E.D.*

Note that the result is equally valid in both finite- and infinite-dimensional vector spaces. Suppose we are trying to solve a linear† system

$$K[\,\mathbf{u}\,] = \mathbf{f}, \tag{11.179}$$

where $K\colon U \to V$ is a linear operator between inner product spaces. Using the lemma, this can be reformulated as requiring

$$\langle\!\langle\, K[\,\mathbf{u}\,]\,;\mathbf{v}\,\rangle\!\rangle = \langle\!\langle\, \mathbf{f}\,;\mathbf{v}\,\rangle\!\rangle \qquad \text{for all} \qquad \mathbf{v} \in V.$$

According to the definition (7.64), one can replace K by its adjoint $K^*\colon W \to V$, and require

$$\langle\, \mathbf{u}\,;K^*[\,\mathbf{v}\,]\,\rangle = \langle\!\langle\, \mathbf{f}\,;\mathbf{v}\,\rangle\!\rangle \qquad \text{for all} \qquad \mathbf{v} \in V. \tag{11.180}$$

The latter is called the *weak formulation* of our original equation. The general philosophy is that one can check whether $\mathbf{u}$ is a weak solution to the system by evaluating it on various *test elements* $\mathbf{v}$ using the weak form (11.180) of the system.

In the finite-dimensional situation, when K is merely multiplication by some matrix, the weak formulation is an unnecessary complication, and not of use. However, in the infinite-dimensional situation, when K is a differential operator, then the original boundary value problem $K[u] = f$ requires that u be sufficiently differentiable, whereas the weak version

$$\langle\, u\,;K^*[\,\varphi\,]\,\rangle = \langle\!\langle\, f\,;\varphi\,\rangle\!\rangle \qquad \text{for all} \qquad \varphi$$

requires only that the *test function* $\varphi(x)$ be smooth. As a result, weak solutions are not restricted to be smooth functions possessing the required number of derivatives.

Example 11.18. Consider the homogeneous Dirichlet boundary value problem

$$K[u] = -\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad 0 < x < \ell, \qquad u(0) = u(\ell) = 0,$$

for a nonuniform bar. Its weak version is obtained by integration by parts. We initially restrict to test functions which vanish at the boundary $\varphi(0) = \varphi(\ell) = 0$. This requirement will eliminate any boundary terms in the integration by parts computation

$$\begin{aligned} \langle\, K[u]\,;\varphi\,\rangle = \int_0^\ell \left[-\frac{d}{dx}\left(\,c(x)\,\frac{du}{dx}\,\right)\,\varphi(x)\,\right] dx &= -\int_0^\ell c(x)\,\frac{du}{dx}\,\frac{d\varphi}{dx}\,dx \\ &= \int_0^\ell f(x)\,\varphi(x)\,dx = \langle\, f\,;\varphi\,\rangle. \end{aligned} \tag{11.181}$$

† The method also straightforwardly extends to nonlinear systems.

This "semi-weak" formulation is known in mechanics as the *principle of virtual work*, [**134**]. For example, the Green's function of the boundary value problem does not qualify as a classical solution since it is not twice continuously differentiable, but can be formulated as a weak solution satisfying the virtual work equation with right hand side defined by the delta forcing function.

A second integration by parts produces the weak form (11.180) of the differential equation:

$$\langle\, u\,;K[\,\varphi\,]\,\rangle = -\int_0^\ell u(x)\,\frac{d}{dx}\left(\,c(x)\,\frac{d\varphi}{dx}\,\right)dx = \int_0^\ell f(x)\,\varphi(x)\,dx = \langle\, f\,;\varphi\,\rangle. \tag{11.182}$$

Now, even discontinuous functions $u(x)$ are allowed as weak solutions. The goal is to find $u(x)$ such that this condition holds for all smooth test functions $\varphi(x)$. For example, any function $u(x)$ which satisfies the differential equation (11.12) except at points of discontinuity qualifies as a weak solution.

In a *finite element* or *Galerkin approximation* to the weak solution, one restricts attention to a finite-dimensional subspace W spanned by functions $\varphi_1,\dots,\varphi_{n-1}$, and requires that the approximate solution

$$w(x) = c_1\,\varphi_1(x) + \,\cdots\, + c_{n-1}\,\varphi_{n-1}(x) \tag{11.183}$$

satisfy the orthogonality condition (11.180) only for elements $\varphi \in W$ of the subspace. As usual, this only needs to be checked on the basis elements. Substituting (11.183) into the semi-weak form of the system, (11.181), produces a linear system of equations of the form

$$\langle\, w\,;K[\,\varphi_i\,]\,\rangle = \sum_{i=1}^{n} m_{ij}\,c_j = b_i = \langle\, f\,;\varphi_i\,\rangle, \qquad\qquad i = 1,\dots,n. \tag{11.184}$$

The reader will recognize this as exactly the same finite element linear system (11.160) derived through the minimization approach. Therefore, for a self-adjoint boundary value problem, the weak formulation and the minimization principle, when restricted to the finite-dimensional subspace W, lead to exactly the same equations for the finite element approximation to the solution.

In non-self-adjoint scenarios, the weak formulation is still applicable even though there is no underlying minimization principle. On the other hand, there is no guarantee that either the original boundary value problem or its finite element approximation have a solution. Indeed, it is entirely possible that the boundary value problem has a solution, but the finite element matrix system does not. Even more worrying are cases in which the finite element system has a solution, but there is, in fact, no actual solution to the boundary value problem! In such cases, one is usually tipped off by the non-convergence of the approximations as the mesh size goes to zero. Nevertheless, in many situations, the weak solution approach leads to a perfectly acceptable numerical approximation to the true solution to the system. Further analytical details and applications of weak solutions can be found in [**63**, **147**].

Chapter 12

Fourier Series

Just before 1800, the French mathematician/physicist/engineer Jean Baptiste Joseph Fourier made an astonishing discovery. As a result of his investigations into the partial differential equations modeling heat propagation in bodies, Fourier was led to claim that "every" function could be represented by an infinite series of elementary trigonometric functions — sines and cosines. Consider the sound produced by a musical instrument, e.g., piano, violin, trumpet, oboe, or drum. Decomposing the signal into a Fourier series reveals the fundamental frequencies (tones, overtones, etc.) that are combined to produce its distinctive tones. The Fourier decomposition lies at the heart of modern electronic music; a synthesizer combines pure sine and cosine tones to reproduce the diverse sounds of instruments, both natural and artificial, according to Fourier's general prescription.

Fourier's claim was so remarkable and unexpected that most of the leading mathematicians of the time did not believe him. Nevertheless, it was not long before scientists came to appreciate its power and far-ranging applicability, opening up vast new realms of physics, engineering, and, later, many other fields, to mathematical analysis. Indeed, Fourier's discovery easily ranks in the "top ten" mathematical advances of all time, a list that would include Newton's invention of the calculus and Riemann's establishment of differential geometry that, 70 years later, formed the foundation of Einstein's theory of relativity. Fourier analysis is an essential component of much of modern applied (and pure) mathematics. It forms an exceptionally powerful analytical tool for solving and analyzing a broad range of partial differential equations. Applications in pure mathematics, physics and engineering are almost too numerous to catalogue — typing in "Fourier" in the subject index of a modern science library will dramatically demonstrate just how ubiquitous these methods are. For instance, modern signal processing, including audio, speech, images, videos, seismic data, radio transmissions, and so on, is based on Fourier analysis and its variants. Many modern technological advances, including television, music CD's and DVD's, video movies, computer graphics, image processing, and fingerprint analysis and storage, are, in one way or another, founded upon the many ramifications of Fourier's discovery. In your career as a mathematician, scientist or engineer, you will find that Fourier theory, like calculus and linear algebra, is one of the most basic and essential tools in your mathematical arsenal. Mastery of the subject will be unavoidable.

In addition, a remarkably large fraction of modern pure mathematics is the result of subsequent attempts to place Fourier's discoveries on a firm mathematical foundation. Thus, all of the student's "favorite" analytical tools, including the modern definition of a function, the ε–δ definition of limit and continuity, convergence properties in function space, including uniform convergence, weak convergence, etc., the modern theory of inte-

gration and measure, generalized functions such as the delta function, and many others, all owe a profound debt to the prolonged struggle to establish the theory of Fourier series and integrals on a rigorous foundation. Even more remarkably, modern set theory, and, as a result, mathematical logic and foundations, can be traced directly back to Cantor's attempts to understand the sets upon which Fourier series converge!

Fourier's ideas are a very natural outgrowth of the same basic constructions that we have already developed in the finite-dimensional linear algebra context for analyzihng discrete dynamical processes. The Fourier representation of a function is a continuous counterpart of the eigenvalue and eigenvector expansions used to solve linear dynamical systems of ordinary differential equations. The fundamental partial differential equations governing heat propagation and vibrations in continuous media are the function space versions of these discrete systems. Replacing the eigenvectors are eigenfunctions. The trigonometric functions in the Fourier series are the eigenfunctions for a certain simple self-adjoint linear boundary value problem of the type considered in the preceding chapter, and hence play the same role as the eigenvectors do in the solution to discrete systems of ordinary differential equations. The key difference, which greatly magnifies the theoretical complications, is that we must deal with infinite series rather than finite sums. Moreover, self-adjointness The Fourier coefficient formulae are the direct analogues of standard orthogonality formulae among the eigenvectors of symmetric matrices. Thus, the Fourier series is the *simplest*, but far from the only, function space analog of the spectral theory of symmetric matrices. Their role in the solution to the partial differential equations of heat diffusion and vibration are obtained by arguing by analogy with the solution methods for the systems of ordinary differential equations modeling vibrations and gradient flows of discrete mechanical systems. the Fourier trigonometric functions are the very simplest example of an orthogonal system of eigenfunctions of a self-adjoint boundary value problem. The orthogonality conditions lead directly to the mysterious formulae for the Fourier coefficients and hence effectively implement a solution to the partial differential equations.

The key to the efficacy of Fourier series rests on the orthogonality properties of the trigonometric functions, which is a direct consequence of their status as eigenfunctions of the most basic self-adjoint boundary value problem. In this manner, Fourier series can also be viewed as a function space version of the finite-dimensional spectral theory of symmetric matrices and orthogonal eigenvector bases. The main complication is that we must now deal with infinite series rather than finite sums, and so convergence issues that do not appear in the finite-dimensional situation become of paramount importance. The Fourier trigonometric series is the *simplest* representative of a broad class of infinite eigenfunction series based on self-adjoint boundary value problems. Other important examples arising in physical applications, including Bessel and Legendre functions, will appear in Chapters 17 and 18.

Once we have established the proper theoretical background, the Fourier series will no longer be a special, isolated phenomenon, but rather appears in its natural context as the simplest of a large class of eigenfunction expansions for solving a wide range of linear problems. Modern and classical extensions of the Fourier method, including Bessel and Legendre functions, Fourier integrals, wavelets, discrete Fourier series, and many others, all rely on the same foundations of complete, orthogonal systems of functions in an infinite-

dimensional vector space. Many of the most important cases will appear in the ensuing chapters.

12.1. Dynamical Equations of Continuous Media.

The purpose of this motivational section is to understand why Fourier series naturally appear when we move from discrete systems of ordinary differential equations to the partial differential equations that govern the dynamics of continuous mechanical systems. As we will return to these issues in full detail in Chapter 14, readers wishing to dive straight into Fourier methods might want to skip ahead to the next section.

In continuum mechanics, we replace a system of discrete masses and springs by a continuum, e.g., a one-dimensional bar or string, a two-dimensional plate or a three-dimensional solid body. Of course, real physical bodies are composed of atoms, and so could, at least in principle, be modeled by discrete mechanical systems. However, the number of atoms is so large that any direct attempt to analyze the resulting system of ordinary differential equations would be completely impractical. Thus, regarding bodies as ideal continuous media not only leads to very accurate physical models, but is absolutely crucial for significant progress towards their mathematical and computational analysis. Paradoxically, the numerical solution of such partial differential equations returns us to the discrete realm, but now to a computationally tractable system. While one might envision going directly from the discrete atomic system to the discrete numerical approximation, this is not such a simple matter. The analytical power and insight offered by calculus in the continuous regime makes this intermediate construction an essential component for the effective modeling of physical phenomena.

The two principal classes that will be treated are the first order systems (9.22) governing gradient flows, and the second order Newtonian vibration systems (9.54). The former will naturally lead to diffusion equations, including the heat equation that models the propagation of heat in a homogeneous body. The latter will lead to general vibration equations such as the wave equation, modeling the vibrational motions of bars, strings, and, in higher dimensions, plates and solid bodies.

As we saw in Chapter 6, the equilibrium configurations of discrete mechanical systems, such as a system of masses and springs, are found by solving a linear algebraic system $K\mathbf{u} = \mathbf{f}$ with positive (semi-)definite coefficient matrix for the displacement vector $\mathbf{u}$. In Chapter 11, we learned that the same abstract formalism applies to the equilibrium equations of one-dimensional media — bars, beams, etc. The solution is now a function $u(x)$ representing, say, displacement of the bar, while the positive (semi-)definite matrix is replaced by a certin positive (semi-)definite boundary value problem $K[u] = f$ involving a linear ordinary differential equation along with a suitable collection of boundary conditions.

The dynamics of discrete systems are governed by initial value problems for linear systems of ordinary differential equations. Thus, we expect that the dynamics of continua will be modeled by partial differential equations involving both space and time variables, along with suitable boundary conditions in space and initial conditions in time. Let us first consider the unforced gradient flow system

$$\frac{d\mathbf{u}}{dt} = -K\mathbf{u}, \qquad \mathbf{u}(t_0) = \mathbf{u}_0, \tag{12.1}$$

associated with the positive (semi-)definite coefficient matrix. Gradient flows are designed to decrease the quadratic energy function $q(\mathbf{u}) = \frac{1}{2}\,\mathbf{u}^T K\,\mathbf{u}$ as rapidly as possible. By analogy, the corresponding continuous gradient flow will take the form[†]

$$\frac{\partial u}{\partial t} = -K[u], \qquad u(t_0, x) = f(x), \tag{12.2}$$

in which the differential operator K incorporates the same equilibrium spatial boundary conditions. Such partial differential equations model diffusion processes in which a quadratic energy functional is decreasing as rapidly as possible. A good physical example is the flow of heat in a body; the heat disperses throughout the body so as to decrease the thermal energy as quickly as it can, tending (in the absence of external heat sources) to thermal equilibrium. Other physical processes modeled by (12.2) include diffusion of chemicals (solvents, pollutants, etc.), and of populations (animals, bacteria, people, etc.) in a medium.

The simplest and most instructive example is the case of a uniform periodic (or circular) bar of length 2π. As we saw in Chapter 11, the equilibrium equation takes the form

$$K[u] = -u'' = f, \qquad u(-\pi) = u(\pi), \qquad u'(-\pi) = u'(\pi), \tag{12.3}$$

associated with the positive semi-definite differential operator

$$K = D^* \circ D = (-D)\,D = -D^2 \tag{12.4}$$

acting on the space of 2π periodic functions. The corresponding gradient flow (12.2) is the partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\,, \qquad u(t, -\pi) = u(t, \pi), \qquad \frac{\partial u}{\partial x}\,(t, -\pi) = \frac{\partial u}{\partial x}\,(t, \pi), \tag{12.5}$$

known as the *heat equation* since it models (among other diffusion processes) thermodynamics in one-dimensional media. The function $u(t, x)$ represents the temperature at position x and time t. Heat naturally flows from hot to cold, and so the fact that it can be described by a gradient flow should not be surprising; a physical derivation of (12.5) will appear in Chapter 14. Solving the periodic heat equation was the seminal problem that led Fourier to develop the profound theory that now bears his name.

Now, to solve a finite-dimensional gradient flow (12.1), we imposed a exponential *ansatz* (or inspired guess) for a solution $\mathbf{u}(t) = e^{-\lambda t}\,\mathbf{v}$. Substituting the formula reduced the differential equation to the eigenvalue equation $K\,\mathbf{v} = \lambda\,\mathbf{v}$. Each eigenvalue and eigenvector yields a basic solution to the dynamical system, and the general solution is obtained by superposition of these fundamental modes. This same idea carries over directly to the continuous realm! To find solutions to the partial differential equation (12.2), we introducing the exponential formula

$$u(t, x) = e^{-\lambda t}\,v(x), \tag{12.6}$$

[†] Since $u(t, x)$ now depends upon time as well as position, we switch from ordinary to partial derivative notation.

in which we replace the eigenvector $\mathbf{v}$ by a function $v(x)$ that satsifies the relevant boundary conditions. We compute

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial t}\left[\, e^{-\lambda t}\, v(x)\,\right] = -\lambda\, e^{-\lambda t}\, v(x), \quad \text{while} \quad -K[u] = -K\left[\, e^{-\lambda t}\, v(x)\,\right] = -e^{-\lambda t}\, K[v],$$

since the exponential factor is a function of t, while the differential operator K only involves differentiation with respect to x. Substituting these two expressions into the dynamical equations (12.2) and cancelling the common exponential factor, we conclude that $v(x)$ must solve a boundary value problem of the form

$$K[v] = \lambda\, v. \tag{12.7}$$

We interpret λ as the *eigenvalue* and $v(x)$ as the corresponding *eigenfunction* for the operator K subject to the relevant boundary conditions. As in the finite-dimensional case, positive definiteness will imply that all eigenvalues are strictly positive definite, $\lambda > 0$; in the unstable, positive semi-definite situation, null eigenvalues $\lambda = 0$ may also appear. Each eigenvalue and eigenfunction pair will produce a solution (12.6) to the partial differential equation, and the general solution can be built up through linear superposition.

For example, substitution of the exponential ansatz (12.6) into the periodic heat equation (12.5) leads to the eigenvalue problem

$$v'' + \lambda\, v = 0, \qquad v(-\pi) = v(\pi), \qquad v'(-\pi) = v'(\pi). \tag{12.8}$$

This constitutes a $2\,\pi$ periodic boundary value problem for the *eigenfunction* $v(x)$. Positive semi-definiteness of the underlying differential operator (12.4) implies that its eigenvalues must be real and non-negative: $\lambda \geq 0$. Indeed, as the reader can verify, if $\lambda < 0$ or λ is complex, then the only periodic solution to (12.8) is the trivial solution $v(x) \equiv 0$. When $\lambda = 0$, the periodic solutions to (12.8) are the constant functions, and so any nonzero constant function $v(x) \equiv c$ is an eigenfunction for the $\lambda = 0$ eigenvalue. For the positive eigenvalues, if we write $\lambda = \omega^2$ with $\omega > 0$, then the general solution to the differential equation (12.8) is a linear combination

$$v(x) = a\, \cos \omega\, x + b\, \sin \omega\, x.$$

A nonzero function of this form will satisfy the $2\,\pi$ periodic boundary conditions if and only if $\omega = k$ is an integer. Therefore, the eigenvalues

$$\lambda = k^2, \qquad 0 \leq k \in \mathbb{N},$$

are the squares of positive integers. Each positive eigenvalue $\lambda = k^2 > 0$ admits two linearly independent eigenfunctions, namely $\sin k\, x$ and $\cos k\, x$, while the zero eigenvalue $\lambda = 0$ has only one independent eigenfunction, the constant function 1. We conclude that the standard trigonometric functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos 2\, x, \qquad \sin 2\, x, \qquad \cos 3\, x, \qquad \dots \tag{12.9}$$

form a complete system of independent eigenfunctions for the periodic boundary value problem (12.8).

Each eigenfunction gives rise to a particular solution to the periodic heat equation (12.5). We have therefore constructed an infinite collection of independent solutions:

$$u_k(x) = e^{-k^2 t} \cos k\, x, \qquad \widetilde{u}_k(x) = e^{-k^2 t} \sin k\, x, \qquad k = 0, 1, 2, 3, \dots .$$

According to our linear superposition principle, any finite linear combination

$$u(t,x) = a_0 + \sum_{k=1}^{n} \Big[a_k\, e^{-k^2 t} \cos k\, x + b_k\, e^{-k^2 t} \sin k\, x \Big] \tag{12.10}$$

of these particular solutions is also a solution. However, finite linear combinations will not suffice to describe the general solution to the problem, and we must replace the finite sum (12.10) by an infinite series. This immediately raises deep and interesting analytical questions. When does such an infinite series converge? Can we represent a given function $f(x)$ as such an infinite series, and if so, how? For the trigonometric eigenfunctions, these are the fundamental questions of Fourier analysis. After we have firmly established the basics of Fourier theory, we shall then return to these questions for both the heat and wave equations in Chapter 14.

A similar analysis applies to a second order system of the Newtonian form

$$\frac{\partial^2 u}{\partial t^2} = -K[u]. \tag{12.11}$$

Such differential equations are used to describe the free vibrations of continuous mechanical systems, such as bars, strings, and, in higher dimensions, membranes, solid bodies, fluids, etc. For example, the vibration system (12.11) corresponding to the differential operator (12.4) is the *wave equation*

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}\,. \tag{12.12}$$

The wave equation models stretching vibrations of a bar, sound vibrations in a column of air, e.g., inside a wind instrument, transverse vibrations of a string, e.g., a violin string, surfaces waves on a fluid, electromagnetic waves, and a wide variety of other vibrational and wave phenomena.

As always, we need to impose suitable boundary conditions in order to proceed. Consider, for example, the wave equation with homogeneous Dirichlet boundary conditions

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}\,, \qquad u(t,0) = 0, \qquad u(t,\ell) = 0, \tag{12.13}$$

that models, for instance, the vibrations of a uniform violin string whose ends are tied down. Adapting our discrete trigonometric ansatz, we are naturally led to look for a separable solution of the form

$$u(t,x) = \cos(\omega t)\, v(x) \tag{12.14}$$

in which ω represents the vibrational frequency. Substituting into the wave equation and the associated boundary conditions, we deduce that $v(x)$ must be a solution to the eigenvalue problem

$$\frac{d^2 v}{dx^2} + \omega^2\, v = 0, \qquad v(0) = 0 = v(\ell), \tag{12.15}$$

in which $\omega^2 = \lambda$ plays the role of the eigenvalue. For $\omega^2 > 0$ — which results from positive definiteness of the underlying system — the general solution to the differential equation is a trigonometric function

$$v(x) = a \cos \omega x + b \sin \omega x.$$

The boundary condition at $x = 0$ requires $a = 0$, and so

$$v(x) = b \sin \omega x.$$

The second boundary condition requires

$$v(\ell) = b \sin \omega \ell = 0.$$

Assuming $b \neq 0$, as otherwise the solution is trivial, $\omega \ell$ must be an integer multiple of π. Thus, the natural frequencies of vibration are

$$\omega_k = \frac{k \pi}{\ell}, \qquad k = 1, 2, 3, \ldots .$$

The corresponding eigenfunctions are

$$v_k(x) = \sin \frac{k \pi x}{\ell}, \qquad k = 1, 2, 3, \ldots . \tag{12.16}$$

Thus, we find the following natural modes of vibration of the wave equation:

$$u_k(t,x) = \cos \frac{k \pi t}{\ell} \, \sin \frac{k \pi x}{\ell}, \qquad \widetilde{u}_k(t,x) = \sin \frac{k \pi t}{\ell} \, \sin \frac{k \pi x}{\ell}.$$

Each solution represents a spatially periodic standing wave form. We expect to write the general solution to the boundary value problem as an infinite series

$$u(t,x) = \sum_{k=1}^{\infty} \left(b_k \cos \frac{k \pi t}{\ell} \, \sin \frac{k \pi x}{\ell} + d_k \sin \frac{k \pi t}{\ell} \sin \frac{k \pi x}{\ell} \right) \tag{12.17}$$

in the natural modes. Interestingly, in this case at each fixed t, there are no cosine terms, and so we have a more specialized type of Fourier series. The same convergence issues for such *Fourier sine series* arise. It turns out that the general theory of Fourier series will also cover Fourier sine series.

We have now completed our brief introduction to the dynamical equations of continuous media and the Fourier series method of solution. The student should now be sufficiently motivated, and it is time to delve into the theory of basic Fourier series. We will not try to deal with more general eigenfunction expansions until Chapter 17, but instead concentrate on the simplest and most important special case, when the eigenfunctions are trigonometric functions, and the series is a classical Fourier series. In Chapter 14 we will return to the applications to the one-dimensional heat and wave equations.

12.2. Fourier Series.

While the need to solve physically interesting partial differential equations served as our (and Fourier's) initial motivation, the remarkable range of applications qualifies Fourier's discovery as one of the most important in all of mathematics. We therefore take some time to properly develop the basic theory of Fourier series and, in the following chapter, a number of important extensions. Then, properly equipped, we will be in a position to return to the source — solving partial differential equations.

We commence the discussion with the fundamental definition.

Definition 12.1. A *Fourier series* is an infinite trigonometric series

$$f(x) \quad \sim \quad \frac{a_0}{2} \; + \; \sum_{k=1}^{\infty} \, [\, a_k \, \cos k\, x + b_k \, \sin k\, x \,] \, . \tag{12.18}$$

The extra factor of $\frac{1}{2}$ is traditionally included in the first term for later convenience.

Of course, without additional assumptions on the coefficients a_k, b_k, the Fourier series (12.18) may not converge. This is the reason that we use the $\sim$ symbol instead of an equals sign. The key questions are

(i) First, when does such an infinite trigonometric series converge?

(ii) Second, what kinds of functions $f(x)$ can be represented by a convergent Fourier series?

(iii) Third, if we have such an f, how do we determine its Fourier coefficients a_k, b_k?

(iv) And lastly, since we are trying to solve differential equations, can we safely differentiate a Fourier series?

The first order of business is to determine the formulae for the Fourier coefficients a_k, b_k. The key is orthogonality. We already observed, in Example 5.12, that the trigonometric functions (12.9) form an orthogonal system of functions with respect to the (rescaled) L^2 inner product

$$\langle \, f \, ; g \, \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \, g(x) \, dx \tag{12.19}$$

on the interval[†] $[\, -\pi, \pi \,]$. The explicit orthogonality relations are

$$\begin{aligned} \langle \cos k\, x \, ; \cos l\, x \, \rangle = \langle \sin k\, x \, ; \sin l\, x \, \rangle = 0, \qquad & \text{for} \qquad k \neq l, \\ \langle \cos k\, x \, ; \sin l\, x \, \rangle = 0, \qquad & \text{for all} \qquad k, l, \\ \| \, 1 \, \| = \sqrt{2} \, , \qquad \| \cos k\, x \, \| = \| \sin k\, x \, \| = 1, \qquad & \text{for} \qquad k \neq 0, \end{aligned} \tag{12.20}$$

whenever k and l are non-negative integers.

[†] We have chosen the interval $[\, -\pi, \pi \,]$ for convenience. A common alternative choice is the interval $[\, 0, 2\pi \,]$. In fact, since the trigonometric functions are 2π periodic, any interval of length 2π will serve equally well. Adapting Fourier series to intervals of other lengths will be discussed in Section 12.4.

Remark: If we were to replace the constant function 1 by $\frac{1}{\sqrt{2}}$, then the resulting functions would form an orthonormal system. However, this extra $\sqrt{2}$ factor turns out to be utterly annoying, and is best omitted from the outset.

If we ignore convergence issues for the moment and treat the Fourier series representation (12.18) as an equality, then the orthogonality relations (12.20) serve to immediately determine the Fourier coefficients. Taking the inner product of both sides with, respectively, $\cos k\,x$ and $\sin k\,x$, we find

$$\begin{aligned} a_k &= \langle\, f\,; \cos k\,x\,\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos k\,x\,dx, \qquad & k &= 0, 1, 2, 3, \ldots, \\ b_k &= \langle\, f\,; \sin k\,x\,\rangle = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin k\,x\,dx, \qquad & k &= 1, 2, 3, \ldots. \end{aligned} \tag{12.21}$$

These fundamental formulae prescribe the *Fourier coefficients* of the function f. The fact that we can also use them as written for a_0 is the reason for including the $\frac{1}{2}$ in the constant term of the Fourier series (12.18).

Example 12.2. Consider the function $f(x) = x$. We may compute its Fourier coefficients directly, using integration by parts to evaluate the integrals:

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\,dx = 0, \qquad a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x\cos k\,x\,dx = \frac{1}{\pi}\left[\,\frac{x\,\sin k\,x}{k} + \frac{\cos k\,x}{k^2}\,\right]\Bigg|_{x=-\pi}^{\pi} = 0, \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin k\,x\,dx = \frac{1}{\pi}\left[-\frac{x\,\cos k\,x}{k} + \frac{\sin k\,x}{k^2}\,\right]\Bigg|_{x=-\pi}^{\pi} = \frac{2}{k}\,(-1)^{k+1}. \end{aligned} \tag{12.22}$$

Therefore, the Fourier cosine coefficients of the function x all vanish, $a_k = 0$, and its Fourier series is

$$x \;\sim\; 2\left(\sin x \;-\; \frac{\sin 2\,x}{2} \;+\; \frac{\sin 3\,x}{3} \;-\; \frac{\sin 4\,x}{4} + \;\cdots\right). \tag{12.23}$$

The convergence of this series is *not* an elementary matter. Standard tests, including the ratio and root tests, that almost always work for power series, fail to apply. Even if we know that the series converges (which it does, for all x) it is certainly not obvious what function it converges to. Indeed, it cannot converge to the function $f(x) = x$ for all values of x. If we substitute $x = \pi$, then every term in the series is zero, and so the Fourier series converges to 0 — which is not the same as $f(\pi) = \pi$.

The n^{th} *partial sum* of a Fourier series is the trigonometric polynomial[†]

$$s_n(x) \;=\; \frac{a_0}{2} + \sum_{k=1}^{n}\,\left[\,a_k\,\cos k\,x + b_k\,\sin k\,x\,\right]. \tag{12.24}$$

[†] The reason for the term "trigonometric polynomial" was discussed at length in Example 2.16(c).

Thus, the Fourier series *converges* at a point x if and only if the partial sums have a limit

$$\lim_{n \to \infty} s_n(x) = \widetilde{f}(x), \tag{12.25}$$

which may or may not equal the value of the original function $f(x)$. Thus, a key requirement is to formulate easily verifiable conditions on the function $f(x)$ that guarantee that the Fourier series converges, and, even more importantly, the limiting sum reproduces the original function: $\widetilde{f}(x) = f(x)$. This will all be done in detail below.

Remark: The passage from trigonometric polynomials to Fourier series is analogous to the passage from polynomials to power series. A power series

$$f(x) \quad \sim \quad c_0 + c_1\, x + \, \cdots \, + c_n\, x^n + \, \cdots \, = \, \sum_{k=0}^{\infty} c_k\, x^k$$

can be viewed as an infinite linear combination of the basic monomials $1, x, x^2, x^3, \ldots$. According to Taylor's formula, (C.8), the coefficients $c_k = \dfrac{f^{(k)}(0)}{k!}$ are given in terms of the derivatives of the function at the origin. The partial sums

$$s_n(x) = c_0 + c_1\, x + \, \cdots \, + c_n\, x^n \; = \; \sum_{k=0}^{n} c_k\, x^k$$

of a power series are ordinary polynomials, and similar convergence issues arise.

Although superficially similar, in actuality the two theories are profoundly different. A power series either converges everywhere, or on an interval centered at 0, or nowhere except at 0. (See Section 16.2 for details.) On the other hand, a Fourier series can converge on quite bizarre sets. In fact, the detailed analysis of the convergence properties of Fourier series led Georg Cantor to formulate modern set theory, and, thus, played a seminal role in the establishment of the foundations of mathematics. Secondly, when a power series converges, it converges to an analytic function, which is infinitely differentiable, and whose derivatives are represented by the power series obtained by termwise differentiation. Fourier series may converge, not only to a periodic continuous function, but also to a wide variety of discontinuous functions and, even, when suitably interpreted, to generalized functions like the delta function! Therefore, the termwise differentiation of a Fourier series is a nontrivial issue. Indeed, while the theory of power series was well established in the early days of the calculus, there remain, to this day, unresolved foundational issues in Fourier theory.

Once one comprehends how different the two subjects are, one begins to understand why Fourier's astonishing claims were intially widely disbelieved. Before the advent of Fourier, mathematicians only viewed analytic functions as genuine. The fact that Fourier series can converge to nonanalytic, even discontinuous functions was extremely disconcerting, and led to a complete re-evaluation of function theory, culminating in the modern definition of function that you now learn in first year calculus. Only through the combined efforts of many of the leading mathematicians of the nineteenth century was a rigorous theory of Fourier series firmly established.

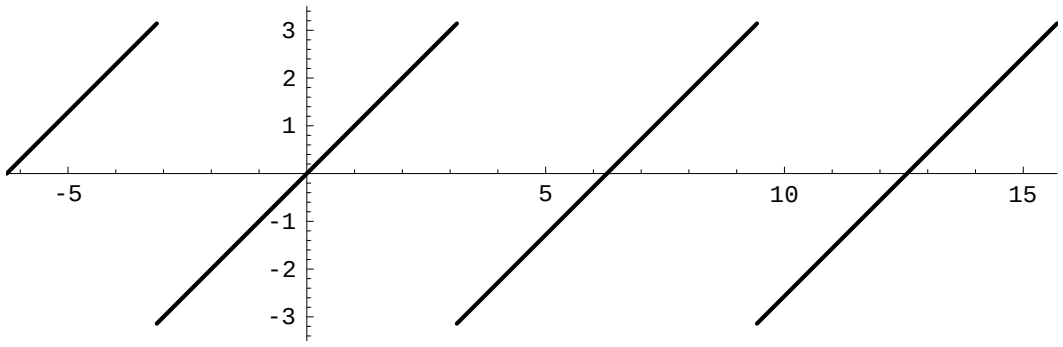


Figure 12.1. Periodic extension of x.

Periodic Extensions

The trigonometric constituents (12.9) of a Fourier series are all periodic functions of period 2π. Therefore, if the series converges, the resulting function $\widetilde{f}(x)$ must also be periodic of period 2π:

$$\widetilde{f}(x + 2\pi) = \widetilde{f}(x) \qquad \text{for all} \qquad x \in \mathbb{R}.$$

A Fourier series can only converge to a 2π periodic function. Therefore, we should not expect the Fourier series (12.23) to converge to $f(x) = x$ everywhere, since it is not a periodic function. Rather, it converges to its periodic extension, as we now define.

Lemma 12.3. *If $f(x)$ is any function defined for $-\pi < x \leq \pi$, then there is a unique 2π periodic function $\widetilde{f}$, known as the 2π* periodic extension *of f, that satisfies $\widetilde{f}(x) = f(x)$ for all $-\pi < x \leq \pi$.*

Proof: Pictorially, the graph of the periodic extension of a function $f(x)$ is obtained by repeatedly copying that part of the graph of f between $-\pi$ and π to all other adjacent intervals of length 2π; see, for instance, Figure 12.1. More formally, given $x \in \mathbb{R}$, there is a unique integer m so that $-\pi < x - 2m\pi \leq \pi$. Periodicity of $\widetilde{f}$ leads us to define

$$\widetilde{f}(x) = \widetilde{f}(x - 2m\pi) = f(x - 2m\pi).$$

In particular, if $-\pi < x \leq \pi$, then $m = 0$ and hence $\widetilde{f}(x) = f(x)$. The proof that the resulting function $\widetilde{f}$ is 2π periodic is left as Exercise ∎. *Q.E.D.*

Remark: The construction of the periodic extension of Lemma 12.3 uses the value $f(\pi)$ at the right endpoint and requires $\widetilde{f}(-\pi) = \widetilde{f}(\pi) = f(\pi)$. One could, alternatively, require $\widetilde{f}(\pi) = \widetilde{f}(-\pi) = f(-\pi)$, which, if $f(-\pi) \neq f(\pi)$, leads to a slightly different 2π periodic extension of the function, differing when x is an odd multiple of π. There is no *a priori* reason to prefer one over the other. In fact, for Fourier theory, as we shall discover, one should use neither, but rather an "average" of the two. Thus, the preferred Fourier periodic extension $\widetilde{f}(x)$ will satisfy

$$\widetilde{f}(\pi) = \widetilde{f}(-\pi) = \tfrac{1}{2}\big[\, f(\pi) + f(-\pi) \,\big], \tag{12.26}$$

which then fixes its values at the odd multiples of π.

Example 12.4. The 2π periodic extension $\widetilde{f}(x)$ of $f(x) = x$ is the "sawtooth" function graphed in Figure 12.1. It agrees with x between $-\pi$ and π. If we adopt the Fourier extension (12.26), then, for any odd integer k, we set $\widetilde{f}(k\pi) = 0$, the average of the values of $f(x) = x$ at the endpoints $\pm\pi$. Thus, explicitly,

$$\widetilde{f}(x) = \begin{cases} x - 2m\pi, & (2m-1)\pi < x < (2m+1)\pi, \\ 0, & x = (2m-1)\pi, \end{cases} \qquad \text{where } m \text{ is an arbitrary integer.}$$

With this convention, it can be proved that the Fourier series (12.23) for $f(x) = x$ converges everywhere to the 2π periodic extension $\widetilde{f}(x)$. In particular,

$$2\sum_{k=1}^{\infty} (-1)^{k+1}\,\frac{\sin k\,x}{k} = \begin{cases} x, & -\pi < x < \pi, \\ 0, & x = \pm\pi. \end{cases} \tag{12.27}$$

Even this very simple example has remarkable and nontrivial consequences. For instance, if we substitute $x = \frac{1}{2}\pi$ in (12.23) and divide by 2, we obtain *Gregory's series*

$$\frac{\pi}{4} \;=\; 1 \,-\, \frac{1}{3} \,+\, \frac{1}{5} \,-\, \frac{1}{7} \,+\, \frac{1}{9} \,-\, \cdots\,. \tag{12.28}$$

While this striking formula predates Fourier theory — it was first discovered by Leibniz — a direct proof is not easy.

Remark: While fascinating from a numerological viewpoint, Gregory's series is of scant practical use for actually computing π since its rate of convergence is painfully slow. The reader may wish to try adding up terms to see how far out one needs to go to accurately compute even the first two decimal digits of π. Round-off errors will eventually interfere with any attempt to compute the complete summation to any reasonable degree of accuracy.

Piecewise Continuous Functions

As we shall see, all continuously differentiable, 2π periodic functions can be represented as convergent Fourier series. More generally, we can allow the function to have some simple discontinuities. Although not the most general class of functions that possess convergent Fourier series, such "piecewise continuous" functions will suffice for all the applications we consider in this text.

Definition 12.5. A function $f(x)$ is said to be *piecewise continuous* on an interval $[a, b]$ if it is defined and continuous except possibly at a finite number of points $a \le x_1 < x_2 < \ldots < x_n \le b$. At each point of discontinuity, the left and right hand limits[†]

$$f(x_k^-) = \lim_{x \to x_k^-} f(x), \qquad\qquad f(x_k^+) = \lim_{x \to x_k^+} f(x),$$

exist. Note that we do not require that $f(x)$ be defined at x_k. Even if $f(x_k)$ is defined, it does not necessarily equal either the left or the right hand limit.

[†] At the endpoints a, b we only require one of the limits, namely $f(a^+)$ and $f(b^-)$, to exist.

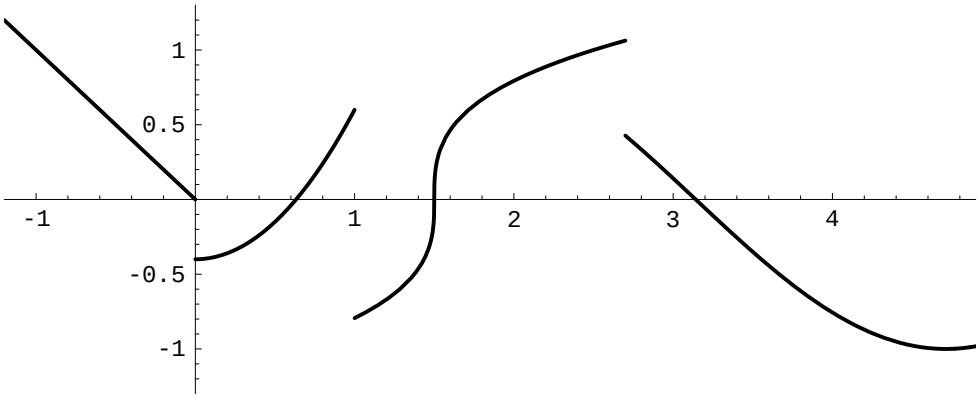


Figure 12.2. Piecewise Continuous Function.

A function $f(x)$ defined for all $x \in \mathbb{R}$ is piecewise continuous provided it is piecewise continuous on every bounded interval. In particular, a 2π periodic function $\widetilde{f}(x)$ is piecewise continuous if and only if it is piecewise continuous on the interval $[-\pi, \pi]$.

A representative graph of a piecewise continuous function appears in Figure 12.2. The points x_k are known as *jump discontinuities* of $f(x)$ and the difference

$$\beta_k = f(x_k^+) - f(x_k^-) = \lim_{x \to x_k^+} f(x) - \lim_{x \to x_k^-} f(x) \tag{12.29}$$

between the left and right hand limits is the *magnitude* of the jump, cf. (11.46). If $\beta_k = 0$, and so the right and left hand limits agree, then the discontinuity is *removable* since redefining $f(x_k) = f(x_k^+) = f(x_k^-)$ makes f continuous at x_k. We will assume, without significant loss of generality, that our functions have no removable discontinuities.

The simplest example of a piecewise continuous function is the step function

$$\sigma(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0. \end{cases} \tag{12.30}$$

It has a single jump discontinuity at $x = 0$ of magnitude 1, and is continuous — indeed, constant — everywhere else. If we translate and scale the step function, we obtain a function

$$h(x) = \beta\, \sigma(x - y) = \begin{cases} \beta, & x > y, \\ 0, & x < y, \end{cases} \tag{12.31}$$

with a single jump discontinuity of magnitude β at the point $x = y$.

If $f(x)$ is any piecewise continuous function, then its Fourier coefficients are well-defined — the integrals (12.21) exist and are finite. Continuity, however, is not enough to ensure convergence of the resulting Fourier series.

Definition 12.6. A function $f(x)$ is called *piecewise* C^1 on an interval $[a, b]$ if it is defined, continuous and continuously differentiable except possibly at a finite number of points $a \leq x_1 < x_2 < \ldots < x_n \leq b$. At each exceptional point, the left and right hand

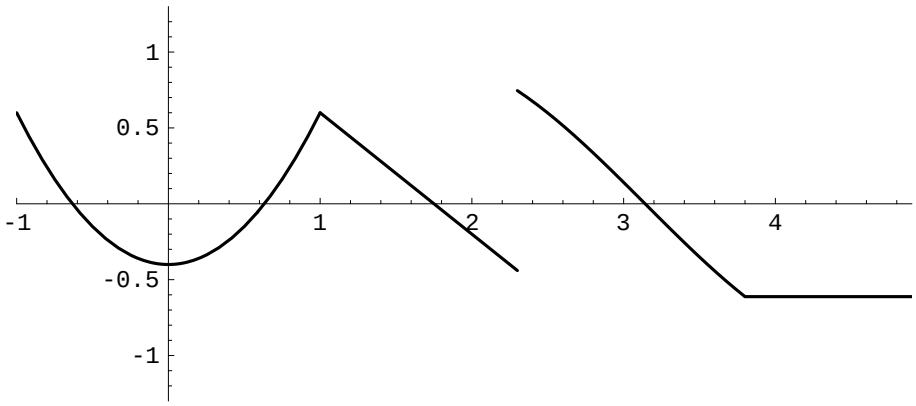


Figure 12.3. Piecewise C^1 Function.

limits† exist:

$$f(x_k^-) = \lim_{x \to x_k^-} f(x), \qquad f(x_k^+) = \lim_{x \to x_k^+} f(x),$$
$$f'(x_k^-) = \lim_{x \to x_k^-} f'(x), \qquad f'(x_k^+) = \lim_{x \to x_k^+} f'(x).$$

See Figure 12.3 for a representative graph. For a piecewise continuous C^1 function, an exceptional point x_k is either

(a) a *jump discontinuity* of f, but where the left and right hand derivatives exist, or

(b) a *corner*, meaning a point where f is continuous, so $f(x_k^-) = f(x_k^+)$, but has different left and right hand derivatives: $f'(x_k^-) \neq f'(x_k^+)$.

Thus, at each point, including jump discontinuities, the graph of $f(x)$ has well-defined right and left tangent lines. For example, the function $f(x) = |x|$ is piecewise C^1 since it is continuous everywhere and has a corner at $x = 0$, with $f'(0^+) = +1$, $f'(0^-) = -1$.

There is an analogous definition of a piecewise C^n function. One requires that the function has n continuous derivatives, except at a finite number of points. Moreover, at every point, the function has well-defined right and left hand limits of all its derivatives up to order n.

The Convergence Theorem

The fundamental convergence theorem for Fourier series can now be stated.

Theorem 12.7. *If $\widetilde{f}(x)$ is any 2π periodic, piecewise C^1 function, then its Fourier series converges for all x to*

$$\begin{cases} \widetilde{f}(x) & \text{if } \widetilde{f} \text{ is continuous at } x, \\ \frac{1}{2}\big[\, \widetilde{f}(x^+) + \widetilde{f}(x^-) \,\big] & \text{if } x \text{ is a jump discontinuity.} \end{cases}$$

† As before, at the endpoints we only require the appropriate one-sided limits, namely $f(a^+)$, $f'(a^+)$ and $f(b^-), f'(b^-)$, to exist.

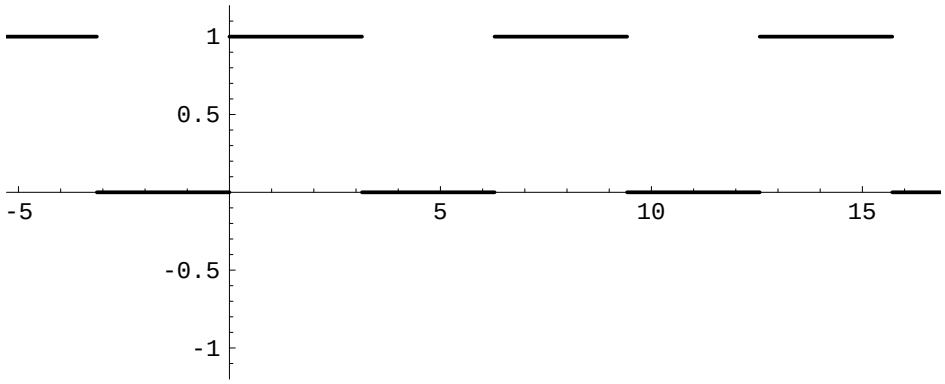


Figure 12.4. Periodic Step Function.

Thus, at discontinuities, the Fourier series "splits the difference" and converges to the average of the right and left hand limits. If we redefine

$$\widetilde{f}(x) = \tfrac{1}{2}\big[\,\widetilde{f}(x^+) + \widetilde{f}(x^-)\,\big] \tag{12.32}$$

to have such a value at its jump discontinuities — an equation that automatically holds at all points of continuity — then Theorem 12.7 would say that the Fourier series converges to $\widetilde{f}(x)$ everywhere. We will discuss the ideas underlyilng the proof of the Convergence Theorem 12.7 at the end of Section 12.5.

Example 12.8. Let $\sigma(x)$ denote the step function (12.30). Its Fourier coefficients are easily computed:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} \sigma(x)\,dx = \frac{1}{\pi}\int_0^{\pi}\, dx = 1,$$

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \sigma(x)\cos k\,x\,dx = \frac{1}{\pi}\int_0^{\pi} \cos k\,x\,dx = 0,$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \sigma(x)\sin k\,x\,dx = \frac{1}{\pi}\int_0^{\pi} \sin k\,x\,dx = \begin{cases} \dfrac{2}{k\,\pi}, & k = 2\,l+1 \text{ odd},\\ 0, & k = 2\,l \text{ even}. \end{cases}$$

Therefore, the Fourier series for the step function is

$$\sigma(x) \;\sim\; \frac{1}{2} \;+\; \frac{2}{\pi}\left(\sin x \;+\; \frac{\sin 3\,x}{3} \;+\; \frac{\sin 5\,x}{5} \;+\; \frac{\sin 7\,x}{7} \;+\; \cdots\right). \tag{12.33}$$

According to Theorem 12.7, the Fourier series will converge to the $2\,\pi$ periodic extension of the step function, as plotted in Figure 12.4:

$$\widetilde{\sigma}(x) = \begin{cases} 0, & (2\,m+1)\,\pi < x < 2\,m\,\pi,\\ 1, & 2\,m\,\pi < x < (2\,m+1)\,\pi,\\ \frac{1}{2}, & x = m\,\pi, \end{cases} \qquad \text{where } m \text{ denotes an arbitrary integer.}$$

In accordance with Theorem 12.7, $\widetilde{\sigma}(x)$ takes the midpoint value $\frac{1}{2}$ at the jump discontinuities $0, \pm\pi, \pm 2\,\pi, \ldots$.

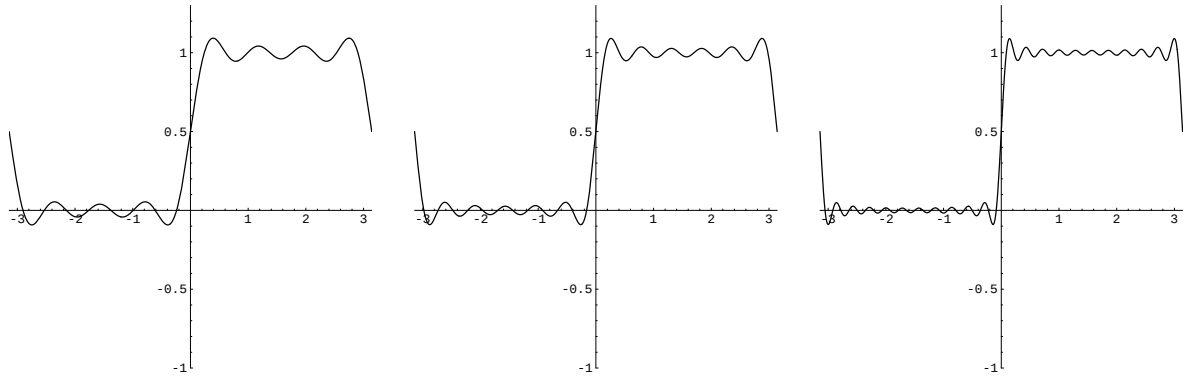


Figure 12.5. Gibbs Phenomenon.

It is instructive to investigate the convergence of this particular Fourier series in some detail. Figure 12.5 displays a graph of the first few partial sums, taking, respectively, $n = 3, 5$, and 10 terms. The reader will notice that away from the discontinuities, the series does appear to be converging, albeit slowly. However, near the jumps there is a consistent overshoot of about 9%. The region where the overshoot occurs becomes narrower and narrower as the number of terms increases, but the magnitude of the overshoot persists no matter how many terms are summed up. This was first noted by the American physicist Josiah Gibbs, and is now known as the *Gibbs phenomenon* in his honor. The Gibbs overshoot is a manifestation of the subtle non-uniform convergence of the Fourier series.

Even and Odd Functions

We already noted that the Fourier cosine coefficients of the function $f(x) = x$ are all 0. This was not an accident, but rather a consequence of the fact that x is an odd function. Recall first the basic definition:

Definition 12.9. A function is called *even* if $f(-x) = f(x)$. A function is *odd* if $f(-x) = -f(x)$.

For example, the functions 1, $\cos k\,x$ and x^2 are all even, whereas $\sin k\,x$ and x are odd. We require two elementary lemmas, whose proofs are left to the reader.

Lemma 12.10. *The sum, $f(x) + g(x)$, of two even functions is even; the sum of two odd functions is odd. The product $f(x)\,g(x)$ of two even functions, or of two odd functions, is an even function. The product of an even and an odd function is odd.*

Remark: *Every* function can be represented as the sum of an even and an odd function; see Exercise ■.

Lemma 12.11. *If $f(x)$ is odd and integrable on the symmetric interval $[-a, a]$, then $\displaystyle\int_{-a}^{a} f(x)\,dx = 0$. If $f(x)$ is even and integrable, then $\displaystyle\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx$.*

The next result is an immediate consequence of applying Lemmas 12.10 and 12.11 to the Fourier integrals (12.21).

Proposition 12.12. *If $f(x)$ is even, then its Fourier sine coefficients all vanish, $b_k = 0$, and so f can be represented by a* Fourier cosine series

$$f(x) \;\sim\; \frac{a_0}{2} \,+\, \sum_{k=1}^{\infty} a_k \cos k\,x\,, \tag{12.34}$$

where

$$a_k = \frac{2}{\pi}\int_0^{\pi} f(x)\cos k\,x\,dx, \qquad\qquad k = 0, 1, 2, 3, \ldots\,. \tag{12.35}$$

If $f(x)$ is odd, then its Fourier cosine coefficients vanish, $a_k = 0$, and so f can be represented by a Fourier sine series

$$f(x) \;\sim\; \sum_{k=1}^{\infty} b_k \sin k\,x\,, \tag{12.36}$$

where

$$b_k = \frac{2}{\pi}\int_0^{\pi} f(x)\sin k\,x\,dx, \qquad\qquad k = 1, 2, 3, \ldots\,. \tag{12.37}$$

Conversely, a convergent Fourier cosine (respectively, sine) series always represents an even (respectively, odd) function.

Example 12.13. The absolute value $f(x) = |\,x\,|$ is an even function, and hence has a Fourier cosine series. The coefficients are

$$a_0 = \frac{2}{\pi}\int_0^{\pi} x\,dx = \pi, \tag{12.38}$$

$$a_k = \frac{2}{\pi}\int_0^{\pi} x\cos k\,x\,dx = \frac{2}{\pi}\left[\,\frac{x\,\sin k\,x}{k} + \frac{\cos k\,x}{k^2}\,\right]_{x=0}^{\pi} = \begin{cases} 0, & 0 \neq k \text{ even}, \\ -\,\dfrac{4}{k^2\,\pi}\,, & k \text{ odd}. \end{cases}$$

Therefore

$$|\,x\,| \;\sim\; \frac{\pi}{2} \,-\, \frac{4}{\pi}\left(\cos x \,+\, \frac{\cos 3\,x}{9} \,+\, \frac{\cos 5\,x}{25} \,+\, \frac{\cos 7\,x}{49} \,+\, \cdots\right). \tag{12.39}$$

According to Theorem 12.7, this Fourier cosine series converges to the $2\,\pi$ periodic extension of $|\,x\,|$, which is graphed in Figure 12.6.

In particular, if we substitute $x = 0$, we obtain another interesting series

$$\frac{\pi^2}{8} = 1 \,+\, \frac{1}{9} \,+\, \frac{1}{25} \,+\, \frac{1}{49} \,+\, \cdots \,=\, \sum_{n=0}^{\infty} \frac{1}{(2\,n+1)^2}\,. \tag{12.40}$$

It converges faster than Gregory's series (12.28), and, while not optimal, can be used to compute reasonable approximations to π. One can further manipulate this result to compute the sum of the series

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2} = 1 \,+\, \frac{1}{4} \,+\, \frac{1}{9} \,+\, \frac{1}{16} \,+\, \frac{1}{25} \,+\, \frac{1}{36} \,+\, \frac{1}{49} \,+\, \cdots\,.$$

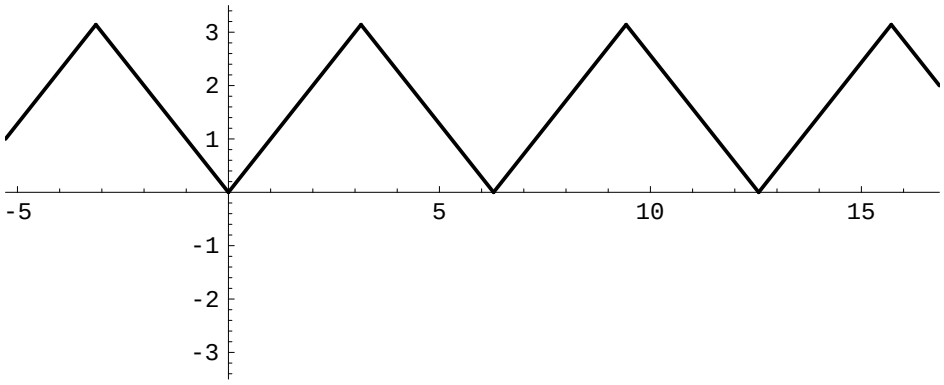


Figure 12.6. Periodic extension of $|\,x\,|$.

We note that

$$\frac{S}{4} = \sum_{n=1}^{\infty} \frac{1}{4n^2} = \sum_{n=1}^{\infty} \frac{1}{(2n)^2} = \frac{1}{4} + \frac{1}{16} + \frac{1}{36} + \frac{1}{64} + \cdots .$$

Therefore, by (12.40),

$$\frac{3}{4}\,S = S - \frac{S}{4} = 1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \cdots = \frac{\pi^2}{8}\,,$$

from which we conclude that

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots = \frac{\pi^2}{6}\,. \tag{12.41}$$

Remark: If $\widetilde{f}(x)$ is either even or odd and 2π periodic, then it is uniquely determined by its values on the interval $[\,0, \pi\,]$. This is proved by a straightforward adaptation of the proof of Lemma 12.3; see Exercise ■.

Complex Fourier Series

An alternative, and often more convenient, approach to Fourier series is to use complex exponentials instead of sines and cosines. Indeed, Euler's formula

$$e^{\,\mathrm{i}\,k\,x} = \cos k\,x + \mathrm{i}\,\sin k\,x, \qquad e^{-\,\mathrm{i}\,k\,x} = \cos k\,x - \mathrm{i}\,\sin k\,x, \tag{12.42}$$

shows how to write the trigonometric functions

$$\cos k\,x = \frac{e^{\,\mathrm{i}\,k\,x} + e^{-\,\mathrm{i}\,k\,x}}{2}\,, \qquad \sin k\,x = \frac{e^{\,\mathrm{i}\,k\,x} - e^{-\,\mathrm{i}\,k\,x}}{2\,\mathrm{i}}\,, \tag{12.43}$$

in terms of complex exponentials. Orthonormality with respect to the rescaled L^2 Hermitian inner product

$$\langle\, f\,;g\,\rangle = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(x)\,\overline{g(x)}\,dx\,, \tag{12.44}$$

was proved by direct computation in Example 3.45:

$$\begin{aligned} \langle\, e^{\,\mathrm{i}\,k\,x}\,;e^{\,\mathrm{i}\,l\,x}\,\rangle &= \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} e^{\,\mathrm{i}\,(k-l)x}\,dx = \begin{cases} 1, & k=l,\\ 0, & k\neq l, \end{cases}\\ \|\,e^{\,\mathrm{i}\,k\,x}\,\|^2 &= \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} |\,e^{\,\mathrm{i}\,k\,x}\,|^2\,dx = 1. \end{aligned} \tag{12.45}$$

The *complex Fourier series* for a (piecewise continuous) real or complex function f is

$$f(x)\;\sim\;\sum_{k=-\infty}^{\infty} c_k\,e^{\,\mathrm{i}\,k\,x} = \;\cdots + c_{-2}\,e^{-2\,\mathrm{i}\,x} + c_{-1}\,e^{-\,\mathrm{i}\,x} + c_0 + c_1\,e^{\,\mathrm{i}\,x} + c_2\,e^{2\,\mathrm{i}\,x} + \;\cdots\;. \tag{12.46}$$

The orthonormality formulae (12.44) imply that the *complex Fourier coefficients* are obtained by taking the inner products

$$c_k = \langle\, f\,;e^{\,\mathrm{i}\,k\,x}\,\rangle = \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} f(x)\,e^{-\,\mathrm{i}\,k\,x}\,dx \tag{12.47}$$

with the associated complex exponential. Pay attention to the minus sign in the integrated exponential — the result of taking the complex conjugate of the second argument in the inner product (12.44). It should be emphasizedthat the real (12.18) and complex (12.46) Fourier formulae are just two different ways of writing the *same* series! Indeed, if we apply Euler's formula (12.42) to (12.47) and compare with the real Fourier formulae (12.21), we find that the real and complex Fourier coefficients are related by

$$\begin{aligned} a_k &= c_k + c_{-k}, & c_k &= \tfrac{1}{2}(a_k - \mathrm{i}\,b_k),\\ b_k &= \mathrm{i}\,(c_k - c_{-k}), & c_{-k} &= \tfrac{1}{2}(a_k + \mathrm{i}\,b_k), \end{aligned} \qquad k = 0, 1, 2, \ldots\;. \tag{12.48}$$

Remark: We already see one advantage of the complex version. The constant function $1 = e^{0\,\mathrm{i}\,x}$ no longer plays an anomalous role — the annoying factor of $\frac{1}{2}$ in the real Fourier series (12.18) has mysteriously disappeared!

Example 12.14. For the step function $\sigma(x)$ considered in Example 12.8, the complex Fourier coefficients are

$$c_k = \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} \sigma(x)\,e^{-\,\mathrm{i}\,k\,x}\,dx = \frac{1}{2\,\pi}\,\int_{0}^{\pi} e^{-\,\mathrm{i}\,k\,x}\,dx = \begin{cases} \frac{1}{2}, & k=0,\\ 0, & 0\neq k \text{ even},\\ \dfrac{1}{\mathrm{i}\,k\,\pi}, & k \text{ odd}. \end{cases}$$

Therefore, the step function has the complex Fourier series

$$\sigma(x) \;\;\sim\;\; \frac{1}{2} \;-\; \frac{\mathrm{i}}{\pi}\,\sum_{l=-\infty}^{\infty} \frac{e^{(2\,l+1)\,\mathrm{i}\,x}}{2\,l+1}\;.$$

You should convince yourself that this is *exactly the same series* as the real Fourier series (12.33). We are merely rewriting it using complex exponentials instead of real sines and cosines.

Example 12.15. Let us find the Fourier series for the exponential function $e^{a\,x}$. It is much easier to evaluate the integral for the complex Fourier coefficients, and so

$$c_k = \langle\, e^{a\,x}\,;e^{\mathrm{i}\,k\,x}\,\rangle = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} e^{(a-\mathrm{i}\,k)\,x}\,dx = \frac{e^{(a-\mathrm{i}\,k)\,x}}{2\,\pi\,(a-\mathrm{i}\,k)}\,\Big|_{x=-\pi}^{\pi}$$
$$= \frac{e^{(a-\mathrm{i}\,k)\,\pi} - e^{-(a-\mathrm{i}\,k)\,\pi}}{2\,\pi\,(a-\mathrm{i}\,k)} = (-1)^k\,\frac{e^{a\,\pi}-e^{-a\,\pi}}{2\,\pi\,(a-\mathrm{i}\,k)} = \frac{(-1)^k(a+\mathrm{i}\,k)\,\sinh a\,\pi}{\pi\,(a^2+k^2)}\,.$$

Therefore, the desired Fourier series is

$$e^{a\,x} \;\sim\; \frac{\sinh a\,\pi}{\pi}\sum_{k=-\infty}^{\infty}\frac{(-1)^k(a+\mathrm{i}\,k)}{a^2+k^2}\,e^{\mathrm{i}\,k\,x}. \tag{12.49}$$

As an exercise, the reader should try writing this as a real Fourier series, either by breaking up the complex series into its real and imaginary parts, or by direct evaluation of the real coefficients via their integral formulae (12.21).

The Delta Function

Fourier series can even be used to represent more general objects than mere functions. The most important example is the delta function $\delta(x)$. Using its characterizing properties (11.37), the real Fourier coefficients are computed as

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi}\delta(x)\cos k\,x\,dx = \frac{1}{\pi}\cos k\,0 = \frac{1}{\pi}\,,\\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi}\delta(x)\sin k\,x\,dx = \frac{1}{\pi}\sin k\,0 = 0. \end{aligned} \tag{12.50}$$

Therefore,

$$\delta(x) \;\sim\; \frac{1}{2\,\pi} + \frac{1}{\pi}\,\big(\cos x + \cos 2\,x + \cos 3\,x + \,\cdots\,\big). \tag{12.51}$$

Since $\delta(x)$ is an even function, it should come as no surprise that it has a cosine series.

To understand in what sense this series converges to the delta function, it will help to rewrite it in complex form

$$\delta(x) \;\sim\; \frac{1}{2\,\pi}\sum_{k=-\infty}^{\infty} e^{\mathrm{i}\,k\,x} = \frac{1}{2\,\pi}\,\big(\,\cdots + e^{-2\,\mathrm{i}\,x} + e^{-\mathrm{i}\,x} + 1 + e^{\mathrm{i}\,x} + e^{2\,\mathrm{i}\,x} + \,\cdots\,\big). \tag{12.52}$$

where the complex Fourier coefficients are computed[†] as

$$c_k = \frac{1}{2\,\pi}\int_{-\pi}^{\pi}\delta(x)\,e^{-\mathrm{i}\,k\,x}\,dx = \frac{1}{2\,\pi}\,.$$

[†] Or, we could use (12.48).

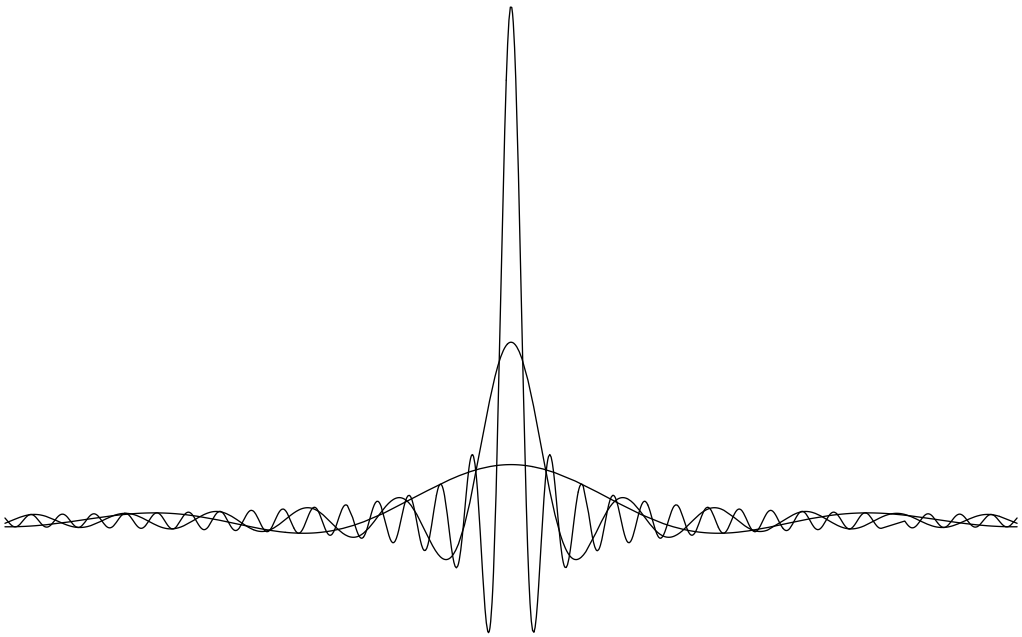


Figure 12.7. Partial Fourier Sums Approximating the Delta Function.

The n^{th} partial sum $s_n(x) = \dfrac{1}{2\pi} \displaystyle\sum_{k=-n}^{n} e^{\,\mathrm{i}\,k\,x}$ can, in fact be explicitly computed. It has the form of an elementary geometric series

$$\sum_{k=0}^{m} a\,r^k = a + a\,r + a\,r^2 + \,\cdots\, + a\,r^m = a\,\left(\frac{r^{m+1}-1}{r-1}\right), \tag{12.53}$$

in which the initial term is $a = e^{-\,\mathrm{i}\,n x}$, the ratio is $r = e^{\,\mathrm{i}\,x}$, while $m = 2n$ indicates the number of terms. We conclude that

$$\begin{aligned} s_n(x) = \frac{1}{2\pi} \sum_{k=-n}^{n} e^{\,\mathrm{i}\,k\,x} &= \frac{1}{2\pi}\, e^{-\,\mathrm{i}\,n x}\left(\frac{e^{\,\mathrm{i}\,(2n+1)x}-1}{e^{\,\mathrm{i}\,x}-1}\right) = \frac{1}{2\pi}\,\frac{e^{\,\mathrm{i}\,(n+1)x}-e^{-\,\mathrm{i}\,n x}}{e^{\,\mathrm{i}\,x}-1} \\ &= \frac{1}{2\pi}\,\frac{e^{\,\mathrm{i}\,\left(n+\frac{1}{2}\right)x}-e^{-\,\mathrm{i}\,\left(n+\frac{1}{2}\right)x}}{e^{\,\mathrm{i}\,x/2}-e^{-\,\mathrm{i}\,x/2}} = \frac{1}{2\pi}\,\frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}\,. \end{aligned} \tag{12.54}$$

To go from the first to the secojnd line, we multiplied numerator and denominator by $e^{-\,\mathrm{i}\,x/2}$, after which we used the formula (3.78) for the sine function in terms of complex exponentials. Incidentally, (12.54) is the same as the intriguing trigonometric summation formula

$$s_n(x) = \frac{1}{2\pi} + \frac{1}{\pi}\,\left(\cos x + \cos 2\,x + \cos 3\,x + \,\cdots\, + \cos n\,x\,\right) = \frac{1}{2\pi}\,\frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}\,. \tag{12.55}$$

Graphs of the partial sums $s_n(x)$ for several values of n are displayed in Figure 12.7. Note that the spike, at $x = 0$, progressively becomes taller and thinner, converging to an

infinitely tall, infinitely thin delta spike. Indeed, by l'Hôpital's Rule,

$$\lim_{x\to 0}\ \frac{1}{2\,\pi}\,\frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}=\lim_{x\to 0}\ \frac{1}{2\,\pi}\,\frac{\left(n+\frac{1}{2}\right)\cos\left(n+\frac{1}{2}\right)x}{\frac{1}{2}\cos\frac{1}{2}\,x}=\frac{n+\frac{1}{2}}{\pi}\ \longrightarrow\ \infty\quad\text{as}\quad n\to\infty.$$

(An elementary proof of this fact is to note that, at $x=0$, every term in the original sum (12.52) is equal to 1.) Furthermore, the integrals remain fixed

$$\frac{1}{2\,\pi}\int_{-\pi}^{\pi}s_n(x)\,dx=\frac{1}{2\,\pi}\int_{-\pi}^{\pi}\frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}\,dx=\frac{1}{2\,\pi}\int_{-\pi}^{\pi}\sum_{k=-n}^{n}e^{\,\mathrm{i}\,k\,x}\,dx=1,\qquad(12.56)$$

as required for convergence to the delta function. However, away from the spike, the partial sums do *not* go to zero. Rather, they oscillates more and more rapidly, maintaining an overall amplitude of $\csc\frac{1}{2}\,x=1/\sin\frac{1}{2}\,x$. As n gets large, the amplitude function appears as an envelope of the increasingly rapid oscillations. Roughly speaking, the fact that $s_n(x)\to\delta(x)$ as $n\to\infty$ means that the "infinitely fast" oscillations somehow cancel each other out, and the net effect is zero away from the spike at $x=0$. Thus, the convergence of the Fourier sums to $\delta(x)$ is much more subtle than in the original limiting definition (11.32). The technical term is "weak convergence", and plays an very important role in advanced mathematical analysis, [**126**].

Remark: Although we stated that the Fourier series (12.51), (12.52) represent the delta function, this is not entirely correct. Remember that a Fourier series converges to the $2\,\pi$ periodic extension of original the function. Therefore, (12.52) actually represents the periodic extension of the delta function

$$\widetilde{\delta}(x)=\ \cdots+\delta(x+4\pi)+\delta(x+2\,\pi)+\delta(x)+\delta(x-2\,\pi)+\delta(x-4\pi)+\delta(x-6\pi)+\cdots\ ,\quad(12.57)$$

consisting of a periodic array of delta spikes concentrated at all integer multiples of $2\,\pi$.

12.3. Differentiation and Integration.

If a series of functions converges — at least in a sufficiently regular manner — then one expects to be able to integrate and differentiate it term by term; the resulting series should converge to the integral and derivative of the original sum. For power series, the implementation of this idea is straightforward, and used extensively in the construction of series solutions of differential equations, series for integrals of non-elementary functions, and so on. Appendix C develops some of the details.

As we now appreciate, the convergence of Fourier series is a much more delicate matter, and so one must take considerably more care in the application of term-wise differentiation and integration. Nevertheless, in favorable situations, both operations lead to valid results, and provide a powerful means of constructing Fourier series of more complicated functions. It is a remarkable, profound fact that Fourier analysis is completely compatible with the calculus of generalized functions that we developed in Chapter 11. In particular, differentiating the Fourier series for a suitably nice function with a jump discontinuity leads to the Fourier series for the differentiated function, with a delta function of the appropriate magnitude appearing at the discontinuity. This fact reassures us that the rather

mysterious construction of delta functions and their generalizations is indeed the right way to extend calculus to functions which do not possess derivatives in the ordinary sense.

Integration of Fourier Series

Integration is a smoothing operation — the integrated function is always nicer than the original. Therefore, we should anticipate being able to integrate Fourier series without difficulty. However, there is one complication: the integral of a periodic function is not necessarily periodic. The simplest example is the constant function 1, which is certainly periodic, but its integral, namely x, is not. On the other hand, integrals of all the other periodic sine and cosine functions appearing in the Fourier series are periodic. Thus, only the constant term might cause us difficulty when we try to integrate a Fourier series (12.18). According to (2.4), the constant term

$$\frac{a_0}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, dx \tag{12.58}$$

is the *mean* or *average* of the function $f(x)$ on the interval $[-\pi, \pi]$. A function has no constant term in its Fourier series if and only if it has zero mean. It is easily shown, cf. Exercise ■, that the mean zero functions are precisely the ones that remain periodic upon integration.

Lemma 12.16. *If $f(x)$ is 2π periodic, then its integral $g(x) = \displaystyle\int_0^x f(y)\, dy$ is 2π periodic if and only if f has mean zero on the interval $[-\pi, \pi]$.*

In particular, Lemma 12.11 implies that all odd functions automatically have mean zero.

Theorem 12.17. *If f is piecewise continuous, 2π periodic, and has mean zero, then its Fourier series can be integrated term by term, to produce the Fourier series*

$$g(x) = \int_0^x f(y)\, dy \quad \sim \quad m \; + \; \sum_{k=1}^{\infty} \left[-\frac{b_k}{k} \cos k\, x + \frac{a_k}{k} \sin k\, x \right], \tag{12.59}$$

for its integral. The constant term

$$m = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x)\, dx$$

is the mean of the integrated function.

In many situations, the integration formula (12.59) provides a very convenient alternative to the direct derivation of the Fourier coefficients.

Example 12.18. The function $f(x) = x$ is odd, and so has mean zero, $\displaystyle\int_{-\pi}^{\pi} x\, dx = 0$. Let us integrate its Fourier series

$$x \quad \sim \quad 2 \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sin k\, x \tag{12.60}$$

that we found in Example 12.2. The result is the Fourier series

$$\begin{aligned} \frac{1}{2}\,x^2 &\sim \frac{\pi^2}{6} - 2\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2}\cos k\,x \\ &\sim \frac{\pi^2}{6} - 2\left(\cos x - \frac{\cos 2\,x}{4} + \frac{\cos 3\,x}{9} - \frac{\cos 4\,x}{16} + \cdots\right), \end{aligned} \tag{12.61}$$

whose the constant term is the mean of the left hand side:

$$\frac{1}{2\,\pi}\int_{-\pi}^{\pi} \frac{x^2}{2}\,dx = \frac{\pi^2}{6}\,.$$

If we were to integrate each trigonometric summand in a Fourier series (12.18) from 0 to x, we would obtain

$$\int_0^x \cos k\,y\,dy = \frac{\sin k\,x}{k}\,, \qquad \text{while} \qquad \int_0^x \sin k\,y\,dy = \frac{1}{k} - \frac{\cos k\,x}{k}\,.$$

The $1/k$ terms arising from the sine integrals do not appear explicitly in (12.59), and so must be hidden in the constant term m. We deduce that the mean value of the integrated function can be computed using the Fourier sine coefficients of f via the formula

$$\frac{1}{2\,\pi}\int_{-\pi}^{\pi} g(x)\,dx = m = \sum_{k=1}^{\infty} \frac{b_k}{k}. \tag{12.62}$$

For example, the result of integrating both sides of the Fourier series (12.60) from 0 to x is

$$\frac{x^2}{2} \sim 2\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2}\,(1-\cos k\,x).$$

The constant terms sum up to yield the mean value of the integrated function:

$$2\left(1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \cdots\right) = 2\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^2} = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} \frac{x^2}{2}\,dx = \frac{\pi^2}{6}\,, \tag{12.63}$$

which reproduces the formula in Exercise ∎.

If $f(x)$ does not have mean zero, its Fourier series has a nonzero constant term,

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[\, a_k\,\cos k\,x + b_k\,\sin k\,x\,\right].$$

In this case, the result of integration will be

$$g(x) = \int_0^x f(y)\,dy \sim \frac{a_0}{2}\,x + m + \sum_{k=1}^{\infty} \left[-\frac{b_k}{k}\,\cos k\,x + \frac{a_k}{k}\,\sin k\,x\right], \tag{12.64}$$

where m is given in (12.62). The right hand side is not, strictly speaking, a Fourier series. There are two ways to interpret this formula within the Fourier framework. Either we can write (12.64) as the Fourier series for the difference

$$g(x) - \frac{a_0}{2}\, x \quad \sim \quad m + \sum_{k=1}^{\infty} \left[-\frac{b_k}{k} \cos k\, x + \frac{a_k}{k} \sin k\, x \right], \tag{12.65}$$

which is a 2π periodic function, cf. Exercise ■. Alternatively, one can replace x by its Fourier series (12.23), and the result will be the Fourier series for the 2π periodic extension of the integral $g(x) = \displaystyle\int_0^x f(y)\, dy$.

Differentiation of Fourier Series

Differentiation has the opposite effect to integration. Differentiation makes a function worse. Therefore, to justify taking the derivative of a Fourier series, we need to know that the differentiated function remains reasonably nice. Since we need the derivative $f'(x)$ to be piecewise C^1 for the convergence Theorem 12.7 to be applicable, we have to require that $f(x)$ itself be continuous and piecewise C^2.

Theorem 12.19. *If f is 2π periodic, continuous, and piecewise C^2, then its Fourier series can be differentiated term by term, to produce the Fourier series for the derivative*

$$h(x) = f'(x) \quad \sim \quad \sum_{k=1}^{\infty} \left[\, k\, b_k \cos k\, x - k\, a_k \sin k\, x \,\right]. \tag{12.66}$$

Example 12.20. If we differentiate the Fourier series (12.39) for $f(x) = |\, x\,|$, we obtain

$$f'(x) \quad \sim \quad \frac{4}{\pi} \left(\sin x \, + \, \frac{\sin 3\, x}{3} \, + \, \frac{\sin 5\, x}{5} \, + \, \frac{\sin 7\, x}{7} \, + \, \cdots \right). \tag{12.67}$$

The derivative (11.52) of the absolute value function is the sign function

$$\frac{d\,|\, x\,|}{dx} = \operatorname{sign} x = \begin{cases} \, +1, & x > 0 \\ \, -1, & x < 0. \end{cases}$$

Note that $\operatorname{sign} x = \sigma(x) - \sigma(-x)$ is the difference of two step functions. Indeed, subtracting the step function Fourier series (12.33) at x from the same series at $-x$ reproduces (12.67).

Example 12.21. If we differentiate the Fourier series

$$x \quad \sim \quad 2 \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sin k\, x \; = \; 2 \left(\sin x \, - \, \frac{\sin 2\, x}{2} \, + \, \frac{\sin 3\, x}{3} \, - \, \frac{\sin 4\, x}{4} \, + \, \cdots \right).$$

for x, we obtain an apparent contradiction:

$$1 \quad \sim \quad 2 \sum_{k=1}^{\infty} (-1)^{k+1} \cos k\, x = 2 - 2 \cos x + 2 \cos 2\, x - 2 \cos 3\, x + \, \cdots \, . \tag{12.68}$$

But the Fourier series for 1 just consists of a single constant term! (Why?)

The resolution of this difficulty is not hard. The Fourier series (12.23) does *not* converge to x, but rather to its periodic extension $\widetilde{f}(x)$, which has a jump discontinuity of magnitude 2π at odd multiples of π. Thus, Theorem 12.19 is not directly applicable. Nevertheless, we can assign a consistent interpretation to the differentiated series. As discussed in Section 11.2, the derivative $\widetilde{f}\,'(x)$ of the periodic extension is *not* equal to the constant function 1, but, rather, has an additional delta function concentrated at each jump discontinuity:

$$\widetilde{f}\,'(x) = 1 - 2\pi \sum_{j=-\infty}^{\infty} \delta\big(x - (2j+1)\pi\big) = 1 - 2\pi\,\widetilde{\delta}(x-\pi),$$

where $\widetilde{\delta}$ denotes the 2π periodic extension of the delta function, cf. (12.57). The differentiated Fourier series (12.68) does, in fact, converge to this modified distributional derivative!

12.4. Change of Scale.

So far, we have only dealt with Fourier series on the standard interval of length 2π. (We chose $[-\pi, \pi]$, but the statements and formulas are easily adapted to any other interval of the same length, e.g., $[0, 2\pi]$.) Since physical objects like bars and strings do not all come in this particular length, we need to understand how to adapt the formulas to more general intervals. The basic idea is to rescale the variable so as to stretch or contract the standard interval, and was already used, in Section 5.4, to adapt the orthogonal Legendre polynomials to other intervals.

Any symmetric interval $[-\ell\,, \ell\,]$ of length 2ℓ can be rescaled to the standard interval $[-\pi, \pi]$ by using the linear change of variables

$$x = \frac{\ell}{\pi}\, y, \qquad \text{so that} \qquad -\pi \le y \le \pi \quad \text{whenever} \quad -\ell \le x \le \ell. \tag{12.69}$$

Given a function $f(x)$ defined on $[-\ell\,, \ell\,]$, the *rescaled function* $F(y) = f\left(\dfrac{\ell}{\pi}\, y\right)$ lives on $[-\pi, \pi]$. Let

$$F(y) \;\sim\; \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} \big[\, a_k \cos k\,y + b_k \sin k\,y \,\big],$$

be the standard Fourier series for $F(y)$, so that

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} F(y)\cos k\,y\,dy, \qquad\qquad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} F(y)\sin k\,y\,dy. \tag{12.70}$$

Then, reverting to the unscaled variable x, we deduce that

$$f(x) \;\sim\; \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} \left[\, a_k \cos\frac{k\pi\,x}{\ell} + b_k \sin\frac{k\pi\,x}{\ell} \,\right]. \tag{12.71}$$

The Fourier coefficients can be computed directly. Indeed, replacing the integration variable by $y = \pi x/\ell$, and noting that $dy = (\pi/\ell)\, dx$, we deduce the modified formulae

$$a_k = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\, \cos\frac{k\pi x}{\ell}\, dx, \qquad\qquad b_k = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\, \sin\frac{k\pi x}{\ell}\, dx, \tag{12.72}$$

for the Fourier coefficients of $f(x)$ on the interval $[\,-\ell\,,\ell\,]$.

All of the convergence results, integration and differentiation formulae, etc., that are valid for the interval $[\,-\pi,\pi\,]$ carry over, essentially unchanged, to Fourier series on nonstandard intervals. In particular, adapting our basic convergence Theorem 12.7, we conclude that if $f(x)$ is piecewise C^1, then its rescaled Fourier series (12.71) converges to its $2\,\ell$ periodic extension $\widetilde{f}(x)$ with the proviso that $\widetilde{f}(x)$ takes on the midpoint values at all jump discontinuities.

Example 12.22. Let us compute the Fourier series for the function $f(x) = x$ on the interval $-1 \le x \le 1$. Since f is odd, only the sine coefficients will be nonzero. We have

$$b_k = \int_{-1}^{1} x \sin k\pi x\, dx = \left[\, -\frac{x\cos k\pi x}{k\pi} + \frac{\sin k\pi x}{(k\pi)^2}\,\right]_{x=-1}^{1} = \frac{2(-1)^{k+1}}{k\,\pi}\,.$$

The resulting Fourier series is

$$x \;\sim\; \frac{2}{\pi}\left(\sin\pi x \,-\, \frac{\sin 2\pi x}{2} \,+\, \frac{\sin 3\pi x}{3} \,-\, \cdots\right)$$

The series converges to the 2 periodic extension of the function x, namely

$$\widetilde{f}(x) = \begin{cases} x - 2m, & 2m-1 < x < 2m+1, \\ 0, & x = m, \end{cases} \qquad \text{where } m \text{ is an arbitrary integer.}$$

We can similarly reformulate complex Fourier series on the nonstandard interval $[\,-\ell\,,\ell\,]$. Scaling the variables in (12.46) in (12.69), we find

$$f(x) \;\sim\; \sum_{k=-\infty}^{\infty} c_k\, e^{\,\mathrm{i}\,k\pi x/\ell}, \qquad \text{where} \qquad c_k = \frac{1}{2\,\ell}\int_{-\ell}^{\ell} f(x)\, e^{-\,\mathrm{i}\,k\pi x/\ell}\, dx. \tag{12.73}$$

Again, this is merely an alternative way of writing the real Fourier series (12.71).

For a more general interval $[a,b]$ there are two options. Either one can take a function $f(x)$ defined for $a \le x \le b$ and periodically extend it to a function $\widetilde{f}(x)$ that has period $b-a$. One can then use the Fourier series (12.71) for the symmetric interval $[\,\frac{1}{2}(a-b), \frac{1}{2}(b-a)\,]$ of width $2\,\ell = b-a$. An alternative approach is to translate the interval by an amount $\frac{1}{2}(a+b)$ to make it symmetric; this is done by the change of variables $\widehat{x} = x - \frac{1}{2}(a+b)$. The two methods are essentially equivalent, and details are left to the reader.

12.5. Convergence of the Fourier Series.

In this final section, we establish the basic convergence results for Fourier series. As a by product, we obtain additional information about the nature of Fourier series that plays

an important role in applications, particularly the interplay between the smoothness of the function and the decay of its Fourier coefficients, a result that is exploited in signal and image denoising and in the analytial properties of solutions to partial differential equations. Be forwarned: this material will be the most theoretical that we cover in this text, and the more applied reader may consider omitting it from a first reading. However, while we may have already gained sufficient experience to conduct basic Fourier analysis, a complete understanding of its range and the limitations does requires some familiarity with the underlying theoretical developments. Moreover, the required techniques and proofs serve as an excellent introduction to some of the most important tools of modern mathematical analysis. Rest assured that the effort expended to assimilate this material will be more than amply rewarded in your subsequent career.

Unlike power series, which converge to analytic functions on the interval of convergence, and diverge elsewhere (the only tricky point being whether or not the series converges at the endpoints), the convergence of a Fourier series is a much more subtle matter, and still not understood in complete generality. A large part of the difficulty stems from the intricacies of convergence in infinite-dimensional function spaces. Let us therefore begin with a brief discussion of the most basic issues.

Convergence in Vector Spaces

In a finite-dimensional vector space, e.g., $\mathbb{R}^m$, convergence of sequences, and hence series, is straightforward: there is essentially only one way for a sequence of vectors $\mathbf{v}^{(0)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots \in \mathbb{R}^m$ to converge, which is guaranteed by any of the following equivalent criteria:

(a) The vectors converge: $\mathbf{v}^{(n)} \longrightarrow \mathbf{v}^\star \in \mathbb{R}^m$ as $n \to \infty$.

(b) All components of $\mathbf{v}^{(n)} = (v_1^{(n)}, \ldots, v_m^{(n)})$ converge, so $v_i^{(n)} \longrightarrow v_i^\star$, for $i = 1, \ldots, m$.

(c) The difference in norms goes to zero: $\| \mathbf{v}^{(n)} - \mathbf{v}^\star \| \longrightarrow 0$ as $n \to \infty$.

The last requirement, known as *convergence in norm*, does not, in fact, depend on which norm is chosen. Indeed, Theorem 3.19 implies that, on a finite-dimensional vector space, all norms are essentially equivalent, and if one norm goes to zero, so does any other norm.

The corresponding convergence criteria are certainly *not* the same on infinite-dimensional vector spaces. There are, in fact, a bewildering variety of diverse convergence mechanisms in function space, including pointwise convergence, uniform convergence, convergence in norm, weak convergence, and so on. All play a significant role in advanced mathematical analysis. For our applications, we shall be content to study just the most basic aspects of convergence of the Fourier series. Much more detail is available in more advanced texts, e.g., [**51**, **159**].

The most basic convergence mechanism for a sequence of functions $v_n(x)$ is called *pointwise convergence*, where we require

$$\lim_{n \to \infty} v_n(x) = v_\star(x) \qquad \text{for all} \quad x. \tag{12.74}$$

In other words, the functions' values at each individual point converge in the usual sense. Pointwise convergence is the function space version of the convergence of the components of

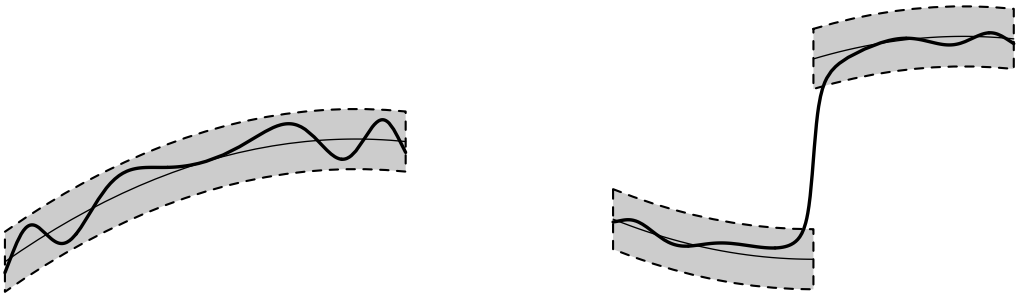


Figure 12.8. Uniform and Non-Uniform Convergence of Functions.

a vector. Indeed, pointwise convergence immediately implies component-wise convergence of the sample vectors $\mathbf{v}^{(n)} = (\, v_n(x_1), \dots, v_n(x_m)\,)^T \in \mathbb{R}^m$ for any choice of sample points.

On the other hand, *convergence in norm* of the function sequence requires

$$\lim_{n \to \infty} \| \, v_n - v_\star \, \| \;=\; 0,$$

where $\| \cdot \|$ is a prescribed norm on the function space. As mentioned earlier, not all norms on an infinite-dimensional function space are equivalent: a function might be small in one norm, but large in another. As a result, convergence in norm *will* depend upon the choice of norm. Moreover, convergence in norm does not necessarily imply pointwise convergence or vice versa. A variety of examples can be found in the exercises.

Uniform Convergence

Proving uniform convergence of a Fourier series is reasonably straightforward, and so we will begin there. You no doubt first saw the concept of a uniformly convergent sequence of functions in your calculus course, although chances are it didn't leave much of an impression. In Fourier analysis, uniform convergence begins to play an increasingly important role, and is worth studying in earnest. For the record, let us restate the basic definition.

Definition 12.23. A sequence of functions $v_n(x)$ is said to converge *uniformly* to a function $v_\star(x)$ on a subset $I \subset \mathbb{R}$ if, for every $\varepsilon > 0$, there exists an integer $N = N(\varepsilon)$ such that

$$| \, v_n(x) - v_\star(x) \, | < \varepsilon \qquad \text{for all } \; x \in I \; \text{ and all } \; n \geq N. \tag{12.75}$$

The key point — and the reason for the term "uniform convergence" — is that the integer N depends only upon ε and not on the point $x \in I$. Roughly speaking, the sequence converges unifomly if and only if for any small ε, the graphs of the functions eventually lie inside a band of width 2ε centered around the graph of the limiting function; see Figure 12.8. Functions may converge pointwise, but non-uniformly. The Gibbs phenomenon is the prototypical example of a nonuniformly convergent sequence. For a fixen $\varepsilon > 0$, the integer n required for (12.75) to hold depends on the point — the closer x is to the discontinuity, the larger n must be chosen. Thus, for a given ε, there is *no* uniformly valid N that fulfills the requirement (12.75) for all points x. A detailed discussion

of these issues, including the proofs of the basic theorems, can be found in any basic real analysis text, e.g., [**9**, **126**].

A key consequence of uniform convergence is that it preserves continuity.

Theorem 12.24. *If $v_n(x) \to v_\star(x)$ converges uniformly, and each $v_n(x)$ is continuous, then $v_\star(x)$ is also a continuous function.*

Intuitively, as sketched in Figure 12.8, a sufficiently small band around the limiting function would not connect up at a discontinuity, and this prevents the graph of any continuous function, such as $v_n(x)$, from remaining entirely within the band.

Warning: A sequence of continuous functions can converge *non-uniformly* to a continuous function. An example is the sequence $v_n(x) = \dfrac{2nx}{1+n^2x^2}$, which converges pointwise to $v_\star(x) \equiv 0$ (why?) but not uniformly since $\max |\, v_n(x) \,| = v_n\big(\tfrac{1}{n}\big) = 1$.

The convergence (pointwise, uniform, in norm, etc.) of a series $\displaystyle\sum_{k=1}^{\infty} u_k(x)$ is governed by the convergence of its sequence of *partial sums*

$$v_n(x) = \sum_{k=1}^{n} u_k(x). \tag{12.76}$$

The most useful test for uniform convergence of series of functions is known as the *Weierstrass M–test*, due to the highly influential nineteenth century German mathematician Karl Weierstrass, the "father of modern analysis".

Theorem 12.25. *Suppose the functions $u_k(x)$ are bounded by*

$$|\, u_k(x) \,| \le m_k \qquad \text{for all} \qquad x \in I, \tag{12.77}$$

where the $m_k \ge 0$ are fixed positive constants. If the series

$$\sum_{k=1}^{\infty} m_k < \infty \tag{12.78}$$

converges, then the series

$$\sum_{k=1}^{\infty} u_k(x) = f(x) \tag{12.79}$$

converges uniformly to a function $f(x)$ for all $x \in I$. In particular, if the summands $u_k(x)$ in Theorem 12.25 are continuous, so is the sum $f(x)$.

With some care, we are allowed to manipulate uniformly convergent series just like finite sums. Thus, if (12.79) is a uniformly convergent series, so is the term-wise product

$$\sum_{k=1}^{\infty} g(x)\, u_k(x) = g(x)\, f(x) \tag{12.80}$$

with any bounded function: $|\, g(x)\,| \leq C$ for $x \in I$. We can also integrate a uniformly convergent series term by term, and the integrated series

$$\int_a^x \left(\sum_{k=1}^{\infty} u_k(y) \right) dy = \sum_{k=1}^{\infty} \int_a^x u_k(y)\, dy = \int_a^x f(y)\, dy \tag{12.81}$$

is uniformly convergent. Differentiation is also allowed — but only when the differentiated series converges uniformly.

Proposition 12.26. *If* $\displaystyle\sum_{k=1}^{\infty} u_k'(x) = g(x)$ *is a uniformly convergent series, then* $\displaystyle\sum_{k=1}^{\infty} u_k(x) = f(x)$ *is also uniformly convergent, and, moreover,* $f'(x) = g(x)$.

We are particularly interested in applying these results to Fourier series, which, for convenience, we take in complex form

$$f(x) \;\sim \sum_{k=-\infty}^{\infty} c_k\, e^{\,\mathrm{i}\,k\,x}. \tag{12.82}$$

Since x is real,

$$\left|\, e^{\,\mathrm{i}\,k\,x} \,\right| \leq 1,$$

and hence the individual summands are bounded by

$$\left|\, c_k\, e^{\,\mathrm{i}\,k\,x} \,\right| \leq |\, c_k\,| \qquad \text{for all} \qquad x.$$

Applying the Weierstrass M–test, we immediately deduce the basic result on uniform convergence of Fourier series.

Theorem 12.27. *If the Fourier coefficients* c_k *satisfy*

$$\sum_{k=-\infty}^{\infty} |\, c_k\,| < \infty, \tag{12.83}$$

then the Fourier series (12.82) *converges uniformly to a continuous function* $\widetilde{f}(x)$ *whose Fourier coefficients are the same* c_k.

Proof: Uniform convergence and continuity of the limiting function follows from Theorem 12.25. To show that the c_k actually are the Fourier coefficients of the sum, we multiply the Fourier series by $e^{-\,\mathrm{i}\,k\,x}$ and integrate term by term from $-\pi$ to π. As noted in (12.80), (12.81), both operations are valid owing to the uniform convergence of the series. *Q.E.D.*

The one thing that the theorem does not guarantee is that the original function $f(x)$ used to compute the Fourier coefficients c_k is the *same* as the function $\widetilde{f}(x)$ obtained by summing the resulting Fourier series! Indeed, this may very well not be the case. As we

know, the function that the series converges to is necessarily 2π periodic. Thus, at the very least, $\widetilde{f}(x)$ will be the 2π periodic extension of $f(x)$. But even this may not suffice.

Two functions $f(x)$ and $\widehat{f}(x)$ that have the same values except for a finite set of points $x_1, \dots, x_m$ have the same Fourier coefficients. (Why?) More generally, two functions which agree everywhere outside a set of "measure zero" will have the same Fourier coefficients. In this way, a convergent Fourier series singles out a distinguished representative from a collection of essentially equivalent 2π periodic functions.

Remark: The term "measure" refers to a rigorous generalization of the notion of the length of an interval to more general subsets $S \subset \mathbb{R}$. In particular, S has *measure zero* if it can be covered by a collection of intervals of arbitrarily small total length. For example, any collection of finitely many points, or even countably many points, e.g., the rational numbers, has measure zero. The proper development of the notion of measure, and the consequential Lebesgue theory of integration, is properly studied in a course in real analysis, [**125**], and will only be touched upon here.

Fourier series cannot converge uniformly when discontinuities are present. However, it can be proved, [**28**, **51**, **159**], that even when the function fails to be everywhere continuous, its Fourier series is uniformly converges on any closed subset of continuity.

Theorem 12.28. *Let $f(x)$ be 2π periodic and piecewise C^1. If f is continuous for $a < x < b$, then its Fourier series converges uniformly to $f(x)$ on any closed subinterval $a + \delta \le x \le b - \delta$, with $\delta > 0$.*

For example, the Fourier series (12.33) for the step function does converge uniformly if we stay away from the discontinuities; for instance, by restriction to a subinterval of the form $[\,\delta, \pi - \delta\,]$ or $[\,-\pi + \delta, -\delta\,]$ for any $0 < \delta < \frac{1}{2}\pi$. This reconfirms our observation that the nonuniform Gibbs behavior becomes progressively more and more localized at the discontinuities.

Smoothness and Decay

The uniform convergence criterion (12.83) requires, at the very least, that the Fourier coefficients decay to zero: $c_k \to 0$ as $k \to \pm\infty$. In fact, the Fourier coefficients cannot tend to zero too slowly. For example, the individual summands of the infinite series

$$\sum_{k=-\infty}^{\infty} \frac{1}{|\,k\,|^{\alpha}} \tag{12.84}$$

go to 0 as $k \to \infty$ for all $\alpha > 0$, but the series converges if and only if $\alpha > 1$. (This follows from the standard integral test for series, [**9**, **126**].) Thus, if we can bound the Fourier coefficients by

$$|\,c_k\,| \;\le\; \frac{M}{|\,k\,|^{\alpha}} \qquad \text{for all} \qquad |\,k\,| \gg 0, \tag{12.85}$$

for some power $\alpha > 1$ and some positive constant $M > 0$, then the Weierstrass M test will guarantee that the Fourier series converges uniformly to a continuous function.

An important consequence of the differentiation formulae (12.66) for Fourier series is the fact that the faster the Fourier coefficients of a function tend to zero as $k \to \infty$, the smoother the function is. Thus, one can detect the degree of smoothness of a function by looking at how rapidly its Fourier coefficients decay to zero. More rigorously, we have:

Theorem 12.29. *If the Fourier coefficients satisfy*

$$\sum_{k=-\infty}^{\infty} k^n \, |\, c_k \,| \; < \; \infty, \tag{12.86}$$

then the Fourier series (12.46) *converges to a* 2π *periodic function which is* n *times continuously differentiable:* $f(x) \in \mathrm{C}^n$. *Moreover, for any* $m \leq n$, *the* m *times differentiated Fourier series converges uniformly to the corresponding derivative* $f^{(m)}(x)$.

Proof: This is an immediate consequence of Proposition 12.26. Application of the Weierstrass M test to the differentiated Fourier series and use of (12.86) completes the proof. *Q.E.D.*

Corollary 12.30. *If the Fourier coefficients satisfy* (12.85) *for some* $\alpha > n+1$, *then the function* $f(x)$ *is* n *times continuously differentiable.*

Thus, the faster the Fourier coefficients go to zero at large frequency k, the smoother the function is. If the Fourier coefficients go to zero faster than any power of k, e.g., exponentially fast, then the function is infinitely differentiable. Analyticity is a little more delicate, and we refer the reader to [**51**, **159**] for details.

An important consequence of the differentiation formulae is the fact that, the smoother the function is, the faster its Fourier coefficients a_k, b_k decay to zero as $k \to \infty$. For a Fourier series (12.18) to converge to a piecewise continuous function, we must, at the very least, have $a_k \to 0$ and $b_k \to 0$ as $k \to \infty$; see Lemma 12.35 below. If we assume that[†] $f(x)$ is 2π periodic, continuous and piecewise C^2, then Theorem 12.19 implies that the Fourier series for $f'(x)$ converges, and so its Fourier coefficients, namely $k\, b_k$ and $-k\, a_k$, must tend to zero as $k \to \infty$. In general, if f is 2π periodic, has $n-1$ continuous derivatives and n^{th} derivative at least piecewise continuous C^1, then the Fourier coefficients of $f^{(n)}(x)$ must tend to zero, which, by a simple induction, implies that

$$k^n \, a_k, \; k^n \, b_k \quad \longrightarrow \quad 0 \qquad \text{as} \qquad k \longrightarrow \infty.$$

In particular, this requires that the Fourier coefficients of f satisfy

$$|\, a_k \,| < \frac{C}{k^n}\,, \qquad\qquad |\, b_k \,| < \frac{C}{k^n}\,, \tag{12.87}$$

for some constant $C > 0$. If f is infinitely differentiable, or, even more restrictively, analytic, then its Fourier coefficients go to zero faster than *any* power of k. For instance, if $|\, a_k \,|, |\, b_k \,| < C\, e^{-k}$, then the Fourier sum is a C^∞ function. Thus, one can detect the degree of smoothness of a function by looking at how rapidly its Fourier coefficients decay to zero. See Theorem 12.29 below for a more precise result.

[†] If the function is not periodic, one must impose the assumptions on its periodic extension for the remarks to be valid.

Example 12.31. The 2π periodic extension of the function $|\,x\,|$ is continuous with piecewise continuous first derivative. Its Fourier coefficients (12.38) satisfy the estimate (12.87) for $n = 2$, which is in accord with the previous remarks. On the other hand, the Fourier coefficients (12.22) of the step function $\sigma(x)$ only tend to zero as $1/k$, reflecting the fact that its periodic extension is only piecewise continuous. Finally, the Fourier coefficients (12.50) for the delta function do not tend to zero at all, indicative of the fact that it is not an ordinary function, and its Fourier series does not converge in the standard sense.

Hilbert Space

In order to make further progress, we must take a little detour. The proper setting for the rigorous theory of Fourier series turns out to be the most important function space in modern physics and modern analysis, known as *Hilbert space*. The precise definition of this infinite-dimensional inner product space is rather technical, but a rough version goes as follows:

Definition 12.32. A complex-valued function $f(x)$ is called *square-integrable* on the interval $[\,-\pi, \pi\,]$ if it satisfies

$$\| f \|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\, f(x) \,|^2 \, dx < \infty. \tag{12.88}$$

The *Hilbert space* $\mathrm{L}^2 = \mathrm{L}^2[\,-\pi, \pi\,]$ is the vector space consisting of all complex-valued *square-integrable* functions on $[\,-\pi, \pi\,]$.

Note that (12.88) is the L^2 norm based on the Hermitian inner product

$$\langle\, f \,;\, g \,\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, \overline{g(x)}\, dx. \tag{12.89}$$

The triangle inequality (3.19), namely

$$\| \,c\, f + d\, g \,\| \;\le\; |\, c \,|\, \| \, f \, \| + |\, d \,|\, \| \, g \, \|,$$

implies that the Hilbert space is a complex vector space, i.e., if $f, g \in \mathrm{L}^2$, so is any linear combination $c\, f + d\, g$. The Cauchy–Schwarz inequality (3.16), namely

$$|\, \langle\, f \,;\, g \,\rangle \,| \le \| \, f \, \| \, \| \, g \, \|,$$

implies that the inner product of two square-integrable functions is well-defined and finite. In particular, the Fourier coefficients of a function $f(x)$ are defined as inner products

$$c_k = \langle\, f \,;\, e^{\,\mathrm{i}\,k\,x} \,\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx$$

of f with the complex exponentials, and hence are well-defined for any $f \in \mathrm{L}^2$.

There are some interesting analytical subtleties that arise when one tries to prescribe precisely which functions are to be admitted to Hilbert space. Every piecewise continuous function belongs to L^2. But some functions with singularities are also members. For example, the power function $|\, x \,|^{-\,\alpha}$ belongs to L^2 for any $\alpha < \frac{1}{2}$, but not if $\alpha \ge \frac{1}{2}$.

Analysis requires limiting procedures, and the Hilbert space must be "complete" in the sense that appropriately convergent† sequences of functions have a limit. The completeness requirement relies on the development of the more sophisticated Lebesgue theory of integration, which was formalized in the early part of the twentieth century by the French mathematician Henri Lebesgue — and just in time for quantum mechanics! Any function which is square-integrable in the Lebesgue sense is admitted into L^2. This includes such non-piecewise continuous functions as $\sin \dfrac{1}{x}$ and $x^{-1/3}$, as well as the strange function

$$r(x) = \begin{cases} 1 & \text{if } x \text{ is a rational number,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases} \tag{12.90}$$

One soon discovers that general square-integrable functions can be quite bizarre.

A second complication is that (12.88) does not, strictly speaking, define a norm once we allow discontinuous functions. For example, the piecewise continuous function

$$f_0(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0, \end{cases} \tag{12.91}$$

has norm zero, $\| f_0 \| = 0$, even though it is not zero everywhere. Indeed, any function which is zero except on a set of measure zero also has norm zero, including the function (12.90). Therefore, in order to make (12.88) into a legitimate norm on Hilbert space, we must agree to identify any two functions which have the same values except on a set of measure zero. For instance, the zero function 0 and the preceding examples $f_0(x)$ and $r(x)$ are all viewed as definining the *same* element of Hilbert space. Thus, although we treat them as if they were ordinary functions, each element of Hilbert space is not, in fact, a function, but, rather, an equivalence class of functions all differing on a set of measure zero. All this might strike the reader as becoming much too abstract and arcane. In fact, the casual reader will not lose much by assuming that the "functions" in L^2 are always piecewise continuous and square-integrable. Nevertheless, the full analytical power of Hilbert space theory is only unleashed by allowing much more general kinds of functions into the fold.

After its invention by David Hilbert around the turn of the twentieth century, physicists in the 1920's suddenly realized that Hilbert space was the correct setting to establish the modern theory of quantum mechanics. A quantum mechanical *wave function* is a element† $\varphi \in \mathrm{L}^2$ that has unit norm: $\| \varphi \| = 1$. Thus, the set of wave functions is merely the unit sphere in Hilbert space. In quantum mechanics, a wave function is endowed with a probabilistic interpretation. The modulus $| \varphi(x) |$ of the wave function at a position x quantifies the probability of finding the corresponding particle (photon, electron, etc.)

† The precise technical requirement is that every *Cauchy sequence* of functions $v_k(x) \in \mathrm{L}^2$ converges to a function $v_\star(x) \in \mathrm{L}^2$; see Exercise ■ for details.

† Here we are considering the physical space to be represented by the one-dimensional interval $[-\pi, \pi]$. The more physically relevant case of three-dimensional space is treated similarly, replacing the single integral by a triple integral over all of $\mathbb{R}^3$.

there. More correctly, the probability that the particle resides in a prescribed interval $[a, b]$ is equal to $\sqrt{\dfrac{1}{2\pi}\displaystyle\int_a^b |\,\varphi(x)\,|^2\,dx}$. In particular, the wave function has unit norm

$$\| \varphi \| = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} |\,\varphi(x)\,|^2\,dx} = 1$$

because the particle must certainly, i.e., with probability 1, be *somewhere*!

Convergence in Norm

We are now in a position to discuss convergence in norm of the Fourier series. We begin with the basic definition, which makes sense for any normed vector space.

Definition 12.33. Let V be an normed vector space. A sequence $\mathbf{v}^{(n)} \in V$ is said to *converge in norm* to $\mathbf{v}^\star \in V$ if $\| \mathbf{v}^{(n)} - \mathbf{v}^\star \| \to 0$ as $n \to \infty$.

Remark: Convergence in norm is very different from pointwise convergence. For instance, it is possible, cf. Exercise ■, to construct a sequence of functions that converges in norm to 0, but does not converge pointwise *anywhere*!

We are particularly interested in the convergence in norm of the Fourier series of a square integrable function $f(x) \in \mathrm{L}^2$. Let

$$s_n(x) = \sum_{k=-n}^{n} c_k\, e^{\,\mathrm{i}\,k\,x} \tag{12.92}$$

be the n^{th} partial sum of its Fourier series (12.46). The partial sum (12.92) is an element of the subspace $\mathcal{T}^{(n)} \subset \mathrm{L}^2$ consisting of all trigonometric polynomials of degree at most n, cf. Example 2.12. It is, in fact, a distinguished element of this subspace — namely, *it is the closest function in* $\mathcal{T}^{(n)}$ *to* $f \in \mathrm{L}^2$, where the distance between functions is measured by the L^2 norm of their difference: $\| f - g \|$. Thus, in the language of Chapter 4, the Fourier partial sum $s_n(x)$ is the best trigonometric polynomial approximation to the given function $f(x)$ in the least squares sense. This important characterization of the Fourier partial sums is, in fact, an immediate consequence of the orthonormality of the trigonometric basis.

Theorem 12.34. *The* n^{th} *order Fourier partial sum* $s_n(x)$ *is the closest approximation to* $f(x)$ *in the space of trigonometric polynomials* $\mathcal{T}^{(n)}$*, meaning that it minimizes the* L^2 *norm of the difference*

$$\| f - p_n \|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |\,f(x) - p_n(x)\,|^2\,dx \tag{12.93}$$

among all possible degree n *trigonometric polynomials*

$$p_n(x) = \sum_{k=-n}^{n} d_k\, e^{\,\mathrm{i}\,k\,x} \in \mathcal{T}^{(n)}. \tag{12.94}$$

Proof: The proof is, in fact, a function space version of the proof of the finite-dimensional Theorems 5.34 and 5.36. Note first that, owing to the orthonormality (12.45)

of the basis exponentials, we can compute the norm of a trigonometric polynomial (12.94) by summing the squared moduli of its Fourier coefficients:

$$\| p_n \|^2 = \langle p_n ; p_n \rangle = \sum_{k,l=-n}^{n} d_k \, \overline{d_l} \, \langle e^{\mathrm{i}\, k x} ; e^{\mathrm{i}\, l x} \rangle = \sum_{k=-n}^{n} | d_k |^2;$$

see also (5.6). Hence, we can compute

$$\| f - p_n \|^2 = \| f \|^2 - 2 \langle f ; p_n \rangle + \| p_n \|^2 = \| f \|^2 - 2 \sum_{k=-n}^{n} \overline{d_k} \, \langle f ; e^{\mathrm{i}\, k x} \rangle + \| p_n \|^2$$

$$= \| f \|^2 - 2 \sum_{k=-n}^{n} c_k \, \overline{d_k} + \sum_{k=-n}^{n} | d_k |^2 = \| f \|^2 - \sum_{k=-n}^{n} | c_k |^2 + \sum_{k=-n}^{n} | d_k - c_k |^2$$

The last equality results from adding and subtracting the norm

$$\| s_n \|^2 = \sum_{k=-n}^{n} | c_k |^2$$

of the Fourier partial sum. Therefore,

$$\| f - p_n \|^2 = \| f \|^2 - \| s_n \|^2 + \sum_{k=-n}^{n} | d_k - c_k |^2.$$

The first and second terms in the right hand expression are uniquely determined by $f(x)$ and hence cannot be altered by the choice of trigonometric polynomial $p_n(x)$, which only affects the final summation. Since it is a sum of nonnegative quantities, the sum is, in fact, minimized by setting all the summands to zero, i.e., setting $d_k = c_k$. We conclude that $\| f - p_n \|$ is minimized if and only if $d_k = c_k$ are the Fourier coefficients, and hence $p_n(x) = s_n(x)$ is the Fourier partial sum. *Q.E.D.*

Setting $p_n = s_n$ in the preceding formula, so $d_k = c_k$, we conclude that

$$\| f - s_n \|^2 = \| f \|^2 - \| s_n \|^2 = \| f \|^2 - \sum_{k=-n}^{n} | c_k |^2. \tag{12.95}$$

Now, the left hand side of this equality is always non-negative

$$\| f - s_n \|^2 \geq 0.$$

Applying this inequality to the right hand side, we conclude that the Fourier coefficients of the function f must satisfy the basic inequality

$$\sum_{k=-n}^{n} | c_k |^2 \leq \| f \|^2.$$

Since we are summing a sequence of non-negative numbers whose partial sums are uniformly bounded, the limiting summation as $n \to \infty$ will exist and also be bounded by the right hand side. We have thus proved *Bessel's inequality*

$$\sum_{k=-\infty}^{\infty} |\, c_k \,|^2 \le \| \, f \, \|^2. \tag{12.96}$$

As noted earlier, if a series is to converge, the individual summands must go to zero: $|\, c_k \,|^2 \to 0$. We therefore deduce an immediate corollary — an easy form of the *Riemann–Lebesgue Lemma*.

Lemma 12.35. *If $f \in \mathrm{L}^2$ is square integrable, then its Fourier coefficients satisfy*

$$c_k = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx \quad \longrightarrow \quad 0, \qquad \text{as} \qquad |\, k \,| \to \infty.$$

This is equivalent to the convergence of the real Fourier coefficients

$$\left.\begin{aligned} a_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos k\,x\, dx \\ b_k &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin k\,x\, dx \end{aligned}\right\} \longrightarrow\ 0, \qquad \text{as} \qquad k \to \infty.$$

Remark: As before, the convergence of the sum (12.96) requires that the coefficients c_k cannot tend to zero too slowly. For instance, if c_k satisfies the power bound

$$|\, c_k \,| \le M\, |\, k \,|^{-\alpha}, \qquad \text{then} \qquad \sum_{k=-\infty}^{\infty} |\, c_k \,|^2 < \infty \qquad \text{provided} \qquad \alpha > \tfrac{1}{2}.$$

Uniform convergence required $\alpha > 1$, cf. (12.85), and hence convergence in norm imposes a less restrictive assumption on the decay of the Fourier coefficients. Indeed, a Fourier series may very well converge in norm to a discontinuous function, which is not possible under uniform convergence. In fact, there are some bizarre continuous functions whose Fourier series do not converge uniformly, failing to converge at all at some points. A deep result says that the Fourier series of a continuous function converges except possibly on a set of measure zero, [**159**]. Again, the subtle details of the convergence of Fourier series are rather delicate, and lack of space and analytical tools prevents us from delving any further into these issues.

Completeness

As we know, specification of a basis allows one to describe all elements of a finite-dimensional vector. The number of basis elements equals the dimension of the vector space. For an infinite-dimensional vector space, there are, by definition, infinitely many linearly independent elements, and no finite collection can serve as a basis. The question then arises to what extent an infinite collection of linearly independent elements can be considered as a basis for the space. Mere counting will no longer suffice, since omitting

one, or two, or any finite number, or even certain infinite subcollections, from a supposed basis will still leave infinitely many linearly independent elements of the vector space; but, clearly, the reduced collection should, in some sense, no longer serve to define a basis. The curse of infinity strikes again! For example, while the complete trigonometric collection $1, \cos x, \sin x, \cos 2x, \sin 2x, \dots$ can represent any 2π periodic L^2 function as a Fourier series, the subcollection $\cos x, \sin x, \cos 2x, \sin 2x, \dots$ will only represent functions with mean zero, while the functions $\sin x, \sin 2x, \dots$ only represent odd functions. All three collections have infinitely many elements, but only the first can be properly called a basis. In general, just because we have found a infinite collection of independent elements in an infinite-dimensional vector space, how do we know that we have enough, and are not missing one or two or 10,000 or even infinitely many additional elements?

The concept of "completeness" serves to properly formalize the notion of a "basis" of an infinite-dimensional vector space. We shall discuss completeness in a general, abstract setting, but the key example is, of course, the Hilbert space L^2 and the system of trigonometric or complex exponential functions forming a Fourier series. Other important examples arising in later applications include Bessel functions, Legendre polynomials, spherical harmonics, and general systems of eigenfunctions of self-adjoint boundary value problems.

For simplicity, we only define completeness in the case of orthonormal systems. Similar arguments will clearly apply to orthogonal systems, but the additional normality condition helps to simplify the formulae. Let V be an infinite-dimensional inner product space. Suppose that $u_1, u_2, u_3, \dots \in V$ form an orthonormal collection of elements of V, so

$$\langle\, u_i \,;\, u_j \,\rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{12.97}$$

A straightforward argument proves that the u_i are linearly independent; see Proposition 5.4. Given a general element $f \in V$, we form its *generalized Fourier series*

$$f \;\sim\; \sum_{k=1}^{\infty} c_k\, u_k, \qquad \text{where} \qquad c_k = \langle\, f \,;\, u_k \,\rangle. \tag{12.98}$$

The Fourier coefficients c_k are given by our usual orthonormal basis formula (5.5), which is obtained by formally taking the inner product of the series with u_k.

Definition 12.36. An orthonormal system of elements $u_1, u_2, u_3, \dots \in V$ is called *complete* if the generalized Fourier series (12.98) of any $f \in V$ converges in norm to f. In other words,

$$\| f - s_n \| \;\longrightarrow\; 0, \qquad \text{where} \qquad s_n = \sum_{k=1}^{n} c_k\, u_k \tag{12.99}$$

is the n^{th} partial sum of the generalized Fourier series (12.98).

Thus, completeness requires that every element can be arbitrarily closely approximated (in norm) by a suitable linear combination of the basis elements. A complete orthonormal system should be viewed as the infinite-dimensional version of an orthonormal basis of a finite-dimensional vector space.

The key result for Fourier series is that the complex exponentials, or, equivalently the trigonometric functions, form a complete system.

Theorem 12.37. *The complex exponentials $e^{\,\mathrm{i}\,k\,x}$, $k = 0, \pm 1, \pm 2, \ldots$, form a complete orthonormal system in $\mathrm{L}^2[-\pi, \pi]$. In other words, if $s_n(x)$ denotes the n^{th} partial sum* (12.99) *of the Fourier series of the square-integrable function $f(x) \in \mathrm{L}^2[-\pi, \pi]$, then*

$$\lim_{n \to \infty} \| f - s_n \| = 0. \tag{12.100}$$

An indication of the proof of completeness will appear below.

Remark: Theorem 12.37 is, in fact, a particular case of a theorem that governs orthogonal eigenfunction expansions arising from quite general positive definite boundary value problems.

In order to understand this result, let us first describe some equivalent characterizations of completeness. The Plancherel formula is the infinite-dimensional counterpart of our formula (5.6) for the norm of a vector in terms of its coordinates with respect to an orthonormal basis.

Theorem 12.38. *The orthonormal system of elements $u_1, u_2, u_3, \ldots \in V$ is complete if and only if the* Plancherel formula

$$\| f \|^2 = \sum_{k=-\infty}^{\infty} | c_k |^2 \qquad \text{where} \qquad c_k = \langle f ; u_k \rangle, \tag{12.101}$$

holds for every $f \in V$.

Proof: We begin by computing[†] the Hermitian norm

$$\| f - s_n \|^2 = \| f \|^2 - \langle f ; s_n \rangle - \langle s_n ; f \rangle + \| s_n \|^2 = \| f \|^2 - 2 \operatorname{Re} \langle f ; s_n \rangle + \| s_n \|^2.$$

Substituting the formula $s_n = \sum_{k=1}^{n} c_k u_k$ for the partial sums, we find, by orthonormality,

$$\| s_n \|^2 = \sum_{k=1}^{n} | c_k |^2, \qquad \text{while} \qquad \langle f ; s_n \rangle = \sum_{k=1}^{n} \overline{c_k} \, \langle f ; u_k \rangle = \sum_{k=1}^{n} | c_k |^2.$$

Therefore,

$$0 \leq \| f - s_n \|^2 = \| f \|^2 - \sum_{k=1}^{n} | c_k |^2. \tag{12.102}$$

[†] We are in essence repeating the proofs of Theorem 12.34 and the subsequent trigonometric Bessel inequality (12.96) in a more abstract setting.

The fact that the left hand side of (12.103) is non-negative for all n implies the general *Bessel inequality*

$$\| f \|^2 \;\geq\; \sum_{k=1}^{\infty} | c_k |^2, \tag{12.103}$$

which is valid for *any* orthonormal system of elements in an inner product space. As we noted above, Bessel's inequality implies that the generalized Fourier coefficients $c_k \to 0$ must tend to zero reasonably rapidly in order that the sum of their squares converges.

Plancherel's formula (12.101), thus, states that, if the system of functions is complete, the Bessel inequality is, in fact, an equality! Indeed, letting $n \to \infty$ in (12.102), we have

$$0 \;=\; \lim_{n \to \infty} \| f - s_n \|^2 \;=\; \| f \|^2 - \sum_{k=1}^{\infty} | c_k |^2.$$

Therefore, the completeness condition (12.100) holds if and only if the right hand side vanishes, which is the Plancherel identity (12.101). *Q.E.D.*

Corollary 12.39. *Let $c_k = \langle f ; \varphi_k \rangle, d_k = \langle g ; \varphi_k \rangle$ are the Fourier coefficients of f, g, respectively, with respect to a complete orthonormal system. Then they satisfy* Parseval's identity

$$\langle f ; g \rangle = \sum_{k=-\infty}^{\infty} c_k \, \overline{d_k} \, . \tag{12.104}$$

Proof: According to Exercise ■,

$$\langle f ; g \rangle = \tfrac{1}{4} \big(\| f + g \|^2 - \| f - g \|^2 + \mathrm{i} \, \| f + \mathrm{i} \, g \|^2 - \mathrm{i} \, \| f - \mathrm{i} \, g \|^2 \big).$$

We apply the Plancherel formula (12.101) to each term on the right hand side, so

$$\langle f ; g \rangle = \frac{1}{4} \sum_{k=-\infty}^{\infty} \big(| c_k + d_k |^2 - | c_k - d_k |^2 + \mathrm{i} \, | c_k + \mathrm{i} \, d_k |^2 - \mathrm{i} \, | c_k - \mathrm{i} \, d_k |^2 \big) = \sum_{k=-\infty}^{\infty} c_k \, \overline{d_k}.$$

Q.E.D.

In particular, in the case of the complex exponential basis of $\mathrm{L}^2[-\pi, \pi]$, the Plancherel and Parseval formulae tell us that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} | f(x) |^2 \, dx = \sum_{k=-\infty}^{\infty} | c_k |^2, \qquad \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, \overline{g(x)} \, dx = \sum_{k=-\infty}^{\infty} c_k \, \overline{d_k} \, , \tag{12.105}$$

in which c_k, d_k are, respectively, the ordinary Fourier coefficients of the complex-valued functions $f(x)$ and $g(x)$. Note that the Plancherel formula is a special case of the Parseval identity obtained by setting $f = g$. In Exercise ■, the reader is asked to rewrite the two formulas in terms of the real Fourier coefficients.

Completeness also tells us that a function is uniquely determined by its Fourier coefficients.

Proposition 12.40. *If the orthonormal system $u_1, u_2, \ldots \in V$ is complete, then the only element $f \in V$ with all zero Fourier coefficients, $0 = c_1 = c_2 = \cdots$, is the zero element: $f = 0$. More generally, two elements $f, g \in V$ have the same Fourier coefficients if and only if they are the same: $f = g$.*

Proof: The proof is an immediate consequence of the Plancherel formula. Indeed, if $c_k = 0$, then (12.101) implies that $\| f \| = 0$. The second statement follows by applying the first to the function $f - g$. *Q.E.D.*

Another way of stating this result is that the only function which is orthogonal to every element of a complete orthonormal system is the zero function†. Interpreting in another way, a complete orthonormal system is maximal in the sense that no further orthonormal elements can be appended to it.

Let us now discuss the completeness of the Fourier trigonometric/complex exponential functions. We shall prove the completeness criterion only for continuous functions, leaving the harder general proof to the references, [**51**, **159**]. According to Theorem 12.27, if $f(x)$ is continuous, 2π periodic, and piecewise C^1, its Fourier series converges uniformly to $f(x)$, so

$$f(x) \;=\; \sum_{k=-\infty}^{\infty} c_k\, e^{\,\mathrm{i}\,k\,x} \qquad \text{for all} \qquad -\,\pi \le x \le \pi.$$

The same holds for its complex conjugate $\overline{f(x)}$. Therefore,

$$|\, f(x)\,|^2 = f(x)\,\overline{f(x)} = f(x) \sum_{k=-\infty}^{\infty} \overline{c}_k\, e^{-\,\mathrm{i}\,k\,x} = \sum_{k=-\infty}^{\infty} \overline{c}_k\, f(x)\, e^{-\,\mathrm{i}\,k\,x},$$

which also converges uniformly by (12.80). Equation (12.81) permits us to integrating both sides from $-\pi$ to π, yielding

$$\| f \|^2 = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} |\, f(x)\,|^2\, dx = \sum_{k=-\infty}^{\infty} \frac{1}{2\,\pi} \int_{-\pi}^{\pi} \overline{c}_k\, f(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx = \sum_{k=-\infty}^{\infty} c_k\, \overline{c}_k = \sum_{k=-\infty}^{\infty} |\, c_k\,|^2.$$

Therefore, Plancherel's identity (12.101) holds for any continuous function. With some additional technical work, this result is used to establish the validity of Plancherel's formula for all $f \in \mathrm{L}^2$, the key step being to suitably approximate f by continuous functions. With this in hand, completeness follows from Theorem 12.38. *Q.E.D.*

Pointwise Convergence

Let us finally turn to the proof of the Pointwise Convergence Theorem 12.7. The goal is to prove that, under the appropriate hypotheses, the limit of the partial Fourier sums is

$$\lim_{n \to \infty} s_n(x) = \tfrac{1}{2}\, \big[\, f(x^+) + f(x^-)\,\big]. \tag{12.106}$$

† Or, to be more technically accurate, any function which is zero outside a set of measure zero.

We begin by substituting the formulae (12.47) for the complex Fourier coefficients into the formula (12.92) for the n^{th} partial sum:

$$s_n(x) = \sum_{k=-n}^{n} c_k\, e^{\,\mathrm{i}\,k\,x} = \sum_{k=-n}^{n} \left(\frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(y)\, e^{-\,\mathrm{i}\,k\,y}\, dy \right) e^{\,\mathrm{i}\,k\,x}$$
$$= \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(y) \sum_{k=-n}^{n} e^{\,\mathrm{i}\,k(x-y)}\, dy.$$

We can then use the geometric summation formula (12.54) to evaluate the result:

$$s_n(x) = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(y)\, \frac{\sin\left(n + \frac{1}{2}\right)(x-y)}{\sin \frac{1}{2}(x-y)}\, dy$$
$$= \frac{1}{2\,\pi} \int_{x-\pi}^{x+\pi} f(x+y)\, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(x+y)\, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy.$$

The next to last equality comes from changing variable in the integral from y to $x + y$. The final equality comes from the fact that the integrand is $2\,\pi$ periodic, and so its integral over any interval of length $2\,\pi$ has the same value; see Exercise ∎.

Thus, to prove (12.106), it suffices to show that

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{\pi} \int_{0}^{\pi} f(x+y)\, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy &= f(x^+), \\ \lim_{n \to \infty} \frac{1}{\pi} \int_{-\pi}^{0} f(x+y)\, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy &= f(x^-). \end{aligned} \tag{12.107}$$

The proofs of the two formulae are identical, and so we concentrate on the first. Equation (12.56) implies that

$$\frac{1}{\pi} \int_{0}^{\pi} \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy = \frac{1}{\pi} \int_{0}^{\pi} \sum_{k=-n}^{n} e^{\,\mathrm{i}\,k\,y}\, dy = 1.$$

Multiplying the right hand side of the first equation in (12.107) by the integral allows us to rewrite it in the form

$$\lim_{n \to \infty} \frac{1}{\pi} \int_{0}^{\pi} \frac{f(x+y) - f(x^+)}{\sin \frac{1}{2} y} \sin\left(n + \tfrac{1}{2}\right) y \; dy = 0. \tag{12.108}$$

We claim that, for each fixed value of x, the function

$$g(y) = \frac{f(x+y) - f(x^+)}{\sin \frac{1}{2} y}$$

is piecewise continuous for all $0 \le y \le \pi$. Owing to our hypothesis on $f(x)$, the only problematic point is when $y = 0$, but then

$$\lim_{y \to 0^+} g(y) = \lim_{y \to 0^+} \frac{f(x+y) - f(x^+)}{y}\, \frac{y}{\sin \frac{1}{2} y} = 2\, f'(x^+)$$

is twice the right hand derivative of f at x. The factor of 2 comes from the elementary calculus limit[†]

$$\lim_{y \to 0^+} \frac{y}{\sin \frac{1}{2} y} = 2 \lim_{y \to 0^+} \frac{\frac{1}{2} y}{\sin \frac{1}{2} y} = 2 \lim_{z \to 0^+} \frac{z}{\sin z} = 2.$$

Thus, formula (12.108) will follow if we can show that

$$\lim_{n \to \infty} \frac{1}{\pi} \int_0^\pi g(y) \sin \left(n + \tfrac{1}{2}\right) y \, dy = 0 \tag{12.109}$$

for any piecewise continuous function g. Were it not for the extra $\frac{1}{2}$, this would immediately follow from Lemma 12.35. More honestly, we use the addition formula for $\sin \left(n + \frac{1}{2}\right) y$ to write

$$\frac{1}{\pi} \int_0^\pi g(y) \sin \left(n + \tfrac{1}{2}\right) y \, dy = \frac{1}{\pi} \int_0^\pi \left[g(y) \sin \tfrac{1}{2} y \right] \cos n y \, dy + \frac{1}{\pi} \int_0^\pi \left[g(y) \cos \tfrac{1}{2} y \right] \sin n y \, dy$$

The first integral is the Fourier cosine coefficient $\widetilde{a}_n$ for the piecewise continuous function $g(y) \sin \frac{1}{2} y$, while the second integral is the Fourier sine coefficient $\widetilde{b}_n$ for the piecewise continuous function $g(y) \cos \frac{1}{2} y$. Lemma 12.35 implies that both of these converge to zero as $n \to \infty$, and hence (12.109) holds. This completes the proof. *Q.E.D.*

Remark: An alternative approach to the last part of the proof is to use the general *Riemann–Lebesgue Lemma*.

Lemma 12.41. *Suppose $g(x)$ is piecewise continuous on $[a, b]$. Then*

$$\lim_{\omega \to \infty} \int_a^b g(x) \, e^{\mathrm{i} \, \omega \, x} \, dx = 0.$$

Intuitively, the lemma says that, as the frequency ω gets larger and larger, the increasingly rapid oscillations in $\sin \omega x$ tend to cancel each other out. A formal proof of the lemma from first principles can be found in [**51**, **159**].

[†] One might argue that this follows from l'Hôpital's rule, but in point of fact this is a fake application: one already needs to evaluate the limit when proving the formula for the derivative of the sine function!

Chapter 13

Fourier Analysis

Fourier series are merely the entry point into the wonderful world of Fourier analysis and its wide-ranging extensions and generalizations. An entire industry is devoted to developing the theory and enlarging the scope of applications of Fourier–inspired methods. New species of Fourier systems continue to be discovered and applied in a broad range of physical, mathematical, engineering, chemical, biological, financial and other systems.

Digital media, such as CD's, DVD's, MP3's, and so on, rely on discrete data, not continuous functions. This inspires the development of a purely discrete version of Fourier series methods, in which one replaces periodic functions representing analog signals by their discrete sample vectors. The resulting discrete Fourier sum can, in fact, be handled by finite-dimensional vector space methods, and so technically belongs in the previous linear algebra portion of this text. However, the insight provided by the classical continuous Fourier theory proves to be essential in understanding and analyzing its discrete digital cousin. An important application of discrete Fourier sums is in signal and image processing. One typically samples a signal at equally spaced time intervals, and then processes the resulting discrete (digital) data. Basic methods of data compression and noise removal are based on the corresponding discrete Fourier coefficients, acting on the observation that noise lies in the high frequency Fourier modes, while most important features are concentrated at low frequencies. In Section 13.1, we develop the basic Fourier theory in this discrete setting, culminating with the *Fast Fourier Transform*, which forms a fast numerical algorithm for passing between a signal and its discrete Fourier coefficients, and is an essential tool in modern signal processing.

One of the inherent limitations of classical Fourier methods, both continuous and discrete, is that they are not well adapted to localized data owing to the nonlocal character of the constituent trigonometric functions. For instance, localized compression algorithms in the physical domain become unmanageable in the Fourier frequency version of the signal. (The physics version of this mathematical fact is known as the Heisenberg Uncertainty Principle.) In the second section, we introduce wavelets, a very recent extension of the Fourier theory that more naturally incorporates the roles of various scales and localization into the analysis. The modern theory of wavelets is playing an increasingly dominant role in many modern applications. For instance, the new JPEG digital image compression format is based on wavelets, as are the current FBI fingerprint files used in law enforcement in the United States.

Analysis of non-periodic functions defined on the entire real line requires replacing the Fourier series by a limiting Fourier integral, leading to the justly famous Fourier transform. Fourier transforms play an essential role in ordinary and partial differential equations,

quantum mechanics, data analysis, and many other areas. In Section 13.3, we introduce the most important features of the Fourier transform in a form suitable for applications, leaving more subtle analytical details to a more advanced treatment of the subject. The real version of the Fourier transform is known as the Laplace transform. Both the Fourier and Laplace transforms change differentiation into multiplication, thereby converting linear differential equations into algebraic equations. The Fourier transform is used for solving boundary value problems on the real line, while initial value problems are most effectively handled by the Laplace transform. Again, our emphasis is on practical features of the method.

13.1. Discrete Fourier Analysis and the Fast Fourier Transform.

In practical computations, one does not deal with continuous functions, but rather with discrete numerical data. For example, even when measuring a continuous signal, one can only perform a finite, discrete set of measurements — leading to a *sample* of the full signal data. In digital media (CD's, DVD's, etc.), or experimental data that is stored on a computer, this is all we have — a signal sampled at discrete time intervals. (And then "quantized" because we cannot store its sample values to infinite precision on a digital computer — but that is part of yet another story.) Therefore, although the Fourier series are of unquestionable theoretical importance, the complications involved in the passage to an infinite-dimensional function space could, from the computer's point of view, be entirely avoided by restricting from the outset to finite-dimensional vector spaces of sampled data. Nevertheless, the insight gained from the classical continuous version of Fourier series in crucial to the proper formulation and analysis of its discrete counterpart.

In general, then, instead of a function $f(x)$ defined on an interval $a \leq x \leq b$, the computer can only store its measured values at a finite number of *sample points* $a \leq x_0 < x_1 < \cdots < x_n \leq b$. In the simplest and, by far, the most common case, the sample points are equally spaced, and so

$$x_j = a + j\,h, \qquad j = 0, \ldots, n, \qquad \text{where} \qquad h = \frac{b-a}{n}$$

indicates the sample rate. In signal processing applications, x represents time instead of space, and the x_j represent the times that we sample the signal $f(x)$. Sample rates can be very high, e.g., every 10–20 milliseconds in current speech recognition systems.

Fourier series apply to periodic functions; discrete Fourier sums apply to sampled periodic signals. (Of course, real life signals are rarely periodic, so, for analytical purposes, the methods rely on an artificial periodic extension of the original signal.) For simplicity, we adopt the standard period of 2π, although one can readily change the results to other intervals by rescaling, as in Section 12.4. Here, though, it will be more convenient to take the basic interval of definition to be $[0, 2\pi]$ instead of $[-\pi, \pi]$. Consider the n equally spaced sample points

$$x_0 = 0, \qquad x_1 = \frac{2\pi}{n}, \qquad x_2 = \frac{4\pi}{n}, \quad \ldots \quad x_j = \frac{2j\pi}{n}, \quad \ldots \quad x_{n-1} = \frac{2(n-1)\pi}{n}. \quad (13.1)$$

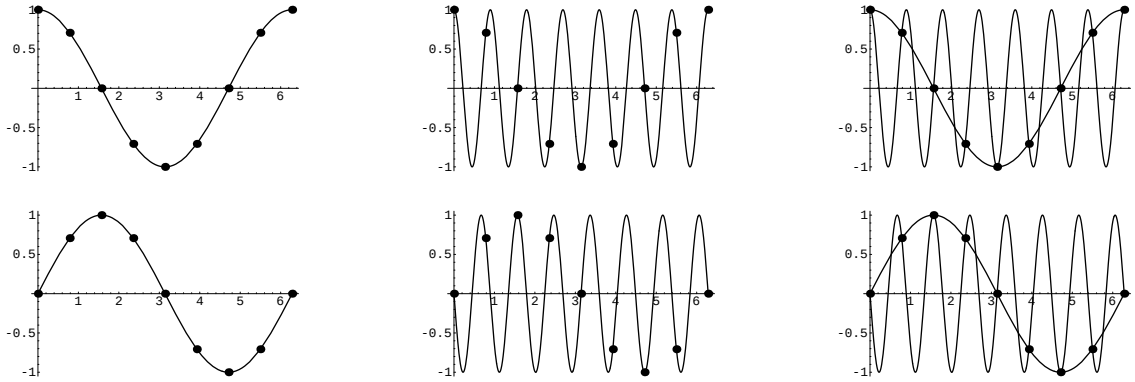


Figure 13.1. Sampling $e^{-\mathrm{i}\,x}$ and $e^{7\,\mathrm{i}\,x}$ on $n = 8$ sample points.

Periodicity requires that $f(0) = f(2\pi)$, and so the final sample point $x_n = 2\pi$ is superfluous and will be omitted. Sampling a (complex-valued) signal or function $f(x)$ at the sample points produces the *sample vector*

$$\mathbf{f} = \big(\, f_0, f_1, \ldots, f_{n-1} \,\big)^T = \big(\, f(x_0), f(x_1), \ldots, f(x_{n-1}) \,\big)^T,$$

where

$$f_j = f(x_j) = f\left(\frac{2\,j\,\pi}{n}\right). \tag{13.2}$$

Sampling cannot distinguish between functions that have the same values at all of the sample points — from the sampler's point of view they are identical. For example, the periodic function

$$f(x) = e^{\,\mathrm{i}\,n\,x} = \cos n\,x + \,\mathrm{i}\,\sin n\,x$$

has sampled values

$$f_j = f\left(\frac{2\,j\,\pi}{n}\right) = \exp\left(\,\mathrm{i}\,n\,\frac{2\,j\,\pi}{n}\right) = e^{2\,j\,\pi\,\mathrm{i}} = 1 \qquad \text{for all} \qquad j = 0, \ldots, n-1,$$

and hence is indistinguishable from the constant function $c(x) \equiv 1$ — both lead to the *same* sample vector $(\,1, 1, \ldots, 1\,)^T$. This has the important implication that sampling at n equally spaced sample points *cannot* detect periodic signals of frequency n. More generally, the two complex exponential signals

$$e^{\,\mathrm{i}\,(k+n)x} \;\sim\; e^{\,\mathrm{i}\,k\,x} \tag{13.3}$$

are also indistinguishable when sampled. This has the important consequence that we need only use the first n periodic complex exponential functions

$$f_0(x) = 1, \qquad f_1(x) = e^{\,\mathrm{i}\,x}, \qquad f_2(x) = e^{2\,\mathrm{i}\,x}, \qquad \ldots \qquad f_{n-1}(x) = e^{(n-1)\,\mathrm{i}\,x}, \tag{13.4}$$

in order to represent an arbitrary 2π periodic sampled signal. In particular, exponentials $e^{-\,\mathrm{i}\,k\,x}$ of "negative" frequency can all be converted into positive versions, namely $e^{\,\mathrm{i}\,(n-k)x}$, by the same sampling argument (13.3). For example,

$$e^{-\,\mathrm{i}\,x} = \cos x - \,\mathrm{i}\,\sin x \qquad \text{and} \qquad e^{(n-1)\,\mathrm{i}\,x} = \cos(n-1)\,x + \,\mathrm{i}\,\sin(n-1)\,x$$

have identical values on the sample points (13.1). However, off of the sample points, they are quite different; the former is of low frequency, while the latter represents a high frequency oscillation. In Figure 13.1, we compare $e^{-\mathrm{i}\,x}$ and $e^{7\,\mathrm{i}\,x}$ when there are $n = 8$ sample points, indicated by the large dots on the graphs. The top row compares the real parts, $\cos x$ and $\cos 7x$, while the bottom row compares the imaginary parts, $\sin x$ and $-\sin 7x$. Note that both functions have the same pattern of sample values, even though their overall behavior is strikingly different; this effect is commonly referred to as *aliasing*[†]. If you view a moving particle under a stroboscopic light that flashes eight times, you would be unable to determine which of the two graphs the particle was following. Aliasing is the cause of a well-known artifact in movies: sometimes spoked wheels appear to be rotating backwards because, owing to the discretizing imposed by the frames of the film, our brain psychologically views a point on the wheel to be following a low freuency backwards motion whereas in reality it is making a high frequency forward motion that has exactly the same sample poistions on each frame of the movie.

Remark: Aliasing also has important implications for the design of music CD's. We must sample at a sufficiently high rate that all audible frequencies can be adequately represented. In fact, human appreciation of music also relies on inaudible high frequency tones, and so a much higher sample rate is actually used in commercial CD design. The chosen sample rate remains controversial; hi fi aficionados complain that it was not set high enough to fully reproduce the musical quality of an analog LP record!

The complex Fourier series (12.46) decomposes a function $f(x)$ into a sum of complex exponentials. In the discrete version, we do the same thing with our sampled function. Since we cannot distinguish sampled exponentials of frequency higher than n, we only need consider a finite sum

$$f(x) \;\sim\; p(x) \;=\; c_0 + c_1\, e^{\,\mathrm{i}\,x} + c_2\, e^{2\,\mathrm{i}\,x} + \;\cdots\; + c_{n-1}\, e^{(n-1)\,\mathrm{i}\,x} \;=\; \sum_{k=0}^{n-1} c_k\, e^{\,\mathrm{i}\,k\,x} \tag{13.5}$$

using the first n exponentials (13.4). Equation (13.5), which has the effect of decomposing the signal $f(x)$ into a linear combination of purely periodic signals, is known as the *discrete Fourier sum* (or sometimes, even though it has only finitely many terms, *series*) of the function $f(x)$. Thus, the $\sim$ in (13.5) will mean that the function $f(x)$ and the sum $p(x)$ agree on the sample points,

$$f(x_j) = p(x_j), \qquad\qquad j = 0, \dots, n-1, \tag{13.6}$$

and hence $p(x)$ is an *interpolating trigonometric polynomial* of degree $\leq n-1$ for the sample data $f_j = f(x_j)$.

Remark: If $f(x)$ is real, then $p(x)$ is also real on the data points, but may very well be complex-valued on intermediate points. Thus, in practice, the real part of $p(x)$ is used as the interpolating trigonometric polynomial.

[†] In computer graphics, the term "aliasing" is used in a much broader sense that covers a variety of artifacts introduced by discretization — particularly, the jagged appearance of lines and smooth curves on a digital monitor.

Since we are working in the finite-dimensional vector space $\mathbb{C}^n$ throughout, we may reformulate the discrete Fourier series in vectorial form. Sampling the basic exponentials (13.4) produces the complex vectors

$$\begin{aligned} \boldsymbol{\omega}_k &= \left(e^{\mathrm{i}\, k x_0}, e^{\mathrm{i}\, k x_1}, e^{\mathrm{i}\, k x_2}, \dots, e^{\mathrm{i}\, k x_n} \right)^T \\ &= \left(1, e^{2 k \pi \mathrm{i}/n}, e^{4 k \pi \mathrm{i}/n}, \dots, e^{2(n-1) k \pi \mathrm{i}/n} \right)^T, \end{aligned} \qquad k = 0, \dots, n-1. \tag{13.7}$$

Thus, the interpolation conditions (13.6) can be recast in the equivalent vector form

$$\mathbf{f} = c_0\, \boldsymbol{\omega}_0 + c_1\, \boldsymbol{\omega}_1 + \, \cdots \, + c_{n-1}\, \boldsymbol{\omega}_{n-1}. \tag{13.8}$$

In other words, to compute the discrete Fourier coefficients $c_0, \dots, c_{n-1}$ of f, all we need to do is rewrite its sample vector $\mathbf{f}$ as a linear combination of the sampled exponential vectors $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$.

Now, as with continuous Fourier series, the crucial property is the orthogonality of the basis elements. Were it not for orthogonality and its simplifying consequences, the proceeding ideas would have remained mere mathematical curiosities, rather than an essential tool for applications.

Proposition 13.1. *The sampled exponential vectors $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$ form an orthonormal basis of $\mathbb{C}^n$ with respect to the inner product*

$$\langle\, \mathbf{f}\,;\mathbf{g}\,\rangle \;=\; \frac{1}{n} \sum_{j=0}^{n-1} f_j\, \overline{g_j} \;=\; \frac{1}{n} \sum_{j=0}^{n-1} f(x_j)\, \overline{g(x_j)}\,, \qquad \mathbf{f}, \mathbf{g} \in \mathbb{C}^n. \tag{13.9}$$

The inner product (13.9) is a rescaled version of the standard Hermitian dot product (3.82) between complex vectors. We can interpret the inner product between the sample vectors $\mathbf{f}, \mathbf{g}$ as the *average* of the sampled values of the product signal $f(x)\, \overline{g(x)}$.

Remark: As usual, orthogonality is no accident. Just as the complex exponentials are eigenfunctions for a self-adjoint boundary value problem, so their discrete sampled counterparts are eigenvectors for a self-adjoint matrix eigenvalue problem; see Exercise ■ for details. Here, to keep the discussion on track, we shall outline a direct proof.

Proof: The crux of the matter relies on properties of the remarkable complex numbers

$$\zeta_n = e^{2\pi \mathrm{i}/n} = \cos \frac{2\pi}{n} + \mathrm{i} \sin \frac{2\pi}{n}\,, \tag{13.10}$$

where $n = 1, 2, 3, \dots$. Particular cases include

$$\zeta_2 = -1, \qquad \zeta_3 = -\frac{\sqrt{3}}{2} + \frac{1}{2}\,\mathrm{i}\,, \qquad \zeta_4 = \mathrm{i}\,, \qquad \text{and} \qquad \zeta_8 = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}\,\mathrm{i}\,. \tag{13.11}$$

The n^{th} power of $\zeta = \zeta_n$ is

$$\zeta^n = \left(e^{2\pi \mathrm{i}/n} \right)^n = e^{2\pi \mathrm{i}} = 1,$$

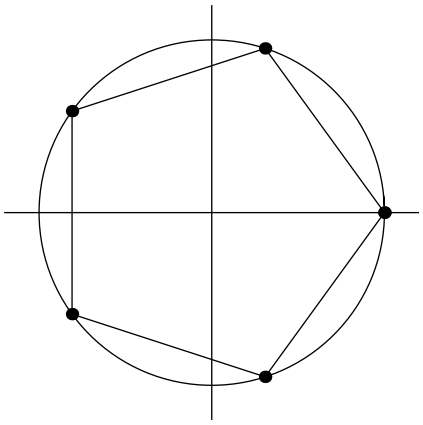


Figure 13.2. The Fifth Roots of Unity.

and hence it is one of the complex n^{th} roots of unity: $\zeta = \zeta_n = \sqrt[n]{1}$. There are, in fact, n different complex n^{th} roots of 1, including 1 itself; these roots are the different powers of ζ, namely

$$\zeta^k = e^{2\,k\,\pi\,\mathrm{i}\,/n} = \cos\frac{2\,k\,\pi}{n} + \mathrm{i}\,\sin\frac{2\,k\,\pi}{n}\,, \qquad k = 0, \dots, n-1. \tag{13.12}$$

For this reason, ζ_n is known as the *primitive* n^{th} *root of unity*. Geometrically, the n^{th} roots of 1 form the vertices of a regular unit n–gon in the complex plane; see Figure 13.2. The primitive root ζ is the first vertex we encounter as we go counterclockwise around the n–gon starting at 1. The other roots appear in their natural order $\zeta^2, \zeta^3, \dots, \zeta^{n-1}$, and finishing back at $\zeta^n = 1$. The complex conjugate of ζ is the "last" n^{th} root

$$e^{-2\,\pi\,\mathrm{i}\,/n} = \overline{\zeta} = \frac{1}{\zeta} = \zeta^{n-1} = e^{2(n-1)\pi\,\mathrm{i}\,/n} \tag{13.13}$$

The complex numbers (13.12) are a complete set of roots of the polynomial $z^n - 1$, which can therefore be factored:

$$z^n - 1 = (z-1)(z-\zeta)(z-\zeta^2)\,\cdots\,(z-\zeta^{n-1}).$$

On the other hand, elementary algebra provides us with the real factorization

$$z^n - 1 = (z-1)(1 + z + z^2 + \,\cdots\, + z^{n-1}).$$

Comparing the two factorizations, we conclude that

$$1 + z + z^2 + \,\cdots\, + z^{n-1} = (z-\zeta)(z-\zeta^2)\cdots(z-\zeta^{n-1}).$$

Substituting $z = \zeta^k$ into both sides of this identity, we deduce a useful formula

$$1 + \zeta^k + \zeta^{2k} + \,\cdots\, + \zeta^{(n-1)\,k} = \begin{cases} n, & k = 0, \\ 0, & 0 < k < n. \end{cases} \tag{13.14}$$

Since $\zeta^{n+k} = \zeta^k$, the preceding formula can be applied to general integers k — the sum is equal to n if n evenly divides k and is 0 otherwise.

Now, let us apply what we've learned to prove Proposition 13.1. First, in view of (13.12), the sampled exponential vectors (13.7) can all be written in terms of the n^{th} roots of unity:

$$\boldsymbol{\omega}_k = \big(\, 1, \zeta^k, \zeta^{2\,k}, \zeta^{3\,k}, \ \ldots \ , \zeta^{(n-1)\,k} \,\big)^T, \qquad k = 0, \ldots, n-1. \tag{13.15}$$

Therefore, applying (13.13), (13.14), we conclude

$$\langle\, \boldsymbol{\omega}_k \,;\boldsymbol{\omega}_l \,\rangle = \frac{1}{n} \sum_{j=0}^{n-1} \zeta^{jk}\, \overline{\zeta^{jl}} = \frac{1}{n} \sum_{j=0}^{n-1} \zeta^{j(k-l)} = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \qquad 0 \leq k, l < n,$$

which proves orthonormality. *Q.E.D.*

Since the sampled exponential vectors are orthonormal, we can immediately compute the Fourier coefficients in the discrete Fourier sum (13.5) by taking inner products:

$$c_k = \langle\, \mathbf{f} \,;\boldsymbol{\omega}_k \,\rangle = \frac{1}{n} \sum_{j=0}^{n-1} f_j\, \overline{e^{\,\mathrm{i}\,k\,x_j}} = \frac{1}{n} \sum_{j=0}^{n-1} f_j\, e^{-\,\mathrm{i}\,k\,x_j} = \frac{1}{n} \sum_{j=0}^{n-1} \zeta^{-jk}\, f_j. \tag{13.16}$$

In other words, the discrete Fourier coefficient c_k is obtained by averaging the sampled values of the product function $f(x)\, e^{-\,\mathrm{i}\,k\,x}$. The passage from the signal to its Fourier coefficients and back is known as the *discrete Fourier transform*. The inverse procedure of reconstructing a signal from its discrete Fourier coefficients via the sum (13.5) (or (13.8)) is known as the *inverse discrete Fourier transform*. The discrete Fourier transform and its inverse define mutually inverse linear transformations on the space $\mathbb{C}^n$, whose matrix representations can be found in Exercise ∎.

Example 13.2. If $n = 4$, then $\zeta = \zeta_4 = \mathrm{i}$. In this case, the sampled exponential vectors (13.15) are

$$\boldsymbol{\omega}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\omega}_1 = \begin{pmatrix} 1 \\ \mathrm{i} \\ -1 \\ -\,\mathrm{i} \end{pmatrix}, \qquad \boldsymbol{\omega}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \qquad \boldsymbol{\omega}_3 = \begin{pmatrix} 1 \\ -\,\mathrm{i} \\ -1 \\ \mathrm{i} \end{pmatrix}.$$

They form an orthonormal basis of $\mathbb{C}^4$ with respect to the averaged inner product

$$\langle\, \mathbf{v} \,;\mathbf{w} \,\rangle = \frac{v_0\, \overline{w_0} + v_1\, \overline{w_1} + v_2\, \overline{w_2} + v_3\, \overline{w_3}}{4}, \qquad \text{where} \qquad \mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix}.$$

Given the sampled function values

$$f_0 = f(0), \qquad f_1 = f\big(\tfrac{1}{2}\pi\big), \qquad f_2 = f(\pi), \qquad f_3 = f\big(\tfrac{3}{2}\pi\big),$$

we construct the discrete Fourier representation

$$\mathbf{f} = c_0\, \boldsymbol{\omega}_0 + c_1\, \boldsymbol{\omega}_1 + c_2\, \boldsymbol{\omega}_2 + c_3\, \boldsymbol{\omega}_3, \tag{13.17}$$

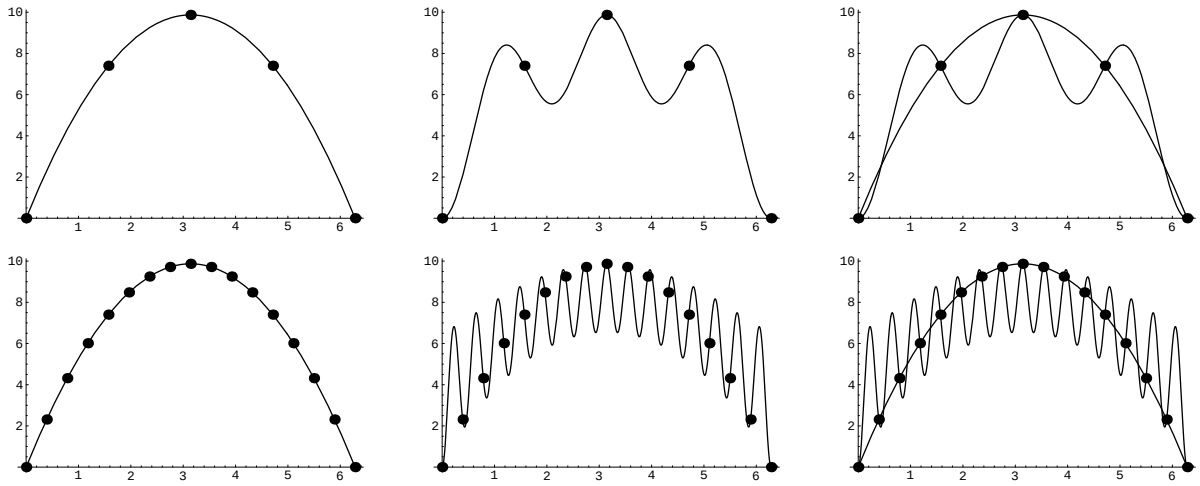


Figure 13.3. The Discrete Fourier Transform of $x^2 - 2\pi x$.

where

$$c_0 = \langle\, \mathbf{f}\,;\boldsymbol{\omega}_0\,\rangle = \tfrac{1}{4}(f_0 + f_1 + f_2 + f_3), \qquad c_1 = \langle\, \mathbf{f}\,;\boldsymbol{\omega}_1\,\rangle = \tfrac{1}{4}(f_0 - \mathrm{i}\, f_1 - f_2 + \mathrm{i}\, f_3),$$
$$c_2 = \langle\, \mathbf{f}\,;\boldsymbol{\omega}_2\,\rangle = \tfrac{1}{4}(f_0 - f_1 + f_2 - f_3), \qquad c_3 = \langle\, \mathbf{f}\,;\boldsymbol{\omega}_3\,\rangle = \tfrac{1}{4}(f_0 + \mathrm{i}\, f_1 - f_2 - \mathrm{i}\, f_3).$$

We interpret this decomposition as the sampled version of the interpolation

$$f(x) \;\sim\; p(x) = c_0 + c_1\, e^{\,\mathrm{i}\, x} + c_2\, e^{2\,\mathrm{i}\, x} + c_3\, e^{3\,\mathrm{i}\, x}$$

of the function $f(x)$ by a trigonometric polynomial, which means that the two functions agree on the sample points.

For instance if

$$f(x) = 2\pi x - x^2,$$

then

$$f_0 = 0., \qquad f_1 = 7.4022, \qquad f_2 = 9.8696, \qquad f_3 = 7.4022,$$

and hence

$$c_0 = 6.1685, \qquad c_1 = -2.4674, \qquad c_2 = -1.2337, \qquad c_3 = -2.4674.$$

Therefore, the interpolating trigonometric polynomial is given by the real part of

$$p(x) = 6.1685 - 2.4674\, e^{\,\mathrm{i}\, x} - 1.2337\, e^{2\,\mathrm{i}\, x} - 2.4674\, e^{3\,\mathrm{i}\, x}, \tag{13.18}$$

namely,

$$\mathrm{Re}\, p(x) = 6.1685 - 2.4674\, \cos x - 1.2337\, \cos 2x - 2.4674\, \cos 3x.$$

In Figure 13.3 we compare the function, with the interpolation points indicated, and discrete Fourier representations for both $n = 4$ and $n = 16$ points. The resulting graphs point out a significant difficulty with the discrete Fourier transform as developed so far. While the trigonometric polynomials do indeed correctly interpolate the sampled function values, their highly oscillatory behavior makes them completely unsuitable for interpolating away from the sample points.

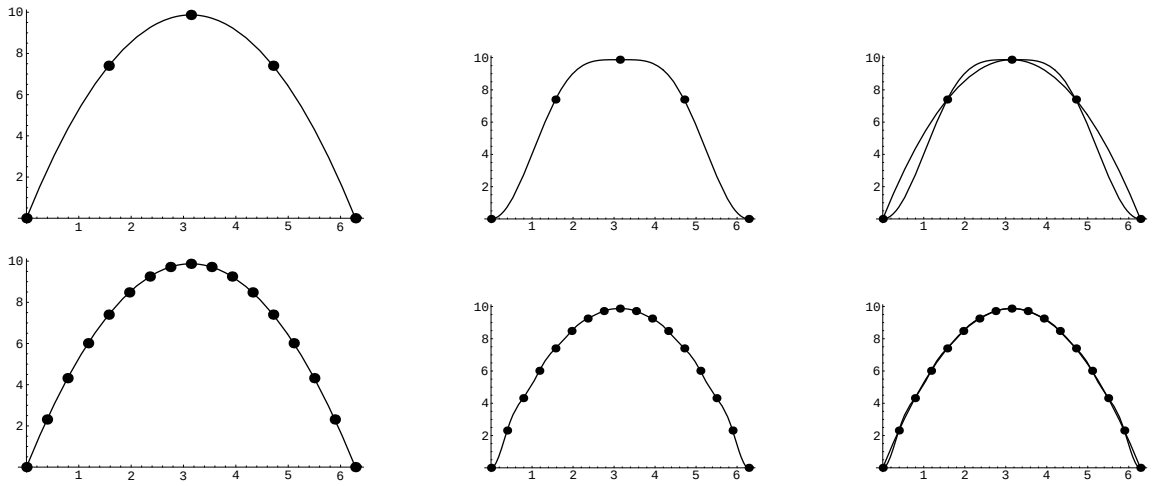


Figure 13.4. The Discrete Fourier Transform of $x^2 - 2\pi x$.

However, this difficulty can be fixed by being a little more clever. The problem is that we have not been paying sufficient attention to the frequencies represented in our Fourier sum (13.5). Indeed, the graphs in Figure 13.3 might remind the reader of our earlier observation, Figure 13.1, that low and high frequency exponentials can have the same sample data, but differ wildly in between the sample points. While the first half of the summands in (13.5) represent relatively low frequencies, the second half do not, and can be replaced by equivalent lower frequency, and hence less oscillatory exponentials. Namely, if $0 < k \leq \frac{1}{2}n$, then $e^{-\mathrm{i}\,k\,x}$ and $e^{\mathrm{i}\,(n-k)\,x}$ have the same sample values, but the former is of lower frequency than the latter. Thus, for interpolatory purposes, we should replace the second half of the summands in the Fourier sum (13.5) by their low frequency counterparts. If $n = 2m + 1$ is an odd number, then we use

$$\widehat{p}(x) \;=\; c_{-m}\,e^{-\mathrm{i}\,m\,x} + \cdots + c_{-1}\,e^{-\mathrm{i}\,x} + c_0 + c_1\,e^{\mathrm{i}\,x} + \cdots + c_m\,e^{\mathrm{i}\,m\,x} \;=\; \sum_{k=-m}^{m} c_k\,e^{\mathrm{i}\,k\,x} \tag{13.19}$$

as the low frequency interpolating polynomial. If $n = 2m$ is even — which is the most common case occurring in applications — then

$$\widehat{p}(x) \;=\; c_{-m}\,e^{-\mathrm{i}\,m\,x} + \cdots + c_{-1}\,e^{-\mathrm{i}\,x} + c_0 + c_1\,e^{\mathrm{i}\,x} + \cdots + c_{m-1}\,e^{\mathrm{i}\,(m-1)\,x} \;=\; \sum_{k=-m}^{m-1} c_k\,e^{\mathrm{i}\,k\,x} \tag{13.20}$$

will be our choice. (It is a matter of personal taste whether to use $e^{-\mathrm{i}\,m\,x}$ or $e^{\mathrm{i}\,m\,x}$ for the terms of frequency m.) In both cases, the coefficients of the negative exponentials are the same as their high frequency counterparts:

$$c_{-k} = c_{n-k} = \langle\, \mathbf{f}\,;\boldsymbol{\omega}_{n-k}\,\rangle = \langle\, \mathbf{f}\,;\boldsymbol{\omega}_{-k}\,\rangle.$$

Returning to the previous example, for interpolating purposes, we should replace (13.18) by the low frequency alternative

$$\widehat{p}(x) = -1.2337\,e^{-2\,\mathrm{i}\,x} - 2.4674\,e^{-\mathrm{i}\,x} + 6.1685 - 2.4674\,e^{\mathrm{i}\,x}, \tag{13.21}$$

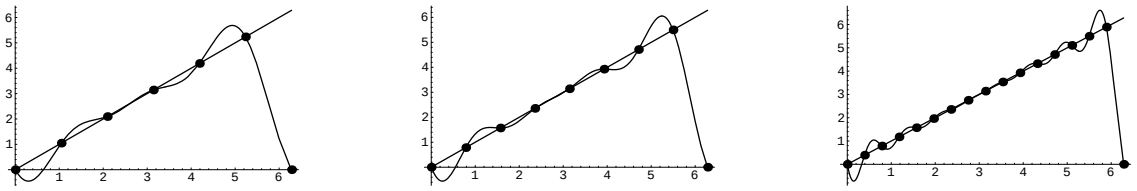


Figure 13.5. The Discrete Fourier Transform of x.

with real part

$$\operatorname{Re} \widehat{p}(x) = 6.1685 - 4.9348 \cos x - 1.2337 \cos 2x.$$

A comparison of the $n = 4$ and 16 low frequency trigonometric interpolants appears in Figure 13.4. Thus, by utilizing only the lowest frequency exponentials, we successfully construct a reasonable trigonometric interpolant to the given function.

Remark: It can be shown, [**27**], that if the function $f(x)$ is continuous, 2π–periodic, and piecewise C^1, then the low frequency trigonometric interpolating polynomials, (13.19), (13.20), converge uniformly to $f(x)$ as the number of sample points $n \to \infty$. On the other hand, if the periodic extension of $f(x)$ is discontinuous, then one observes a discrete version of the Gibbs phenomenon at the points of discontinuity. An illustration appears in Figure 13.5, which shows the trigonometric interpolants to $f(x) = x$ based on $4, 8$ and 16 sample points.

Compression and Noise Removal

In a typical experimental signal, noise has a tendency to primarily affect the high frequency modes, while the significant features tend to accumulate at low frequencies. Think of the hiss and static you hear on the radio or a low quality audio tape. Thus, a very simple, but effective, method for cleaning a noisy signal is to decompose it into its Fourier modes, as in (13.5), and then discard the high frequency constituents. The main issue is the specification of the cut-off between low and high frequency, that is, between signal and noise. This choice will depend upon the properties of the measured signal, and is left to the discretion of the signal processor.

To correctly implement such a denoising procedure, it is better to use the less oscillatory forms (13.19), (13.20) of the trigonometric interpolant, in which the low frequency summands appear when $|\, k \,|$ is small. Therefore, to eliminate high frequency components, we replace the full summation by

$$q(x) \;=\; \sum_{k=-l}^{l} c_k\, e^{\,\mathrm{i}\,k\,x} \tag{13.22}$$

where the number $l < \frac{1}{2}\,(n+1)$ specifies the cut-off point between low and high frequencies. In other words, rather than keep all the n constituents, the $2l+1 \ll n$ low frequency Fourier modes will often suffice to encapsulate a denoised version of the original signal.

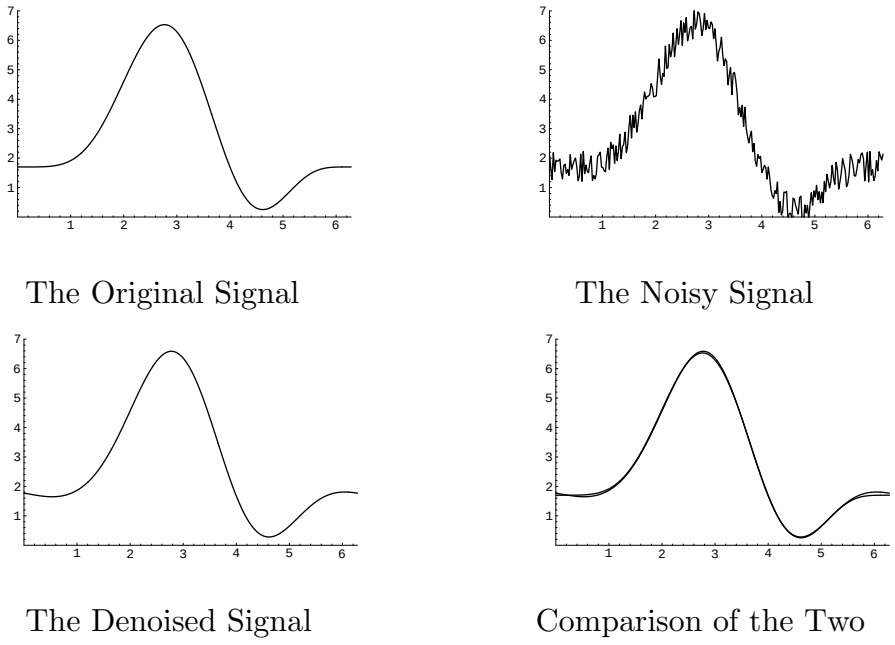


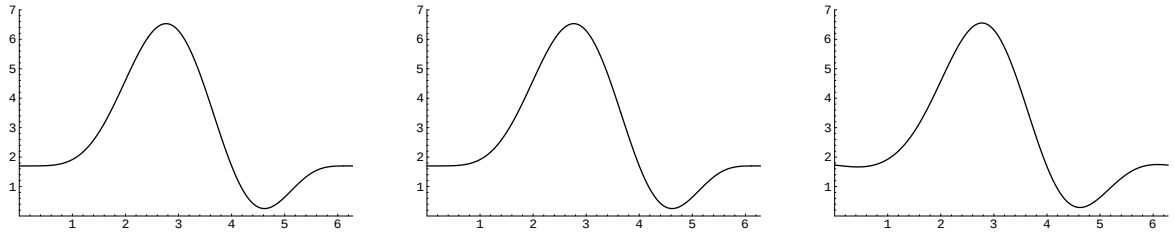
Figure 13.6. Denoising a Signal.

Remark: For the original form (13.5) of the discrete Fourier transform, the denoising algorithm will retain the $2l+1$ summands with $0 \le k \le l$ and $n-l \le k \le n-1$. The latter correspond to low frequency modes with negative exponentials $e^{\,\mathrm{i}\,k\,x} \sim e^{\,\mathrm{i}\,(k-n)\,x}$.

In Figure 13.6 we display an original signal followed by the same signal corrupted by adding in random noise. We use $n = 2^8 = 512$ sample points in the computation. To remove the noise, we retain only the $2l+1 = 11$ lowest frequencies. In other words, instead of all $n = 512$ Fourier coefficients $c_{-256}, \dots, c_{-1}, c_0, c_1, \dots, c_{255}$, we only compute the 11 lowest order ones $c_{-5}, \dots, c_5$. Summing up just those 11 exponentials, $\displaystyle\sum_{k=-5}^{5} c_k\, e^{\,\mathrm{i}\,k\,x}$, produces the cleaned signal. The final graph combines the original signal and the denoised version on the same graph. In this case, the maximal deviation between the original and the cleaned version is less than .15 over the entire interval $[\,0, 2\pi\,]$.

The same idea works in data compression. Efficient storage and transmission of audio recordings, digital images and, particularly, video often requires a compression algorithm that does not significantly alter the signal. According to Theorem 12.29, the Fourier coefficients of smooth functions tend rapidly to zero — the smoother the function, the faster the decay rate. A similar result holds in the discrete case. Thus, we expect all the important features of the signal to be contained in the low frequency constituents, and so discarding all of the small, high frequency terms will, in favorable situations, not lead to any noticeable degradation of the signal or image. Thus, to compress a signal (and, simultaneously, remove high frequency noise), we retain only its low frequency discrete Fourier coefficients (13.16). The signal is reconstructed by summing the associated truncated discrete Fourier series.

In Figure 13.7 ,the same signal is compressed by retaining, respectively, $2l+1 = 21$ and $2l+1 = 7$ Fourier coefficients only instead of all $n = 512$ that would be required for



The Original Signal Moderate Compression High Compression

Figure 13.7. Compressing a Signal.

complete accuracy. For the case of moderate compression, the maximal deviation between the original and the compressed version is less than 1.5×10^{-4} over the entire interval, while the highly compressed version is still everywhere within .05 of the original signal. Of course, the lack of any fine scale features in this particular signal means that a very high compression can be achieved — the more complicated or detailed the original signal is, the more Fourier modes need to be retained for accurate reproduction.

The Fast Fourier Transform

While we may appreciate a theoretical algorithm for its intrinsic elegance, in the real world, the bottom line is always efficiency of implementation: the less total computation, the faster the processing, and hence the more extensive the range of applications. Orthogonality is the first and most important feature of any linear algebra algorithm; were it not for the dramatic simplification afforded by the orthogonal basis formulae, Fourier analysis would not have evolved into today's essential tool. Still even these formulae have their limits when it comes to dealing with truly large scale problems such as three-dimensional medical imaging or video processing. In the early 1960's, James Cooley and John Tukey, [**36**], discovered† a much more efficient approach to the discrete Fourier transform, exploiting the rather special structure of the sampled exponential vectors. The resulting algorithm is known as the Fast Fourier Transform, often abbreviated FFT, and its discovery launched the modern revolution in digital signal and data processing.

In general, computing all the discrete Fourier coefficients (13.16) of an n times sampled signal requires a total of n^2 complex multiplications and $n^2 - n$ complex additions. Note also that each complex addition

$$z + w = (x + \mathrm{i}\, y) + (u + \mathrm{i}\, v) = (x + u) + \mathrm{i}\,(y + v)$$

requires two real additions, while each complex multiplication

$$z\, w = (x + \mathrm{i}\, y)\,(u + \mathrm{i}\, v) = (x\, u - y\, v) + \mathrm{i}\,(x\, v + y\, u)$$

requires 4 real multiplications and 2 real additions, or, by using the alternative formula

$$x\, v + y\, u = (x + y)\,(u + v) - x\, u - y\, v$$

† In fact, the key ideas can be found in Gauss' computations, but his insight was not appreciated until modern computers arrived on the scene.

for the imaginary part, 3 real multiplications and 5 real additions. Similarly, given the Fourier coefficients $c_0, \dots, c_{n-1}$, reconstruction of the sampled signal via (13.5) requires $n^2 - n$ complex multiplications and $n^2 - n$ complex additions. Both computations become quite labor intensive for large n. Extending these ideas to multi-dimensional data only exacerbates the problem.

The key observation is that if the number of sample points $n = 2m$ is even, then the square of the primitive n^{th} root of unity $\zeta_n = \sqrt[n]{1} = \sqrt[2m]{1}$ is equal to the primitive m^{th} root $\zeta_m = \sqrt[m]{1}$, so

$$\zeta_m = \zeta_n^2.$$

We use this fact to split the summation (13.16) for the order n discrete Fourier coefficients into two parts, involving the even and the odd powers of ζ_n^k:

$$\begin{aligned} c_k &= \frac{1}{n}\left[\, f_0 + f_1\,\zeta_n^{-k} + f_2\,\zeta_n^{-2\,k} + \,\cdots\, + f_{n-1}\,\zeta_n^{-(n-1)\,k}\,\right] \\ &= \frac{1}{n}\left[\, f_0 + f_2\,\zeta_n^{-2\,k} + f_4\,\zeta_n^{-4\,k} + \,\cdots\, + f_{2m-2}\,\zeta_n^{-(2\,m-2)\,k}\,\right] + \\ &\qquad + \zeta_n^{-k}\,\frac{1}{n}\left[\, f_1 + f_3\,\zeta_n^{-2\,k} + f_5\,\zeta_n^{-4\,k} + \,\cdots\, + f_{2m-1}\,\zeta_n^{-(2\,m-2)\,k}\,\right] \\ &= \frac{1}{2\,m}\left[\, f_0 + f_2\,\zeta_m^{-k} + f_4\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-2}\,\zeta_m^{-(m-1)\,k}\,\right] + \\ &\qquad + \zeta_n^{-k}\,\frac{1}{2\,m}\left[\, f_1 + f_3\,\zeta_m^{-k} + f_5\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-1}\,\zeta_m^{-(m-1)\,k}\,\right]. \end{aligned} \tag{13.23}$$

The two expressions in brackets are the order m Fourier coefficients corresponding to the sampled functions

$$\begin{aligned} \mathbf{f}^e &= \big(\, f_0, f_2, f_4, \dots, f_{2m-2}\,\big)^T = \big(\, f(x_0), f(x_2), f(x_4), \dots, f(x_{2m-2})\,\big)^T, \\ \mathbf{f}^o &= \big(\, f_1, f_3, f_5, \dots, f_{2m-1}\,\big)^T = \big(\, f(x_1), f(x_3), f(x_5), \dots, f(x_{2m-1})\,\big)^T. \end{aligned} \tag{13.24}$$

Note that $\mathbf{f}^e$ is obtained by sampling $f(x)$ on the *even* sample points x_{2j} while $\mathbf{f}^o$ is obtained by sampling the same function $f(x)$, but now at the *odd* sample points x_{2j+1}. In other words, we are splitting the original sampled signal into two "half-sampled" signals obtained by sampling on every other sample point. The even and odd Fourier coefficients are given as

$$\begin{aligned} c_k^e &\equiv \frac{1}{m}\left[\, f_0 + f_2\,\zeta_m^{-k} + f_4\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-2}\,\zeta_m^{-(m-1)\,k}\,\right], \\ c_k^o &= \frac{1}{m}\left[\, f_1 + f_3\,\zeta_m^{-k} + f_5\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-1}\,\zeta_m^{-(m-1)\,k}\,\right], \end{aligned} \qquad k = 0, \dots, m-1. \tag{13.25}$$

Both the even and odd signals only have m distinct Fourier coefficients, and we use the identification

$$c_k^e = c_{k-m}^e, \qquad c_k^o = c_{k-m}^o,$$

in formulating (13.23) when $k \geq m$. Therefore, the order $n = 2m$ discrete Fourier coefficients (13.23) can be constructed from a pair of order m discrete Fourier coefficients

via

$$c_k = \tfrac{1}{2}\big(\, c_k^e + \zeta_n^{-k}\, c_k^o \,\big), \qquad k = 0, \dots, m-1. \tag{13.26}$$

Now if $m = 2l$ is also even, then one can play the same game on the order m Fourier coefficients (13.25), reconstructing each of them from a pair of order l discrete Fourier coefficients — which are obtained by sampling the signal on every fourth sample point. If $n = 2^r$ is a power of 2, then this game can be played all the way back to the start, beginning with the trivial order 1 discrete Fourier series which just samples the function at a single point. The result is the desired algorithm. After some rearrangement of the basic steps, we arrive the Fast Fourier Transform, which we now present in its final form.

We begin with a sampled signal on $n = 2^r$ sample points. To efficiently program the algorithm, it helps to write out each index $0 \le j < 2^r$ in binary (as opposed to decimal) representation

$$j = j_{r-1}\, j_{r-2}\, \cdots\, j_2\, j_1\, j_0, \qquad \text{where} \qquad j_\nu = 0 \text{ or } 1, \tag{13.27}$$

where the notation is shorthand for its r digit binary expansion

$$j = j_0 + 2\, j_1 + 2^2\, j_2 + \ \cdots\ 2^{r-1}\, j_{r-1}.$$

We define the *bit reversal* map

$$\rho(j_{r-1}\, j_{r-2}\, \cdots\, j_2\, j_1\, j_0) = j_0\, j_1\, \cdots\, j_{r-2}\, j_{r-1}. \tag{13.28}$$

For instance, if $r = 5$, and $j = 13$, with 5 digit binary expansion 01101, then $\rho(j) = 22$ has the reversed binary expansion 10110. Note particularly that the bit reversal map depends upon the original choice of $r = \log_2 n$.

Secondly, for each $0 \le k \le r-1$, define the maps

$$\begin{aligned} \alpha_k(j) &= j_{r-1}\, \cdots\, j_{k+1}\, 0\, j_{k-1}\, \cdots\, j_0, \\ \beta_k(j) &= j_{r-1}\, \cdots\, j_{k+1}\, 1\, j_{k-1}\, \cdots\, j_0 = \alpha_k(j) + 2^k, \end{aligned} \qquad \text{for} \qquad j = j_{r-1}\, j_{r-2}\, \cdots\, j_1\, j_0. \tag{13.29}$$

In other words, $\alpha_k(j)$ makes the k^{th} binary digit of j equal to 0, while $\beta_k(j)$ makes it 1. In the preceding example, $\alpha_2(13) = 9$, with binary representation 01001, while $\beta_2(13) = 13$ with binary form 01101. The bit operations (13.28), (13.29) are especially easy to implement on modern binary computers.

Given a sampled signal $f_0, \dots, f_{n-1}$, its discrete Fourier coefficients $c_0, \dots, c_{n-1}$ are computed using the following iterative algorithm:

$$c_j^{(0)} = f_{\rho(j)}, \qquad c_j^{(k+1)} = \tfrac{1}{2}\big(\, c_{\alpha_k(j)}^{(k)} + \zeta_{2^{k+1}}^{-j}\, c_{\beta_k(j)}^{(k)} \,\big), \qquad \begin{matrix} j = 0, \dots, n-1, \\ k = 0, \dots, r-1. \end{matrix} \tag{13.30}$$

Here $\zeta_{2^{k+1}}$ is the primitive 2^{k+1} root of unity. The final output of the iterative procedure, namely

$$c_j = c_j^{(r)}, \qquad j = 0, \dots, n-1, \tag{13.31}$$

are the discrete Fourier coefficients of our signal. The preprocessing step of the algorithm, where we define $c_j^{(0)}$, produces a more convenient rearrangement of the sample values. The

subsequent steps successively combine the Fourier coefficients of appropriate even and odd sampled subsignals together, as in (13.23). The following example should help make the overall process clearer.

Example 13.3. Consider the case $r = 3$, and so our signal has $n = 2^3 = 8$ sampled values $f_0, f_1, \dots, f_7$. We begin the process by rearranging the sample values

$$c_0^{(0)} = f_0,\ \ c_1^{(0)} = f_4,\ \ c_2^{(0)} = f_2,\ \ c_3^{(0)} = f_6,\ \ c_4^{(0)} = f_1,\ \ c_5^{(0)} = f_5,\ \ c_6^{(0)} = f_3,\ \ c_7^{(0)} = f_7,$$

in the order specified by the bit reversal map ρ. For instance $\rho(3) = 6$, or, in binary notation, $\rho(011) = 110$.

The first stage of the iteration is based on $\zeta_2 = -1$. Equation (13.30) gives

$$\begin{aligned} &c_0^{(1)} = \tfrac{1}{2}(c_0^{(0)} + c_1^{(0)}), \quad c_1^{(1)} = \tfrac{1}{2}(c_0^{(0)} - c_1^{(0)}), \quad c_2^{(1)} = \tfrac{1}{2}(c_2^{(0)} + c_3^{(0)}), \quad c_3^{(1)} = \tfrac{1}{2}(c_2^{(0)} - c_3^{(0)}),\\ &c_4^{(1)} = \tfrac{1}{2}(c_4^{(0)} + c_5^{(0)}), \quad c_5^{(1)} = \tfrac{1}{2}(c_4^{(0)} - c_5^{(0)}), \quad c_6^{(1)} = \tfrac{1}{2}(c_6^{(0)} + c_7^{(0)}), \quad c_7^{(1)} = \tfrac{1}{2}(c_6^{(0)} - c_7^{(0)}), \end{aligned}$$

where we combine successive pairs of the rearranged sample values. The second stage of the iteration is based on $k = 1$ with $\zeta_4 = \mathrm{i}$. We find

$$\begin{aligned} &c_0^{(2)} = \tfrac{1}{2}(c_0^{(1)} + c_2^{(1)}), \quad c_1^{(2)} = \tfrac{1}{2}(c_1^{(1)} - \mathrm{i}\, c_3^{(1)}), \quad c_2^{(2)} = \tfrac{1}{2}(c_0^{(1)} - c_2^{(1)}), \quad c_3^{(2)} = \tfrac{1}{2}(c_1^{(1)} + \mathrm{i}\, c_3^{(1)}),\\ &c_4^{(2)} = \tfrac{1}{2}(c_4^{(1)} + c_6^{(1)}), \quad c_5^{(2)} = \tfrac{1}{2}(c_5^{(1)} - \mathrm{i}\, c_7^{(1)}), \quad c_6^{(2)} = \tfrac{1}{2}(c_4^{(1)} - c_6^{(1)}), \quad c_7^{(2)} = \tfrac{1}{2}(c_5^{(1)} + \mathrm{i}\, c_7^{(1)}). \end{aligned}$$

Note that the indices of the combined pairs of coefficient differ by 2. In the last step, based on $\zeta_8 = \frac{\sqrt{2}}{2}(1 + \mathrm{i})$, we combine coefficients whose indices differ by 4; the final output

$$\begin{aligned} c_0 &= \tfrac{1}{2}(c_0^{(3)} + c_4^{(3)}), & c_4 &= \tfrac{1}{2}(c_0^{(3)} - c_4^{(3)}),\\ c_1 &= \tfrac{1}{2}\left(c_1^{(3)} + \tfrac{\sqrt{2}}{2}(1 - \mathrm{i})\, c_5^{(3)}\right), & c_5 &= \tfrac{1}{2}\left(c_1^{(3)} - \tfrac{\sqrt{2}}{2}(1 - \mathrm{i})\, c_5^{(3)}\right),\\ c_2 &= \tfrac{1}{2}\left(c_2^{(3)} - \mathrm{i}\, c_6^{(3)}\right), & c_6 &= \tfrac{1}{2}\left(c_2^{(3)} + \mathrm{i}\, c_6^{(3)}\right),\\ c_3 &= \tfrac{1}{2}\left(c_3^{(3)} - \tfrac{\sqrt{2}}{2}(1 + \mathrm{i})\, c_7^{(3)}\right), & c_7 &= \tfrac{1}{2}\left(c_3^{(3)} + \tfrac{\sqrt{2}}{2}(1 + \mathrm{i})\, c_7^{(3)}\right), \end{aligned}$$

is the complete set of discrete Fourier coefficients.

Let us count the number of arithmetic operations required in the Fast Fourier Transform algorithm. At each stage in the computation, we must perform $n = 2^r$ complex additions/subtractions and the same number of complex multiplications. (Actually, the number of multiplications is slightly smaller since multiplication by ± 1 and $\pm \mathrm{i}$ are extremely simple. However, this does not significantly alter the final operation count.) There are $r = \log_2 n$ different stages, and so we require a total of $r\, n = n \log_2 n$ complex additions/subtractions and the same number of multiplications. Now, when n is large, $n \log_2 n$ is *significantly* smaller than n^2, which is the number of operations required for the direct algorithm. For instance, if $n = 2^{10} = 1,024$, then $n \log_2 n = 10,240$, while $n^2 = 1,048,576$. As a result, large scale computations that would be intractable using the direct approach are brought into the realm of feasibility. This is the reason why all modern implementations of the discrete Fourier transform are based on the FFT.

The reconstruction of the signal from the discrete Fourier coefficients $c_0, \dots, c_{n-1}$, as in (13.5), is speeded up in exactly the same manner. The only differences are that we

replace $\zeta_n^{-1} = \overline{\zeta_n}$ by ζ_n, and drop the factors of $\frac{1}{2}$ since there is no need to divide by n in the final result (13.5). Therefore, we apply the iterative procedure

$$f_j^{(0)} = c_{\rho(j)}, \qquad f_j^{(k+1)} = f_{\alpha_k(j)}^{(k)} + \zeta_{2^{k+1}}^{j}\, f_{\beta_k(j)}^{(k)}, \qquad \begin{array}{l} j = 0, \dots, n-1, \\ k = 0, \dots, r-1, \end{array} \tag{13.32}$$

and finishing with

$$f(x_j) = f_j = f_j^{(r)}, \qquad j = 0, \dots, n-1. \tag{13.33}$$

Example 13.4. The reconstruction formulae in the case of $n = 8 = 2^3$ Fourier coefficients $c_0, \dots, c_7$, which were computed in Example 13.3, can be implemented as follows. First, we rearrange the Fourier coefficients in bit reversed order:

$$f_0^{(0)} = c_0, \;\; f_1^{(0)} = c_4, \;\; f_2^{(0)} = c_2, \;\; f_3^{(0)} = c_6, \;\; f_4^{(0)} = c_1, \;\; f_5^{(0)} = c_5, \;\; f_6^{(0)} = c_3, \;\; f_7^{(0)} = c_7,$$

Then we begin combining them in successive pairs:

$$\begin{array}{llll} f_0^{(1)} = f_0^{(0)} + f_1^{(0)}, & f_1^{(1)} = f_0^{(0)} - f_1^{(0)}, & f_2^{(1)} = f_2^{(0)} + f_3^{(0)}, & f_3^{(1)} = f_2^{(0)} - f_3^{(0)}, \\ f_4^{(1)} = f_4^{(0)} + f_5^{(0)}, & f_5^{(1)} = f_4^{(0)} - f_5^{(0)}, & f_6^{(1)} = f_6^{(0)} + f_7^{(0)}, & f_7^{(1)} = f_6^{(0)} - f_7^{(0)}, \end{array}$$

Next,

$$\begin{array}{llll} f_0^{(2)} = f_0^{(1)} + f_2^{(1)}, & f_1^{(2)} = f_1^{(1)} + \mathrm{i}\, f_3^{(1)}, & f_2^{(2)} = f_0^{(1)} - f_2^{(1)}, & f_3^{(2)} = f_1^{(1)} - \mathrm{i}\, f_3^{(1)}, \\ f_4^{(2)} = f_4^{(1)} + f_6^{(1)}, & f_5^{(2)} = f_5^{(1)} + \mathrm{i}\, f_7^{(1)}, & f_6^{(2)} = f_4^{(1)} - f_6^{(1)}, & f_7^{(2)} = f_5^{(1)} - \mathrm{i}\, f_7^{(1)}. \end{array}$$

Finally, the sampled signal values are

$$\begin{array}{ll} f(x_0) = f_0 = f_0^{(3)} + f_4^{(3)}, & f(x_4) = f_4 = f_0^{(3)} - f_4^{(3)}, \\ f(x_1) = f_1 = f_1^{(3)} + \frac{\sqrt{2}}{2}\,(1 + \mathrm{i}\,)\, f_5^{(3)}, & f(x_5) = f_5 = f_1^{(3)} - \frac{\sqrt{2}}{2}\,(1 + \mathrm{i}\,)\, f_5^{(3)}, \\ f(x_2) = f_2 = f_2^{(3)} + \mathrm{i}\, f_6^{(3)}, & f(x_6) = f_6 = f_2^{(3)} - \mathrm{i}\, f_6^{(3)}, \\ f(x_3) = f_3 = f_3^{(3)} - \frac{\sqrt{2}}{2}\,(1 - \mathrm{i}\,)\, f_7^{(3)}, & f(x_7) = f_7 = f_3^{(3)} + \frac{\sqrt{2}}{2}\,(1 - \mathrm{i}\,)\, f_7^{(3)}. \end{array}$$

13.2. Wavelets.

Trigonometric Fourier series, both continuous and discrete, are amazingly powerful, but they do suffer from one significant defect. The basis functions $e^{\,\mathrm{i}\,k\,x} = \cos k\,x + \mathrm{i}\,\sin k\,x$ are spread out over all of the interval $[-\pi, \pi]$ and so are not well-adapted to localized signals — meaning data that are concentrated in a relatively small region. Indeed, the most concentrated data of all — a single delta function — has every Fourier component of equal magnitude in its Fourier series (12.52) and the fact that the delta function is highly localized is completely lost in its series formulation.

Ideally, one would like to construct a system of functions that is orthogonal, and so has all the advantages of the Fourier trigonometric functions, but, in addition, preserves the localized structure of signals. This dream was the inspiration for the development of the modern theory of wavelets. Just as the trigonometric functions can be discretized, each system of wavelet functions has a discrete counterpart obtained by sampling. The

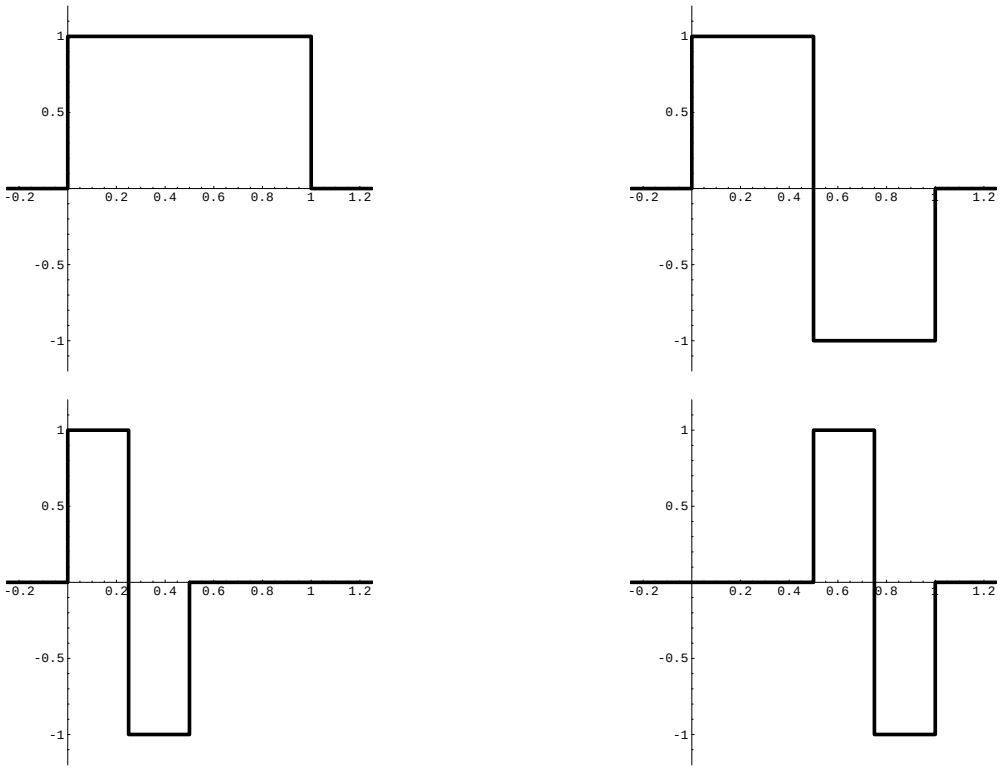


Figure 13.8. The First Four Haar Wavelets.

resulting discrete wavelet transform can be rapidly computed, which makes them ideally suited to processing complicated signals and multi-dimensional image data.

The Haar Wavelets

Let us begin with the simplest example of a wavelet basis, discovered by the Hungarian mathematician Alfréd Haar in 1910, [**70**]. We consider the space of functions (signals) defined the interval $[\,0,1\,]$, equipped with the standard L^2 inner product

$$\langle\, f\,;g\,\rangle = \int_0^1 f(x)\, g(x)\, dx. \tag{13.34}$$

This choice is merely for convenience, being slightly better suited to our construction than $[\,-\pi,\pi\,]$ or $[\,0,2\,\pi\,]$. Moreover, the usual scaling arguments can be used to adapt the wavelet formulas to any other interval. The first four *Haar wavelets* are graphed in Figure 13.8. The first, which is the constant function

$$\varphi_1(x) = \varphi(x) \equiv 1, \qquad\qquad 0 \le x \le 1,$$

is known as the *scaling function*, for reasons which shall appear shortly. The second Haar function

$$\varphi_2(x) = w(x) = \begin{cases} 1, & 0 < x < \frac{1}{2}, \\ -1, & \frac{1}{2} < x < 1, \end{cases}$$

is known as the *mother wavelet*. The value of $w(x)$ at the points of discontinuity is not very important, but for specificity we can set it to take the value $w(x) = 0$ at $x = 0, \frac{1}{2}$ and 1. The third and fourth Haar functions are compressed versions of the mother wavelet:

$$\varphi_3(x) = \begin{cases} 1, & 0 < x < \frac{1}{4}, \\ -1, & \frac{1}{4} < x < \frac{1}{2}, \\ 0 & \frac{1}{2} < x < 1, \end{cases} \qquad \varphi_4(x) = \begin{cases} 0 & 0 < x < \frac{1}{2}, \\ 1, & \frac{1}{2} < x < \frac{3}{4}, \\ -1, & \frac{3}{4} < x < 1, \end{cases}$$

the "daughter wavelets". We extend the definition of the mother wavelet to all of $\mathbb{R}$ by setting it equal to zero outside the basic interval, so

$$w(x) = \begin{cases} 1, & 0 < x < \frac{1}{2}, \\ -1, & \frac{1}{2} < x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{13.35}$$

With this convention,

$$\varphi_3(x) = w(2\,x), \qquad \varphi_4(x) = w(2\,x - 1).$$

The scaling transformation $x \mapsto 2\,x$ serves to compress the wavelet function, while the translation $2\,x \mapsto 2\,x - 1$ moves the compressed version to the right by a unit distance. Furthermore, we can represent the mother wavelet itself by compressing and translating the scaling function, which we extend off the basic interval by setting

$$\varphi(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{13.36}$$

The function $\varphi(x) = \sigma(x) - \sigma(x - 1)$ is merely a difference of two step functions, known as a *box function* due to the shape of its graph. The mother wavelet

$$w(x) = \varphi(2\,x) - \varphi(2\,x - 1) = \sigma(x) - 2\,\sigma\big(x - \tfrac{1}{2}\big) + \sigma(x - 1) \tag{13.37}$$

is, in turn, a *difference* of two compressed versions of the box function. It is these two operations of scaling and compression — coupled with the all-important orthogonality — that underlies the power of wavelets. Indeed, one can easily check by direct integration that the four Haar wavelet functions are orthogonal with respect to the L^2 inner product (13.34).

The Haar wavelets have an evident discretization. If we decompose the interval $[\,0, 1\,]$ into the four subintervals

$$\big(\,0\,,\tfrac{1}{4}\,\big), \qquad \big(\,\tfrac{1}{4}\,,\tfrac{1}{2}\,\big), \qquad \big(\,\tfrac{1}{2}\,,\tfrac{3}{4}\,\big), \qquad \big(\,\tfrac{3}{4}\,,1\,\big), \tag{13.38}$$

on which the four wavelet functions are constant, then we can represent each of them by a vector in $\mathbb{R}^4$ whose entries are the values of each wavelet function sampled on each

subinterval[†]. In this manner, we obtain the wavelet sample vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \tag{13.39}$$

that formed the orthogonal wavelet basis of $\mathbb{R}^4$ introduced in Example 2.33. Orthogonality of the vectors (13.39) with respect to the standard Euclidean dot product is equivalent to the orthogonality of the Haar wavelet functions with respect to the inner product (13.34). If

$$f(x) \sim \mathbf{f} = (f_1, f_2, f_3, f_4) \qquad \text{and} \qquad g(x) \sim \mathbf{g} = (g_1, g_2, g_3, g_4)$$

are *piecewise constant* real functions that achieve the indicated values on the four subintervals (13.38), then their L^2 inner product

$$\langle\, f\,;g\,\rangle = \int_0^1 f(x)\, g(x)\, dx = \tfrac{1}{4}\,\big(\, f_1\, g_1 + f_2\, g_2 + f_3\, g_3 + f_4\, g_4\,\big) = \tfrac{1}{4}\,\mathbf{f}\cdot\mathbf{g},$$

is equal to the averaged dot product of their sample values — the same inner product (13.9) that was used in the discrete Fourier transform.

Since the vectors (13.39) form an orthogonal basis of $\mathbb{R}^4$, we can uniquely decompose any such piecewise constant function as a linear combination of wavelets

$$f(x) = c_1\,\varphi_1(x) + c_2\,\varphi_2(x) + c_3\,\varphi_3(x) + c_4\,\varphi_4(x),$$

or, in terms of the sample vectors,

$$\mathbf{f} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4.$$

The required coefficients

$$c_k = \frac{\langle\, f\,;\varphi_k\,\rangle}{\|\,\varphi_k\,\|^2} = \frac{\mathbf{f}\cdot\mathbf{v}_k}{\|\,\mathbf{v}_k\,\|^2}$$

are computed using our usual orthogonality formula (5.8). Explicitly,

$$\begin{aligned} c_1 &= \tfrac{1}{4}\,(f_1 + f_2 + f_3 + f_4), & c_3 &= \tfrac{1}{2}\,(f_1 - f_2),\\ c_2 &= \tfrac{1}{4}\,(f_1 + f_2 - f_3 - f_4), & c_4 &= \tfrac{1}{2}\,(f_3 - f_4). \end{aligned}$$

Before proceeding to the more general case, let us introduce an important analytical definition that quantifies precisely how localized a function is.

Definition 13.5. The *support* of a function $f(x)$, written $\operatorname{supp} f$, is the closure of the set where $f(x) \neq 0$.

[†] Because the discontinuities of the Haar wavelets occur at the endpoints of the intervals, it is better to sample at the midpoints, $\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}$.

Thus, a point will belong to the support of $f(x)$, provided f is not zero there, or at least is not zero at nearby points. More precisely:

Lemma 13.6. *If $f(a) \neq 0$, then $a \in \operatorname{supp} f$. More generally, a point $a \in \operatorname{supp} f$ if and only if there exist a convergent sequence $x_n \to a$ such that $f(x_n) \neq 0$. Conversely, $a \notin \operatorname{supp} f$ if and only if $f(x) \equiv 0$ on a small interval $a - \delta < x < a + \delta$ for some $\delta > 0$.*

Intuitively, the smaller the support of a function, the more localized it is. For example, the support of the fundamental Haar wavelet (13.35) is $\operatorname{supp} w = [0, 1]$ — the points, $0, \frac{1}{2}, 1$ are included, even though $w = 0$ there, because they are limits of points where $w \neq 0$. The two daughter wavelets have smaller support

$$\operatorname{supp} \varphi_3 = \left[0\,, \tfrac{1}{2}\right], \qquad \operatorname{supp} \varphi_4 = \left[\tfrac{1}{2}\,, 1\right],$$

and so are twice as locallized. An extreme case is the delta function, whose support is a single point. In contrast, the support of the Fourier trigonometric basis functions is the entire interval $[-\pi, \pi]$, since they are zero only at isolated points.

The effect of our translation and scaling processes on the support of a function is easy to discern.

Lemma 13.7. *If* $\operatorname{supp} f = [a, b]$, *and*

$$g(x) = f(r\,x - \delta), \qquad \text{then} \qquad \operatorname{supp} g = \left[\frac{a+\delta}{r}\,, \frac{b+\delta}{r}\right].$$

Therefore, scaling x by a factor r compresses the support of the function by a factor $1/r$, while translating x translates the support of the function.

The key requirement of a wavelet basis is that it contains functions with arbitrarily small support. To this end, the full Haar wavelet basis is obtained from the mother wavelet by iterating the scaling and translation processes. We begin with the scaling function

$$\varphi(x). \tag{13.40}$$

For any nonnegative integer $j \geq 0$, we first compress the mother wavelet so that its support fits into an interval of length 2^{-j}:

$$w_{j,0}(x) = w(2^j\,x), \qquad \text{so that} \qquad \operatorname{supp} w_{j,0} = [0, 2^{-j}\,].$$

We then translate $w_{j,0}$ so as to fill up the entire interval $[0, 1]$ by 2^j subintervals, each of length 2^{-j}, by defining

$$w_{j,k}(x) = w_{j,0}(x - k) = w(2^j\,x - k), \qquad \text{where} \qquad k = 0, 1, \dots 2^j - 1.$$

Lemma 13.7 implies that $\operatorname{supp} w_{j,k} = [\,2^{-j}\,k, 2^{-j}\,(k+1)\,]$, and so the combined supports of all the daughter wavelets $w_{j,k}$ for each fixed j is the entire interval:

$$\bigcup_{k=0}^{2^j-1} \operatorname{supp} w_{j,k} = [0, 1].$$

The case $j = 0$ just consists of the mother wavelet

$$w_{0,0}(x) = w(x).$$

When $j = 1$ we pick up the next two functions already introduced as φ_3 and φ_4, namely

$$w_{1,0}(x) = w(2x), \qquad w_{1,1}(x) = w(2x-1).$$

For $j = 2$, we append four additional wavelets to our basis:

$$w_{2,0}(x) = w(4x), \quad w_{2,1}(x) = w(4x-1), \quad w_{2,2}(x) = w(4x-2), \quad w_{2,3}(x) = w(4x-3).$$

The 8 Haar wavelets $\varphi, w_{0,0}, w_{1,0}, w_{1,1}, w_{2,0}, w_{2,1}, w_{2,2}, w_{2,3}$ are constant on the 8 subintervals of length $\frac{1}{8}$, taking the successive sample values given by the columns of the matrix

$$W_8 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

As the reader can verify, the columns of W_8 are mutually orthogonal vectors. (Unfortunately, the usual terminological constraints, cf. Definition 5.18, prevent us from calling W_8 an orthogonal matrix because its columns are not orthonormal!)

At stage n there are 2^{n+1} different wavelet functions: $w_0(x) = \varphi(x)$ and $w_{j,k}(x)$ for $0 \le j \le n$ and $0 \le k < 2^j$. They are all constant on each subinterval of length 2^{-n-1}.

Theorem 13.8. *The wavelet functions $\varphi(x)$, $w_{j,k}(x)$ are orthogonal with respect to the inner product* (13.34).

Proof: First, note that each wavelet $w_{j,k}(x)$ is equal to $+1$ on an interval of length 2^{-j-1} and equal to -1 on an adjacent interval of the same length. Therefore,

$$\langle\, w_{j,k} \,;\, \varphi \,\rangle = \int_0^1 w_{j,k}(x)\, dx = 0, \tag{13.41}$$

since the $+1$ and -1 contributions cancel each other. If two different wavelets $w_{j,k}$ and $w_{l,m}$ with, say $j \le l$, have supports which are either disjoint, or just overlap at a single point, then their product $w_{j,k}(x)\, w_{l,m}(x) \equiv 0$, and so their inner product is clearly zero:

$$\langle\, w_{j,k} \,;\, w_{l,m} \,\rangle = \int_0^1 w_{j,k}(x)\, w_{l,m}(x)\, dx = 0.$$

Otherwise, except in the case when the two wavelets are identical, the support of $w_{l,m}$ is entirely contained in an interval where $w_{j,k}$ is constant and so $w_{j,k}(x)\, w_{l,m}(x) = \pm\, w_{l,m}(x)$. Therefore, by (13.41),

$$\langle\, w_{j,k} \,;\, w_{l,m} \,\rangle = \int_0^1 w_{j,k}(x)\, w_{l,m}(x)\, dx = \pm \int_0^1 w_{l,m}(x)\, dx = 0.$$

Finally, we compute

$$\| \varphi \|^2 = \int_0^1 dx = 1, \qquad \| w_{j,k} \|^2 = \int_0^1 w_{j,k}(x)^2 \, dx = 2^{-j}. \tag{13.42}$$

The second equality follows from the fact that $| w_{j,k}(x) | = 1$ on an interval of length 2^{-j} and is 0 elsewhere. *Q.E.D.*

In direct analogy with the Fourier series, the *wavelet series* of a signal $f(x)$ is given by

$$f(x) \;\sim\; c_0 \, \varphi(x) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} c_{j,k} \, w_{j,k}(x). \tag{13.43}$$

Orthogonality implies that the wavelet coefficients $c_0, c_{j,k}$ can be immediately computed using the standard inner product formula coupled with (13.42):

$$\begin{aligned} c_0 &= \frac{\langle f \, ; \varphi \rangle}{\| \varphi \|^2} = \int_0^1 f(x) \, dx, \\ c_{j,k} &= \frac{\langle f \, ; w_{j,k} \rangle}{\| w_{j,k} \|^2} = 2^j \int_{2^{-j}k}^{2^{-j}k + 2^{-j-1}} f(x) \, dx \;-\; 2^j \int_{2^{-j}k + 2^{-j-1}}^{2^{-j}(k+1)} f(x) \, dx. \end{aligned} \tag{13.44}$$

The convergence properties of the wavelet series (13.43) are similar to those of Fourier series, and will not be developed in any detail here; see [**43**].

Example 13.9. Haar wavelet example ■

Remark: There appear to be many more wavelets than trigonometric functions. But this is just another illusion in the magic show of infinite dimensional space. The point is that they both form a countably infinite set of functions, and so could, if necessary but less conveniently, be numbered in order $0, 1, 2, \ldots$. On the other hand, accurate representation of reasonable functions does require many more Haar wavelets than trigonometric functions. This motivates the search for a more sophisticated choice of wavelet basis.

Modern Wavelets

The main defect of the Haar wavelets is that they do not provide a very efficient means of representing even very simple functions — it takes quite a large number of wavelets to reproduce signals with any degree of accuracy. The reason for this is that the Haar wavelets are piecewise constant, and so even a simple affine function $y = \alpha \, x + \beta$ requires many sample values, and hence a relatively extensive collection of Haar wavelets, to be everywhere accurately reproduced. In particular, compression and denoising algorithms based on Haar wavelets are either insufficiently accurate or rather inefficient, and hence a minor practical import.

For a long time it was thought that the requirements of localization, orthogonality and accurate reproduction of simple functions were incompatible. The breakthrough came in 1988, when, in her thesis, the Dutch mathematician Ingrid Daubechies produced the first

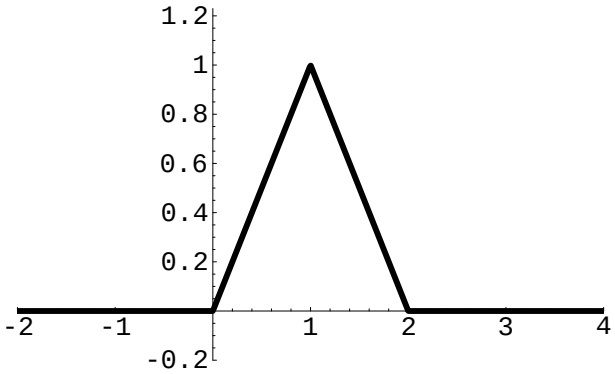


Figure 13.9. The Hat Function.

examples of wavelet bases that realized all three basic criteria. In the intervening years, wavelets has developed into a sophisticated and burgeoning industry. Significant applications include the compression of the FBI fingerprint data base, and the new JPEG2000 image format, which, unlike earlier JPEG standards which were based on Fourier methods, will incorporate wavelet technology in its image compression and reconstruction algorithms. In this section, we give a brief outline of the basic ideas underlying Daubechies' remarkable construction.

The recipe for any wavelet system involves two basic ingredients — a scaling function and a mother wavelet. The latter can be constructed from the scaling function by a prescription similar to that in (13.37), and therefore we first concentrate on the properties of the scaling function. The key requirement is that a scaling function must solve a *dilation equation* of the form

$$\varphi(x) = \sum_{k=0}^{p} c_k \, \varphi(2\,x - k) = c_0 \, \varphi(2\,x) + c_1 \, \varphi(2\,x - 1) + \, \cdots \, + c_p \, \varphi(2\,x - p) \tag{13.45}$$

for some collection of constants $c_0, \ldots, c_p$. The dilation equation relates the function to a finite linear combination of its compressed translates. The coefficients $c_0, \ldots, c_p$ are not arbitrary, since the properties of orthogonality and localization will impose certain rather stringent constraints.

Example 13.10. The Haar or box scaling function (13.36) satisfies the dilation equation (13.45) with $c_0 = c_1 = 1$, namely

$$\varphi(x) = \varphi(2\,x) + \varphi(2\,x - 1). \tag{13.46}$$

We recommend that the reader explicitly verify this identity before continuing. Another example is provided by the "hat function"

$$\varphi(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2 - x, & 1 \le x \le 2, \\ 0, & \text{otherwise}, \end{cases} \tag{13.47}$$

graphed in Figure 13.9, whose variants played a starring role in the finite element method,

cf. (11.164). The hat function satisfies

$$\varphi(x) = \tfrac{1}{2}\,\varphi(2\,x) + \varphi(2\,x-1) + \tfrac{1}{2}\,\varphi(2\,x-2), \tag{13.48}$$

which is (13.45) with $c_0 = \frac{1}{2}, c_1 = 1, c_2 = \frac{1}{2}$. Again, the reader should be able to validate this identity.

The dilation equation (13.45) is a type of "functional equation", and, as such, is not so easy to solve. Indeed, the theory of functional equations remains much less well developed than differential equations or even integral equations. Even to prove that solutions exist is a nontrivial analytical problem. Since we already know two explicit examples, let us defer the discussion of solution techniques until we understand how the dilation equation can be used to construct a wavelet basis.

Given a solution to the dilation equation, we define the *mother wavelet* to be

$$\begin{aligned} w(x) &= \sum_{k=0}^{p} (-1)^k c_{p-k}\,\varphi(2\,x-k) \\ &= c_p\,\varphi(2\,x) - c_{p-1}\,\varphi(2\,x-1) + c_{p-2}\,\varphi(2\,x-2) + \,\cdots\, \pm c_0\,\varphi(2\,x-p), \end{aligned} \tag{13.49}$$

which generalizes the Haar wavelet relation (13.37). The daughter wavelets are then all found, as in the Haar basis, by iteratively compressing and translating the mother wavelet:

$$w_{j,k}(x) = w(2^j\,x - k). \tag{13.50}$$

In the general framework, we do not necessarily restrict our attention to the interval $[\,0,1\,]$ and so $j \geq 0$ and k can, potentially, be arbitrary integers.

Let us investigate what sort of conditions should be imposed on the coefficients $c_0, \ldots, c_p$ in the dilation equation in order that we obtain a viable wavelet basis by this construction. First, localization of the wavelets requires that the scaling function has bounded support, and so $\varphi(x) \equiv 0$ when x lies outside some bounded interval $[\,a,b\,]$. If we integrate both sides of (13.45), we find

$$\int_a^b \varphi(x)\,dx = \int_{-\infty}^{\infty} \varphi(x)\,dx = \sum_{k=0}^{p} c_k \int_{-\infty}^{\infty} \varphi(2\,x-k)\,dx. \tag{13.51}$$

Now using the change of variables $y = 2\,x - k$ in the latter integrals, we find

$$\int_{-\infty}^{\infty} \varphi(2\,x-k)\,dx = \frac{1}{2}\int_{-\infty}^{\infty} \varphi(y)\,dy = \frac{1}{2}\int_a^b \varphi(x)\,dx, \tag{13.52}$$

where we revert to x as our (dummy) integration variable. We substitute this result back into (13.51). Assuming that[†] $\displaystyle\int_a^b \varphi(x)\,dx \neq 0$, , we discover that the dilation coefficients must satisfy

$$c_0 + \,\cdots\, + c_p = 2. \tag{13.53}$$

[†] This constraint holds in all standard examples.

Example 13.11. Once we impose the constraint (13.53), the very simplest version of the dilation equation is

$$\varphi(x) = 2\,\varphi(2\,x) \tag{13.54}$$

where $c_0 = 2$ is the only (nonzero) coefficient. Up to constant multiple, the only "solution" of the functional equation (13.54) with bounded support is the delta function $\delta(x)$; see Exercise ∎ for a justification. Other solutions, such as $\varphi(x) = 1/x$, are not localized, and thus not useful for constructing a wavelet basis.

The second condition we require is orthogonality of the wavelets. For simplicity, we only consider the standard L^2 inner product†

$$\langle\, f\,;g\,\rangle = \int_{-\infty}^{\infty} f(x)\,g(x)\,dx.$$

It turns out that the orthogonality of the complete wavelet system is guaranteed once we know that the scaling function $\varphi(x)$ is orthogonal to all its integer translates:

$$\langle\, \varphi(x)\,;\varphi(x-m)\,\rangle = 0 \qquad \text{for} \qquad m \neq 0. \tag{13.55}$$

We first note the formula

$$\begin{aligned}\langle\, \varphi(2\,x-k)\,;\varphi(2\,x-l)\,\rangle &= \int_{-\infty}^{\infty} \varphi(2\,x-k)\,\varphi(2\,x-l)\,dx \\ &= \frac{1}{2}\int_{-\infty}^{\infty} \varphi(x)\,\varphi(x+k-l)\,dx = \frac{1}{2}\,\langle\, \varphi(x)\,;\varphi(x+k-l)\,\rangle\end{aligned} \tag{13.56}$$

follows from using the previous change of variables $2\,x-k \mapsto x$ in the integral. Therefore, since φ satisfies the dilation equation (13.45),

$$\begin{aligned}\langle\, \varphi(x)\,;\varphi(x-m)\,\rangle &= \left\langle \sum_{j=0}^{p} c_j\,\varphi(2\,x-j)\,;\, \sum_{k=0}^{p} c_k\,\varphi(2\,x-2\,m-k)\,\right\rangle \\ &= \sum_{j,k=0}^{p} c_j\,c_k\,\langle\, \varphi(2\,x-j)\,;\varphi(2\,x-2\,m-k)\,\rangle = \frac{1}{2}\sum_{j,k=0}^{p} c_j\,c_k\,\langle\, \varphi(x)\,;\varphi(x+j-2\,m-k)\,\rangle.\end{aligned} \tag{13.57}$$

If we require orthogonality (13.55) of all the translates of φ, then the left hand side of this identity will be 0 unless $m = 0$, while only the summands with $j = 2\,m+k$ will be nonzero on the right. Therefore, orthogonality requires that

$$\sum_{0\,\le\,k\,\le\,p-2\,m} c_{2\,m+k}\,c_k = \begin{cases} 2, & m = 0, \\ 0, & m \neq 0. \end{cases} \tag{13.58}$$

Equations (13.53), (13.58) are the basic requirements for the construction of an orthogonal wavelet basis.

† In all instances, the functions have bounded support, and so the inner product integral can be reduced to an integral over a finite interval where both f and g are nonzero.

For example, if we have two nonzero coefficients c_0, c_1, then (13.53), (13.58) require

$$c_0 + c_1 = 2, \qquad c_0^2 + c_1^2 = 2,$$

and so $c_0 = c_1 = 1$ is the only solution, leading to the Haar dilation equation (13.46). If we have three coefficients c_0, c_1, c_2, then (13.53), (13.58) require

$$c_0 + c_1 + c_2 = 2, \qquad c_0^2 + c_1^2 + c_2^2 = 2, \qquad c_0\, c_2 = 0.$$

Thus either $c_2 = 0$, $c_0 = c_1 = 1$, and we are back to the Haar case, or $c_0 = 0$, $c_1 = c_2 = 1$, and the resulting dilation equation is a simple reformulation of the Haar case; see Exercise ■. In particular, the hat function (13.47) does *not* give rise to orthogonal wavelets.

The remarkable fact, discovered by Daubechies, is that there *is* a nontrivial solution for four (and, indeed, any even number) of coefficients c_0, c_1, c_2, c_3. The basic equations (13.53), (13.58) require

$$c_0 + c_1 + c_2 + c_3 = 2, \qquad c_0^2 + c_1^2 + c_2^2 + c_3^2 = 2, \qquad c_0\, c_2 + c_1\, c_3 = 0. \tag{13.59}$$

The particular values

$$c_0 = \frac{1+\sqrt{3}}{4}, \qquad c_1 = \frac{3+\sqrt{3}}{4}, \qquad c_2 = \frac{3-\sqrt{3}}{4}, \qquad c_3 = \frac{1-\sqrt{3}}{4}, \tag{13.60}$$

solve (13.59). These coefficients correspond to the *Daubechies dilation equation*

$$\varphi(x) = \frac{1+\sqrt{3}}{4}\,\varphi(2x) + \frac{3+\sqrt{3}}{4}\,\varphi(2x-1) + \frac{3-\sqrt{3}}{4}\,\varphi(2x-2) + \frac{1-\sqrt{3}}{4}\,\varphi(2x-3). \tag{13.61}$$

Any solution to this particular dilation equation whose support is contained in a bounded interval will give rise to a scaling function and an associated system of orthogonal wavelets.

Before explaining how to solve the Daubechies dilation equation, let us complete the proof of orthogonality of the wavelets. It is easy to see that, by translation invariance, since $\varphi(x)$ and $\varphi(x-m)$ are orthogonal for any $m \neq 0$, so are $\varphi(x-k)$ and $\varphi(x-l)$ for any $k \neq l$. Next we prove orthogonality of $\varphi(x-m)$ and $w(x)$:

$$\begin{aligned}
\langle\, w(x)\,;\varphi(x-m)\,\rangle &= \left\langle\, \sum_{j=0}^{p} (-1)^{j+1}\, c_j\, \varphi(2x-1+j)\,;\, \sum_{k=0}^{p} c_k\, \varphi(2x-2m-k)\,\right\rangle \\
&= \sum_{j,k=0}^{p} (-1)^{j+1}\, c_j\, c_k\, \langle\, \varphi(2x-1+j)\,;\varphi(2x-2m-k)\,\rangle \\
&= \frac{1}{2} \sum_{j,k=0}^{p} (-1)^{j+1}\, c_j\, c_k\, \langle\, \varphi(x)\,;\varphi(x-1+j-2m-k)\,\rangle,
\end{aligned}$$

using (13.56). By orthogonality (13.55) of the translates of φ, the only summands that are nonzero are when $j = 2m+k+1$; the resulting coefficient of $\|\,\varphi(x)\,\|^2$ is

$$\sum_k (-1)^{j+1}\, c_{1-2m-k}\, c_k = 0,$$

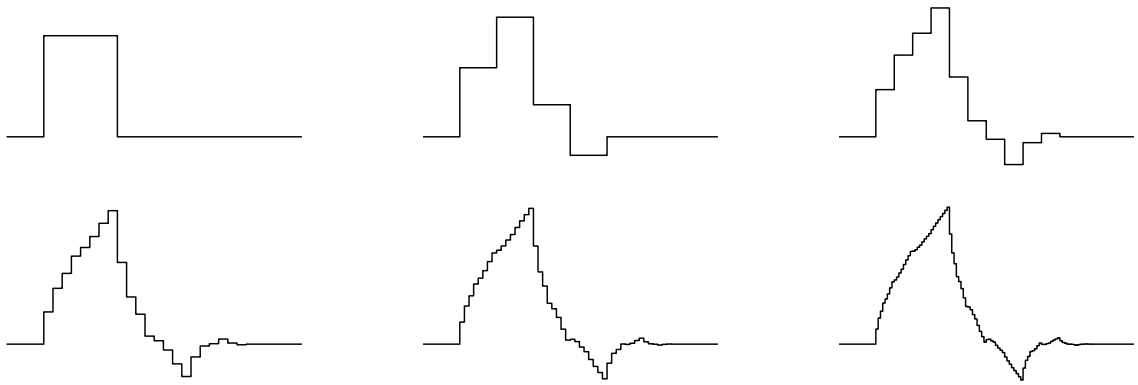


Figure 13.10. Approximating the Daubechies Wavelet.

where the sum is over all $0 \le k \le p$ such that $0 \le 1 - 2m - k \le p$. Each term in the sum appears twice, with opposite signs, and hence the result is always zero — no matter what the coefficients $c_0, \dots, c_p$ are! The proof that the translates $w(x - m)$ of the mother wavelet, along with all the daughter wavelets $w(2^j x - k)$, are orthogonal is done by a similar argument, and the details are left as an exercise for the reader.

Solving the Dilation Equation

Let us now discuss how to solve the dilation equation (13.45). The key is to observe that it has the form of a fixed point equation, but not in ordinary Euclidean space, but now in infinite-dimensional function space: $\varphi = F[\varphi]$. With luck, the solution or fixed point is stable, and so starting with a nearby initial guess $\varphi_0(x)$, the successive iterates $\varphi_{n+1} = F[\varphi_n]$ will converge to the fixed function. In detail, the iterative version of the dilation equation reads

$$\varphi_{n+1}(x) = \sum_{k=0}^{p} c_k \, \varphi_n(2x - k), \qquad n = 0, 1, 2, \dots . \tag{13.62}$$

Before attempting to prove convergence of this iterative procedure to the Daubechies scaling function, let us experimentally investigate what happens.

A good choice is to take our initial guess to be the the Haar scaling function or box function

$$\varphi_0(x) = \begin{cases} 1, & 0 < t \le 1. \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 13.10 we graph the next 5 iterates $\varphi_1(x), \dots, \varphi_5(x)$. There clearly appears to be convergence to some function $\varphi(x)$, although the final result does look a littlee bizarre. Bolstered by this experimental evidence, we can then try to prove convergence of the iterative scheme. This turns out to be true; the rigorous proof relies on the Fourier transform, [**43**], but is a little too advanced for this text and will be omitted.

Theorem 13.12. *The functions converge $\varphi_n(x)$ defined by converge uniformly to a continuous function $\varphi(x)$, called the* Daubechies scaling function.

Assuming convergence, the resulting scaling function and consequentila wavelets, in fact, for an orthonormal system of functions.

Proposition 13.13. *The integer translates $\varphi(x-k)$, for $k = 0, \pm 1, \pm 2, \dots$, of the Daubechies scaling function, and all wavelets $w_{j,k}(x) = w(2^j x - k)$, $j \geq 0$, form an orthogonal system of continuous functions in L^2; Moreover, $\| \varphi \|^2 = 1$, while $\| w_{j,k} \|^2 = 2^{-j}$.*

Proof: As noted earlier, the orthogonality of the entire wavelet system will follow once we know the orthogonality (13.55) of the scaling function and its integer translates. We use induction to prove that this holds for all the iterates $\varphi_n(x)$, and so, in view of uniform convergence, it also holds for the limiting scaling function. We already know that the orthogonality property holds for the Haar scaling function $\varphi_0(x)$. To demonstrate the induction step, we repeat the computation in (13.57), but now the left hand side is $\langle \varphi_{n+1}(x) \,;\, \varphi_{n+1}(x-m) \rangle$, while all other terms involve the previous iterate φ_n. In view of the the algebraic connstraints (13.58) on the wavelet coefficients and the induction hypothesis, we deduce that $\langle \varphi_{n+1}(x) \,;\, \varphi_{n+1}(x-m) \rangle = 0$ whenever $m \neq 0$, while when $m = 0$, $\| \varphi_{n+1} \|^2 = \| \varphi_n \|^2$. Since $\| \varphi_0 \| = 1$, we furuther conclude that all the iterates, and hence the limiting scaling function, all have unit L^2 norm. The proof of formula for the norms of the mother and daughter wavelets is left as an exercise for the reader. *Q.E.D.*

There is no explicit formula for the Daubechies scaling function. In practical computations, the limiting procedure is not so convenient, and an alternative means of computing its values is employed. The starting point is to determine what it is at integer values. First, we have[†] $\varphi_0(m) = 0$ for all integers m except $\varphi_0(1) = 1$. Now, according to the iterative equation (13.62), when $p = 3$, the value of φ_{n+1} at an integer m is a linear combination of the values of φ_n at the integers $2m, 2m-1, 2m-2$ and $2m-3$. A simple induction should convince you that $\varphi_n(m) = 0$ at all integers m except for $m = 1$ and $m = 2$; moreover, for the Daubechies coefficients (13.60)

$$\varphi_{n+1}(1) = \frac{3+\sqrt{3}}{4}\,\varphi_n(1) + \frac{1+\sqrt{3}}{4}\,\varphi_n(2), \qquad \varphi_{n+1}(2) = \frac{1-\sqrt{3}}{4}\,\varphi_n(1) + \frac{3-\sqrt{3}}{4}\,\varphi_n(2),$$

since all other terms are 0. Thus, the vectors $\mathbf{v}^{(n)} = \big(\, \varphi_n(1), \varphi_n(2) \,\big)^T$ satisfy a linear iterative system

$$\mathbf{v}^{(n+1)} = A\,\mathbf{v}^{(n)} \qquad \text{where} \qquad A = \begin{pmatrix} \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} \\ \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} \end{pmatrix}. \tag{13.63}$$

Now, according to Chapter 10, the solution to such an iterative system can be found by looking at the eigenvalues and eigenvector of the coefficient matrix, which are

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} \frac{1+\sqrt{3}}{4} \\ \frac{1-\sqrt{3}}{4} \end{pmatrix}, \qquad \lambda_2 = \tfrac{1}{2}, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

[†] To matintain consistency, we use the left hand limiting values at the points of discontinuity.

Writing the initial condition as a linear combination of the eigenvectors

$$\mathbf{v}^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2\,\mathbf{v}_1 - \frac{1-\sqrt{3}}{2}\,\mathbf{v}_2 = \begin{pmatrix} \frac{1+\sqrt{3}}{2} \\ \frac{1-\sqrt{3}}{2} \end{pmatrix} + \begin{pmatrix} \frac{1-\sqrt{3}}{2} \\ \frac{-1+\sqrt{3}}{2} \end{pmatrix};$$

Since $A\,\mathbf{v}_1 = \mathbf{v}_1, A\,\mathbf{v}_2 = \frac{1}{2}\,\mathbf{v}_2$, the solution is

$$\mathbf{v}^{(n)} = A^n\,\mathbf{v}^{(0)} = 2\,\mathbf{v}_1 - \frac{1}{2^n}\,\frac{1-\sqrt{3}}{2}\,\mathbf{v}_2.$$

The limiting vector

$$\mathbf{v}^\star = \lim_{n\to\infty} \mathbf{v}^{(n)} = 2\,\mathbf{v}_1 = \begin{pmatrix} \frac{1+\sqrt{3}}{2} \\ \frac{1-\sqrt{3}}{2} \end{pmatrix}$$

gives the desired integer values of the scaling function:

$$\varphi(1) = \frac{1-\sqrt{3}}{2} = 1.366025\ldots\,, \qquad \varphi(2) = \frac{-1+\sqrt{3}}{2} = -.366025\ldots\,, \qquad \varphi(m) = 0, \qquad m \neq 1, 2. \tag{13.64}$$

Now that we know the values of $\varphi(x)$ when $x = m$ is an integer, the Daubechies dilation equation

$$\varphi(x) = \frac{1+\sqrt{3}}{4}\,\varphi(2\,x) + \frac{3+\sqrt{3}}{4}\,\varphi(2\,x-1) + \frac{3-\sqrt{3}}{4}\,\varphi(2\,x-2) + \frac{1-\sqrt{3}}{4}\,\varphi(2\,x-3) \tag{13.65}$$

then prescribes its values at all half integers because when $x = \frac{1}{2}\,m$ then $2\,x - k = m - k$ is an integer. Once we know its values at the half integers, we can use equation (13.45) again to give its values at quarter integers. Continuing onwards, we determine the values of $\varphi(x)$ at all *dyadic points*, meaning rational numbers of the form $x = k/2^j$, i.e., those with a finite binary expansion. One can then use an interpolation scheme to approximate the values of $\varphi(x)$ at non-dyadic points. Or, since all computers are ultimately based on the binary number system, only dyadic points actually exist in a computer's memory, and so there is no need to determine the value of φ at non-dyadic points. Indeed, any real number can be approximated arbitrarily closely by a dyadic number — just truncate its binary expansion sufficiently far beyond the decimal (or, rather "binary") point, which means that we can approximate $\varphi(x)$ at all points of continuity by computing its values at nearby dyadic points. This scheme was used to produce the graphs of the Daubechies scaling function in Figure 13.11. It is continuous, but non-differentiable function — and has a very jagged, fractal-like appearance when viewed closely. The Daubechies scaling function is, in fact, a close cousin of the famous example of a continuous, nowhere differentiable function originally due to Weierstrass, cf. [**102**, **125**], whose construction also relies on a similar scaling argument.

With the values of the Daubechies scaling function on a sufficiently dense set of dyadic points in hand, the consequential values of the mother wavelet are given by (13.49), which,

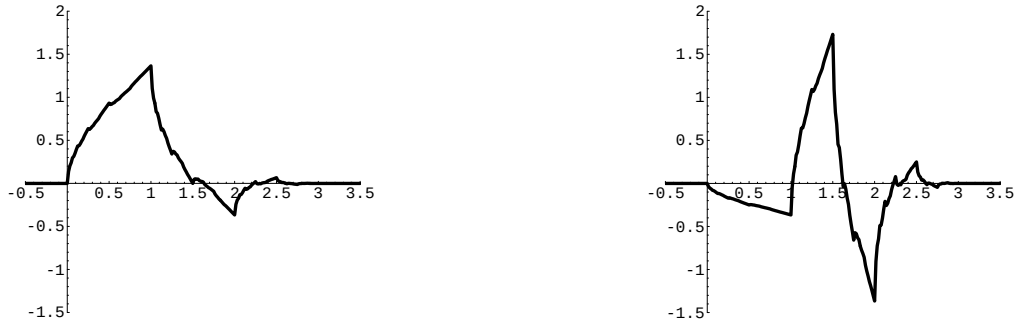


Figure 13.11. The Daubechies Scaling Function and Mother Wavelet.

in this instance, has the form

$$w(x) = \frac{1-\sqrt{3}}{4}\,\varphi(2x) - \frac{3-\sqrt{3}}{4}\,\varphi(2x-1) + \frac{3+\sqrt{3}}{4}\,\varphi(2x-2) - \frac{1+\sqrt{3}}{4}\,\varphi(2x-3). \tag{13.66}$$

Note that $\operatorname{supp}\varphi = \operatorname{supp} w = [0,3]$. The daughter wavelets are then found by the usual compression and translation formula (13.50).

The Daubechies wavelet expansion of a function whose support is contained in† $[0,3]$ is then given by

$$f(x) \;\sim\; c_0\,\varphi(x) + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} c_{j,k}\,w_{j,k}(x). \tag{13.67}$$

The wavelet coefficients $c_0, c_{j,k}$ are computed by the usual orthonormality formula

$$\begin{aligned} c_0 &= \langle\, f\,;\varphi\,\rangle = \int_0^3 f(x)\,\varphi(x)\,dx, \\ c_{j,k} &= \langle\, f\,; w_{j,k}\,\rangle = 2^j \int_{2^{-j}\,k}^{2^{-j}\,(k+3)} f(x)\,w_{j,k}(x)\,dx = \int_0^3 f\big(\,2^{-j}\,(x+k)\,\big)\,w(x)\,dx. \end{aligned} \tag{13.68}$$

In practice, one uses a basic numerical integration routine, e.g., the trapezoid rule, [**9**], with dyadic nodes to speedily evaluate the integrals (13.68). Proof of completeness of the resulting wavelet basis functions can be found in [**43**]. Compression and denoising algorithms based on retaining only low frequency modes proceed as before, and are left for the reader to implement.

Example 13.14. Daubechies wavelet example ■

13.3. The Fourier Transform.

Fourier series and their generalizations were originally designed to solve boundary value problems on bounded intervals. The extension of Fourier methods to boundary value

† For functions with larger support, one should include additional terms in the expansion corresponding to further translates of the wavelets so as to cover the entire support of the function. Alternatively, one can translate and rescale x to fit the function's support inside $[0,3]$.

problems on the entire real line $-\infty < x < \infty$ leads naturally to the *Fourier transform*. The Fourier transform is a powerful tool for the analysis of general functions, and plays an essential role in a broad range of applications, including both ordinary and partial differential equations, quantum mechanics, signal processing, control theory, and many others areas of both pure and applied mathematics. To mathematicians, the Fourier transform is more fundamental than the Laplace transform, which, as discussed in the following section, is a particular real form.

We begin by motivating the Fourier transform as a limiting case of Fourier series. Although the rigorous details can be quite intricate, the underlying idea is fairly simple. One begins with Fourier series for a fixed function $f(x)$ on progressively larger intervals, and then takes the limit as the intervals' length becomes infinite. The result is a "Fourier series" for $f(x)$ on the entire real line. The limiting process converts the Fourier sums into integrals, and the resulting representation of a function is renamed the Fourier transform. Since we are dealing with an infinite interval, there are no longer any periodicity requirements on the function $f(x)$, which is defined for all $-\infty < x < \infty$. Moreover, the frequencies represented in the Fourier transform are no longer dictated by the length of the interval, and so we are effectively decomposing a quite general, non-periodic function, into a "continuous linear combination" or, rather, integral, of trigonometric functions of arbitrary frequency.

Let us present the details of this construction in a more concrete form. The computations will be significantly simpler if we work with the complex version of the Fourier series from the outset. The starting point is the rescaled Fourier series (12.73) on a symmetric interval $[\,-\ell\,,\ell\,]$ of length $2\,\ell$, which we rewrite in the following adapted form

$$f(x) \quad \sim \quad \sum_{\nu=-\infty}^{\infty} \frac{\widehat{f}(k_\nu)}{2\,\ell}\, e^{2\,\pi\,\mathrm{i}\,k_\nu x}. \tag{13.69}$$

The sum is over the discrete collection of frequencies

$$2\,\pi\,k_\nu = \frac{\pi\,\nu}{\ell}, \qquad\qquad \nu = 0, \pm 1, \pm 2, \dots\,, \tag{13.70}$$

corresponding to all trigonometric functions of period $2\,\ell$. For reasons that will soon become apparent, the Fourier coefficients of f are now denoted as

$$c_\nu = \langle\, f\,; e^{2\,\pi\,\mathrm{i}\,k_\nu\,x}\,\rangle = \frac{1}{2\,\ell}\int_{-\ell}^{\ell} f(x)\, e^{-\,2\,\pi\,\mathrm{i}\,k_\nu\,x}\,dx = \frac{\widehat{f}(k_\nu)}{2\,\ell},$$

so that

$$\widehat{f}(k_\nu) = \int_{-\ell}^{\ell} f(x)\, e^{-\,2\,\pi\,\mathrm{i}\,k_\nu\,x}\,dx. \tag{13.71}$$

This reformulation of the basic Fourier series formula allows us to immediately pass to the limit $\ell \to \infty$.

For an interval of length ℓ, the frequencies (13.70) required to represent a function are equally spaced, with

$$\Delta k = k_{\nu+1} - k_\nu = \frac{1}{2\,\ell}\,.$$

As $\ell \to \infty$, the interfrequency spacing $\Delta k \to 0$, and so the required frequencies become more and more densely packed in the space of all possible frequencies $-\infty < k < \infty$. In the limit, we anticipate *all* possible frequencies to be represented. The resulting infinite integral

$$\widehat{f}(k) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi \mathrm{i}\, k x}\, dx \tag{13.72}$$

is known as the *Fourier transform*, and will be well-defined for a wide range of functions.

In order to reconstruct the function from its Fourier transform, we use a similar limiting procedure on the Fourier series (13.69), which we rewrite in the even more suggestive form

$$f(x) \quad \sim \quad \sum_{\nu=-\infty}^{\infty} \widehat{f}(k_\nu)\, e^{2\pi \mathrm{i}\, k_\nu x}\, \Delta k. \tag{13.73}$$

The right hand side takes the form of a Riemann sum, [**9**, **126**], over the entire frequency space $-\infty < k < \infty$, for the function

$$g(k) = \widehat{f}(k)\, e^{\mathrm{i}\, k x}.$$

Under reasonable hypotheses, the Riemann sums converge, in the limit $\Delta k \to 0$, to the integral

$$f(x) \quad \sim \quad \int_{-\infty}^{\infty} \widehat{f}(k)\, e^{2\pi \mathrm{i}\, k x}\, dk. \tag{13.74}$$

In this manner, the Fourier series (13.73) becomes a Fourier integral that reconstructs the function $f(x)$ as a (continuous) superposition of complex exponentials $e^{2\pi \mathrm{i}\, k x}$ over *all* possible frequencies $2\pi k$. The coefficient or weight of each such exponential is given by the Fourier transform $\widehat{f}(k)$.

Remark: There are several different versions of the Fourier transform in the literature, and so the reader needs to pay very careful attention as to which convention is being used when consulting any particular reference. We have chosen to adopt a modern convention that is gaining popularity, [**90**]. Including an extra 2π factor in the frequency variable k has the desirable effect of eliminating many later occurrences of 2π that tend to plague the classical forms of the Fourier transform. The other common versions, and their advantages and disadvantages, are discussed in the exercises at the end of the section.

Recapitulating, by letting the length of the interval go to ∞, the discrete Fourier series has become a continuous Fourier integral, while the Fourier coefficients, which were defined only at a discrete collection of possible frequencies, have become an entire function $\widehat{f}(k)$ defined on all of frequency space $k \in \mathbb{R}$, known as the Fourier transform of the function $f(x)$. The reconstruction of $f(x)$ from its Fourier transform via (13.74) can be rigorously justified under suitable hypotheses. For example, if $f(x)$ is piecewise C^1 on $-\infty < x < \infty$ and decays reasonably fast as $|\, x\,| \to \infty$, then the inverse Fourier integral will converge to $f(x)$ at all points of continuity, and to the midpoint $\frac{1}{2}(f(x^-) + f(x^+))$ at

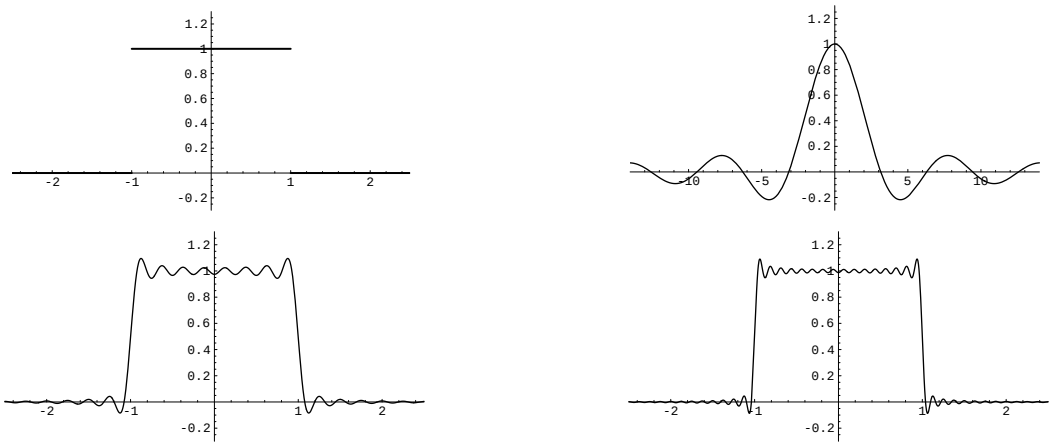


Figure 13.12. Fourier Transform of Rectangular Pulse.

jump discontinuities — just like a Fourier series. The decay conditions are assured if, for instance, f satisfies

$$| f(x) | \leq \frac{M}{| x |^{1/2+\delta}}, \qquad \text{for all sufficiently large} \qquad | x | \gg 0, \tag{13.75}$$

for some $M > 0$, $\delta > 0$. The decay conditions are the only remnant of the original periodicity requirement for the Fourier series. A more precise, general result will be formulated in Theorem 13.27 below.

Example 13.15. The Fourier transform of a rectangular pulse or box function

$$f(x) = \sigma(x+a) - \sigma(x-a) = \begin{cases} 1, & -a < x < a, \\ 0, & | x | > a. \end{cases} \tag{13.76}$$

of width $2a$ is easily computed:

$$\widehat{f}(k) = \int_{-a}^{a} e^{-2\pi \mathrm{i} k x}\, dx = \frac{e^{2\pi \mathrm{i} k a} - e^{-2\pi \mathrm{i} k a}}{\mathrm{i}\, k} = \frac{2 \sin 2\pi k a}{k}\,. \tag{13.77}$$

On the other hand, the reconstruction of the pulse via the inverse transform (13.74) tells us that

$$2 \int_{-\infty}^{\infty} \frac{e^{2\pi \mathrm{i} k x} \sin 2\pi k a}{k}\, dk = f(x) = \begin{cases} 1, & -a < x < a, \\ \frac{1}{2}\,, & x = \pm a, \\ 0, & | x | > a. \end{cases} \tag{13.78}$$

Note the convergence to the middle of the jump discontinuities at $x = \pm a$. Splitting this complex integral into its real and imaginary parts, we deduce a pair of interesting real integral identities

$$\int_{-\infty}^{\infty} \frac{\cos 2\pi k x \, \sin 2\pi k a}{k}\, dk = \begin{cases} \frac{1}{2}\,, & -a < x < a, \\ \frac{1}{4}\,, & x = \pm a, \\ 0, & | x | > a, \end{cases} \qquad \int_{-\infty}^{\infty} \frac{\sin 2\pi k x \, \sin 2\pi k a}{k}\, dk = 0. \tag{13.79}$$

Just as many Fourier series yield nontrivial summation formulae, the reconstruction of a function from its Fourier transform often leads to nontrivial integral identities. One *cannot* compute the integral (13.78) by the Fundamental Theorem of Calculus, since is no elementary function[†] whose derivative equals the integrand. Moreover, it is not even clear that the integral converges; indeed, the amplitude of the integrand decays like $1/|\,k\,|$, but the latter function does not have a convergent integral, and so the usual comparison test for infinite integrals, [**9**, **38**, **136**], fails to apply. Thus the convergence of the integral is marginal at best: the trigonometric oscillations somehow overcome the slow rate of decay of $1/k$ and thereby induce the (conditional) convergence of the integral! In Figure 13.12 we display the box function, its Fourier transform, along with a reconstruction obtained by numerically integrating (13.79). Since we are dealing with an infinite integral, we must break off the numerical integrator by restriction to a finite interval. The first graph is obtained by integrating from $-25 < k < 25$ while the second is from $-50 < k < 50$. The non-uniform convergence of the integral leads to the appearance of a Gibbs phenomenon at the two discontinuities, just as with the Fourier series.

It is also worth pointing out that both the Fourier transform (13.72) and its inverse (13.74) define linear maps on function space. This means that if $f(x)$ has Fourier transform $\widehat{f}(k)$, while $g(x)$ has Fourier transform $\widehat{g}(k)$, then the Fourier transform of the linear combination $c\,f(x) + d\,g(x)$ is $c\,\widehat{f}(k) + d\,\widehat{g}(k)$, for any complex constants c, d.

Example 13.16. Consider an exponentially decaying pulse[†]

$$f_r(x) = \begin{cases} e^{-a\,x}, & x > 0, \\ 0, & x < 0, \end{cases} \tag{13.80}$$

where $a > 0$. We compute

$$\widehat{f}(k) = \int_0^\infty e^{-a\,x}\, e^{-2\,\pi\,\mathrm{i}\,k\,x}\, dx = -\,\frac{e^{-(a+2\,\pi\,\mathrm{i}\,k)x}}{a + 2\,\pi\,\mathrm{i}\,k}\,\Big|_{x=0}^{\infty} = \frac{1}{a + 2\,\pi\,\mathrm{i}\,k}\,.$$

As in the preceding example, the inverse Fourier transform for this function produces a nontrivial complex integral identity:

$$\int_{-\infty}^{\infty} \frac{e^{2\,\pi\,\mathrm{i}\,k\,x}}{a + 2\,\pi\,\mathrm{i}\,k}\, dk = \begin{cases} e^{-a\,x}, & x > 0, \\ \frac{1}{2}\,, & x = 0, \\ 0, & x < 0. \end{cases} \tag{13.81}$$

Similarly, a pulse that decays to the left

$$f_l(x) = \begin{cases} e^{a\,x}, & x < 0, \\ 0, & x > 0, \end{cases} \tag{13.82}$$

[†] One can use Euler's formula (3.76) to reduce the integrand to one of the form $e^{\alpha\,k}/k$, but it can be proved that there is no formula for $\int (e^{\alpha\,k}/k)\, dk$ in terms of elementary functions.

[†] Note that we can't Fourier transform the entire exponential function $e^{-a\,x}$ because it does not go to zero at both $\pm\infty$, which is required for the integral (13.72) to converge.

where $a > 0$ is still positive, has Fourier transform

$$\widehat{f}_l(k) = \frac{1}{a - 2\pi\,\mathrm{i}\,k}\,. \tag{13.83}$$

This also follows from the general fact that the Fourier transform of $f(-x)$ is $\widehat{f}(-k)$; see Exercise ■. The bidirectional decaying pulse

$$f_e(x) = e^{-a\,|\,x\,|} \tag{13.84}$$

is merely the sum of left and right pulses: $f_e = f_r + f_l$. Thus, by linearity,

$$\widehat{f}_e(k) = \widehat{f}_r(k) + \widehat{f}_l(k) = \frac{1}{a + 2\pi\,\mathrm{i}\,k} + \frac{1}{a - 2\pi\,\mathrm{i}\,k} = \frac{2a}{4\pi^2\,k^2 + a^2}\,, \tag{13.85}$$

which is real and even because f_e is an even function; see Exercise ■. The inverse Fourier transform (13.74) yields another nontrivial integral identity:

$$e^{-a\,|\,x\,|} = \int_{-\infty}^{\infty} \frac{2a\,e^{2\pi\,\mathrm{i}\,k\,x}}{4\pi^2\,k^2 + a^2}\,dk = \int_{-\infty}^{\infty} \frac{2a\cos 2\pi\,k\,x}{4\pi^2\,k^2 + a^2}\;dk. \tag{13.86}$$

The imaginary part of the integral vanishes because the integrand is odd. On the other hand, the odd exponentially decaying pulse,

$$f_o(x) = \begin{cases} e^{-a\,x}, & x > 0, \\ -\,e^{a\,x}, & x < 0, \end{cases} \tag{13.87}$$

is the difference of the right and left pulses, $f_o = f_r - f_l$, and has purely imaginary and odd Fourier transform

$$\widehat{f}_o(k) = \widehat{f}_r(k) - \widehat{f}_l(k) = \frac{1}{a + 2\pi\,\mathrm{i}\,k} \;-\; \frac{1}{a - 2\pi\,\mathrm{i}\,k} = -\,\mathrm{i}\;\frac{4\pi\,k}{4\pi^2\,k^2 + a^2}\,. \tag{13.88}$$

The inverse transform is

$$(\operatorname{sgn} x)e^{-a\,|\,x\,|} = -\,4\pi\,\mathrm{i}\;\int_{-\infty}^{\infty} \frac{k\,e^{2\pi\,\mathrm{i}\,k\,x}}{4\pi^2\,k^2 + a^2}\,dk = 4\pi\;\int_{-\infty}^{\infty} \frac{k\sin 2\pi\,k\,x}{4\pi^2\,k^2 + a^2}\;dk. \tag{13.89}$$

As a final example, consider the rational function

$$f(x) = \frac{1}{x^2 + c^2}\,, \qquad \text{where} \qquad c > 0. \tag{13.90}$$

Its Fourier transform requires integrating

$$\widehat{f}(k) = \int_{-\infty}^{\infty} \frac{e^{-2\pi\,\mathrm{i}\,k\,x}}{x^2 + c^2}\;dx. \tag{13.91}$$

The indefinite integral does not appear in integration tables, and, in fact, cannot be done in terms of elementary functions. However, we have just evaluated this particular integral! Look at (13.86). Changing x to $-k$ and k to x yields

$$e^{-a\,|\,k\,|} = \int_{-\infty}^{\infty} \frac{2a\,e^{-2\pi\,\mathrm{i}\,k\,x}}{4\pi^2\,x^2 + a^2}\,dx = \frac{a}{2\pi^2}\;\int_{-\infty}^{\infty} \frac{e^{-2\pi\,\mathrm{i}\,k\,x}}{x^2 + (a/2\pi)^2}\;dx.$$

Identifying $c = a/2\pi$ leads to the formula for (13.91), and we conclude that the Fourier transform of (13.90) is

$$\widehat{f}(k) = \frac{\pi}{c}\, e^{-2\pi c|k|}. \tag{13.92}$$

This last example is indicative of an important general fact. The reader has no doubt already noted the remarkable similarity between the Fourier transform (13.72) and its inverse (13.74). Indeed, the only difference is that the former has a minus sign in the exponential. Let us state a formal result.

Theorem 13.17. *If the Fourier transform of the function $f(x)$ is $\widehat{f}(k)$, then the Fourier transform of $\widehat{f}(x)$ is $f(-k)$.*

The symmetry property allows us to reduce the tabulation of Fourier transforms by half. For instance, referring back to Example 13.15, we deduce that the Fourier transform of the function $f(x) = \dfrac{2\sin 2\pi a x}{x}$ is

$$\widehat{f}(k) = \sigma(-k+a) - \sigma(-k-a) = \sigma(k+a) - \sigma(k-a) = \begin{cases} 1, & -a < k < a, \\ \frac{1}{2}, & k = \pm a, \\ 0, & |k| > a. \end{cases}$$

Replacing $2\pi a$ by λ and dividing by 2, we deduce that

$$f(x) = \frac{\sin \lambda x}{x} \quad \text{has Fourier transform} \quad \widehat{f}(k) = \frac{1}{2}\left[\sigma\left(k + \frac{\lambda}{2\pi}\right) - \sigma\left(k - \frac{\lambda}{2\pi}\right)\right], \tag{13.93}$$

a useful result that we cannot obtain by direct evaluation of the integral.

All of the Examples 13.16 required $a > 0$ for the Fourier integrals to converge. The functions that emerge in the limit as a goes to 0 are of fundamental importance. Let us start with the odd exponential pulse (13.87). In the limit $a \to 0$, the function $f_o(x)$ converges to the *sign function*

$$f(x) = \operatorname{sgn} x = \sigma(x) - \sigma(-x) = \begin{cases} +1, & x > 0, \\ -1, & x < 0. \end{cases} \tag{13.94}$$

Taking the limit of the Fourier transform (13.88) leads to

$$\widehat{f}(k) = -\,\mathrm{i}\,\frac{4\pi k}{4\pi^2 k^2} = -\,\frac{\mathrm{i}}{\pi k}\,. \tag{13.95}$$

The nonintegrable singularity of $\widehat{f}(k)$ at $k = 0$ is indicative of the fact that the sign function does *not* decay as $|x| \to \infty$. In this case, neither the Fourier transform integral nor its inverse are well-defined as standard (Riemann, or even Lesbesgue, [**126**]) integrals. Nevertheless, it is possible to rigorously justify these results within the framework of generalized functions.

More interesting are the even pulse functions $f_e(x)$, which, in the limit $a \to 0$, become the constant function

$$f(x) \equiv 1. \tag{13.96}$$

The limit of the Fourier transform (13.85) is

$$\lim_{a \to 0} \frac{2\,a}{4\,\pi^2\,k^2 + a^2} = \begin{cases} 0, & k \neq 0, \\ \infty, & k = 0. \end{cases} \tag{13.97}$$

This limiting behavior should remind the reader of our construction (11.32) of the delta function as the limit of the functions

$$\delta(x) = \lim_{n \to \infty} \frac{n}{\pi\,(1 + n^2\,x^2)} = \lim_{a \to 0} \frac{a}{\pi\,(a^2 + x^2)}\,, \qquad \text{where} \qquad n = \frac{1}{a}\,.$$

Comparing with (13.97), we conclude that the Fourier transform of the constant function (13.96) is a multiple of the delta function in the frequency variable:

$$\widehat{f}(k) = 2\,\pi\,\delta(2\,\pi\,k) = \delta(k), \tag{13.98}$$

where we used Exercise ■ to simplify the final formula. Thus, the Fourier transform of a constant function $f(x) \equiv c$ is the same constant multiple, $\widehat{f}(k) = c\,\delta(k)$, of the delta function in frequency space!

We remark that the direct transform integral

$$\delta(k) = \int_{-\infty}^{\infty} e^{-2\,\pi\,\mathrm{i}\,k\,x}\,dx = \int_{-\infty}^{\infty} \big(\cos 2\,\pi\,k\,x + \mathrm{i}\,\sin 2\,\pi\,k\,x\,\big)\,dx$$

is, strictly speaking, not defined because the infinite integral of the oscillatory sine and cosine functions doesn't converge! However, this identity can be validly interpreted within the framework of generalized functions. On the other hand, the inverse transform formula (13.74) yields

$$\int_{-\infty}^{\infty} \delta(k)\,e^{2\,\pi\,\mathrm{i}\,k\,x}\,dk = e^{2\,\pi\,\mathrm{i}\,k\,0} = 1,$$

which is in accordance with the basic definition of the delta function. As in the previous case, the delta function singularity at $k = 0$ reflects the lack of decay of the constant function.

Conversely, the delta function $\delta(x)$ has unit Fourier transform

$$\widehat{\delta}(k) = \int_{-\infty}^{\infty} \delta(x)\,e^{-2\,\pi\,\mathrm{i}\,k\,x}\,dx = e^{-2\,\pi\,\mathrm{i}\,k\,0} = 1\,. \tag{13.99}$$

To determine the Fourier transform of a delta spike $\delta_y(x) = \delta(x - y)$ concentrated at position $x = y$, we compute

$$\widehat{\delta}_y(k) = \int_{-\infty}^{\infty} \delta(x - y)\,e^{-2\,\pi\,\mathrm{i}\,k\,x}\,dx = e^{-2\,\pi\,\mathrm{i}\,k\,y}. \tag{13.100}$$

The result is a pure exponential in frequency space. Applying the inverse Fourier transform (13.74) leads, formally, to the remarkable identity

$$\delta_y(x) = \delta(x - y) = \frac{1}{2\,\pi} \int_{-\infty}^{\infty} e^{-2\,\pi\,\mathrm{i}\,k\,(x-y)}\,dk = \langle\, e^{2\,\pi\,\mathrm{i}\,k\,y}\,; e^{2\,\pi\,\mathrm{i}\,k\,x}\,\rangle\,. \tag{13.101}$$

Since the delta function vanishes for $x \neq y$, this identity implies that the complex exponential functions are mutually orthogonal. However, the latter statement must be taken with a grain of salt, since the integral does not converge in the normal (Riemann or Lebesgue) sense. But it is possible to make sense of this identity within the language of generalized functions. Indeed, multiplying both sides by $f(x)$, and then integrating with respect to x, we find

$$f(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)\, e^{-2\pi \mathrm{i}\, k(x-y)}\, dx\, dk. \tag{13.102}$$

This *is* a perfectly valid formula, being a restatement (or, rather, combination) of the basic formulae (13.72), (13.74) connecting the direct and inverse Fourier transforms of the function $f(x)$.

Conversely, the Symmetry Theorem 13.17 tells us that the Fourier transform of a pure exponential $e^{2\pi \mathrm{i}\, l x}$ will be a shifted delta spike $\delta(k-l)$, concentrated in frequency space. Both results are particular cases of the general Shift Theorem, whose proof is left as an exercise for the reader.

Proposition 13.18. *If $f(x)$ has Fourier transform $\widehat{f}(k)$, then the Fourier transform of the shifted function $f(x-y)$ is $e^{-2\pi \mathrm{i}\, k y}\widehat{f}(k)$. Similarly, the transform of the product function $e^{2\pi \mathrm{i}\, l x} f(x)$ is the shifted transform $\widehat{f}(k-l)$.*

Derivatives and Integrals

One of the most remarkable and important properties of the Fourier transform is that it converts calculus into algebra! More specifically, the two basic operations in calculus — differentiation and integration of functions — are realized as algebraic operations on their Fourier transforms. (On the other hand, algebraic operations become more complicated in the transform domain.)

Let us begin with derivatives. If we differentiate[†] the basic inverse Fourier transform formula

$$f(x) \quad \sim \quad \int_{-\infty}^{\infty} \widehat{f}(k)\, e^{2\pi \mathrm{i}\, k x}\, dk.$$

with respect to x, we obtain

$$f'(x) \quad \sim \quad \int_{-\infty}^{\infty} 2\pi \mathrm{i}\, k\, \widehat{f}(k)\, e^{2\pi \mathrm{i}\, k x}\, dk. \tag{13.103}$$

The resulting integral is itself in the form of a Fourier transform, which immediately implies the following basic result, which is analogous to our earlier rule, (12.66), for differentiating a Fourier series.

[†] We are assuming the integrand is sufficiently nice so that we can move the derivative past the integral sign; see [**51**, **159**] for a fully rigorous derivation.

Proposition 13.19. *The Fourier transform of the derivative $f'(x)$ of a function is obtained by multiplication of its Fourier transform by $2\pi\,\mathrm{i}\,k$:*

$$\widehat{f\,'}(k) = 2\pi\,\mathrm{i}\,k\,\widehat{f}(k). \tag{13.104}$$

Example 13.20. The derivative of the even exponential pulse $f_e(x) = e^{-a|x|}$ is **a**multiple of the odd exponential pulse $f_o(x) = \operatorname{sgn} x\, e^{-a|x|}$:

$$f_e'(x) = -a \operatorname{sgn} x\, e^{-a|x|} = -a f_o(x).$$

According to Proposition 13.19, their Fourier transforms are related by multiplication by $2\pi\,\mathrm{i}\,k$, so

$$\widehat{f}_o(k) = -\,\mathrm{i}\,\frac{4\pi\,k}{4\pi^2\,k^2 + a^2} = -\,\frac{2\pi\,\mathrm{i}\,k}{a}\,\widehat{f}_e(k),$$

as previously noted in (13.85), (13.88). On the other hand, since the odd exponential pulse has a jump discontinuity of magnitude 2 at $x = 0$, its derivative contains a delta function, and is equal to

$$f_o'(x) = -\,a\,e^{-a|x|} + 2\,\delta(x) = -\,a\,f_e(x) + 2\,\delta(x).$$

This is reflected in the relation between their Fourier transforms. If we multiply (13.88) by $2\pi\,\mathrm{i}\,k$ we obtain

$$2\pi\,\mathrm{i}\,k\,\widehat{f}_o(k) = \frac{8\pi^2\,k^2}{4\pi^2\,k^2 + a^2} = 2 - \frac{2a^2}{4\pi^2\,k^2 + a^2} = 2\,\widehat{\delta}(k) - a\,\widehat{f}_e(k).$$

Higher order derivatives are handled by iterating formula (13.104), and so:

Corollary 13.21. *The Fourier transform of $f^{(n)}(x)$ is $(2\pi\,\mathrm{i}\,k)^n\,\widehat{f}(k)$.*

This result has an important consequence: the smoothness of $f(x)$ is manifested in the rate of decay of its Fourier transform $\widehat{f}(k)$. Because, if the n^{th} derivative $f^{(n)}(x)$ is a reasonable function, then its Fourier transform $\widehat{f^{(n)}}(k) = (2\pi\,\mathrm{i}\,k)^n\,\widehat{f}(k)$ must go to zero as $|\,k\,| \to \infty$, which in turn requires very rapid decay of $\widehat{f}(k)$ itself at large $|\,k\,|$. As a general rule of thumb, local features of f, such as smoothness, are manifested by global features of $\widehat{f}(k)$, such as decay for large k. The reverse is also true: global features of f correspond to local features of $\widehat{f}$. This local-global duality, which we also encountered in the series version, is one of the major themes of Fourier theory.

Integration of functions is the inverse operation to differentiation, and so should correspond to division by $2\pi\,\mathrm{i}\,k$ in frequency space. As with Fourier series, this is not quite correct. There is an extra constant involved, and this contributes an extra delta function in frequency space.

Proposition 13.22. *If $g(x) = \displaystyle\int_{-\infty}^{x} f(y)\,dy$, then*

$$\widehat{g}(k) = \frac{\widehat{f}(k)}{2\pi\,\mathrm{i}\,k} + \frac{c}{2}\,\delta(k), \qquad \textit{where} \qquad c = \int_{-\infty}^{\infty} f(x)\,dx. \tag{13.105}$$

Table of Fourier Transforms

$f(x)$	$\widehat{f}(k)$
1	$\delta(k)$
$\operatorname{sgn} x$	$-\dfrac{\mathrm{i}}{\pi\, k}$
$\sigma(x)$	$\dfrac{1}{2}\,\delta(k) - \dfrac{\mathrm{i}}{2\pi\, k}$
$\sigma(x+a) - \sigma(x-a)$	$\dfrac{\sin 2\pi\, k\, a}{\pi\, k}$
$e^{-a\,x}\,\sigma(x)$	$\dfrac{1}{2\pi\,\mathrm{i}\,k + a}$
$e^{-a\,\lvert\,x\,\rvert}$	$\dfrac{2\,a}{4\pi^2\,k^2 + a^2}$
$e^{-a\,x^2}$	$\sqrt{\dfrac{\pi}{a}}\; e^{-\pi^2\,k^2/a}$
$\tan^{-1} x$	$-\,\mathrm{i}\;\dfrac{e^{-2\,\pi\,\lvert\,k\,\rvert}}{2\,k} \;+\; \dfrac{\pi}{2}\,\delta(k)$
$f(c\,x + d)$	$\dfrac{e^{2\,\pi\,\mathrm{i}\,k\,d/c}}{\lvert\,c\,\rvert}\,\widehat{f}\left(\dfrac{k}{c}\right)$
$\overline{f(x)}$	$\overline{\widehat{f}(-k)}$
$\widehat{f}(x)$	$f(-k)$
$f'(x)$	$2\pi\,\mathrm{i}\,k\,\widehat{f}(k)$
$f^{(n)}(x)$	$(2\pi\,\mathrm{i}\,k)^n\,\widehat{f}(k)$
$x^n\,f(x)$	$\left(\dfrac{\mathrm{i}}{2\pi}\;\dfrac{d}{dk}\right)^n\,\widehat{f}(k)$

Note: The parameter $a > 0$ is always positive, while $c \neq 0$.

Proof: First notice that

$$\lim_{x \to -\infty} g(x) = 0, \qquad \lim_{x \to +\infty} g(x) = c = \int_{-\infty}^{\infty} f(x)\, dx.$$

Therefore, the function

$$h(x) = g(x) - c\,\sigma(x)$$

obtained by subtracting a step function from the integral, decays to 0 at both $\pm\infty$. Moreover, according to the accompanying Table of Fourier Transforms,

$$\widehat{h}(k) = \widehat{g}(k) - \frac{c}{2}\,\delta(k) + \frac{\mathrm{i}\, c}{2\pi\, k}\,. \tag{13.106}$$

On the other hand,

$$h'(x) = f(x) - c\,\delta(x).$$

Since $h(x) \to 0$ as $|\,x\,| \to \infty$, we can apply our rule (13.104) for differentiation, and conclude that

$$2\pi\,\mathrm{i}\,k\,\widehat{h}(k) = \widehat{f}(k) - c. \tag{13.107}$$

Combining (13.106) and (13.107) proves (13.105). *Q.E.D.*

Example 13.23. The Fourier transform of the inverse tangent function

$$f(x) = \tan^{-1} x = \int_0^x \frac{dy}{1+y^2} = \int_{-\infty}^x \frac{dy}{1+y^2} \;-\; \frac{\pi}{2}$$

can be computed via Proposition 13.22 and (13.92): $\widehat{f}(k) = -\,\mathrm{i}\,\dfrac{e^{-2\pi\,|\,k\,|}}{2\,k} + \dfrac{\pi}{2}\,\delta(k)$, where we use the fact that $\displaystyle\int_{-\infty}^{\infty} \tan^{-1} x\, dx = \pi$.

Since the Fourier transform uniquely associates a function $\widehat{f}(k)$ on frequency space with each (reasonable) function $f(x)$ on physical space, one can characterize functions by their transforms. Practical applications rely on tables (or, even better, computer algebra systems) that recognize a wide variety of transforms for basic functions of importance in applications. The accompanying table lists some of the most important examples of functions and their Fourier transforms. Note that, according to the symmetry Theorem 13.17, each tabular entry can be used to deduce two different Fourier transforms. A more extensive list can be found in [**114**].

Applications to Differential Equations

The fact that the Fourier transform changes differentiation in the physical domain into multiplication in the frequency domain is one of its most attractive features. A particularly important consequence is that the Fourier transform effectively converts differential equations into algebraic equations, and thereby opens the door to their solution by elementary algebra! One begins by applying the Fourier transform to both sides of the differential

equation under consideration. Solving the resulting algebraic equation will produce a formula for the Fourier transform of the desired solution, which can then be immediately reconstructed via the inverse Fourier transform.

The Fourier transform is particularly well adapted to boundary value problems on the entire real line. In place of the boundary conditions used on finite intervals, we look for solutions that decay to zero sufficiently rapidly as $|\,x\,| \to \infty$ — in order that their Fourier transform be well-defined. In quantum mechanics, these are known as the *bound states* of the system, and correspond to subatomic particles that are trapped or localized in a region of space by some sort of force field. For example, the bound electrons in an atom are localized by the electrostatic attraction of the nucleus.

As a specific example, consider the boundary value problem

$$-\frac{d^2u}{dx^2} + \omega^2\, u = h(x), \qquad\qquad -\infty < x < \infty, \tag{13.108}$$

where $\omega > 0$ is a positive constant. In lieu of boundary conditions, we require that the solution $u(x) \to 0$ as $|\,x\,| \to \infty$. We will solve this problem by applying the Fourier transform to both sides of the differential equation. In view of Corollary 13.21, the result is a linear algebraic equation

$$4\,\pi^2\, k^2\, \widehat{u}(k) + \omega^2\, \widehat{u}(k) = \widehat{h}(k),$$

relating the Fourier transforms of u and h. Unlike the differential equation, the transformed equation can be immediately solved:

$$\widehat{u}(k) = \frac{\widehat{h}(k)}{4\,\pi^2\, k^2 + \omega^2\,.} \tag{13.109}$$

Therefore, we can reconstruct the solution by applying the inverse Fourier transform formula (13.74). We conclude that the solution to the boundary value problem is given by the following integral formula

$$u(x) = \int_{-\infty}^{\infty} \frac{\widehat{h}(k)\, e^{2\,\pi\,\mathrm{i}\,k\,x}}{4\,\pi^2\, k^2 + \omega^2}\, dk \tag{13.110}$$

involving the Fourier transform of the forcing function. For example, if the forcing function is an even exponential pulse,

$$h(x) = e^{-|\,x\,|} \qquad \text{with} \qquad \widehat{h}(k) = \frac{2}{4\,\pi^2\, k^2 + 1}\,,$$

then our formula (13.110) gives the solution in the form of an integral:

$$u(x) = \int_{-\infty}^{\infty} \frac{2\, e^{2\,\pi\,\mathrm{i}\,k\,x}}{(4\,\pi^2\, k^2 + \omega^2)(4\,\pi^2\, k^2 + 1)}\, dk = \int_{-\infty}^{\infty} \frac{\cos 2\,\pi\, k\, x}{(4\,\pi^2\, k^2 + \omega^2)(4\,\pi^2\, k^2 + 1)}\, dk\,.$$

The imaginary part of the complex integral vanishes because the integrand is an odd function. The remaining integral can be evaluated using a partial fraction decomposition

$$\widehat{u}(k) = \frac{2}{(4\,\pi^2\, k^2 + \omega^2)(4\,\pi^2\, k^2 + 1)} = \frac{2}{\omega^2 - 1}\left(\frac{1}{4\,\pi^2\, k^2 + 1} - \frac{1}{4\,\pi^2\, k^2 + \omega^2}\right),$$

provided $\omega \neq 1$. Thus, referring to our Table of Fourier transforms, we conclude that the solution to this boundary value problem is

$$u(x) = \frac{e^{-|x|} - \dfrac{1}{\omega}\, e^{-\omega|x|}}{\omega^2 - 1}\,. \tag{13.111}$$

The reader may wish to verify that this function is indeed a solution, meaning that it is twice continuously differentiable (which is not so immediately apparent from the formula), decays to 0 as $|x| \to \infty$, and satisfies the differential equation everywhere.

Remark: The method of partial fractions that you learned in first year calculus is an effective tool for constructing the Fourier transforms and inverse Fourier transforms of general rational functions.

A particularly important case is when the forcing function $h(x) = \delta_y(x) = \delta(x - y)$ represents a unit impulse concentrated at $x = y$. The resulting square-integrable solution is the Green's function $G(x, y)$ for the boundary value problem. According to (13.109), its Fourier transform with respect to x is

$$\widehat{G}(k, y) = \frac{e^{-2\pi i k y}}{4\pi^2 k^2 + \omega^2}\,,$$

which is the product of an exponential factor $e^{-2\pi i k y}$ times the Fourier transform of the even exponential pulse $e^{-\omega|x|}$. We apply Proposition 13.18, and conclude that the Green's function for this boundary value problem is an exponential pulse centered at y, namely

$$G(x, y) = \frac{1}{2\,\omega}\, e^{-\omega|x-y|}\,.$$

As with other self-adjoint boundary value problems, the Green's function is symmetric under interchange of x and y. As a function of x, it satisfies the homogeneous differential equation $-u'' + \omega^2 u = 0$, except at the point $x = y$ when its derivative has a jump discontinuity of unit magnitude. It also decays as $|x| \to \infty$, as required by the boundary conditions. The general superposition principle for the Green's function tells us that the solution to the general forced boundary value problem (13.108) can be represented in the integral form

$$u(x) = \int_{-\infty}^{\infty} G(x, y)\, h(y)\, dy = \frac{1}{2\,\omega} \int_{-\infty}^{\infty} e^{-\omega|x-y|}\, h(y)\, dy. \tag{13.112}$$

The reader may enjoy recovering the particular exponential solution (13.111) from this integral formula.

Convolution

The Green's function formula (13.112) that we derived at the end of the preceding subsection is indicative of a general property of Fourier transforms. The right hand side is a particular case of the general convolution product between functions.

Definition 13.24. Let $f(x)$ and $g(x)$ be scalar functions. Their *convolution* is the scalar function $h = f * g$ defined by the formula

$$h(x) = f(x) * g(x) = \int_{-\infty}^{\infty} f(x-y)\, g(y)\, dy. \tag{13.113}$$

We record the basic properties of the convolution product, leaving their verification to the reader. All of these assume that the implied convolution integrals converge.

$$\begin{array}{ll}
(a) & f * g = g * f, \\
(b) & \begin{array}{l} f * (a\,g + b\,h) = a\,(f * g) + b\,(f * h), \\ (a\,f + b\,g) * h = a\,(f * h) + b\,(g * h), \end{array} \qquad a, b \in \mathbb{C}, \\
(c) & f * (g * h) = (f * g) * h, \\
(d) & f * 0 = 0, \\
(e) & f * \delta = f.
\end{array}$$

Warning: The constant function 1 is not a unit for the convolution product; indeed,

$$f * 1 = \int_{-\infty}^{\infty} f(y)\, dy$$

is a constant function, not the original function $f(x)$. In fact, according to the last property, the delta function plays the role of the "convolution unit":

$$f(x) * \delta(x) = \int_{-\infty}^{\infty} f(x-y)\, \delta(y)\, dy = f(x)$$

by the fundamental property (11.37) of the delta function. Proofs of the other convolution identities are left to the exercises.

In particular, the solution formula (13.112) is equal to the convulution of an even exponential pulse $g(x) = \dfrac{1}{2\omega}\, e^{-\omega |\, x\, |}$ with the forcing function:

$$u(x) = g(x) * h(x) = \int_{-\infty}^{\infty} g(x-y)\, h(y)\, dy = g * h(x)$$

On the other hand, its Fourier transform (13.109) is the ordinary multiplicative product

$$\widehat{u}(k) = \widehat{g}(k)\, \widehat{h}(k)$$

of the Fourier transforms of g and h. This is a special case of a general fact: convolution in the physical domain corresponds to multiplication in the frequency domain, and converserly. More explicitly:

Theorem 13.25. *The Fourier transform of the convolution $u = f * g$ of two functions is the product of their Fourier transforms: $\widehat{u} = \widehat{f} \cdot \widehat{g}$. Vice versa, the Fourier transform of their product $h = f \cdot g$ is the convolution of their Fourier transforms $\widehat{h} = \widehat{f} * \widehat{g}$.*

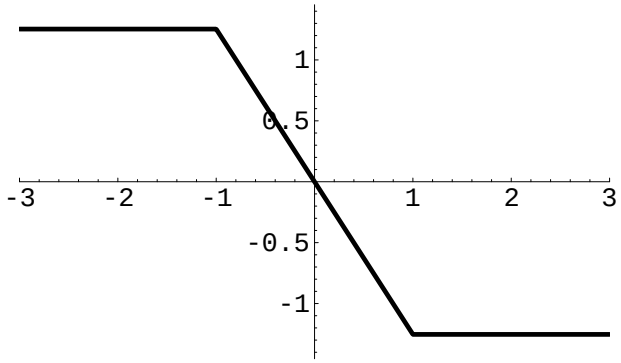


Figure 13.13. The Fourier transform of $\dfrac{\sin x}{x^2}$.

Proof: From the definition of the Fourier transform,

$$\widehat{u}(k) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y)\, g(y)\, e^{-2\pi \mathrm{i}\, k x}\, dx\, dy.$$

Applying the change of variables $z = x - y$ in the inner integral produces

$$\begin{aligned} \widehat{u}(k) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(z)\, g(y)\, e^{-2\pi \mathrm{i}\, k(y+z)}\, dz\, dy \\ &= \left(\int_{-\infty}^{\infty} f(z)\, e^{-2\pi \mathrm{i}\, k z}\, dz \right) \left(\int_{-\infty}^{\infty} g(y)\, e^{-2\pi \mathrm{i}\, k y}\, dy \right) = \widehat{f}(k)\, \widehat{g}(k). \end{aligned}$$

The second statement follows directly from the Symmetry Theorem 13.17. *Q.E.D.*

Example 13.26. We already know that the Fourier transform of $f(x) = \dfrac{\sin x}{x}$ is the box function

$$\widehat{f}(k) = \pi \left[\sigma\left(k + \frac{1}{2\pi} \right) - \sigma\left(k - \frac{1}{2\pi} \right) \right] = \begin{cases} \pi, & |\, k\,| < \frac{1}{2}\pi, \\ 0, & \text{otherwise}, \end{cases}$$

cf. (13.93). We also know that the Fourier transform of $g(x) = \dfrac{1}{x}$ is $\widehat{g}(k) = -\pi \mathrm{i}\, \operatorname{sgn} k$. Therefore, the Fourier transform of their product $h(x) = f(x)\, g(x) = \dfrac{\sin x}{x^2}$ can be obtained by convolution:

$$\begin{aligned} \widehat{h}(k) &= \widehat{f} * \widehat{g}(k) \\ &= \int_{-\infty}^{\infty} \widehat{f}(l)\, \widehat{g}(k-l)\, dl = -\mathrm{i} \int_{-1/2\pi}^{1/2\pi} \operatorname{sgn}(k-l)\, dl = \begin{cases} \pi\, \mathrm{i}\,, & k < -\frac{1}{2}\pi, \\ -2\pi^2\, \mathrm{i}\, k, & -\frac{1}{2}\pi < k < \frac{1}{2}\pi, \\ -\pi\, \mathrm{i}\,, & k > \frac{1}{2}\pi. \end{cases} \end{aligned}$$

A graph of $\widehat{h}(k)$ appears in Figure 13.13.

Fourier Transform on Hilbert Space

While we do not have the space to embark on a fully rigorous treatment of the theory underlying the Fourier transform, it is worth outlining a few of the most basic ideas and results. We have already noted that the Fourier transform, when defined, is a linear map, taking functions $f(x)$ on physical space to functions $\widehat{f}(k)$ on frequency space. A critical question is precisely which function space should the theory be applied to. Not every function admits a Fourier transform — the integral in (13.72) must converge.

It turns out the proper setting for the rigorous theory is the *Hilbert space* of complex-valued square-integrable functions, the same infinite-dimensional vector space that lies at the foundation of modern quantum mechanics. We adapt Definition 12.32 to the entire real line, and so define the Hilbert space $\mathrm{L}^2 = \mathrm{L}^2(-\infty, \infty)$ to be the infinite-dimensional vector space consisting of all complex-valued functions $f(x)$ which are defined for all $x \in \mathbb{R}$ and have finite L^2 norm

$$\| f \|^2 = \int_{-\infty}^{\infty} | f(x) |^2 \, dx < \infty. \tag{13.114}$$

The inner product on L^2 is defined in the usual manner,

$$\langle f \, ; g \rangle = \int_{-\infty}^{\infty} f(x) \, \overline{g(x)} \, dx,$$

with the Cauchy–Schwarz inequality (3.16) ensuring that the integral is finite whenever $f, g \in \mathrm{L}^2$. The decay criterion (13.75) is sufficient to ensure that a piecewise continuous function belongs to the Hilbert space L^2. However, as discussed in Section 12.5, Hilbert space contains many more functions, and the precise definitions and identification of functions is quite subtle. In quantum mechanics, the Hilbert space functions $\varphi \in \mathrm{L}^2$ that have unit norm, $\| \varphi \| = 1$, are known as *wave functions*. The modulus $| \varphi(x) |$ of the wave function at a position x indicates the probability of finding the physical system there.

Let us state the fundamental theorem governing the convergence of the Fourier transform on Hilbert space. A rigorous proof can be found in [**141**].

Theorem 13.27. *If $f(x) \in \mathrm{L}^2$ is square-integrable, then its Fourier transform $\widehat{f}(k) \in \mathrm{L}^2$ is well-defined and square-integrable function of the frequency variable k. If $f(x)$ is continuously differentiable at a point x, then its inverse Fourier integral* (13.74) *equals the value $f(x)$. More generally, if the right and left hand limits $f(x^-)$, $f'(x^-)$, and $f(x^+)$, $f'(x^+)$ exist, then the Fourier integral converges to the average value $\frac{1}{2}\big[\, f(x^-) + f(x^+) \,\big]$.*

Thus, the Fourier transform $\widehat{f} = \mathcal{F}[\, f \,]$ defines a linear transformation from L^2 functions of x to L^2 functions of k. In fact, the Fourier transform preserves inner products, and hence defines an *isometry* or norm-preserving transformation on Hilbert space. This important result is known as *Parseval's formula*, whose Fourier series counterpart appeared in (12.104).

Theorem 13.28. *If $\widehat{f}(k) = \mathcal{F}[\, f(x) \,]$ and $\widehat{g}(k) = \mathcal{F}[\, g(x) \,]$, then $\langle f \, ; g \rangle = \langle \widehat{f} \, ; \widehat{g} \, \rangle$, i.e.,*

$$\int_{-\infty}^{\infty} f(x) \, \overline{g(x)} \, dx = \int_{-\infty}^{\infty} \widehat{f}(k) \, \overline{\widehat{g}(k)} \, dk. \tag{13.115}$$

Proof: Let us sketch a formal proof that serves to motivate why this result is valid. We use the definition (13.72) of the Fourier transform in the formula

$$\int_{-\infty}^{\infty} \widehat{f}(k)\,\overline{\widehat{g}(k)}\,dk = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x)\, e^{-2\pi i k x}\, dx \right) \left(\int_{-\infty}^{\infty} \overline{g(y)}\, e^{2\pi i k y}\, dy \right) dk$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)\,\overline{g(y)} \left(\int_{-\infty}^{\infty} e^{-2\pi i k(x-y)}\, dk \right) dx\, dy.$$

Now according to (13.101), the inner k integral can be replaced by a delta function $\delta(x-y)$, and hence

$$\int_{-\infty}^{\infty} \widehat{f}(k)\,\overline{\widehat{g}(k)}\,dk = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)\,\overline{g(y)}\,\delta(x-y)\,dx\,dy = \int_{-\infty}^{\infty} f(x)\,\overline{g(x)}\,dx.$$

This completes the "proof". *Q.E.D.*

Choosing $f = g$ in (13.115) results in the *Plancherel formula* $\| f \| = \| \widehat{f} \|$, or, explicitly,

$$\int_{-\infty}^{\infty} | f(x) |^2\, dx = \int_{-\infty}^{\infty} | \widehat{f}(k) |^2\, dk. \tag{13.116}$$

The Heisenberg Uncertainty Principle

In its popularized form, the Heisenberg Uncertainty Principle is a familiar philosophical concept from quantum mechanics. The principle, first formulated by the twentieth century German physicist Werner Heisenberg, one of the founders of modern quantum mechanics, states that certain pairs of physical quantities cannot be simultaneously determined to complete accuracy by an experimental measurement. For instance, the more precisely one determines the position of a particle, the less accuracy there will be in the measurement of its momentum, and conversely. Similarly, the smaller the error in the energy, the larger the error in the time. Experimental verification of the uncertainty principle can be found even in fairly simple situations. Consider a light beam passing through a small hole. The position of the photons is constrained by the hole; the effect of their momenta is in the pattern of light diffused on a screen placed beyond the hole. The smaller the hole, the more constrained the position, and the wider the image on the screen, meaning the less certainty there is in the momentum.

This is not the place to discuss the philosophical and experimental consequences of Heisenberg's principle. What we will show is that the Uncertainty Principle is, in fact, a rigorous theorem concerning the Fourier transform! While we do not have the space to discuss why the physical underpinnings of quantum mechanics reduce to this statement, we can state and prove the mathematical version of the basic uncertainty principle. In quantum theory, each of the paired quantities, e.g., position and momentum, are interrelated by the Fourier transform.

In quantum mechanics, measurable quantities are interpreted as linear operators $A\colon \mathrm{L}^2 \to \mathrm{L}^2$ on the underlying Hilbert space. In particular, *position*, usually denoted by Q, is identified with the operation of multiplication by x, so $Q[\varphi] = x\,\varphi(x)$, whereas *momentum*, denoted by P for historical reasons, is identified as the differentiation operator

$P = d/dx$, so that $P[\varphi] = \varphi'(x)$. If we apply the Fourier transform to our wave function, then, as we saw, the differentiation or momentum operator becomes a multiplication operator $\widehat{P}[\widehat{\varphi}] = 2\pi\,\mathrm{i}\,k\,\widehat{\varphi}(k)$. On the other hand, the momentum operator transforms into differentiation $\widehat{Q}[\widehat{\varphi}] = -\,\mathrm{i}\,\widehat{\varphi}'(k)$ in the frequency domain. This duality between position and momentum, or multiplication and differentiation, via the Fourier transform lies at the heart of the Uncertainty Principle.

Let A be a linear operator on Hilbert space representing a physical quantity. If the quantum system is in a state represented by a particular wave function φ, then the localization of the quantity A is measured by the norm $\| A[\varphi] \|$. The smaller this norm, the more accurate the measurement. For instance, $\| Q[\varphi] \| = \| x\,\varphi(x) \|$ measures the localization of the position of the particle represented by φ; the smaller $\| Q[\varphi] \|$, the more concentrated the probability of finding the particle near[†] $x = 0$ and hence the smaller the error in the measurement of its position. Similarly, by Plancherel's formula (13.116),

$$\| P[\varphi] \| = \| \varphi'(x) \| = \| 2\pi\,\mathrm{i}\,k\,\widehat{\varphi}(k) \|,$$

measures the localization in the momentum of the particle, which is small if and only if the Fourier transform is concentrated near $k = 0$, and hence the smaller the error in its measured momentum. With this interpretation, the Uncertainty Principle states that one cannot simultaneously make these two quantities arbitrarily small.

Theorem 13.29. *If $\varphi(x)$ is a wave function, so $\| \varphi \| = 1$, then*

$$\| Q[\varphi] \|\ \| P[\varphi] \| \geq \tfrac{1}{2}. \tag{13.117}$$

Proof: The proof rests on the Cauchy–Schwarz inequality

$$\Big| \langle\, x\,\varphi(x)\,;\varphi'(x)\,\rangle \Big| \leq \| x\,\varphi(x) \|\ \| \varphi'(x) \| = \| Q[\varphi] \|\ \| P[\varphi] \|. \tag{13.118}$$

On the other hand, writing out the inner product term

$$\langle\, x\,\varphi(x)\,;\varphi'(x)\,\rangle = \int_{-\infty}^{\infty} x\,\varphi(x)\,\varphi'(x)\,dx.$$

Let us integrate by parts, using the fact that

$$\varphi(x)\,\varphi'(x) = \frac{d}{dx}\,\big[\,\tfrac{1}{2}\,\varphi(x)^2\,\big].$$

Since $\varphi(x) \to 0$ as $|\,x\,| \to \infty$, the boundary terms vanish, and hence

$$\langle\, x\,\varphi(x)\,;\varphi'(x)\,\rangle = \int_{-\infty}^{\infty} x\,\varphi(x)\,\varphi'(x)\,dx = -\int_{-\infty}^{\infty} \tfrac{1}{2}\,\varphi(x)^2\,dx = -\,\tfrac{1}{2},$$

since $\| \varphi \| = 1$. Substituting back into (13.118) completes the proof. *Q.E.D.*

The inequality (13.117) quantifies the statement that the more accurately we measure the momentum Q, the less accurately we are able to measure the position P, and vice versa. For more details and physical consequences, you should consult an introductory text on mathematical quantum mechanics, e.g., [**100**, **104**].

[†] For another position, one replaces x by $x - a$.

13.4. The Laplace Transform.

In engineering applications, the Fourier transform is often overshadowed by a close relative — the Laplace transform. The Laplace transform plays an essential role in control theory, linear systems analysis, electronics, and many other fields of practical engineering and science. However, the Laplace transform is most properly interpreted as a particular real form of the more fundamental Fourier transform. When the Fourier transform is evaluated along the imaginary axis, the complex exponential factor turns into a real exponential, and the resulting Laplace transform maps real-valued functions to real-valued functions. The Laplace transform is one-sided; it only looks forward in time, while the Fourier transform looks in both directions in space. Since it is so closely allied to the Fourier transform, the Laplace transform enjoys many of its featured properties, including linearity. Moreover, derivatives are transformed into algebraic operations, which underlies its applications to solving differential equations. The key difference is that the Fourier transform is designed for boundary value problems on the real line, whereas the Laplace transform is more suitable for initial value problems.

Since we will be applying the Laplace transform to initial value problems, we switch our notation from x to t to emphasize this fact. Suppose $f(t)$ is a (reasonable) function with $f(t) = 0$ for all $t < 0$. Then the Fourier transform of f is

$$\widehat{f}(k) = \int_0^\infty f(t)\, e^{-2\pi \mathrm{i}\, k t}\, dt,$$

since, by our assumption, the negative t values makes no contribution to the integral. The *Laplace transform* $\mathcal{L}$ is obtained by replacing $2\pi\,\mathrm{i}\,k$ by a real[†] variable s, leading to

$$F(s) = \mathcal{L}[\,f(t)\,] \equiv \int_0^\infty f(t)\, e^{-s\,t}\, dt. \tag{13.119}$$

In particular, if $f(t)$ is real, then its Laplace transform $F(s)$ is also a real function. By allowing complex values of the Fourier frequency variable k, we identify the Laplace transform with the evaluation of the Fourier transform on the imaginary axis:

$$F(s) = \widehat{f}\left(-\frac{\mathrm{i}\,s}{2\pi}\right). \tag{13.120}$$

Using real exponentials has the advantage that the transform takes real functions to real functions. Moreover, since the integral kernel $e^{-s\,t}$ is exponentially decaying for $s > 0$, we are no longer required to restrict our attention to functions that decay to zero as $t \to \infty$.

Example 13.30. Consider an exponential function $f(t) = e^{\alpha\,t}$, where the exponent α is allowed to be complex. Its Laplace transform is

$$F(s) = \int_0^\infty e^{(\alpha - s)\,t}\, dt = \frac{1}{s-\alpha}\,. \tag{13.121}$$

[†] One can also define the Laplace transform at complex values of s, but this will not be required in the applications discussed here.

Note that the integrand is exponentially decaying, and hence the integral converges, if and only if $\operatorname{Re}(\alpha - s) < 0$. Therefore, the Laplace transform (13.121) is, strictly speaking, only defined when $s > \operatorname{Re}\alpha$ is sufficiently large. In particular, for an oscillatory exponential,

$$\mathcal{L}[e^{\mathrm{i}\,\omega\,t}] = \frac{1}{s - \mathrm{i}\,\omega} = \frac{s + \mathrm{i}\,\omega}{s^2 + \omega^2} \qquad \text{provided} \qquad s > 0.$$

Taking real and imaginary parts of this identity, we discover the formulae for the Laplace transforms of the basic trigonometric functions:

$$\mathcal{L}[\cos\omega\,t] \;=\; \frac{s}{s^2 + \omega^2}\,, \qquad\qquad \mathcal{L}[\sin\omega\,t] \;=\; \frac{\omega}{s^2 + \omega^2}\,. \tag{13.122}$$

Two further important cases are

$$\mathcal{L}[1] = \int_0^\infty e^{-\,s\,t}\,dt = \frac{1}{s}\,, \qquad \mathcal{L}[t] = \int_0^\infty t\,e^{-\,s\,t}\,dt = \frac{1}{s}\int_0^\infty e^{-\,s\,t}\,dt = \frac{1}{s^2}\,. \tag{13.123}$$

The middle step in the second computation is an integration by parts, making sure that the boundary terms vanish.

Remark: In every case, we really mean the Laplace transform of the function whose values are given for $t > 0$ and is equal to 0 for all negative t. Therefore, the function 1 in reality signifies the step function

$$\sigma(t) = \begin{cases} 1, & t > 0, \\ 0, & t < 0, \end{cases} \tag{13.124}$$

and so the first formula in (13.123) should really be written

$$\mathcal{L}[\sigma(t)] = \frac{1}{s}\,. \tag{13.125}$$

However, in the traditional approach to the Laplace transform, one only considers the functions on the positive t axis, and so the step function and the constant function are, from this viewpoint, indistinguishable. However, once one moves beyond a purely mechanistic viewpoint, any deeper understanding of the properties of the Laplace transform relies on keeping this distinction firmly in mind.

Let us now pin down the precise class of functions to which the Laplace transform can be applied.

Definition 13.31. A function $f(t)$ is said to have *exponential growth* of *order* a if it satisfies the inequality

$$|\,f(t)\,| < M\,e^{a\,t}, \qquad \text{for all} \qquad t > t_0. \tag{13.126}$$

Here $M > 0$ is a positive constant and $t_0 > 0$.

Table of Laplace Transforms

$f(t)$	$F(s)$
1	$\dfrac{1}{s}$
t	$\dfrac{1}{s^2}$
t^n	$\dfrac{n!}{s^{n+1}}$
$\delta(t-c)$	e^{-sc}
$\sigma(t-c)$	$\dfrac{e^{-sc}}{s}$
$e^{\alpha t}$	$\dfrac{1}{s-\alpha}$
$\cos \omega t$	$\dfrac{s}{s^2+\omega^2}$
$\sin \omega t$	$\dfrac{\omega}{s^2+\omega^2}$
$e^{ct}\, f(t)$	$F(s-c)$
$\dfrac{f(t)}{t}$	$\displaystyle\int_s^\infty F(r)\, dr$
$f(t-c)$	$e^{-sc}\, F(s)$
$f'(t)$	$s\, F(s) - f(0)$
$f^{(n)}(t)$	$s^n\, F(s) - s^{n-1}\, f(0) - - s^{n-2}\, f'(0) - \cdots - f^{(n-1)}(0)$
$f(t) * g(t)$	$F(s)\, G(s)$

Note that the exponential growth condition only depends upon the function's behavior for large values of t. If $a < 0$, then f is, in fact, exponentially decaying as $x \to \infty$. Since $e^{at} < e^{bt}$ for $a < b$ and all $t > 0$, if $f(t)$ has exponential growth of order a, it automat-

ically has exponential growth of any higher order $b > a$. All polynomial, trigonometric, and exponential functions (with linear argument) have exponential growth. The simplest example of a function that does not admit an exponential growth bound is $f(t) = e^{t^2}$, which grow faster than any simple exponential $e^{a\,t}$.

The following result guarantees the existence of the Laplace transform, at least for sufficiently large values of the transform variable s, for a rather broad class of functions that includes almost all of the functions that arise in applications.

Theorem 13.32. *If $f(t)$ is piecewise continuous and has exponential growth of order a, then its Laplace transform $F(s) = \mathcal{L}[\,f(t)\,]$ is defined for all $s > a$.*

Proof: The exponential growth inequality (13.126) implies that we can bound the integrand in (13.119) by

$$\big|\, f(t)\, e^{-s\,t} \,\big| < M\, e^{(a-s)\,t}.$$

Therefore, as soon as $s > a$, the integrand is exponentially decaying as $t \to \infty$, and this suffices to prove that the Laplace transform integral converges. *Q.E.D.*

Theorem 13.32 is an existential result, and, in practice, we may not be able to explicitly evaluate the Laplace transform integral. Nevertheless, the Laplace transforms of most common functions are not hard to find, and extensive lists of known Laplace transforms have been tabulated, [**115**]. Nowadays, the best source of transform formulas are computer algebra packages, including MATHEMATICA and MAPLE.

A key fact of Fourier analysis is that the Fourier transform uniquely specifies the function, except possibly at jump discontinuities where the limiting value must be half way in between. The following result, which follows from the Fourier transform Theorem 13.27, says that the Laplace transform also uniquely determines the function.

Lemma 13.33. *If f and g are piecewise continuous functions that are of exponential growth, and $\mathcal{L}[\,f(t)\,] = \mathcal{L}[\,g(t)\,]$ for all s sufficiently large, then $f(t) = g(t)$ at all points of continuity $t > 0$.*

In fact, there is an explicit formula for the inverse Laplace transform, which follows from its identification (13.120) as the Fourier transform along the imaginary axis. Under suitable hypotheses, given the Laplace transform $F(s)$, the original function can be found by a complex integral formula[†]

$$f(t) = \frac{1}{2\,\pi\,\mathrm{i}} \int_{-\,\mathrm{i}\,\infty}^{\,\mathrm{i}\,\infty} F(s)\, e^{s\,t}\, ds, \qquad\qquad t > 0. \tag{13.127}$$

In practice, one hardly ever uses this complicated formula to compute the inverse Laplace transform. Rather, one simply relies on tables of known Laplace transforms, coupled with a few basic rules to be covered in the following subsection.

[†] See Section 16.5 for details on complex integration. The stated formula doesn't apply to all functions of exponential growth. A more universally valid inverse Laplace transform formula is obtained by shifting the complex contour to run from $b - \mathrm{i}\,\infty$ to $b + \mathrm{i}\,\infty$ for some $b > a$, the order of exponential growth of f.

Just like its Fourier cousin, the Laplace transform converts calculus into algebra. In particular, in the transform domain, differentiation turns into multiplication, but with one additional term that depends upon the value of the function at $t = 0$.

Theorem 13.34. *Let $f(t)$ have exponential growth of order a. If $\mathcal{L}[f(t)] = F(s)$ then, for $s > a$,*

$$\mathcal{L}[f'(t)] = s\,F(s) - f(0). \tag{13.128}$$

Proof: The proof relies on an integration by parts:

$$\begin{aligned}\int_0^\infty f'(t)\,e^{-st}\,dt &= f(t)\,e^{-st}\,\Big|_{t=0}^{\infty} + s\int_0^\infty f(t)\,e^{-st}\,dt\\ &= \lim_{t\to\infty} f(t)\,e^{-st} - f(0) + s\,F(s).\end{aligned}$$

The exponential growth inequality (13.126) implies that first term vanishes when $s > a$. *Q.E.D.*

Example 13.35. According to Example 13.30, $\mathcal{L}[\sin\omega t] = \dfrac{\omega}{s^2+\omega^2}$. The derivative is $\dfrac{d}{dt}\sin\omega t = \omega\cos\omega t$, and therefore $\mathcal{L}[\omega\cos\omega t] = \dfrac{\omega\,s}{s^2+\omega^2}$, since $\sin\omega t$ vanishes at $t = 0$. The reuslt agrees with (13.122). On the other hand, $\dfrac{d}{dt}\cos\omega t = -\omega\sin\omega t$, and so

$$\mathcal{L}[-\omega\sin\omega t] \;=\; s\,\mathcal{L}[\cos\omega t] - 1 = \frac{s^2}{s^2+\omega^2} - 1 = -\,\frac{\omega^2}{s^2+\omega^2}\,,$$

again in agreement with the known formula.

Remark: In the language of Fourier transforms, the final term in (13.128) is, in fact, due to a discontinuity in the function at $t = 0$. Keep in mind that the Laplace transform only applies to functions with $f(t) = 0$ for all $t < 0$. If $f(0) \neq 0$, then $f(t)$ has a jump discontinuity of magnitude $f(0)$ at $t = 0$. Therefore, according to the calculus of generalized functions, its derivative $f'(t)$ should include a delta function term, namely $f(0)\,\delta(0)$, which would account for an additional constant term in its transform. Thus, in the example, the derivative of the (continuous) function that equals $\sin\omega t$ for $t > 0$ is the function that equals $\omega\cos\omega t$ for $t > 0$, while the derivative of the latter has an extra delta function resulting from its discontinuity at $t = 0$. In the usual working version of the Laplace transform calculus, one suppresses the delta function when computing the derivative $f'(t)$. However, its effect must reappear on the other side of the differentiation formula (13.128), and this exactly accounts for the extra term $-f(0)$.

Laplace transforms of higher order derivatives are computed by iterating the basic formula (13.128). For example,

$$\mathcal{L}[f''(t)] = s\,\mathcal{L}[f'(t)] - f'(0) = s^2\,F(s) - s\,f(0) - f'(0). \tag{13.129}$$

In general,

$$\mathcal{L}\big[\, f^{(n)} \,\big] = s^n F(s) - s^{n-1} f(0) - s^{n-2} f'(0) - \, \cdots \, - f^{(n-1)}(0). \tag{13.130}$$

Conversely, integration corresponds to dividing the Laplace transform by s, so

$$\mathcal{L}\left[\int_0^t f(\tau)\, d\tau \right] = \frac{F(s)}{s}\,. \tag{13.131}$$

Unlike the Fourier transform, there are no additional terms in the integration formula as long as we start the integral at $t = 0$. In particular,

$$\mathcal{L}[t^2\,] = \frac{1}{s}\,\mathcal{L}[2\,t] = \frac{2}{s^3}\,, \qquad \text{and, more generally,} \qquad \mathcal{L}[t^n\,] = \frac{n!}{s^{n+1}}\,. \tag{13.132}$$

There is also a shift formula, analogous to Proposition 13.18 for Fourier transforms, but with one important caveat. Since all functions must vanish for $t < 0$, we are only allowed to shift the functions to the right, as otherwise they would assume nonzero values on the negative t axis. The Laplace transform of the function $f(t-c)$, which is shifted to the right by an amount $c > 0$, is

$$\begin{aligned} \mathcal{L}[\, f(t-c)\,] &= \int_0^\infty f(t-c)\, e^{-\,s\,t}\, dt = \int_{-\,c}^\infty f(t)\, e^{-\,s\,(t+c)}\, dt \\ &= \int_{-\,c}^0 f(t)\, e^{-\,s\,(t+c)}\, dt + \int_0^\infty f(t)\, e^{-\,s\,(t+c)}\, dt = e^{-\,s\,c} \int_0^\infty f(t)\, e^{-\,s\,t}\, dt = e^{-\,s\,c}\, F(s). \end{aligned} \tag{13.133}$$

In this computation, we first employed a change of variables, replacing $t - c$ by t; then, the fact that $f(t) \equiv 0$ for $t < 0$ was used to eliminate the integral from $-c$ to 0.

Example 13.36. Consider the square wave pulse

$$f(t) = \begin{cases} 1, & b < t < c, \\ 0, & \text{otherwise}, \end{cases}$$

where $0 < b < c$. To compute its Laplace transform, we write it as the difference

$$f(t) = \sigma(t-b) - \sigma(t-c)$$

of shifted versions of the step function (13.124). Combining the shift formula (13.133) and the formula (13.125) for the Laplace transform of the step function, we find

$$\mathcal{L}[\, f(t)\,] = \mathcal{L}[\,\sigma(t-b)\,] - \mathcal{L}[\,\sigma(t-c)\,] = \frac{e^{-\,s\,b} - e^{-\,s\,c}}{s}\,. \tag{13.134}$$

Applications to Initial Value Problems

A key application of the Laplace transform is to aid in the solution of initial value problems for linear, constant coefficient ordinary differential equations. As a prototypical example, consider the second order initial value problem

$$a\,\ddot{u} + b\,\dot{u} + c\,u = f(t), \qquad u(0) = \alpha, \qquad \dot{u}(0) = \beta, \tag{13.135}$$

in which a, b, c are constant. We apply the Laplace transform to both sides of the differential equation. In view of the differentiation formulae (13.128), (13.129), we obtain

$$a\big[\, s^2 \mathcal{L}[\, u(t)\,] - s\, u(0) - \dot u(0)\,\big] + b\big[\, s\, \mathcal{L}[\, u(t)\,] - u(0)\,\big] + c\, \mathcal{L}[\, u(t)\,] = \mathcal{L}[\, f(t)\,].$$

Setting $\mathcal{L}[\, u(t)\,] = U(s)$ and $\mathcal{L}[\, f(t)\,] = F(s)$, the preceding equation takes the form

$$(a\, s^2 + b\, s + c)\, U(s) = F(s) + (a\, s + b)\, \alpha + a\, \beta. \tag{13.136}$$

Thus, by applying the Laplace transform, we have effectively reduced the differential equation to an elementary algebraic equation. The solution $u(t)$ to the initial value problem is found by taking the inverse Laplace transform of the solution

$$U(s) = \frac{F(s) + (a\, s + b)\, \alpha + a\, \beta}{a\, s^2 + b\, s + c}\,. \tag{13.137}$$

As remarked above, in practice the inverse transform $u(t)$ is found by manipulating the answer (13.137) into a sum of terms all of which appear in a table of transforms.

Example 13.37. Consider the initial value problem

$$\ddot u + u = 10\, e^{-3\, t}, \qquad\qquad u(0) = 1, \qquad \dot u(0) = 2.$$

Taking the Laplace transform of the differential equation as above, we find

$$(s^2 + 1)\, U(s) - s - 2 = \frac{10}{s + 3}\,, \qquad \text{and so} \qquad U(s) = \frac{s + 2}{s^2 + 1} + \frac{10}{(s + 3)(s^2 + 1)}.$$

The second summand does not directly correspond to any of the entries in our table of Laplace transforms. However, we can use the method of partial fractions to write it as a sum

$$U(s) = \frac{s + 2}{s^2 + 1} + \frac{1}{s + 3} + \frac{3 - s}{s^2 + 1} = \frac{1}{s + 3} + \frac{5}{s^2 + 1}$$

of terms appearing in the table. Linearity of the Laplace transform tells us that the solution to our initial value problem is

$$u(t) = e^{-3\, t} + 5 \sin t.$$

Of course, the last example is a problem that you can easily solve directly. The standard method learned in your first course on differential equations is just as effective in finding the final solution, and does not require all the extra Laplace transform machinery! The Laplace transform method is, however, particularly effective for dealing with complications that arise in cases of discontinuous forcing functions.

Example 13.38. Consider a mass vibrating on a spring with fixed stiffness $c = 4$. Assume that the mass starts at rest, is then subjected to a unit force over time interval $\frac{1}{2}\, \pi < t < 2\, \pi$, after which it left to vibrate on its own. The initial value problem is

$$\ddot u + 4\, u = f(t) = \begin{cases} 1, & \frac{1}{2}\, \pi < t < 2\, \pi, \\ 0, & \text{otherwise}, \end{cases} \qquad\qquad u(0) = \dot u(0) = 0.$$

Taking the Laplace transform, and using (13.134), we find

$$(s^2+4)\,U(s) = \frac{e^{-\pi s/2} - e^{-2\pi s}}{s}, \qquad \text{and so} \qquad U(s) = \frac{e^{-\pi s/2} - e^{-2\pi s}}{s(s^2+4)}.$$

Therefore, by the shift formula (13.133)

$$u(t) = h\big(t - \tfrac{1}{2}\pi\big) - h(t - 2\pi),$$

where $h(t)$ is the function with Laplace transform

$$\mathcal{L}[\,h(t)\,] = H(s) = \frac{1}{s(s^2+4)} = \frac{1}{4}\left(\frac{1}{s} - \frac{s}{s^2+4}\right),$$

which is conveniently rewritten using partial fractions. According to our table of Laplace transforms,

$$h(t) = \tfrac{1}{4} - \tfrac{1}{4}\cos 2t.$$

Therefore, our desired solution is

$$u(t) = \begin{cases} 0, & 0 \le t \le \frac{1}{2}\pi, \\ \frac{1}{4} + \frac{1}{4}\cos 2t, & \frac{1}{2}\pi \le t \le 2\pi, \\ \frac{1}{2}\cos 2t, & 2\pi \le t. \end{cases}$$

Note that $u(t)$ is only C^1 at the points of discontinuity of the forcing function.

Remark: A direct solution of this problem would proceed as follows. One solves the differential equation on each interval of continuity of the forcing function, leading to a solution on that interval depending upon two integration constants. The integration constants are then adjusted so that the solution satisfies the initial conditions and is continuous and has continuous first derivative at each point of discontinuity of the forcing function. The details are straightforward, but messy. The Laplace transform method successfully bypasses the intervening manipulations required in the direct method.

Convolution

We already noted that the Fourier transform of the convolution product of two functions is realized as the ordinary product of their individual transforms. A similar result holds for the Laplace transform, as we now demonstrate.

Let $f(t), g(t)$ be given functions. A key point is that, since we are implicitly assuming that the functions vanish at all negative values of t, their convolution product (13.113) reduces to a finite integral

$$h(t) = f(t) * g(t) = \int_0^t f(t-\tau)\, g(\tau)\, d\tau. \tag{13.138}$$

In particular $h(t) = 0$ for all $t < 0$ also. Further, it is not hard to show that the convolution of two functions of exponential growth also has exponential growth.

The proof of the convolution theorem for the Laplace transform proceeds along the same lines as its Fourier transform version Theorem 13.25, and is left as an exercise for the reader.

Theorem 13.39. *If $\mathcal{L}[f(t)] = F(s)$ and $\mathcal{L}[g(t)] = G(s)$, then the convolution $h(t) = f(t) * g(t)$ has Laplace transform given by the product $H(s) = F(s)\,G(s)$.*

The Convolution Theorem 13.39 has useful applications to differential equations. Consider the initial value problem

$$\ddot{u} + \omega^2 u = f(t), \qquad u(0) = \dot{u}(0) = 0,$$

Applying the Laplace transform to the differential equation and using the initial conditions,

$$(s^2 + \omega^2)\,U(s) = F(s), \qquad \text{and hence} \qquad U(s) = \frac{F(s)}{s^2 + \omega^2}\,.$$

The right hand side is the product of the Laplace transform of the forcing function $f(t)$ and that of the trigonometric function $\dfrac{\sin \omega t}{\omega}$; therefore, Theorem 13.39 implies that

$$u(t) = f(t) * \frac{\sin \omega t}{\omega} = \int_0^t \frac{\sin \omega (t-\tau)}{\omega}\, f(\tau)\, d\tau. \tag{13.139}$$

The integral kernel

$$k(t) = \begin{cases} \dfrac{\sin \omega t}{\omega}\,, & t > 0, \\ 0, & t < 0, \end{cases}$$

is known as the *fundamental solution* for the initial value problem. The function $k(t - \tau)$ gives the response of the system to a unit impulse force that is applied instantanteously at the time $t = \tau$. Note particularly that (unlike boundary value problems) the impulse only affect the solutions at later times $t > \tau$. The fundamental solution plays a role similar to that of a Green's function in a boundary value problem. A general external force can be viewed as a superpostion of individual impulses over time, and the integral formula (13.139) expresses the response of the system as the same superposition of the fundamental solution responses.

This concludes our brief introduction to the Laplace transform and a few of its many applications to physical problems. More details can be found in almost all applied texts on mechanics, electronical circuits, signal processing, control theory, and many other areas.

Chapter 14

Vibration and Diffusion in One–Dimensional Media

In this chapter, we study the solutions, both analytical and numerical, to the two most important equations of one-dimensional continuum dynamics. The *heat equation* models the diffusion of thermal energy in a body; here, we analyze the case of a one-dimensional bar. The *wave equation* describes vibrations and waves in continuous media, including sound waves, water waves, elastic waves, electromagnetic waves, and so on. Again, we restrict our attention here to the case of waves in a one-dimensional medium, e.g., a string, or a bar, or a column of air. The two- and three-dimensional versions of these fundamental equations will be analyzed in the later Chapters 17 and 18.

As we saw in Section 12.1, the basic solution technique is inspired by our eigenvalue-based methods for solving linear systems of ordinary differential equations. Substituting the appropriate exponential or trigonometric ansatz will effectively reduce the partial differential equation to a one-dimensional boundary value problem. The linear superposition principle implies that general solution can then be expressed as a infinite series in the resulting eigenfunction solutions. In both cases considered here, the eigenfunctions of the one-dimensional boundary value problem are trigonometric, and so the solution to the partial differential equation takes the form of a time-dependent Fourier series. Although we cannot, in general, analytically sum the Fourier series to produce a simpler formula for the solution, there are a number of useful observations that can be gleaned from this representation.

In the case of the heat equation, the solutions decay exponentially fast to thermal equilibrium, at a rate governed by the smallest positive eigenvalue of the associated boundary value problem. The higher order Fourier modes damp out very rapidly, which makes the heat equation a means of automatically smoothing and denoising functions representing signals and images. In the case of the wave equation, each Fourier mode vibrates with a natural frequency, and, in a stable situation, the full solution is a linear combination of these fundamental vibrational modes. For one-dimensional media, the natural frequencies are integral multiples of a single lowest frequency, and hence the solution is periodic, which, in particular, explains the tonal qualities of string and wind instruments. There is an alternative solution technique for the one-dimensional wave equation, due to d'Alembert, which leads to an explicit formula for the solution that points out the role of characteristics for signal propagation and the behavior of solutions. Both the explicit and series solution methods are useful, and shed complementary lights on the physical phenomena of vibration.

The repose of the system to a concentrated unit impulse leads to the fundamental solution, which can then be used to construct integral representations of the solution to the inhomogeneous system subject to external forcing. We will also show how to exploit the symmetry properties of the differential equation in order to construct new solutions from known solutions. While less powerful than separation of variables ofr linear partial differential equations, the symmetry method is, in fact, the one technique that can be directly applied to nonlinear problems, where it often assumes a central role in the construction of explicit solutions.

Finally, we present and analyze several basic numerical solution methods for both the heat and the wave equation. We begin with a general discussion of finite difference formulae that are used to numerically approximate derivatives of. Many important numerical solution algorithms for differential equations are obtained by replacing the derivatives by the appropriate numerical differentiation formulae. However, the resulting *finite difference scheme* is not necessarily guaranteed to accurately approximate the actual analytic solution to the differential equation; further analysis is required to elicit bona fide, convergent numerical algorithms. In the cases considered here, the finite difference schemes replace the partial differential equation by an iterative linear matrix system. The basic results from Chapter 10 are then brought to bear on understanding convergence and stability of the different numerical solution algorithms.

14.1. The Diffusion and Heat Equations.

Let us begin with a physical derivation of the heat equation from first principles of thermodynamics. The reader solely interested in the mathematical developments can skip ahead to the following subsection. However, physical insight can often play an critical role in understanding the underlying mathematics, and is neglected at one's peril.

We consider a bar — meaning a thin, heat-conducting body of length ℓ. "Thin" means that we can regard the bar as a one-dimensional continuum with no significant transverse temperature variation. We use $0 \le x \le \ell$ to denote the position along the bar. Our goal is to find the temperature $u(t,x)$ of the bar at position x and time t. The dynamical equations governing the temperature are based on three fundamental physical laws.

The first law is that, in the absence of external sources, thermal energy can only enter the bar through its ends. In physical terms, we are assuming that the bar is fully insulated along its length. Let $\varepsilon(t,x)$ to denote the thermal energy in the bar at position x and time t. Consider a small section of the bar lying between x and $x+\Delta x$. The total amount of heat energy contained in this section is obtained by integrating (summing): $\displaystyle\int_x^{x+\Delta x} \varepsilon(t,y)\,dy$. Further, let $w(t,x)$ denote the *heat flux*, i.e., the rate of flow of thermal energy along the bar. We use the convention that $w(t,x) > 0$ means that the energy is moving to the right, while $w(t,x) < 0$ if it moves to the left. The first law implies that the rate of change in the thermal energy in any section of the bar is equal to the total heat flux, namely the amount of the heat passing through its ends. Therefore, in view of our sign convention on the flux,

$$\frac{\partial}{\partial t}\int_x^{x+\Delta x} \varepsilon(t,y)\,dy = -\,w(t,x+\Delta x) + w(t,x),$$

the latter two terms denoting the respective flux of heat *into* the section of the bar at its right and left ends. Assuming sufficient regularity of the integrand, we are permitted to bring the derivative inside the integral. Thus, dividing both sides of the resulting equation by Δx,

$$\frac{1}{\Delta x}\int_x^{x+\Delta x} \frac{\partial \varepsilon}{\partial t}\,(t,y)\,dy = -\,\frac{w(t,x+\Delta x)-w(t,x)}{\Delta x}\,.$$

In the limit as the length $\Delta x \to 0$, the right hand side of this equation converges to minus the x derivative of $w(t,x)$, while, by the Fundamental Theorem of Calculus, the left hand side converges to the integrand $\partial\varepsilon/\partial t$ at the point x; the net result is the fundamental differential equation

$$\frac{\partial \varepsilon}{\partial t} = -\,\frac{\partial w}{\partial x} \tag{14.1}$$

relating thermal energy ε and heat flux w. A partial differential equation of this particular form is known as a *conservation law*, and, in this instance, formulates the law of conservation of thermal energy. See Exercise ■ for details.

The second physical law is a *constitutive assumption*, based on experimental evidence. In most physical materials, thermal energy is found to be proportional to temperature,

$$\varepsilon(t,x) = \sigma(x)\,u(t,x). \tag{14.2}$$

The factor

$$\sigma(x) = \rho(x)\,\chi(x) > 0$$

is the product of the *density* ρ of the material and its *specific heat* χ, which is the amount of heat energy required to raise the temperature of a unit mass of the material by one unit. Note that we are assuming the bar is not changing in time, and so physical quantities such as density and specific heat depend only on position x. We also assume, perhaps with less physical justification, that the material properties do not depend upon the temperature; otherwise, we would be led to a much more difficult nonlinear diffusion equation.

The third physical law relates the heat flux to the temperature. Physical experiments in a wide variety of materials indicate that the heat energy moves from hot to cold at a rate that is in direct proportion to the rate of change — meaning the derivative — of the temperature. The resulting linear constitutive relation

$$w(t,x) = -\,\kappa(x)\,\frac{\partial u}{\partial x} \tag{14.3}$$

is known as *Fourier's Law of Cooling*. The proportionality factor $\kappa(x) > 0$ is called the *thermal conductivity* of the bar at position x. A good heat conductor, e.g., silver, will have high conductivity, while a poor conductor, e.g., glass, will have low conductivity. The minus sign tells us that heat energy moves from hot to cold; if $\dfrac{\partial u}{\partial x}(t,x) > 0$ the temperature is increasing from left to right, and so the heat energy moves back to the left, with consequent flux $w(t,x) < 0$.

Combining the three laws (14.1), (14.2) and (14.3) produces the basic partial differential equation

$$\frac{\partial}{\partial t}\big[\,\sigma(x)\,u\,\big] = \frac{\partial}{\partial x}\left(\kappa(x)\,\frac{\partial u}{\partial x}\right), \qquad 0 < x < \ell, \tag{14.4}$$

governing the diffusion of heat in a non-uniform bar. The resulting *linear diffusion equation* is used to model a variety of diffusive processes, including heat flow, chemical diffusion, population dispersion, and the spread of infectious diseases. If, in addition, we allow external heat sources $h(t,x)$, then the linear diffusion equation acquires an inhomogeneous term:

$$\frac{\partial}{\partial t}\big[\,\sigma(x)\,u\,\big] = \frac{\partial}{\partial x}\left(\kappa(x)\,\frac{\partial u}{\partial x}\right) + h(t,x), \qquad 0 < x < \ell. \tag{14.5}$$

In order to uniquely prescribe the solution $u(t,x)$ to the diffusion equation, we need to specify the initial temperature distribution

$$u(t_0,x) = f(x), \qquad 0 \le x \le \ell, \tag{14.6}$$

along the bar at an initial time t_0. In addition, we mujst impose suitable boundary conditions at the two ends of the bar. As with the equilibrium equations discussed in Chapter 11, there are three common physical types. The first is a *Dirichlet boundary condition*, where an end of the bar is held at prescribed temperature. Thus, the boundary condition

$$u(t,0) = \alpha(t) \tag{14.7}$$

fixes the temperature at the left hand end of the bar. Alternatively, the *Neumann boundary condition*

$$\frac{\partial u}{\partial x}\,(t,0) = \xi(t) \tag{14.8}$$

prescribes the heat flux $w(t,0) = -\kappa(0)\,\dfrac{\partial u}{\partial x}\,(t,0)$ at the left hand end. In particular, the homogeneous Neumann condition with $\xi(t) \equiv 0$ corresponds to an insulated end, where no heat can flow in or out. Each end of the bar should have one or the other of these boundary conditions. For example, a bar with both ends having prescribed temperatures is governed by the pair of Dirichlet boundary conditions

$$u(t,0) = \alpha(t), \qquad u(t,\ell) = \beta(t), \tag{14.9}$$

whereas a bar with two insulated ends requires two homogeneous Neumann boundary conditions

$$\frac{\partial u}{\partial x}\,(t,0) = 0, \qquad \frac{\partial u}{\partial x}\,(t,\ell) = 0. \tag{14.10}$$

The mixed case, with one end fixed and the other insulated, is similarly formulated. Finally, the *periodic boundary conditions*

$$u(t,0) = u(t,\ell), \qquad \frac{\partial u}{\partial x}\,(t,0) = \frac{\partial u}{\partial x}\,(t,\ell), \tag{14.11}$$

correspond to a circular *ring* of length ℓ. As before, we are assuming the heat is only allowed to flow around the ring — insulation prevents any radiation of heat from one side of the ring to the other.

The Heat Equation

In this book, we will retain the term "heat equation" to refer to the homogeneous case, in which the bar is made of a uniform material, and so its density ρ, conductivity κ,

and specific heat χ are all positive constants. Under these assumptions, the homogeneous diffusion equation (14.4) reduces to the *heat equation*

$$\frac{\partial u}{\partial t} = \gamma \, \frac{\partial^2 u}{\partial x^2} \tag{14.12}$$

for the temperature $u(t,x)$ in the bar. The constant

$$\gamma = \frac{\kappa}{\sigma} = \frac{\kappa}{\rho \, \chi} \tag{14.13}$$

is called the *thermal diffusivity* of the bar, and incorporates all of its relevant physical properties. The solution $u(t,x)$ will be uniquely prescribed once we specify initial conditions (14.6) and a suitable pair of boundary conditions at the ends of the bar.

As we learned in Section 12.1, the elementary, separable solutions to the heat equation are based on the exponential ansatz

$$u(t,x) = e^{-\lambda t} \, v(x), \tag{14.14}$$

where $v(x)$ is a time-independent function. Substituting the solution formula (14.14) into (14.12) and canceling the common exponential factors, we find that $v(x)$ must solve the ordinary differential equation

$$-\gamma \, v'' = \lambda \, v.$$

In other words, v is an *eigenfunction* with *eigenvalue* λ, for the second derivative operator $K = -\gamma \, D^2$. Once we determine the eigenvalues and eigenfunctions, we will be able to reconstruct the solution $u(t,x)$ as a linear combination, or, rather, infinite series in the corresponding separable eigenfunction solutions.

Let us consider the simplest case of a uniform bar held at zero temperature at each end. For simplicity, we take the initial time to be $t_0 = 0$, and so the initial and boundary conditions are

$$\begin{aligned} u(t,0) = 0, \qquad u(t,\ell) = 0, \qquad & t \geq 0, \\ u(0,x) = f(x), \qquad & 0 < x < \ell. \end{aligned} \tag{14.15}$$

According to the general prescription, we need to solve the eigenvalue problem

$$\gamma \, \frac{d^2 v}{dx^2} + \lambda \, v = 0, \qquad v(0) = 0, \qquad v(\ell) = 0. \tag{14.16}$$

As noted in Exercise ■, positive definiteness of the underlying differential operator $K = -\gamma \, D^2$ when subject to Dirichlet boundary conditions implies that we need only look for positive eigenvalues: $\lambda > 0$. In Exercises ■, ■, the skeptical reader is asked to check explicitly that if $\lambda \leq 0$ or λ is complex, then the boundary value problem (14.16) admits only the trivial solution $v(x) \equiv 0$.

Setting $\lambda = \gamma \, \omega^2$ with $\omega > 0$, the general solution to the differential equation is a trigonometric function

$$v(x) = a \cos \omega \, x + b \sin \omega \, x,$$

where a, b are arbitrary constants whose values are specified by the boundary conditions. The boundary condition at $x = 0$ requires $a = 0$. The second boundary condition requires

$$v(\ell) = b \sin \omega \ell = 0.$$

Therefore, $\omega \ell$ must be an integer multiple of π, and so

$$\omega \;=\; \frac{\pi}{\ell}, \qquad \frac{2\pi}{\ell}, \qquad \frac{3\pi}{\ell}, \qquad \dots .$$

We conclude that the eigenvalues and eigenfunctions of the boundary value problem (14.16) are

$$\lambda_n = \gamma \left(\frac{n\,\pi}{\ell}\right)^2, \qquad v_n(x) = \sin \frac{n\,\pi\,x}{\ell}, \qquad n = 1, 2, 3, \dots . \tag{14.17}$$

The corresponding separable solutions (14.14) to the heat equation with the given boundary conditions are

$$u_n(t,x) = \exp\left(-\frac{\gamma\, n^2\,\pi^2\, t}{\ell^2}\right) \sin \frac{n\,\pi\,x}{\ell}, \qquad n = 1, 2, 3, \dots . \tag{14.18}$$

Each represents a trigonometrically oscillating temperature profile that maintains its form while decaying at an exponential rate to zero. The first of these,

$$u_1(t,x) = \exp\left(-\frac{\gamma\,\pi^2\, t}{\ell^2}\right) \sin \frac{\pi\,x}{\ell},$$

experiences the slowest decay. The higher "frequency" modes $u_n(t,x)$, $n \geq 2$, all go to zero at a faster rate, with those having a highly oscillatory temperature profile, where $n \gg 0$, effectively disappearing almost instantaneously. Thus, small scale temperature fluctuations tend to rapidly cancel each other out through diffusion of heat energy.

Linear superposition is used to assemble the general series solution

$$u(t,x) = \sum_{n=1}^{\infty} b_n\, u_n(t,x) = \sum_{n=1}^{\infty} b_n \exp\left(-\frac{\gamma\, n^2\,\pi^2\, t}{\ell^2}\right) \sin \frac{n\,\pi\,x}{\ell} \tag{14.19}$$

as a combination of the separable solutions. Assuming that the series converges, the initial temperature profile is

$$u(0,x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\,\pi\,x}{\ell} = f(x). \tag{14.20}$$

This has the form of a Fourier sine series (12.36) on the interval $[0, \ell]$ for the initial temperature profile $f(x)$. By orthogonality of the eigenfunctions — which is a direct consequence of the self-adjointness of the underlying boundary value problem (14.16) — the coefficients are determined by the inner product formulae (12.37), and so

$$b_n = \frac{2}{\ell} \int_0^{\ell} f(x) \sin \frac{n\,\pi\,x}{\ell}\, dx, \qquad n = 1, 2, 3, \dots . \tag{14.21}$$

The resulting solution (14.19) describes the Fourier sine series for the temperature $u(t,x)$ of the bar at each later time $t \geq 0$. It can be rigorously proved that, for quite general initial conditions, the Fourier series does indeed converge to a solution to the initial-boundary value problem, [**146**].

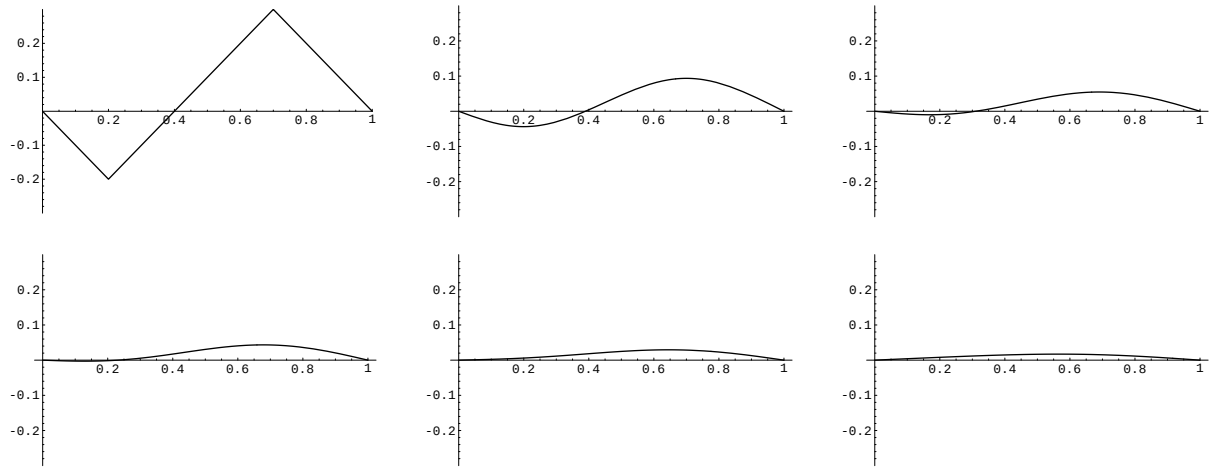


Figure 14.1. A Solution to the Heat Equation.

Example 14.1. Consider the initial temperature profile

$$u(0,x) = f(x) = \begin{cases} -x, & 0 \le x \le \frac{1}{5}, \\ x - \frac{2}{5}, & \frac{1}{5} \le x \le \frac{7}{10}, \\ 1 - x, & \frac{7}{10} \le x \le 1, \end{cases} \tag{14.22}$$

on a bar of length 1, plotted in the first graph in Figure 14.1. Using (14.21), the first few Fourier coefficients of $f(x)$ are computed as

$$\begin{array}{llll} b_1 = .0448\ldots, & b_2 = -.096\ldots, & b_3 = -.0145\ldots, & b_4 = 0, \\ b_5 = -.0081\ldots, & b_6 = .0066\ldots, & b_7 = .0052\ldots, & b_8 = 0, \end{array} \quad \ldots \ .$$

Setting $\gamma = 1$, the resulting Fourier series solution to the heat equation is

$$\begin{aligned} u(t,x) &= \sum_{n=1}^{\infty} b_n\, u_n(t,x) = \sum_{n=1}^{\infty} b_n\, e^{-n^2\pi^2 t}\, \sin n\pi x \\ &= .0448\, e^{-\pi^2 t}\, \sin \pi x - .096\, e^{-4\pi^2 t}\, \sin 2\pi x - .0145\, e^{-9\pi^2 t}\, \sin 3\pi x - \ \cdots \ . \end{aligned}$$

In Figure 14.1, the solution is plotted at the successive times $t = ., .02, .04, \ldots, .1$. Observe that the corners in the initial data are immediately smoothed out. As time progresses, the solution decays at an exponential rate of $\pi^2 \approx 9.87$ to a uniform, zero temperature, which is the equilibrium temperature distribution for the homogeneous Dirichlet boundary conditions. As the solution decays to thermal equilibrium, it also assumes the progressively more symmetric shape of a single sine arc, of exponentially decreasing amplitude, which is merely the profile of the first term in its Fourier series.

Smoothing and Long Time Behavior

The fact that we can write the solution to an initial-boundary value problem in the form of an infinite series is progress of a sort. However, because it cannot be summed in closed form, this "solution" is considerably less satisfying than having a direct, explicit

formula. Nevertheless, there are important qualititative and quantitative features of the solution that can be easily gleaned from such series expansions.

If the initial data $f(x)$ is piecewise continuous, then its Fourier coefficients are uniformly bounded; indeed, for any $n \geq 1$,

$$| b_n | \;\leq\; \frac{2}{\ell} \int_0^\ell \left| f(x) \sin \frac{n \pi x}{\ell} \right| dx \;\leq\; \frac{2}{\ell} \int_0^\ell | f(x) | \, dx \equiv M. \tag{14.23}$$

This property holds even for quite irregular data; for instance, the Fourier coefficients (12.50) of the delta function are also uniformly bounded. Under these conditions, each term in the series solution (14.19) is bounded by an exponentially decaying function

$$\left| \; b_n \exp\left(- \frac{\gamma \, n^2 \pi^2}{\ell^2} \, t \right) \sin \frac{n \pi x}{\ell} \; \right| \;\leq\; M \exp\left(- \frac{\gamma \, n^2 \pi^2}{\ell^2} \, t \right).$$

This means that, as soon as $t > 0$, most of the high frequency terms, $n \gg 0$, will be extremely small. Only the first few terms will be at all noticeable, and so the solution essentially degenerates into a finite sum over the first few Fourier modes. As time increases, more and more of the Fourier modes will become negligible, and the sum further degenerates into progressively fewer significant terms. Eventually, as $t \to \infty$, *all* of the Fourier modes will decay to zero. Therefore, the solution will converge exponentially fast to a zero temperature profile: $u(t, x) \to 0$ as $t \to \infty$, representing the bar in its final uniform thermal equilibrium. The fact that its equilibrium temperature is zero is a direct consequence of the fact that we are holding both ends of the bar fixed at zero temperature — any initial heat in the bar will eventually be dissipated away through its two ends. The last term to disappear is the one with the slowest decay, namely

$$u(t, x) \;\approx\; b_1 \exp\left(- \frac{\gamma \, \pi^2}{\ell^2} \, t \right) \sin \frac{\pi x}{\ell}, \qquad \text{where} \qquad b_1 = \frac{1}{\pi} \int_0^\pi f(x) \sin x \, dx. \tag{14.24}$$

Generically, $b_1 \neq 0$, and the solution approaches thermal equilibrium exponentially fast with rate equal to the smallest eigenvalue, $\lambda_1 = \gamma \, \pi^2 / \ell^2$, which is proportional to the thermal diffusivity divided by the square of the length of the bar. The longer the bar, or the smaller the diffusivity, the longer it takes for the effect of holding the ends at zero temperature to propagate along the entire bar. Also, again provided $b_1 \neq 0$, the asymptotic shape of the temperature profile is a small sine arc, just as we observed in Example 14.1. In exceptional situations, namely when $b_1 = 0$, the solution decays even faster, at a rate equal to the eigenvalue $\lambda_k = \gamma \, k^2 \pi^2 / \ell^2$ corresponding to the first nonzero term, $b_k \neq 0$, in the series; its asymptotic shape now oscillates k times over the interval.

The heat equation's smoothing effect on irregular initial data by fast damping of the high frequency modes underlies its effectiveness for smoothing out and denoising signals. We take the initial data $u(0, x) = f(x)$ to be a noisy signal, and then evolve the heat equation forward to a prescribed time $t_\star > 0$. The resulting function $g(x) = u(t_\star, x)$ will be a smoothed version of the original signal $f(x)$ in which most of the high frequency noise has been eliminated. Of course, if we run the heat flow for too long, all of the low frequency features will be also be smoothed out and the result will be a uniform, constant signal. Thus, the choice of stopping time $t_\star$ is crucial to the success of this

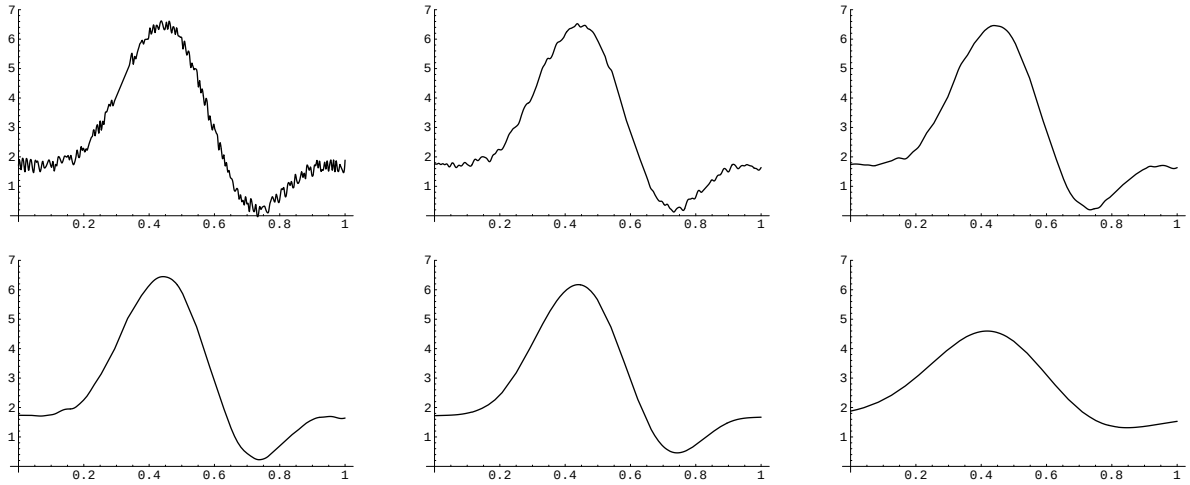


Figure 14.2. Denoising a Signal Using the Heat Equation.

method. Figure 14.2 shows the effect running the heat equation, with $\gamma = 1$, to times $t = 0., .00001, .00005, .0001, .001, .01$ on the same signal from Figure 13.6. Observe how quickly the noise is removed. By the final time, the overall smoothing effect of the heat flow has caused significant degradation (blurring) of the original signal. The heat equation approach to denoising has the advantage that no Fourier coefficients need be explicitly computed, nor does one need to reconstruct the smoothed signal from its remaining Fourier coefficients. The final section discusses some numerical methods that can be used to solve the heat equation directly.

Another, closely related observation is that, for any fixed time $t > 0$ after the initial moment, the coefficients in the Fourier series (14.19) decay exponentially fast as $n \to \infty$. According to the discussion at the end of Section 12.3, this implies that the solution $u(t, x)$ is a very smooth, infinitely differentiable function of x at each positive time t, *no matter how unsmooth the initial temperature profile*. We have discovered the basic smoothing property of heat flow.

Theorem 14.2. *If $u(t, x)$ is a solution to the heat equation with piecewise continuous initial data $f(x) = u(0, x)$, or, more generally, initial data satisfying* (14.23)*, then, for any $t > 0$, the solution $u(t, x)$ is an infinitely differentiable function of x.*

After even a very short amount of time, the heat equation smoothes out most, and, eventually, all of the fluctuations in the initial temperature profile. As a consequence, it becomes impossible to reconstruct the initial temperature $u(0, x) = f(x)$ by measuring the temperature distribution $h(x) = u(t, x)$ at a later time $t > 0$. *Diffusion is irreversible* — we cannot run the heat equation backwards in time! Indeed, if the initial data $u(0, x) = f(x)$ is not smooth, there is *no* function $u(t, x)$ for $t < 0$ that could possibly yield such an initial distribution because all corners and singularities are smoothed out by the diffusion process as t goes forward! Or, to put it another way, the Fourier coefficients (14.21) of any purported solution will be exponentially growing when $t < 0$, and so high frequency noise will completely overwhelm the solution. For this reason, the backwards heat equation is said to be *ill-posed*.

On the other hand, the unsmoothing effect of the backwards heat equation does have potential benefits. For example, in image processing, diffusion will gradually blur an image. Image enhancement is the reverse process, and so can be done by running the heat flow backwards in some well-prescribed manner. For instance, one can restrict to the first few Fourier modes, and then the corresponding backwards evolution is well-defined. Similar problems occur in the reconstruction of subterranean profiles from seismic data, a problem of great concern in the oil and gas industry. In forensics, determining the time of death based on the current temperature of a corpse also requires running the equations governing the dissipation of body heat backwards in time. For these and other applications, a key issue in contemporary research is how to cleverly circumventing the ill-posedness of the backwards heat flow.

Remark: The irreversibility of the heat equation points out a crucial distinction between partial differential equations and ordinary differential equations. Ordinary differential equations are always reversible — unlike the heat equation, existence, uniqueness and continuous dependence properties of solutions are all equally valid in reverese time (although the detailed qualitative and quantitative properties of solutions can very well depend upon whether time is running forwards or backwards). The irreversibility of partial differential equations modeling the diffusive processes in our universe may well be why Time's Arrow points only to the future.

The Heated Ring

Let us next consider the periodic boundary value problem modeling heat flow in an insulated circular ring. Let us fix the length of the ring to be $\ell = 2\pi$, with $-\pi < x < \pi$ representing "angular" coordinate around the ring. For simplicity, we also choose units in which the thermal diffusivity is $\gamma = 1$. Thus, we seek to solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\ , \qquad -\pi < x < \pi, \qquad t > 0, \tag{14.25}$$

subject to periodic boundary conditions

$$u(t, -\pi) = u(t, \pi), \qquad \frac{\partial u}{\partial x}\,(t, -\pi) = \frac{\partial u}{\partial x}\,(t, \pi), \qquad t \geq 0. \tag{14.26}$$

The initial temperature distribution is

$$u(0, x) = f(x), \qquad -\pi < x < \pi. \tag{14.27}$$

The resulting temperature $u(t, x)$ will be a periodic function in x of period 2π.

Substituting the separable solution ansatz $u(t, x) = e^{-\lambda t} v(x)$ into the heat equation and the boundary conditions leads to the periodic eigenvalue problem

$$\frac{d^2 v}{dx^2} + \lambda\, v = 0, \qquad v(-\pi) = v(\pi), \qquad v(-\pi) = v(\pi). \tag{14.28}$$

As we know, in this case the eigenvalues are $\lambda_n = n^2$ where $n = 0, 1, 2, \ldots$ is a non-negative integer, and the corresponding eigenfunction solutions are the trigonometric functions

$$v_n(x) = \cos n\, x, \qquad \widetilde{v}_n(x) = \sin n\, x, \qquad n = 0, 1, 2, \ldots\ .$$

Note that $\lambda_0 = 0$ is a simple eigenvalue, with constant eigenfunction $\cos 0x = 1$ — the sine solution $\sin 0x \equiv 0$ is trivial — while the positive eigenvalues are, in fact, double, each possessing two linearly independent eigenfunctions. The corresponding separable solutions to the heated ring equation are

$$u_n(t,x) = e^{-n^2 t}\cos n\,x, \qquad \widetilde{u}_n(t,x) = e^{-n^2 t}\sin n\,x, \qquad n = 0, 1, 2, 3, \dots .$$

The resulting infinite series solution is

$$u(t,x) = \tfrac{1}{2}\,a_0 + \sum_{n=1}^{\infty} \big(a_n\, e^{-n^2 t}\cos n\,x + b_n\, e^{-n^2 t}\sin n\,x \big). \tag{14.29}$$

The initial conditions require

$$u(0,x) = \tfrac{1}{2}\,a_0 + \sum_{n=1}^{\infty} \big(a_n \cos n\,x + b_n \sin n\,x \big) = f(x), \tag{14.30}$$

which is precisely the Fourier series of the initial temperature profile $f(x)$. Consequently,

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos n\,x\,dx, \qquad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin n\,x\,dx, \tag{14.31}$$

are the usual Fourier coefficients of $f(x)$.

As in the Dirichlet problem, after the initial instant, the high frequency terms in the series (14.29) become extremely small, since $e^{-n^2 t} \ll 1$ for $n \gg 0$. Therefore, as soon as $t > 0$, the solution essentially degenerates into a finite sum over the first few Fourier modes. Moreover, as $t \to \infty$, *all* of the Fourier modes will decay to zero with the exception of the constant one, with null eigenvalue $\lambda_0 = 0$. Therefore, the solution will converge exponentially fast to a constant temperature profile:

$$u(t,x) \;\longrightarrow\; \frac{1}{2}\,a_0 \;=\; \frac{1}{2\,\pi}\int_{-\pi}^{\pi} f(x)\,dx,$$

which equals the *average* of the initial temperature profile. Physically, we observe that the heat energy is redistributed so that the ring achieves a uniform constant temperature and is in thermal equilibrium. Indeed, the total heat energy

$$E = \int_{-\pi}^{\pi} u(t,x)\,dx = \text{constant} \tag{14.32}$$

is conserved, meaning constant, for all time; the proof of this fact is left as an Exercise .

Prior to equilibrium, only the lowest frequency Fourier modes will still be noticeable, and so the solution will asymptotically look like

$$u(t,x) \approx \tfrac{1}{2}\,a_0 + e^{-t}(a_1\cos x + b_1 \sin x) = \tfrac{1}{2}\,a_0 + r_1\, e^{-t}\cos(x + \delta_1), \tag{14.33}$$

where

$$a_1 = r_1\cos\delta_1 = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} f(x)\cos x\,dx, \qquad b_1 = r_1\sin\delta_1 = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} f(x)\sin x\,dx.$$

Thus, for most initial data, the solution approaches thermal equilibrium exponentially fast, at a unit rate. The exceptions are when $r_1 = \sqrt{a_1^2 + b_1^2} = 0$, for which the rate of convergence is even faster, namely at a rate e^{-kt} where k is the smallest integer such that $r_k = \sqrt{a_k^2 + b_k^2} \neq 0$.

Inhomogeneous Boundary Conditions

So far, we have concentrated our attention on homogeneous boundary conditions. There is a simple trick that will convert a boundary value problem with inhomogeneous but constant Dirichlet boundary conditions,

$$u(t,0) = \alpha, \qquad u(t,\ell) = \beta, \qquad t \geq 0, \tag{14.34}$$

into a homogeneous Dirichlet problem. We begin by solving for the equilibrium temperature profile, which is the affine function

$$u_\star(x) = \alpha + \frac{\beta - \alpha}{\ell}\, x. \tag{14.35}$$

The difference

$$\widetilde{u}(t,x) = u(t,x) - u_\star(x) = u(t,x) - \alpha - \frac{\beta - \alpha}{\ell}\, x \tag{14.36}$$

measures the deviation of the solution from equilibrium. It clearly satisfies the homogeneous boundary conditions at both ends:

$$\widetilde{u}(t,0) = 0 = \widetilde{u}(t,\ell).$$

Moreover, by linearity, since both $u(t,x)$ and $u_\star(x)$ are solutions to the heat equation, so is $\widetilde{u}(t,x)$. The initial data must be similarly adapted:

$$\widetilde{u}(0,x) = \widetilde{f}(x) = f(x) - u_\star(x) = f(x) - \alpha - \frac{\beta - \alpha}{\ell}\, x.$$

Solving the resulting homogeneous initial value problem, we write $\widetilde{u}(t,x)$ in Fourier series form (14.19), where the Fourier coefficients are computed from the modified initial data $\widetilde{f}(x)$. The solution to the inhomogeneous boundary value problem thus has the series form

$$u(t,x) = \alpha + \frac{\beta - \alpha}{\ell}\, x \;+\; \sum_{n=1}^{\infty} \widetilde{b}_n \exp\left(-\frac{\gamma\, n^2 \pi^2}{\ell^2}\, t \right) \sin \frac{n\pi x}{\ell}\,, \tag{14.37}$$

where

$$\widetilde{b}_n = \frac{2}{\ell} \int_0^\ell \widetilde{f}(x) \sin \frac{n\pi x}{\ell}\, dx, \qquad n = 1, 2, 3, \dots\ . \tag{14.38}$$

Since, for any reasonable initial data, $\widetilde{u}(t,0) \to 0$ will decay to zero at an exponential rate as $t \to \infty$, the actual temperature profile (14.37) will asymptotically decay to the equilibrium profile,

$$u(t,x) \quad \longrightarrow \quad u_\star(x) = \alpha \;+\; \frac{\beta - \alpha}{\ell}\, x$$

at the same exponentially fast rate.

This method does not apply when the boundary conditions are time-dependent: $u(t,0) = \alpha(t)$, $u(t,\ell) = \beta(t)$. Attempting to mimic the preceding technique, we discover that the deviation

$$\widetilde{u}(t,x) = u(t,x) - u_\star(t,x), \qquad \text{where} \qquad u_\star(t,x) = \alpha(t) + \frac{\beta(t) - \alpha(t)}{\ell}\, x, \tag{14.39}$$

does satisfy the homogeneous boundary conditions, but solves an inhomogeneous version of the heat equation:

$$\frac{\partial \widetilde{u}}{\partial t} = \frac{\partial^2 \widetilde{u}}{\partial x^2} - h(t,x), \qquad \text{where} \qquad h(t,x) = \frac{\partial h_\star}{\partial t}\,(t,x). \tag{14.40}$$

Solution techniques in this case will be discussed below.

14.2. Symmetry and Similarity.

So far we have relied almost exclusively on the method of separation of variables to construct explicit solutions to partial differential equations. Beyond this, the most useful solution technique relies on exploiting inherent symmetry properties of the differential equation. Unlike separation of variables†, symmetry methods can be also successfully applied to produce solutions to a broad range of nonlinear partial differential equations; some simple examples can be found in Chapter 22. While we do not have the space or required mathematical tools to develop the full apparatus of symmetry techniques, we can introduce the important concept of a *similarity solution*, applied in the particular context of the heat equation.

In general, by a *symmetry* of an equation, we mean a transformation, either lineqr as in Section 7.2, affine, as in Section 7.3, or even nonlinear, that takes solutions to solutions. Thus, if we know a symmetry, and one solution, then we can construct a second solution by applying the symmetry. And, possibly, a third solution by applying the symmetry yet again. And so on. If we know lots of symmetries, then we can produce lots and lots of solutions by this simple device.

Remark: General symmetry techniques are founded on the theory of Lie groups, named after the influential nineteenth century Norwegian mathematician Sophus Lie (pronounced "Lee"). Lie's theory provides an algorithm for completely determining all the symmetries of a given differential equation, but this is beyond the scope of this introductory text. However, direct inspection and/or physical intuition will often detect the most important symmetries without appealing to such a sophisticated theory. Modern applications of Lie's symmetry methods to partial differential equations arising in physics and engineering can be traced back to the influential book of G. Birkhoff, [**18**], on hydrodynamics. A complete and comprehensive treatment of symmetry methods can be found in the first author's book [**117**], and, at a more introductory level, in the recent books by Cantwell, [**32**], with particular emphasis on the equations of fluid mechanics, and Hydon, [**84**].

† This is not quite fair: separation of variables can be applied to a few very special partial differential equations such as Hamilton–Jacobi equations, [**105**].

The heat equation serves as an excellent testing ground for the general symmetry methodology, as it admits a rich variety of symmetry transformations that take solutions to solutions. The simplest are the translations. Moving the space and time coordinates by a fixed amount,

$$t \mapsto t - a, \qquad x \mapsto x - b, \tag{14.41}$$

where a, b are constants, changes the function $u(t, x)$ into the translated function

$$U(t, x) = u(t - a, x - b). \tag{14.42}$$

A simple application of the chain rule proves that the partial derivatives of U with respect to t and x agree with the corresponding partial derivatives of u, so

$$\frac{\partial U}{\partial t} = \frac{\partial u}{\partial t}\,, \qquad \frac{\partial U}{\partial x} = \frac{\partial u}{\partial x}\,, \qquad \frac{\partial^2 U}{\partial x^2} = \frac{\partial^2 u}{\partial x^2}\,,$$

and so on. In particular, the function $U(t, x)$ is a solution to the heat equation $U_t = \gamma\, U_{xx}$ whenever $u(t, x)$ also solves $u_t = \gamma\, u_{xx}$. Physically, the translational symmetry formalizes the property that the heat equation models a homogeneous medium, and hence the solution does not depend on the choice of reference point, i.e., the origin of our coordinate system.

As a consequence, each solution to the heat equation will produce an infinite family of translated solutions. For example, starting with the separable solution

$$u(t, x) = e^{-\gamma\, t} \sin x,$$

we immediately produce the additional solutions

$$u(t, x) = e^{-\gamma\,(t-a)} \sin \pi\,(x - b),$$

valid for any choice of constants a, b.

Warning: Typically, the symmetries of a differential equation do not respect initial or boundary conditions. For instance, if $u(t, x)$ is defined for $t > 0$ and in the domain $0 \le x \le \ell$, then its translated version $U(t, x)$ is defined for $t > a$ and in the translated domain $b \le x \le \ell + b$, and so will solve an appropriately translated initial-boundary value problem.

A second, even more important class of symmetries are the scaling invariances. We already know that if $u(t, x)$ is a solution, so is any scalar multiple $c\, u(t, x)$; this is a simple consequence of linearity of the heat equation. We can also add an arbitrary constant to the temperature, noting that

$$U(t, x) = c\, u(t, x) + k \tag{14.43}$$

is a solution for any choice of constants c, k. Physically, the transformation (14.43) amounts to a change in the scale for measuring temperature. For instance, if u is measured degrees Celsius, and we set $c = \frac{9}{5}$ and $k = 32$, then $U = \frac{9}{5}\, u + 32$ will be measured in degrees Fahrenheit. Thus, reassuringly, the physical processes described by the heat equation do not depend upon our choice of thermometer.

More interestingly, suppose we rescale the space and time variables:

$$t \mapsto \alpha\, t, \qquad\qquad x \mapsto \beta\, x, \tag{14.44}$$

where $\alpha, \beta > 0$ are positive constants. The effect of such a scaling transformation is to change the function $u(t,x)$ into a rescaled function

$$U(t,x) = u(\alpha\, t, \beta\, x). \tag{14.45}$$

The derivatives of U are related to those of u according to the following formulae, which are direct consequences of the multi-variable chain rule:

$$\frac{\partial U}{\partial t} = \alpha\, \frac{\partial u}{\partial t}\,, \qquad\qquad \frac{\partial U}{\partial x} = \beta\, \frac{\partial u}{\partial x}\,, \qquad\qquad \frac{\partial^2 U}{\partial x^2} = \beta^2\, \frac{\partial^2 u}{\partial x^2}\,.$$

Therefore, if u satisfies the heat equation $u_t = \gamma\, u_{xx}$, then U satisfies the rescaled heat equation

$$U_t = \alpha\, u_t = \alpha\, \gamma\, u_{xx} = \frac{\alpha\, \gamma}{\beta^2}\, U_{xx},$$

which we rewrite as

$$U_t = \Gamma\, U_{xx}, \qquad \text{where} \qquad \Gamma = \frac{\gamma\, \alpha}{\beta^2}\,. \tag{14.46}$$

Thus, the net effect of scaling space and time is merely to rescale the diffusion coefficient in the heat equation.

Remark: Physically, the scaling symmetry (14.44) corresponds to a change in the physical units used to measure time and distance. For instance, to change from seconds to minutes, set $\alpha = 60$, and from meters to yards, set $\beta = 1.0936$. The net effect (14.46) on the diffusion coefficient is a reflection of its physical units, namely distance^2/time.

In particular, if we choose

$$\alpha = \frac{1}{\gamma}\,, \qquad\qquad \beta = 1,$$

then the rescaled diffusion coefficient becomes $\Gamma = 1$. This observation has the following important consequence. If $U(t,x)$ solves the heat equation for a unit diffusivity, $\Gamma = 1$, then

$$u(t,x) = U(\gamma\, t, x) \tag{14.47}$$

solves the heat equation for the diffusivity γ. Thus, the only effect of the diffusion coefficient γ is to speed up or slow down time! A body with diffusivity $\gamma = 2$ will cool down twice as fast as a body (of the same shape subject to the same boundary conditions and initial conditions) with diffusivity $\gamma = 1$. Note that this particular rescaling has not altered the space coordinates, and so $U(t,x)$ is defined on the same domain as $u(t,x)$.

On the other hand, if we set $\alpha = \beta^2$, then the rescaled diffusion coefficient is exactly the same as the original: $\Gamma = \gamma$. Thus, the transformation

$$t \mapsto \beta^2\, t, \qquad\qquad x \mapsto \beta\, x, \tag{14.48}$$

does not alter the equation, and hence defines a *scaling symmetry*, also known as a *similarity transformation*, for the heat equation. Combining (14.48) with the linear rescaling $u \mapsto c\,u$, we make the elementary, but important observation that if $u(t,x)$ is any solution to the heat equation, then so is the function

$$U(t,x) = c\,u(\beta^2\,t, \beta\,x), \tag{14.49}$$

for the *same* diffusion coefficient γ.

Warning: As in the case of translations, rescaling space by a factor $\beta \neq 1$ will alter the domain of definition of the solution. If $u(t,x)$ is defined for $0 \leq x \leq \ell$, then $U(t,x)$ is defined for $0 \leq x \leq \ell/\beta$.

In particular, suppose that we have solved the heat equation for the temperature $u(t,x)$ on a bar of length 1, subject to certain initial and boundary conditions. We are then given a bar composed of the same material of length 2. Since the diffusivity coefficient has not changed, and we can directly construct the new solution $U(t,x)$ by rescaling. Setting $\beta = \frac{1}{2}$ will serve to double the length. If we also rescale time by a factor $\alpha = \beta^2 = \frac{1}{4}$, then the rescaled function $U(t,x) = u\big(\frac{1}{4}\,t, \frac{1}{2}\,x\big)$ will be a solution of the heat equation on the longer bar with the same diffusivity constant. The net effect is that the rescaled solution will be evolving four times as slowly as the original solution $u(t,x,y)$. Thus, it effectively takes a bar that is double the size four times as long to cool down.

14.3. The Fundamental Solution.

One disadvantage of the Fourier series solution to the heat equation is that it is not nearly as explicit as one might desire for either practical applications, numerical computations, or even further theoretical investigations and developments. An alternative, and quite useful approach is based on the idea of the *fundamental solution*, which derives its inspiration from the Green's function method for solving boundary value problems. For the heat equation, the fundamental solution measures the effect of an initial concentrated heat source.

Let us initially restrict our attention to homogeneous boundary conditions. The idea is to first analyze the case when the initial data $u(0,x) = \delta_y(x) = \delta(x-y)$ is a delta function, which we can interpret as a highly concentrated unit heat source, e.g., a soldering iron, that is instantaneously applied at the position y along the bar. The heat will diffuse away from its initial concentration, and the resulting *fundamental solution* is denoted by

$$u(t,x) = F(t,x;y), \qquad \text{with} \qquad F(0,x;y) = \delta(x-y). \tag{14.50}$$

For each fixed y, the fundamental solution, as a function of $t > 0$ and x, must satisfy the differential equation as well as the specified homogeneous boundary conditions.

Once we have found the fundamental solution, we can then use linear superposition to reconstruct the general solution to the initial-boundary value problem. Namely, we first write the initial data

$$u(0,x) = f(x) = \int_0^\ell \delta(x-y)\,f(y)\,dy \tag{14.51}$$

as a linear superposition of delta functions, as in (11.37). Linearity implies that the solution is then the same superposition of the responses to those concentrated delta profiles:

$$u(t,x) = \int_0^\ell F(t,x;y)\, f(y)\, dy. \tag{14.52}$$

Assuming that we can differentiate under the integral sign, the fact that $F(t,x;y)$ satisfies the differential equation and the homogeneous boundary conditions for each fixed y immediately implies that the integral (14.52) is also a solution, and, moreover, has the correct initial and (homogeneous) boundary conditions.

Unfortunately, most boundary value problems do not have fundamental solutions that can be written down in closed form. An important exception is the case of an infinitely long homogeneous bar, which requires solving the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad \text{for} \qquad -\infty < x < \infty, \qquad t > 0. \tag{14.53}$$

For simplicity, we have chosen units in which the thermal diffusivity is $\gamma = 1$. The solution $u(t,x)$ is defined for all $x \in \mathbb{R}$, and has initial conditions

$$u(0,x) = f(x) \qquad \text{for} \qquad -\infty < x < \infty.$$

In order to specify the solution uniquely, we require that the temperature be square-integrable at all times, so that

$$\int_{-\infty}^{\infty} |\, u(t,x)\,|^2\, dx < \infty \qquad \text{for all} \qquad t \geq 0. \tag{14.54}$$

Thus, roughly speaking, the temperature should be small at large distances, which are the relevant boundary conditions for this situation.

On an infinite interval, the Fourier series solution to the heat equation becomes a Fourier integral. We write the initial temperature distribution as a superposition

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(k)\, e^{2\pi\, \mathrm{i}\, k\, x}\, dk,$$

of complex exponentials $e^{2\pi\, \mathrm{i}\, k\, x}$, where $\widehat{f}(k)$ is the Fourier transform (13.72) of $f(x)$. The corresponding separable solutions to the heat equation are

$$u_k(t,x) = e^{-4\pi^2 k^2 t}\, e^{2\pi\, \mathrm{i}\, k\, x} = e^{-4\pi^2 k^2 t}\, \big(\cos 2\pi\, \mathrm{i}\, k\, x + \mathrm{i}\, \sin 2\pi\, \mathrm{i}\, k\, x\,\big), \tag{14.55}$$

where the frequency variable k is allowed to assume any real value. We invoke linear superposition to combine these complex solutions into a Fourier integral

$$u(t,x) = \int_{-\infty}^{\infty} e^{-4\pi^2 k^2 t}\, e^{2\pi\, \mathrm{i}\, k\, x}\, \widehat{f}(k)\, dk \tag{14.56}$$

to form the solution to the initial value problem for the heat equation.

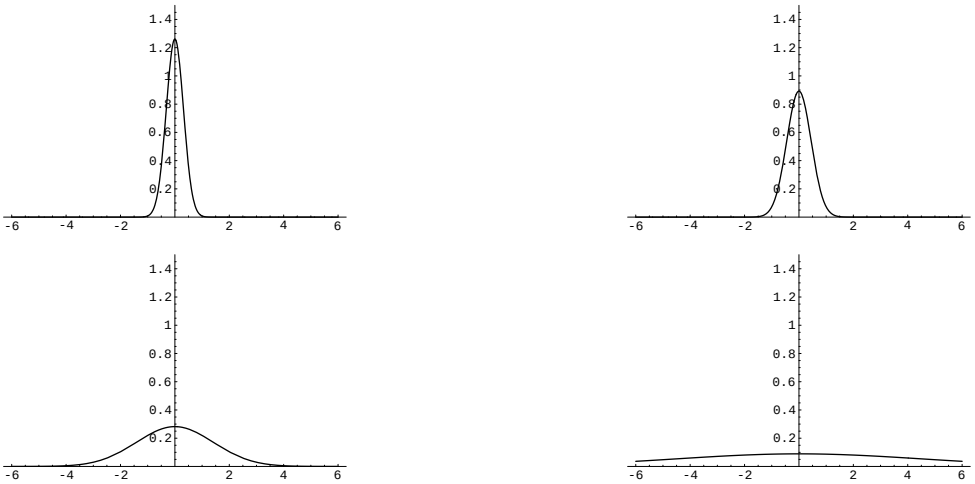


Figure 14.3. The Fundamental Solution to the Heat Equation.

In particular, to recover the fundamental solution, we take the initial temperature profile to be a delta function $\delta_y(x) = \delta(x - y)$ concentrated at $x = y$. According to (13.100), its Fourier transform is

$$\widehat{\delta}_y(k) = e^{-2\pi i k y}.$$

Plugging this into (14.56), and then referring to our table of Fourier transforms, we find the following explicit formula for the fundamental solution

$$F(t, x; y) = \int_{-\infty}^{\infty} e^{-4\pi^2 k^2 t}\, e^{2\pi i k (x-y)}\, dk = \frac{1}{2\sqrt{\pi\, t}}\, e^{-(x-y)^2/(4t)}. \tag{14.57}$$

As you can verify, for each fixed y, the function $F(t, x; y)$ is indeed a solution to the heat equation for all $t > 0$. In addition,

$$\lim_{t \to 0^+} F(t, x; y) = \begin{cases} 0, & x \neq y, \\ \infty, & x = y. \end{cases}$$

Furthermore, its integral

$$E = \int_{-\infty}^{\infty} F(t, x; y)\, dx = 1, \tag{14.58}$$

which represents the total heat energy is constant — in accordance with the law of conservation of energy, cf. Exercise ■. We conclude that, at the initial instant $t = 0$, the fundamental solution satisfies the original limiting definition (11.30), (11.31) of the delta function, and so $F(0, x; y) = \delta_y(x)$ has the desired initial temperature profile. As graphed in Figure 14.3 at times $t = .05, .1, 1., 10.$, $F(t, x; y)$ starts out as a delta spike at $x = y$ and then immediately smoothes out into a tall and narrow bell-shaped curve, centered at $x = y$. As time increases, the solution shrinks and widens, decaying everywhere to zero. Its maximal amplitude is proportional to $t^{-1/2}$, while its overall width is proportional to $t^{1/2}$. The total heat energy (14.58), which is the area under the graph, remains fixed while gradually spreading out over the entire real line.

Remark: In probability, these exponentially bell-shaped curves are known as *normal* or *Gaussian distributions*. The width of the bell curve corresponds to the *standard deviation*. For this reason, the fundamental solution to the heat equation sometimes referred to as a "Gaussian filter".

Remark: One of the non-physical artifacts of the heat equation is that the heat energy propagates with infinite speed. Indeed, the effect of any initial concentration of heat energy will immediately be felt along the entire length of an infinite bar, because, at any $t > 0$, the fundamental solution is nonzero for all x. (The graphs in Figure 14.3 are a little misleading because they fail to show the extremely small, but still positive exponentially decreasing tails in the solution.) This effect, while more or less negligible at large distances, is nevertheless in clear violation of physical intuition — not to mention relativity that postulates that signals cannot propagate faster than the speed of light. Despite this non-physical property, the heat equation remains an extremely accurate model for heat propagation and similar diffusive phenomena.

With the fundamental solution in hand, we can then adapt the linear superposition formula (14.52) to reconstruct the general solution

$$u(t,x) = \frac{1}{2\sqrt{\pi\, t}} \int_{-\infty}^{\infty} e^{-(x-y)^2/(4t)}\, f(y)\, dy \tag{14.59}$$

to our initial value problem (14.53). In other words, the solutions are obtained by convolution, cf. (13.113),

$$u(t,x) = g(t,x) * f(x), \qquad \text{where} \qquad g(t,x) = F(t,x;0) = e^{-x^2/(4t)},$$

of the initial data with a one-parameter family of progressively wider and shorter Gaussian filters. since it coincides with the heat equation solution, Gaussian filter convolution has the same smoothing effect on the initial signal $f(x)$. Indeed, the convolution integral (14.59) serves to replace each initial value $f(x)$ by a weighted average of nearby values, the weight being determined by the Gaussian distribution. The weighted averaging operation has the effect of smoothing out high frequency variations in the signal, and, consequently, the Gaussian convolution formula (14.59) provides an effective method of signal and image denoising. In fact, for practical reasons, the graphs displayed earlier in Figure 14.2 were computed by using a standard numerical integration routine to evaluate the convolution (14.59), rather than a numerical solution scheme for the heat equation.

Example 14.3. An infinite bar is initially heated to unit temperature along a finite interval. This corresponds to an initial temperature profile

$$u(0,x) = f(x) = \sigma(x-a) - \sigma(x-b) = \begin{cases} 1, & a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding solution to the heat equation is obtained by the integral formula (14.59), producing

$$u(t,x) = \frac{1}{2\sqrt{\pi\, t}} \int_a^b e^{-(x-y)^2/(4t)}\, dy = \frac{1}{2}\left[\operatorname{erf}\left(\frac{x-a}{2\sqrt{t}} \right) - \operatorname{erf}\left(\frac{x-b}{2\sqrt{t}} \right) \right], \tag{14.60}$$

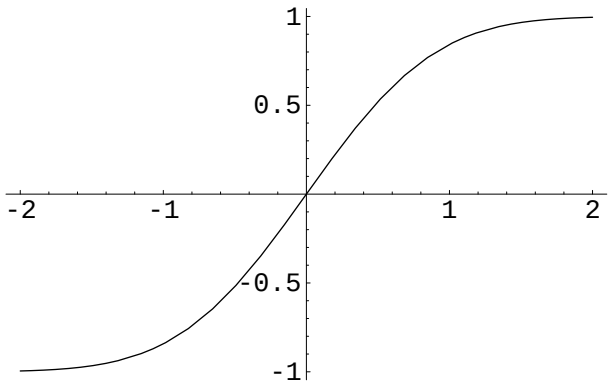


Figure 14.4. The Error Function.

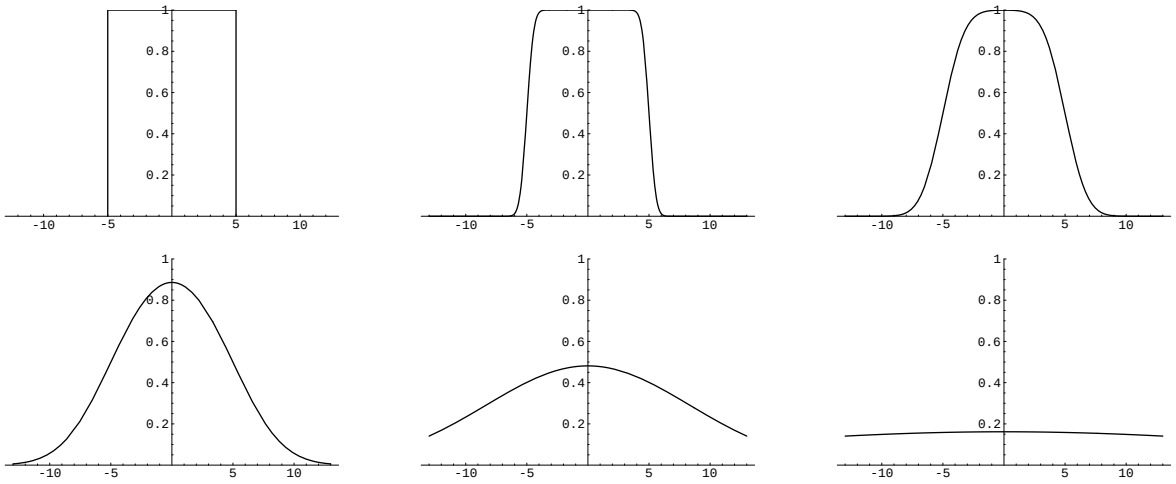


Figure 14.5. Error Function Solution to the Heat Equation.

where

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2}\, dz \tag{14.61}$$

is known as the *error function* due to its applications in probability and statistics, [**Feller**]. A graph appears in Figure 14.4. The error function integral cannot be written in terms of elementary functions. Nevertheless, its importance in various applications means that its properties have been well studied, and its values tabulated, [**48**]. In particular, it has asymptotic values

$$\lim_{x \to \infty} \operatorname{erf} x = 1, \qquad \lim_{x \to -\infty} \operatorname{erf} x = -1. \tag{14.62}$$

A graph of the solution (14.60) when $a = -5$, $b = 5$, at successive times $t = 0., .1, 1, 5, 30, 300$, is displayed in Figure 14.5. Note the initial smoothing or blurring of the sharp interface, followed by a gradual decay to thermal equilibrium.

The Inhomogeneous Heat Equation

The fundamental solution can be also used to solve the inhomogeneous heat equation

$$u_t = u_{xx} + h(t,x), \tag{14.63}$$

that models a bar under an external heat source $h(t,x)$, that might depend upon both position and time. We begin by solving the particular case

$$u_t = u_{xx} + \delta(t-s)\,\delta(x-y), \tag{14.64}$$

whose inhomogeneity represents a heat source of unit magnitude that is concentrated at a position $0 < y < \ell$ and applied instantaneously at a single time $t = s > 0$. Physically, we apply a soldering iron or laser beam to a single spot on the bar for a brief moment. Let us also impose homogeneous initial conditions

$$u(0,x) = 0 \tag{14.65}$$

as well as homogeneous boundary conditions of one of our standard types. The resulting solution

$$u(t,x) = G(t,x;s,y) \tag{14.66}$$

will be referred to as the *general fundamental solution* to the heat equation. Since a heat source which is applied at time s will only affect the solution at later times $t \geq s$, we expect that

$$G(t,x;s,y) = 0 \qquad \text{for all} \qquad t < s. \tag{14.67}$$

Indeed, since $u(t,x)$ solves the unforced heat equation at all times $t < s$ subject to homogeneous boundary conditions and has zero initial temperature, this follows immediately from the uniqueness of solutions to the initial-boundary value problem.

Once we know the general fundamental solution (14.66), we are able to solve the problem for a general external heat source (14.63) by appealing to linearity. We first write the forcing as a superposition

$$h(t,x) = \int_0^\infty \int_0^\ell h(s,y)\,\delta(t-s)\,\delta(x-y)\,dy\,ds \tag{14.68}$$

of concentrated instantaneous heat sources. Linearity allows us to conclude that the solution is given by the self-same superposition formula

$$u(t,x) = \int_0^t \int_0^\ell h(s,y)\,G(t,x;s,y)\,dy\,ds. \tag{14.69}$$

The fact that we only need to integrate over times $0 \leq s \leq t$ follows from (14.67).

Remark: If we have a nonzero initial condition, $u(0,x) = f(x)$, then we appeal to linear superposition to write the solution

$$u(t,x) = \int_0^\ell F(t,x;y)\,f(y)\,dy + \int_0^t \int_0^\ell h(s,y)\,G(t,x;s,y)\,dy\,ds \tag{14.70}$$

as a combination of (a) the solution with no external heat source, but inhomogeneous initial conditions, plus (b) the solution with homogeneous initial conditions but nonzero heat source.

Let us solve the forced heat equation in the case of a infinite bar, so $-\infty < x < \infty$. We begin by computing the general fundamental solution to (14.64), (14.65). As before, we take the Fourier transform of both sides of the partial differential equation with respect to x. In view of (13.100), (13.104), we find

$$\frac{\partial \widehat{u}}{\partial t} + 4\pi^2 k^2 \widehat{u} = e^{-2\pi i k y}\, \delta(t-s), \tag{14.71}$$

which is an inhomogeneous first order ordinary differential equation for the Fourier transform $\widehat{u}(t,k)$ of $u(t,x)$. Assuming $s > 0$, by (14.67), the initial condition is

$$\widehat{u}(0,k) = 0. \tag{14.72}$$

We solve the initial value problem by the usual method, [**24**]. Multiplying (14.71) by the integrating factor $e^{4\pi^2 k^2 t}$ yields

$$\frac{\partial}{\partial t}\left(e^{4\pi^2 k^2 t}\, \widehat{u}\right) = e^{4\pi^2 k^2 t - 2\pi i k y}\, \delta(t-s).$$

Integrating both sides from 0 to t and using the initial condition (14.72), we

$$\widehat{u}(t,k) = e^{4\pi^2 k^2 (s-t) - 2\pi i k y}\, \sigma(t-s),$$

where $\sigma(t)$ is the usual step function (11.42). Finally, we apply the inverse Fourier transform formula (13.74) and then (14.57), we deduce that

$$\begin{aligned} G(t,x;s,y) = u(t,x) &= \sigma(t-s) \int_{-\infty}^{\infty} e^{4\pi^2 k^2 (s-t) + 2\pi i k (x-y)}\, dk \\ &= \frac{\sigma(t-s)}{2\sqrt{\pi(t-s)}}\, \exp\left[-\frac{(x-y)^2}{4(t-s)} \right] = \sigma(t-s)\, F(t-s,x;y)\,. \end{aligned}$$

Thus, the general fundamental solution is obtained by translating the fundamental solution $F(t,x;y)$ for the initial value problem to a starting time of $t = s$ instead of $t = 0$. Thus, an initial condition has the same aftereffect on the temperature as an instantaneous applied heat source of the same magnitude. Finally, the superposition principle (14.69) produces the solution

$$u(t,x) = \int_0^t \int_{-\infty}^{\infty} \frac{h(s,y)}{2\sqrt{\pi(t-s)}}\, \exp\left[-\frac{(x-y)^2}{4(t-s)} \right] dy\, ds. \tag{14.73}$$

to the heat equation with source term on an infinite bar.

Example 14.4. ∎

The Root Cellar Problem

As a final example, we discuss a problem that involves analysis of the heat equation on a semi-infinite interval. The question is: how deep should you dig a root cellar? In the prerefrigeration era, a root cellar was used to keep food cool in the summer, but not freeze in the winter. We assume that the temperature in the earth only depends on the depth

and the time of yea. Let $u(t,x)$ denote the deviation in the temperature in the earth at depth $x > 0$ and time t from its annual mean. We shall assume that the temperature at the earth's surface, $x = 0$, fluctuates in a periodic manner; specifically, we set

$$u(t,0) = a\,\cos\omega\,t, \tag{14.74}$$

where the oscillatory frequency

$$\omega = \frac{2\,\pi}{365.25\text{ days}} = 2.0 \times 10^{-7}\text{sec}^{-1} \tag{14.75}$$

refers to yearly temperature variations. In this model, we shall ignore daily temperature fluctuations as their effect is not significant below a thin surface layer. At large depth the temperature is assumed to be unvarying:

$$u(t,x) \;\longrightarrow\; 0 \qquad \text{as} \qquad x \;\longrightarrow\; \infty, \tag{14.76}$$

where 0 refers to the mean temperature.

Thus, we must solve the heat equation on a semi-infinite bar $0 < x < \infty$, with time-dependent boundary conditions (14.74), (14.76) at the ends. The analysis will be simplified a little if we replace the cosine by a complex exponential, and so look for a complex solution with boundary conditions

$$u(t,0) = a\,e^{\,\mathrm{i}\,\omega\,t}, \qquad \lim_{x\,\to\,\infty} u(t,x) = 0. \tag{14.77}$$

Let us try a separable solution of the form

$$u(t,x) = v(x)\,e^{\,\mathrm{i}\,\omega\,t}. \tag{14.78}$$

Substituting this expression into the heat equation $u_t = \gamma\,u_{xx}$ leads to

$$\mathrm{i}\,\omega\,v(x)\,e^{\,\mathrm{i}\,\omega\,t} = \gamma\,v''(x)\,e^{\,\mathrm{i}\,\omega\,t}.$$

Canceling the common exponential factors, we conclude that $v(x)$ should solve the boundary value problem

$$\gamma\,v''(x) = \,\mathrm{i}\,\omega\,v, \qquad v(0) = a, \qquad \lim_{x\,\to\,\infty} v(x) = 0.$$

The solutions to the ordinary differential equation are

$$v_1(x) = e^{\sqrt{\,\mathrm{i}\,\omega/\gamma}\,x} = e^{\sqrt{\omega/2\,\gamma}\,(1+\mathrm{i}\,)\,x} \qquad \text{and} \qquad v_2(x) = e^{-\sqrt{\,\mathrm{i}\,\omega/\gamma}\,x} = e^{-\sqrt{\omega/2\,\gamma}\,(1+\mathrm{i}\,)\,x}.$$

The first solution is exponentially growing as $x \to \infty$, and so not appropriate to our problem. The solution to the boundary value problem must therefore be a multiple,

$$v(x) = a\,e^{-\sqrt{\omega/2\,\gamma}\,(1+\mathrm{i}\,)\,x}$$

of the exponentially decaying solution. Substituting back into (14.78), we find the (complex) solution to the root cellar problem to be

$$u(t,x) = a\,e^{-x\,\sqrt{\omega/2\,\gamma}}\,e^{\,\mathrm{i}\,\omega\left(\,t-\sqrt{\omega/2\,\gamma}\,\right)\,x}. \tag{14.79}$$

The corresponding real solution is obtained by taking the real part,

$$u(t,x) = a\, e^{-x\sqrt{\omega/2\gamma}}\, \cos\left(\omega\, t - \sqrt{\frac{\omega}{2\gamma}}\, x\right). \tag{14.80}$$

The first term in (14.80) is exponentially decaying as a function of the depth. Thus, the further down one goes, the less noticeable the effect of the surface temperature fluctuations. The second term is periodic with the same annual frequency ω. The interesting feature is the phase lag in the response. The temperature at a given depth x is out of phase with respect to the surface temperature fluctuations, with the phase lag

$$\delta = \sqrt{\frac{\omega}{2\gamma}}\, x$$

depending linearly on depth. In particular, a cellar built at a depth where δ is an odd multiple of π will be completely out of phase, being hottest in the winter, and coldest in the summer. Thus, the (shallowest) ideal depth at which to build a root cellar would take $\delta = \pi$, corresponding to a depth of

$$x = \pi \sqrt{\frac{2\gamma}{\omega}}\,.$$

For typical soils in the earth, $\gamma \approx 10^{-6}$ meters^2 sec^{-1}, [**X**], and hence, by (14.75), $x \approx 9.9$ meters. However, at this depth, the relative amplitude of the oscillations is

$$e^{-x\sqrt{\omega/2\gamma}} = e^{-\pi} = .04$$

and hence there is only a 4% temperature fluctuation. In Minnesota, the temperature varies, roughly, from $-40°$C to $+40°$C, and hence our 10 meter deep root cellar would experience only a 3.2°C annual temperature deviation from the winter, when it is the warmest, to the summer, where it is the coldest. Building the cellar twice as deep would lead to a temperature fluctuation of .2%, now in phase with the surface variations, which means that the cellar is, for all practical purposes, at constant temperature year round.

14.4. The Wave Equation.

The second important class of dynamical partial differential equations are those modeling vibrations of continuous media. As we saw in Chapter 9, Newton's Law implies that the free vibrations of a discrete mechanical system are governed by a second order system of ordinary differential equations of the form

$$M\, \frac{d^2\mathbf{u}}{dt^2} = -K\,\mathbf{u},$$

in which M is the positive definite, diagonal mass matrix, while $K = A^*A = A^TCA$ is the positive definite (or semi-definite in the case of an unstable system) stiffness matrix.

The corresponding dynamical equations describing the small vibrations of continuous media take an entirely analogous form

$$\rho\, \frac{\partial^2 u}{\partial t^2} = -K[u]. \tag{14.81}$$

In this framework, ρ describes the density of the medium, while $K = L^* \circ L$ is the same self-adjoint differential operator, with appropriate boundary conditions, that appears in the equilibrium equations. For one-dimensional media, such as a vibrating bar or string, we are lead to a partial differential equation in the particular form

$$\rho(x)\,\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(\kappa(x)\,\frac{\partial u}{\partial x}\right), \qquad 0 < x < \ell, \tag{14.82}$$

where $\rho(x)$ is the density of the bar or string at position x, while $\kappa(x) > 0$ denotes its stiffness or tension. The second order partial differential equation (14.82) models the dynamics of vibrations and waves in a broad range of continuous media, including elastic vibrations of a bar, sound vibrations in a column of air, e.g., inside a wind instrument, and also transverse vibrations of a string, e.g., a violin string. (However, bending vibrations of a beam lead to a fourth order partial differential equation; see Exercise ■.) It also be used to model small amplitude water waves, electromagnetic waves, including light, radio and microwaves, and many others. A detailed derivation of the model from first principles in the case of a vibrating string can be found in [**146**].

In addition, we must impose suitable boundary conditions. The usual suspects — Dirichlet, Neumann, mixed, and periodic boundary conditions — continue to play a central role, and have immediate physical interpretations. Tying down an end of string imposes a Dirichlet condition $u(t, 0) = \alpha$. A free end is prescribed by a homogeneous Neumann boundary condition $u_x(t, 0) = 0$. Periodic boundary conditions, as in (14.11), correspond to the vibrations of a circular ring. As with all second order Newtonian systems of ordinary differential equations, the solution to the full boundary value problem for the second order partial differential equation will be uniquely specified by its initial displacement and initial velocity:

$$u(0, x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0, x) = g(x). \tag{14.83}$$

The simplest situation occurs when the medium is homogeneous, and so both its density and stiffness are constant. Then the general vibration equation (14.82) reduces to the one-dimensional *wave equation*

$$\frac{\partial^2 u}{\partial t^2} = c^2\,\frac{\partial^2 u}{\partial x^2}\,. \tag{14.84}$$

The constant

$$c = \sqrt{\frac{\kappa}{\rho}} \;>\; 0 \tag{14.85}$$

is known as the *wave speed*, for reasons that will soon become apparent.

The method for solving such second order systems is motivated by our solution in the discrete case discussed in Section 9.5. To keep matters simple, we shall concentrate on the homogeneous wave equation (14.84), although the method is easily extended to the general system (14.82). Since we anticipate solutions that are time periodic, we will try a separable solution of the special form

$$u(t, x) = \cos(\omega t)\, v(x). \tag{14.86}$$

with trigonometric time dependence. Differentiating (14.86), we find

$$\frac{\partial^2 u}{\partial t^2} = -\omega^2 \cos(\omega t)\, v(x), \qquad \frac{\partial^2 u}{\partial x^2} = \cos(\omega t)\, v''(x).$$

Substituting these formulae into the wave equation (14.84) and canceling the common cosine factors, we deduce that $v(x)$ must satisfy the ordinary differential equation

$$c^2 \, \frac{d^2 v}{dx^2} + \omega^2 \, v = 0, \tag{14.87}$$

in which $\lambda = \omega^2$ represents an eigenvalue for the second order differential operator $K = -c^2 D^2$. Ignoring the boundary conditions for the moment, if $\omega > 0$, the solutions are the trigonometric functions $\cos \dfrac{\omega x}{c}$, $\sin \dfrac{\omega x}{c}$, and so we have constructed the explicit solutions

$$\cos \omega t \, \cos \frac{\omega x}{c}\,, \qquad \cos \omega t \, \sin \frac{\omega x}{c}\,,$$

to the wave equation. Now, in the original ansatz (14.86), the cosine could just as well be a sine, and the same computation will apply. Therefore, we deduce two additional solutions

$$\sin \omega t \, \cos \frac{\omega x}{c}\,, \qquad \sin \omega t \, \sin \frac{\omega x}{c}\,.$$

Each of these four solutions represents a spatially periodic standing wave form of period $2\pi c/\omega$, that is vibrating with frequency ω. Note particularly that the smaller scale waves vibrate faster.

On the other hand, if $\omega = 0$, then (14.87) has the solution $v = \alpha x + \beta$, leading to the solutions

$$u(t,x) = 1, \qquad \text{and} \qquad u(t,x) = x. \tag{14.88}$$

The first is a constant, nonvibrating solution, while the second is also constant in time, but will typically not satisfy the boundary conditions and so can be discarded. As we learned in Chapter 9, the existence of a zero eigenvalue corresponds to an unstable mode in the physical system, in which the displacement grows linearly in time. In the present situation, these correspond to the two additional solutions

$$u(t,x) = t, \qquad \text{and} \qquad u(t,x) = x\,t, \tag{14.89}$$

both of which satisfy the wave equation. Again, the second solution will typically not satisfy the homogeneous boundary conditions, and can usually be safely ignored. These null eigenfunction modes will only arise in unstable configurations.

The boundary conditions will serve to specify the particular eigenvalues and natural frequencies of vibration. Consider first the case of a string of length ℓ with two fixed ends, and thus subject to homogeneous Dirichlet boundary conditions

$$u(t,0) = 0 = u(t,\ell).$$

This constitutes a positive definite boundary value problem, and so there is no unstable mode. Indeed, the eigenfunctions of the boundary value problem (14.87) with Dirichlet boundary conditions $v(0) = 0 = v(\ell)$ were found in (14.17):

$$v_n(x) = \sin \frac{n\pi x}{\ell} \qquad \text{with} \qquad \omega_n = \frac{n\pi c}{\ell}, \qquad n = 1, 2, 3, \dots .$$

Therefore, we can write the general solution as a Fourier sine series

$$u(t,x) = \sum_{n=1}^{\infty} \left(b_n \cos \frac{n\pi c t}{\ell} \, \sin \frac{n\pi x}{\ell} + d_n \sin \frac{n\pi c t}{\ell} \, \sin \frac{n\pi x}{\ell} \right). \tag{14.90}$$

The solution is thus a linear combination of the natural Fourier modes vibrating with frequencies

$$\omega_n = \frac{n\pi c}{\ell} = \frac{n\pi}{\ell} \sqrt{\frac{\kappa}{\rho}}, \qquad n = 1, 2, 3, \dots . \tag{14.91}$$

Note that the longer the length ℓ of the string, or the higher its density ρ, the slower the vibrations, whereas increasing its stiffness or tension κ speeds them up — in exact accordance with physical intuition.

The Fourier coefficients b_n and d_n in (14.90) will be uniquely determined by the initial conditions (14.83). Differentiating the series term by term, we discover that we must represent the initial displacement and velocity as Fourier sine series

$$u(0,x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{\ell} = f(x), \qquad \frac{\partial u}{\partial t}(0,x) = \sum_{n=1}^{\infty} d_n \frac{n\pi c}{\ell} \sin \frac{n\pi x}{\ell} = g(x).$$

Therefore,

$$b_n = \frac{2}{\ell} \int_0^{\ell} f(x) \sin \frac{n\pi x}{\ell} \, dx, \qquad n = 1, 2, 3, \dots .$$

are the Fourier sine coefficients (12.72) of the initial displacement $f(x)$, while

$$d_n = \frac{2}{n\pi c} \int_0^{\ell} g(x) \sin \frac{n\pi x}{\ell} \, dx, \qquad n = 1, 2, 3, \dots .$$

are the Fourier sine coefficients of the initial velocity $g(x)$, rescaled by the wave speed.

Example 14.5. A string of unit length is held taut in the center and then released. Our goal is to describe the ensuing vibrations. Let us assume the physical units are chosen so that $c^2 = 1$, and so we are asked to solve the initial-boundary value problem

$$u_{tt} = u_{xx}, \qquad u(0,x) = f(x), \qquad u_t(0,x) = 0, \qquad u(t,0) = u(t,1) = 0. \tag{14.92}$$

To be specific, we assume that the center of the string has been displaced by half a unit, and so the initial displacement is

$$f(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2}, \\ 1 - x, & \frac{1}{2} \le x \le 1. \end{cases}$$

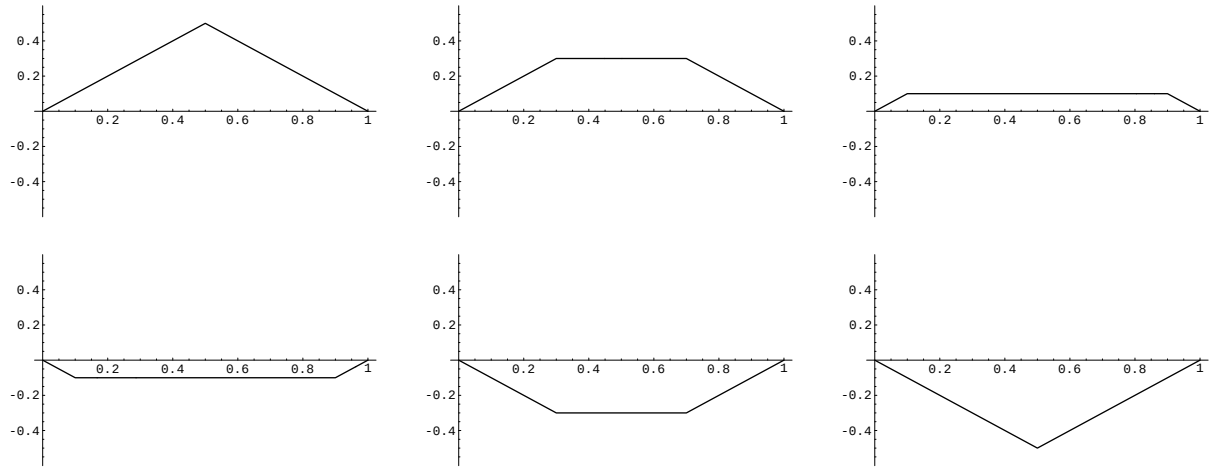


Figure 14.6. Plucked String Solution of the Wave Equation.

The vibrational frequencies are the integral multiples $\omega_n = n\pi$, and so the natural modes of vibration are

$$\cos n\pi t \ \sin n\pi x \qquad \text{and} \qquad \sin n\pi t \ \sin n\pi x \qquad \text{for} \qquad n = 1, 2, \dots .$$

Consequently, the general solution to the boundary value problem is

$$u(t,x) = \sum_{n=1}^{\infty} \big(b_n \cos n\pi t \ \sin n\pi x + d_n \sin n\pi t \ \sin n\pi x \big),$$

where

$$b_n = 2\int_0^1 f(x) \sin n\pi x \, dx = \begin{cases} 4\displaystyle\int_0^{1/2} x \sin n\pi x \, dx = \frac{4\,(-1)^k}{(2k+1)^2\pi^2}, & n = 2k+1, \\ 0, & n = 2k, \end{cases}$$

are the Fourier sine coefficients of the initial displacement, while $d_n = 0$ are the Fourier sine coefficients of the initial velocity. Therefore, the solution takes the form of a single Fourier sine series

$$u(t,x) = 4\sum_{k=0}^{\infty} (-1)^k \, \frac{\cos(2k+1)\pi t \ \sin(2k+1)\pi x}{(2k+1)^2\pi^2}, \tag{14.93}$$

whose graph is depicted in Figure 14.6 at times $t = 0, .2, .4, .6, .8, 1.$. At this point in time, the original displacement is reproduced exactly, but upside down. The subsequent dynamics proceeds as before, but in mirror image form. The original displacement reappears at time $t = 2$, after which time the motion is periodically repeated. Interestingly, at times $t_k = .5, 1.5, 2.5, \dots$, the displacement is identically zero: $u(t_k, x) \equiv 0$, although the velocity $u_t(t_k, x)$ is nonzero. When summed to a sufficinetly high order, the solution appears to be piecewise affine, i.e., its graph is a collection of straight lines. This fact, which is in stark contrast to the smoothing effect of the heat equation, will be verified in Exercise ■, where you are asked to construct an exact analytical formula for this solution.

While the series form (14.90) of the solution is not entirely satisfying, we can still use it to deduce important qualitative properties. First of all, since each term is periodic in t with period $2\,\ell/c$, the entire solution is time periodic with that period: $u(t + 2\,\ell/c, x) = u(t, x)$. In fact, after half the period, at time $t = \ell/c$, the solution reduces to

$$u\left(\frac{\ell}{c}\,, x\right) = \sum_{n=1}^{\infty} (-1)^n\, b_n\, \sin\frac{n\,\pi\,x}{\ell} = -\sum_{n=1}^{\infty} b_n\, \sin\frac{n\,\pi\,(\ell - x)}{\ell} = -\,u(0, \ell - x) = -\,f(\ell - x).$$

In general,

$$u\left(t + \frac{\ell}{c}, x\right) = -\,u(t, \ell - x), \qquad\qquad u\left(t + \frac{2\,\ell}{c}, x\right) = u(t, x). \tag{14.94}$$

Therefore, the initial wave form is reproduced, first as an upside down mirror image of itself at time $t = \ell/c$, and then identical to its original form at time $t = 2\,\ell/c$. This has the important consequence that vibrations of (homogeneous) one-dimensional media are purely periodic phenomena! There is no quasi-periodicity because the fundamental frequencies are all integer multiples of each other.

Remark: The preceding analysis has important musical consequences. To the human ear, sonic vibrations that are integral multiples of a single frequency are harmonic, whereas those that admit quasi-periodic vibrations, with irrationally related frequencies, sound percussive. This is why most tonal instruments rely on vibrations in one dimension, be it a violin string, a column of air in a wind instrument (flute, clarinet, trumpet or saxophone), a xylophone bar or a triangle. On the other hand, most percussion instruments rely on the vibrations of two-dimensional media, e.g., drums and cymbals, or three-dimensional solid bodies, e.g., blocks, which, as we shall see in Chapters 17 and 18, admit frequencies with irrational ratios.

A bar with both ends left free, and so subject to the Neumann boundary conditions

$$\frac{\partial u}{\partial x}\,(t, 0) = 0 = \frac{\partial u}{\partial x}\,(t, \ell), \tag{14.95}$$

will have a slightly different behavior, owing to the instability of the underlying equilibrium equations. The eigenfunctions of (14.87) with Neumann boundary conditions $v'(0) = 0 = v'(\ell)$ are now

$$v_n(x) = \cos\frac{n\,\pi\,x}{\ell} \qquad \text{with} \qquad \omega_n = \frac{n\,\pi\,c}{\ell}, \qquad n = 0, 1, 2, 3, \ldots\,.$$

The resulting solution takes the form of a Fourier cosine series

$$u(t, x) = a_0 + c_0\,t + \sum_{n=1}^{\infty} \left(a_n\, \cos\frac{n\,\pi\,c\,t}{\ell}\, \cos\frac{n\,\pi\,x}{\ell} + c_n\, \sin\frac{n\,\pi\,c\,t}{\ell}\, \cos\frac{n\,\pi\,x}{\ell} \right). \tag{14.96}$$

In accordance with (14.88), the first two terms come from the null eigenfunction $v_0(x) = 1$ with $\omega_0 = 0$. The bar vibrates with the same fundamental frequencies (14.91) as in the fixed end case, but there is now an additional unstable mode $c_0\,t$ that is no longer periodic, but grows linearly in time.

Substituting (14.96) into the initial conditions (14.83), we find the Fourier coefficients are prescribed, as before, by the initial displacement and velocity,

$$a_n = \frac{2}{\ell}\int_0^\ell f(x)\,\cos\frac{n\,\pi\,x}{\ell}\,dx, \qquad c_n = \frac{2}{n\,\pi\,c}\int_0^\ell g(x)\,\cos\frac{n\,\pi\,x}{\ell}\,dx, \qquad n = 1, 2, 3, \dots .$$

The order zero coefficients[†],

$$a_0 = \frac{1}{\ell}\int_0^\ell f(x)\,dx, \qquad\qquad c_0 = \frac{1}{\ell}\int_0^\ell g(x)\,dx,$$

are equal to the average initial displacement and average initial velocity of the bar. In particular, when $c_0 = 0$ there is no net initial velocity, and the unstable mode is not excited. In this case, the solution is time-periodic, oscillating around the position given by the average initial displacement. On the other hand, if $c_0 \neq 0$, then the unstable mode will be excited. Since there is nothing to restrain its motion, the bar will move off with constant average speed c_0, all the while vibrating at the same fundamental frequencies.

Similar considerations apply to the periodic boundary value problem for the wave equation on a circular ring. The details are left as Exercise ■ for the reader.

Forcing and Resonance

In Section 9.6, we learned that periodically forcing an undamped mechanical structure (or a resistanceless electrical circuit) at a frequency that is distinct from its natural vibrational frequencies leads, in general, to a quasi-periodic response. The solution is a sum of the unforced vibrations superimposed with an additional vibrational mode at the forcing frequency. However, if forced at (or very near) one of the natural frequencies, the system may go into a catastrophic resonance.

The exact same quasiperiodic and resonant responses are also observed in the corresponding continuum partial differential equations governing periodic vibrations. To keep the analysis as simple as possible, we restrict our attention to the forced wave equation for a homogeneous bar

$$\frac{\partial^2 u}{\partial t^2} = c^2\,\frac{\partial^2 u}{\partial x^2} + F(t, x). \tag{14.97}$$

The external forcing function $F(t, x)$ may depend upon both time t and position x. We will be particularly interested in a periodically varying external force of the form

$$F(t, x) = \cos(\omega\,t)\,h(x), \tag{14.98}$$

where the function $h(x)$ is fixed.

As always — see Theorem 7.37 — the solution to an inhomogeneous linear system can be written as a combination,

$$u(t, x) = u_\star(t, x) + z(t, x) \tag{14.99}$$

[†] Note that, unlike the usual Fourier series, we have not included the $\frac{1}{2}$ factor in the constant terms in (14.96).

of a particular solution $u_\star(t,x)$ along with the general solution $z(t,x)$ to the homogeneous equation, namely

$$\frac{\partial^2 z}{\partial t^2} = c^2 \, \frac{\partial^2 z}{\partial x^2} \, . \tag{14.100}$$

The boundary and initial conditions will serve to uniquely prescribe the solution $u(t,x)$, but there is some flexibility in its two constituents $u_\star, z$. For instance, we may ask that the particular solution $u_\star$ satisfy the homogeneous boundary conditions along with zero (homogeneous) initial conditions, and thus represents the pure response of the system to the forcing. The homogeneous solution $z(t,x)$ will then reflect the effect of the initial and boundary conditions unadulterated by the external forcing. The final solution is then a sum of the two individual responses.

In the case of periodic forcing (14.98), we look for a particular solution

$$u_\star(t,x) = \cos(\omega\, t)\, v_\star(x) \tag{14.101}$$

that vibrates at the forcing frequency. Substituting the ansatz (14.101) into the equation (14.97), and canceling the common cosine factors, we discover that $v_\star(x)$ must satisfy the boundary value problem prescribed by

$$-c^2\, v_\star'' - \omega^2\, v_\star = h(x), \tag{14.102}$$

supplemented by therelevant homogeneous boundary conditions — Dirichlet, Neumann, mixed, or periodic.

At this juncture, there are two possibilities. If the unforced, homogeneous boundary value problem has only the trivial solution $v \equiv 0$, then there is a solution to the forced boundary value problem for any form of the forcing function $h(x)$. On the other hand, the homogeneous boundary value problem has a nontrivial solution $v(x)$ if and only if $\omega^2 = \lambda$ is an eigenvalue, and so ω is a natural frequency of vibration to the homogeneous problem; the solution $v(x)$ is the corresponding eigenfunction appearing in the solution series (14.90). In this case, the boundary value problem (14.102) has a solution if and only if the forcing function $h(x)$ is orthogonal to the eigenfunction(s):

$$\langle\, h\,;v\,\rangle = 0. \tag{14.103}$$

This result is a manifestation of the Fredholm alternative, Theorem 5.51, and the self-adjointness of the boundary value problem; see Example 11.3 and Exercise ■ for details. If we force in a resonant manner — meaning that (14.103) is not satisfied — then the solution will have the form of a resonantly growing vibration

$$u_\star(t,x) = t\, \sin\omega\, t\, v_\star(x)$$

that will, if unchecked, eventually lead to a catastrophic breakdown of the system, e.g., the bar breaks or the string snaps.

Remark: In a real-world situation, the large resonant or near resonant vibrations will either cause a catastrophic breakdown, e.g., the string snaps, or send the system into a different, nonlinear regime that helps mitigate the resonant effects, but is no longer modeled

by the simple linear wave equation. There are, indeed, familiar physical systems where resonance is desirable! In a microwave oven, the microwaves are tuned to the resonant frequencies of water molecules, and thus excite them into large vibrations, thereby heating up your dinner. Blowing into a clarinet or other wind instrument excites the resonant frequencies in the column of air contained within it, and this produces the musical sound vibrations that we hear.

Example 14.6. As a specific example, consider the forced vibrations of a uniform bar that is fixed at both ends:

$$\begin{aligned} u_{tt} &= c^2\, u_{xx} + \cos(\omega\, t)\, h(x), \\ u(t,0) = 0 = u(t,1), \qquad u(0,x) &= f(x), \qquad u_x(0,x) = g(x). \end{aligned} \tag{14.104}$$

(We take the length $\ell = 1$ to simplify the formulas.) The particular solution will have the nonresonant form (14.101) provided we can find a solution $v_\star(x)$ to the boundary value problem

$$c^2\, v_\star'' + \omega^2\, v_\star = -\,h(x), \qquad\qquad v_\star(0) = 0 = v_\star(1). \tag{14.105}$$

The resonant frequencies and corresponding eigenfunctions in this particular case are

$$\omega_n = n\, c\, \pi, \qquad\qquad v_n(x) = \sin n\,\pi\, x, \qquad\qquad n = 1, 2, 3, \dots\ .$$

The boundary value problem (14.105) will have a solution, and hence the forcing is not resonant, provided either $\omega \neq \omega_n$ is not an eigenvalue, or, $\omega = \omega_n$ is an eigenvalue, but

$$0 = \langle\, h\,;v_n\,\rangle = \int_0^1 h(x) \sin n\,\pi\, x\, dx \tag{14.106}$$

is orthogonal to the associated eigenfunction. The remaining (generic) case, where the forcing profile is not orthogonal to the eigenfunction, induces a resonance whose amplitude grows linearly in time.

For example, under periodic forcing of frequency ω with trigonometric profile $h(x) \equiv \sin k\,\pi\, x$, the particular solution to (14.105) is

$$v_\star(x) = \frac{\sin k\,\pi\, x}{\omega^2 - k^2\,\pi^2\, c^2}, \qquad \text{so that} \qquad u_\star(t,x) = \frac{\cos \omega\, t\ \sin k\,\pi\, x}{\omega^2 - k^2\,\pi^2\, c^2}, \tag{14.107}$$

which is a valid solution as long as $\omega \neq \omega_k = k\,\pi\, c$. Note that we may allow the forcing frequency $\omega = \omega_n$ to coincide with any other resonant forcing frequency, $n \neq k$, because the sine profiles are mutually orthogonal and so the nonresonance condition (14.106) is satisfied. On the other hand, if $\omega = \omega_k = k\,\pi\, c$, then the particular solution

$$u_\star(t,x) = \frac{t\, \sin k\,\pi\, c t\ \sin k\,\pi\, x}{2\, k\,\pi\, c}\,, \tag{14.108}$$

is resonant, and grows linearly in time, in precise analogy with the ordinary differential equation case discussed in Section 9.6.

To obtain the actual solution to the initial-boundary value problem, we write $u = u_\star + z$ where $z(t,x)$ must satisfy

$$z_{tt} - c^2 z_{xx} = 0, \qquad z(t,0) = 0 = z(t,1),$$

along with the modified initial conditions

$$z(0,x) = f(x) \, - \, \frac{\sin k\pi x}{\omega^2 - k^2\pi^2 c^2}\,, \qquad \frac{\partial u}{\partial x}\,(0,x) = g(x),$$

stemming from the fact that the particular solution (14.107) has non-trivial initial data. (In the resonant case (14.108), there is no extra term in the initial data.) As before, the solution $z(t,x)$ to the homogeneous equation can be written as a Fourier sine series (14.90). The final formulae are left to the reader to complete.

14.5. d'Alembert's Solution.

In the case of the one-dimensional wave equation, there is an alternative explicit solution formula due to the eighteenth century French mathematician Jean d'Alembert. His solution avoids the complicated Fourier series formulae, and thereby provides additional insight into the behavior of the solutions. Unfortunately, unlike the series method, that has very broad applicability, d'Alembert's approach only succeeds in this one very special situation: the homogeneous wave equation in a single space variable.

The method begins by writing the wave equation (14.84) in the suggestive form

$$\Box u = (\partial_t^2 - c^2\,\partial_x^2)\,u = u_{tt} - c^2\,u_{xx} = 0. \tag{14.109}$$

Here $\Box = \partial_t^2 - c^2\,\partial_x^2$ is a common mathematical notation for the linear *wave differential operator*, while ∂_t, ∂_x are convenient shorthands for the partial derivative operators with respect to t and x. In analogy with the elementary polynomial factorization

$$t^2 - c^2\,x^2 = (t - c\,x)(t + c\,x),$$

we can factor the second order wave operator into a product of two first order partial differential operators:

$$\Box = \partial_t^2 - c^2\,\partial_x^2 = (\partial_t - c\,\partial_x)\,(\partial_t + c\,\partial_x). \tag{14.110}$$

If the second factor annihilates the function $u(t,x)$, meaning

$$(\partial_t + c\,\partial_x)\,u = u_t + c\,u_x = 0, \tag{14.111}$$

then u is automatically a solution to the wave equation:

$$\Box u = (\partial_t - c\,\partial_x)\,(\partial_t + c\,\partial_x)\,u = (\partial_t - c\,\partial_x)\,0 = 0.$$

In other words, every solution to the simpler first order partial differential equation (14.111) is a solution to the wave equation (14.84). (The converse is, of course, not true.)

It is relatively easy to solve linear† first order partial differential equations.

† See Chapter 22 for more details, including the extension of the method of characteristics to first order nonlinear partial differential equations.

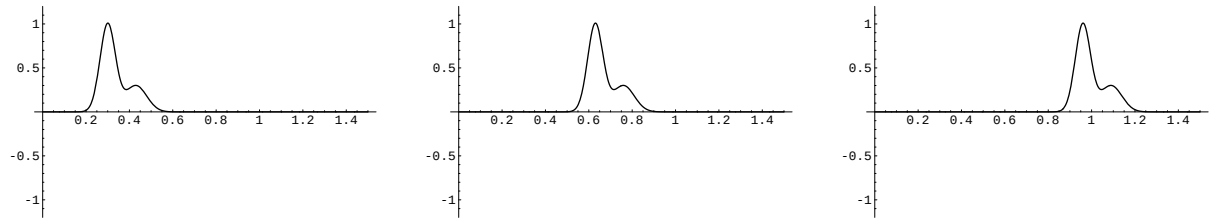


Figure 14.7. Traveling Wave.

Proposition 14.7. *Every solution $u(t,x)$ to the partial differential equation*

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = 0 \tag{14.112}$$

has the form

$$u(t,x) = p(x - c\,t), \tag{14.113}$$

where $p(\xi)$ is an arbitrary function of the single characteristic variable $\xi = x - c\,t$.

Proof: We adopt a linear change of variables to rewrite the solution

$$u(t,x) = p(t, x - c\,t) = p(t,\xi)$$

in terms of the characteristic variable ξ and the time t. Applying the chain rule, we express the derivatives of u in terms of the derivatives of p as follows:

$$\frac{\partial u}{\partial t} = \frac{\partial p}{\partial t} - c\,\frac{\partial p}{\partial \xi}\,, \qquad\qquad \frac{\partial u}{\partial x} = \frac{\partial p}{\partial \xi}\,,$$

and hence

$$\frac{\partial u}{\partial t} + c\,\frac{\partial u}{\partial x} = \frac{\partial p}{\partial t} - c\,\frac{\partial p}{\partial \xi} + c\,\frac{\partial p}{\partial \xi} = \frac{\partial p}{\partial t}\,.$$

Therefore, u is a solution to (14.112) if and only if $p(t,\xi)$ is a solution to the very simple partial differential equation

$$\frac{\partial p}{\partial t} = 0.$$

This clearly[†] implies that $p = p(\xi)$ does not depend on the variable t, and hence

$$u = p(\xi) = p(x - c\,t)$$

is of the desired form. *Q.E.D.*

Therefore, *any* function of the characteristic variable, e.g., $\xi^2 + 1$ or $\cos\xi$ or e^{ξ}, will produce a corresponding solution, $(x - c\,t)^2 + 1$ or $\cos(x - c\,t)$ or $e^{x - c\,t}$, to the first order partial differential equation (14.112), and hence a solution to the wave equation (14.84). The functions of the form (14.113) are known as *traveling waves*. At $t = 0$ the wave has the initial profile $u(0,x) = p(x)$. As t progresses, the wave moves to the *right* with constant speed $c > 0$, unchanged in form; see Figure 14.7. For this reason, (14.112) is

[†] More rigorously, one must also assume that, at each time t, the domain of definition of $p(\xi)$ is a connected interval. A similar technical restriction should be imposed upon the solutions in the statement of Proposition 14.7.

sometimes referred to as the *one-way* or *unidirectional wave equation*. We conclude that every traveling wave solution to the unidirectional wave equation (14.112) is a solution to the wave equation (14.84).

Now, since c is constant, the factorization (14.110) can be written equally well in the reverse order:

$$\Box = \partial_t^2 - c^2\,\partial_x^2 = (\partial_t + c\,\partial_x)\,(\partial_t - c\,\partial_x). \tag{14.114}$$

The same argument tells us that any solution to the alternative first order partial differential equation

$$\frac{\partial u}{\partial t} - c\,\frac{\partial u}{\partial x} = 0, \tag{14.115}$$

also provides a solution to the wave equation. This is also a one-way wave equation, having the opposite wave speed $-c$. Applying Proposition 14.7, now with c replaced by $-c$, we conclude that the general solution to (14.115) has the form

$$u(t,x) = q(x + c\,t) \tag{14.116}$$

where $q(\eta)$ is an arbitrary differentiable function of the second characteristic variable $\eta = x + c\,t$. The solutions (14.116) represent traveling waves moving to the *left* with constant speed $c > 0$ and unchanged in form.

Thus, we have uncovered two different classes of solutions to the full wave equation (14.84). One class consists of traveling waves moving to the right with speed c, while the other class consists of traveling waves moving to the left with the same speed c. Thus, the wave equation is *bidirectional* and has both left and right traveling wave solutions. (Of course, such solutions do not necessarily respect the boundary conditions, which, when present, will affect their ultimate behavior.)

Linearity of the wave equation implies that the sum of solutions is again a solution. In this way, we can produce solutions which are superpositions of left and right traveling waves. The remarkable fact, due to d'Alembert, is that *every* solution to the wave equation can be so represented.

Theorem 14.8. *The general solution to the wave equation* (14.84) *is a combination*

$$u(t,x) = p(\xi) + q(\eta) = p(x - c\,t) + q(x + c\,t) \tag{14.117}$$

of right and left traveling waves, depending on their respective characteristic variables

$$\xi = x - c\,t, \qquad\qquad \eta = x + c\,t. \tag{14.118}$$

Proof: The key is to use a linear changes of variables to rewrite the wave equation entirely in terms of the characteristic variables ξ, η defined by (14.118). We set

$$u(t,x) = w(x - c\,t, x + c\,t) = w(\xi,\eta), \qquad \text{or} \qquad w(\xi,\eta) = u\left(\frac{\xi+\eta}{2}\,,\,\frac{\eta-\xi}{2\,c}\right).$$

Then, invoking to the chain rule to compute partial derivatives,

$$\frac{\partial u}{\partial t} = c\left(\frac{\partial w}{\partial \xi} - \frac{\partial w}{\partial \eta}\right), \qquad\qquad \frac{\partial u}{\partial x} = \frac{\partial w}{\partial \xi} + \frac{\partial w}{\partial \eta}\,.$$

and hence

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 w}{\partial \xi^2} - 2\,\frac{\partial^2 w}{\partial \xi\,\partial \eta} + \frac{\partial^2 w}{\partial \eta^2} \right), \qquad \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 w}{\partial \xi^2} + 2\,\frac{\partial^2 w}{\partial \xi\,\partial \eta} + \frac{\partial^2 w}{\partial \eta^2}\,.$$

Therefore

$$\Box\, u = \frac{\partial^2 u}{\partial t^2} - c^2\,\frac{\partial^2 u}{\partial x^2} = -4\,c^2\,\frac{\partial^2 w}{\partial \xi\,\partial \eta}\,.$$

We conclude that $u(t,x)$ solves the wave equation $\Box\, u = 0$ if and only if $w(\xi,\eta)$ solves the second order partial differential equation

$$\frac{\partial^2 w}{\partial \xi\,\partial \eta} = 0, \qquad \text{which we write in the form} \qquad \frac{\partial}{\partial \xi}\left(\frac{\partial w}{\partial \eta} \right) = 0.$$

As before, this partial differential equation can be integrated once with respect to ξ, resulting in

$$\frac{\partial w}{\partial \eta} = r(\eta),$$

where r is an arbitrary function of the characteristic vvaraible η. Integrating both sides of the latter partial differential equation with respect to η, we find

$$w(\xi,\eta) = p(\xi) + q(\eta), \qquad \text{where} \qquad q'(\eta) = r(\eta),$$

while $p(\xi)$ represents the integration "constant". Replacing the characteristic variables by their formulae in terms of t and x completes the proof. *Q.E.D.*

Remark: As above, we have been a little cavalier with our specification of the domain of definition of the functions and the differentiability assumptions required. Sorting out the precise technical details is left to the dedicated reader.

Remark: The general solution to a second order *ordinary* differential equation depends on two arbitrary constants. Here we observe that the general solution to a second order *partial* differential equation depends on two arbitrary functions — in this case $p(\xi)$ and $q(\eta)$.

Let us now see how this new form of solution to the wave equation can be used to effectively solve initial value problems. The simplest case is that of a bar or string of infinite length, in which case we have a pure initial value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2\,\frac{\partial^2 u}{\partial x^2}\,, \qquad u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x), \qquad \text{for} \qquad -\infty < x < \infty. \tag{14.119}$$

Substituting the solution formula (14.117) into the initial conditions, we find

$$u(0,x) = p(x) + q(x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = -\,c\,p'(x) + c\,q'(x) = g(x).$$

To solve this pair of linear equations for p and q, we differentiate the first equation:

$$p'(x) + q'(x) = f'(x).$$

Subtracting the second equation divided by c, we find

$$2\,p'(x) = f'(x) - \frac{1}{c}\,g(x).$$

Therefore,

$$p(x) = \frac{1}{2}\,f(x) - \frac{1}{2\,c}\int_0^x g(z)\,dz + a,$$

where a is an integration constant. The first equation then yields

$$q(x) = f(x) - p(x) = \frac{1}{2}\,f(x) + \frac{1}{2\,c}\int_0^x g(z)\,dz - a.$$

Substituting these two expressions back into (14.117), we find

$$\begin{aligned} u(t,x) = p(\xi) + q(\eta) &= \frac{f(\xi)+f(\eta)}{2} + \frac{1}{2\,c}\left[-\int_0^{\xi} + \int_0^{\eta}\right] g(z)\,dz \\ &= \frac{f(\xi)+f(\eta)}{2} + \frac{1}{2\,c}\int_{\xi}^{\eta} g(z)\,dz, \end{aligned}$$

where ξ, η are the characteristic variables (14.118). In this fashion, we have derived d'Alembert's solution to the wave equation on the entire line $-\infty < x < \infty$.

Theorem 14.9. *The solution to the initial value problem*

$$\frac{\partial^2 u}{\partial t^2} = c^2\,\frac{\partial^2 u}{\partial x^2}\,, \qquad u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x), \qquad -\infty < x < \infty. \tag{14.120}$$

is given by

$$u(t,x) = \frac{f(x-c\,t) + f(x+c\,t)}{2} + \frac{1}{2\,c}\int_{x-c\,t}^{x+c\,t} g(z)\,dz. \tag{14.121}$$

Let us investigate the implications of d'Alembert's formula (14.121). First, suppose there is no initial velocity, so $g(x) \equiv 0$, and the motion is purely the result of the initial displacement $u(0,x) = f(x)$. In this case, the solution (14.121) reduces to

$$u(t,x) = \tfrac{1}{2}\,f(x-c\,t) + \tfrac{1}{2}\,f(x+c\,t).$$

The basic effect is that the initial displacement $f(x)$ splits into two waves, one traveling to the right and one traveling to the left, each with exactly the same shape as the initial displacement $f(x)$, but only half as tall. For example, if the initial displacement is a localized pulse centered at the origin, say

$$u(0,x) = e^{-x^2}, \qquad \frac{\partial u}{\partial t}\,(0,x) = 0,$$

then the solution

$$u(t,x) = \tfrac{1}{2}\,e^{-(x-c\,t)^2} + \tfrac{1}{2}\,e^{-(x+c\,t)^2}$$

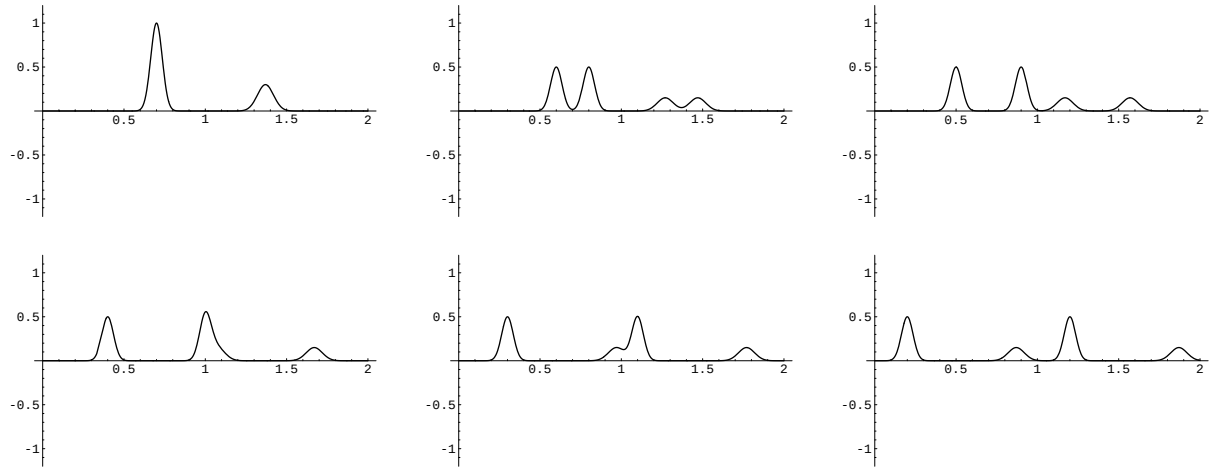


Figure 14.8. Interaction of Waves.

consists of two half size copies of the initial pulse running away from the origin in opposite directions with equal speed c. If we take two separated pulses, say

$$u(0,x) = e^{-x^2} + 2\,e^{-(x-1)^2}, \qquad \frac{\partial u}{\partial x}\,(0,x) = 0,$$

centered at $x = 0$ and $x = 1$, then the solution

$$u(t,x) = \tfrac{1}{2}\,e^{-(x-c\,t)^2} + e^{-(x-1-c\,t)^2} + \tfrac{1}{2}\,e^{-(x+c\,t)^2} + e^{-(x-1+c\,t)^2}$$

will consist of four pulses, two moving to the right and two to the left, all with the same speed, as pictured in Figure 14.8.

Remark: If the initial displacement has compact support, and so $f(x) = 0$ if $x < a$ or $x > b$ for some $a < b$, then after a short time the right and left-moving waves will completely disengage and the observer will see two exact half size replicas running away, with speed c, in opposite directions. If the displacement is not localized, then the left and right traveling waves will never fully disengage, and one might be hard pressed (just as in our earlier discussion of quasi-periodic phenomena) in recognizing that a complicated solution pattern is, in reality, just the superposition of two very simple traveling waves.

An important observation is that when a right-moving pulse collides with a left-moving pulse, they emerge from the collision unchanged — a consequence of the linearity of the wave equation. The first picture shows the initial displacement. In the second and third pictures, the two localized bumps have each split into two copies moving in opposite directions. In the fourth and fifth, the larger right moving bump is in the process of interacting with the smaller left moving bump. Finally, in the last picture the interaction is complete, and the two left moving bumps and two right moving bumps travel in tandem with no further collisions.

The lines in the (t,x)–plane where the characteristic variables are constant,

$$\xi = x - c\,t = a, \qquad \eta = x + c\,t = b, \tag{14.122}$$

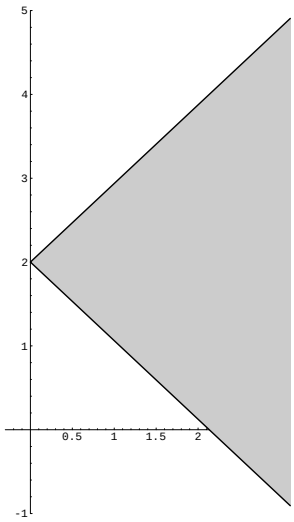


Figure 14.9. Characteristic Lines for the Wave Equation.

have slope $\pm c$, and are known as the *characteristics* of the wave equation. The two characteristics emanating from a point on the x axis, where the initial data is prescribed, are illustrated in Figure 14.9. The reader should note that, in this figure, the t axis is horizontal, while the x axis is vertical.

In general, signals propagate along characteristics. More specifically, if we start out with an initial displacement concentrated very close to a point $x = a$, then the solution will be concentrated along the two characteristic lines through the point $x = a$, $t = 0$, namely $x - c\,t = a$ and $x + c\,t = a$. In the limit, a unit impulse or delta function displacement at $x = a$, corresponding to the initial condition

$$u(0, x) = \delta(x - a), \qquad \frac{\partial u}{\partial t}\,(0, x) = 0, \tag{14.123}$$

will result in a solution

$$u(t, x) = \tfrac{1}{2}\,\delta(x - c\,t - a) + \tfrac{1}{2}\,\delta(x + c\,t - a) \tag{14.124}$$

consisting of two half-strength delta spikes traveling away from the starting position along the two characteristic lines.

Let us return to the general initial value problem (14.120). Suppose now that there is no initial displacement, $u(0, x) = f(x) \equiv 0$, but rather a concentrated initial velocity, say a delta function

$$\frac{\partial u}{\partial t}\,(0, x) = \delta_a(x) = \delta(x - a).$$

Physically, this would correspond to striking the string with a concentrated blow at the point $x = a$. The d'Alembert solution (14.121) is

$$u(t, x) = \frac{1}{2\,c}\int_{x-c\,t}^{x+c\,t} \delta_a(z)\,dz = \begin{cases} \dfrac{1}{2\,c}\,, & x - c\,t < a < x + c\,t, \\ 0, & \text{otherwise}, \end{cases} \tag{14.125}$$

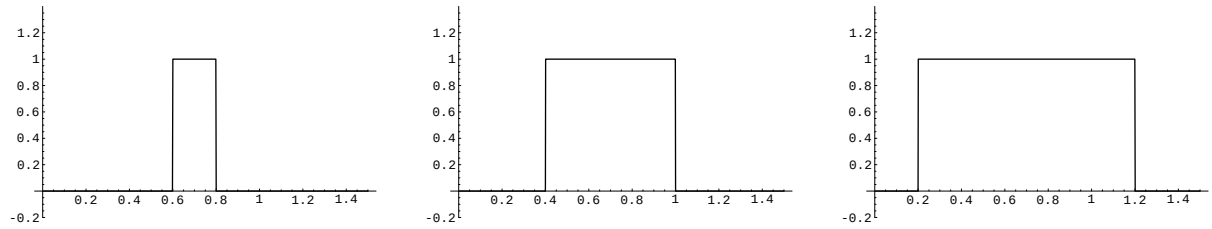


Figure 14.10. Concentrated Initial Velocity for Wave Equation.

and consists of a constant displacement, of magnitude $1/(2\,c)$, between the two characteristic lines $x - c\,t = a = x + c\,t$ emanating from the point $x = a$, $t = 0$ — the shaded region of Figure 14.9. The solution, which is plotted in Figure 14.10, has two jump discontinuities between the undisturbed state and the displaced state, each propagating along its characteristic line with speed c, but in opposite directions. Note that, unlike a concentrated initial displacement, where the signal remains concentrated and each point along the bar is temporarily displaced, eventually returning to its undisturbed state, a concentrated initial velocity has a lasting effect, and the bar remains permanently deformed by an amount $1/(2\,c)$.

Solutions on Bounded Intervals

So far, we have been looking at the solutions to the initial value problem for the wave equation on an infinite interval. The d'Alembert formula can still be used on bounded intervals, but in a suitably modified format so as to respect the boundary conditions. The easiest to deal with is the periodic case. If the boundary conditions are periodic on $0 \le x \le \ell$, then the solution $u(t, x)$ must also be periodic as a functiuon of x. one extends to the initial displacement and velocity, $f(x)$ and $g(x)$, to also be periodic of period ℓ. If the initial velocity has mean zero, then the resulting d'Alembert solution (14.121) will remain periodic. Otherwise, the solution will not be periodic in time, owing to the excitation of the unstable mode. Exercise ∎ contains the necessary details.

If we have fixed (Dirichlet) boundary conditions, say

$$u(t, 0) = 0, \qquad u(t, \ell) = 0, \tag{14.126}$$

then, motivated by the fact that the solution can be written as a Fourier sine series (14.90), one takes the initial displacement $f(x)$ and velocity $g(x)$ and extends them to be odd, periodic functions of period $2\,\ell$:

$$f(-x) = -f(x), \qquad f(x + 2\,\ell) = f(x), \qquad g(-x) = -g(x), \qquad g(x + 2\,\ell) = g(x).$$

This will ensure that the d'Alembert solution also remains odd, periodic, and hence the boundary conditions (14.126) remain valid for all t. Keep in mind that, while the solution $u(t, x)$ is defined for all x, the only physically relevant values occur on the interval $0 \le x \le \ell$. Nevertheless, the effects of displacements in the unphysical regime will eventually be "felt" as the propagating waves pass through the physical interval.

For example, consider an initial displacement which is concentrated near $x = a$ for some $0 < a < \ell$. Its odd, periodic extension consists of two sets of replicas: those of the

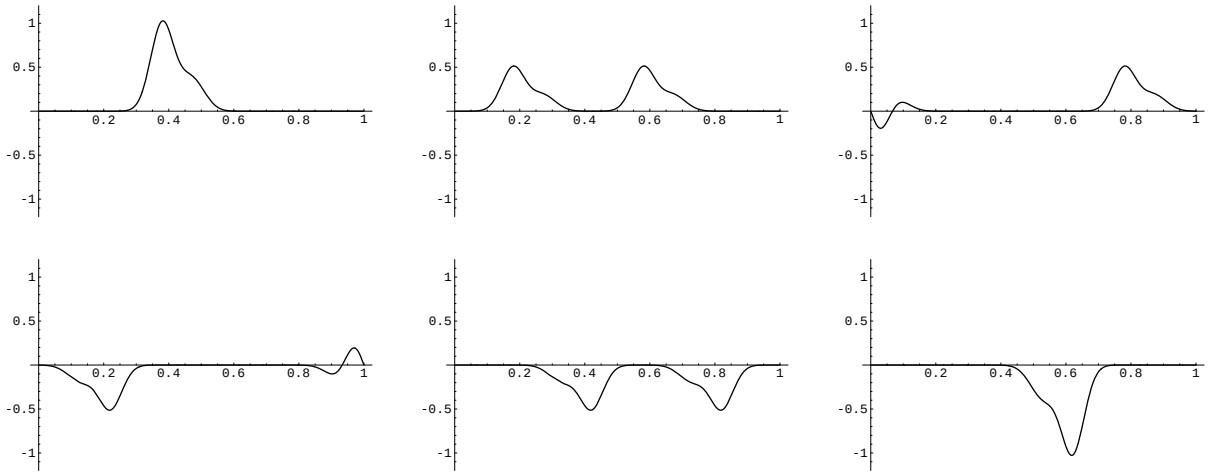


Figure 14.11. Solution to Wave Equation with Fixed Ends.

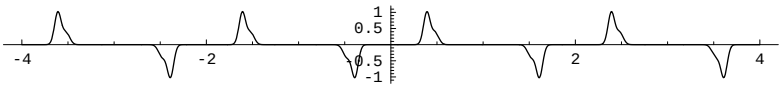


Figure 14.12. Odd Periodic Extension of a Concentrated Pulse.

same form occurring at positions $a \pm 2\,\ell, a \pm 4\,\ell, \ \dots$, and mirror image versions, resulting from the oddness of the function, at intermediate positions $-a, -a \pm 2\,\ell, -a \pm 4\,\ell, \ \dots$; see Figure 14.12. The resulting solution begins by each of the pulses, positive and negative, splitting into two half-size replicas that propagate with speed c in opposite directions. As the individual pulses meet, they interact as they pass through each other, eventually emerging unchanged. The process repeats periodically, with infinite rows of pulses moving to the right continually interacting with infinite rows moving to the left.

However, only the part of this solution that lies on $0 \leq x \leq \ell$ is actually realized on the physical bar. The net effect is as if we were viewing the solution passing by a stationary window, of length ℓ, that blocks out all other regions of the real axis. What the viewer effectively sees assumes a somewhat different interpretation. Namely, the original pulse at position $0 < a < \ell$ splits up into two half size replicas that move off in opposite directions. As each half-size pulse reaches an end of the bar, it meets a mirror image pulse that has been propagating in the opposite direction from the non-physical regime. The effect is that the pulse appears to be reflected at the end of the interval, and changes into an upside down mirror image of itself moving in the opposite direction. The original positive pulse has moved off the end of the bar just as its mirror image has moved into the physical regime. A good physical illustration is a wave propagating down a jump rope that is held fixed at its end; the reflected wave is upside down. A similar reflection occurs as the other half-size pulse hits the other end of the physical interval, after which the solution consists of two upside down half-size pulses moving back towards each other. At time $t = \ell/c$ they recombine at the point $\ell - a$ to instantaneously form a full-sized, but upside-down mirror image of the original disturbance, in accordance with (14.94). This in turn splits apart into two upside down bumps that, when they collide with the ends,

reflect and become right side up. At time $t = 2\,\ell/c$ they recombine to exactly reproduce the original displacement. The process then repeats, and the solution is periodic in time with period $2\,\ell/c$. In Figure 14.11, the first picture gives the initial displacement, which splits into left and right moving, half-size clones. In the third picture, the left moving bump is in the process of emerging from its collision with the end. In the fourth picture, it has emerged from its collision with the end, and is now upside down, reflected, and moving to the right. Meanwhile, the right moving pulse is starting to collide with the right hand end. In the fifth picture, both pulses have completed their collisions and are now moving back towards each other, where, in the last picture, they recombine into an upside-down version of the original pulse. The process then repeats itself in mirror image, finally recombining to the original pulse after the same length of time.

The Neumann (free) boundary value problem

$$\frac{\partial u}{\partial x}\,(t,0) = 0, \qquad \frac{\partial u}{\partial x}\,(t,\ell) = 0, \tag{14.127}$$

is handled similarly. Here, inspired by the Fourier cosine series form of the solution, one extends the initial conditions to be *even*, $2\,\ell$ periodic functions

$$f(-x) = f(x), \qquad f(x + 2\,\ell) = f(x), \qquad g(-x) = g(x), \qquad g(x + 2\,\ell) = g(x).$$

If the initial velocity has mean zero,

$$c_0 = \frac{1}{\ell}\int_0^{\ell} g(x)\,dx = 0, \tag{14.128}$$

then the solution remains periodic of period $2\,\ell/c$. In this case, when a bump hits one of the ends, the reflected bumps remains upright, but a mirror image of the original traveling in the opposite direction. A familiar physical illustration is a water wave that reflects off a solid wall. After an elapsed time of $t = \ell/c$, the individual reflected bumps recombine to form a positive mirror image of the original displacement, i.e., $u(t + \ell/c, x) = u(t, \ell - x)$. After a further time lapse of $t = 2\,\ell/c$, the original displacement reappears, and the solution is time periodic with period $2\,\ell/c$, i.e., $u(t + 2\,\ell/c, x) = u(t, x)$. On the other hand, if there is a net initial velocity, so $c_0 \neq 0$, then, as noted above, the solution is no longer periodic, but is a linear combination of periodic vibrations with the steadily increasing unstable mode $c_0\,t$.

In summary, we have now learned two different versions of the solution to the one-dimensional wave equation. The first, based on Fourier analysis, emphasizes the vibrational or wave character of the solutions. The second, based on the d'Alembert formula, emphasizes the "particle" aspects of the solutions, where individual wave packets collide with each other, or reflect at the boundary, but maintain their overall form. Some solutions look like vibrating waves, while others are much more like interacting particles. The Fourier series shows how every particle-like solution can be decomposed into its constituent vibrational modes, while the d'Alembert formula shows how vibrating waves can be viewed as moving particles.

The coexistence of particle and wave features is reminiscent of the long running historical debate over the nature of light, with Newton and his disciples advocating its particle

basis in the form of photons, while until the beginning of the twentieth century most physicists advocated the wave and vibrational viewpoint. Einstein's explanation of the photoelectric effect served to resurrect the particle interpretation of light. Only with the establishment of quantum mechanics was the debate resolved — light, and, indeed, all subatomic particles are both, manifesting both particle and wave features depending upon the experiment and the physical situation. But the evidence for a wave-particle duality already existed in the classical wave equation!

14.6. Numerical Methods.

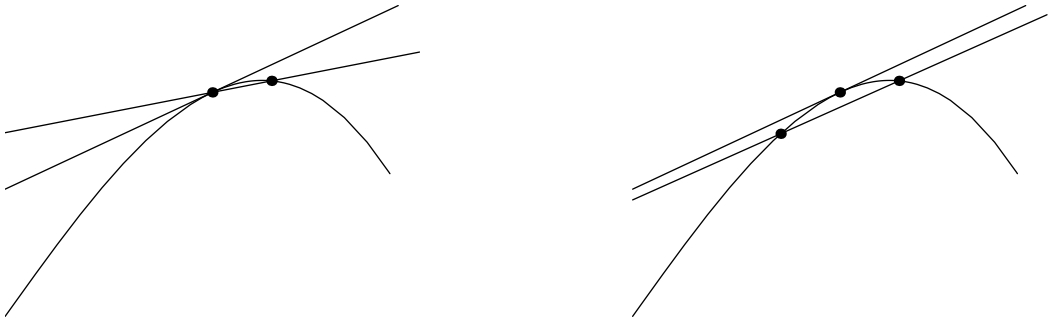
As we know, most differential equations are too complicated to solved analytically, and so one is usually forced to resort to numerical solution methods. Even in cases, like the heat and wave equations, where explicit solution formulas (either closed form or infinite series) exist, the numerical methods still can be profitably applied to solve particular initial-boundary value problems. Moreover, verification that the numerical algorithm produces a reasonable approximation to the true solution is much easier if one has an alternative solution formula in hand. The lessons learned in the design of numerical algorithms for solved problems prove to be of immense value when one is confronted with more complicated problems for which solution formulas no longer exist.

In this final section we present some of the most basic numerical solution techniques for the heat and wave equations. We just consider the simplest cases, leaving variations and extensions to a more thorough treatment as found in basic numerical analysis texts, [**30**].

Numerical solution methods for differential equations can be partitioned into two principal classes. (In this oversimplified presentation, we are leaving out more specialized methods of less general applicability.) The first category, already introduced in Section 11.6, are the *finite element methods*. Finite elements are designed for the differential equations describing equilibrium configurations, since they rely on minimizing a functional. The alternative approach is to directly approximate the derivatives appearing in the differential equation, through use of numerical differentiation formulae. In general, to approximate the derivative of a function, one constructs a suitable combination of sampled function values at nearby points. The underlying formalism used to construct these approximation formulae is know as the *calculus of finite differences*, and has a long history, dating back to Newton, that includes many prominent mathematicians in its development and application. The resulting *finite difference methods* have extremely broad applicability, and can, with proper care, be designed to solve most differential equations arising in mathematics, physics, engineering, biology, finance, and elsewhere.

Finite Differences

In this section, we give a brief introduction to the most basic finite difference approximations for derivatives of a function of one variable. In this presentation, we concentrate on the simplest version of the *calculus of finite differences*, based on equally spaced sample points. In the exercises, the reader is asked to generalize the difference formulae to non-equally spaced points.



One-Sided Difference Central Difference

Figure 14.13. Finite Difference Approximations.

The simplest finite difference approximation is the ordinary difference quotient

$$\frac{u(x+h)-u(x)}{h} \;\approx\; u'(x), \tag{14.129}$$

used to approximate the first derivative of the function $u(x)$. Throughout our discussion, the *step size* h, which may be either positive or negative, is assumed to be small: $|\,h\,| \ll 1$. The difference quotient can be interpreted as a linear combination of the sampled function values at the two nearby points x and $x+h$. Geometrically, the difference quotient equals the slope of the secant line through the two points $(x, u(x))$ and $(x+h, u(x+h))$ on the graph of the function. For small h, this should be a reasonably good approximation to the slope of the tangent line, as illustrated in the first picture in Figure 14.13. Indeed, if u is differentiable at x, then $u'(x)$ is, by definition, the limit, as $h \to 0$ of the finite difference quotients.

How close an approximation is (14.129)? To answer this question, we use the first order Taylor expansion

$$u(x+h) = u(x) + u'(x)\,h + \tfrac{1}{2}\,u''(\xi)\,h^2, \tag{14.130}$$

where we assume that $u(x)$ is at least twice continuously differentiable. Here ξ represents a point lying between x and $x+h$, which follows from the Cauchy form of the remainder term (C.2) in the Taylor expansion; see Appendix C for full details. Therefore, the difference quotient is given by the formula

$$\frac{u(x+h)-u(x)}{h} \;=\; u'(x) + \tfrac{1}{2}\,u''(\xi)\,h.$$

The *error* is the difference between the finite difference formula and the derivative being approximated, namely

$$\frac{u(x+h)-u(x)}{h} \;-\; u'(x) = \tfrac{1}{2}\,u''(\xi)\,h. \tag{14.131}$$

We say that the finite difference approximation (14.131) is *first order* because the error is proportional to h. Indeed, the error can be bounded by $\frac{1}{2}\,M\,h$, where $|\,u''\,| < M$ is an

overall bound on the second derivative of the function near the point x. If the precise formula for the error is not so important, we will write

$$u'(x) = \frac{u(x+h) - u(x)}{h} + \mathrm{O}(h). \tag{14.132}$$

The "big Oh" notation $\mathrm{O}(h)$ refers to a term proportional to h, or, more correctly, a term that is bounded by a constant multiple of h as $h \to 0$.

Example 14.10. Let $u(x) = \sin x$. Let us compute $u'(1) = \cos 1 = 0.5403023\ \ldots$ by using the finite difference quotient (14.129), and so

$$\cos 1 \ \approx \ \frac{\sin(1+h) - \sin 1}{h}\,.$$

The result for different values of h is listed in the following table.

h	1	.1	.01	.001	.0001
approximation	0.067826	0.497364	0.536086	0.539881	0.540260
error	−0.472476	−0.042939	−0.004216	−0.000421	−0.000042

We observe that reducing the step size by a factor of $\frac{1}{10}$ reduces the size of the error by approximately the same factor. Thus, to obtain 10 decimal digits of accuracy, we anticipate needing a step size of about $h = 10^{-11}$. The fact that the error is more of less proportional to the step size tells us that we are using a first order numerical approximation.

To approximate higher order derivatives, we need to evaluate the function at more than two points. In general, an approximation to the n^{th} order derivative $u^{(n)}(x)$ requires at least $n+1$ distinct sample points. For example, let us try to approximate $u''(x)$ by using the particular sample points x, $x+h$ and $x-h$. Which combination of the function values $u(x-h), u(x), u(x+h)$ can be used to approximate the derivative $u''(x)$? The answer to such a question can be found by consideration of the relevant Taylor expansions

$$\begin{aligned} u(x+h) &= u(x) + u'(x)\,h + u''(x)\,\frac{h^2}{2} + u'''(x)\,\frac{h^3}{6} + \mathrm{O}(h^4), \\ u(x-h) &= u(x) - u'(x)\,h + u''(x)\,\frac{h^2}{2} - u'''(x)\,\frac{h^3}{6} + \mathrm{O}(h^4), \end{aligned} \tag{14.133}$$

where the error terms are proportional to h^4. Adding the two formulae together gives

$$u(x+h) + u(x-h) = 2\,u(x) + u''(x)\,h^2 + \mathrm{O}(h^4).$$

Rearranging terms, we conclude that

$$u''(x) = \frac{u(x+h) - 2\,u(x) + u(x-h)}{h^2} + \mathrm{O}(h^2), \tag{14.134}$$

The result is is the simplest finite difference approximation to the second derivative of a function. The error is of order h^2, and depends upon the magnitude of the fourth order derivative of u near x; see Exercise ∎.

Example 14.11. Let $u(x) = e^{x^2}$, with $u''(x) = (4x^2 + 2)\, e^{x^2}$. Let us approximate $u''(1) = 6e = 16.30969097\ \ldots$ by using the finite difference quotient (14.134):

$$6e \;\approx\; \frac{e^{(1+h)^2} - 2e + e^{(1-h)^2}}{h^2}\,.$$

The results are listed in the following table.

h	1	.1	.01	.001	.0001
approximation	50.16158638	16.48289823	16.31141265	16.30970819	16.30969115
error	33.85189541	0.17320726	0.00172168	0.00001722	0.00000018

Each reduction in step size by a factor of $\frac{1}{10}$ reduces the size of the error by a factor of $\frac{1}{100}$ and a gain of two new decimal digits of accuracy, which is a reflection of the fact that the finite difference formula (14.134) is of second order, with error proportional to h^2. However, this prediction is not entirely borne out in practice. If we take $h = .00001$ then the formula produces the approximation 16.3097002570, with an error of 0.0000092863 — which is *less* accurate that the approximation with $h = .0001$. The problem is that round-off errors have now begun to affect the computation, and underscores a significant difficulty with numerical differentiation formulae. Such finite difference formulae involve dividing very small quantities, and this can lead to high numerical errors due to round-off. As a result, while they typically produce reasonably good approximations to the derivatives for moderately small step sizes, to achieve high accuracy, one must employ high precision arithmetic. A similar comment applied to the previous Example 14.10, and our expectations about the error for a very small step size were not, in fact justified as the reader may have discovered.

We can improve the order of accuracy of finite difference approximations to derivatives by employing more sample points to form an appropriate linear combination of the function values. For instance, if the first order approximation (14.132) to the first derivative based on the two points x and $x + h$ is not sufficiently accurate, one can try combining the function values at three points x, $x + h$ and $x - h$. To find the appropriate combination of $u(x - h), u(x), u(x + h)$, we return to the Taylor expansions (14.133). To solve for $u'(x)$, we subtract[†] the two formulae, and so

$$u(x + h) - u(x - h) = 2\, u'(x)\, h + u'''(x)\, \frac{h^3}{3} + \mathrm{O}(h^4).$$

Rearranging the terms, we are led to the well-known *centered difference formula*

$$u'(x) = \frac{u(x + h) - u(x - h)}{2h} + \mathrm{O}(h^2), \tag{14.135}$$

which is a second order approximation to the first derivative. Geometrically, the centered difference quotient represents the slope of the secant line through the two points $(x - h, u(x - h))$ and $(x + h, u(x + h))$ on the graph of u centered symmetrically about

[†] The terms $\mathrm{O}(h^4)$ do *not* cancel, since they represent potentially different multiples of h^4.

the point x. Figure 14.13 illustrates the geometry behind the two approximations; the advantages in accuracy in the centered difference version are graphically evident. Higher order approximations can be found by evaluating u at additional points, including, say, $x + 2\,h, x - 2\,h$, and so on.

Example 14.12. Return to the function $u(x) = \sin x$ considered in Example 14.10. The centered difference approximation to its derivative $u'(1) = \cos 1 = 0.5403023\ \ldots$ is

$$\cos 1 \ \approx \ \frac{\sin(1+h) - \sin(1-h)}{2\,h}\,.$$

The results are tabulated as follows:

h	.1	.01	.001	.0001
approximation	0.53940225217	0.54029330087	0.54030221582	0.54030230497
error	−0.00090005370	−0.00000900499	−0.00000009005	−0.00000000090

As advertised, the results are much more accurate than the one-sided finite difference approximation used in Example 14.10 at the same step size. As in Example 14.11, we see that each reduction in the step size by a factor of $\frac{1}{10}$ adds two more decimal places of accuracy, which is a consequence of the second order accuracy in the centered difference approximation.

Many more finite difference formulae can be constructed by similar manipulations of Taylor expansions, but these will suffice for our purposes. Let us now apply these basic formulas to construct numerical solution algorithms for the heat and wave equations.

Numerical Solution Methods for the Heat Equation

Consider the heat equation

$$\frac{\partial u}{\partial t} = \gamma\, \frac{\partial^2 u}{\partial x^2}\,, \qquad\qquad 0 < x < \ell, \qquad t \geq 0, \tag{14.136}$$

on a bar of length ℓ, where $\gamma > 0$ represents the thermal diffusivity, which is assumed to be constant. To be specific, we impose Dirichlet boundary conditions

$$u(t,0) = \alpha(t), \qquad\qquad u(t,\ell) = \beta(t), \qquad\qquad t \geq 0. \tag{14.137}$$

at the ends of the bar, along with the initial conditions

$$u(0,x) = f(x), \qquad\qquad 0 \leq x \leq \ell. \tag{14.138}$$

In order to effect a numerical approximation to the solution to this initial-boundary value problem, we begin by introducing a *rectangular mesh* consisting of points (t_i, x_j) with $0 = x_0 < x_1 < \cdots < x_n = \ell$ and $0 = t_0 < t_1 < t_2 < \ \cdots$. For simplicity, we maintain a fixed, regular mesh spacing, with

$$h = x_{j+1} - x_j = \frac{\ell}{n}\,, \qquad\qquad k = t_{i+1} - t_i,$$

representing, respectively, the spatial mesh size and the time step size. It is important that the two step sizes are *not* necessarily the same. Note that

$$t_i = i\,k, \qquad\qquad x_j = j\,h.$$

We shall use the notation

$$u_{i,j} \;\approx\; u(t_i, x_j) \tag{14.139}$$

to denote our numerical approximation to the value of the solution at a given mesh point.

As a first try at designing a numerical method, we shall use the simplest finite difference approximations to the derivatives. The second order space derivative is approximated by (14.134), and hence

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2}\,(t_i, x_j) \;&\approx\; \frac{u(t_i, x_{j+1}) - 2\,u(t_i, x_j) + u(t_i, x_{j-1})}{h^2} + \mathrm{O}(h^2) \\ &\approx\; \frac{u_{i,j+1} - 2\,u_{i,j} + u_{i,j-1}}{h^2} + \mathrm{O}(h^2), \end{aligned} \tag{14.140}$$

where the error in the approximation is proportional to h^2. Similarly, the one-sided finite difference approximation (14.132) is used for the time derivative, and so

$$\frac{\partial u}{\partial t}\,(t_i, x_j) \;\approx\; \frac{u(t_{i+1}, x_j) - u(t_i, x_j)}{k} + \mathrm{O}(k) \;\approx\; \frac{u_{i+1,j} - u_{i,j}}{k} + \mathrm{O}(k), \tag{14.141}$$

where the error is proportion to k. In practice, it is important to ensure that the approximations have similar orders of accuracy, which tells us to choose

$$k \approx h^2.$$

Assuming the step size $h < 1$, this requirement has the important consequence that the time steps must be *much* smaller than the space mesh size.

Remark: At this stage, the reader might be tempted to replace (14.141) by the second order central difference approximation (14.135). However, this produces significant complications in the implementation of the method, and is not suitable for a practical numerical algorithm for the heat equation. We shall subsequently see how to construct a practical numerical method that *is* second order in the time step k.

Substituting equations (14.140), (14.141) into the partial differential equation (14.142), and rearranging terms, we find

$$u_{i+1,j} = u_{i,j} + \mu\,\big(\,u_{i,j+1} - 2\,u_{i,j} + u_{i,j-1}\,\big), \qquad \begin{aligned} &i = 0, 1, 2, \dots\,, \\ &j = 1, \dots, n-1, \end{aligned} \tag{14.142}$$

where

$$\mu = \frac{\gamma\,k}{h^2}\,. \tag{14.143}$$

The numerical scheme takes the form of an iterative linear system for the solution values $u_{i,j} \approx u(t_i, x_j)$ at each time step t_i.

The initial condition (14.138) means that we should initialize our numerical data by sampling the initial temperature at the mesh points:

$$u_{0,j} = f_j = f(x_j), \qquad j = 1, \dots, n-1. \tag{14.144}$$

Similarly, the boundary conditions (14.137) require that

$$u_{i,0} = \alpha_i = \alpha(t_i), \qquad u_{i,n} = \beta_i = \beta(t_i), \qquad i = 0, 1, 2, \dots . \tag{14.145}$$

In addition, we assume consistency of the initial and boundary conditions at the corners of the domain:

$$f_0 = f(0) = u(0,0) = \alpha(0) = \alpha_0, \qquad f_n = f(\ell) = u(0,\ell) = \beta(0) = \beta_0.$$

The three equations (14.142), (14.144), (14.145) completely prescribe the numerical approximation scheme for solving the initial-boundary value problem (14.136), (14.137), (14.138) for the heat equation.

Let us rewrite this discrete dynamical system in a more transparent matrix form. First, let

$$\mathbf{u}^{(i)} = \big(u_{i,1}, u_{i,2}, \dots, u_{i,n-1} \big)^T \approx \big(u(t_i, x_1), u(t_i, x_2), \dots, u(t_i, x_{n-1}) \big)^T \tag{14.146}$$

be the vector whose entries are the numerical approximations to the solution values at the *interior* nodes — omitting the boundary nodes $x_0 = 0, x_n = \ell$, where the values of u are fixed by the boundary conditions (14.137). Then (14.142) takes the matrix form

$$\mathbf{u}^{(i+1)} = A\,\mathbf{u}^{(i)} + \mathbf{b}^{(i)}, \tag{14.147}$$

where

$$A = \begin{pmatrix} 1-2\mu & \mu & & & & & \\ \mu & 1-2\mu & \mu & & & & \\ & \mu & 1-2\mu & \mu & & & \\ & & \mu & \ddots & \ddots & & \\ & & & \ddots & \ddots & \mu & \\ & & & & \mu & 1-2\mu \end{pmatrix}, \qquad \mathbf{b}^{(i)} = \begin{pmatrix} \mu\,\alpha_i \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mu\,\beta_i \end{pmatrix}. \tag{14.148}$$

The coefficient matrix A is symmetric and tridiagonal. The contributions (14.145) of the boundary nodes are found in the vector $\mathbf{b}^{(i)}$. This numerical method is known as an "explicit scheme" since each iterate is computed explicitly without relying on solving an auxiliary equation — unlike the "implicit schemes" discussed below. The method is not guaranteed to work, and indeed does not unless the mesh sizes are chosen appropriately.

Example 14.13. Let us fix the diffusivity $\gamma = 1$ and the bar length $\ell = 1$. For illustrative purposes, we fix the spatial step size to be $h = .1$. In Figure 14.14 we compare two (slightly different time step sizes on the same initial data as used in (14.22). The first sequence takes $k = h^2 = .01$ and plots the solution at times $t = 0., .02, .04$. The solution is already starting to show signs of instability, and indeed soon becomes completely wild.

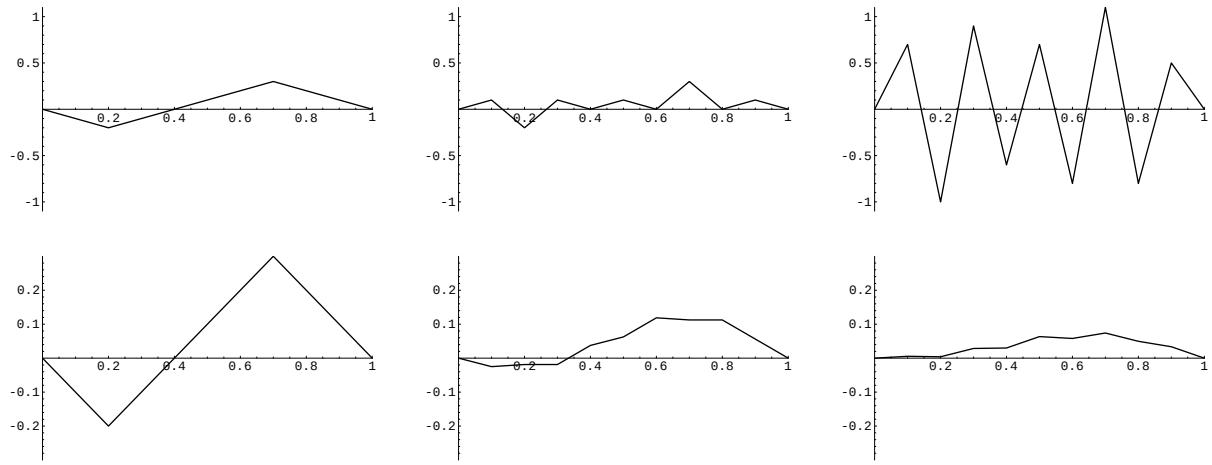


Figure 14.14. Numerical Solutions for the Heat Equation Based on the Explicit Scheme.

The second sequence takes $k = .005$ and plots the solution at times $t = 0., .025, .05$. (Note that we are using different vertical scales for the two sequences of plots.) Even though we are employing a rather coarse mesh, the numerical solution is not too far away from the true solution to the initial value problem, which can be found in Figure 14.1.

In order to understand the underlying issues, let us concentrate on homogeneous boundary conditions

$$u(t, 0) = 0 = u(t, \ell)$$

whereby $\alpha_i = \beta_i = 0$ for all i and so (14.147) reduces to a homogeneous, linear iterative system

$$\mathbf{u}^{(i+1)} = A\,\mathbf{u}^{(i)}. \tag{14.149}$$

The solution will converge to zero, $\mathbf{u}^{(i)} \to \mathbf{0}$, as it is supposed to (why?), if and only if A is a convergent matrix. But convergence will depend on the step sizes. For instance, if $\gamma = 1$, choosing a spatial step size of $h = .1$ and a time step size of $k = h^2 = .01$ gives a non-convergent matrix, and an invalid numerical scheme, while a much smaller step size, e.g., $k = .0005$, gives a convergent matrix and a valid numerical scheme.

As we learned in Chapter 10, the convergence property of a matrix is fixed by its spectral radius, i.e., the largest eigenvalue in magnitude; see Theorem 10.13. There is, in fact, an explicit formula for the eigenvalues of the particular tridiagonal matrix (14.148).

Lemma 14.14. *The eigenvalues of the* $(n-1) \times (n-1)$ *matrix* A *in* (14.148) *are*

$$\lambda_k = 1 - 4\,\mu\,\sin^2\frac{\pi\,k}{2\,n}, \qquad\qquad k = 1, \dots, n-1.$$

A proof of this fact, including an explicit formula for the associated eigenfunctions, is outlined in Exercise ∎. The matrix is convergent if and only if all its eigenvalues are less than 1 in absolute value. Here, convergence requires

$$\left| 1 - 4\,\mu\,\sin^2\frac{\pi\,k}{2\,n} \right| < 1, \qquad \text{for all} \qquad k = 1, \dots, n-1.$$

Since $0 \leq \sin^2 x \leq 1$, the convergence inequality will be valid as long as

$$|\,\mu\,| < \tfrac{1}{2}.$$

In this way, we have deduced the basic stability criterion for the linear iterative system (14.149). Referring to the formula (14.143), we find the condition

$$\frac{\gamma\, k}{h^2} < \frac{1}{2}\,, \qquad \text{or} \qquad k < \frac{h^2}{2\gamma}\,, \tag{14.150}$$

required for the coefficient matrix to be convergent.

As a result, this numerical method is called *conditionally stable*, which means that not all choices of space and time steps lead to a convergence scheme. The convergence criterion (14.150) places a rather severe restriction on the time step size. For instance, if we have $h = .01$, and $\gamma = 1$, then we can only use a time step size of $k < .00005$, which is minuscule. It would take a huge number of time steps to compute the value of the solution at even a moderate times, e.g., $t = 1$. Moreover, owing to the limited accuracy of computers, the propagation of round-off errors might then become a significant issue in reducing the overall accuracy of the final solution values.

An unconditionally stable method — one that does not restrict the time step — can be constructed by using the backwards difference formula

$$\frac{\partial u}{\partial t}\,(t_i, x_j) \approx \frac{u(t_i, x_j) - u(t_{i-1}, x_j)}{k} + \mathrm{O}(h^k) \tag{14.151}$$

to the temporal derivatives instead. Substituting (14.151) and the same approximation (14.140) for u_{xx} into the heat equation, and then replacing i by $i+1$, leads to the iterative system

$$u_{i+1,j} - \mu\,\big(\,u_{i+1,j+1} - 2\,u_{i,j} + u_{i+1,j-1}\,\big) = u_{i,j}, \qquad \begin{array}{l} i = 0, 1, 2, \dots\,, \\ j = 1, \dots, n-1, \end{array} \tag{14.152}$$

where the parameter $\mu = \gamma\, k/h^2$ is as above. The initial and boundary conditions also have the same form (14.144), (14.145). The system has the matrix form

$$\widehat{A}\,\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \mathbf{b}^{(i+1)}, \tag{14.153}$$

where $\widehat{A}$ is obtained from the matrix A in (14.148) by replacing μ by $-\mu$. This defines an *implicit method* since we have to solve a tridiagonal linear system at each step in order to compute the next iterate $\mathbf{u}^{(i+1)}$. However, as we learned in Section 1.7, a tridiagonal linear system can be solved quite rapidly, and so this does not become a significant issue in the practical implementation.

Let us look at the convergence of the implicit scheme. For homogeneous Dirichlet boundary conditions, the system takes the form

$$\mathbf{u}^{(i+1)} = \widehat{A}^{-1}\,\mathbf{u}^{(i)},$$

and the convergence is now governed by the eigenvalues of $\widehat{A}^{-1}$. Lemma 14.14 tells us that the eigenvalues of $\widehat{A}$ are

$$\lambda_k = 1 + 4\,\mu \sin^2 \frac{\pi\, k}{2\,n}\,, \qquad\qquad k = 1, \dots, n-1.$$

As a result, its inverse $\widehat{A}^{-1}$ has eigenvalues

$$\frac{1}{\lambda_k} = \frac{1}{1 + 4\,\mu \sin^2 \dfrac{\pi\, k}{2\,n}}\,, \qquad\qquad k = 1, \dots, n-1.$$

Since $\mu > 0$, the latter are *always* less than 1 in absolute value, and so $\widehat{A}$ is a convergent matrix for *any* $\mu > 0$. Therefore, the implicit scheme (14.153) is convergent for any choice of step sizes h, k.

Compute previous example ∎

An even better numerical scheme is obtained by averaging the explicit and implicit schemes (14.142), (14.152). The result is known as the *Crank–Nicholson scheme*, and takes the form

$$u_{i+1,j} - u_{i,j} = \frac{\mu}{2}\,\big(\,u_{i+1,j+1} - 2\,u_{i+1,j} + u_{i+1,j-1} + u_{i,j+1} - 2\,u_{i,j} + u_{i,j-1}\,\big). \tag{14.154}$$

We can write the system in matrix form

$$B\,\mathbf{u}^{(i+1)} = C\,\mathbf{u}^{(i)} + \tfrac{1}{2}\big(\,\mathbf{b}^{(i)} + \mathbf{b}^{(i+1)}\,\big),$$

where

$$B = \begin{pmatrix} 1+\mu & -\frac{1}{2}\,\mu & & \\ -\frac{1}{2}\,\mu & 1+\mu & -\frac{1}{2}\,\mu & \\ & -\frac{1}{2}\,\mu & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \qquad C = \begin{pmatrix} 1-\mu & \frac{1}{2}\,\mu & & \\ \frac{1}{2}\,\mu & 1-\mu & \frac{1}{2}\,\mu & \\ & \frac{1}{2}\,\mu & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix}. \tag{14.155}$$

The convergence is governed by the generalized eigenvalues of the matrix pair B, C, or, equivalently, the eigenvalues of the product $B^{-1}C$. According to Exercise ∎, these are

$$\lambda_k = \frac{1 - 2\,\mu \sin^2 \dfrac{\pi\, k}{2\,n}}{1 + 2\,\mu \sin^2 \dfrac{\pi\, k}{2\,n}}\,, \qquad\qquad k = 1, \dots, n-1. \tag{14.156}$$

If $\mu > 0$, are these eigenvalues are less than 1 in absolute value, so that the Crank–Nicholson scheme is also unconditionally stable. A detailed analysis based on the Taylor expansions will show that the errors are of the order of k^2 and h^2, and so it is reasonable to choose the time step to have the same order of magnitude as the space step, $k \approx h$. This gives the Crank–Nicholson scheme a considerable advantage over the other two schemes.

Example 14.15. ∎

Let us now look at numerical solution techniques for the wave equation. Although this is in a sense unnecessary, owing to the explicit d'Alembert formula (14.121) for the solution, the experience we gain in designing a suitable method will serve us well in more complicated situations, when there is no explicit formula, including one-dimensional inhomogeneous media, and higher dimensional problems.

Consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} \, , \qquad 0 < x < \ell, \qquad t \geq 0, \tag{14.157}$$

on a homogeneous bar of length *ell* with constant wave speed $c > 0$. To be specific, we impose Dirichlet boundary conditions

$$u(t,0) = \alpha(t), \qquad u(t,\ell) = \beta(t), \qquad t \geq 0. \tag{14.158}$$

and initial conditions

$$u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x), \qquad 0 \leq x \leq \ell. \tag{14.159}$$

We adopt the same uniformly spaced mesh

$$t_i = i\,k, \qquad x_j = j\,h,$$

where $h = \ell/n$, as in the heat equation.

In order to discretize the wave equation, we replace the second order derivatives by their standard finite difference approximations (14.134), namely

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2}\,(t_i,x_j) &\approx \frac{u(t_{i+1},x_j) - 2\,u(t_i,x_j) + u(t_{i-1},x_j)}{k^2} + \mathrm{O}(h^2), \\ \frac{\partial^2 u}{\partial x^2}\,(t_i,x_j) &\approx \frac{u(t_i,x_{j+1}) - 2\,u(t_i,x_j) + u(t_i,x_{j-1})}{h^2} + \mathrm{O}(k^2), \end{aligned} \tag{14.160}$$

Since the errors are of orders of k^2 and h^2, we expect to be able to choose the space and time step sizes of comparable magnitude:

$$k \approx h.$$

Substituting the finite difference formulae (14.160) into the partial differential equation (14.157), and rearranging terms, we are led to the iterative system

$$u_{i+1,j} = u_{i,j} + \sigma^2\,u_{i,j+1} + 2\,(1-\sigma^2)\,u_{i,j} + \sigma^2\,u_{i,j-1} - u_{i-1,j}, \qquad \begin{matrix} i = 1,2,\dots\,, \\ j = 1,\dots,n-1, \end{matrix} \tag{14.161}$$

for the numerical approximations $u_{i,j} \approx u(t_i,x_j)$, with parameter

$$\sigma = \frac{c\,k}{h} > 0. \tag{14.162}$$

The boundary conditions (14.158) require that

$$u_{i,0} = \alpha_i = \alpha(t_i), \qquad u_{i,n} = \beta_i = \beta(t_i), \qquad i = 0, 1, 2, \dots . \tag{14.163}$$

This allows us to rewrite the system in matrix form

$$\mathbf{u}^{(i+1)} = B\,\mathbf{u}^{(i)} - \mathbf{u}^{(i-1)} + \mathbf{b}^{(i)}, \tag{14.164}$$

where

$$B = \begin{pmatrix} 2\,(1-\sigma^2) & \sigma^2 & & & & \\ \sigma^2 & 2\,(1-\sigma^2) & \sigma^2 & & & \\ & \sigma^2 & \ddots & \ddots & & \\ & & \ddots & \ddots & \sigma^2 & \\ & & & \sigma^2 & 2\,(1-\sigma^2) \end{pmatrix}, \quad \mathbf{u}^{(j)} = \begin{pmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{n-2,j} \\ u_{n-1,j} \end{pmatrix}, \quad \mathbf{b}^{(j)} = \begin{pmatrix} \sigma^2\,\alpha_j \\ 0 \\ \vdots \\ 0 \\ \sigma^2\,\beta_j \end{pmatrix}. \tag{14.165}$$

The entries $u_{i,j}$ of $\mathbf{u}^{(i)}$ are, as in (14.146), the numerical approximations to the solution values $u(t_i, x_j)$ at the *interior* nodes. Note that the system (14.164) is a second order iterative scheme, since computing the $(i+1)^{\text{st}}$ iterate $\mathbf{u}^{(i+1)}$ requires the value of the preceding two iterates $u^{(i)}$ and $\mathbf{u}^{(i-1)}$.

The one difficulty is getting the method started. We know $\mathbf{u}^{(0)}$ since $u_{0,j} = f_j = f(x_j)$ is determined by the initial position. However, we also need to find $\mathbf{u}^{(1)}$ with entries $u_{1,j} \approx u(k, x_j)$ at time $t_1 = k$ in order to get off the ground, but the initial velocity $u_t(0,x) = g(x)$ prescribes the derivatives $u_t(0, x_j) = g(x_j) = g_j$ at time $t_0 = 0$ instead. One way to approach this would be to use the finite difference approximation

$$g_j = \frac{\partial u}{\partial t}\,(0, x_j) \approx \frac{u(k, x_j) - u(0, x_j)}{k} \approx \frac{u_{1,j} - g_j}{k} \tag{14.166}$$

to compute the required values

$$u_{1,j} = f_j + k\,g_j.$$

However, the approximation (14.166) is only accurate to order k, whereas the rest of the scheme has error proportional to k^2. Therefore, we would introduce a significantly larger error at the initial step, and the resulting solution would not have the desired order of accuracy.

In order to compute an initial approximation to $\mathbf{u}^{(1)}$ with error on the order of k^2, we need to analyze the local error in more details. Note that, by Taylor's theorem,

$$\frac{u(k, x_j) - u(0, x_j)}{k} \approx \frac{\partial u}{\partial t}\,(0, x_j) + \frac{k}{2}\,\frac{\partial^2 u}{\partial t^2}\,(0, x_j) = \frac{\partial u}{\partial t}\,(0, x_j) + \frac{c^2\,k}{2}\,\frac{\partial^2 u}{\partial x^2}\,(0, x_j)\,,$$

where the error is now of order k^2, and we have used the fact that u is a solution to the wave equation. Therefore, we find

$$\begin{aligned} u(k, x_j) &\approx u(0, x_j) + k\,\frac{\partial u}{\partial t}\,(0, x_j) + \frac{c^2\,k^2}{2}\,\frac{\partial^2 u}{\partial x^2}\,(0, x_j) \\ &= f(x_j) + k\,g(x_j) + \frac{c^2\,k^2}{2}\,f''(x_j) \approx f_j + k\,g_j + \frac{c^2\,k^2}{2\,h^2}(f_{j+1} - 2\,f_j + f_{j-1})\,, \end{aligned}$$

Figure 14.15. The Courant Condition.

where we can use the finite difference approximation (14.134) for the second derivative of $f(x)$ if no explicit formula is known. Therefore, we can initiate the scheme by setting

$$u_{1,j} = \tfrac{1}{2}\,\sigma^2\, f_{j+1} + (1 - \sigma^2) f_j + \tfrac{1}{2}\,\sigma^2\, f_{j-1} + k\, g_j, \tag{14.167}$$

maintain order k^2 (and h^2) accuracy.

Example 14.16. Consider the particular initial value problem ∎

The stability analysis of the numerical scheme proceeds as follows. We first need to recast the second order iterative system (14.164) into a first order system. As in Exercise ∎, this is accomplished by introducing the vector $\mathbf{z}^{(i)} = \begin{pmatrix} \mathbf{u}^{(i)} \\ \mathbf{u}^{(i-1)} \end{pmatrix} \in \mathbb{R}^{2n-2}$. Then

$$\mathbf{z}^{(i+1)} = C\,\mathbf{z}^{(i)} + \mathbf{c}^{(i)}, \qquad \text{where} \qquad C = \begin{pmatrix} B & -\mathrm{I} \\ \mathrm{I} & \mathrm{O} \end{pmatrix}. \tag{14.168}$$

Therefore, the stability of the method will be determined by the eigenvalues of the coefficient matrix C. The eigenvector equation $C\,\mathbf{z} = \lambda\,\mathbf{z}$, can be written out in components

$$B\,\mathbf{u} - \mathbf{v} = \lambda\,\mathbf{u}, \qquad \mathbf{u} = \lambda\,\mathbf{v}, \qquad \text{where} \qquad \mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}.$$

Substituting the second equation into the first, we find

$$(\lambda\, B - \lambda^2 - 1)\,\mathbf{v} = \mathbf{0}, \qquad \text{or} \qquad B\,\mathbf{v} = \left(\lambda + \frac{1}{\lambda}\right)\,\mathbf{v}.$$

The latter equation means that $\lambda + \lambda^{-1}$ is an eigenvalue of B and $\mathbf{v}$ the corresponding eigenvector. There is a straightforward connection between b and the matrix B, (14.148), with parameter $\mu = \sigma^2$ appearing the the numerical scheme for the heat equation. Using Lemma 14.14 and Exercise ∎, the eigenvalues of B are given by

$$\lambda + \frac{1}{\lambda} = 1 + 2\,\sigma^2\,\sin^2\frac{\pi\,k}{n}\,, \qquad k = 1, \dots, n-1.$$

Fixing k for the moment, we rewrite the eigenvalue equation in the form

$$\lambda^2 - 2\,a_k\,\lambda + 1 = 0, \qquad \text{where} \qquad a_k = \sigma^2\,\sin^2\frac{\pi\,k}{n}\,.$$

Each pair of solutions to this quadratic equation,

$$\lambda_k^{\pm} = a_k \pm \sqrt{a_k^2 - 1}\,, \tag{14.169}$$

gives two eigenvalues of the matrix C. If $a_k > 1$, then one of the two eigenvalues will be larger than one in magnitude, and hence the linear iterative system will have an exponentially growing mode, and hence $\| \mathbf{u}^{(i)} \| \to \infty$ as $i \to \infty$ for almost all choices of initial data. This is clearly incompatible with the wave equation solution that we are trying to approximate, which is periodic and hence remains bounded.

On the other hand, if $|\,a_k\,| < 1$, then the eigenvalues (14.169) are complex numbers of modulus 1, indicated stability (but not convergence) of the matrix C. Therefore, we should require that all $]a_1, \dots, a_{n-1}$ are less than 1 in magnitude, which is guaranteed provided

$$\sigma = \frac{c\,k}{h} < 1, \qquad \text{or} \qquad k < \frac{h}{c}\,. \tag{14.170}$$

This places a restriction on the relative sizes of the time and space steps, and hence the numerical scheme is conditionally stable.

The stability criterion (14.170) is known as the *Courant condition*, and can be assigned a simple geometric interpretation. Recall that the wave speed c is the slope of the characteristic lines for the wave equation. The Courant condition requires that the *mesh slope*, which is defined to be the ratio of the space step size to the time step size, namely h/k, must be strictly greater than the characteristic slope c. This implies that a signal starting at a mesh point (t_i, x_j) will reach positions $x_j \pm k/c$ at the next time $t_{i+1} = t_i + k$, which are still between the mesh points x_{j-1} and x_{j+1}. Thus, characteristic lines that start at a mesh point are not allowed to reach beyond the neighboring mesh points at the next time step.

For instance, in Figure 14.15, the wave speed is $c = 1.25$. The first figure has equal mesh spacing $k = h$, and does not satisfy the Courant condition (14.170), whereas the second figure has $k = \frac{1}{2}\,h$, which does. Note how the characteristic lines starting at a given mesh point have progressed beyond the neighboring mesh points after one time step in the first case, but not in the second.

Chapter 15

The Laplace Equation

The fundamental partial differential equations that govern the equilibrium mechanics of multi-dimensional media are the Laplace equation and its inhomogeneous counterpart, the Poisson equation. The Laplace equation is arguably the most important differential equation in all of applied mathematics. It arises in an astonishing variety of mathematical and physical systems, ranging through fluid mechanics, electromagnetism, potential theory, solid mechanics, heat conduction, geometry, probability, number theory, and on and on. The solutions to the Laplace equation are known as "harmonic functions", and the discovery of their many remarkable properties forms one of the most significant chapters in the history of mathematics.

In this chapter, we concentrate on the Laplace and Poisson equations in a two-dimensional (planar) domain. Their status as equilibrium equations implies that the solutions are determined by their values on the boundary of the domain. As in the one-dimensional equilibrium boundary value problems, the principal cases are Dirichlet or fixed, Neumann or free, and mixed boundary conditions arise. In the introductory section, we shall briefly survey the basic boundary value problems associated with the Laplace and Poisson equations. We also take the opportunity to summarize the crucially important tripartite classification of planar second order partial differential equations: *elliptic*, such as the Laplace equation; *parabolic*, such as the heat equation; and *hyperbolic*, such as the wave equation. Each species has quite distinct properties, both analytical and numerical, and each forms an essentially distinct discipline. Thus, by the conclusion of this chapter, you will have encountered all three of the most important genres of partial differential equations.

The most important general purpose method for constructing explicit solutions of linear partial differential equations is the method of separation of variables. The method will be applied to the Laplace and Poisson equations in the two most important coordinate systems — rectangular and polar. Linearity implies that we may combine the separable solutions, and the resulting infinite series expressions will play a similar role as for the heat and wave equations. In the polar coordinate case, we can, in fact, sum the infinite series in closed form, leading to the explicit Poisson integral formula for the solution. More sophisticated techniques, relying on complex analysis, but (unfortunately) only applicable to the two-dimensional case, will be deferred until Chapter 16.

Green's formula allows us to properly formulate the Laplace and Poisson equations in self-adjoint, positive definite form, and thereby characterize the solutions via a minimization principle, first proposed by the nineteenth century mathematician Lejeune Dirichlet, who also played a crucial role in putting Fourier analysis on a rigorous foundation. Minimization forms the basis of the most important numerical solution technique — the finite

element method that we first encountered in Chapter 11. In the final section, we discuss numerical solution techniques based on finite element analysis for the Laplace and Poisson equations and their elliptic cousins, including the Helmholtz equation and more general positive definite boundary value problems.

15.1. The Laplace Equation in the Plane.

The two-dimensional *Laplace equation* is the second order linear partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{15.1}$$

Along with the heat and wave equations, it completes the trinity of truly fundamental partial differential equations. A real-valued solution $u(x, y)$ to the Laplace equation is known as a *harmonic function*. The space of harmonic functions can thus be identified as the kernel of the second order linear partial differential operator

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\,, \tag{15.2}$$

known as the *Laplace operator*, or *Laplacian* for short. The inhomogeneous or forced version, namely

$$-\Delta[u] = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y) \tag{15.3}$$

is known as *Poisson's equation*. It forms the two-dimensional analogue of the basic equilibrium equation (11.12) for a bar, with the overall minus sign playing an analogous role.

The Laplace and Poisson equations arise as the basic equilibrium equations in a remarkable variety of physical systems. For example, we may interpret $u(x, y)$ as the displacement of a *membrane*, e.g., a drum skin. The inhomogeneity $f(x, y)$ in the Poisson equation represents an external forcing of the membrane. Another example is in the thermal equilibrium of planar bodies; here $u(x, y)$ represents the temperature and $f(x, y)$ an external heat source. In fluid mechanics and electrostatics, $u(x, y)$ represents the potential function whose gradient ∇u generates the corresponding flow; see below for details. The dynamical counterparts to the Laplace equation are multi-dimensional versions of the heat and wave equations, to be analyzed in Chapter 17.

Since both the Laplace and Poisson equations describe equilibria, they arise in most physical situations in the context of boundary value problems. We seek a solution $u(x, y)$ to the partial differential equation defined on a fixed bounded, open domain[†] $(x, y) \in \Omega \subset \mathbb{R}^2$. The solution will be required to satisfy suitable conditions on the boundary of the domain, denoted $\partial\Omega$, which will consist of one or more simple, closed curves, as illustrated in

[†] See Appendix A for the precise definitions of the terms "domain", "bounded", "boundary", etc.

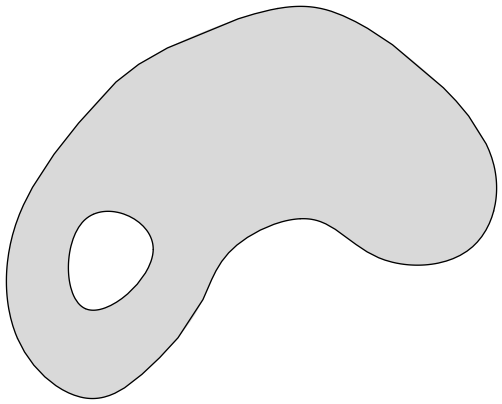


Figure 15.1. Planar Domain.

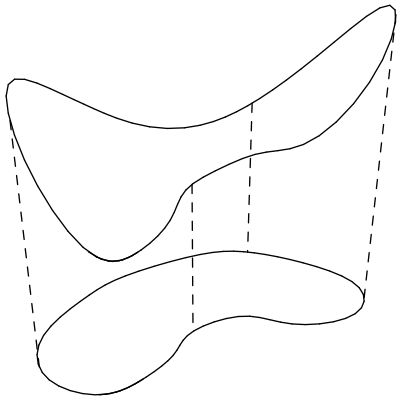


Figure 15.2. Dirichlet Boundary Conditions.

Figure 15.1. As in the one-dimensional case, there are several important types of boundary conditions.

The first are the *fixed* or *Dirichlet boundary conditions*, which specify the value of the function u on the boundary:

$$u(x,y) = h(x,y) \qquad \text{for} \qquad (x,y) \in \partial\Omega. \tag{15.4}$$

Under reasonable conditions on the type of domain, the Dirichlet conditions (15.4) serve to uniquely specify the solution $u(x,y)$ to the Laplace or Poisson equation. Physically, in the case of a free or forced membrane, the Dirichlet boundary conditions correspond to gluing the edge of the membrane to a wire at height $h(x,y)$ over each boundary point $(x,y) \in \partial\Omega$, as illustrated in Figure 15.2. Uniqueness means that the shape of the boundary wire will uniquely specify the vertical displacement of the membrane in equilibrium. Similarly, in the modeling of thermal equilibrium, a Dirichlet boundary condition represents the imposition of a prescribed temperature distribution, represented by the function h, along the boundary of the plate.

The second type of boundary conditions are the *Neumann boundary conditions*

$$\frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n} = k(x,y) \qquad \text{on} \qquad \partial\Omega, \tag{15.5}$$

in which the normal derivative of the solution u on the boundary is prescribed. For example, in thermal equilibrium, a Neumann boundary condition specifies the heat flux into the domain through its boundary. The most important are the "no-flux" or homogeneous Neumann boundary conditions, where $k(x,y) \equiv 0$. In thermomechanics, this corresponds to an insulated boundary. In the case of a membrane, it corresponds to the edge of the drum being left free. In fluid mechanics, where u represents the fluid potential, the no-flux conditions imply that the normal component of the velocity vector vanishes, and so corresponds to a solid boundary that does not allow the fluid to flow across it.

Finally, one can mix the boundary conditions, imposing Dirichlet conditions on part of the boundary, and Neumann on the complementary part. The general *mixed boundary value problem* has the form

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad\qquad u = h \quad \text{on} \quad D, \qquad \frac{\partial u}{\partial \mathbf{n}} = k \quad \text{on} \quad N, \tag{15.6}$$

with the boundary $\partial\Omega = D \cup N$ being the disjoint union of a "Dirichlet part", denoted by D, and a "Neumann part" N. For example, in heat conduction, if we want to find the equilibrium temperature distribution over a planar body, the Dirichlet part of the boundary is where we fix the temperature, while the Neumann part is insulated, or, more generally, has prescribed heat flux. Similarly, for displacement of a membrane, the Dirichlet part is where the edge of the drum is attached to a support, while the homogeneous Neumann part is where it is left hanging free.

Classification of Linear Partial Differential Equations in the Plane

We have, at last, encountered all three of the fundamental linear, second order, partial differential equations for functions of two variables. The homogeneous versions of the trinity are

a)	The wave equation:	$u_{tt} - c^2\, u_{xx} = 0,$	"hyperbolic",
b)	The heat equation:	$u_t - \gamma\, u_{xx} = 0,$	"parabolic",
c)	Laplace's equation:	$u_{xx} + u_{yy} = 0,$	"elliptic".

The last column is the equations' "type", according to a general taxonomy of partial differential equations. An explanation of the choice of terminology will appear later.

The wave, heat and Laplace equations are the prototypical representatives of the three most important genres of partial differential equations. The student should understand that there are fundamental differences. Equations governing vibrations, such as the wave equation, are typically hyperbolic. Equations governing diffusion, such as the heat equation, are parabolic. Hyperbolic and parabolic equations are dynamical processes, and one of the variables is identified with the time. On the other hand, equations of equilibrium, including the Laplace and Poisson equations, are typically elliptic, and only involve spatial

variables. Elliptic partial differential equations are associated with boundary value problems, whereas parabolic and hyperbolic equations lead to initial-boundary value problems, with, respectively, one or two required initial conditions. Furthermore, numerical solution techniques and requirements are of a fundamentally different character in all three cases.

While the initial tripartite classification is most evident in partial differential equations in two variables, the terminology and underlying properties of these three fundamental genres carries over to equations in higher dimensions. Most of the important partial differential equations arising in applications appear in one of these three general classes, and it is fair to say that the field of partial differential equations breaks into three major, disjoint subfields. Or, rather four subfields, the last being all the equations, including higher order equations, that do not fit into this neat categorization, which is, of course, yet further subdivided into a variety of subspecies.

The classification of linear, second order partial differential equations for a scalar-valued function $u(x, y)$ of two variables[†] proceeds in the following manner. The most general such equation has the form

$$L[u] = A\,u_{xx} + B\,u_{xy} + C\,u_{yy} + D\,u_x + E\,u_y + F\,u = f, \tag{15.7}$$

where the coefficients A, B, C, D, E, F are all allowed to be functions of (x, y), as is the inhomogeneity or forcing function $f = f(x, y)$. The equation is *homogeneous* if and only if $f \equiv 0$. We assume that at least one of the leading coefficients A, B, C is nonzero, as otherwise the equation is of first order.

The key quantity that determines the *type* of such a partial differential equation is its *discriminant*

$$\Delta(x, y) = B^2 - 4\,AC. \tag{15.8}$$

This should (and for good reason) remind the reader of the discriminant of the quadratic equation

$$Q(\xi, \eta) = A\,\xi^2 + B\,\xi\,\eta + C\,\eta^2 + D\,\xi + E\,\eta + F = 0. \tag{15.9}$$

The set of solutions (ξ, η) to such an equation describes a curve; namely, a conic section. In the nondegenerate cases, the discriminant $\Delta = B^2 - 4\,AC$ determines its geometrical type; it is an ellipse when $\Delta > 0$, a parabola when $\Delta = 0$, or a hyperbola when $\Delta < 0$. This tripartite classification provides the underlying motivation for the terminology used to classify second order partial differential equations.

Definition 15.1. A linear, second order partial differential equation (15.7) at a point (x, y) is called

a)	*elliptic*		$\Delta(x, y) < 0$,
b)	*parabolic*	if	$\Delta(x, y) = 0$,
c)	*hyperbolic*		$\Delta(x, y) > 0$.

[†] For dynamical equations, we will identify y as the time variable t.

In particular, the wave equation $u_{xx} - u_{yy} = 0$ has discriminant $\Delta = -4$, and is hyperbolic. The heat equation $u_{xx} - u_y = 0$ has discriminant $\Delta = 0$, and is parabolic. Finally, the Poisson equation $u_{xx} + u_{yy} = -f$ has discriminant $\Delta = 4$, and is elliptic.

Example 15.2. Since the coefficients in the partial differential equation are allowed to vary over the domain, the type of an equation can vary from point to point. Equations that change type are much less common, as well as being much harder to handle. One example arising in the theory of supersonic aerodynamics is the *Tricomi equation*

$$y\, u_{xx} - u_{yy} = 0. \tag{15.10}$$

Comparing with (15.7), we find that $A = y$, $C = -1$ and $B = D = E = F = f = 0$. The discriminant in this particular case is $\Delta = 4y$, and hence the equation is hyperbolic when $y > 0$, elliptic when $y < 0$, and parabolic on the transition line $y = 0$. The hyperbolic region corresponds to subsonic fluid flow, while the supersonic regions are of elliptic type. The transitional parabolic boundary represents the shock line between the sub- and supersonic regions.

Characteristics

Certain curves play a distinguished role in the analysis of second order, linear partial differential equations. A smooth curve $\mathbf{x}(t)$ is called a *characteristic curve* for the second order partial differential equation (15.7) if its tangent vector $\dot{\mathbf{x}} = \begin{pmatrix} \dot{x} & \dot{y} \end{pmatrix}^T \neq \mathbf{0}$ satisfies the quadratic *characteristic equation*

$$A(x,y)\, \dot{y}^2 - B(x,y)\, \dot{x}\, \dot{y} + C(x,y)\, \dot{x}^2 = 0. \tag{15.11}$$

Pay careful attention to the form of the characteristic equation; in particular, the first and zero$^{\text{th}}$ order terms in the original partial differential equation play no role.

For example, consider the wave equation[†]

$$u_{yy} - c^2\, u_{xx} = 0.$$

In this case, $A = -c^2, B = 0, C = 1$, and so (15.11) takes the form

$$\dot{x}^2 - c^2\, \dot{y}^2 = 0, \qquad \text{and so} \qquad \dot{x} = \pm\, c\, \dot{y}.$$

The solutions to the resulting ordinary differential equation are

$$x(t) = \pm\, c\, y(t) + k, \tag{15.12}$$

where k is an integration constant. Therefore, the wave equation has two characteristic curves passing through each point (a, b), namely the straight lines (15.12) of slope $\pm\, 1/c$. Thus, the general definition of characteristic curve is in accordance with our earlier definition (14.122) of the characteristic lines for the wave equation. (In our earlier discussion,

[†] Here we are using y as the "time" variable, rather than t, which is now playing the role of the curve parameter.

the geometrical roles of the x and $y = t$ variables were reversed, which is why we now find the reciprocal characteristic value of the slope.)

On the other hand, the Laplace equation

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

has no (real) characteristic curves since the characteristic equation (15.11) reduces to $\dot{x}^2 + \dot{y}^2 = 0$. Finally, for the heat equation

$$u_{xx} - u_y = 0,$$

the characteristic equation is simply $\dot{y}^2 = 0$, and so there is only one characteristic curve through each point (a, b), namely the horizontal line $y = b$. In this manner, one distinguishes elliptic, parabolic, and hyperbolic partial differential equations by the number of (real) characteristic curves passing through a point — namely, zero, one and two, respectively. Further discussion of characteristics for nonlinear partial differential equations can be found in Section 22.1.

Some general remarks on the role of characteristic curves follow, albeit without proof. As with the wave equation, signals and localized waves in a partial differential equation tend to propagate along the characteristic curves. This fact lies at the foundation of geometric optics. Light rays move along characteristic curves, and are thereby subject to the optical phenomena of refraction and focusing. Similarly, since the characteristic curves for the heat equation are the horizontal lines parallel to the x axis, the signals propagate instantaneously, in accordance with our observation that the effect on an initial concentrated heat source is immediately felt all along the bar. Finally, elliptic equations have no characteristics, and as a consequence, do not admit propagating signals; the effect of a localized disturbance, say on a membrane, is immediately felt everywhere.

15.2. Separation of Variables.

One of the earliest — and still most widely used — techniques for constructing explicit analytical solutions to partial differential equations is the method of *separation of variables*. We have, in fact, already applied the separation of variables method to construct particular solutions to the heat and wave equations. In each case, we looked for a solution in the form of a product $u(t, x) = h(t)\, v(x)$. In the general separation of variables method, one does not know either factor in advance. If the method succeeds (which is not guaranteed), both will be determined as solutions to associated ordinary differential equations.

For the Laplace equation, the solution depends on x and y, and so the separation of variables ansatz becomes

$$u(x, y) = v(x)\, w(y). \tag{15.13}$$

Let us substitute this expression into the Laplace equation. First of all,

$$\frac{\partial^2 u}{\partial x^2} = v''(x)\, w(y), \qquad \frac{\partial^2 u}{\partial y^2} = v(x)\, w''(y),$$

where the primes indicate ordinary derivatives, and so

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = v''(x)\, w(y) + v(x)\, w''(y) = 0.$$

The method of separation of variables will succeed if we are able to manipulate the resulting equation so as to place all of the terms involving x on one side of the equation and all the terms involving y on the other. Here, we first write

$$v''(x)\, w(y) = -\, v(x)\, w''(y).$$

Dividing both sides by $v(x)\, w(y)$ (which we assume is not identically zero as otherwise the solution would be trivial) yields

$$\frac{v''(x)}{v(x)} = -\,\frac{w''(y)}{w(y)} \equiv \lambda, \tag{15.14}$$

and effectively separates the x and y variables on each side of the equation. Now, how could a function of x alone be equal to a function of y alone? A moment's reflection should convince the reader that this can happen if and only if the two functions are constant[†]. We use λ, known as the *separation constant*, to designate this common value. Then (15.14) reduces to a pair of ordinary differential equations

$$v'' - \lambda\, v = 0, \qquad\qquad w'' + \lambda\, w = 0,$$

for the individual factors $v(x)$ and $w(y)$. We already know how to solve both of these ordinary differential equations by elementary techniques. There are three different cases, depending on the sign of the separation constant λ. Each case leads to four different solutions, and we collect the entire family of separable solutions together in the following table.

Separable Solutions to Laplace's Equation

λ	$v(x)$	$w(y)$	$u(x,y) = v(x)\, w(y)$
$\lambda = -\,\omega^2 < 0$	$\cos \omega x,\ \sin \omega x$	$e^{-\omega y},\ e^{\omega y},$	$e^{\omega y} \cos \omega x,\quad e^{\omega y} \sin \omega x,$ $e^{-\omega y} \cos \omega x,\quad e^{-\omega y} \sin \omega x$
$\lambda = 0$	$1,\ x$	$1,\ y$	$1,\ x,\ y,\ x\, y$
$\lambda = \omega^2 > 0$	$e^{-\omega x},\ e^{\omega x}$	$\cos \omega y,\ \sin \omega y$	$e^{\omega x} \cos \omega y,\quad e^{\omega x} \sin \omega y,$ $e^{-\omega x} \cos \omega y,\quad e^{-\omega x} \sin \omega y$

[†] Technical detail: one should assume that the underlying domain is connected for this to be true; however, in practical analysis, this technicality is irrelevant.

Since Laplace's equation is linear, we can utilize superposition to combine these types of solutions together, either as finite linear combinations, or, provided we pay proper attention to convergence issues, as infinite series.

For boundary value problems, the applicability of such separable solutions imposes fairly severe restrictions on the geometry of the domain. The ansatz (15.13) effectively requires that the domain be rectangular. Thus, we are led to consider the boundary value problem for Laplace's equation

$$\Delta u = 0 \qquad \text{on a rectangle} \qquad R = \{\, 0 < x < a, \;\; 0 < y < b \,\}. \tag{15.15}$$

To illustrate the method, consider the following Dirichlet boundary conditions

$$u(x,0) = f(x), \qquad u(x,b) = 0, \qquad u(0,y) = 0, \qquad u(a,y) = 0. \tag{15.16}$$

We are only allowing a nonzero boundary condition on one of the four sides of the rectangle, in order to simplify the analysis. The Dirichlet boundary value problem can then be solved by adding together the solutions to the four boundary value problems which only have nonzero boundary conditions on one side of the rectangle; see Exercise ∎.

Of the variety of solutions available through separation of variables, the only ones that will play a role are those that respect the boundary conditions. Putting the nonzero boundary condition aside for the moment, we ask that $u(x,y) = v(x)\,w(y)$ be zero on the top, right and left sides of the rectangle. This requires

$$v(0) = v(a) = 0, \qquad\qquad w(b) = 0.$$

Referring to the above table, the first condition $v(0) = 0$ requires

$$v(x) = \begin{cases} \sin \omega x, & \lambda = \omega^2 > 0, \\ x, & \lambda = 0, \\ \sinh \omega x, & \lambda = -\,\omega^2 < 0, \end{cases}$$

where $\sinh z = \frac{1}{2}(e^z - e^{-z})$ is the usual hyperbolic sine function. The second and third cases cannot satisfy the second boundary condition $v(a) = 0$, and so we discard them. The first case leads to the condition

$$v(a) = \sin \omega\, a = 0, \qquad \text{and hence} \qquad \omega\, a = \pi, 2\pi, 3\pi, \dots \,.$$

Therefore, the separation constant has the form

$$\lambda = \omega^2 = \frac{n^2 \pi^2}{a^2}\,, \tag{15.17}$$

with the corresponding solutions

$$v(x) = \sin \frac{n \pi x}{a}, \qquad\qquad n = 1, 2, 3, \dots \,. \tag{15.18}$$

Remark: We have just recomputed the known eigenvalues and eigenfunctions of the familiar boundary value problem

$$v'' + \lambda\, v = 0, \qquad\qquad v(0) = v(a) = 0.$$

The third boundary condition $w(b) = T0$ requires that, up to constant multiple,

$$w(y) = \sinh \omega\,(b-y) = \sinh \frac{n\,\pi\,(b-y)}{a}. \tag{15.19}$$

Therefore, each of the separable solutions

$$u_n(x,y) = \sin \frac{n\,\pi\,x}{a}\ \sinh \frac{n\,\pi\,(b-y)}{a}\,, \qquad\qquad n = 1, 2, 3, \dots\,, \tag{15.20}$$

satisfies the three homogeneous boundary conditions. It remains to analyze the boundary condition along the bottom edge of the rectangle. We try a linear superposition of the separable solutions in the form of an infinite series

$$u(x,y) = \sum_{n=1}^{\infty}\ c_n\,u_n(x,y) = \sum_{n=1}^{\infty}\ c_n\,\sin \frac{n\,\pi\,x}{a}\ \sinh \frac{n\,\pi\,(b-y)}{a}\,,$$

where the coefficients $c_1, c_2, \dots$ are to be determined by the remaining inhomogeneous boundary condition. At the bottom edge $y = 0$ we find

$$u(x,0) = \sum_{n=1}^{\infty}\ c_n\,\sinh \frac{n\,\pi\,b}{a}\ \sin \frac{n\,\pi\,x}{a} = f(x), \tag{15.21}$$

which takes the form of a Fourier sine series for the function $f(x)$. According to (12.72), for the interval $0 < x < a$, the coefficients b_n of the Fourier sine series

$$f(x) = \sum_{n=1}^{\infty}\ b_n\,\sin \frac{n\,\pi\,x}{a} \qquad \text{are given by} \qquad b_n = \frac{2}{a}\int_0^a f(x)\sin \frac{n\,\pi\,x}{a}\,dx. \tag{15.22}$$

Comparing (15.21), (15.22), we see that

$$c_n \sinh \frac{n\,\pi b}{a} = b_n \qquad\qquad \text{or} \qquad\qquad c_n = \frac{b_n}{\sinh \dfrac{n\,\pi\,b}{a}}\,.$$

Therefore, the solution to the boundary value problem takes the form of an infinite series

$$u(x,y) = \sum_{n=1}^{\infty}\ b_n\,\sin \frac{n\,\pi\,x}{a}\ \frac{\sinh \dfrac{n\,\pi\,(b-y)}{a}}{\sinh \dfrac{n\,\pi\,b}{a}}\,, \tag{15.23}$$

where b_n are the Fourier sine coefficients (15.22) of $f(x)$. It can be shown, cf. Exercise ■, that if f is integrable, $\displaystyle\int_0^a |\,f(x)\,|\,dx < \infty$, then the series solution converges on the entire rectangle R. Moreover, if $y > 0$, the go to zero exponentially fast, and so the solution can be well approximated by partial summation. The exponentially fast decay of the Fourier coefficients implies that $u(x,y)$ is an infinitely differentiable function of x at each $y > 0$. In fact, as we shall see, the solutions to the Laplace equation are always *analytic* functions inside the domain — even when the boundary conditions are quite unsmooth.

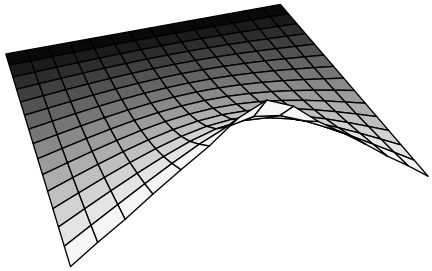


Figure 15.3. Square Membrane on a Wire.

Example 15.3. A membrane is stretched over a wire in the shape of a unit square with one side bent in half, as graphed in Figure 15.3. The precise boundary conditions are

$$u(x,y) = \begin{cases} x, & 0 \le x \le \frac{1}{2}, & y = 0, \\ 1-x, & \frac{1}{2} \le x \le 1, & y = 0, \\ 0, & 0 \le x \le 1, & y = 1, \\ 0, & x = 0, & 0 \le y \le 1, \\ 0, & x = 1, & 0 \le y \le 1. \end{cases}$$

The Fourier sine series of the inhomogeneous boundary function is readily computed:

$$f(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2}, \\ 1-x, & \frac{1}{2} \le x \le 1, \end{cases}$$
$$= \frac{4}{\pi^2}\left(\sin \pi x - \frac{\sin 3\pi x}{9} + \frac{\sin 5\pi x}{25} - \cdots\right) = \frac{4}{\pi^2}\sum_{m=0}^{\infty} (-1)^m \frac{\sin(2m+1)\pi x}{(2m+1)^2}.$$

Therefore, the solution to the boundary value problem is given by the Fourier series

$$u(x,y) = \frac{4}{\pi^2}\sum_{m=0}^{\infty} (-1)^m \frac{\sin(2m+1)\pi x \, \sinh(2m+1)\pi(1-y)}{(2m+1)^2 \sinh(2m+1)\pi}.$$

For $y > 0$, the series converges rapidly owing to the exponential decay of its terms, and so can be well approximated by its first few summands. In Figure 15.3 we graph the sum of the first 10 terms in the series, which is a reasonably good approximation except when we are very close to the raised corner of the wire, which is the point of maximal displacement of the membrane. This is indicative of a very general and important fact: a harmonic function achieves its maximum and minimum values only on the boundary of its domain; see Corollary 15.8 for details.

Polar Coordinates

The method of separation of variables can be used in certain other very special geometries. One particularly important case is a circular disk. Let us take the disk to have

radius 1, centered at the origin. Consider the Dirichlet boundary value problem

$$\Delta u = 0, \qquad x^2 + y^2 < 1, \qquad \text{and} \qquad u = h, \qquad x^2 + y^2 = 1, \tag{15.24}$$

so that the function $u(x, y)$ satisfies the Laplace equation on the unit disk and has Dirichlet boundary conditions on the unit circle. For example, $u(x, y)$ might represent the displacement of a circular drum that is attached to a wire of height

$$h(x, y) = h(\cos\theta, \sin\theta) \equiv h(\theta), \qquad 0 \le \theta \le 2\,\pi, \tag{15.25}$$

above each point $(x, y) = (\cos\theta, \sin\theta)$ on the unit circle.

The rectangular separable solutions are not particularly helpful in this situation. The fact that we are dealing with a circular geometry inspires us to adopt polar coordinates

$$x = r\cos\theta, \qquad y = r\sin\theta, \qquad \text{or} \qquad r = \sqrt{x^2 + y^2}\,, \qquad \theta = \tan^{-1}\frac{y}{x}\,,$$

and write the solution as a function of r, θ. We also need to relate derivatives with respect to x and y to those with respect to r and θ. Performing a standard chain rule computation, we find

$$\begin{aligned} \frac{\partial}{\partial r} &= \cos\theta\ \frac{\partial}{\partial x} + \sin\theta\ \frac{\partial}{\partial y}\,, \\ \frac{\partial}{\partial \theta} &= -\,r\sin\theta\ \frac{\partial}{\partial x} + r\cos\theta\ \frac{\partial}{\partial y}\,, \end{aligned} \qquad \text{so} \qquad \begin{aligned} \frac{\partial}{\partial x} &= \cos\theta\ \frac{\partial}{\partial r} - \frac{\sin\theta}{r}\ \frac{\partial}{\partial \theta}\,, \\ \frac{\partial}{\partial y} &= \sin\theta\ \frac{\partial}{\partial r} + \frac{\cos\theta}{r}\ \frac{\partial}{\partial \theta}\,. \end{aligned} \tag{15.26}$$

These formulae allow us to rewrite the Laplace equation in polar coordinates; after some calculation in which many of the terms cancel, we find

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \theta^2} = 0. \tag{15.27}$$

The boundary conditions are on the unit circle $r = 1$, and so, by (15.25), take the form

$$u(1, \theta) = h(\theta).$$

Note especially that $u(r, \theta)$ and the boundary value $h(\theta)$ are 2π periodic functions in the angular coordinate:

$$u(r, \theta + 2\,\pi) = u(r, \theta), \qquad h(\theta + 2\,\pi) = h(\theta). \tag{15.28}$$

Polar separation of variables is based on the product ansatz

$$u(r, \theta) = v(r)\, w(\theta). \tag{15.29}$$

Substituting (15.29) into the polar form (15.27) of Laplace's equation, we find

$$v''(r)\, w(\theta) + \frac{1}{r}\, v'(r)\, w(\theta) + \frac{1}{r^2}\, v(r)\, w''(\theta) = 0.$$

We now separate variables by moving all the terms involving r onto one side of the equation and all the terms involving θ onto the other. This is accomplished by first rewriting the equation in the form

$$\left(v''(r) + \frac{1}{r}\, v'(r) \right) w(\theta) = -\, \frac{1}{r^2}\, v(r)\, w''(\theta),$$

and then dividing by the product $v(r)\, w(\theta)$, whence

$$\frac{r^2\, v''(r) + r\, v'(r)}{v(r)} = -\, \frac{w''(\theta)}{w(\theta)} = \lambda.$$

As in the rectangular case, a function of r can equal a function of θ if and only if both are equal to a common separation constant λ. Therefore, the partial differential equation reduces to a pair of ordinary differential equations

$$r^2\, v'' + r\, v' - \lambda\, r = 0, \qquad\qquad w'' + \lambda\, w = 0, \tag{15.30}$$

for the components of the separable solution (15.29). These take the form of eigenfunction equations in which the separation constant λ plays the role of the eigenvalue.

We have already solved the eigenvalue problem for $w(\theta)$. According to (15.28), $w(\theta + 2\,\pi) = w(\theta)$ must be a $2\,\pi$ periodic eigenfunction. Therefore, the eigenvalues (separation constants) are $\lambda = n^2$, with associated eigenfunctions

$$1, \qquad\qquad \sin n\,\theta, \qquad\qquad \cos n\,\theta, \qquad\qquad n = 0, 1, 2, \dots\ . \tag{15.31}$$

Using the value $\lambda = n^2$, the remaining ordinary differential equation

$$r^2\, v'' + r\, v' - n^2\, r = 0. \tag{15.32}$$

has the form of a second order Euler equation for $v(r)$. As discussed in Example 7.34, the solutions are obtained by substituting the power ansatz $v(r) = r^k$ into the equation. The resulting "characteristic equation" requires

$$k^2 - n^2 = 0, \qquad \text{and hence} \qquad k = \pm\, n.$$

Therefore, if $n \neq 0$, we find two linearly independent solutions,

$$v_1(r) = r^n, \qquad\qquad v_2(r) = r^{-n}, \qquad\qquad n = 1, 2, \dots. \tag{15.33}$$

If $n = 0$, we have an additional logarithmic solution

$$v_1(r) = 1, \qquad\qquad v_2(r) = \log r, \qquad\qquad n = 0, \tag{15.34}$$

as in Exercise ∎. Combining (15.31) and (15.33), (15.34), we recover the following separable polar coordinate solutions to the Laplace equation

$$\begin{array}{lll} 1, & r^n \cos n\,\theta, & r^n \sin n\,\theta, \\ \log r, & r^{-n} \cos n\,\theta, & r^{-n} \sin n\,\theta, \end{array} \qquad n = 1, 2, 3, \dots. \tag{15.35}$$

Now, the solutions in the top row of (15.35) are continuous (in fact analytic) at the origin, whereas the solutions in the bottom row have singularities as $r \to 0$. The latter

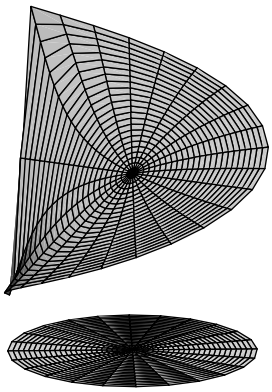


Figure 15.4. Membrane Attached to Helical Wire.

are not relevant since we require the solution $u(x, y)$ to remain bounded and smooth — even at the center of the disk. Thus, we should only use the former in concocting a series solution

$$u(r, \theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n r^n \cos n\theta + b_n r^n \sin n\theta \big). \tag{15.36}$$

At the boundary $r = 1$, we must have

$$u(1, \theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n \cos n\theta + b_n \sin n\theta \big) = h(\theta).$$

Therefore,

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} h(\theta) \cos n\theta \, d\theta, \qquad\qquad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} h(\theta) \sin n\theta \, d\theta, \tag{15.37}$$

are precisely the Fourier coefficients (12.21) of the boundary value function $h(\theta)$.

Remark: Introducing the complex variable $z = r\, e^{\,\mathrm{i}\,\theta} = x + \mathrm{i}\, y$ allows us to write

$$z^n = r^n\, e^{\,\mathrm{i}\, n\theta} = r^n \, \cos n\theta + \mathrm{i}\, r^n \, \sin n\theta. \tag{15.38}$$

Therefore, the separable solutions for $n \geq 1$ are nothing but the harmonic polynomial solutions derived in Example 7.51, namely

$$r^n \cos n\theta = \operatorname{Re} z^n, \qquad\qquad r^n \sin n\theta = \operatorname{Im} z^n. \tag{15.39}$$

Exploitation of the remarkable connections between the solutions to the Laplace equation and complex functions will form the focus of Chapter 16.

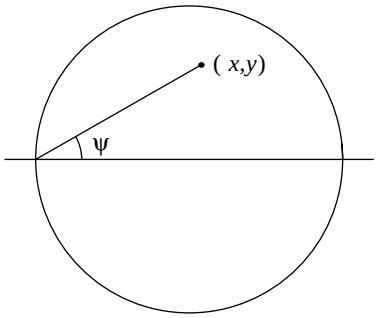


Figure 15.5. Geometrical Construction of the Solution.

Example 15.4. Consider the boundary value problem on the unit disk with

$$u(1,\theta) = \theta \qquad \text{for} \qquad -\pi < \theta < \pi. \tag{15.40}$$

The boundary data can be interpreted as attaching a circular membrane to a wire in the shape of a single turn of a spiral helix bent over the unit circle. The wire has a jump discontinuity at $(-1, 0)$. The Fourier series for $h(\theta) = \theta$ was computed in Example 12.2, namely

$$h(\theta) = \theta \quad \sim \quad 2\left(\sin\theta - \frac{\sin 2\theta}{2} + \frac{\sin 3\theta}{3} - \frac{\sin 4\theta}{4} + \cdots\right).$$

Therefore, the solution to the Laplace equation with these boundary conditions is

$$u(r,\theta) = 2\left(r\sin\theta - \frac{r^2\sin 2\theta}{2} + \frac{r^3\sin 3\theta}{3} - \frac{r^4\sin 4\theta}{4} + \cdots\right). \tag{15.41}$$

In fact, this series can be explicitly summed. Using (15.38), we find

$$u(x,y) = 2\ \mathrm{Im}\left(z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots\right) = 2\ \mathrm{Im}\ \log(1+z)$$

is twice the imaginary part of the Taylor series for $\log(1+z)$. If we write $1 + z = \rho\, e^{\,\mathrm{i}\,\psi} = \exp(\log\rho + \,\mathrm{i}\,\psi)$, then the solution (15.41) is given in the explicit form

$$u(x,y) = 2\,\psi = 2\,\tan^{-1}\frac{y}{1+x}\,, \tag{15.42}$$

and is plotted in Figure 15.4. The quantity ψ is the angle that the line passing through the two points (x, y) and $(-1, 0)$ makes with the x-axis, as in Figure 15.5. You should try to convince yourself that, on the unit circle, $2\,\psi = \theta$ has the correct boundary values! Moreover, even though the boundary values are discontinuous, the solution is an analytic function inside the disk.

Unlike the rectangular series solution, the general Fourier series solution (15.36) for a disk can, in fact, be summed in closed form! If we substitute the Fourier formulae

(15.37) into (15.36) — remembering to change the integration variable to, say, ϕ to avoid a notational conflict — we find

$$\begin{aligned} u(r,\theta) &= \frac{1}{\pi}\int_{-\pi}^{\pi} h(\phi)\left(\frac{1}{2}+\sum_{n=1}^{\infty} r^n\big[\cos n\theta\cos n\phi+\sin n\theta\sin n\phi\big]\right)d\phi \\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} h(\phi)\left(\frac{1}{2}+\sum_{n=1}^{\infty} r^n\cos n(\theta-\phi)\right)d\phi. \end{aligned} \tag{15.43}$$

We next show how to sum the series in brackets. Using (15.38), we can write it as the real part of a geometric series:

$$\begin{aligned} &\frac{1}{2}+\sum_{n=1}^{\infty} r^n\cos n\theta = \mathrm{Re}\left(\frac{1}{2}+\sum_{n=1}^{\infty} z^n\right) = \mathrm{Re}\left(\frac{1}{2}+\frac{z}{1-z}\right) = \mathrm{Re}\left(\frac{1+z}{2(1-z)}\right) \\ &= \mathrm{Re}\left(\frac{(1+z)(1-\overline{z})}{2\,|\,1-z\,|^2}\right) = \frac{\mathrm{Re}\,(1+z-\overline{z}-|\,z\,|^2)}{2\,|\,1-z\,|^2} = \frac{1-|\,z\,|^2}{2\,|\,1-z\,|^2} = \frac{1-r^2}{2\,(1+r^2-2\,r\cos\theta)}\,. \end{aligned}$$

Substituting this formula back into (15.43), we have deduced the important *Poisson Integral Formula* for the solution to the boundary value problem, named after the French mathematician Siméon–Denis Poisson.

Theorem 15.5. *The solution $u(r,\theta)$ to the Laplace equation in the unit disk with Dirichlet boundary conditions $u(1,\theta)=h(\theta)$ is*

$$u(r,\theta) = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} h(\phi)\,\frac{1-r^2}{1+r^2-2\,r\cos(\theta-\phi)}\,d\phi. \tag{15.44}$$

Example 15.6. A particularly important case is when the boundary value

$$h(\theta)=\delta(\theta-\phi)$$

is a delta function concentrated at the point $(\cos\phi,\sin\phi)$ on the unit circle. The solution to the resulting boundary value problem is the *Poisson integral kernel*

$$u(r,\theta) = \frac{1-r^2}{2\,\pi\big[\,1+r^2-2\,r\cos(\theta-\phi)\,\big]} = \frac{1-|\,z\,|^2}{2\,\pi\,|\,1-z\,e^{-\,\mathrm{i}\,\phi}\,|^2}\,. \tag{15.45}$$

The reader may enjoy verifying that this function does indeed, solve the Laplace equation and has the correct boundary values in the limit as $r\to 1$. Physically, if $u(r,\theta)$ represents the equilibrium temperature of the disk, then the delta function boundary values correspond to a unit concentrated heat source being applied at a single point on the boundary. The solution is sketched in Figure 15.6. The general Poisson integral formula (15.44) results from our general superposition principle, based on the fact that general boundary data can be written as a superposition,

$$h(\theta)=\int_{-\pi}^{\pi} h(\phi)\,\delta(\phi-\theta)\,d\phi,$$

of delta functions.

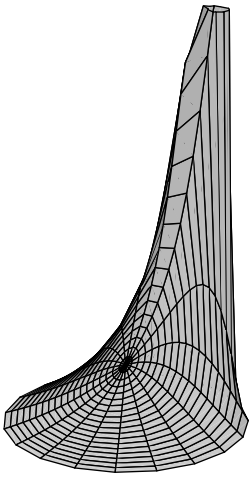


Figure 15.6. The Poisson Kernel.

If we set $r = 0$ in the Poisson formula (15.44), then we obtain

$$u(0, \theta) = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} h(\phi)\, d\phi. \tag{15.46}$$

The left hand side is the value of u at the origin; the right hand side is the average of its boundary values around the unit circle. This is a particular case of an important general fact.

Theorem 15.7. *The value of a harmonic function u at a point (x_0, y_0) is equal to the average of its values on any circle centered at the point:*

$$u(x_0, y_0) = \frac{1}{2\,\pi\, r} \oint_C u\, ds = \frac{1}{2\,\pi} \int_0^{2\,\pi} u(x_0 + r\cos\theta, y_0 + r\sin\theta)\, d\theta. \tag{15.47}$$

The result requires that u be harmonic on the entire closed disk bounded by this circle.

Proof: We use a scaling and translation to map the disk of radius r centered at (x_0, y_0) to the unit disk centered at the origin. Specifically, we set

$$U(x, y) = u(x_0 + r\,x, y_0 + r\,y). \tag{15.48}$$

An easy chain rule computation proves that $U(x, y)$ is harmonic on the unit disk, with boundary values

$$h(\theta) = U(\cos\theta, \sin\theta) = u(x_0 + r\cos\theta, y_0 + r\sin\theta).$$

Therefore, by (15.46) ,

$$U(0, 0) = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} h(\theta)\, d\theta = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} U(\cos\theta, \sin\theta)\, d\theta.$$

Replacing U by its formula (15.48) produces the desired result. *Q.E.D.*

An important consequence of the integral formula (15.47) is the *Maximum Principle* for harmonic functions.

Corollary 15.8. *If $u(x, y)$ is a nonconstant harmonic function defined on a domain Ω, then u does not have a local maximum or local minimum at any interior point of Ω.*

Proof: The average of a real function lies strictly between its maximum and minimum values (except in the trivial case when the function is constant). Theorem 15.7 therefore implies that $u(x, y)$ lies strictly between its maximal and minimal values on any small circle centered at (x, y). But if $u(x, y)$ had a local maximum (minimum), then it would be larger (smaller) than its values at all nearby points, and, in particular, its values on a small circle around the point. This contradiction proves the theorem. *Q.E.D.*

As a consequence, harmonic functions achieve their maxima and minima only at boundary points of a domain. Any interior critical point, where $\nabla u = \mathbf{0}$, must be a saddle point. Physically, if we interpret $u(x, y)$ as the vertical displacement of an unforced membrane, then Corollary 15.8 says that the membrane cannot have any internal bumps — its highest and lowest points are necessarily on the boundary of the domain. This reconfirms our physical intuition: the restoring force exerted by the stretched membrane will serve to flatten any bump, and hence a membrane with a local maximum or minimum cannot be in equilibrium. A similar interpretation holds for heat conduction. A body in thermal equilibrium can achieve its maximum and minimum temperature only on the boundary of the domain. Again, physically, heat energy would flow away from any internal maximum, or towards any local minimum, and so if the body contained a local maximum or minimum on its interior, it could not be in thermal equilibrium.

This concludes our discussion of the method of separation of variables and series solutions to the planar Laplace equation. The method of separation of variables does apply in a few other special coordinate systems. See Exercise ■ for one example, and [**105**, **108**, **110**] for a complete account, including connections with underlying symmetries of the equation.

15.3. The Green's Function.

Now we turn to the Poisson equation (15.3), which is the inhomogeneous form of the Laplace equation. In Section 11.2, we learned how to solve one-dimensional boundary value problems by use of the Green's function. This important technique can be adapted to solve inhomogeneous boundary value problems for elliptic partial differential equations in higher dimensions, including Poisson's equation. As in the one-dimensional situation, the Green's function is the solution to the homogeneous boundary value problem in which the inhomogeneity is a concentrated unit impulse — a delta function. The solution to the general forced boundary value problem is then obtained via linear superposition, that is, as a convolution integral with the Green's function.

The first order of business is to establish the proper form for the delta function or unit impulse in our two-dimensional situation. We denote the *delta function* concentrated at position $\boldsymbol{\xi} = (\xi, \eta) \in \mathbb{R}^2$ by

$$\delta_{\boldsymbol{\xi}}(\mathbf{x}) = \delta_{(\xi,\eta)}(x, y) = \delta(\mathbf{x} - \boldsymbol{\xi}). \tag{15.49}$$

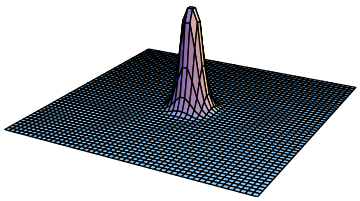


Figure 15.7. Gaussian Distributions Converging to the Delta Function.

The delta function can be viewed as the limit, as $n \to \infty$, of a sequence of more and more highly concentrated unit sources $f_n(x,y)$, which are required to satisfy

$$\lim_{n\to\infty} f_n(x,y) = 0, \quad \text{for} \quad (x,y) \neq (0,0), \qquad \text{while} \qquad \iint_\Omega f_n(x,y)\,dx\,dy = 1.$$

A good example of a suitable sequence are the *radial Gaussian distributions*

$$f_n(x,y) = \frac{e^{-(x^2+y^2)/n}}{n\,\pi}\,, \tag{15.50}$$

which relies on the fact that

$$\iint_{\mathbb{R}^2} e^{-(x^2+y^2)/n}\,dx\,dy = n\,\pi,$$

established in Exercise ∎. Note in Figure 15.7 how the Gaussian profiles become more and more concentrated at the origin, while maintaining a unit volume underneath their graphs.

Alternatively, one can assign the delta function a dual interpretation as a linear functional on the vector space of continuous scalar-valued functions. We formally prescribe the delta function by the integral formula

$$\langle\, \delta_{(\xi,\eta)}\,;f\,\rangle = \iint_\Omega \delta_{(\xi,\eta)}(x,y) f(x,y)\,dx\,dy = \begin{cases} f(\xi,\eta), & (\xi,\eta) \in \Omega, \\ 0, & (\xi,\eta)\ /\ \overline{\Omega}, \end{cases} \tag{15.51}$$

for any continuous function $f(x,y)$ and any domain $\Omega \subset \mathbb{R}^2$. As in the one-dimensional situation, we will avoid defining the integral when the delta function is concentrated at a boundary point $(\xi,\eta) \in \partial\Omega$ of the domain of integration.

Since double integrals can be evaluated as repeated one-dimensional integrals, (A.48), we can conveniently view

$$\delta_{(\xi,\eta)}(x,y) = \delta_\xi(x)\,\delta_\eta(y) = \delta(x-\xi)\,\delta(y-\eta) \tag{15.52}$$

as the product of a pair of one-dimensional delta functions. Indeed, if the domain

$$\Omega = R = \big\{\, a < x < b,\ c < y < d \,\big\}$$

is a rectangle, then

$$\begin{aligned} \iint_R \delta_{(\xi,\eta)}(x,y) f(x,y)\,dx\,dy &= \int_a^b \int_a^b \delta(x-\xi)\,\delta(y-\eta)\,f(x,y)\,dy\,dx \\ &= \int_a^b \delta(x-\xi)\,f(x,\eta)\,dx = f(\xi,\eta), \end{aligned}$$

provided $a < \xi < b$ and $c < \eta < d$, i.e., $(\xi, \eta) \in R$; otherwise, for $(\xi, \eta) \not\in \overline{R}$, the integral is 0, in accordance with (15.51).

To find the Green's function, we must solve the equilibrium equation subject to a concentrated unit delta force at a prescribed point $\boldsymbol{\xi} = (\xi, \eta) \in \Omega$ in the domain. In the case of Poisson's equation, this takes the form

$$-\Delta u = \delta_{\boldsymbol{\xi}}, \qquad \text{or} \qquad -\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial y} = \delta(x-\xi)\,\delta(y-\eta), \qquad (x,y) \in \Omega, \tag{15.53}$$

along with homogeneous boundary conditions, either Dirichlet or mixed. (The nonuniqueness of solutions to the pure Neumann boundary value problem precludes the existence of a Green's function.) The resulting solution to the particular Poisson boundary value problem is denoted

$$G(\mathbf{x}; \boldsymbol{\xi}) = G(x, y; \xi, \eta), \tag{15.54}$$

and called the *Green's function* associated with the given boundary value problem. For each fixed value of $\boldsymbol{\xi} = (\xi, \eta)$, the function (15.54) measures the effect, at position $\mathbf{x} = (x, y)$ of a concentrated force applied at position $\boldsymbol{\xi} = (\xi, \eta)$.

Once we know the Green's function, the solution to the general Poisson boundary value problem

$$-\Delta u = f \qquad \text{in} \qquad \Omega, \qquad\qquad u = 0 \qquad \text{on} \qquad \partial\Omega \tag{15.55}$$

is reconstructed by using a superposition principle. We regard the forcing function

$$f(x,y) = \iint_\Omega \delta(x-\xi)\,\delta(y-\eta) f(\xi,\eta)\,d\xi\,d\eta$$

as a linear combination of delta impulses, whose strength at each point equals the value of f. Linearity implies that the solution to the boundary value problem is the same combination of Green's function responses to each of the constituent impulses. The net result is the fundamental *superposition formula*

$$u(x,y) = \iint_\Omega G(x,y;\xi,\eta)\, f(\xi,\eta)\,d\xi\,d\eta \tag{15.56}$$

for the solution to the general inhomogeneous boundary value problem.

As in the one-dimensional situation, self-adjointness of the boundary value problem will imply that the Green's function is symmetric under interchange of its arguments:

$$G(\xi,\eta;x,y) = G(x,y;\xi,\eta). \tag{15.57}$$

(The proof of this fact is not hard, but will not be given in detail here.) Symmetry has the following interesting physical interpretation: Let $(x,y), (\xi,\eta) \in \Omega$ be any pair of points in the domain. If we apply a unit impulse at the first point, and measure the effect at the second, the result is exactly the same as if we apply the impulse at the second point, and measure the effect at the first! The reader may wish to reflect on whether this is physically plausible: if we push a membrane, of arbitrary shape, with unit force concentrated at $\boldsymbol{\xi}$ and measure the deflection at position $\mathbf{x}$ the result is the same as if we apply our force at

position $\mathbf{x}$ and measure the deflection at $\boldsymbol{\xi}$. (The deflections at other points in the domain will typically bear very little connection with each other.) Similarly, in electrostatics, the solution $u(x, y)$ is interpreted as the electrostatic potential for a system in equilibrium. A delta function corresponds to a point charge, e.g., an electron. The symmetry property says that the electrostatic potential at $\mathbf{x}$ due to a point charge placed at position $\boldsymbol{\xi}$ is the same as the potential at $\boldsymbol{\xi}$ due to a point charge at $\mathbf{x}$.

Unfortunately, most Green's functions, with a few exceptions, cannot be written down in closed form. However, their fundamental properties can be based on the following construction. As usual, the general solution to an inhomogeneous linear equation is a sum

$$u(x, y) = u_\star(x, y) + z(x, y) \tag{15.58}$$

of a particular solution $u_\star$ and the general solution z to the associated homogeneous equation, namely

$$-\Delta z = 0.$$

Thus, $z(x, y)$ is an arbitrary harmonic function. We shall assume that the particular solution $u_\star(x, y)$ is due to the effect of the unit impulse, irrespective of any imposed boundary conditions. Once we have determined $u_\star$, we shall use the freedom inherent in the harmonic constituent $z(x, y)$ to ensure that the sum (15.58) satisfies the required boundary conditions.

In order to find a particular solution $u_\star$, we may appeal to physical intuition. First, since the delta function is concentrated at the point $\boldsymbol{\xi}$, the solution $u_\star$ must solve the homogeneous Laplace equation $\Delta u_\star = 0$ except at the point $\mathbf{x} = \boldsymbol{\xi}$, where we expect it to have some sort of discontinuity. Second, since the Poisson equation is modeling a homogeneous, uniform medium (membrane, plate, etc.), in the absence of boundary conditions, the effect of a unit impulse should only depend upon on the distance away from the source of the impulse. Therefore, we expect that the desired particular solution $u_\star = u_\star(r)$ will depend only on the radial variable

$$r = \| \mathbf{x} - \boldsymbol{\xi} \| = \sqrt{(x - \xi)^2 + (y - \eta)^2}\,.$$

According to (15.34), the only radially symmetric solutions to the Laplace equation are

$$u(r) = a + b \log r, \tag{15.59}$$

where a and b are constants. The constant term a is smooth and harmonic everywhere, and so cannot contribute to a delta function singularity. Therefore, our only chance to produce a particular solution with such a singularity at the point $\boldsymbol{\xi}$ is if we take a multiple of the logarithmic potential:

$$u_\star = b \, \log r.$$

By construction, this function solves the Laplace equation for $r \neq 0$, i.e., for $\mathbf{x} \neq \boldsymbol{\xi}$, and has a singularity at $r = 0$. But we need to see whether, for some choice of b, it satisfies the Poisson equation

$$-\Delta(b \, \log r) = -b \, \Delta \log r = \delta(\mathbf{x} - \boldsymbol{\xi}) \tag{15.60}$$

for some choice of the constant b? There are two possible approaches to resolving this problem, corresponding to the two interpretations of the delta function. One way would be to approximate the delta function on the right hand side of (15.60) by a limit of highly concentrated unit sources, e.g., the Gaussian distributions $g_n(x,y)$ as given in (15.50). We then solve the Poisson equation $-\Delta u_n = g_n$, and prove that, in the limit, $\lim_{n\to\infty} u_n(x,y) = b\log r$ for a suitable b. The details are worked out in Exercise ■.

Alternatively, we interpret both sides of (15.60) as defining linear functionals on the space of smooth scalar functions $f(x,y)$ by taking the L^2 inner product

$$\langle\, -b\,\Delta\log r\,;f\,\rangle = \langle\,\delta_{\boldsymbol{\xi}}\,;f\,\rangle = f(\xi,\eta),$$

where we use the defining property (15.51) of the delta function to evaluate the right hand side. As for the left hand side, since

$$\Delta\log r = 0 \qquad \text{for all} \qquad r>0,$$

we only need integrate

$$\langle\,\Delta\log r\,;f\,\rangle = \iint_{D_\varepsilon}(\Delta\log r) f(x,y)\,dx\,dy = f(\xi,\eta)\iint_{D_\varepsilon}\Delta\log r\,dx\,dy,$$

over a small disk $D_\varepsilon = \big\{\,0\le r<\varepsilon\,\big\} = \big\{\,\|\,\mathbf{x}-\boldsymbol{\xi}\,\|<\varepsilon\,\big\}$ centered at the singularity $\mathbf{x}=\boldsymbol{\xi}$. Applying the divergence form (A.57) of Green's Theorem to evaluate the latter integral, we find

$$\begin{aligned}\iint_{D_\varepsilon}\Delta\log r\,dx\,dy &= \iint_{D_\varepsilon}\nabla\cdot\nabla\log r\,dx\,dy\\ &= \oint_{C_\varepsilon}\frac{\partial\log r}{\partial\mathbf{n}}\,ds = \oint_{C_\varepsilon}\frac{1}{r}\,ds = \int_{-\pi}^{\pi}d\theta = 2\pi,\end{aligned}$$

for all $\varepsilon>0$. Substituting this result back into (15.60), we find

$$\langle\,\Delta\log r\,;f\,\rangle = 2\pi\,f(\xi,\eta), \qquad \text{and hence} \qquad \Delta\log r = 2\pi\delta_{\boldsymbol{\xi}}. \tag{15.61}$$

Therefore, the value $b=-1/(2\pi)$ leads to our desired formula (15.60), and proves that the logarithmic potential function

$$u_\star(x,y) = -\frac{1}{2\pi}\,\log r = -\frac{1}{2\pi}\log\|\,\mathbf{x}-\boldsymbol{\xi}\,\| = -\frac{1}{4\pi}\log\big[\,(x-\xi)^2+(y-\eta)^2\,\big] \tag{15.62}$$

is a particular solution to the Poisson equation (15.53) with a unit impulse force.

The logarithmic potential function (15.62) represents the gravitational force field in empty space due to a unit point mass at position $\boldsymbol{\xi}$, or, equivalently, the electrostatic potential due to a point charge. It should be emphasized that this is in a two-dimensional universe; the three-dimensional versions in our physical universe is proportional to $1/r$ — even when restricted to a two-dimensional plane. See Section 18.1 for further details.

The gravitational or electrostatic potential due to a mass, e.g., a plate, in the shape of a domain $\Omega\subset\mathbb{R}^2$ is given by superimposing delta function sources at each point, whose strength is the density of the material. The result is the potential

$$u(x,y) = -\frac{1}{4\pi}\iint_\Omega \rho(\xi,\eta)\,\log\big[\,(x-\xi)^2+(y-\eta)^2\,\big]\,d\xi\,d\eta, \tag{15.63}$$

in which $\rho(\xi, \eta)$ is the density of the body at position (ξ, η). For example, the gravitational force due to a disk of radius 1, so $D = \{ x^2 + y^2 \leq 1 \}$, and unit density is

$$u(x,y) = -\frac{1}{4\pi} \iint_D \log\left[(x-\xi)^2 + (y-\eta)^2 \right] d\xi \, d\eta,$$

which evaluates to ∎

Returning to our boundary value problem, the general solution to (15.53) is, therefore,

$$u(x,y) = -\frac{1}{2\pi} \log \| \mathbf{x} - \boldsymbol{\xi} \| + z(x,y), \tag{15.64}$$

where $z(x,y)$ is an arbitrary harmonic function. To construct the Green's function for a given domain $\Omega \subset \mathbb{R}^2$ with prescribed homogeneous boundary conditions on $\partial\Omega$, we need to choose the harmonic function $z(x,y)$ so that $u(x,y)$, as given in (15.64), satisfies the boundary conditions. Let us state this result in the case of Dirichlet boundary conditions.

Theorem 15.9. *The Green's function for the Dirichlet boundary value problem for the Poisson equation $-\Delta u = f$ on Ω, and $u = 0$ on $\partial\Omega$ has the form*

$$G(x,y;\xi,\eta) = -\frac{1}{4\pi} \log\left[(x-\xi)^2 + (y-\eta)^2 \right] + z(x,y) \tag{15.65}$$

where

$$z(x,y) = \frac{1}{4\pi} \log\left[(x-\xi)^2 + (y-\eta)^2 \right], \qquad (x,y) \in \partial\Omega,$$

is the harmonic function that has the same boundary values as the logarithmic potential function.

The Method of Images

The preceding analysis uncovers the basic form of the Green's function, but we are still left with the determination of the harmonic component required to match the logarithmic potential boundary values. There are three principal techniques used to determine explicit formulas. The first is an adaptation of the method of separation of variables, and leads to infinite series expressions, similar to those of the fundamental solution for the heat equation derived in Chapter 14. We will not dwell on this technique here, although a couple of the exercises ask the reader to fill in the details. The second method is called the "method of images" and will be described in this section. The most powerful method is based on the theory of conformal mappings, and will be presented in Section 16.3 in the subsequent chapter. While the first two methods only apply to a fairly limited class of domains, they do adapt straightforwardly to higher dimensional problems, as well as certain other types of elliptic partial differential equations, whereas the method of conformal mapping is, unfortunately, only applicable to two-dimensional problems involving the Laplace and Poisson equations.

We already know that the singular part of the Green's function for the two-dimensional Poisson equation is provided by a logarithmic potential. The problem, then, is to construct the harmonic part, called $z(x,y)$ in (15.65), so that the sum has the correct homogeneous

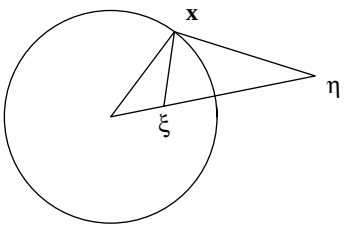


Figure 15.8. Method of Images for the Unit Disk.

boundary values, or, equivalently, that $z(x, y)$ has the same boundary values as the logarithmic potential.

In certain cases, $z(x, y)$ can be thought of as the potential induced by one or more electric charges (or, equivalently, gravitational point masses) that are located outside the domain Ω, arranged in such a manner that their electrostatic potential coincides with the logarithmic potential on the boundary of the domain. The goal, then, is to place the image charges in the proper positions.

We only consider the simplest case of a single image, located at a position $\boldsymbol{\eta} \notin \Omega$. We slightly generalize the logarithmic potential (15.62) by allowing an arbitrary scalar multiple and also an extra constant:

$$z(x, y) = a, \log \| \mathbf{x} - \boldsymbol{\eta} \| + b.$$

This function is harmonic inside Ω since the logarithmic potential is harmonic everywhere except at the singularity $\boldsymbol{\eta}$, which is assumed to lies outside the domain. For the Dirichlet boundary value problem, then, for each point $\boldsymbol{\xi} \in \Omega$ we require an image point $\boldsymbol{\eta} \notin \Omega$ and constants $a, b \in \mathbb{R}$, such that

$$\log \| \mathbf{x} - \boldsymbol{\xi} \| = a \log \| \mathbf{x} - \boldsymbol{\eta} \| + b \qquad \text{for all} \qquad \mathbf{x} \in \partial\Omega. \tag{15.66}$$

To simplify the formulas, we have omitted the $1/(2\pi)$ factor, which can easily be reinstated at the end of the analysis.

In order to make further progress, we make some simplifying assumptions. First, we assume that $a = 1$, and so (15.66) can be rewritten as

$$\| \mathbf{x} - \boldsymbol{\xi} \| = \lambda \| \mathbf{x} - \boldsymbol{\eta} \|,$$

where $\lambda = \log b$. We now use a geometrical construction based upon similar triangles. We choose $\boldsymbol{\eta} = a\,\boldsymbol{\xi}$ to be a point lying on the ray through $\boldsymbol{\xi}$, chosen so that the triangle with vertices $\mathbf{0}, \mathbf{x}, \boldsymbol{\eta}$ is similar to the triangle with vertices $\mathbf{0}, \boldsymbol{\xi}, \mathbf{x}$, noting that they have the same angle at the common vertex $\mathbf{0}$, as illustrated in Figure 15.8. Similarity requires that the triangles' sides be in a common ratio, and so

$$\frac{\| \boldsymbol{\xi} \|}{\| \mathbf{x} \|} = \frac{\| \mathbf{x} \|}{\| \boldsymbol{\eta} \|} = \frac{\| \mathbf{x} - \boldsymbol{\xi} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|} = \lambda. \tag{15.67}$$

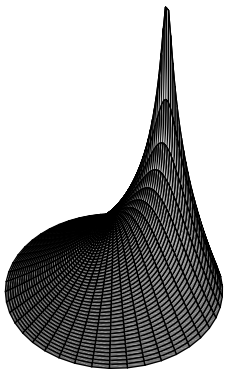


Figure 15.9. Green's Function for the Unit Disk.

Thus, if we choose

$$\| \boldsymbol{\xi} \| = \frac{1}{\| \boldsymbol{\eta} \|}, \qquad \text{then} \qquad \| \mathbf{x} \|^2 = \| \boldsymbol{\xi} \| \, \| \boldsymbol{\eta} \| = 1,$$

and hence $\mathbf{x}$ lies on the boundary of the unit disk. Given $\boldsymbol{\xi}$ inside the disk, its image point $\boldsymbol{\eta}$ will be at the reciprocal radius, with

$$\boldsymbol{\eta} = \frac{\boldsymbol{\xi}}{\| \boldsymbol{\xi} \|^2} \, . \tag{15.68}$$

The map taking the point $\boldsymbol{\xi}$ to the point $\boldsymbol{\eta}$ defined by (15.68) is known as *inversion* with respect to the unit circle. The final equation in (15.67) implies that

$$\| \mathbf{x} - \boldsymbol{\xi} \| = \| \boldsymbol{\xi} \| \, \| \mathbf{x} - \boldsymbol{\eta} \| \; = \; \frac{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|}{\| \boldsymbol{\xi} \|} \, .$$

Consequently, the functions

$$\frac{1}{2\,\pi} \, \log \frac{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|}{\| \boldsymbol{\xi} \|} = \frac{1}{2\,\pi} \, \log \| \mathbf{x} - \boldsymbol{\xi} \|, \tag{15.69}$$

has the same boundary values on the unit circle $\| \mathbf{x} \| = 1$. Consequently, their difference

$$G(\mathbf{x}; \boldsymbol{\xi}) = - \frac{1}{2\,\pi} \log \| \mathbf{x} - \boldsymbol{\xi} \| + \frac{1}{2\,\pi} \log \frac{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|}{\| \boldsymbol{\xi} \|} = \frac{1}{2\,\pi} \log \frac{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|}{\| \boldsymbol{\xi} \| \, \| \boldsymbol{\xi} - \mathbf{x} \|}$$

has the required properties of the Green's function for the unit disk. In terms of polar coordinates

$$\mathbf{x} = (r \cos \theta, r \sin \theta), \qquad \boldsymbol{\xi} = (\rho \cos \varphi, \rho \sin \varphi),$$

the Law of Cosines leads to the explicit formula

$$G(r, \theta; \rho, \varphi) = \frac{1}{4\,\pi} \, \log \left(\frac{1 + r^2 \rho^2 - 2\, r\, \rho \cos(\theta - \varphi)}{r^2 + \rho^2 - 2\, r\, \rho \cos(\theta - \varphi)} \right). \tag{15.70}$$

In Figure 15.9 we sketch the Green's function corresponding to a unit impulse being applied at a point half way between the center and the edge of the disk.

Applying the general superposition formula (15.56), we arrive at a general formula for the solution to the Dirichlet boundary value problem for the Poisson equation in the unit disk.

Theorem 15.10. *The solution $u(r,\theta)$ to the homogeneous Dirichlet boundary value problem*

$$-\Delta u = f, \qquad r = \| \mathbf{x} \| < 1, \qquad\qquad u = 0, \qquad r = 1$$

is, in polar coordinates,

$$u(r,\theta) = \frac{1}{4\pi} \int_0^{2\pi} \int_0^1 f(\rho,\varphi)\, \log\left(\frac{1 + r^2\rho^2 - 2\,r\,\rho\cos(\theta-\varphi)}{r^2 + \rho^2 - 2\,r\,\rho\cos(\theta-\varphi)} \right) \rho\, d\rho\, d\varphi.$$

The Green's function can also be used to solve the inhomogeneous boundary value problem.

Theorem 15.11. *Let $G(\mathbf{x};\boldsymbol{\xi})$ denote the Green's function for the homogeneous Dirichlet boundary value problem for the Poisson equation on a domain $\Omega \subset \mathbb{R}^2$. Then the solution to the inhomogeneous Dirichlet problem*

$$-\Delta u = f, \qquad \mathbf{x} \in \Omega, \qquad\qquad u = h, \qquad \mathbf{x} \in \partial\Omega, \tag{15.71}$$

is given by

$$u(\mathbf{x}) = v(\mathbf{x}) + \psi(\mathbf{x}) = \iint_\Omega G(\mathbf{x};\boldsymbol{\xi})\, f(\boldsymbol{\xi})\, d\xi\, d\eta - \oint_{\partial\Omega} \frac{\partial G(\mathbf{x};\boldsymbol{\xi})}{\partial \mathbf{n}}\, h(\boldsymbol{\xi})\, ds. \tag{15.72}$$

Proof: Let $\psi(\mathbf{x})$ be any function such that

$$\psi = h, \qquad \text{for} \qquad \mathbf{x} \in \partial\Omega.$$

Set $v = u - \psi$, so that v satisfies the homogeneous boundary value problem

$$-\Delta v = f + \Delta\psi, \qquad \mathbf{x} \in \Omega, \qquad\qquad v = 0, \qquad \mathbf{x} \in \partial\Omega.$$

We can therefore express

$$v(\mathbf{x}) = \iint_\Omega G(\mathbf{x};\boldsymbol{\xi}) \big[\, f(\boldsymbol{\xi}) + \Delta\psi(\boldsymbol{\xi}) \,\big]\, d\xi\, d\eta.$$

The second integral can be simplified using the integration by parts formula (Greenudeltauv■):

$$\iint_\Omega G(\mathbf{x};\boldsymbol{\xi})\, \Delta\psi(\boldsymbol{\xi})\, d\xi\, d\eta = \iint_\Omega \Delta G(\mathbf{x};\boldsymbol{\xi})\, \psi(\boldsymbol{\xi})\, d\xi\, d\eta + \\ + \oint_{\partial\Omega} \left(G(\mathbf{x};\boldsymbol{\xi})\, \frac{\partial\psi(\boldsymbol{\xi})}{\partial \mathbf{n}} - \frac{\partial G(\mathbf{x};\boldsymbol{\xi})}{\partial \mathbf{n}}\, \psi(\boldsymbol{\xi}) \right) ds.$$

Since the Green's function solves $-\Delta G = \delta_{\boldsymbol{\xi}}$, the first term reproduces $-\psi(\mathbf{x})$. Moreover, $G = 0$ and $\psi = h$ on $\partial\Omega$, and so this reduces to (15.72). *Q.E.D.*

For example, applying (15.72) to the Green's function (15.70) for the unit disk recovers the Poisson integral formula (15.44).

15.4. Adjoints and Minimum Principles.

We shall now explain how the Laplace and Poisson equations fit into our universal self-adjoint equilibrium framework. The one-dimensional version of the Poisson equation,

$$-\frac{d^2u}{dx^2} = f,$$

is the equilibrium equation for a uniform elastic bar. In Section 11.3, we wrote the underlying boundary value problems in self-adjoint form $D^* \circ D[u] = f$ based on the derivative operator $Du = u'$ and its adjoint $D^* = -D$ with respect to the standard L^2 inner product.

For the two-dimensional Poisson equation

$$-\Delta[u] = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x,y)$$

the role of the one-dimensional derivative D will be played by the *gradient* operator

$$\nabla u = \operatorname{grad} u = \begin{pmatrix} u_x \\ u_y \end{pmatrix}.$$

The gradient ∇ maps a scalar-valued function $u(x,y)$ to the vector-valued function consisting of its two first order partial derivatives. Thus, its domain is the vector space $U = \mathrm{C}^1(\Omega, \mathbb{R})$ consisting of all continuously differentiable functions $u(x,y)$ defined for $(x,y) \in \Omega$. The target space $V = \mathrm{C}^0(\Omega, \mathbb{R}^2)$ consists of all continuous vector-valued functions $\mathbf{v}(x,y) = (\, v_1(x,y), v_2(x,y)\,)^T$, known as *vector fields*. (By way of analogy, scalar-valued functions are sometimes referred to as *scalar fields*.) The gradient defines a *linear map*

$$\nabla : U \longrightarrow V$$

from scalar fields to vector fields. Indeed, if $u_1, u_2 \in U$ are any two scalar functions and $c_1, c_2 \in \mathbb{R}$ any constants, then

$$\nabla(c_1\, u_1 + c_2\, u_2) = c_1\, \nabla u_1 + c_2\, \nabla u_2,$$

which is the requirement for linearity of Definition 7.1.

In accordance with the general Definition 7.52, the adjoint of the gradient must go in the reverse direction,

$$\nabla^* : V \longrightarrow U,$$

mapping vector fields $\mathbf{v}(x,y)$ to scalar functions $z(x,y) = \nabla^* \mathbf{v}$. The defining equation for the adjoint

$$\langle\!\langle\, \nabla u \,;\, \mathbf{v} \,\rangle\!\rangle = \langle\, u \,;\, \nabla^* \mathbf{v} \,\rangle \tag{15.73}$$

requires inner products on the two vector spaces. The simplest inner product between real-valued scalar functions $u(x,y), v(x,y)$ defined on a domain $\Omega \subset \mathbb{R}^2$ is given by the double integral

$$\langle\, u \,;\, v \,\rangle = \iint_\Omega u(x,y)\, v(x,y)\, dx\, dy. \tag{15.74}$$

As in the one-dimensional case (3.11), this is often referred to as the L^2 *inner product* between scalar fields, with associated norm

$$\| u \|^2 = \langle u \,;\, u \rangle = \iint_\Omega u(x,y)^2 \, dx \, dy.$$

More generally, the L^2 inner product between vector-valued functions (vector fields) defined on Ω is obtained by integrating their usual dot product:

$$\langle\!\langle \mathbf{v} \,;\, \mathbf{w} \rangle\!\rangle = \iint_\Omega \mathbf{v}(x,y) \cdot \mathbf{w}(x,y) \, dx \, dy = \iint_\Omega \big[\, v_1(x,y) \, w_1(x,y) + v_2(x,y) \, w_2(x,y) \, \big] \, dx \, dy. \tag{15.75}$$

These form the two most basic inner products on the spaces of scalar and vector fields, and are the ones required to place the Laplace and Poisson equations in self-adjoint form.

The adjoint identity (15.73) is supposed to hold for all appropriate scalar fields u and vector fields $\mathbf{v}$. For the L^2 inner products (15.74), (15.75), the two sides of the identity read

$$\langle\!\langle \nabla u \,;\, \mathbf{v} \rangle\!\rangle = \iint_\Omega \nabla u \cdot \mathbf{v} \, dx \, dy = \iint_\Omega v_1 \, \frac{\partial u}{\partial x} + v_2 \, \frac{\partial u}{\partial y} \, dx \, dy,$$

$$\langle \, u \,;\, \nabla^* \mathbf{v} \, \rangle = \iint_\Omega u \nabla^* \mathbf{v} \, dx \, dy.$$

Thus, to equate these two double integrals, we need to remove the derivatives from the scalar field u. As in the one-dimensional computation (11.74), the secret is integration by parts.

As the student may recall, integration by parts is an immediate consequence of the Fundamental Theorem of Calculus when applied to Leibniz's rule for the derivative of the product of two functions. Now, according to Appendix A, Green's Theorem A.25 plays the role of the fundamental theorem in two-dimensional calculus. We will find the divergence form

$$\iint_\Omega \nabla \cdot \mathbf{v} \, dx \, dy = \oint_{\partial\Omega} \mathbf{v} \cdot \mathbf{n} \, ds. \tag{15.76}$$

the more convenient for the present purposes. In analogy with the one-dimensional argument, we now replace the vector field $\mathbf{v}$ by the product $u \, \mathbf{v}$ of a scalar field u and a vector field $\mathbf{v}$. An elementary computation proves that

$$\nabla \cdot (u \, \mathbf{v}) = u \, \nabla \cdot \mathbf{v} + \nabla u \cdot \mathbf{v}. \tag{15.77}$$

As a result, we deduce what is usually known as *Green's formula*

$$\iint_\Omega \big[\, u \, \nabla \cdot \mathbf{v} + \nabla u \cdot \mathbf{v} \, \big] \, dx \, dy = \oint_{\partial\Omega} u \, (\mathbf{v} \cdot \mathbf{n}) \, ds, \tag{15.78}$$

which is valid for arbitrary bounded domains Ω, and arbitrary scalar and vector fields defined thereon. Rearranging the terms in this integral identity produces the required integration by parts formula for double integrals:

$$\iint_\Omega \nabla u \cdot \mathbf{v} \, dx \, dy = \oint_{\partial\Omega} u \, (\mathbf{v} \cdot \mathbf{n}) \, ds - \iint_\Omega u \, \nabla \cdot \mathbf{v} \, dx \, dy. \tag{15.79}$$

The first term on the right hand side of this identity is a boundary term, just like the first terms on the right hand side of the one-dimensional integration by parts formula (11.77). Moreover, the derivative operation has moved from a gradient on the scalar field to a divergence on the vector field in the double integral on the right — even the minus sign is there!

Now, The left hand side in the integration by parts formula (15.79) is the same as the left hand side of (15.73). If the boundary integral vanishes,

$$\oint_{\partial\Omega} u\,\mathbf{v}\cdot\mathbf{n}\,ds = 0, \tag{15.80}$$

then the right hand side of formula (15.79) also reduces to an L^2 inner product

$$-\iint_\Omega u\,\nabla\cdot\mathbf{v}\,dx\,dy = \iint_\Omega u\,(-\nabla\cdot\mathbf{v})\,dx\,dy = \langle\, u\,;-\nabla\cdot\mathbf{v}\,\rangle$$

between the scalar field u and minus the divergence of the vector field $\mathbf{v}$. Therefore, subject to the boundary constraint (15.80), the integration by parts formula reduces to the inner product identity

$$\langle\!\langle\, \nabla u\,;\mathbf{v}\,\rangle\!\rangle = \langle\, u\,;\,-\nabla\cdot\mathbf{v}\,\rangle. \tag{15.81}$$

Comparing (15.73), (15.81), we conclude that $\nabla^*\mathbf{v} = -\nabla\cdot\mathbf{v}$, and hence the adjoint of the gradient operator is minus the divergence, $\nabla^* = -\nabla\cdot$. In this manner, we are able to write the two-dimensional Poisson equation in the standard self-adjoint form

$$-\Delta u = \nabla^*\circ\nabla u = -\nabla\cdot(\nabla u) = f \tag{15.82}$$

subject to an appropriate system of boundary conditions that justify (15.81).

The vanishing of the boundary integral (15.80) will be ensured by the imposition of suitable homogeneous boundary conditions on the scalar field u and/or the vector field $\mathbf{v}$. Clearly the line integral will vanish if either $u = 0$ or $\mathbf{v}\cdot\mathbf{n} = 0$ at each point on the boundary. These lead immediately to the three principle types of boundary conditions. The first are the fixed or *Dirichlet boundary conditions*, which require

$$u = 0 \qquad \text{on} \qquad \partial\Omega. \tag{15.83}$$

Alternatively, we can require

$$\mathbf{v}\cdot\mathbf{n} = 0 \qquad \text{on} \qquad \partial\Omega, \tag{15.84}$$

which requires that $\mathbf{v}$ be tangent to $\partial\Omega$ at each point, and so there is no net flux across the (solid) boundary. If we identify $\mathbf{v} = \nabla u$, then the no flux boundary condition (15.84) translates into the *Neumann boundary conditions*

$$\frac{\partial u}{\partial\mathbf{n}} = \nabla u\cdot\mathbf{n} = 0 \qquad \text{on} \qquad \partial\Omega. \tag{15.85}$$

One can evidently also mix the boundary conditions, imposing Dirichlet conditions on part of the boundary, and Neumann on the complementary part:

$$u = 0 \quad \text{on} \quad D, \qquad \frac{\partial u}{\partial\mathbf{n}} = 0 \quad \text{on} \quad N, \qquad \text{where} \qquad \partial\Omega = D\cup N \tag{15.86}$$

is the disjoint union of the Dirichlet and Neumann parts.

To model inhomogeneous membranes, heat flow through inhomogeneous media, and similar physical equilibria, we replace the L^2 inner product between vector fields by the weighted version

$$\langle\!\langle\, \mathbf{v}\,;\mathbf{w}\,\rangle\!\rangle = \iint_\Omega \big[\, p(x,y)\, v_1(x,y)\, w_1(x,y) + q(x,y)\, v_2(x,y)\, w_2(x,y)\,\big]\, dx\, dy, \tag{15.87}$$

in which $p(x,y), q(x,y) > 0$ are strictly positive functions on the domain $(x,y) \in \Omega$. Retaining the usual L^2 inner product (15.74) between scalar fields, let us compute the weighted adjoint of the gradient operator. Using the same basic defining formula (15.73), we compute

$$\langle\!\langle\, \nabla u\,;\mathbf{v}\,\rangle\!\rangle = \iint_\Omega p\, v_1\, \frac{\partial u}{\partial x} + q\, v_2\, \frac{\partial u}{\partial y}\, dx\, dy.$$

We then apply the same integration by parts formula (15.79) to remove the derivatives from the scalar field u, leading to

$$\begin{aligned} &\iint_\Omega p\, v_1\, \frac{\partial u}{\partial x} + q\, v_2\, \frac{\partial u}{\partial y}\, dx\, dy \\ &\qquad = \oint_{\partial\Omega} \big[\, -u\, q\, v_2\, dx + u\, p\, v_1\, dy\,\big] - \iint_\Omega u\, \left[\, \frac{\partial (p\, v_1)}{\partial x} + \frac{\partial (q\, v_2)}{\partial y}\,\right] dx\, dy. \end{aligned} \tag{15.88}$$

Equating this to the right hand side $\langle\, u\,;\nabla^*\mathbf{v}\,\rangle$, we deduce that, provided the boundary integral vanishes, the weighted adjoint of the gradient operator with respect to (15.87) is given by

$$\nabla^*\mathbf{v} = -\,\frac{\partial (p\, v_1)}{\partial x} - \frac{\partial (q\, v_2)}{\partial y} = -\,p\,\frac{\partial v_1}{\partial x} - q\,\frac{\partial v_2}{\partial y} - v_1\,\frac{\partial p}{\partial x} - v_2\,\frac{\partial q}{\partial y}\,. \tag{15.89}$$

The boundary integral in (15.88) vanishes provided either $u = 0$ or $\mathbf{v} = 0$ on $\partial\Omega$. Therefore, the same homogeneous boundary conditions — Dirichlet, Neumann or mixed — are still applicable.

The corresponding self-adjoint boundary value problem takes the form

$$\nabla^* \circ \nabla u = -\,\frac{\partial}{\partial x}\left(\, p(x,y)\,\frac{\partial u}{\partial x}\,\right) - \frac{\partial}{\partial x}\left(\, q(x,y)\,\frac{\partial u}{\partial x}\,\right) = f(x,y), \qquad (x,y) \in \Omega, \tag{15.90}$$

along with either homogeneous or inhomogeneous boundary conditions of either Dirichlet, Neumann or mixed type. The weight functions p, q are prescribed by the physical inhomogeneities in the body.

Remark: In electrostatics, the gradient equation $\mathbf{v} = \nabla u$ relates the voltage drop to the electrostatic potential u, and is the continuous analog of the circuit formula (6.17) relating potentials to voltages. The continuous version of Kirchhoff's Voltage Law (6.19) that the net voltage drop around any loop is zero is the fact that any gradient vector has zero curl, $\nabla \wedge \mathbf{v} = \mathbf{0}$, i.e., the flow is irrotational. Ohm's law (6.22) has the form $\mathbf{y} = C\,\mathbf{v}$ where the vector field $\mathbf{y}$ represents the current, while $C = \mathrm{diag}\,(p(x,y), q(x,y))$ represents

th conductance of the medium; in the case of Laplace's equation, we a re assuming a uniform unit conductance. Finally, the equation $f = \nabla \cdot \mathbf{y} = \nabla^* \mathbf{v}$ relating current and external current sources forms the continuous analog of Kirchhoff's Current Law (6.25) — the transpose of the discrete incidence matrix translates into the adjoint of the gradient operator is the divergence. Thus, our discrete electro-mechanical analogy carries over, in the continuous realm, to a tripartite electro-mechanical-fluid analogy, with all three physical systems leading to the same very general mathematical structure.

Positive Definiteness and the Dirichlet Principle

In conclusion, as a result of the integration by parts calculation, we have successfully formulated the Poisson and Laplace equations (as well as their weighted counterparts) in self-adjoint form

$$-\Delta u = \nabla^* \circ \nabla u = f,$$

including either Dirichlet, Neumann, or mixed boundary conditions. A key benefit of the formulation of a system in self-adjoint form is, in the positive definite cases, the characterization of the solutions by a minimization principle.

According to Theorem 7.59, the self-adjoint operator $\nabla^* \circ \nabla$ is positive definite if and only if the kernel of the underlying gradient operator — restricted to the appropriate space of scalar fields — is trivial: $\ker \nabla = \{0\}$. The determination of the kernel of the gradient operator relies on the following elementary fact, which is the multi-variable version of the result that the only function with zero derivative is a constant.

Lemma 15.12. *If $u(x, y)$ is a C^1 function defined on a connected domain, then $\nabla u \equiv 0$ if and only if $u \equiv c$ is a constant.*

This result is a simple consequence of Theorem A.20; see Exercise ∎. Therefore, the only functions which could show up in $\ker \nabla$, and thus prevent positive definiteness, are the constants. The boundary conditions will tell us whether or not this occurs. The only constant function that satisfies either homogeneous Dirichlet or homogeneous mixed boundary conditions is the zero function, and thus, just as in the one-dimensional case, the boundary value problem for the Poisson equation with Dirichlet or mixed boundary conditions is positive definite. On the other hand, any constant function satisfies the homogeneous Neumann boundary conditions, and hence such boundary value problems are only positive semi-definite.

In the positive definite definite cases, when $\ker \nabla = \{0\}$ — as dictated by the boundary conditions — the equilibrium solution can be characterized by our basic minimization principle based on the general formula (7.71). For the Poisson equation, the resulting quadratic functional is the justly famous *Dirichlet principle*.

Theorem 15.13. *The solution $u(x, y)$ to the Poisson equation* (15.3) *subject to either homogeneous Dirichlet or mixed boundary conditions is characterized as the unique function that minimizes the* Dirichlet integral

$$\tfrac{1}{2} \| \nabla u \|^2 - \langle u ; f \rangle = \iint_\Omega \left(\tfrac{1}{2} u_x^2 + \tfrac{1}{2} u_y^2 - f u \right) dx\, dy \tag{15.91}$$

among all C^1 functions that satisfy the prescribed boundary conditions.

In physical applications the Dirichlet integral (15.91) represents the energy in the system. Hence, just as in discrete and one-dimensional mechanics, Nature chooses the equilibrium configuration so as to minimize the energy. The application of this minimum principle for numerical approximation to the solutions based on the finite element approach will form the subject of Section 15.5.

Remark: Theorem 15.13 only says that *if* a minimum is achieved, *then* it must satisfy the boundary value problem. It does not actually guarantee the existence of a minimizer, and hence a solution to the boundary value problem. Dirichlet originally thought this to be self-evident, but it was later realized that the proof of existence is a rather difficult analytical theorem. It took about 50 years from Dirichlet's statement of his principle until Hilbert supplied the first rigorous existence proof. In applications, it is certainly comforting to know that there is a solution to the boundary value problem. In this introductory treatment, we adopt a more pragmatic approach, concentrating on the computation of the solution — reassured, if necessary, by the theoreticians' efforts in establishing the existence of the solution.

The Dirichlet minimization principle (15.91) was derived under the assumption that the boundary conditions are homogeneous — either pure Dirichlet or mixed. As it turns out, the principle also applies to inhomogeneous Dirichlet boundary conditions as stated. However, if we have inhomogeneous Neumann conditions on part of the boundary, then we must include an additional boundary term in the minimizing functional. The general result can be stated as follows:

Theorem 15.14. *The solution $u(x,y)$ to the boundary value problem*

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad u = h \quad \text{on} \quad D, \qquad \frac{\partial u}{\partial \mathbf{n}} = k \quad \text{on} \quad N,$$

with $\partial\Omega = D \cup N$, and $D \neq \varnothing$, is characterized as the unique function that minimizes the modified Dirichlet integral

$$\iint_\Omega \left(\tfrac{1}{2} \| \nabla u \|^2 - f\, u \right) dx\, dy \; + \; \int_N u\, k\, ds \tag{15.92}$$

among all C^1 *functions that satisfy the prescribed boundary conditions.*

The inhomogeneous Dirichlet problem has $N = \varnothing$ and $D = \partial\Omega$, in which case the boundary integral does not appear. An outline of the proof of this result appears in the exercises.

While the Dirichlet and mixed boundary value problems are positive definite, any constant function satisfies the homogeneous Neumann boundary conditions, and so in this case $\ker \nabla$ consists of all constant functions. Therefore, just as in the one-dimensional bar, the Neumann boundary value problem is only positive semi-definite, and so we cannot construct a minimization principle. Indeed, when the system is only positive semi-definite, the solution is not unique: if $u(x,y)$ is a solution, so is $u(x,y) + c$ for any constant c.

As we know, positive definiteness is directly related to the stability of the physical system. The Dirichlet and mixed boundary value problems are stable, and can support

any imposed force. On the other hand, the pure Neumann boundary value problem is unstable, owing to the existence of a nontrivial kernel — the constant functions. Physically, the unstable mode represents a rigid translation of the entire membrane in the vertical direction. Indeed, the Neumann problem leaves the entire boundary of the membrane unattached to any support, and so the unforced membrane is free to move up or down without affecting its equilibrium status.

Furthermore, non-uniqueness and non-existence of solutions go hand in hand. As we learned in Section 11.3, the existence of a solution to a Neumann boundary value problem relies on the *Fredholm alternative*, suitably adapted to this multi-dimensional situation. A necessary condition for the existence of a solution is that the forcing function be orthogonal to the elements of the kernel of the underlying self-adjoint linear operator, which, in the present situation requires that f be orthogonal to the subspace consisting of all constant functions. In practical terms, we only need to check orthogonality with respect to a basis for the subspace, which in this situation consists of the constant function 1. The fact that, under such conditions, a solution actually exists is harder, and we refer to [**39**] for details of the existence part of the following result.

Theorem 15.15. *The Neumann boundary value problem*

$$-\Delta u = f, \qquad \text{in} \qquad \Omega \qquad\qquad \frac{\partial u}{\partial \mathbf{n}} = 0, \qquad \text{on} \qquad \partial\Omega, \tag{15.93}$$

has a solution $u(x,y)$ if and only if

$$\langle\, 1\,;f\,\rangle = \iint_\Omega f(x,y)\,dx\,dy = 0. \tag{15.94}$$

Moreover, the solution is not unique since any function of the form $u(x,y)+c$, where $c \in \mathbb{R}$ is an arbitrary constant, is also a solution.

Forcing functions $f(x,y)$ which do not satisfy the orthogonality constraint (15.94) will excite the translational instability, and no equilibrium configuration is possible. For example, if we force a free membrane, (15.94) requires that the net force in the vertical direction be zero; otherwise, the membrane will start moving and cannot be in an equilibrium.

15.5. Finite Elements.

As the reader has no doubt already guessed, explicit solutions to boundary value problems for the Laplace and Poisson equations are few and far between. In most cases, exact solution formulae are not available, or are so complicated as to be of scant utility. To proceed further, one is forced to design suitable numerical approximation schemes that can accurately evaluate the desired solution and thereby aid in the study of its behavior.

The most powerful class of numerical algorithms for solving general elliptic boundary value problems are the finite element methods. We have already learned, in Section 11.6, the key underlying idea. One begins with a minimization principle, prescribed by a quadratic functional defined on a suitable vector space of functions U that serves to incorporate the (homogeneous) boundary conditions. The desired solution is characterized as the

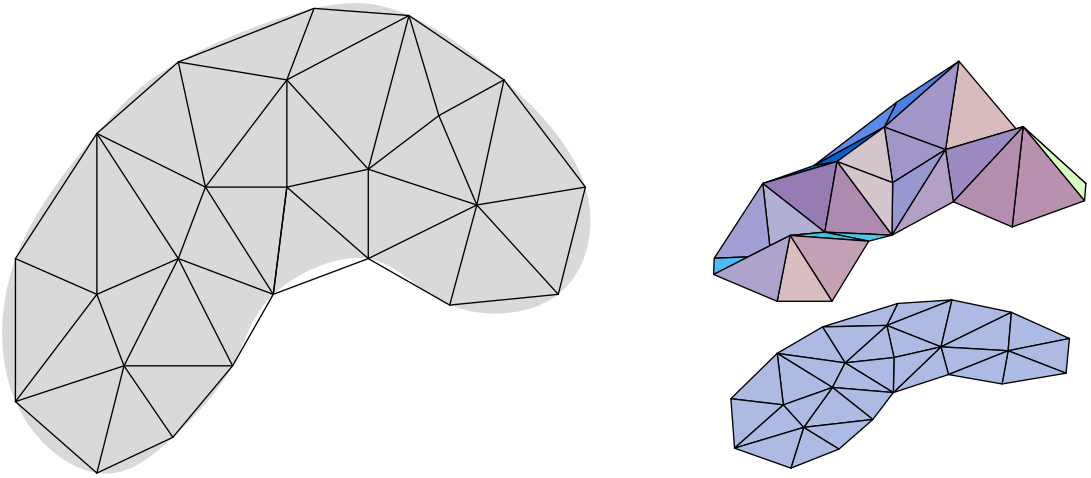


Figure 15.10. Triangulation of a Planar Domain and Piecewise Affine Function.

unique minimizer $u_\star \in U$. One then restricts the functional to a suitably chosen finite-dimensional subspace $W \subset U$, and seeks a minimizer $w_\star \in W$. Finite-dimensionality of W has the effect of reducing the infinite-dimensional minimization problem to a finite-dimensional problem, which can then be solved by numerical linear algebra. The resulting minimizer $w_\star$ will — provided the subspace W has been cleverly chosen — provide a good approximation to the true minimizer $u_\star$ on the entire domain. Here we concentrate on the practical design of the finite element procedure, and refer the reader to a more advanced text, e.g., [**FE**], for the analytical details and proofs of convergence. Most of the multi-dimensional complications are not in the underlying theory, but rather in the realms of data management and organizational details.

In this section, we first concentrate on applying these ideas to the two-dimensional Poisson equation. For specificity, we first treat the homogeneous Dirichlet boundary value problem

$$-\Delta u = f \quad \text{in} \quad \Omega \qquad\qquad u = 0 \quad \text{on} \quad \partial\Omega. \tag{15.95}$$

According to Theorem 15.13, the solution $u = u_\star$ is characterized as the unique minimizing function for the Dirichlet functional (15.91) among all smooth functions $u(x, y)$ that satisfy the prescribed boundary conditions. In the finite element approximation, we restrict the Dirichlet functional to a suitably chosen finite-dimensional subspace. As in the one-dimensional situation, the most convenient finite-dimensional subspaces consist of functions that may lack the requisite degree of smoothness that qualifies them as possible solutions to the partial differential equation. Nevertheless, they do provide good approximations to the actual solution. An important practical consideration is to use functions with small support, cf. Definition 13.5. The resulting finite element matrix will then be sparse and the solution to the linear system can be relatively rapidly calculate, usually by application of an iterative numerical scheme such as the Gauss–Seidel or SOR methods discussed in Chapter 10.

Finite Elements and Triangulation

For one-dimensional boundary value problems, the finite element construction rests on the introduction of a mesh $a = x_0 < x_1 < \cdots < x_n = b$ on the interval of definition. The mesh nodes x_k break the interval into a collection of small subintervals. In two-dimensional problems, a *mesh* consists of a finite number of points $\mathbf{x}_k = (x_k, y_k)$, $k = 1, \dots, m$, known as *nodes*, usually lying inside the domain $\Omega \subset \mathbb{R}^2$. As such, there is considerable freedom in the choice of mesh nodes, and completely uniform spacing is often not possible. We regard the nodes as forming the vertices of a *triangulation* of the domain Ω, consisting of a finite number of small triangles, which we denote by $T_1, \dots, T_N$. The nodes are split into two categories — *interior nodes* and *boundary nodes*, the latter lying on or close to the boundary of the domain. A curved boundary is approximated by the polygon through the boundary nodes formed by the sides of the triangles lying on the edge of the domain; see Figure 15.10 for a typical example. Thus, in computer implementations of the finite element method, the first ingredient is a routine that will automatically triangulate a specified domain in some reasonable manner; see below for details on what "reasonable" entails.

As in our one-dimensional finite element construction, the functions $w(x, y)$ in the finite-dimensional subspace W will be continuous and *piecewise affine*. "Piecewise affine" means that, on each triangle, the graph of w is flat, and so has the formula[†]

$$w(x, y) = \alpha^\nu + \beta^\nu x + \gamma^\nu y, \qquad \text{for} \qquad (x, y) \in T_\nu. \tag{15.96}$$

Continuity of w requires that its values on a common edge between two triangles must agree, and this will impose certain compatibility conditions on the coefficients $\alpha^\mu, \beta^\mu, \gamma^\mu$ and $\alpha^\nu, \beta^\nu, \gamma^\nu$ associated with adjacent pairs of triangles T_μ, T_ν. The graph of $z = w(x, y)$ forms a connected polyhedral surface whose triangular faces lie above the triangles in the domain; see Figure 15.10 for an illustration.

The next step is to choose a basis of the subspace of piecewise affine functions for the given triangulation. As in the one-dimensional version, the most convenient basis consists of *pyramid functions* $\varphi_k(x, y)$ which have the value 1 at a single node $\mathbf{x}_k$, and zero at all the other nodes; thus

$$\varphi_k(x_i, y_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases} \tag{15.97}$$

Note that φ_k will be nonzero only on those triangles which have the node $\mathbf{x}_k$ as one of their vertices, and hence the graph of φ_k looks like a pyramid of unit height sitting on a flat plane, as illustrated in Figure 15.11.

The pyramid functions $\varphi_k(x, y)$ corresponding to the *interior nodes* $\mathbf{x}_k$ satisfy the homogeneous Dirichlet boundary conditions on the boundary of the domain — or, more correctly, on the polygonal boundary of the triangulated domain, which is supposed to be a good approximation to the curved boundary of the original domain Ω. Thus, the finite-dimensional finite element subspace W will be spanned by the interior node pyramid

[†] Here and subsequently, the index ν is a superscript, not a power!

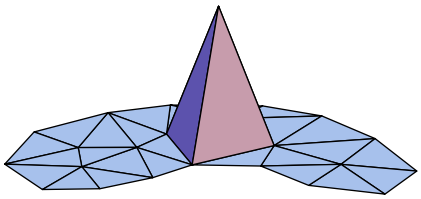


Figure 15.11. Finite Element Pyramid Function.

functions. A general element $w \in W$ is a linear combination thereof, so

$$w(x,y) = \sum_{k=1}^{n} c_k \, \varphi_k(x,y), \tag{15.98}$$

where the sum ranges over the n interior nodes of the triangulation. Owing to the original specification (15.97) of the pyramid functions, the coefficients

$$c_k = w(x_k, y_k) \;\approx\; u(x_k, y_k), \qquad\qquad k = 1, \dots, n, \tag{15.99}$$

are the *same* as the values of the finite element approximation $w(x,y)$ at the interior nodes. This immediately implies linear independence of the pyramid functions, since the only linear combination that vanishes at all nodes is the trivial one $c_1 = \cdots = c_n = 0$. Thus, the interior node pyramid functions $\varphi_k 1, \dots \varphi_n$ form a basis for finite element subspace W, which therefore has dimension equal to n, the number of interior nodes.

The explicit formulae for the finite element basis functions are not difficult to determine. On one of the triangles T_ν that has $\mathbf{x}_k$ as a vertex, $\varphi_k(x,y)$ will be the unique affine function (15.96) that takes the value 1 at the vertex $\mathbf{x}_k$ and 0 at the other two vertices $\mathbf{x}_l, \mathbf{x}_m$. Thus, we need a formula for an affine function or *element*

$$\omega_k^\nu(x,y) = \alpha_k^\nu + \beta_k^\nu \, x + \gamma_k^\nu \, y, \qquad\qquad (x,y) \in T_\nu, \tag{15.100}$$

that takes the prescribed values

$$\omega_k^\nu(x_i, y_i) = \omega_k^\nu(x_j, y_j) = 0, \qquad\qquad \omega_k^\nu(x_k, y_k) = 1,$$

at three specified points. These three conditions lead to the linear system

$$\begin{aligned} \omega_k^\nu(x_i, y_i) &= \alpha_k^\nu + \beta_k^\nu \, x_i + \gamma_k^\nu \, y_i \;\; = 0, \\ \omega_k^\nu(x_j, y_j) &= \alpha_k^\nu + \beta_k^\nu \, x_j + \gamma_k^\nu \, y_j \;\; = 0, \\ \omega_k^\nu(x_k, y_k) &= \alpha_k^\nu + \beta_k^\nu \, x_k + \gamma_k^\nu \, y_k = 1. \end{aligned} \tag{15.101}$$

The solution[†] produces the explicit formulae

$$\alpha_k^\nu = \frac{x_i \, y_j - x_j \, y_i}{\Delta_\nu}, \qquad\qquad \beta_k^\nu = \frac{y_i - y_j}{\Delta_\nu}, \qquad\qquad \gamma_k^\nu = \frac{x_j - x_i}{\Delta_\nu}, \tag{15.102}$$

[†] Cramer's Rule (Cramer3■) comes in handy here

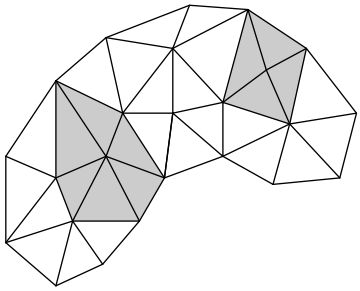


Figure 15.12. Vertex Polygons.

for the coefficients. The denominator

$$\Delta_\nu = \det \begin{pmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{pmatrix} = \pm 2 \,\text{area}\, T_\nu \tag{15.103}$$

is, up to sign, twice the area of the triangle T_ν; see Exercise ∎.

Example 15.16. Consider an isoceles right triangle T with vertices

$$\mathbf{x}_1 = (0,0), \qquad \mathbf{x}_2 = (1,0), \qquad \mathbf{x}_3 = (0,1).$$

Using equations (15.102), (15.103) (or solving the linear systems (15.101) directly), we immediately produce the three affine elements

$$\omega_1(x,y) = 1 - x - y, \qquad \omega_2(x,y) = x, \qquad \omega_3(x,y) = y. \tag{15.104}$$

They are defined so that ω_k equals 1 at the vertex $\mathbf{x}_k$ and is zero at the other two vertices.

The finite element pyramid function is then obtained by piecing together the individual affine elements:

$$\varphi_k(x,y) = \begin{cases} \omega_k^\nu(x,y), & \text{if } (x,y) \in T_\nu \text{ which has } \mathbf{x}_k \text{ as a vertex,} \\ 0, & \text{otherwise.} \end{cases} \tag{15.105}$$

Continuity of $\varphi_k(x,y)$ is ensured by the fact that the constituent affine elements have the same values at common vertices. The support of the finite element basis function (15.105) is the polygon

$$\operatorname{supp} \varphi_k = P_k = \bigcup_\nu T_\nu \tag{15.106}$$

consisting of all the triangles T_ν that have the node $\mathbf{x}_k$ as a vertex. Thus, $\varphi_k(x,y) = 0$ whenever $(x,y) \notin P_k$. We will call P_k the k^{th} *vertex polygon*. The node $\mathbf{x}_k$ lies on the interior of its vertex polygon P_k, while the vertices of P_k are all the adjacent vertices that are connected to $\mathbf{x}_k$ by an edge of the triangulation. In Figure 15.12 the shaded regions are two of the vertex polygons for the triangulation in Figure 15.10.

Example 15.17. The simplest, and most common triangulations are based on regular meshes. Suppose that the nodes lie on a square grid, and so are of the form $\mathbf{x}_{i,j} = (i\,h + a, j\,h + b)$ where $h > 0$ is the inter-node spacing, and (a,b) is an overall

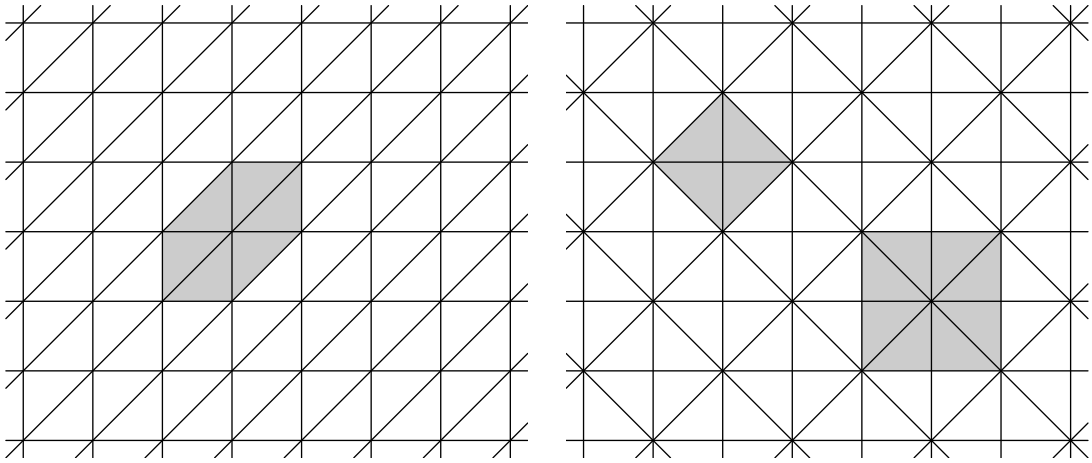


Figure 15.13. Square Mesh Triangulations.

offset. If we choose the triangles to all have the same orientation, as in the first picture in Figure 15.13, then the vertex polygons all have the same shape, consisting of 6 triangles of total area $3h^2$ — the shaded region. On the other hand, if we choose an alternating, perhaps more æsthetically pleasing triangulation as in the second picture, then there are two types of vertex polygons. The first, consisting of four triangles, has area $2h^2$, while the second, containing 8 triangles, has twice the area, $4h^2$. In practice, there are good reasons to prefer the former triangulation; see below.

In general, in order to ensure convergence of the finite element solution to the true minimizer, one should choose a triangulation with the following properties:

(a) The triangles are not too long and skinny. In other words, the sides should have comparable lengths. In particular, obtuse triangles should be avoided.

(b) The areas of nearby triangles T_ν should not vary too much.

(c) The areas of nearby vertex polygons P_k should also not vary too much.

For adaptive or variable meshes, one might very well have wide variations in area over the entire grid, with small triangles in regions of rapid change in the solution, and large ones in less interesting regions. But, overall, the sizes of the triangles and vertex polygons should not dramatically vary as one moves across the domain.

The Finite Element Equations

We now seek to approximate the solution to the homogeneous Dirichlet boundary value problem by restricting the Dirichlet functional to the finite element subspace W. Substituting the formula (15.98) for a general element of W into the quadratic Dirichlet

functional (15.91) and expanding, we find

$$\mathcal{P}[w] = \mathcal{P}\left[\sum_{i=1}^{n} c_i\,\varphi_i\right] = \iint_\Omega \left[\left(\sum_{i=1}^{n} c_i\,\nabla\varphi_i\right)^2 - f\left(\sum_{i=1}^{n} c_i\,\varphi_i\right)\right] dx\,dy$$
$$= \frac{1}{2}\sum_{i,j=1}^{n} k_{ij}\,c_i\,c_j - \sum_{i=1}^{n} b_i\,c_i = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c}.$$

Here $K = (k_{ij})$ is a symmetric $n \times n$ matrix, while $\mathbf{b} = (\,b_1, b_2, \ldots, b_n\,)^T$ is a vector with respective entries

$$\begin{aligned} k_{ij} &= \langle\,\nabla\varphi_i\,;\nabla\varphi_j\,\rangle = \iint_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,dx\,dy,\\ b_i &= \langle\,f\,;\varphi_i\,\rangle = \iint_\Omega f\,\varphi_i\,dx\,dy. \end{aligned} \tag{15.107}$$

Thus, to determine the finite element approximation, we need to minimize the quadratic function

$$P(\mathbf{c}) = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c} \tag{15.108}$$

over all possible choices of coefficients $\mathbf{c} = (\,c_1, c_2, \ldots, c_n\,)^T \in \mathbb{R}^n$, i.e., over all possible function values at the interior nodes.

Restricting to the finite element subspace has reduced us to a standard finite-dimensional quadratic minimization problem. First, the coefficient matrix $K > 0$ is positive definite due to the positive definiteness of the original functional; the proof in Section 11.6 is easily adapted to the present situation. Theorem 4.1 tells us that the minimizer is obtained by solving the associated linear system

$$K\,\mathbf{c} = \mathbf{b}. \tag{15.109}$$

The solution to (15.109) is effected by either Gaussian elimination or an iterative technique.

To find explicit formulae for the matrix coefficients k_{ij} in (15.107), we begin by noting that the gradient of the affine element (15.100) is equal to

$$\nabla\omega_k^\nu(x,y) = \mathbf{a}_k^\nu = \begin{pmatrix}\beta_k^\nu\\ \gamma_k^\nu\end{pmatrix} = \frac{1}{\Delta_\nu}\begin{pmatrix} y_i - y_j\\ x_j - x_i\end{pmatrix}, \qquad (x,y)\in T_\nu, \tag{15.110}$$

which is a constant vector within the triangle T_ν; outside it, $\nabla\omega_k^\nu = \mathbf{0}$ is zero. Therefore,

$$\nabla\varphi_k(x,y) = \begin{cases} \nabla\omega_k^\nu = \mathbf{a}_k^\nu, & \text{if } (x,y)\in T_\nu \text{ which has } \mathbf{x}_k \text{ as a vertex},\\ \mathbf{0}, & \text{otherwise.}\end{cases} \tag{15.111}$$

Actually, (15.111) is not quite right since if (x,y) lies on the boundary of a triangle T_ν, then the gradient does not exist. However, this technicality will not cause any difficulty in evaluating the ensuing integral. Thus, $\nabla\varphi_k$ reduces to a piecewise constant function on the triangulation.

We will approximate integrals over the domain Ω by integrals over the triangles, which assumes that the polygonal boundary of the triangulation is a reasonably close approximation to the true boundary $\partial\Omega$. In particular,

$$k_{ij} \;\approx\; \sum_{\nu} \iint_{T_\nu} \nabla\varphi_i \cdot \nabla\varphi_j \, dx\, dy \;\equiv\; \sum_{\nu} k_{ij}^{\nu}. \tag{15.112}$$

Now, according to (15.111), one or the other gradient in the integrand will vanish on the entire triangle T_ν *unless* both $\mathbf{x}_i$ and $\mathbf{x}_j$ are vertices. Therefore, the only terms contributing to the sum are those triangles T_ν that have both $\mathbf{x}_i$ and $\mathbf{x}_j$ as vertices. If $i \neq j$ there are only two such triangles, while if $i = j$ every triangle in the i^{th} vertex polygon P_i contributes. The individual summands are easily evaluated, since the gradients are constant on the triangles, and so

$$k_{ij}^{\nu} = \iint_{T_\nu} \mathbf{a}_i^{\nu} \cdot \mathbf{a}_j^{\nu} \; dx\, dy = \mathbf{a}_i^{\nu} \cdot \mathbf{a}_j^{\nu} \text{ area } T_\nu \,.$$

Let T_ν have vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$. Then, by (15.110), (15.111) and (15.103),

$$k_{ij}^{\nu} = \frac{(y_j - y_k)(y_k - y_i) + (x_k - x_j)(x_i - x_k)}{(\Delta_\nu)^2} \text{ area } T_\nu = -\,\frac{(\mathbf{x}_i - \mathbf{x}_k)\cdot(\mathbf{x}_j - \mathbf{x}_k)}{4 \text{ area } T_\nu}\,, \quad i \neq j,$$
$$k_{ii}^{\nu} = \frac{(y_j - y_k)^2 + (x_k - x_j)^2}{(\Delta_\nu)^2} \text{ area } T_\nu = \frac{\| \mathbf{x}_j - \mathbf{x}_k \|^2}{4 \text{ area } T_\nu}\,, \quad \text{where} \quad \text{area } T_\nu = \tfrac{1}{2}\,|\,\Delta_\nu\,|. \tag{15.113}$$

In this manner, each triangle T_ν is associated with a collection of 6 different coefficients, $k_{ij}^{\nu} = k_{ji}^{\nu}$, known as the *elemental stiffnesses* of T_ν. The indices i, j range over the three different vertices of the triangle T_ν. In practice, one assembles the elemental stiffnesses into a symmetric 3×3 matrix S_ν, known as the *elemental stiffness matrix* of the triangle, whose rows and columns are labeled by its vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$.

Interestingly, the elemental stiffnesses depend only on the *angles* of the triangle and not on its size. Thus, similar triangles have the *same* elemental stiffness matrix — provided their vertices are labeled in the same order. Indeed, if we denote the angle in T_ν at the vertex $\mathbf{x}_k$ by θ_k^{ν}, then, according to Exercise ■,

$$k_{ii}^{\nu} = \tfrac{1}{2}\big(\cot\theta_k^{\nu} + \cot\theta_j^{\nu}\big), \qquad \text{while} \qquad k_{ij}^{\nu} = k_{ji}^{\nu} = -\,\tfrac{1}{2}\cot\theta_k^{\nu}, \quad i \neq j, \tag{15.114}$$

depend only upon the cotangents of the angles, and hence the elemental stiffness matrix has the form

$$S_\nu = \frac{1}{2}\begin{pmatrix} \cot\theta_j^{\nu} + \cot\theta_k^{\nu} & -\cot\theta_k^{\nu} & -\cot\theta_j^{\nu} \\ -\cot\theta_k^{\nu} & \cot\theta_i^{\nu} + \cot\theta_k^{\nu} & -\cot\theta_i^{\nu} \\ -\cot\theta_j^{\nu} & -\cot\theta_i^{\nu} & \cot\theta_i^{\nu} + \cot\theta_j^{\nu} \end{pmatrix}. \tag{15.115}$$

One can use either formula for the elemental stiffness matrix. Equation (15.113) is more convenient when one is given the coordinates of its vertices, while (15.115) should be used if one knows its angles.

Figure 15.14. Triangles.

Example 15.18. The right triangle with vertices $(0,0), (1,0), (0,1)$ has elemental stiffness matrix

$$S = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}. \tag{15.116}$$

The same holds for any other isoceles right triangle, as long as we chose the first vertex to be at the right angle and the other two at the 45° angles. A different ordering of the vertices will serve to permute the rows and columns of S. Similarly, an equilateral triangle has all 60° angles, and so its elemental stiffness matrix is

$$S = \begin{pmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}} \\ -\frac{1}{2\sqrt{3}} & \frac{1}{\sqrt{3}} & -\frac{1}{2\sqrt{3}} \\ -\frac{1}{2\sqrt{3}} & -\frac{1}{2\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \approx \begin{pmatrix} 0.577350 & -0.288675 & -0.288675 \\ -0.288675 & 0.577350 & -0.288675 \\ -0.288675 & -0.288675 & 0.577350 \end{pmatrix}. \tag{15.117}$$

Assembling the Elements

Each elemental stiffness matrix will contribute, through the summation (15.112), to the finite element coefficient matrix K. We begin by constructing a larger matrix K^*, which we call the *full finite element matrix*, of size $m \times m$ where m is the total number of nodes in our triangulation, including both interior and boundary nodes. The rows and columns of K^* are labeled by the nodes $\mathbf{x}_i$. On the other hand, the three rows and columns of an individual elemental stiffness matrix S_ν are labeled by the vertices of its triangle T_ν. We let $K_\nu = (k_{ij}^\nu)$ denote the corresponding $m \times m$ matrix containing the 9 entries of S_ν which are placed in the rows and columns corresponding to the vertices of the triangle T_ν; all other entries of K_ν are 0. For example, if T_3 has vertices $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_6$, then the entries of its elemental stiffness matrix S_3 will appear in rows and columns with labels $2, 3, 6$ of the full matrix K_3. The resulting $m \times m$ matrices are all summed together over all the triangles,

$$K^* = \sum_{\nu=1}^{N} K_\nu, \tag{15.118}$$

to produce the full finite element matrix. As in (15.112), each entry $k_{ij} = \sum k_{ij}^\nu$ of K^* will be a sum of elemental stiffnesses corresponding to all the triangles that have $\mathbf{x}_i$ and $\mathbf{x}_j$ as vertices.

Figure 15.15. The Oval Plate.

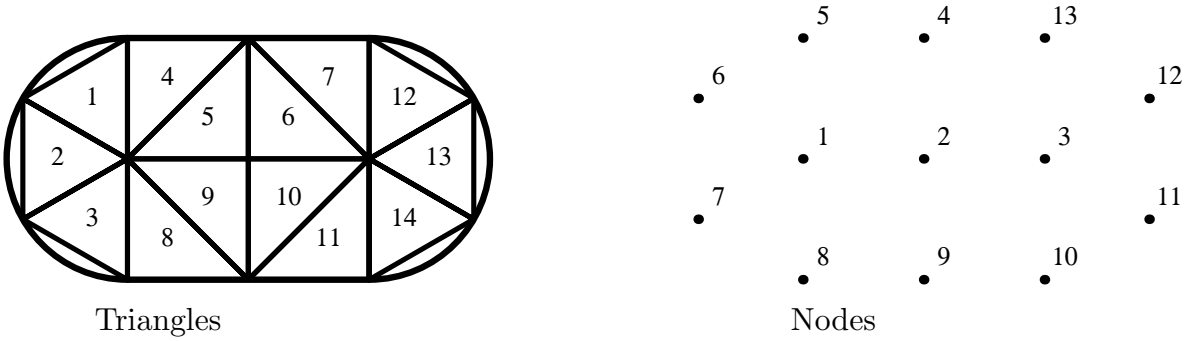


Figure 15.16. A Coarse Triangulation of the Oval Plate.

The full finite element matrix K^* is too large for our linear system (15.109) since its rows and columns include all the nodes, whereas the finite element matrix K appearing in (15.109) only refers to the n interior nodes. The reduced $n \times n$ *finite element matrix* K is simply obtained from K^* by deleting all rows and columns indexed by boundary nodes, retaining only the elements k_{ij} when both $\mathbf{x}_i$ and $\mathbf{x}_j$ are interior nodes. (This may remind the reader of our construction of the reduced incidence matrix for a structure in Chapter 6.) For the homogeneous boundary value problem, this is all we require. However, as we shall see, inhomogeneous boundary conditions are most easily handled by retaining (part of) the full matrix K^*.

The easiest way to understand the construction is through a particular example.

Example 15.19. A metal plate has the shape of an oval running track, consisting of a rectangle, with side lengths 1 m by 2 m, and two semicircular disks glued onto the shorter ends, as sketched in Figure 15.15. The plate is subject to a heat source while its edges are held at a fixed temperature. The problem is to find the equilibrium temperature distribution within the plate. Mathematically, we must solve the Poisson equation $-\Delta u = f$ with prescribed Dirichlet boundary conditions, for the equilibrium temperature $u(x, y)$. Let us describe how to set up the finite element approximation to such a boundary value problem.

We begin with a very coarse triangulation of the plate, which will not give particularly accurate results, but does serve to illustrate how to go about assembling the finite element matrix. We divide the rectangular part of the plate into 8 right triangles, while each semicircular end will be approximated by three equilateral triangles. The triangles are numbered from 1 to 14 as indicated in Figure 15.16. There are 13 nodes in all, numbered

as in the second figure. Only nodes $1, 2, 3$ are interior, while the boundary nodes are labeled 4 through 13, going counterclockwise around the boundary starting at the top. Therefore, the full finite element matrix K^* will have size 13×13, its rows and columns labeled by all the nodes. The reduced matrix K appearing in the finite element equations (15.109) consists of the upper left 3×3 submatrix of K^*.

Each triangle T_ν will contribute the summand K_ν to the matrix K^*, modifying the nine entries k_{ij} indexed by the vertices of T_ν. The values are extracted from the nine entries of the elemental stiffness matrix for that triangle. For example, the first triangle T_1 is equilateral, and so has elemental stiffness matrix (15.117). Its vertices are labeled 1, 5, and 6, and therefore we place the entries of (15.117) in the rows and columns numbered $1, 5, 6$ to form the summand

$$K_1 = \begin{pmatrix} 0.577350 & 0. & 0. & 0. & -0.288675 & -0.288675 & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ -0.288675 & 0. & 0. & 0. & 0.577350 & -0.2886750. & 0. & 0. & \dots \\ -0.288675 & 0. & 0. & 0. & -0.288675 & 0.577350 & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \end{pmatrix},$$

where all the undisplayed entries in the full 13×13 matrix are 0. The next triangle T_2 has the same equilateral elemental stiffness matrix (15.117), but now its vertices are $1, 6, 7$, and so it will contribute

$$K_2 = \begin{pmatrix} 0.577350 & 0. & 0. & 0. & 0. & -0.288675 & -0.288675 & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ -0.288675 & 0. & 0. & 0. & 0. & 0.577350 & -0.2886750. & 0. & \dots \\ -0.288675 & 0. & 0. & 0. & 0. & -0.288675 & 0.5773500. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ \end{pmatrix}.$$

Similarly for K_3, with vertices $1, 7, 8$. On the other hand, triangle T_4 is an isoceles right triangle, and so has elemental stiffness matrix (15.116). Its vertices are labeled 1, 4, and

5, with vertex 5 at the right angle. Therefore, its contribution is

$$K_4 = \begin{pmatrix} 0.5 & 0. & 0. & 0. & -0.5 & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0.5 & -0.5 & 0. & 0. & 0. & \dots \\ -0.5 & 0. & 0. & -0.5 & 1.0 & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ & & & & & & & & \end{pmatrix}.$$

Note particularly how we need to permute the rows and columns of (15.116) in order to have the vertices in the correct order. Continuing in this manner, we assemble 14 contributions $K_1, \dots, K_{14}$, each with (at most) 9 nonzero entries. The full finite element matrix is the sum

$$K^* = K_1 + \cdots + K_{14},$$

and equals

$$K^* = \begin{pmatrix} 3.732 & -1. & 0 & 0 & -0.7887 & -0.5774 & -0.5774 \\ -1. & 4. & -1. & -1. & 0 & 0 & 0 \\ 0 & -1. & 3.732 & 0 & 0 & 0 & 0 \\ 0 & -1. & 0 & 2. & -0.5 & 0 & 0 \\ -0.7887 & 0 & 0 & -0.5 & 1.577 & -0.2887 & 0 \\ -0.5774 & 0 & 0 & 0 & -0.2887 & 1.155 & -0.2887 \\ -0.5774 & 0 & 0 & 0 & 0 & -0.2887 & 1.155 \\ -0.7887 & 0 & 0 & 0 & 0 & 0 & -0.2887 \\ 0 & -1. & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.7887 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.7887 & -0.5 & 0 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} -0.7887 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1. & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.7887 & -0.5774 & -0.5774 & -0.7887 \\ 0 & 0 & 0 & 0 & 0 & -0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -0.2887 & 0 & 0 & 0 & 0 & 0 \\ 1.577 & -0.5 & 0 & 0 & 0 & 0 \\ -0.5 & 2. & -0.5 & 0 & 0 & 0 \\ 0 & -0.5 & 1.577 & -0.2887 & 0 & 0 \\ 0 & 0 & -0.2887 & 1.155 & -0.2887 & 0 \\ 0 & 0 & 0 & -0.2887 & 1.155 & -0.2887 \\ 0 & 0 & 0 & 0 & -0.2887 & 1.577 \end{pmatrix}. \tag{15.119}$$

Since only nodes $1, 2, 3$ are interior nodes, the reduced finite element matrix only uses the

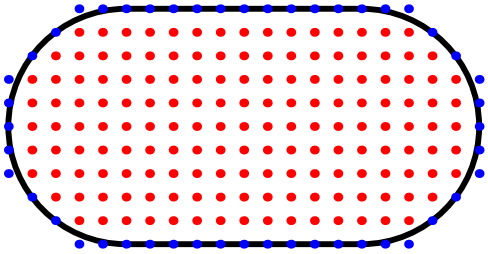


Figure 15.17. A Square Mesh for the Oval Plate.

upper left 3×3 block of K^*, so

$$K = \begin{pmatrix} 3.732 & -1. & 0 \\ -1. & 4. & -1.0 \\ 0 & -1. & 3.732 \end{pmatrix}. \tag{15.120}$$

With some practice, one can learn how to directly construct K, bypassing K^* entirely.

For a finer triangulation, the construction is similar, but the matrices become much larger. The procedure can, of course, be automated. Fortunately, if we choose a very regular triangulation, then we do not need to be nearly as meticulous in assembling the stiffness matrices, since many of the entries are the same. The simplest case is when we use a uniform square mesh, and so triangulate the domain into isoceles right triangles. This is accomplished by laying out a relatively dense square grid over the domain $\Omega \subset \mathbb{R}^2$. The interior nodes are the grid points that fall inside the oval domain, while the boundary nodes are all those grid points lying adjacent to one or more of the interior nodes. The interior nodes will be near but not necessarily on the boundary $\partial\Omega$. Figure 15.17 shows the nodes in a square grid with intermesh spacing $h = .2$. While a bit crude in its approximation of the boundary of the domain, this procedure does have the advantage of making the construction of the associated finite element matrix relatively painless.

For such a mesh, all the triangles are isoceles right triangles, with elemental stiffness matrix (15.116). Summing the corresponding matrices K_ν over all the triangles, as in (15.118), the rows and columns of K^* corresponding to the interior nodes are seen to all have the same form. Namely, if i labels an interior node, then the corresponding diagonal entry is $k_{ii} = 4$, while the off-diagonal entries $k_{ij} = k_{ji}$, $i \neq j$, are equal to either -1 when node i is adjacent to node j on the grid, and is equal to 0 in all other cases. Node j is allowed to be a boundary node. (Interestingly, the result does not depend on how one orients the pair of triangles making up each square of the grid, as in Figure 15.13; the orientation only plays a role in the computation of the right hand side of the finite element equation.) The same computation applies even to our coarse triangulation. The interior node 2 belongs to all right isoceles triangles, and the corresponding entries in (15.119) are $k_{22} = 4$, and $k_{2j} = -1$ for the four adjacent nodes $j = 1, 3, 4, 9$.

Remark: Interestingly, the coefficient matrix arising from the finite element method on a square (or even rectangular) grid is the same as the coefficient matrix arising from a finite difference solution to the Laplace equation; see Exercise ∎. The finite element approach has the advantage of applying to much more general triangulations.

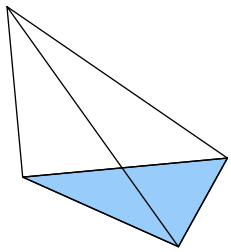


Figure 15.18. Finite Element Tetrahedron.

In general, while the finite element matrix K for a two-dimensional boundary value problem is not as nice as the tridiagonal matrices we obtained in our one-dimensional problems, it is still very sparse and, on regular grids, highly structured. This makes solution of the resulting linear system particularly amenable to an iterative matrix solver such as Gauss–Seidel, Jacobi, or, best of all, successive over-relaxation (SOR).

The Coefficient Vector and the Boundary Conditions

So far, we have been concentrating on assembling the finite element coefficient matrix K. We also need to compute the forcing vector $\mathbf{b} = (\, b_1, b_2, \dots, b_n \,)^T$ appearing on the right hand side of the fundamental linear equation (15.109). According to (15.107), the entries b_i are found by integrating the product of the forcing function and the finite element basis function. As before, we will approximate the integral over the domain Ω by an integral over the triangles, and so

$$b_i = \iint_\Omega f\, \varphi_i \, dx\, dy \;\approx\; \sum_\nu \iint_{T_\nu} f\, \omega_i^\nu \, dx\, dy \equiv \sum_\nu b_i^\nu. \tag{15.121}$$

Typically, the exact computation of the various triangular integrals is not convenient, and so we resort to a numerical approximation. Since we are assuming that the individual triangles are small, we can adopt a very crude numerical integration scheme. If the function $f(x, y)$ does not vary much over the triangle T_ν — which will certainly be the case if T_ν is sufficiently small — we may approximate $f(x, y) \approx c_i^\nu$ for $(x, y) \in T_\nu$ by a constant. The integral (15.121) is then approximated by

$$b_i^\nu = \iint_{T_\nu} f\, \omega_i^\nu \, dx\, dy \;\approx\; c_i^\nu \iint_{T_\nu} \omega_i^\nu(x, y)\, dx\, dy = \tfrac{1}{3}\, c_i^\nu \,\text{area}\, T_\nu = \tfrac{1}{6}\, c_i^\nu \,|\, \Delta_\nu \,|. \tag{15.122}$$

The formula for the integral of the affine element $\omega_i^\nu(x, y)$ follows from solid geometry. Indeed, it equals the volume of the solid under its graph, which is a tetrahedron of height 1 and base T_ν; see Figure 15.18.

How to choose the constant c_i^ν? In practice, the simplest choice is to let $c_i^\nu = f(x_i, y_i)$ be the value of the function at the i^{th} vertex. With this choice, the sum in (15.121) becomes

$$b_i \;\approx\; \sum_\nu \tfrac{1}{3}\, f(x_i, y_i) \,\text{area}\, T_\nu = \tfrac{1}{3}\, f(x_i, y_i) \,\text{area}\, P_i, \tag{15.123}$$

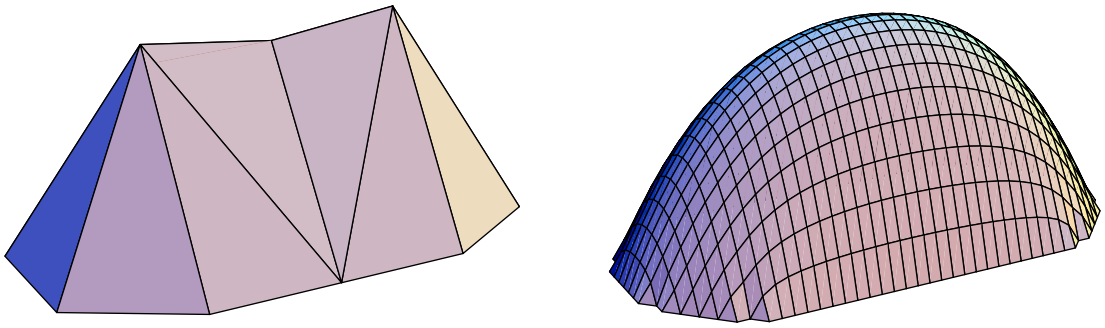


Figure 15.19. Finite Element Solutions to Poisson's Equation for an Oval Plate.

where $P_i = \bigcup T_\nu$ is the vertex polygon (15.106) corresponding to the node $\mathbf{x}_i$. In particular, for the square mesh with the uniform choice of triangles, as in Example 15.17,

$$\text{area } P_i = 3\,h^2 \qquad \text{for all } i, \text{ and so} \qquad b_i \;\approx\; f(x_i, y_i)\,h^2 \tag{15.124}$$

is well approximated by just h^2 times the value of the forcing function at the node. In this case, the finite element equations (15.109) are *identical* with the finite difference equations based on the square grid; see Exercise ■. This is the underlying reason to choose the uniform triangulation for the square mesh; the alternating version would give unequal values for the b_i over adjacent nodes, and this would introduce additional errors into the final approximation.

Example 15.20. For the coarsely triangulated oval plate, the reduced stiffness matrix is (15.120). The Poisson equation $-\Delta u = 4$ models a constant external heat source of magnitude 4° over the entire plate. If we keep the edges of the plate fixed at 0°, then we need to solve the finite element equation $K\mathbf{c} = \mathbf{b}$, where K is the coefficient matrix (15.120), while

$$\mathbf{b} = \tfrac{4}{3}\left(2 + \tfrac{3\sqrt{3}}{4},\ 2,\ 2 + \tfrac{3\sqrt{3}}{4}\right)^T = (\,4.39872, 2.66667, 4.39872\,)^T.$$

The entries of $\mathbf{b}$ are, by (15.123), equal to $4 = f(x_i, y_i)$ times one third the area of the corresponding vertex polygon, which for node 2 is the square consisting of 4 right triangles, each of area $\frac{1}{2}$, whereas for nodes 1 and 3 it consists of 4 right triangles of area $\frac{1}{2}$ plus three equilateral triangles, each of area $\frac{\sqrt{3}}{4}$; see Figure 15.16.

The solution to the final linear system is easily found:

$$\mathbf{c} = (\,1.56724, 1.45028, 1.56724\,)^T.$$

Its entries are the values of the finite element approximation at the three interior nodes. The finite element solution is plotted in the first illustration in Figure 15.19. A more accurate solution, based on a square grid triangulation of size $h = .1$ is plotted in the second figure.

Inhomogeneous Boundary Conditions

So far, we have restricted our attention to problems with homogeneous Dirichlet boundary conditions. According to Theorem 15.14, the solution to the inhomogeneous Dirichlet problem

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad\qquad u = h \quad \text{on} \quad \partial\Omega,$$

is also obtained by minimizing the Dirichlet functional (15.91). However, now the minimization takes place over the affine subspace consisting of all functions that satisfy the inhomogeneous boundary conditions. It is not difficult to fit this problem into the finite element scheme.

The elements corresponding to the interior nodes of our triangulation remain as before, but now we need to include additional elements to ensure that our approximation satisfies the boundary conditions. Note that if $\mathbf{x}_k$ is a boundary node, then the corresponding *boundary element* $\varphi_k(x,y)$ satisfies the interpolation condition (15.97), and so has the same piecewise affine form (15.105). The corresponding finite element approximation

$$w(x,y) = \sum_{i=1}^{m} c_i\, \varphi_i(x,y), \tag{15.125}$$

has the same form as before, (15.98), but now the sum is over all nodes, both interior and boundary. As before, the coefficients $c_i = w(x_i, y_i) \approx u(x_i, y_i)$ are the values of the finite element approximation at the nodes. Therefore, in order to satisfy the boundary conditions, we require

$$c_j = h(x_j, y_j) \qquad \text{whenever} \qquad \mathbf{x}_j = (x_j, y_j) \qquad \text{is a boundary node.} \tag{15.126}$$

Remark: If the boundary node $\mathbf{x}_j$ does not lie precisely on the boundary $\partial\Omega$, we need to approximate the value $h(x_j, y_j)$ appropriately, e.g., by using the value of $h(x,y)$ at the nearest boundary point $(x,y) \in \partial\Omega$.

The derivation of the finite element equations proceeds as before, but now there are additional terms arising from the nonzero boundary values. Leaving the intervening details to the reader, the final outcome can be written as follows. Let K^* denote the full $m \times m$ finite element matrix constructed as above. The reduced coefficient matrix K is obtained by retaining the rows and columns corresponding to only interior nodes, and so will have size $n \times n$, where n is the number of interior nodes. The *boundary coefficient matrix* $\widetilde{K}$ is the $n \times (m-n)$ matrix consisting of the entries of the interior rows that do not appear in K, i.e., those lying in the columns indexed by the boundary nodes. For instance, in the the coarse triangulation of the oval plate, the full finite element matrix is given in (15.119), and the upper 3×3 subblock is the reduced matrix (15.120). The remaining entries of the first three rows form the boundary coefficient matrix

$$\widetilde{K} = \begin{pmatrix} 0 & -0.7887 & -0.5774 & -0.5774 & -0.7887 & 0 & 0 & 0 & 0 & 0 \\ -1. & 0 & 0 & 0 & 0 & -1. & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.7887 & -0.5774 & -0.5774 & -0.7887 \end{pmatrix}. \tag{15.127}$$

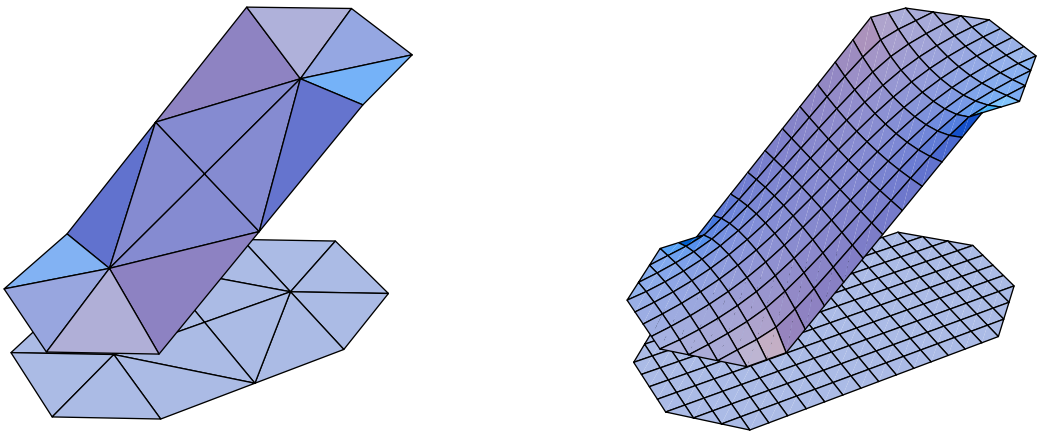


Figure 15.20. Solution to the Dirichlet Problem for the Oval Plate.

We similarly split the coefficients c_i of the finite element function (15.125) into two groups. We let $\mathbf{c} \in \mathbb{R}^n$ denote the as yet unknown coefficients c_i corresponding to the values of the approximation at the interior nodes $\mathbf{x}_i$, while $\mathbf{h} \in \mathbb{R}^{m-n}$ will be the vector of boundary values (15.126). The solution to the finite element approximation (15.125) is obtained by solving the associated linear system

$$K\,\mathbf{c} + \widetilde{K}\,\mathbf{h} = \mathbf{b}, \qquad \text{or} \qquad K\,\mathbf{c} = \mathbf{f} = \mathbf{b} - \widetilde{K}\,\mathbf{h}. \tag{15.128}$$

The full justification of this system is left as an exercise for the reader.

Example 15.21. For the oval plate discussed in Example 15.19, suppose the right hand semicircular edge is held at 10°, the left hand semicircular edge at -10°, while the two straight edges have a linearly varying temperature distribution ranging from -10° at the left to 10° at the right, as illustrated in Figure 15.20. Thus, for the coarse triangulation we have the boundary nodes values

$$\mathbf{h} = (\,h_4, \dots, h_{13}\,)^T = (\,0, -1, -1, -1, -1, 0, 1, 1, 1, 1, 0\,)^T.$$

Using the previously computed formulae (15.120), (15.127) for the interior coefficient matrix K and boundary coefficient matrix $\widetilde{K}$, we approximate the solution to the Laplace equation by solving (15.128). Since there is no external forcing function, $f(x, y) \equiv 0$, the right hand side is $\mathbf{b} = \mathbf{0}$, and so we must solve $K\,\mathbf{c} = \mathbf{f} = -\,\widetilde{K}\,\mathbf{h} = (\,2.18564, 3.6, 7.64974\,)^T$. The finite element function corresponding to the solution $\mathbf{c} = (\,1.06795, 1.8, 2.53205\,)^T$ is plotted in the first illustration in Figure 15.20. Even on such a coarse mesh, the approximation is not too bad, as evidenced by the second illustration, which plots the finite element solution for a square mesh with spacing $h = .2$ between nodes.

Second Order Elliptic Boundary Value Problems

While the Laplace and Poisson equations are by far the most important elliptic partial differential equations, they only model homogeneous media, e.g., membranes made out of

a uniform material, or heated plates with uniform (constant) heat capacity. Inhomogeneous media lead to more general self-adjoint differential operators, leading to variable coefficient second order elliptic boundary value problems. Even more generally, *elastic shells*, meaning bendable plates, lead to fourth order two-dimensional elliptic boundary value problems similar to the one-dimensional beam equation (11.111). And, these are in turn only linear approximations to the fully nonlinear elliptic boundary value problems occurring in elasticity theory, [**69**]. The latter are beyond the scope of this text, although some of the required mathematical tools appear in Chapter 21.

The most important class of linear, self-adjoint, second order, elliptic partial differential equations in two space variables take the form

$$-\frac{\partial}{\partial x}\left(p(x,y)\,\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial x}\left(q(x,y)\,\frac{\partial u}{\partial x}\right)+r(x,y)\,u=f(x,y),\qquad (x,y)\in\Omega,\quad (15.129)$$

where $p(x,y), q(x,y) > 0$ are strictly positive functions, while $r(x,y) \geq 0$ is non-negative. For simplicity, we also impose homogeneous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. Note that the positivity conditions ensure that the partial differential equation is *elliptic* in accordance with the classification of Definition 15.1.

The reader may notice that (15.129) is a two-dimensional version of the Sturm–Liouville ordinary differential equation (11.141). The self-adjoint formulation (11.152) of a Sturm–Liouville boundary value problem serves to inspire the self-adjoint form

$$L^*\circ L[u]=f,\qquad \text{by setting}\qquad L[u]=\begin{pmatrix}u_x\\ u_y\\ u\end{pmatrix},\qquad (15.130)$$

of the boundary value problem for (15.129). Note that the linear operator $L\colon U\to V$ maps the vector space U consisting of all smooth functions $u(x,y)$ satisfying the homogeneous Dirichlet boundary conditions to the vector space V consisting of all vector-valued functions $\mathbf{v} = (\,v_1(x,y), v_2(x,y), v_3(x,y)\,)^T$. We adopt the usual L^2 inner product (15.74) on U, but introduce a weighted inner product[†]

$$\langle\!\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle\!\rangle=\iint_\Omega\,\big(\,p\,v_1\,\widetilde{v}_1+q\,v_2\,\widetilde{v}_2+r\,v_3\,\widetilde{v}_3\,\big)\,dx\,dy$$

on the vector space V. A straightforward computation based on Green's formula (15.78) produces the "weighted adjoint"

$$L^*[\mathbf{v}]=-\frac{\partial}{\partial x}\big[\,p(x,y)\,v_1(x,y)\,\big]-\frac{\partial}{\partial x}\big[\,q(x,y)\,v_2(x,y)\,\big]+r(x,y)\,v_3(x,y)\qquad (15.131)$$

of the operator L. Therefore, the formula for the self-adjoint product

$$L^*\circ L[u]=L^*\begin{pmatrix}u_x\\ u_y\\ u\end{pmatrix}=-\frac{\partial}{\partial x}\left(p(x,y)\,\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial x}\left(q(x,y)\,\frac{\partial u}{\partial y}\right)+r(x,y)\,u(x,y)$$

[†] Technically, we should require that $r(x,y)\not\equiv 0$ not vanish on any open subdomain in order that this define a nondegenerate inner product.

proves the identification of (15.130) and (15.129). Positive definiteness follows from the observation that $\ker L = \{0\}$. The minimization principle associated with the operator L is, as usual,

$$\mathcal{P}[u] = \tfrac{1}{2} \| L[u] \|^2 - \langle f ; u \rangle = \iint_\Omega \left[\tfrac{1}{2} p u_x^2 + \tfrac{1}{2} q u_y^2 + \tfrac{1}{2} r u^2 - f u \right] dx\, dy. \tag{15.132}$$

As always, the solution to our boundary value problem is the unique minimizing function for $\mathcal{P}[u]$ among all functions $u \in U$ satisfying the homogeneous boundary conditions.

Remark: Interestingly, in contrast to the Poisson equation, if $r > 0$ the boundary value problem for (15.129) is positive definite with minimization principle (15.132) even in the case of pure Neumann boundary conditions. This is because the operator L always has trivial kernel.

The finite element approximation is constructed as in the Poisson version — by restricting the minimization principle to the finite-dimensional subspace spanned by the finite element basis functions (11.158). This requires the solution of a linear system of the same form (15.128), in which

$$\begin{aligned} k_{ij} &= \langle L[\varphi_i] ; L[\varphi_j] \rangle = \iint_\Omega \left[p \, \frac{\partial \varphi_i}{\partial x} \frac{\partial \varphi_j}{\partial x} + q \, \frac{\partial \varphi_i}{\partial y} \frac{\partial \varphi_j}{\partial y} + r \, \varphi_i \, \varphi_j \right] dx\, dy, \\ b_i &= \langle f ; \varphi_i \rangle = \iint_\Omega f \, \varphi_i \, dx\, dy. \end{aligned} \tag{15.133}$$

As before, the double integrals are approximated by a sum of integrals over the triangles T_ν. The only triangles that contribute to the final result for k_{ij} are the ones that have both $\mathbf{x}_i$ and $\mathbf{x}_j$ as vertices. When the triangles are small, the integrals can be approximated by fairly crude numerical integration formulae. This completes our brief outline of the method; full details are left to the reader.

Example 15.22. The *Helmholtz equation* is

$$\Delta u + \lambda u = 0, \tag{15.134}$$

along with suitable boundary conditions. As we shall learn in Chapter 17, the Helmholtz equation governs the eigenvalues of the Laplacian, and as such forms the fundamental modes of vibration of a wide variety of mechanical system, including the vibration of plates, scattering of acoustic and electromagnetic waves, and many others.

If $\lambda < 0$, then the Helmholtz equation fits into the positive definite framework (15.129), with $p = q = 1$ and $r = -\lambda$. To solve the problem by finite elements, we restrict the minimization principle

$$\mathcal{P}[u] = \iint_\Omega \left(\tfrac{1}{2} \| \nabla u \|^2 - \tfrac{1}{2} \lambda u^2 - f u \right) dx\, dy. \tag{15.135}$$

to the finite-dimensional finite element subspace determined by a triangulation of the underlying domain. The resulting coefficient matrix has the form

$$\begin{aligned} k_{ij} &= \iint_\Omega \left(\nabla \varphi_i \cdot \nabla \varphi_j - \tfrac{1}{2}\, \lambda\, \varphi_i \varphi_j \right) dx\, dy \\ &\approx \sum_\nu \iint_{T_\nu} \left(\nabla \omega_i^\nu \cdot \nabla \omega_j^\nu - \lambda\, \omega_i^\nu\, \omega_j^\nu \right) dx\, dy \equiv \sum_\nu k_{ij}^\nu. \end{aligned} \tag{15.136}$$

The explicit formulae for the k_{ij}^ν are left as an exercise for the reader. The forcing vector $\mathbf{b}$ has exactly the same form (15.107) as in the Poisson example.

Unfortunately, the most interesting cases are when $\lambda > 0$ and the boundary value problem is not positive definite; nevertheless, the finite element approach can still give quite respectable answers, even though it lacks a simple theoretical justification.

Chapter 16

Complex Analysis

The term "complex analysis" refers to the calculus of complex-valued functions $f(z)$ depending on a complex variable z. On the surface, it may seem that this subject should merely be a simple reworking of standard real variable theory that you learned in first year calculus. However, this naïve first impression could not be further from the truth! Complex analysis is the culmination of a deep and far-ranging study of the fundamental notions of complex differentiation and complex integration, and has an elegance and beauty not found in the more familiar real arena. For instance, complex functions are always *analytic*, meaning that they can be represented as convergent power series. As an immediate consequence, a complex function automatically has an *infinite* number of derivatives, and difficulties with degree of smoothness, strange discontinuities, delta functions, and other forms of pathological behavior of real functions never arise in the complex realm.

There is a remarkable, profound connection between harmonic functions (solutions of the Laplace equation) of two variables and complex-valued functions. Namely, the real and imaginary parts of a complex analytic function are automatically harmonic. In this manner, complex functions provide a rich lode of new solutions to the two-dimensional Laplace equation to help solve boundary value problems. One of the most useful practical consequences arises from the elementary observation that the composition of two complex functions is also a complex function. We interpret this operation as a complex changes of variables, also known as a *conformal mapping* since it preserves angles. Conformal mappings can be effectively used for constructing solutions to the Laplace equation on complicated planar domains, and play a particularly important role in the solution of physical problems. and so on.

Complex integration also enjoys many remarkable properties not found in its real sibling. Integrals of complex functions are similar to the line integrals of planar multivariable calculus. The remarkable theorem due to Cauchy implies that complex integrals are generally path-independent — provided one pays proper attention to the complex singularities of the integrand. In particular, an integral of a complex function around a closed curve can be directly evaluated through the "calculus of residues", which effectively bypasses the Fundamental Theorem of Calculus. Surprisingly, the method of residues can even be applied to evaluate certain types of definite real integrals.

In this chapter, we shall introduce the basic techniques and theorems in complex analysis, paying particular attention to those aspects which are required to solve boundary value problems associated with the planar Laplace and Poisson equations. Complex analysis is an essential tool in a surprisingly broad range of applications, including fluid flow, elasticity, thermostatics, electrostatics, and, in mathematics, geometry, and even number

theory. Indeed, the most famous unsolved problem in all of mathematics, the Riemann hypothesis, is a conjecture about a specific complex function that has profound consequences for the distribution of prime numbers†.

16.1. Complex Variables.

In this section we shall develop the basics of complex analysis — the calculus of complex functions $f(z)$. Here $z = x + \mathrm{i}\, y$ is a single complex variable and $f\colon \Omega \to \mathbb{C}$ is a complex-valued function defined on a domain $z \in \Omega \subset \mathbb{C}$ in the complex plane. Before diving into this material, the reader should first review the basic material on complex numbers in Section 3.6.

Any complex function can be written as a complex combination

$$f(z) = f(x + \mathrm{i}\, y) = u(x,y) + \mathrm{i}\, v(x,y), \tag{16.1}$$

of two real functions u, v of two real variables x, y, called, respectively, its *real* and *imaginary parts*, and written

$$u(x,y) = \operatorname{Re} f(z), \qquad \text{and} \qquad v(x,y) = \operatorname{Im} f(z). \tag{16.2}$$

For example, the monomial function $f(z) = z^3$ is written as

$$z^3 = (x + \mathrm{i}\, y)^3 = (x^3 - 3xy^2) + \mathrm{i}\,(3x^2y - y^3),$$

and so

$$\operatorname{Re} z^3 = x^3 - 3xy^2, \qquad \operatorname{Im} z^3 = 3x^2y - y^3.$$

As we know, we can identify $\mathbb{C}$ with the real, two-dimensional plane $\mathbb{R}^2$, so that the complex number $z = x + \mathrm{i}\, y \in \mathbb{C}$ is identified with the real vector $(x, y)^T \in \mathbb{R}^2$. Based on the identification $\mathbb{C} \simeq \mathbb{R}^2$, we shall adopt the usual terminology from planar vector calculus, e.g., domain, curve, etc., without alteration; see Appendix A for details. In this manner, we may regard a complex function as particular type of real vector field that maps

$$\begin{pmatrix} x \\ y \end{pmatrix} \in \Omega \subset \mathbb{R}^2 \qquad \text{to the vector} \qquad \mathbf{v}(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} \in \mathbb{R}^2. \tag{16.3}$$

Not every real vector field qualifies as a complex function; the components $u(x,y), v(x,y)$ must satisfy certain fairly stringent requirements; see Theorem 16.3 below.

Many of the well-known functions appearing in real-variable calculus — polynomials, rational functions, exponentials, trigonometric functions, logarithms, and many others — have natural complex extensions. For example, complex polynomials

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \tag{16.4}$$

are complex linear combinations (meaning that the coefficients a_k are allowed to be complex numbers) of the basic monomial functions $z^k = (x + \mathrm{i}\, y)^k$. Similarly, we have already made

† Not to mention that a solution will net you a cool \$1,000,000.00. For details on how to claim your prize, check out the web site `http://www.claymath.org`.

sporadic use of complex exponentials such as $e^z = e^{x+\mathrm{i}\,y}$ for solving differential equations. Other examples will appear shortly.

There are several ways to motivate[†] the link between harmonic functions $u(x,y)$, meaning solutions of the two-dimensional Laplace equation

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\ , = 0 \tag{16.5}$$

and complex functions. One natural starting point is to return to the d'Alembert solution (14.121) of the one-dimensional wave equation, which was based on the factorization

$$\Box = \partial_t^2 - c^2\,\partial_x^2 = (\partial_t - c\,\partial_x)\,(\partial_t + c\,\partial_x)$$

of the linear wave operator (14.109). The two-dimensional Laplace operator $\Delta = \partial_x^2 + \partial_y^2$ has essentially the same form, except for a "minor" change in sign[†]. We cannot produce a real factorization of the Laplace operator, but there is a complex factorization,

$$\Delta = \partial_x^2 + \partial_y^2 = (\partial_x - \mathrm{i}\,\partial_y)\,(\partial_x + \mathrm{i}\,\partial_y),$$

into a product of two complex first order differential operators. The wave speed has now become complex: $c = \mathrm{i}$. Mimicking the solution formula (14.117) for the wave equation, we expect that the solutions to the Laplace equation (16.5) should be expressed in the form

$$u(x,y) = f(x + \mathrm{i}\,y) + g(x - \mathrm{i}\,y), \tag{16.6}$$

i.e., a linear combination of functions of the complex variable $z = x + \mathrm{i}\,y$ and its complex conjugate $\overline{z} = x - \mathrm{i}\,y$. The functions f and g satisfy the first order complex partial differential equations

$$\frac{\partial f}{\partial x} = -\,\mathrm{i}\,\frac{\partial f}{\partial y}\,, \qquad\qquad \frac{\partial g}{\partial x} = \mathrm{i}\,\frac{\partial g}{\partial y}\,, \tag{16.7}$$

and hence (16.6) does indeed define a complex-valued solution to the Laplace equation.

In most applications, we are searching for a real solution to the Laplace equation, and so our d'Alembert-type formula (16.6) is not entirely satisfactory. As we know, a complex number $z = x + \mathrm{i}\,y$ is real if and only if it equals its own conjugate, $z = \overline{z}$. Thus, the solution (16.6) will be real if and only if

$$f(x + \mathrm{i}\,y) + g(x - \mathrm{i}\,y) = u(x,y) = \overline{u(x,y)} = \overline{f(x + \mathrm{i}\,y)} + \overline{g(x - \mathrm{i}\,y)}.$$

Now, the complex conjugation operation switches $x + \mathrm{i}\,y$ and $x - \mathrm{i}\,y$, and so we expect the first term $\overline{f(x + \mathrm{i}\,y)}$ to be a function of $x - \mathrm{i}\,y$, while the second term $\overline{g(x - \mathrm{i}\,y)}$ will be a

---

[†] A reader uninterested in the motivation can skip ahead to Proposition 16.1 at this point.

[†] However, the change in sign has dramatic consequences for the analytical properties of solutions to the two equations. According to Section 15.1, there is a vast difference between the elliptic Laplace equation and the hyperbolic wave equation.

function of $x + \mathrm{i}\, y$. Therefore[‡], to equate the two sides of this equation, we should require

$$g(x - \mathrm{i}\, y) = \overline{f(x + \mathrm{i}\, y)},$$

and so

$$u(x, y) = f(x + \mathrm{i}\, y) + \overline{f(x + \mathrm{i}\, y)} = 2 \operatorname{Re} f(x + \mathrm{i}\, y).$$

Dropping the inessential factor of 2, we conclude that a real solution to the two-dimensional Laplace equation can be written as the *real part* of a complex function. A direct proof of the following key result will appear below.

Proposition 16.1. *If $f(z)$ is a complex function, then its real part*

$$u(x, y) = \operatorname{Re} f(x + \mathrm{i}\, y) \tag{16.8}$$

is a harmonic function.

The *imaginary part* of a complex function is also harmonic. This is because

$$\operatorname{Im} f(z) = \operatorname{Re} (-\mathrm{i}\, f(z))$$

is the real part of the complex function

$$-\mathrm{i}\, f(z) = -\mathrm{i}\, [\, u(x, y) + \mathrm{i}\, v(x, y)\,] = v(x, y) - \mathrm{i}\, u(x, y).$$

Therefore, if $f(z)$ is any complex function, we can write it as a complex combination

$$f(z) = f(x + \mathrm{i}\, y) = u(x, y) + \mathrm{i}\, v(x, y),$$

of two real harmonic functions $u(x, y) = \operatorname{Re} f(z)$ and $v(x, y) = \operatorname{Im} f(z)$.

Before delving into the many remarkable properties of complex functions, we look at some of the most basic examples. In each case, the reader can check directly that the harmonic functions given as the real and imaginary parts of the complex function are indeed solutions to the Laplace equation.

Examples of Complex Functions

(a) *Harmonic Polynomials*: The simplest examples of complex functions are polynomials. Any polynomial is a complex linear combinations, as in (16.4), of the basic complex monomials

$$z^n = (x + \mathrm{i}\, y)^n = u_n(x, y) + \mathrm{i}\, v_n(x, y). \tag{16.9}$$

The real and imaginary parts of a complex polynomial are known as *harmonic polynomials*, and we list the first few below. The general formula for the basic harmonic polynomials $u_n(x, y)$ and $v_n(x, y)$ is easily found by applying the binomial theorem, as in Exercise ∎.

[‡] We are ignoring the fact that f and g are not quite uniquely determined since one can add and subtract a constant from them. This does not affect the argument in any significant way.

Figure 16.1. Real and Imaginary Parts of z^2 and z^3.

Harmonic Polynomials

n	z^n	$u_n(x,y)$	$v_n(x,y)$
0	1	1	0
1	$x + \mathrm{i}\, y$	x	y
2	$(x^2 - y^2) + 2\,\mathrm{i}\, x\, y$	$x^2 - y^2$	$2\,x\, y$
3	$(x^3 - 3\,x\, y^2) + \mathrm{i}\,(3\,x^2 y - y^3)$	$x^3 - 3\,x\, y^2$	$3\,x^2 y - y^3$
4	$(x^4 - 6\,x^2 y^2 + y^4) + \mathrm{i}\,(4\,x^3 y - 4\,x\, y^3)$	$x^4 - 6\,x^2 y^2 + y^4$	$4\,x^3 y - 4\,x\, y^3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

We have, in fact, already encountered these polynomial solutions to the Laplace equation. If we write

$$z = r\, e^{\,\mathrm{i}\,\theta}, \tag{16.10}$$

where

$$r = |\, z\,| = \sqrt{x^2 + y^2}, \qquad \theta = \mathrm{ph}\, z = \tan^{-1}\frac{y}{x}\,,$$

are the usual polar coordinates (modulus and phase) of $z = x + \mathrm{i}\, y$, then Euler's formula (3.76) yields

$$z^n = r^n\, e^{\,\mathrm{i}\, n\,\theta} = r^n \cos n\,\theta + \mathrm{i}\, r^n \sin n\,\theta,$$

and so

$$u_n = r^n \cos n\,\theta, \qquad v_n = r^n \sin n\,\theta.$$

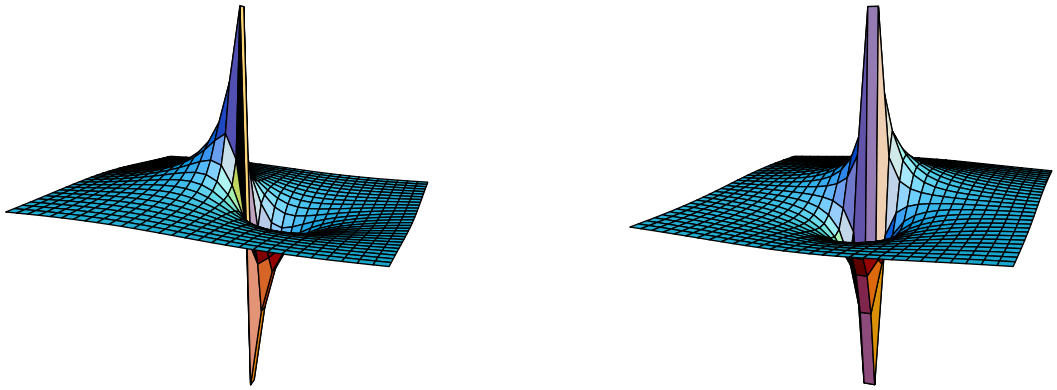


Figure 16.2. Real and Imaginary Parts of $f(z) = \dfrac{1}{z}$.

Therefore, the harmonic polynomials are just the polar coordinate solutions (15.35) to the Laplace equation we obtained previously by the method of separation of variables. In Figure 16.1 we plot† the real and imaginary parts of the monomials z^2 and z^3.

(*b*) *Rational Functions*: Ratios

$$f(z) = \frac{p(z)}{q(z)} \tag{16.11}$$

of complex polynomials provide a large variety of harmonic functions. The simplest case is

$$\frac{1}{z} = \frac{\overline{z}}{z\,\overline{z}} = \frac{\overline{z}}{|\,z\,|^2} = \frac{x}{x^2+y^2} - \mathrm{i}\,\frac{y}{x^2+y^2}\,. \tag{16.12}$$

Its real and imaginary parts are graphed in Figure 16.2. Note that these functions have an interesting singularity at the origin $x = y = 0$, and are harmonic everywhere else.

A slightly more complicated example is the useful function

$$f(z) = \frac{z-1}{z+1}\,. \tag{16.13}$$

To write out (16.13) in real form, we multiply both numerator and denominator by the complex conjugate of the denominator, leading to

$$f(z) = \frac{z-1}{z+1} = \frac{(z-1)(\overline{z}+1)}{(z+1)(\overline{z}+1)} = \frac{|\,z\,|^2 - 1 + z - \overline{z}}{|\,z+1\,|^2} = \frac{x^2+y^2-1}{(x+1)^2+y^2} + \mathrm{i}\,\frac{2\,y}{(x+1)^2+y^2}\,. \tag{16.14}$$

This manipulation can always be used to find the real and imaginary parts of general rational functions.

---

† Graphing a complex function $f\colon \mathbb{C} \to \mathbb{C}$ is problematic. The identification (16.3) of f with a real vector-valued function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ implies that one requires four real dimensions to display the complete graph.

If we assume that the rational function (16.11) is written in lowest terms, so p and q have no common factors, then $f(z)$ will have a singularity, known as a *pole*, wherever the denominator vanishes: $q(z_0) = 0$. The order† of the root z_0 of $q(z)$ tells us the *order* of the pole of $f(z)$. For example, the rational function

$$f(z) = \frac{z+2}{z^5 + z^3} = \frac{z+2}{(z+\mathrm{i}\,)(z-\mathrm{i}\,)z^3}$$

has three poles: a simple (of order 1) pole at $z = +\mathrm{i}\,$, another simple pole at $z = -\mathrm{i}\,$ and a triple (order 3) pole at $z = 0$.

(c) *Complex Exponentials*: Euler's formula

$$e^z = e^x \cos y + \mathrm{i}\, e^x \sin y \tag{16.15}$$

for the complex exponential, cf. (3.76), yields two important harmonic functions: $e^x \cos y$ and $e^x \sin y$, which are graphed in Figure 3.7. More generally, writing out e^{cz} for a complex constant $c = a + \mathrm{i}\, b$ produces the general complex exponential function

$$e^{cz} = e^{ax-by}\cos(bx + ay) + \mathrm{i}\, e^{ax-by}\sin(bx + ay). \tag{16.16}$$

Its real and imaginary parts are harmonic for arbitrary $a, b \in \mathbb{R}$. We already encountered some of these solutions to the Laplace equation when we used the separation of variables method in Cartesian coordinates; see the table in Section 15.2.

(d) *Complex Trigonometric Functions*: The complex trigonometric functions are defined in terms of the complex exponential by adapting our earlier formulae (3.78):

$$\begin{aligned} \cos z &= \frac{e^{\mathrm{i}\, z} + e^{-\mathrm{i}\, z}}{2} = \cos x \cosh y - \mathrm{i}\, \sin x \sinh y, \\ \sin z &= \frac{e^{\mathrm{i}\, z} - e^{-\mathrm{i}\, z}}{2\,\mathrm{i}} = \sin x \cosh y + \mathrm{i}\, \cos x \sinh y. \end{aligned} \tag{16.17}$$

The resulting harmonic functions are products of trigonometric and hyperbolic functions. They can all be written as linear combinations of the harmonic functions (16.16) derived from the complex exponential. Note that when $z = x$ is real, so $y = 0$, these functions reduce to the usual real trigonometric functions $\cos x$ and $\sin x$.

(e) *Complex Logarithm*: In a similar fashion, the complex (natural) logarithm $\log z$ is a complex extension of the usual real natural (i.e., base e) logarithm. In terms of polar coordinates (16.10), the complex logarithm has the form

$$\log z = \log(r\, e^{\mathrm{i}\,\theta}) = \log r + \log e^{\mathrm{i}\,\theta} = \log r + \mathrm{i}\,\theta, \tag{16.18}$$

Thus, the logarithm of a complex number has real part

$$\operatorname{Re} \log z = \log r = \tfrac{1}{2}\, \log(x^2 + y^2),$$

† Recall that the *order* of a root z_0 of a polynomial $q(z)$ is the number times $z - z_0$ occurs as a factor of $q(z)$.

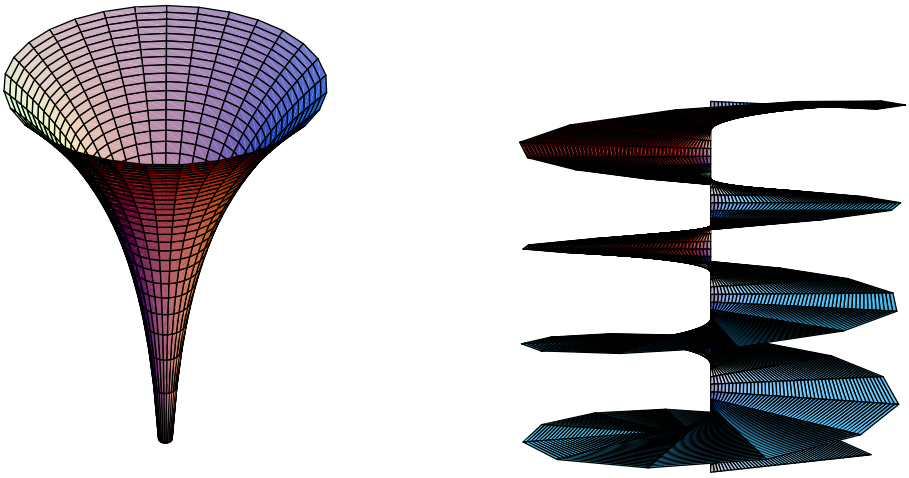


Figure 16.3. Real and Imaginary parts of $\log z = \log r + \mathrm{i}\,\theta$.

which is a well-defined harmonic function on all of $\mathbb{R}^2$ except at the origin $x = y = 0$, where it has a logarithmic singularity. It is, in fact, the logarithmic potential corresponding to a delta function forcing concentrated at the origin that played a key role in the construction of the Green's function for the Poisson equation in Section 15.3.

The imaginary part

$$\operatorname{Im} \log z = \theta = \operatorname{ph} z = \tan^{-1} \frac{y}{x}$$

of the complex logarithm is the *phase* or polar angle of z. The phase is also not defined at the origin $x = y = 0$. Moreover, it is a multiply-valued harmonic function elsewhere, since it is only specified up to integer multiples of 2π. Thus, a given nonzero complex number $z \neq 0$ has an infinite number of possible values for its phase, and hence an infinite number of possible complex logarithms $\log z$, each differing by an integer multiple of $2\pi\,\mathrm{i}$, reflecting the fact that $e^{2\pi\,\mathrm{i}} = 1$. In particular, if $z = x > 0$ is real and positive, then $\log z = \log x$ agrees with the real logarithm, *provided* we choose the angle $\operatorname{ph} z = 0$. Alternative choices for the phase include a multiple of $2\pi\,\mathrm{i}$, and so ordinary real, positive numbers $x > 0$ also have complex logarithms! On the other hand, if $z = x < 0$ is real and negative, then $\log z = \log |\, x \,| + (2k+1)\pi\,\mathrm{i}$ is complex no matter which value of $\operatorname{ph} z$ is chosen. (This explains why we didn't attempt to define the logarithm of a negative number in first year calculus!) In general, as we circle around the origin in a counterclockwise direction, $\operatorname{Im} \log z = \operatorname{ph} z = \theta$ increases by 2π, and so its graph can be likened to an infinitely tall parking ramp with infinitely many levels, spiraling upwards as one goes around the origin, as sketched in Figure 16.3. For the complex logarithm, the origin is a type of singularity known as a *logarithmic branch point*, indicating that there are an infinite number of possible "branches" meaning values that can be assigned to $\log z$ at any nonzero point.

Although the complex logarithm $\log z$ *not* a single-valued complex function on all of $\mathbb{C} \setminus \{0\}$, it *can* be continuously and unambiguously defined when restricted to any simply connected domain $\Omega \subset \mathbb{C} \setminus \{0\}$ that does not include the origin. Essentially, the specification

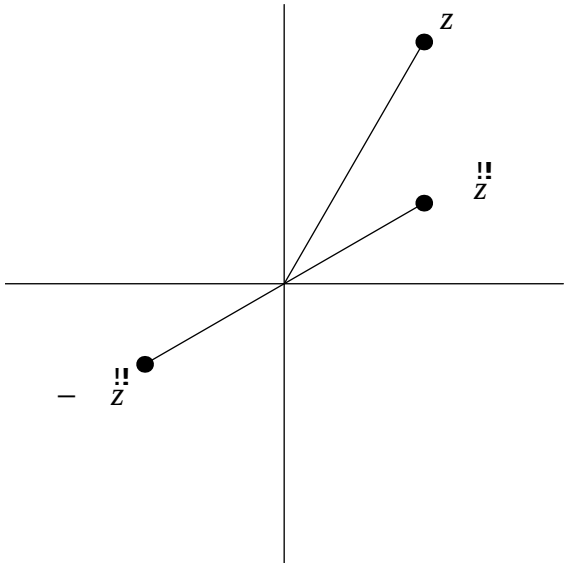


Figure 16.4. Square Roots of a Complex Number.

of logarithm amounts to an unambiguous choice of level of our parking ramp sitting over the domain Ω. For instance, if we restrict our attention to points in the domain

$$\Omega_* = \mathbb{C} \setminus \big\{\, x = \operatorname{Re} z \le 0,\ y = \operatorname{Im} z = 0 \,\big\} = \big\{\, -\pi < \operatorname{ph} z < \pi \,\big\}$$

obtained by "cutting" the complex plane along the negative real axis, then we can uniquely specify an angle by requiring that it lie between $-\pi$ and π. This in turn produces a unique, continuous specification of $\log z$ for all $z \in \Omega_*$. However, other choices are possible, and, indeed, may be required for a given application.

(f) *Roots and Fractional Powers*: A similar branching phenomenon occurs with the fractional powers and roots of complex numbers. The simplest case is the square root function $\sqrt{z}$. Every nonzero complex number $z \neq 0$ has two different possible square roots: $\sqrt{z}$ and $-\sqrt{z}$. As illustrated in Figure 16.4, the two square roots lie on opposite sides of the origin, and are obtained by multiplying by -1. Writing $z = r\, e^{\,\mathrm{i}\,\theta}$ in polar coordinates, we see that

$$\sqrt{z} = \sqrt{r\, e^{\,\mathrm{i}\,\theta}} = \sqrt{r}\, e^{\,\mathrm{i}\,\theta/2} = \sqrt{r}\, \left(\cos \frac{\theta}{2} + \mathrm{i} \sin \frac{\theta}{2} \right), \tag{16.19}$$

i.e., we take the square root of the modulus and halve the phase:

$$\big|\, \sqrt{z}\, \big| = \sqrt{|\, z\,|} = \sqrt{r}\,, \qquad\qquad \operatorname{ph} \sqrt{z} = \tfrac{1}{2} \operatorname{ph} z = \tfrac{1}{2}\, \theta.$$

since θ is only defined up to an integer multiple of 2π, the angle $\frac{1}{2}\,\theta$ is only defined up to an integer multiple of π. The odd and even multiples yield different values for (16.19), which accounts for the two possible values of the square root. For example, since $\operatorname{ph} 4\,\mathrm{i} = \frac{1}{2}\,\pi$ or $\frac{5}{2}\,\pi$, we find

$$\sqrt{4\,\mathrm{i}} = 2\sqrt{\mathrm{i}} = \pm\, 2\, e^{\pi\,\mathrm{i}\,/4} = \pm\, 2 \left(\cos \frac{\pi\,\mathrm{i}}{4} + \mathrm{i} \sin \frac{\pi\,\mathrm{i}}{4} \right) = \pm \big(\sqrt{2} + \mathrm{i}\, \sqrt{2} \big).$$

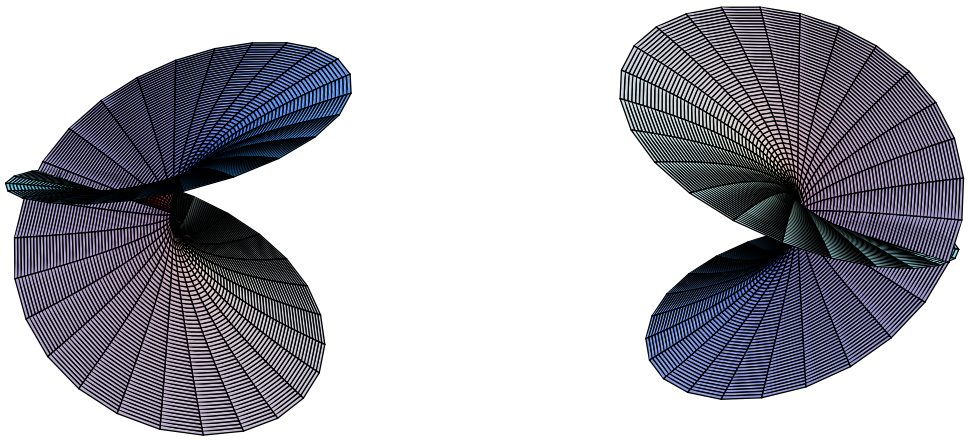


Figure 16.5. Real and Imaginary Parts of $\sqrt{z}$.

If we start at some $z \neq 0$ and circle once around the origin, we increase $\operatorname{ph} z$ by 2π, but $\operatorname{ph} \sqrt{z}$ only increases by π. Thus, at the end of our circumambulation, we arrive at the other square root $-\sqrt{z}$. Circling the origin again increases $\operatorname{ph} z$ by a further 2π, and hence brings us back to the original square root $\sqrt{z}$. Therefore, the graph of the multiply-valued square root function will look like a weirdly interconnected parking ramp with only two levels, as in† Figure 16.5.

Similar remarks apply to the n^{th} root

$$\sqrt[n]{z} = \sqrt[n]{r}\, e^{\,\mathrm{i}\,\theta/n} = \sqrt[n]{r}\, \left(\cos \frac{\theta}{n} + \mathrm{i} \sin \frac{\theta}{n} \right), \tag{16.20}$$

which, except for $z = 0$, has n possible values, depending upon which multiple of 2π is used in the assignment of $\operatorname{ph} z = \theta$. The n different n^{th} roots are obtained by multiplying any one of them by the different n^{th} roots of unity, $\zeta_n^k = e^{2\,k\,\pi\,\mathrm{i}\,/n}$ for $k = 0, \dots, n-1$, as defined in (13.12). In this case, the origin $z = 0$ is called a *branch point* of *order* n since there are n different branches for the function $\sqrt[n]{z}$. Circling around the origin leads to successive branches, returning after circling n times, to the original branch of $\sqrt[n]{z}$.

The preceding list of examples is far from exhausting the range and variety of complex functions. Lack of space will preclude us from studying the remarkable properties of complex versions of the gamma function, Airy functions, Bessel functions, and Legendre functions that appear in Appendix C, as well as elliptic functions, the Riemann zeta function, modular functions, and many, many other important and fascinating functions arising in complex analysis and its manifold applications; see [**141**, **149**].

† These graphs are best appreciated in a fully functional three-dimensional graphics viewer.

16.2. Complex Differentiation.

Complex function theory is founded upon the notion of the complex derivative. Complex differentiation is defined in a direct analogy with the usual calculus limit definition of the derivative of a real function. Yet, despite a superficial similarity, the resulting theory of complex differentiation is profoundly different, and has an elegance and depth not shared by its real progenitor.

Definition 16.2. A complex function $f(z)$ is *differentiable* at a point $z \in \mathbb{C}$ if and only if the limiting difference quotient exists:

$$f'(z) = \lim_{w \to z} \frac{f(w) - f(z)}{w - z}\,. \tag{16.21}$$

The key feature of this definition is that the limiting value $f'(z)$ of the difference quotient must be *independent* of how the point w converges to z. On the real line, there are only two basic directions to approach a limiting point — either from the left or from the right. These lead to the concepts of left and right handed derivatives and their equality is required for the existence of the usual derivative of a real function. In the complex plane, there are an infinite variety of directions[†] in which one can approach the point z, and the definition requires that all of these "directional derivatives" must agree. This is the reason for the more severe restrictions on complex derivatives, and, in consequence, the source of their remarkable properties.

Let us see what happens when we approach z along the two simplest directions — horizontal and vertical. If we set

$$w = z + h = (x + h) + \mathrm{i}\, y, \qquad \text{where } h \text{ is real},$$

then $w \to z$ along a horizontal line as $h \to 0$, as in Figure 16.6. If we write out

$$f(z) = u(x, y) + \mathrm{i}\, v(x, y)$$

in terms of its real and imaginary parts, then we must have

$$\begin{aligned} f'(z) &= \lim_{h \to 0} \frac{f(z + h) - f(z)}{h} = \lim_{h \to 0} \frac{f(x + h + \mathrm{i}\, y) - f(x + \mathrm{i}\, y)}{h} \\ &= \lim_{h \to 0} \left[\frac{u(x + h, y) - u(x, y)}{h} + \mathrm{i}\, \frac{v(x + h, y) - v(x, y)}{h} \right] = \frac{\partial u}{\partial x} + \mathrm{i}\, \frac{\partial v}{\partial x} = \frac{\partial f}{\partial x}\,, \end{aligned}$$

which follows from the usual definition of the (real) partial derivative. On the other hand, if we set

$$w = z + \mathrm{i}\, k = x + \mathrm{i}\, (y + k), \qquad \text{where } k \text{ is real},$$

[†] Not to mention other approaches along parabolas, spirals, etc., although, as it turns out, these more exotic routes do not lead to any further restrictions on the function.

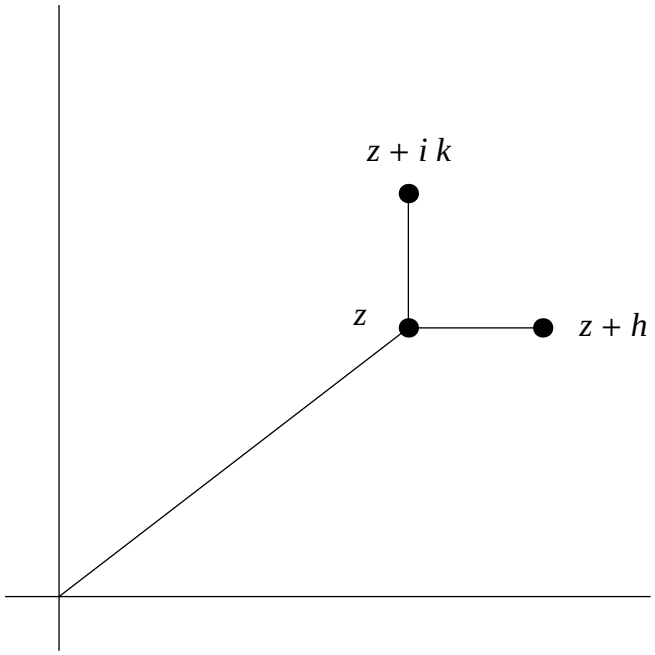


Figure 16.6. Complex Derivative Directions.

then $w \to z$ along a vertical line as $k \to 0$. Therefore, we must also have

$$f'(z) = \lim_{k \to 0} \frac{f(z + \mathrm{i}\, k) - f(z)}{\mathrm{i}\, k} = \lim_{k \to 0} -\mathrm{i}\, \frac{f(x + \mathrm{i}\,(y + k)) - f(x + \mathrm{i}\, y)}{k}$$
$$= \lim_{h \to 0} \left[\frac{v(x, y + k) - v(x, y)}{k} - \mathrm{i}\, \frac{u(x, y + k) - u(x, y)}{k} \right] = \frac{\partial v}{\partial y} - \mathrm{i}\, \frac{\partial u}{\partial y} = -\mathrm{i}\, \frac{\partial f}{\partial y}\,.$$

When we equate the real and imaginary parts of these two distinct formulae for the complex derivative $f'(z)$, we discover that the real and imaginary components of $f(z)$ must satisfy a certain homogeneous linear system of partial differential equations, named after Augustin–Louis Cauchy and Bernhard Riemann, two of the principal founders of modern complex analysis.

Theorem 16.3. *A function $f(z)$ has a complex derivative $f'(z)$ if and only if its real and imaginary parts are continuously differentiable and satisfy the* Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}\,, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}\,. \tag{16.22}$$

In this case, the complex derivative of $f(z)$ is equal to any of the following expressions:

$$f'(z) = \frac{\partial f}{\partial x} = \frac{\partial u}{\partial x} + \mathrm{i}\, \frac{\partial v}{\partial x} = -\mathrm{i}\, \frac{\partial f}{\partial y} = \frac{\partial v}{\partial y} - \mathrm{i}\, \frac{\partial u}{\partial y}\,. \tag{16.23}$$

The proof of the converse — that any function whose real and imaginary components satisfy the Cauchy–Riemann equations is differentiable — will be omitted, but can be found in any basic text on complex analysis, e.g., [**4**, **127**].

Remark: It is worth pointing out that equation (16.23) tells us that f satisfies $\partial f/\partial x = -\,\mathrm{i}\,\partial f/\partial y$, which, reassuringly, agrees with the first equation in (16.7).

Example 16.4. Consider the elementary function

$$z^3 = (x^3 - 3\,x\,y^2) + \mathrm{i}\,(3\,x^2 y - y^3).$$

Its real part $u = x^3 - 3\,x\,y^2$ and imaginary part $v = 3\,x^2 y - y^3$ satisfy the Cauchy–Riemann equations (16.22), namely

$$\frac{\partial u}{\partial x} = 3\,x^2 - 3\,y^2 = \frac{\partial v}{\partial y}\,, \qquad\qquad \frac{\partial u}{\partial y} = -6\,x\,y = -\,\frac{\partial v}{\partial x}\,.$$

This proves that $f(z) = z^3$ is complex differentiable. Not surprisingly, its derivative turns out to be

$$f'(z) = 3\,z^2 = (3\,x^2 - 3\,y^2) + \mathrm{i}\,(6\,x\,y) = \frac{\partial u}{\partial x} + \mathrm{i}\,\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - \mathrm{i}\,\frac{\partial u}{\partial y}\,.$$

Fortunately, the complex derivative obeys all of the usual rules that you learned in real-variable calculus. For example,

$$\frac{d}{dz}\,z^n = n\,z^{n-1}, \qquad\qquad \frac{d}{dz}\,e^{c\,z} = c\,e^{c\,z}, \qquad\qquad \frac{d}{dz}\log z = \frac{1}{z}\,, \tag{16.24}$$

and so on. The power n can even be non-integral or, in view of the identity $z^n = e^{n \log z}$, complex, while c is any complex constant. The exponential formulae (16.17) for the complex trigonometric functions implies that they also satisfy the standard rules

$$\frac{d}{dz}\cos z = -\sin z, \qquad\qquad \frac{d}{dz}\sin z = \cos z. \tag{16.25}$$

The formulae for differentiating sums, products, ratios, inverses, and compositions of complex functions are all the same as their real counterparts. Thus, thankfully, you don't need to learn any new rules for performing complex differentiation!

There are many examples of quite reasonable functions which do *not* have a complex derivative. The simplest is the complex conjugate function

$$f(z) = \overline{z} = x - \mathrm{i}\,y.$$

Its real and imaginary parts do *not* satisfy the Cauchy–Riemann equations, and hence $\overline{z}$ does *not* have a complex derivative. More generally, any function $f(x,y) = h(z,\overline{z})$ that explicitly depends on the complex conjugate variable $\overline{z}$ is *not* complex-differentiable.

Power Series and Analyticity

The most remarkable feature of complex analysis, which completely distinguishes it from real function theory, is that the existence of one complex derivative automatically implies the existence of infinitely many! All complex functions $f(z)$ are infinitely differentiable and, in fact, analytic where defined. The reason for this surprising and profound fact will, however, not become evident until we learn the basics of complex integration in Section 16.5. In this section, we shall take analyticity as a given, and investigate some of its principal consequences.

Definition 16.5. A complex function $f(z)$ is called *analytic* at a point z_0 if it has a power series expansion

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + a_3(z - z_0)^3 + \cdots = \sum_{n=0}^{\infty} a_n (z - z_0)^n, \quad (16.26)$$

which converges for all z sufficiently close to z_0.

Typically, the standard ratio or root tests for convergence of (real) series that you learned in ordinary calculus, [**9**, **136**], can be applied to determine where a given (complex) power series converges. We note that if $f(z)$ and $g(z)$ are analytic at a point z_0, so is their sum $f(z) + g(z)$, product $f(z)\, g(z)$ and, provided $g(z_0) \neq 0$, ratio $f(z)/g(z)$.

Example 16.6. All of the real power series from elementary calculus carry over to the complex versions of the standard functions. For example,

$$e^z = 1 + z + \frac{1}{2} z^2 + \frac{1}{6} z^3 + \cdots = \sum_{n=0}^{\infty} \frac{z^n}{n!} \quad (16.27)$$

is the Taylor series for the exponential function based at $z_0 = 0$. A simple application of the ratio test proves that the series converges for all z. On the other hand, the power series

$$\frac{1}{z^2 + 1} = 1 - z^2 + z^4 - z^6 + \cdots = \sum_{k=0}^{\infty} (-1)^k z^{2k}, \quad (16.28)$$

converges inside the unit disk, where $|\, z \,| < 1$, and diverges outside, where $|\, z \,| > 1$. Again, convergence is established through the ratio test. The ratio test is inconclusive when $|\, z \,| = 1$, and we shall leave the much harder question of precisely where on the unit disk this complex series converges to a more advanced text, e.g., [**4**].

In general, there are three possible options for the domain of convergence of a complex power series (16.26):

(a) The series converges for all z.

(b) The series converges inside a disk $|\, z - z_0 \,| < \rho$ of radius $\rho > 0$ centered at z_0 and diverges for all $|\, z - z_0 \,| > \rho$ outside the disk. The series may converge at some (but not all) of the points on the boundary of the disk where $|\, z - z_0 \,| = \rho$.

(c) The series only converges, trivially, at $z = z_0$.

The number ρ is known as the *radius of convergence* of the series. In case (a), we say $\rho = \infty$, while in case (c), $\rho = 0$, and the series does *not* represent an analytic function. An example when $\rho = 0$ is the power series $\sum n!\, z^n$. In the intermediate case, determining precisely where on the boundary of the convergence disk the power series converges is quite delicate, and will not be pursued here. The proof of this result can be found in Exercise ∎. See [**4**, **76**] for further details.

Remarkably, the radius of convergence for the power series of a known analytic function $f(z)$ can be determined by inspection, without recourse to any fancy convergence tests! Namely, ρ is equal to the distance from z_0 to the nearest *singularity* of $f(z)$, meaning a

point where the function fails to be analytic. This explains why the Taylor series of e^z converges everywhere, while that of $(z^2+1)^{-1}$ only converges inside the unit disk. Indeed e^z is analytic for all z and has no singularities; therefore the radius of convergence of its power series — centered at any point z_0 — is equal to $\rho = \infty$. On the other hand, the function

$$f(z) = \frac{1}{z^2+1} = \frac{1}{(z+\mathrm{i}\,)(z-\mathrm{i}\,)}$$

has singularities (poles) at $z = \pm \mathrm{i}\,$, and so the series (16.28) has radius of convergence $\rho = 1$, which is the distance from $z_0 = 0$ to the singularities. Therefore, the extension of the theory of power series to the complex plane serves to explain the apparent mystery of why, as a real function, $(1+x^2)^{-1}$ is well-defined and analytic for all real x, but its power series only converges on the interval $(-1,1)$. It is the *complex* singularities that prevent its convergence when $|\,x\,| > 1$! If we expand $(z^2+1)^{-1}$ in a power series at some other point, say $z_0 = 1+2\,\mathrm{i}\,$, then we need to determine which singularity is closest. We compute $|\,\mathrm{i} - z_0\,| = |\,-1 - \mathrm{i}\,| = \sqrt{2}$, while $|\,-\mathrm{i} - z_0\,| = |\,-1 - 3\,\mathrm{i}\,| = \sqrt{10}$, and so $\rho = \sqrt{2}$ is the smaller of these two numbers. Thus we can determine the radius of convergence without any explicit formula for its (rather complicated) Taylor expansion at $z_0 = 1 + 2\,\mathrm{i}\,$.

There are, in fact, only three possible types of singularities of a complex function $f(z)$:

(i) *Pole*. A singular point $z = z_0$ is called a *pole* of *order* $n > 0$ if and only if the function

$$h(z) = (z - z_0)^n\, f(z) \tag{16.29}$$

is analytic and nonzero, $h(z_0) \neq 0$, at $z = z_0$. The simplest example of such a function is $f(z) = a\,(z - z_0)^{-n}$ for $a \neq 0$ a complex constant.

(ii) *Branch point*. We have already encountered the two basic types: *algebraic branch points*, such as the function $\sqrt[n]{z}$ at $z_0 = 0$, and *logarithmic branch points* such as $\log z$ at $z_0 = 0$. The *degree* of the branch point is n in the first case and ∞ in the second.

(iii) *Essential singularity*. By definition, a singularity is *essential* if it is not a pole or a branch point. The simplest example is the essential singularity at $z_0 = 0$ of the function $e^{1/z}$. Details are left as an Exercise ∎.

Example 16.7. For example, the function

$$f(z) = \frac{e^z}{z^3 - z^2 - 5\,z - 3}$$

has a simple (order 1) pole at $z = 3$ and a double (order 2) pole at $z = -1$. Indeed, factorizing the denominator $z^3 - z^2 - 5\,z - 3 = (z+1)^2(z-3)$, we see that the functions

$$h_1(z) = (z-3)\, f(z) = \frac{e^z}{(z+1)^2}\,, \qquad\qquad h_2(z) = (z+1)^2\, f(z) = \frac{e^z}{z-3}\,,$$

are analytic and non-zero at, respectively, $z = 3$ and $z = -1$.

A complex function can have a variety of singularities. For example, the function

$$f(z) = \frac{\sqrt[3]{z+2}\; e^{-1/z}}{z^2+1} \tag{16.30}$$

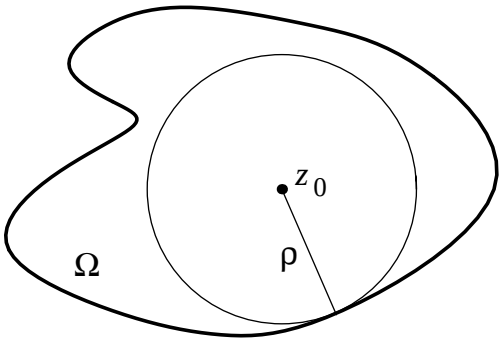


Figure 16.7. Radius of Convergence.

has simple poles at $z = \pm \mathrm{i}$, a branch point of degree 3 at $z = -2$ and an essential singularity at $z = 0$.

As in the real case, and unlike Fourier series, convergent power series can always be repeatedly term-wise differentiated. Therefore, given the convergent series (16.26), we have the corresponding series

$$f'(z) = a_1 + 2\,a_2(z - z_0) + 3\,a_3(z - z_0)^2 + 4\,a_4(z - z_0)^3 + \cdots = \sum_{n=0}^{\infty} (n+1)\,a_{n+1}(z - z_0)^n,$$

$$f''(z) = 2\,a_2 + 6\,a_3(z - z_0) + 12\,a_4(z - z_0)^2 + 20\,a_5(z - z_0)^3 + \cdots = \sum_{n=0}^{\infty} (n+1)(n+2)\,a_{n+2}(z - z_0)^n, \qquad (16.31)$$

and so on, for its derivatives. The proof that the differentiated series have the same radius of convergence can be found in [**4**, **127**]. As a consequence, we deduce the following important result.

Theorem 16.8. *Any analytic function is infinitely differentiable.*

In particular, when we substitute $z = z_0$ into the successively differentiated series, we discover that

$$a_0 = f(z_0), \qquad a_1 = f'(z_0), \qquad a_2 = \tfrac{1}{2}\, f''(z_0),$$

and, in general,

$$a_n = \frac{f^{(n)}(z)}{n!}\,. \qquad (16.32)$$

Therefore, a convergent power series (16.26) is, inevitably, the usual *Taylor series*

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}\,(z - z_0)^n\,, \qquad (16.33)$$

for the function $f(z)$ at the point z_0.

Let us conclude this section by summarizing the fundamental theorem that characterizes complex functions. A complete, rigorous proof relies on complex integration theory, which is the topic of Section 16.5.

Theorem 16.9. *Let $\Omega \subset \mathbb{C}$ be an open set. The following properties are equivalent:*

(a) *The function $f(z)$ has a continuous complex derivative $f'(z)$ for all $z \in \Omega$.*

(b) *The real and imaginary parts of $f(z)$ have continuous partial derivatives and satisfy the Cauchy–Riemann equations* (16.22) *in Ω.*

(c) *The function $f(z)$ is analytic for all $z \in \Omega$, and so is infinitely differentiable and has a convergent power series expansion at each point $z_0 \in \Omega$. The radius of convergence ρ is at least as large as the distance from z_0 to the boundary $\partial\Omega$; see Figure 16.7.*

Any function that satisfies the conditions of Theorem 16.9 will be referred to as a *complex function*. Sometimes one of the equivalent adjectives "analytic" or "holomorphic", is added for emphasis. From now on, all complex functions are assumed to be analytic everywhere on their domain of definition, except, possibly, at certain isolated singularities.

16.3. Harmonic Functions.

We began this section by motivating the analysis of complex functions through applications to the solution of the two-dimensional Laplace equation. Let us now formalize the precise relationship between the two subjects.

Theorem 16.10. *If $f(z) = u(x,y) + \mathrm{i}\, v(x,y)$ is any complex analytic function, then its real and imaginary parts, $u(x,y), v(x,y)$, are both harmonic functions.*

Proof: Differentiating[†] the Cauchy–Riemann equations (16.22), and invoking the equality of mixed partial derivatives, we find that

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial v}{\partial y}\right) = \frac{\partial^2 v}{\partial x\,\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial v}{\partial x}\right) = \frac{\partial}{\partial y}\left(-\frac{\partial u}{\partial y}\right) = -\frac{\partial^2 u}{\partial y^2}\,.$$

Therefore, u is a solution to the Laplace equation $u_{xx} + u_{yy} = 0$. The proof for v is similar. *Q.E.D.*

Thus, every complex function $f = u + \mathrm{i}\, v$ gives rise to two harmonic functions. It is, of course, of interest to know whether we can invert this procedure. Given a harmonic function $u(x,y)$, does there exist a harmonic function $v(x,y)$ such that $f = u + \mathrm{i}\, v$ is a complex analytic function? If so, the harmonic function $v(x,y)$ is known as a *harmonic conjugate* to u. The harmonic conjugate is found by solving the Cauchy–Riemann equations

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}\,, \qquad \frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}\,, \tag{16.34}$$

which, for a prescribed function $u(x,y)$, constitutes an inhomogeneous linear system of partial differential equations for $v(x,y)$. As such, it is usually not hard to solve, as the following example illustrates.

[†] Theorem 16.9 allows us to differentiate u and v as often as desired.

Example 16.11. As the reader can verify, the harmonic polynomial

$$u(x,y) = x^3 - 3x^2y - 3xy^2 + y^3$$

satisfies the Laplace equation everywhere. To find a harmonic conjugate, we solve the Cauchy–Riemann equations (16.34). First of all,

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} = 3x^2 + 6xy - 3y^2,$$

and hence, by direct integration with respect to x,

$$v(x,y) = x^3 + 3x^2y - 3xy^2 + h(y),$$

where $h(y)$ — the "constant of integration" — is a function of y alone. To determine h we substitute our formula into the second Cauchy–Riemann equation:

$$3x^2 - 6xy + h'(y) = \frac{\partial v}{\partial y} = \frac{\partial u}{\partial x} = 3x^2 - 6xy - 3y^2.$$

Therefore, $h'(y) = -3y^2$, and so $h(y) = -y^3 + c$, where c is a real constant. We conclude that every harmonic conjugate to $u(x,y)$ has the form

$$v(x,y) = x^3 + 3x^2y - 3xy^2 - y^3 + c.$$

Note that the corresponding complex function

$$\begin{aligned} u(x,y) + \mathrm{i}\, v(x,y) &= (x^3 - 3x^2y - 3xy^2 + y^3) + \mathrm{i}\,(x^3 + 3x^2y - 3xy^2 - y^3 + c) \\ &= (1 - \mathrm{i}\,)z^3 + c \end{aligned}$$

is a particular complex cubic polynomial.

Remark: On a connected domain, all harmonic conjugates to a given function $u(x,y)$ only differ by a constant: $\widetilde{v}(x,y) = v(x,y) + c$; see Exercise ■.

Although most harmonic functions have harmonic conjugates, unfortunately this is not always the case. Interestingly, the existence or non-existence of a harmonic conjugate can depend on the underlying geometry of the domain of definition of the function. If the domain is simply-connected, and so contains no holes, then one can *always* find a harmonic conjugate. In fact, this is an immediate consequence of our characterization of potential functions in Appendix A. Otherwise, if the domain of definition Ω of our harmonic function $u(x,y)$ is not simply-connected, then there may not exist a single-valued harmonic conjugate $v(x,y)$ to serve as the imaginary part of a complex function $f(z)$.

Example 16.12. The simplest example where the latter possibility occurs is the logarithmic potential

$$u(x,y) = \log r = \tfrac{1}{2}\,\log(x^2 + y^2).$$

This function is harmonic on the non-simply-connected domain $\Omega = \mathbb{C} \setminus \{0\}$, but it is not the real part of any single-valued complex function. Indeed, according to (16.18), the logarithmic potential is the real part of the multiply-valued complex logarithm $\log z$, and

so its harmonic conjugate[†] is $\operatorname{ph} z = \theta$, which cannot be consistently and continuously defined on all of Ω. On the other hand, it is possible to choose a continuous, single-valued branch of the angle $\theta = \operatorname{ph} z$ if z is restricted to a simply connected subdomain $\widetilde{\Omega}$, and so $\log r$ does have a genuine harmonic conjugate on $\widetilde{\Omega}$.

The harmonic function

$$u(x,y) = \frac{x}{x^2+y^2}$$

is also defined on the same non-simply-connected domain $\Omega = \mathbb{C} \setminus \{0\}$ with a singularity at $x = y = 0$. In this case, there is a single valued harmonic conjugate, namely

$$v(x,y) = -\frac{y}{x^2+y^2}\,,$$

which is defined on all of Ω. Indeed, according to (16.12), these functions define the real and imaginary parts of the complex function $u + \mathrm{i}\, v = 1/z$. Alternatively, one can directly check that they satisfy the Cauchy–Riemann equations (16.22).

Remark: On the "punctured" plane $\Omega = \mathbb{C} \setminus \{0\}$, the logarithmic potential is, in a sense, the only counterexample that prevents a harmonic conjugate from being constructed. It can be shown, [**XC**], that if $u(x,y)$ is a harmonic function defined on a punctured disk $\Omega_R = \big\{\, 0 < |\, z \,| < R \,\big\}$, where $0 < R \leq \infty$, then there exists a constant c such that $\widetilde{u}(x,y) = u(x,y) - c \log r$ is also harmonic and possess a single-valued harmonic conjugate $\widetilde{v}(x,y)$. As a result, the function $\widetilde{f} = \widetilde{u} + \mathrm{i}\,\widetilde{v}$ is analytic on all of Ω_R, and so our original function $u(x,y)$ is the real part of the multiply-valued analytic function $f(z) = \widetilde{f}(z) + c \log z$. We shall use this fact in our later analysis of airfoils.

Theorem 16.13. *Every harmonic function $u(x,y)$ defined on a simply-connected domain Ω is the real part of a complex valued function $f(z) = u(x,y) + \mathrm{i}\, v(x,y)$ which is defined for all $z = x + \mathrm{i}\, y \in \Omega$.*

Proof: We first rewrite the Cauchy–Riemann equations (16.34) in vectorial form as an equation for the gradient of v:

$$\nabla v = \nabla u^{\perp}, \qquad \text{where} \qquad \nabla u^{\perp} = \begin{pmatrix} -u_y \\ u_x \end{pmatrix} \tag{16.35}$$

is the vector field that is everywhere orthogonal to the gradient of u and of the same length[†]:

$$\nabla u^{\perp} \cdot \nabla u = 0, \qquad\qquad \| \nabla u^{\perp} \| = \| \nabla u \|.$$

Thus, we have established the important observation that the gradient of a harmonic function and that of its harmonic conjugate are mutually orthogonal vector fields:

$$\nabla v \cdot \nabla u \equiv 0. \tag{16.36}$$

[†] We can, by a previous remark, add in any constant to the harmonic conjugate, but this does not affect the subsequent argument.

[†] Since we are working in $\mathbb{R}^2$, these properties along with the right hand rule serve to uniquely characterize $\nabla u^{\perp}$.

Now, according to Theorem A.8, provided we work on a simply-connected domain, the gradient equation

$$\nabla v = \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

has a solution if and only if the vector field $\mathbf{f}$ satisfies the curl-free constraint

$$\nabla \wedge \mathbf{f} = \frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y} \equiv 0.$$

IN our specific case, the curl of the perpendicular vector field $\nabla u^\perp$ coincides with the divergence of ∇u itself, which, in turn, coincides with the Laplacian:

$$\nabla \wedge \nabla u^\perp = \nabla \cdot \nabla u = \Delta u = 0, \quad \text{i.e.,} \qquad \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(-\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

The result is zero because we are assuming that u is harmonic. Equation (22.26) permits us to reconstruct the harmonic conjugate $v(x, y)$ from its gradient ∇v through line integration

$$v(x, y) = \int_C \nabla v \cdot d\mathbf{x} = \int_C \nabla u^\perp \cdot d\mathbf{x} = \int_C \nabla u \cdot \mathbf{n}\, ds, \tag{16.37}$$

where C is any curve connecting a fixed point (x_0, y_0) to (x, y). Therefore, the harmonic conjugate to a given potential function u can be obtained by evaluating its (path-independent) flux integral (16.37). *Q.E.D.*

Remark: As a consequence of (16.23) and the Cauchy–Riemann equations (16.34),

$$f'(z) = \frac{\partial u}{\partial x} - \mathrm{i}\, \frac{\partial u}{\partial y} = \frac{\partial v}{\partial y} + \mathrm{i}\, \frac{\partial v}{\partial x}\,. \tag{16.38}$$

Thus, the components of the gradients ∇u and ∇v appear as the real and imaginary parts of the complex derivative $f'(z)$.

The orthogonality (16.35) of the gradient of a function and of its harmonic conjugate has the following important geometric consequence. Recall, Theorem A.14, that the gradient ∇u of a function points in the normal direction to its level curves $\{\, u(x, y) = c \,\}$. Since ∇v is orthogonal to ∇u, this must mean that ∇v is *tangent* to the level curves of u. Vice versa, ∇v is normal to its level curves, and so ∇u is tangent to the level curves of its harmonic conjugate v. Since their tangent directions ∇u and ∇v are orthogonal, the level curves of the real and imaginary parts of a complex function form a mutually orthogonal system of plane curves — but with one key exception. If we are at a *critical point*, where $\nabla u = \mathbf{0}$, then $\nabla v = \nabla u^\perp = \mathbf{0}$, and the vectors do not define tangent directions. Therefore, the orthogonality of the level curves does not necessarily hold at critical points. It is worth pointing out that, in view of (16.38), the critical points of u are the same as those of v and also the same as the critical points of the corresponding complex function $f(z)$, i.e., where its complex derivative vanishes: $f'(z) = 0$.

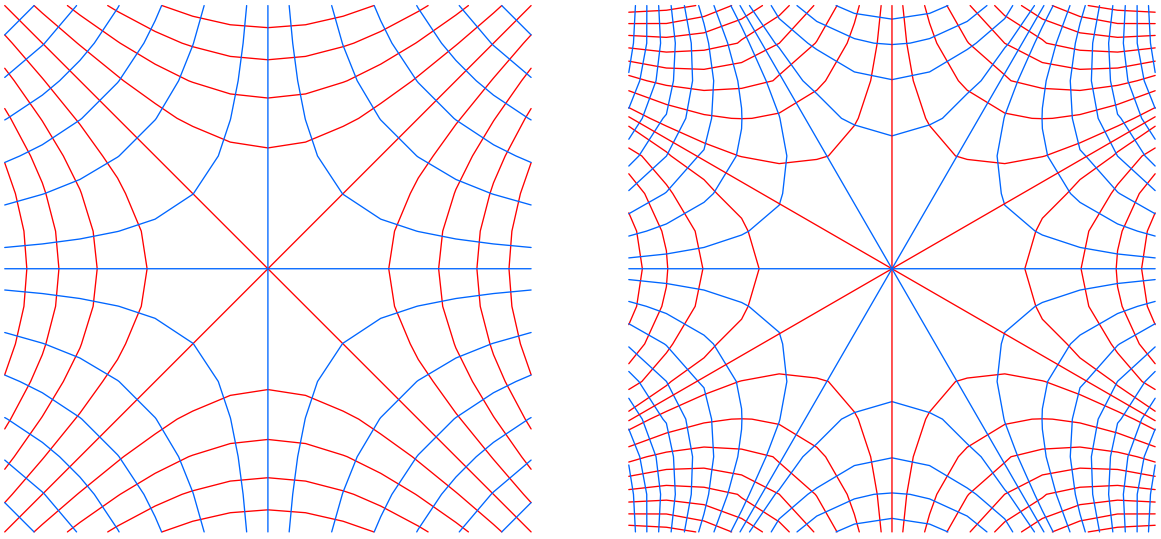


Figure 16.8. Level Curves of the Real and Imaginary Parts of z^2 and z^3.

In Figure 16.8, we illustrate the preceding discussion by plotting the level curves of the real and imaginary parts of the monomials z^2 and z^3. Note that, except at the origin, where the derivative vanishes, the level curves intersect everywhere at right angles.

Applications to Fluid Mechanics

Consider a planar[†] steady state fluid flow, with velocity vector field

$$\mathbf{v}(\mathbf{x}) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} \qquad \text{at the point} \quad \mathbf{x} = (x,y) \in \Omega.$$

Here $\Omega \subset \mathbb{R}^2$ is the domain occupied by the fluid, while the vector $\mathbf{v}(\mathbf{x})$ represents the instantaneous velocity of the fluid at the point $\mathbf{x}$. In many physical situations, the flow of liquids (and, although less often, gases) is both incompressible and irrotational, which for short, are known as *ideal fluid flows*. Recall that the flow is *incompressible* if and only if it has vanishing divergence:

$$\nabla \cdot \mathbf{v} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0. \tag{16.39}$$

On the other hand, the flow is *irrotational* if and only if it has vanishing curl:

$$\nabla \wedge \mathbf{v} = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = 0. \tag{16.40}$$

The two constraints (16.39), (16.40) are almost identical to the Cauchy–Riemann equations (16.22)! The only difference is the sign in front of the derivatives of v, but this can be easily remedied by replacing v by its negative $-v$. As a result, we deduce the connection between ideal planar fluid flows and complex functions.

[†] See the remarks in Appendix A on the interpretation of a planar fluid flow as the cross-section of a fully three-dimensional fluid motion that does not depend upon the vertical coordinate.

Theorem 16.14. *The vector field* $\mathbf{v} = (\, u(x,y), v(x,y)\,)^T$ *is the velocity vector of an ideal fluid flow if and only if*

$$f(z) = u(x,y) - \mathrm{i}\, v(x,y) \tag{16.41}$$

is a complex analytic function of $z = x + \mathrm{i}\, y$.

Therefore, the components $u(x,y)$ and $-v(x,y)$ of the velocity vector field for an ideal fluid are harmonic conjugates. The complex function (16.41) is known as the *complex velocity* of the fluid flow. When using this result, *do not forget* the minus sign that appears in front of the imaginary part of $f(z)$.

As in Example A.7, the fluid particles will follow the curves[†] $z(t) = x(t) + \mathrm{i}\, y(t)$ obtained by integrating the differential equations

$$\frac{dx}{dt} = u(x,y), \qquad \frac{dy}{dt} = v(x,y), \tag{16.42}$$

which, in view of (16.41), we can rewrite in complex form

$$\frac{dz}{dt} = \overline{f(z)}\,. \tag{16.43}$$

Each fluid particle's motion $z(t)$ is uniquely prescribed by its initial position $z(0) = z_0 = x_0 + \mathrm{i}\, y_0$ at time $t = 0$. The curves parametrized by $z(t)$ are the paths followed by the particles, i.e., the *streamlines* of the flow. In particular, if the complex velocity vanishes, $f(z_0) = 0$, then the solution $z(t) \equiv z_0$ to (16.43) is constant, and hence z_0 is a *stagnation point* of the flow.

Example 16.15. The simplest example is when the velocity is constant, corresponding to a uniform steady flow. Consider first the case

$$f(z) = 1,$$

which corresponds to the horizontal velocity vector field $\mathbf{v} = (\, 1, 0\,)^T$. The actual fluid flow is found by integrating the system

$$\dot{z} = 1, \qquad \text{or} \qquad \dot{x} = 1, \quad \dot{y} = 0.$$

Thus, the solution $z(t) = t + z_0$ represents a uniform horizontal fluid motion whose streamlines are straight lines parallel to the real axis; see Figure 16.9.

Consider next a more general constant velocity

$$f(z) = c = a + \mathrm{i}\, b.$$

The fluid particles will solve the ordinary differential equation

$$\dot{z} = \overline{c} = a - \mathrm{i}\, b, \qquad \text{so that} \qquad z(t) = \overline{c}\, t + z_0.$$

[†] See below for more details on complex curves.

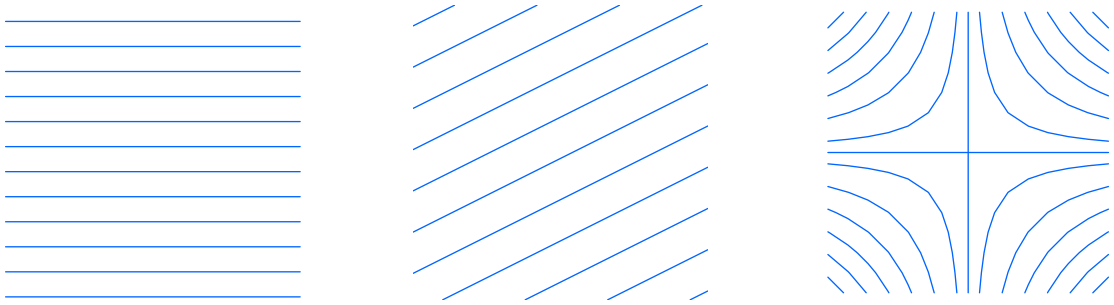


Figure 16.9. Complex Fluid Flows.

The streamlines remain parallel straight lines, but now at an angle $\theta = \operatorname{ph} \overline{c} = -\operatorname{ph} c$ with the horizontal. The fluid particles move along the streamlines at constant speed $|\,\overline{c}\,| = |\,c\,|$.

The next simplest complex velocity function is

$$f(z) = z = x + \mathrm{i}\, y. \tag{16.44}$$

The corresponding fluid flow is found by integrating the system

$$\dot{z} = \overline{z}, \qquad \text{or, in real form,} \qquad \dot{x} = x, \qquad \dot{y} = -y.$$

The origin $x = y = 0$ is a stagnation point. The trajectories of the nonstationary solutions

$$z(t) = x_0\, e^{t} + \mathrm{i}\, y_0\, e^{-t} \tag{16.45}$$

are the hyperbolas $x\,y = c$ and the positive and negative semi-axes, as illustrated in Figure 16.9.

On the other hand, if we choose

$$f(z) = -\,\mathrm{i}\, z = y - \mathrm{i}\, x,$$

then the flow is the solution to

$$\dot{z} = \mathrm{i}\,\overline{z}, \qquad \text{or, in real form,} \qquad \dot{x} = y, \qquad \dot{y} = x.$$

The solutions

$$z(t) = (x_0 \cosh t + y_0 \sinh t) + \mathrm{i}\,(x_0 \sinh t + y_0 \cosh t),$$

move along the hyperbolas (and rays) $x^2 - y^2 = c^2$. Thus, this flow is obtained by rotating the preceding example by 45°.

Example 16.16. A solid object in a fluid flow is characterized by the no-flux condition that the fluid velocity $\mathbf{v}$ is everywhere tangent to the boundary, and hence no fluid flows into or out of the object. As a result, the boundary will consist of streamlines and stagnation points of the idealized fluid flow. For example, the boundary of the upper right quadrant $Q = \{\, x > 0, y > 0 \,\} \subset \mathbb{C}$ consists of the positive x and y axes (along with the origin). Since these are streamlines of the flow with complex velocity (16.44), its restriction

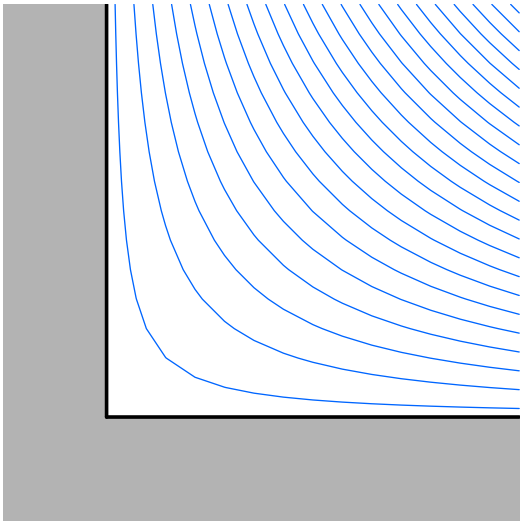


Figure 16.10. Flow Inside a Corner.

to Q represents the flow past a 90° interior corner, which appears in Figure 16.10. The fluid particles move along hyperbolas as they flow past the corner.

Remark: We could also restrict this flow to the domain $\Omega = \mathbb{C} \setminus \{ x < 0, y < 0 \}$ consisting of three quadrants, and corresponding to a 90° exterior corner. However, the restricted flow is not as relevant in this case since it does not have a physically realizable asymptotic behavior at large distances. See Exercise ■ for the "correct" physical flow around an exterior corner.

Now, suppose that the complex velocity $f(z)$ admits a complex anti-derivative, i.e., a complex analytic function

$$\chi(z) = \varphi(x,y) + \mathrm{i}\,\psi(x,y) \qquad \text{that satisfies} \qquad \frac{d\chi}{dz} = f(z). \tag{16.46}$$

Using the formula (16.23) for the complex derivative, we see that

$$\frac{d\chi}{dz} = \frac{\partial \varphi}{\partial x} - \mathrm{i}\,\frac{\partial \varphi}{\partial y} = u - \mathrm{i}\,v, \qquad \text{so} \qquad \frac{\partial \varphi}{\partial x} = u, \qquad \frac{\partial \varphi}{\partial y} = v.$$

Thus, $\nabla \varphi = \mathbf{v}$, and hence the real part $\varphi(x,y)$ of the complex function $\chi(z)$ defines a *velocity potential* for the fluid flow. For this reason, the anti-derivative (16.46) is known as the *complex potential function* for the given fluid velocity field.

Since the complex potential is analytic, its real part, the potential function, is harmonic and therefore satisfies the Laplace equation $\Delta \varphi = 0$. Conversely, any harmonic function can be viewed as the potential function for some fluid flow. The real fluid velocity is its gradient $\mathbf{v} = \nabla \varphi$. The harmonic conjugate $\psi(x,y)$ to the velocity potential also plays an important role, and, in fluid mechanics, is known as the *stream function* for the fluid flow. It also satisfies the Laplace equation $\Delta \psi = 0$, and the potential and stream function are related by the Cauchy–Riemann equations (16.22).

The level curves of the velocity potential, $\varphi(x,y) = c$, are known as *equipotential curves* for the flow. The velocity vector $\mathbf{v} = \nabla \varphi$ points in the normal direction to the

equipotentials. On the other hand, as we noted above, $\mathbf{v} = \nabla\varphi$ is tangent to the level curves $\psi(x,y) = d$ of its harmonic conjugate stream function. But $\mathbf{v}$ is the velocity field, and so tangent to the streamlines followed by the fluid particles. Thus, these two systems of curves must coincide, and we infer that *the level curves of the stream function are the streamlines of the flow*, whence the name "stream function"! Summarizing, for an ideal fluid flow, the equipotentials $\{\,\varphi = c\,\}$ and streamlines $\{\,\psi = d\,\}$ form mutually orthogonal systems of plane curves. The fluid velocity $\mathbf{v} = \nabla\varphi$ is tangent to the stream lines and normal to the equipotentials, whereas the gradient of the stream function $\nabla\psi$ is tangent to the equipotentials and normal to the streamlines.

The discussion in the preceding paragraph implicitly relied on the fact that the velocity is nonzero, $\mathbf{v} = \nabla\varphi \neq 0$, which means we are not at a *stagnation point*, where the fluid is not moving. While streamlines and equipotentials might begin or end at a stagnation point, there is no guarantee, and, indeed, in general it is not the case that they meet at mutually orthogonal directions there.

Example 16.17. The simplest example of a complex potential function is

$$\chi(z) = z = x + \mathrm{i}\, y.$$

Thus, the velocity potential is $\varphi(x,y) = x$, while its harmonic conjugate stream function is $\psi(x,y) = y$. The complex derivative of the potential is the complex velocity,

$$f(z) = \frac{d\chi}{dz} = 1,$$

which corresponds to the uniform horizontal fluid motion considered first in Example 16.15. Note that the horizontal stream lines coincide with the level sets $y = k$ of the stream function, whereas the equipotentials $\varphi = x = c$ are the orthogonal system of vertical lines.

Next, consider the complex potential function

$$\chi(z) = \tfrac{1}{2}\, z^2 = \tfrac{1}{2}\,(x^2 - y^2) + \mathrm{i}\, x\, y.$$

The complex velocity function

$$f(z) = \chi'(z) = z = x + \mathrm{i}\, y$$

leads to the hyperbolic flow (16.45). The hyperbolic streamlines $x\, y = d$ are the level curves of the stream function $\psi(x,y) = x\, y$. The equipotential lines $\frac{1}{2}\,(x^2 - y^2) = c$ form a system of orthogonal hyperbolas. A picture of the equipotentials and stream lines in this particular case can be found in the first plot in Figure 16.8.

Example 16.18. *Flow Around a Disk.* Consider the complex potential function

$$\chi(z) = z + \frac{1}{z} = \left(x + \frac{x}{x^2 + y^2} \right) + \mathrm{i}\, \left(y - \frac{y}{x^2 + y^2} \right). \tag{16.47}$$

The corresponding complex fluid velocity is

$$f(z) = \frac{d\chi}{dz} = 1 - \frac{1}{z^2} = 1 - \frac{x^2 - y^2}{(x^2 + y^2)^2} \;+\; \mathrm{i}\, \frac{2\, x\, y}{(x^2 + y^2)^2}\,. \tag{16.48}$$

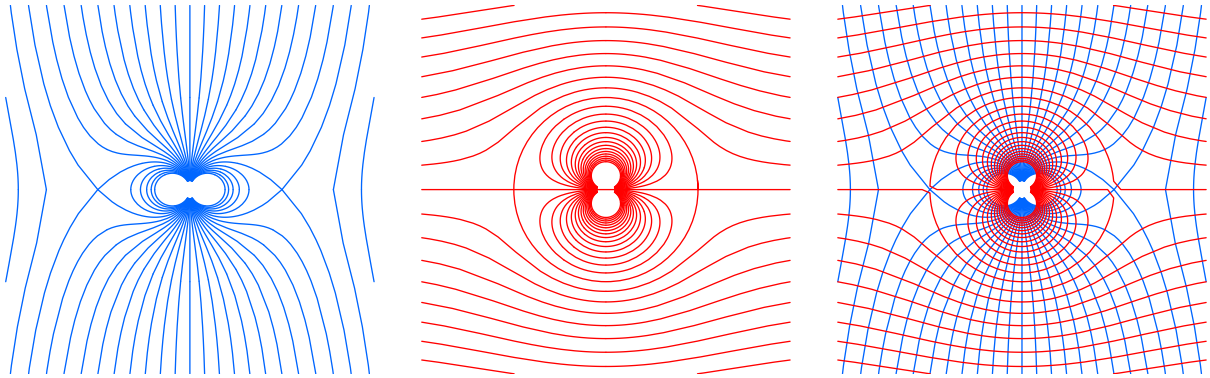


Figure 16.11. Equipotentials and Streamlines for $z + \dfrac{1}{z}$.

The equipotential curves and streamlines are plotted in Figure 16.11. The points $z = \pm 1$ are stagnation points of the flow, while $z = 0$ is a singularity. In particular, fluid particles that move along the positive x axis approach the leading stagnation point $z = 1$, but take an infinite amount of time to reach it. Note that at large distances, the streamlines

$$\psi(x, y) = y - \frac{y}{x^2 + y^2} = d$$

are asymptotically horizontal, and hence, far away from the origin, the flow is indistinguishable from uniform horizontal motion with complex velocity $f(z) \equiv 1$. The level curve for the particular value $d = 0$ consists of the unit circle $|\, z \,| = 1$ and the real axis $y = 0$. In particular, the unit circle $|\, z \,| = 1$ consists of two stream lines and the two stagnation points. Therefore, the flow velocity vector field $\mathbf{v} = \nabla \varphi$ is everywhere tangent to the unit circle, and hence satisfies the no flux condition on the boundary of the unit disk. Thus, we can interpret (16.48), when restricted to the domain $\Omega = \big\{\, |\, z \,| > 1 \,\big\}$, as the complex velocity of a uniformly moving fluid around the outside of a solid circular disk of radius 1. In three dimensions, this would correspond to the steady flow of a fluid around a solid cylinder.

In this section, we have focussed on the fluid mechanical roles of a harmonic function and its conjugate. An analogous interpretation applies when $\varphi(x, y)$ represents an electromagnetic potential function; the level curves of its harmonic conjugate $\psi(x, y)$ are the paths followed by charged particles under the electromotive force field $\mathbf{v} = \nabla \varphi$. Similarly, if $\varphi(x, y)$ represents the equilibrium temperature distribution in a planar domain, its level lines represent the isotherms or curves of constant temperature, while the level lines of its harmonic conjugate are the curves of heat flow, whose mutual orthogonality was already noted in Appendix A. Finally, if $\varphi(x, y)$ represents the height of a deformed membrane, then its level curves are the contour lines of elevation. The level curves of its harmonic conjugate are the curves of steepest descent along the membrane, i.e., the routes followed by, say, water flowing down the membrane.

16.4. Conformal Mapping.

As we now know, complex functions provide an almost inexhaustible source of harmonic functions, i.e., solutions to the Laplace equation. Thus, to solve a boundary value

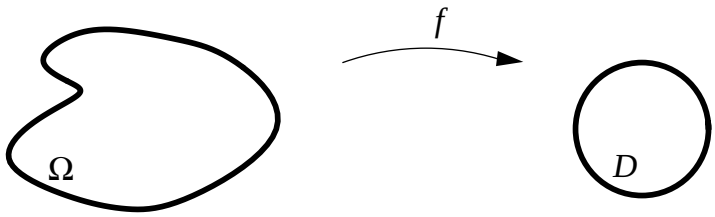


Figure 16.12. Mapping to the Unit Disk.

problem for Laplace's equation we merely need to find the right complex function whose real part matches the prescribed boundary conditions. Unfortunately, even for relatively simple domains, this is still not a particularly easy task. The one case where we do have an explicit solution is that of a circular disk, where the Poisson integral formula (15.44) provides a complete solution to the Dirichlet boundary value problem. (See Exercise ■ for the Neumann and mixed boundary value problems.) However, determining the corresponding integral formula or Green's function for a more complicated domain remains a daunting task, even with the relatively powerful tools of complex analysis at our disposal.

There is, however, a wonderful idea that will go very far towards this general goal. Given that we know how to solve a boundary value problem on one particular domain, the unit disk

$$D = \big\{\, \zeta = \xi + \mathrm{i}\,\eta \;\big|\; \xi^2 + \eta^2 < 1 \,\big\},$$

perhaps we can make an inspired change of variables that will convert the unsolved boundary value problem on Ω into one that we know how to solve on D. In other words, we seek a pair of functions

$$\xi = p(x,y), \qquad\qquad \eta = q(x,y), \tag{16.49}$$

that maps each point $(x,y) \in \Omega$ to a point $(\xi,\eta) \in D$ in the unit disk, as illustrated in Figure 16.12. The desired mapping must satisfy fairly stringent requirements.

(*a*) First of all, it should be one-to-one, and so each point $(x,y) \in \Omega$ maps to a unique point in $(\xi,\eta) = (p(x,y), q(x,y)) \in D$. Under these conditions, each function $U(\xi,\eta)$ defined on the unit disk will correspond to a unique function

$$u(x,y) = U\big(\, p(x,y), q(x,y) \,\big) \tag{16.50}$$

defined on the domain Ω, whose value at the point (x,y) to equal the value of U at the image point $(\xi,\eta) = (p(x,y), q(x,y))$.

(*b*) Secondly, both the map (16.49) and its inverse

$$x = P(\xi,\eta), \qquad\qquad y = Q(\xi,\eta), \tag{16.51}$$

should be sufficiently smooth so as to allow us to take derivatives of the functions $u(x,y)$ and $U(\xi,\eta)$. The Inverse Function Theorem, cf. [**9**], requires that the Jacobian determinant

$$\frac{\partial(\xi,\eta)}{\partial(x,y)} = \det \begin{pmatrix} \dfrac{\partial \xi}{\partial x} & \dfrac{\partial \xi}{\partial y} \\[2ex] \dfrac{\partial \eta}{\partial x} & \dfrac{\partial \eta}{\partial y} \end{pmatrix} \neq 0 \tag{16.52}$$

be everywhere non-zero in the domain Ω. Incidentally, the Jacobian condition is enough to ensure that the map is locally (but not necessarily globally) one-to-one.

(*c*) Moreover, the map (16.49) should extend continuously to the boundary $\partial\Omega$, mapping it to the boundary of the unit disk $\partial D = C = \big\{\, \xi^2 + \eta^2 = 1 \,\big\}$, which is the unit circle. This will ensure that a boundary value problem for $u(x,y)$ on Ω is mapped to a boundary value problem for $U(\xi,\eta)$ on D.

(*d*) Finally, we must ensure that if $U(\xi,\eta)$ satisfies the Laplace equation

$$\Delta U = U_{\xi\xi} + U_{\eta\eta} = 0 \qquad \text{on} \qquad D,$$

then $u(x,y)$ as given by (16.50) will satisfy the Laplace equation

$$\Delta u = u_{xx} + u_{yy} = 0 \qquad \text{on} \qquad \Omega.$$

Otherwise, the proposed mapping will be of scant help for solving the boundary value problem under consideration. The latter requirement is, without extra insight, quite hard to ensure.

Example 16.19. The scaling change of variables

$$\xi = a\,x, \qquad\qquad \eta = b\,y \tag{16.53}$$

changes the elliptical domain $\Omega = \big\{\, a^2\,x^2 + b^2\,y^2 < 1 \,\big\}$ to the unit disk $D = \{\, \xi^2 + \eta^2 < 1 \,\}$. However, it is not of much help for solving the Laplace equation on the elliptical domain. Indeed, when we relate a function $U(\xi,\eta)$ on D to

$$u(x,y) = U(a\,x, b\,y)$$

on Ω, the partial derivatives are related by

$$\frac{\partial^2 u}{\partial x^2} = a^2\,\frac{\partial^2 U}{\partial \xi^2}\,, \qquad\qquad \frac{\partial^2 u}{\partial y^2} = b^2\,\frac{\partial^2 U}{\partial \eta^2}\,.$$

If U is harmonic, so $\Delta U = U_{\xi\xi} + U_{\eta\eta} = 0$, then $u(x,y)$ satisfies the partial differential equation

$$\frac{1}{a^2}\,\frac{\partial^2 u}{\partial x^2} + \frac{1}{b^2}\,\frac{\partial^2 u}{\partial y^2} = 0. \tag{16.54}$$

Unless $a = b$ — in which case the domain Ω is a circle and we are performing a simple scaling transformation — the function $u(x,y)$ is *not* a solution to the Laplace equation on Ω. Be that as it may, this change of variables *does* provide a means of solving the Dirichlet boundary value problem for the elliptic partial differential equation (16.54) on the elliptical domain Ω.

Analytic Maps

The crucial insight that makes the change of variables idea so effective is that complex analytic functions not only provide harmonic functions as candidate solutions to the Laplace equation, they also provide a large class of mappings that accomplish the desired goals. The method rests on the simple fact that the composition of two complex analytic functions is also complex analytic.

Lemma 16.20. *If $w = F(\zeta)$ is an analytic function of the complex variable $\zeta = \xi + \mathrm{i}\,\eta$ and $\zeta = g(z)$ is an analytic function of the complex variable $z = x + \mathrm{i}\,y$, then the composition*[†] *$w = f(z) \equiv F \circ g(z) = F(g(z))$ is an analytic function of z.*

Proof: The proof that the composition of two differentiable functions is differentiable is identical to the real variable version, [**9**, **136**], and need not be reproduced here. The derivative of the composition is explicitly given by the usual chain rule:

$$\frac{d}{dz}\, F \circ g(z) = F'(g(z))\, g'(z), \quad \text{or, in Leibnizian notation,} \quad \frac{dw}{dz} = \frac{dw}{d\zeta}\,\frac{d\zeta}{dz}\,. \qquad Q.E.D.$$

We interpret a complex function

$$\zeta = g(z) \qquad \text{or} \qquad \xi + \mathrm{i}\,\eta = p(x,y) + \mathrm{i}\,q(x,y) \tag{16.55}$$

as a *mapping*, as in (16.49), that takes a point $z = x + \mathrm{i}\,y \in \Omega$ belonging to a prescribed domain $\Omega \subset \mathbb{C}$ to a point $\zeta = \xi + \mathrm{i}\,\eta \in D$ belonging to the image domain $D = g(\Omega) \subset \mathbb{C}$. Based on our earlier comments, we will make three important assumptions:

(*a*) The analytic mapping is one-to-one. In other words, we assume that each point $\zeta \in D$ comes from a unique point $z \in \Omega$, and so the inverse function $z = g^{-1}(\zeta)$ is a well-defined map from D back to Ω.

(*b*) The inverse mapping $g^{-1}(\zeta)$ is analytic on all of D. Recall that the derivative of the inverse function is given by

$$\frac{d}{d\zeta}\, g^{-1}(\zeta) = \frac{1}{g'(z)} \qquad \text{at} \qquad \zeta = g(z). \tag{16.56}$$

This formula, which is equally valid for complex functions, implies that the derivative of $g(z)$ must be nonzero everywhere in order that $g^{-1}(\zeta)$ be differentiable. This condition

$$g'(z) \neq 0 \qquad \text{at every point} \qquad z \in \Omega, \tag{16.57}$$

will play a crucial role in the development of the method.

(*c*) The mapping extends continuously to the boundary $\delta\Omega$ and maps it to the boundary ∂D of the image domain.

Before trying to apply these techniques to solve boundary value problems for the Laplace equation, we consider some of the most important examples of analytic maps.

Example 16.21. The simplest nontrivial analytic maps are the *translations*

$$\zeta = z + c = (x + a) + \mathrm{i}\,(y + b), \tag{16.58}$$

which translates the entire complex plane in the direction given by $c = a + \mathrm{i}\,b$. These are the complex counterparts of the affine translations (7.25) of a vector space. The effect is to map a disk $|\, z + c\,| < 1$ of radius 1 and center at $-c$ to the unit disk $|\,\zeta\,| < 1$.

[†] Of course, to properly define the composition, we need to ensure that the range of the function $\zeta = g(z)$ is contained in the domain of the function $w = f(\zeta)$.

There are two types of linear analytic transformations. First, we have the scaling map

$$\zeta = \rho\, z = \rho\, x + \mathrm{i}\, \rho\, y, \tag{16.59}$$

where $\rho \neq 0$ is a fixed nonzero real number. These map the disk $|\, z \,| < 1/|\, \rho \,|$ to the unit disk $|\, \zeta \,| < 1$. Second are the rotations

$$\zeta = e^{\,\mathrm{i}\,\varphi}\, z = (x \cos \varphi - y \sin \varphi) + \mathrm{i}\, (x \sin \varphi + y \cos \varphi) \tag{16.60}$$

around the origin by a fixed (real) angle φ. These map the unit disk to itself.

Any non-constant *affine transformation*

$$\zeta = \alpha\, z + \beta, \qquad\qquad \alpha \neq 0, \tag{16.61}$$

defines an invertible analytic map on all of $\mathbb{C}$, whose inverse map $z = \dfrac{\zeta - \beta}{\alpha}$ is also affine. Writing $\alpha = \rho\, e^{\,\mathrm{i}\,\varphi}$ in polar coordinates, we see that the affine map (16.61) can be built up from a translation, a scaling and a rotation. As such, it takes the disk $|\, \alpha\, z + \beta \,| < 1$ of radius $1/|\, \alpha \,| = 1/|\, \rho \,|$ and center $-\beta/\alpha$ to the unit disk $|\, \zeta \,| < 1$. As such, none of these maps take us to a radically new class of boundary value problems.

Example 16.22. A more interesting complex function is

$$\zeta = g(z) = \frac{1}{z}\,, \qquad \text{or} \qquad \xi = \frac{x}{x^2 + y^2}\,, \qquad \eta = -\,\frac{y}{x^2 + y^2}\,, \tag{16.62}$$

which is known as an *inversion*[†] of the complex plane. It defines a one-to-one analytic map everywhere except at the origin $z = 0$; indeed $g(z)$ is its own inverse: $g^{-1}(\zeta) = 1/\zeta$. Note that $g'(z) = -1/z^2$ is never zero, and so the derivative condition (16.57) is satisfied everywhere. Thus, any domain $\Omega \subset \mathbb{C} \setminus \{0\}$ will be mapped in a one-to-one manner onto an image domain $D = g(\Omega) \subset \mathbb{C} \setminus \{0\}$.

Note that $|\, \zeta \,| = 1/|\, z \,|$, while $\mathrm{ph}\, \zeta = -\mathrm{ph}\, z$. Thus, if $\Omega = \big\{\, |\, z \,| > \rho \,\big\}$ denotes the exterior of the circle of radius ρ, then the image points $\zeta = 1/z$ satisfy $|\, \zeta \,| = 1/|\, z \,|$, and hence the image domain is the *punctured disk* $D = \big\{\, 0 < |\, \zeta \,| < 1/\rho \,\big\}$. In particular, the inversion maps the outside of the unit disk to its inside, but with the origin removed, and vice versa. The reader may enjoy seeing what the inversion does to other domains, e.g., the unit square.

Example 16.23. The complex exponential

$$\zeta = g(z) = e^z\,, \qquad\qquad \text{or} \qquad\qquad \xi = e^x \cos y, \qquad \eta = e^x \sin y, \tag{16.63}$$

satisfies the condition $g'(z) = e^z \neq 0$ everywhere. Nevertheless, it is *not* one-to-one because $e^{z + 2\,\pi\,\mathrm{i}} = e^z$, and so all points differing by an integer multiple of $2\,\pi\,\mathrm{i}$ are mapped to the same point.

Under the exponential map (16.63), the horizontal line $\mathrm{Im}\; z = b$ is mapped to the curve $\zeta = e^{x + \mathrm{i}\, b} = e^x(\cos b + \mathrm{i}\, \sin b)$, which, as x varies from $-\infty$ to ∞, traces out the ray

[†] This is slightly different than the real inversion (15.68); see Exercise ■.

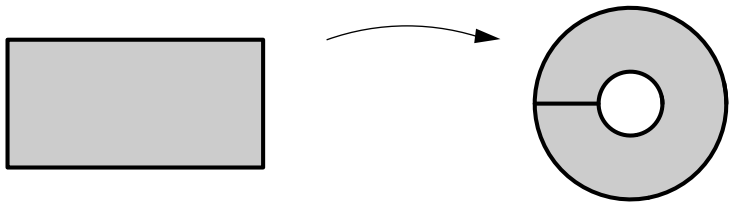


Figure 16.13. The mapping $\zeta = e^z$.

emanating from the origin that makes an angle $\operatorname{ph} \zeta = b$ with the real axis. Therefore, the exponential map will map a horizontal strip $S_{a,b} = \{\, a < \operatorname{Im} z < b \,\}$ to a wedge-shaped domain $\Omega_{a,b} = \{\, a < \operatorname{ph} \zeta < b \,\}$, and is one-to-one provided $|\, b - a \,| < 2\pi$. In particular, the horizontal strip $S_{-\pi/2,\pi/2} = \{\, -\frac{1}{2}\pi < \operatorname{Im} z < \frac{1}{2}\pi \,\}$ of width π centered around the real axis is mapped, in a one-to-one manner, to the right half plane

$$R = \Omega_{-\pi/2,\pi/2} = \{\, -\tfrac{1}{2}\pi < \operatorname{ph} \zeta < \tfrac{1}{2}\pi \,\} = \{\, \operatorname{Im} \zeta > 0 \,\},$$

while the horizontal strip $S_{-\pi,\pi} = \{\, -\pi < \operatorname{Im} z < \pi \,\}$ of width 2π is mapped onto the domain

$$\Omega_* = \Omega_{-\pi,\pi} = \{\, -\pi < \operatorname{ph} \zeta < \pi \,\} = \mathbb{C} \setminus \{\, \operatorname{Im} z = 0,\ \operatorname{Re} z \leq 0 \,\}$$

obtained by cutting the complex plane along the negative real axis.

On the other hand, vertical lines $\operatorname{Re} z = a$ are mapped to circles $|\, \zeta \,| = e^a$. Thus, a vertical strip $a < \operatorname{Re} z < b$ is mapped to an annulus $e^a < |\, \zeta \,| < e^b$, albeit many-to-one, since the strip is effectively wrapped around and around the annulus. The rectangle $R = \{\, a < x < b, -\pi < y < \pi \,\}$ of height 2π is mapped in a one-to-one fashion on an annulus that has been cut along the negative real axis. See Figure 16.13 for an illustration.

Example 16.24. The squaring map

$$\zeta = g(z) = z^2, \qquad \text{or} \qquad \xi = x^2 - y^2, \quad \eta = 2\,x\,y, \tag{16.64}$$

is analytic on all of $\mathbb{C}$, but is not one-to-one. Its inverse is the square root function $z = \sqrt{\zeta}$, which, as discussed in Section 16.1, is double-valued. Furthermore, the derivative $g'(z) = 2\,z$ vanishes at $z = 0$, violating the invertibility condition (16.57). However, once we restrict to a simply connected subdomain Ω that does not contain 0, the function $g(z) = z^2$ does define a one-to-one mapping, whose inverse $z = g^{-1}(\zeta) = \sqrt{\zeta}$ is a well-defined, analytic and single-valued branch of the square root function.

The effect of the squaring map on a point z is to square its modulus, $|\, \zeta \,| = |\, z \,|^2$, while doubling its angle, $\operatorname{ph} \zeta = \operatorname{ph} z^2 = 2 \operatorname{ph} z$. Thus, for example, the upper right quadrant

$$Q = \{\, x > 0, y > 0 \,\} = \{\, 0 < \operatorname{ph} z < \tfrac{1}{2}\pi \,\}$$

is mapped by (16.64) onto the upper half plane

$$U = g(Q) = \{\, \eta = \operatorname{Im} \zeta > 0 \,\} = \{\, 0 < \operatorname{ph} \zeta < \pi \,\}.$$

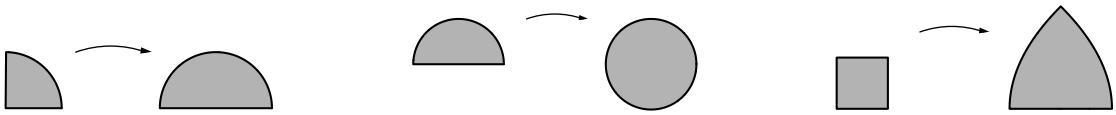


Figure 16.14. The Effect of $\zeta = z^2$ on Various Domains.

The inverse function maps a point $\zeta \in D$ back to its unique square root $z = \sqrt{\zeta}$ that lies in the quadrant Q. Similarly, a quarter disk

$$Q_\rho = \left\{ 0 < |\, z \,| < \rho,\ 0 < \operatorname{ph} z < \tfrac{1}{2}\,\pi \right\}$$

of radius ρ is mapped to a half disk

$$U_{\rho^2} = g(\Omega) = \left\{ 0 < |\, \zeta \,| < \rho^2,\ \operatorname{Im} \zeta > 0 \right\}$$

of radius ρ^2. On the other hand, the unit square $\Omega = \left\{ 0 < x < 1, 0 < y < 1 \right\}$ is mapped to a curvilinear triangular domain, as indicated in Figure 16.14; the edges of the square on the real and imaginary axes map to the two halves of the straight base of the triangle, while the other two edges become its curved sides.

Example 16.25. A particularly important example is the analytic map

$$\zeta = \frac{z-1}{z+1} = \frac{x^2 + y^2 - 1 + 2\,\mathrm{i}\,y}{(x+1)^2 + y^2}\,, \tag{16.65}$$

where we used (16.14) to derive the formulae for its real and imaginary parts. The map is one-to-one with analytic inverse

$$z = \frac{1+\zeta}{1-\zeta} = \frac{1 - \xi^2 - \eta^2 + 2\,\mathrm{i}\,\eta}{(1-\xi)^2 + \eta^2}\,, \tag{16.66}$$

provided $z \neq -1$ and $\zeta \neq 1$. This particular analytic map has the important property of mapping the right half plane $R = \left\{ x = \operatorname{Re} z > 0 \right\}$ to the unit disk $D = \left\{ |\, \zeta \,|^2 < 1 \right\}$. Indeed, by (16.66)

$$|\, \zeta \,|^2 = \xi^2 + \eta^2 < 1 \qquad \text{if and only if} \qquad x = \frac{1 - \xi^2 - \eta^2}{(1-\xi)^2 + \eta^2} > 0.$$

Note that the denominator does not vanish on the interior of the disk.

The complex functions (16.61), (16.62), (16.65) are particular examples of one of the most important class of analytic maps. A general *linear fractional transformation* has the form

$$\zeta = \frac{\alpha\, z + \beta}{\gamma\, z + \delta}\,, \tag{16.67}$$

where $\alpha, \beta, \gamma, \delta$ are arbitrary complex constants, subject to the restriction

$$\alpha\,\delta - \beta\,\gamma \neq 0,$$

since otherwise (16.67) reduces to a trivial constant (and non-invertible) map.

Example 16.26. The linear fractional transformation

$$\zeta = \frac{z - \alpha}{\overline{\alpha}\, z - 1} \qquad \text{where} \qquad |\, \alpha \,| < 1, \tag{16.68}$$

maps the unit disk to itself, moving the origin $z = 0$ to the point $\zeta = \alpha$. To prove this, we note that

$$|\, z - \alpha \,|^2 = (z - \alpha)(\overline{z} - \overline{\alpha}) = |\, z \,|^2 - \alpha\, \overline{z} - \overline{\alpha}\, z + |\, \alpha \,|^2,$$
$$|\, \overline{\alpha}\, z - 1 \,|^2 = (\overline{\alpha}\, z - 1)(\alpha\, \overline{z} - 1) = |\, \alpha \,|^2\, |\, z \,|^2 - \alpha\, \overline{z} - \overline{\alpha}\, z + 1.$$

Subtracting these two formulae, and using the assumptions that $|\, z \,| < 1$, $|\, \alpha \,| < 1$, we find

$$|\, z - \alpha \,|^2 - |\, \overline{\alpha}\, z - 1 \,|^2 = \big(1 - |\, \alpha \,|^2\big)\, \big(|\, z \,|^2 - 1\big) < 0, \qquad \text{so} \qquad |\, z - \alpha \,| < |\, \overline{\alpha}\, z - 1 \,|.$$

The latter inequality implies that

$$|\, \zeta \,| = \frac{|\, z - \alpha \,|}{|\, \overline{\alpha}\, z - 1 \,|} < 1 \qquad \text{provided} \qquad |\, z \,| < 1, \qquad |\, \alpha \,| < 1,$$

and hence ζ lies within the unit disk.

The rotations (16.60) also map the unit disk to itself, preserving the origin. It can be proved, [**4**], that the only invertible analytic mappings that take the unit disk to itself are obtained by composing such a linear fractional transformation with a rotation.

Proposition 16.27. *If $\zeta = g(z)$ is a one-to-one analytic map that takes the unit disk to itself, then*

$$g(z) = e^{\,\mathrm{i}\,\varphi}\, \frac{z - \alpha}{\overline{\alpha}\, z - 1} \tag{16.69}$$

for some $|\, \alpha \,| < 1$ and $0 \le \varphi < 2\,\pi$.

Additional specific properties of linear fractional transformations are outlined in the exercises. The most important is that they map circles to circles, where, to be completely accurate, one should view a straight line as a "circle of infinite radius". Details can be found in Exercise ■.

Conformality

A remarkable geometrical characterization of complex analytic functions is the fact that, at non-critical points, they preserve angles. The mathematical term for this property is *conformal mapping*. Conformality makes sense for any inner product space, although in practice one usually deals with Euclidean space equipped with the standard dot product.

Definition 16.28. A function $g\colon \mathbb{R}^n \to \mathbb{R}^n$ is called *conformal* if it preserves angles.

What does it mean to "preserve angles"? For the Euclidean norm, the angle between two vectors is defined by their dot product, as in (3.18). However, most analytic maps are nonlinear, and so will not map vectors to vectors since they will typically map straight lines to curves. However, if we interpret "angle" to mean the angle between two curves,

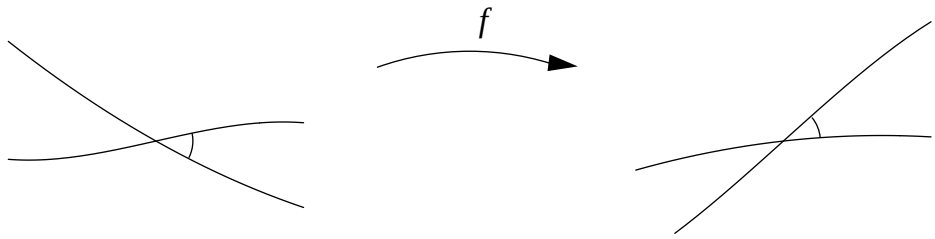


Figure 16.15. A Conformal Map.

as illustrated in Figure 16.15, then we can make sense of the conformality requirement. Consequently, in order to realize complex functions as conformal maps, we first need to understand what they do to curves.

In general, a *curve* $C \in \mathbb{C}$ in the complex plane is parametrized by a complex-valued function

$$z(t) = x(t) + \mathrm{i}\, y(t), \qquad a < t < b, \tag{16.70}$$

that depends on a real parameter t. Note that there is no essential difference between a complex plane curve (16.70) and a real plane curve (A.1) — we have merely switched from vector notation $\mathbf{x}(t) = (\, x(t), y(t)\,)^T$ to complex notation $z(t) = x(t) + \mathrm{i}\, y(t)$. All the vectorial curve terminology (closed, simple, piecewise smooth, etc.) we learned in Appendix A is used without any modification here. In particular, the *tangent vector* to the curve can be identified as the complex number $\dot z(t) = \dot x(t) + \mathrm{i}\, \dot y(t)$. Smoothness of the curve is guaranteed by the requirement that $\dot z(t) \neq 0$.

Example 16.29.

(*a*) The curve

$$z(t) = e^{\,\mathrm{i}\, t} = \cos t + \mathrm{i}\, \sin t, \qquad \text{for} \qquad 0 \leq t \leq 2\pi,$$

parametrizes the unit circle $|\, z\,| = 1$ in the complex plane, which is a simple closed curve. Its complex tangent is $\dot z(t) = \mathrm{i}\, e^{\,\mathrm{i}\, t} = \mathrm{i}\, z(t)$, which is obtained by rotating z through 90°.

(*b*) The complex curve

$$z(t) = \cosh t + \mathrm{i}\, \sinh t = \frac{1 + \mathrm{i}}{2}\, e^t + \frac{1 - \mathrm{i}}{2}\, e^{-t}, \qquad -\infty < t < \infty,$$

parametrizes the right hand branch of the hyperbola

$$\mathrm{Re}\, z^2 = x^2 - y^2 = 1.$$

The complex tangent vector is $\dot z(t) = \sinh t + \mathrm{i}\, \cosh t = \mathrm{i}\, \overline{z}(t)$.

In order to better understand the geometry, it will help to rewrite the tangent $\dot z$ in polar coordinates. We interpret the curve as the motion of a particle in the complex plane, so that $z(t)$ is the position of the particle at time t, and the tangent $\dot z(t)$ its instantaneous velocity. The modulus of the tangent, $|\, \dot z\,| = \sqrt{\dot x^2 + \dot y^2}$, indicates the particle's speed,

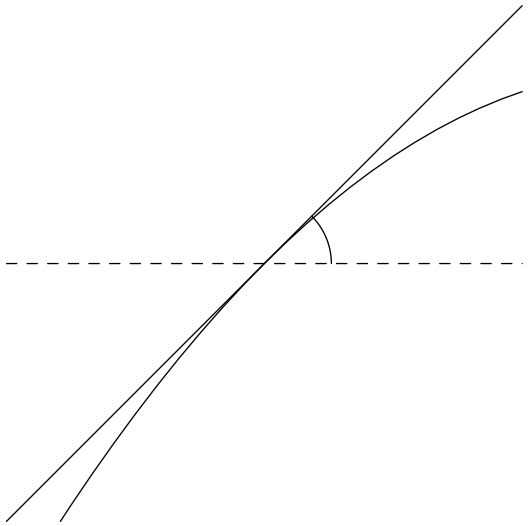


Figure 16.16. Complex Curve and Tangent.

while its phase $\operatorname{ph} \dot{z} = \tan^{-1}(\dot{y}/\dot{x})$ measures the direction of motion, or, more precisely, the angle that the curve makes with the horizontal; see Figure 16.16.

The angle between two curves is defined as the angle between their tangents at the point of intersection. If the curve C_1 makes an angle $\theta_1 = \operatorname{ph} \dot{z}_1(t_1)$ while the curve C_2 has angle $\theta_2 = \operatorname{ph} \dot{z}_2(t_2)$ at the common point $z = z_1(t_1) = z_2(t_2)$, then the angle θ between the two curves at z is the difference

$$\theta = \theta_2 - \theta_1 = \operatorname{ph} \dot{z}_2 - \operatorname{ph} \dot{z}_1 = \operatorname{ph} \frac{\dot{z}_2}{\dot{z}_1}\,. \tag{16.71}$$

Now, suppose we are given an analytic map $\zeta = g(z)$. A curve C parametrized by $z(t)$ will be mapped to a curve $\Gamma = g(C)$ parametrized by the composition $\zeta(t) = g(z(t))$. The tangent to the image curve is related to that of the original curve by the chain rule:

$$\frac{d\zeta}{dt} = \frac{dg}{dz}\,\frac{dz}{dt}\,, \qquad \text{or} \qquad \dot{\zeta}(t) = g'(z(t))\,\dot{z}(t). \tag{16.72}$$

Therefore, the effect of the analytic map on the tangent vector $\dot{z}$ at the point $z \in C$ is to multiply it by the complex number $g'(z)$. If the analytic map satisfies our key assumption $g'(z) \neq 0$, then $\dot{\zeta} \neq 0$, and so the image curve will remain smooth.

According to equation (16.72),

$$|\,\dot{\zeta}\,| = |\,g'(z)\,\dot{z}\,| = |\,g'(z)\,|\,|\,\dot{z}\,|. \tag{16.73}$$

Thus, the speed of motion along the new curve $\zeta(t)$ is multiplied by a factor $\rho = |\,g'(z)\,| > 0$. The magnification factor ρ depends only upon the point z and not how the curve passes through it. All curves passing through the point z are speeded up (or slowed down if $\rho < 1$) by the same factor! Similarly, the angle that the new curve makes with the horizontal is given by

$$\operatorname{ph} \dot{\zeta} = \operatorname{ph}\big(\,g'(z)\,\dot{z}\,\big) = \operatorname{ph} g'(z) + \operatorname{ph} \dot{z}, \tag{16.74}$$

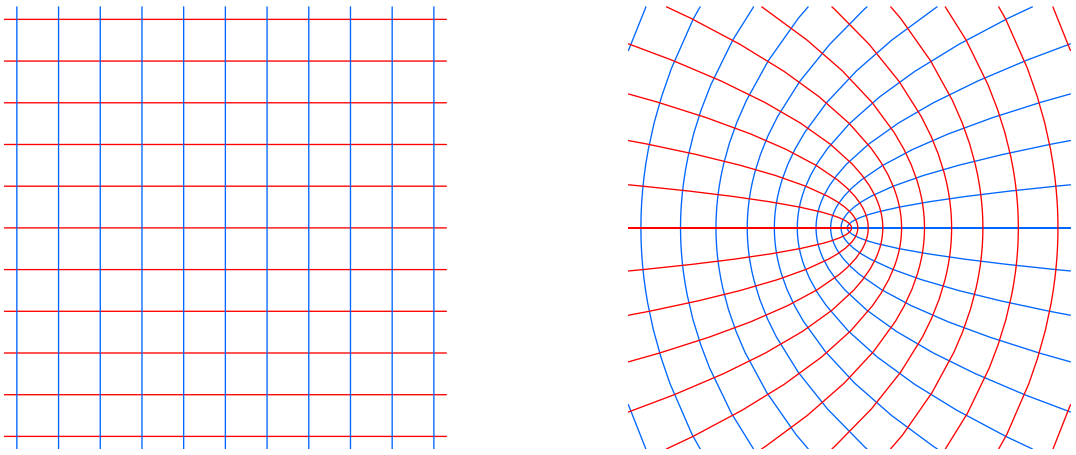


Figure 16.17. Conformality of z^2.

where we use the fact that the phase of the product of two complex numbers is the sum of their individual phases, (3.74). Therefore, the tangent angle of the curve is increased by an amount $\phi = \operatorname{ph} g'(z)$. Geometrically, this means that the tangent to the curve has been rotated through an angle ϕ. Again, the increase in tangent angle only depends on the point z, and all curves passing through z are rotated by the same amount ϕ. As a result, the angle between any two curves is preserved. More precisely, if C_1 is at angle θ_1 and C_2 at angle θ_2 at a point of intersection, then their images $\Gamma_1 = g(C_1)$ and $\Gamma_2 = g(C_2)$ are at angles $\psi_1 = \theta_1 + \phi$ and $\psi_2 = \theta_2 + \phi$. The angle between the two image curves is the difference

$$\psi_2 - \psi_1 = (\theta_2 + \phi) - (\theta_1 + \phi) = \theta_2 - \theta_1,$$

which is *the same* as the angle between the original curves. This proves the conformality or angle-preservation property of analytic maps.

Theorem 16.30. *If $\zeta = g(z)$ is an analytic function and $g'(z) \neq 0$, then g defines a conformal map.*

Remark: The converse is also valid every planar conformal map comes from a complex analytic function with nonvanishing derivative. A proof is outlined in Exercise ■.

The conformality of a analytic functions is all the more surprising when one reconsiders elementary examples. In Example 16.24, we discovered that the function $w = z^2$ maps a quarter plane to a half plane, and therefore *doubles* the angle at the origin! Thus $g(z) = z^2$ is most definitely not conformal at $z = 0$. The explanation is, of course, that it has zero derivative at $z = 0$, and Theorem 16.30 only guarantees conformality when the derivative is nonzero. Amazingly, the map preserves angles everywhere else! Somehow, the angle at the origin is doubled, while the angles at all nearby points are preserved. Figure 16.17 illustrates this remarkable and counter-intuitive feat. The left hand figure shows the coordinate grid, while on the right are the images of the horizontal and vertical lines under the map z^2. Note that, except at the origin, the image curves continue to meet at 90° angles, in accordance with conformality.

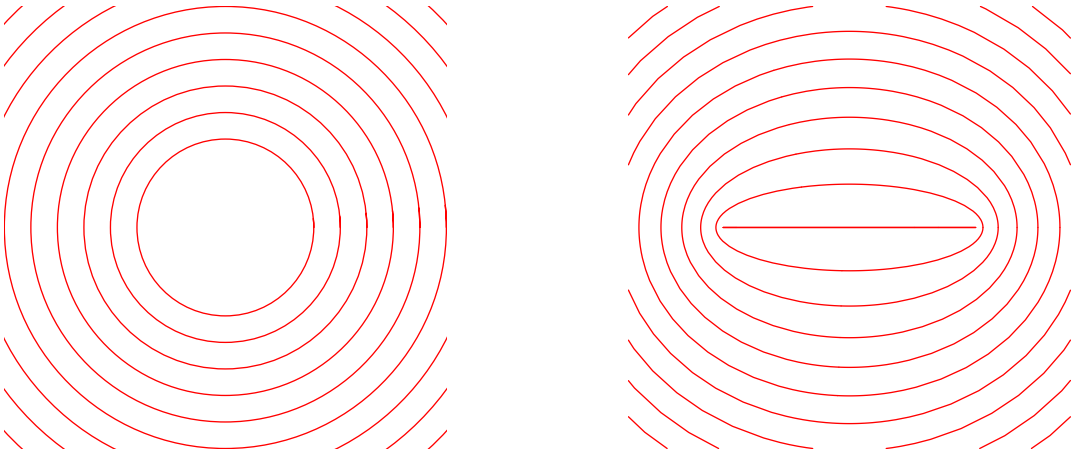


Figure 16.18. The Joukowski Map.

Example 16.31. A particularly interesting conformal transformation is given by the function

$$\zeta = \frac{1}{2}\left(z + \frac{1}{z} \right). \tag{16.75}$$

The *Joukowski map* arises in the study of flows around airplane wings, since it maps circles to a variety of airfoil shapes whose aerodynamic properties can be analyzed exactly, and is named after the pioneering Russian aero- and hydro-dynamics researcher Nikolai Zhukovskii (Joukowski). Since

$$\frac{d\zeta}{dz} = \frac{1}{2}\left(1 - \frac{1}{z^2} \right) = 0 \qquad \text{if and only if} \qquad z = \pm 1,$$

the Joukowski map is conformal except at the critical points $z = \pm 1$, as well as the singularity $z = 0$, where it is not defined.

If $z = e^{\,\mathrm{i}\,\theta}$ lies on the unit circle, then

$$\zeta = \tfrac{1}{2}\big(e^{\,\mathrm{i}\,\theta} + e^{-\,\mathrm{i}\,\theta} \big) = \cos\theta,$$

lies on the real axis, with $-1 \le \zeta \le 1$. Thus, the Joukowski map squashes the unit circle down to the real line segment $[\,-1, 1\,]$. The points outside the unit circle fill the rest of the ζ plane, as do the (nonzero) points inside the unit circle. Indeed, if we solve (16.75) for

$$z = \zeta \,\pm\, \sqrt{\zeta^2 - 1}\,, \tag{16.76}$$

we see that every ζ except ± 1 comes from two different points z; for ζ not on the line segment $[\,-1, 1\,]$ the points inside and outside the unit circle, whereas if $-1 < \zeta < 1$, the points lie directly above and below it on the circle. Therefore, (16.75) defines a one-to-one conformal map from the exterior of the unit circle $\big\{\, |\, z\,| > 1 \,\big\}$ onto the exterior of the unit line segment $\mathbb{C} \setminus [\,-1, 1\,]$.

Under the Joukowski map, the concentric circles $|\, z\,| = r$ for $r \neq 1$ are mapped to ellipses with foci at ± 1 in the ζ plane, as illustrated in Figure 16.18. The effect on circles not centered at the origin is quite interesting. The image curves take on a wide variety

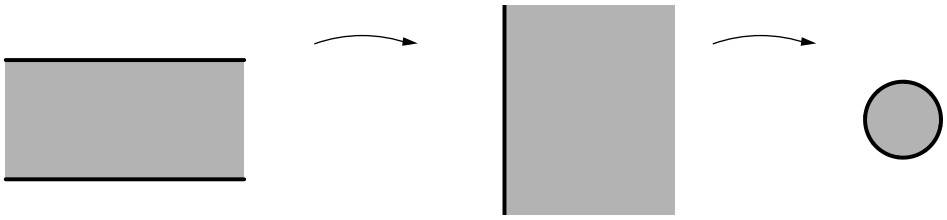


Figure 16.19. Composition of Conformal Maps.

of shapes; several examples are plotted in Figure airfoil. If the circle passes through the singular point $z = 1$, then its image is no longer smooth, but has a cusp at $\zeta = 1$. Some of the image curves have the shape of the cross-section through an airplane wing or *airfoil*. Later we will see how to apply the Joukowski map to construct the physical fluid flow around such an airfoil, which proved to be a critical step in early airplane design.

Composition and The Riemann Mapping Theorem

One of the strengths of the method of conformal mapping is that one can build up lots of complicated examples by simply composing elementary mappings. According to Lemma 16.20, if $w = h(z)$ and $\zeta = k(w)$ are analytic functions, their composition $\zeta = g(z) = k \circ h(z) = k(h(z))$ is also analytic. If both h and k are one-to-one, so is the composition $g = k \circ h$. Moreover, the composition of two conformal maps is also conformal. Indeed, by the chain rule,

$$g'(z) = k'(h(z))\, h'(z) \neq 0 \qquad \text{provided} \qquad k'(h(z)) \neq 0 \qquad \text{and} \qquad h'(z) \neq 0,$$

and so if h and k satisfy the conformality condition (16.57), so does $g = k \circ h$.

Example 16.32. As we learned in Example 16.23, the exponential function

$$w = e^z$$

maps the horizontal strip $S = \{ -\frac{1}{2}\pi < \operatorname{Im} z < \frac{1}{2}\pi \}$ conformally onto the right half plane $R = \{ \operatorname{Re} w > 0 \}$. On the other hand, Example 16.25 tells us that the linear fractional transformation

$$\zeta = \frac{w - 1}{w + 1}$$

maps the right half plane R conformally to the unit disk $D = \{ |\, \zeta \,| < 1 \}$, as in Figure 16.19. Therefore, the composition

$$\zeta = \frac{e^z - 1}{e^z + 1} \tag{16.77}$$

is a one-to-one conformal map from the horizontal strip S to the unit disk D.

Recall that our motivating goal is to use analytic/conformal maps to transform a boundary value problem for the Laplace equation on a complicated domain Ω to a boundary value problem on the unit disk. Since we already know how to solve the latter, the method effectively constructs a solution to the original problem. Of course, the key question the student should be asking at this point is: Can you construct a conformal map $\zeta = g(z)$ from a given domain Ω to the unit disk $D = g(\Omega)$?

The theoretical answer to this question is the celebrated *Riemann Mapping Theorem*.

Theorem 16.33. *If $\Omega \subsetneq \mathbb{C}$ is any simply connected open subset,* not *equal to the entire complex plane, then there exists a one-to-one analytic function $\zeta = g(z)$ that maps Ω to the unit disk $D = \{\, |\, \zeta\, | < 1 \,\}$.*

Thus, *any* simply connected open set, including all domains, can be conformally mapped the unit disk — the one exception is the entire complex plane. (See Exercise ■ for a reason for this exception.) Note that the domain Ω does *not* have to be bounded for this result to hold. For example, the conformal map (16.65) takes the unbounded right half plane $R = \{\, \mathrm{Re}\, z > 0 \,\}$ to the unit disk. The proof of this important theorem is not easy and relies some more advanced results in complex analysis, [**4**].

The Riemann Mapping Theorem guarantees the existence of a conformal map from any simply connected domain to the unit disk, but it is an existential result, and gives no clue as to how to actually construct the desired mapping. And, in general, this is not an easy task. In practice, one assembles a repertoire of useful conformal maps that apply to particular domains of interest. One extensive catalog can be found in [**Cmap**]. More complicated maps can then be built up by composition of the basic examples. Ultimately, though, the determination of a suitable conformal map is often more an art than a systematic science.

Let us consider a few additional examples beyond those already encountered:

Example 16.34. Suppose we are asked to conformally map the upper half plane $U = \big\{\, \mathrm{Im}\, z > 0 \,\big\}$ to the unit disk $D = \big\{\, |\, \zeta\, | < 1 \,\big\}$. We already know that the linear fractional transformation

$$\zeta = g(z) = \frac{z-1}{z+1}$$

maps the right half plane $R = \big\{\, \mathrm{Re}\, z > 0 \,\big\}$ to $D = g(R)$. On the other hand, multiplication by $\mathrm{i} = e^{\,\mathrm{i}\,\pi/2}$, with $z = h(w) = \mathrm{i}\, w$, rotates the complex plane by 90° and so maps the right half plane R to the upper half plane $U = h(R)$. Its inverse $h^{-1}(z) = -\,\mathrm{i}\, z$ will therefore map U to $R = h^{-1}(U)$. Therefore, to map the upper half plane to the unit disk, we compose these two maps, leading to the conformal map

$$\zeta = g \circ h^{-1}(z) = \frac{-\,\mathrm{i}\, z - 1}{-\,\mathrm{i}\, z + 1} = \frac{\mathrm{i}\, z + 1}{\mathrm{i}\, z - 1} \tag{16.78}$$

from U to D.

As a second example, we already know that the squaring map $w = z^2$ maps the upper right quadrant $Q = \big\{\, 0 < \mathrm{ph}\, z < \frac{1}{2}\,\pi \,\big\}$ to the upper half plane U. Composing this with our previously constructed map (16.78) leads to the conformal map

$$\zeta = \frac{\mathrm{i}\, z^2 + 1}{\mathrm{i}\, z^2 - 1} \tag{16.79}$$

that maps the quadrant Q to the unit disk D.

Example 16.35. The goal of this example is to construct an conformal map that takes a half disk

$$D_+ = \big\{\, |\, z\, | < 1,\ y = \mathrm{Im}\, z > 0 \,\big\} \tag{16.80}$$

to the full unit disk D. The answer is *not* $\zeta = z^2$ because the image omits the positive real axis, and so is a disk with a slit cut out of it. The first observation is that the map $z = (w-1)/(w+1)$ that we analyzed in Example 16.25 takes the right half plane $R = \{ \operatorname{Re} w > 0 \}$ to the unit disk. Moreover, it maps the upper right quadrant $Q = \{ 0 < \operatorname{ph} w < \frac{1}{2}\pi \}$ to the half disk (16.80). Its inverse,

$$w = \frac{z+1}{z-1}$$

will therefore map the half disk to the upper right quadrant.

On the other hand, we just constructed a conformal map (16.79) that takes the upper right quadrant Q to the unit disk. Therefore, if compose the two maps (replacing z by w in (16.79)), we obtain the desired conformal map

$$\zeta = \frac{\mathrm{i}\, w^2 + 1}{\mathrm{i}\, w^2 - 1} = \frac{\mathrm{i}\left(\dfrac{z+1}{z-1}\right)^2 + 1}{\mathrm{i}\left(\dfrac{z+1}{z-1}\right)^2 - 1} = \frac{(\mathrm{i}+1)(z^2+1) + 2(\mathrm{i}-1)z}{(\mathrm{i}-1)(z^2+1) + 2(\mathrm{i}+1)z}\,.$$

The formula can be further simplified by multiplying numerator and denominator by $\mathrm{i}+1$, and so

$$\zeta = -\,\mathrm{i}\,\frac{z^2 + 2\,\mathrm{i}\,z + 1}{z^2 - 2\,\mathrm{i}\,z + 1}\,. \tag{16.81}$$

The leading factor $-\,\mathrm{i}$ is unimportant and can be omitted, since it merely rotates the disk by -90°.

Finally, we remark that the conformal map guaranteed by the Riemann Mapping Theorem is *not* unique. Since the linear fractional transformations (16.68) map the unit disk to itself, we can compose them with any Riemann mapping to produce additional maps from a simply-connected domain to the unit disk. For example, composing (16.68) with (16.77) produces a family of mappings

$$\zeta = \frac{1 + e^z - \alpha(1 - e^z)}{\overline{\alpha}\,(1 + e^z) - 1 + e^z}\,, \tag{16.82}$$

which, for any $|\,\alpha\,| < 1$, maps the strip $S = \{ -\frac{1}{2}\pi < \operatorname{Im} z < \frac{1}{2}\pi \}$ onto the unit disk.

Annular Domains

The Riemann Mapping Theorem does not apply directly to non-simply connected domains. For purely topological reasons, a hole cannot be made to disappear under a one-to-one continuous mapping — much less a conformal map!

The simplest non-simply connected domains is an *annulus* consisting of the points between two concentric circles

$$A_{r,R} = \{ r < |\,\zeta\,| < R \}, \tag{16.83}$$

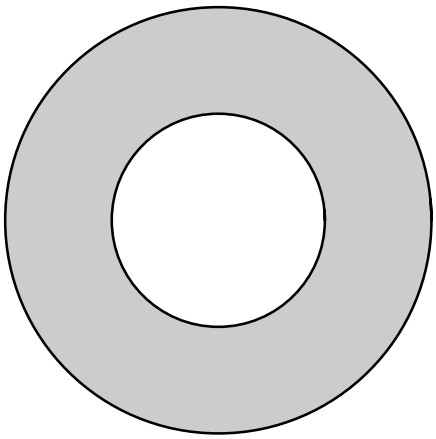


Figure 16.20. An Annulus.

which, for simplicity, is centered at the origin; see Figure 16.20. It can be proved, [**Cmap**], that any other domain with a single hole can be mapped to an annulus. The annular radii r, R are not uniquely specified; indeed the linear map $\zeta = \alpha\, z$ maps the annulus (16.83) to a rescaled annulus $A_{\rho\, r, \rho\, R}$ whose inner and outer radii have both been scaled by the factor $\rho = |\, \alpha\, |$. The ratio[†] r/R of the inner to outer radius of the annulus is uniquely specified; annuli with different ratios *cannot* be mapped to each other by a conformal map. Thus, unlike simply connected domains, there are many "standard" multiply connected domains.

Example 16.36. Consider the domain

$$\Omega = \big\{\, |\, z\, | < 1 \quad \text{and} \quad |\, z - c\, | > c \,\big\}$$

contained between two nonconcentric circles. To keep the computations simple, we take the outer circle to have radius 1 (which can always be arranged by scaling, anyway) while the inner circle has center at the point $z = c$ on the real axis and radius c, which means that it passes through the origin. We must restrict $c < \frac{1}{2}$ in order that the inner circle not overlap with the outer circle. Our goal is to conformally map this non-concentric annular domain to a concentric annulus of the form

$$A_{r,1} = \big\{\, r < |\, \zeta\, | < 1 \,\big\}$$

by a conformal map $\zeta = g(z)$.

Now, according, to Example 16.26, a linear fractional transformation of the form

$$\zeta = g(z) = \frac{z - \alpha}{\overline{\alpha}\, z - 1} \qquad \text{with} \qquad |\, \alpha\, | < 1 \tag{16.84}$$

maps the unit disk to itself. Moreover, as remarked earlier, and demonstrated in Exercise ■, linear fractional transformations always map circles to circles. Therefore, we seek a

[†] If $r = 0$ or $R = \infty$, then $r/R = 0$ by convention.

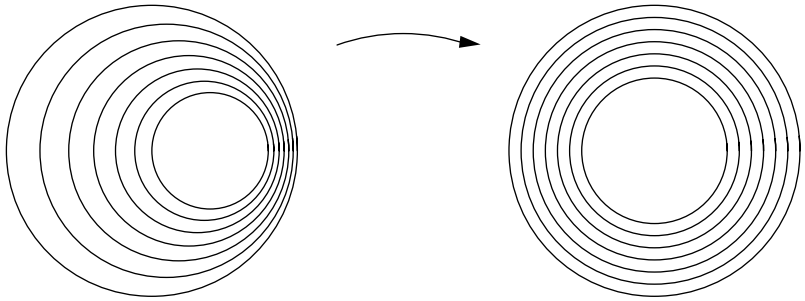


Figure 16.21. Conformal Map for Non-concentric Annulus.

particular value of α that maps the inner circle $|\, z - c \,| = c$ to a circle of the form $|\, \zeta \,| = r$ centered at the origin. We choose α real and try to map the points 0 and $2c$ on the inner circle to the points r and $-r$ on the circle $|\, \zeta \,| = r$. This requires

$$g(0) = \alpha = r, \qquad g(2c) = \frac{2c - \alpha}{2c\alpha - 1} = -r. \tag{16.85}$$

Substituting the first into the second leads to the quadratic equation

$$c\alpha^2 - \alpha + c = 0.$$

There are two real solutions:

$$\alpha = \frac{1 - \sqrt{1 - 4c^2}}{2c} \qquad \text{and} \qquad \alpha = \frac{1 + \sqrt{1 - 4c^2}}{2c}\,. \tag{16.86}$$

Since $0 < c < \frac{1}{2}$, the second solution has $\alpha > 1$, and hence is inadmissible. Therefore, the first solution gives the required conformal map

$$\zeta = \frac{z - 1 + \sqrt{1 - 4c^2}}{(1 - \sqrt{1 - 4c^2}\,)\, z - 2c}\,.$$

Note in particular that the radius $r = \alpha$ of the inner circle in $A_{r,1}$ is *not the same* as the radius c of the inner circle in Ω.

For example, taking $c = \frac{2}{5}$, we find $\alpha = \frac{1}{2}$, and hence the linear fractional transformation $\zeta = \dfrac{2z - 1}{z - 2}$ maps the annular domain $\Omega = \left\{ |\, z \,| < 1 \quad \text{and} \quad \left|\, z - \tfrac{2}{5} \,\right| > \tfrac{2}{5} \right\}$ to the concentric annulus $A = A_{.5,1} = \left\{ \tfrac{1}{2} < |\, \zeta \,| < 1 \right\}$. In Figure 16.21, we plot the non-concentric circles in Ω that map to concentric circles in the annulus A. In Exercise ∎ the reader is asked to adapt this construction to a general non-concentric annular domain.

Applications to Harmonic Functions and Laplace's Equation

Let us now apply what we have learned about analytic/conformal maps to the study of harmonic functions and boundary value problems for the Laplace equation. Suppose $\zeta = g(z)$ defines a one-to-one conformal map from the domain $z \in \Omega$ onto the domain $\zeta \in D$. In many applications, the target domain D is the unit disk $|\, \zeta \,| < 1$, but this is

not necessary for the time being. According to Lemma 16.20, composing the conformal map g takes analytic functions $F(\zeta)$ defined on D to analytic functions $f(z) = F(g(z))$ on Ω, and hence defines a change of variables between their harmonic real and imaginary parts. In fact, this property does not even require the harmonic function to be the real part of an analytic function, i.e., we are not required to assume the existence of a harmonic conjugate.

Proposition 16.37. *If $U(\xi, \eta)$ is a harmonic function of ξ, η, and*

$$\xi + \mathrm{i}\,\eta = p(x, y) + \mathrm{i}\,q(x, y) \tag{16.87}$$

is any analytic mapping, then the composition

$$u(x, y) = U(p(x, y), q(x, y)) \tag{16.88}$$

is a harmonic function of x, y.

Proof: This is a straightforward application of the chain rule:

$$\frac{\partial u}{\partial x} = \frac{\partial U}{\partial \xi}\,\frac{\partial \xi}{\partial x} + \frac{\partial U}{\partial \eta}\,\frac{\partial \eta}{\partial x}, \qquad \frac{\partial u}{\partial y} = \frac{\partial U}{\partial \xi}\,\frac{\partial \xi}{\partial y} + \frac{\partial U}{\partial \eta}\,\frac{\partial \eta}{\partial y},$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 U}{\partial \xi^2}\left(\frac{\partial \xi}{\partial x}\right)^2 + 2\,\frac{\partial^2 U}{\partial \xi\,\partial \eta}\,\frac{\partial \xi}{\partial x}\,\frac{\partial \eta}{\partial x} + \frac{\partial^2 U}{\partial \eta^2}\left(\frac{\partial \eta}{\partial x}\right)^2 + \frac{\partial U}{\partial \xi}\,\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial U}{\partial \eta}\,\frac{\partial^2 \eta}{\partial x^2},$$

$$\frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 U}{\partial \xi^2}\left(\frac{\partial \xi}{\partial y}\right)^2 + 2\,\frac{\partial^2 U}{\partial \xi\,\partial \eta}\,\frac{\partial \xi}{\partial y}\,\frac{\partial \eta}{\partial y} + \frac{\partial^2 U}{\partial \eta^2}\left(\frac{\partial \eta}{\partial y}\right)^2 + \frac{\partial U}{\partial \xi}\,\frac{\partial^2 \xi}{\partial y^2} + \frac{\partial U}{\partial \eta}\,\frac{\partial^2 \eta}{\partial y^2}.$$

Using the Cauchy–Riemann equations

$$\frac{\partial \xi}{\partial x} = -\,\frac{\partial \eta}{\partial y}, \qquad \frac{\partial \xi}{\partial y} = \frac{\partial \eta}{\partial x},$$

for the analytic function $\zeta = \xi + \mathrm{i}\,\eta$, we find, after some algebra,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \left[\left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial x}\right)^2\right]\left[\frac{\partial^2 U}{\partial \xi^2} + \frac{\partial^2 U}{\partial \eta^2}\right].$$

Therefore,

$$\Delta u = |\,g'(z)\,|^2\,\Delta U \qquad \text{where} \qquad |\,g'(z)\,|^2 = \left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial x}\right)^2.$$

We conclude that whenever $U(\xi, \eta)$ is any harmonic function, and so a solution to the Laplace equation $\Delta U = 0$ (in the ξ, η variables), then $u(x, y)$ is a solution to the Laplace equation $\Delta u = 0$ in the x, y variables, and is thus also harmonic. *Q.E.D.*

This observation has profound consequences for boundary value problems arising in physical applications. Suppose we wish to solve the Dirichlet problem

$$\Delta u = 0 \quad \text{in} \quad \Omega, \qquad u = h \quad \text{on} \quad \partial\Omega,$$

on a simply connected domain $\Omega \subsetneq \mathbb{C}$. (The Riemann Mapping Theorem 16.33 tells us to exclude the case $\Omega = \mathbb{C}$. Indeed, this case is devoid of boundary conditions, and so the problem does not admit a unique solution.) If we can find a complex function $\zeta = g(z) = p(x,y) + \mathrm{i}\, q(x,y)$ that defines a one-to-one conformal mapping from the domain Ω to the unit disk D, then we can use the change of variables formula (16.88) to map the harmonic function $u(x,y)$ on Ω to a harmonic function $U(\xi,\eta)$ on D. Moreover, the boundary values of $U = H$ on the unit circle ∂D correspond to those of $u = h$ on $\partial\Omega$ by the same change of variables formula:

$$h(x,y) = H(p(x,y), q(x,y)), \qquad \text{for} \qquad (x,y) \in \partial\Omega. \tag{16.89}$$

We conclude that $U(\xi,\eta)$ solves the Dirichlet problem

$$\Delta U = 0 \qquad \text{in} \qquad D, \qquad\qquad U = H \qquad \text{on} \qquad \partial D.$$

But we already know how to solve the Dirichlet problem on the unit disk! Namely, the Poisson integral formula (15.44) gives $U(\xi,\eta)$. The corresponding solution to our original boundary value problem is given by the composition formula $u(x,y) = U\big(p(x,y), q(x,y)\big)$. Thus, the solution to the Dirichlet problem on a unit disk can be used to solve the Dirichlet problem on a more complicated planar domain — provided we know the conformal map whose existence is guaranteed by the Riemann Mapping Theorem 16.33.

Example 16.38. According to Example 16.25, the analytic function

$$\xi + \mathrm{i}\,\eta = \zeta = \frac{z-1}{z+1} = \frac{x^2+y^2-1}{(x+1)^2+y^2} + \mathrm{i}\,\frac{2y}{(x+1)^2+y^2} \tag{16.90}$$

maps the right half plane $R = \big\{\, x = \operatorname{Re} z > 0 \,\big\}$ to the unit disk $D = \big\{\, |\,\zeta\,| < 1 \,\big\}$. Proposition 16.37 implies that if $U(\xi,\eta)$ is a harmonic function in the unit disk, then

$$u(x,y) = U\left(\frac{x^2+y^2-1}{(x+1)^2+y^2}\,,\ \frac{2y}{(x+1)^2+y^2} \right) \tag{16.91}$$

is a harmonic function on the right half plane.

To solve the Dirichlet boundary value problem

$$\Delta u = 0, \qquad x > 0, \qquad\qquad u(0,y) = h(y), \tag{16.92}$$

on the right half plane, we adopt the change of variables (16.90) and use the Poisson integral formula to construct the solution to the transformed Dirichlet problem

$$\Delta U = 0, \qquad \xi^2 + \eta^2 < 1, \qquad\qquad U(\cos\varphi, \sin\varphi) = H(\varphi), \tag{16.93}$$

on the unit disk. The boundary conditions are found as follows. Using the explicit form

$$x + \mathrm{i}\, y = z = \frac{1+\zeta}{1-\zeta} = \frac{(1+\zeta)(1-\overline{\zeta})}{|\,1-\zeta\,|^2} = \frac{1+\zeta-\overline{\zeta}-|\,\zeta\,|^2}{|\,1-\zeta\,|^2} = \frac{1-\xi^2-\eta^2+2\,\mathrm{i}\,\eta}{(\xi-1)^2+\eta^2}$$

for the inverse map, we see that the boundary point the boundary point $\zeta = \xi + \mathrm{i}\,\eta = e^{\,\mathrm{i}\,\varphi}$ on the unit circle ∂D will correspond to the boundary point

$$\mathrm{i}\,y = \frac{2\,\eta}{(\xi-1)^2+\eta^2} = \frac{2\,\mathrm{i}\,\sin\varphi}{(\cos\varphi-1)^2+\sin^2\varphi} = \frac{\mathrm{i}}{1-\cot\varphi} = \mathrm{i}\,\cot\frac{\varphi}{2} \tag{16.94}$$

on the imaginary axis $\partial R = \big\{\,\mathrm{Re}\,z = 0\,\big\}$. Thus, the boundary data $h(y)$ on ∂R corresponds to the boundary data

$$H(\varphi) = h\left(\cot\tfrac{1}{2}\,\varphi\right)$$

on the unit circle. The Poisson integral formula (15.44) can then be applied to solve the problem (16.93), from which we reconstruct the solution (16.91) to the boundary value problem (16.91) on the half plane.

For example, to solve the problem with the step function

$$u(0,y) = h(y) \equiv \begin{cases} 1, & y > 0, \\ 0, & y < 0, \end{cases}$$

as boundary data, the corresponding boundary data on the unit disk is a (periodic) step function

$$H(\varphi) = \begin{cases} 1, & 0 < \varphi < \pi, \\ 0, & \pi < \varphi < 2\,\pi, \end{cases}$$

with values $+1$ on the upper semicircle, -1 on the lower semicircle, and jump discontinuities at $\zeta = \pm 1$. According to the Poisson formula (15.44), the solution to the latter boundary value problem is given by

$$\begin{aligned} U(\xi,\eta) &= \frac{1}{2\,\pi}\int_0^{\pi} \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\varphi-\phi)}\,d\phi \\ &= \frac{1}{\pi}\left[\tan^{-1}\left(\frac{1+\rho}{1-\rho}\,\cot\frac{\varphi}{2}\right) + \tan^{-1}\left(\frac{1+\rho}{1-\rho}\,\tan\frac{\varphi}{2}\right)\right] \end{aligned} \qquad \text{where} \quad \begin{aligned} \xi &= \rho\,\cos\varphi, \\ \eta &= \rho\,\sin\varphi. \end{aligned}$$

Finally, we use (16.91) to construct the solution on the upper half plane. We shall spare the reader the messy details of the final formula. The result is depicted in Figure zpm1h■.

Remark: The solution to the preceding Dirichlet boundary value problem is not, in fact, unique, owing to the unboundedness of the domain. The solution that we pick out by using the conformal map to the unit disk is the one that remains bounded at ∞. There are other solutions, but they are unbounded as $|\,z\,| \to \infty$ and would correspond to solutions on the unit disk that have some form of delta function singularity in their boundary data at the point -1; see Exercise ■.

Example 16.39. *A non-coaxial cable*. The goal of this example is to determine the electrostatic potential inside a non-coaxial cylindrical cable with prescribed constant potential values on the two bounding cylinders; see Figure c2■. Assume for definiteness that the larger cylinder has radius 1, and centered at the origin, while the smaller cylinder has radius $\frac{2}{5}$, and is centered at $z = \frac{2}{5}$. The resulting electrostatic potential will be independent of the longitudinal coordinate, and so can be viewed as a planar potential in

the annular domain contained between two circles representing the cross-sections of our cylinders. The desired potential must satisfy the Dirichlet boundary value problem

$$\begin{aligned} \Delta u = 0, \qquad & |\, z \,| < 1 \quad \text{and} \quad \left| \, z - \tfrac{2}{5} \, \right| > \tfrac{2}{5}, \\ u = a, \quad |\, z \,| = 1, \qquad & u = b, \quad \left| \, z - \tfrac{2}{5} \, \right| = \tfrac{2}{5}. \end{aligned}$$

According to Example 16.36, the linear fractional transformation $\zeta = \dfrac{2\,z - 1}{z - 2}$ will map this non-concentric annular domain to the annulus $A_{.5,1} = \left\{ \, \frac{1}{2} < |\, \zeta \,| < 1 \, \right\}$, which is the cross-section of a coaxial cable. The corresponding transformed potential $U(\xi, \eta)$ has the given Dirichlet boundary conditions $U = a$ on $|\, \zeta \,| = \frac{1}{2}$ and $U = b$ on $|\, \zeta \,| = 1$. Clearly the coaxial potential U must be a radially symmetric solution to the Laplace equation, and hence, according to (15.59), of the form

$$U(\xi, \eta) = \alpha \, \log |\, \zeta \,| + \beta,$$

for constants α, β. A short computation shows that the particular potential function

$$U(\xi, \eta) = \frac{b - a}{\log 2} \, \log |\, \zeta \,| + b = \frac{b - a}{2 \log 2} \, \log(\xi^2 + \eta^2) + b$$

satisfies the prescribed boundary conditions. Therefore, the desired non-coaxial electrostatic potential

$$u(x, y) = \frac{b - a}{\log 2} \, \log \left| \, \frac{2\,z - 1}{z - 2} \, \right| + b = \frac{b - a}{2 \log 2} \, \log \left(\, \frac{(2x - 1)^2 + y^2}{(x - 2)^2 + y^2} \, \right) + b. \tag{16.95}$$

is given by composition with the given linear fractional transformation. The particular case $a = 0, b = 1$ is plotted in Figure coax■.

Remark: The same harmonic function solves the problem of determining the equilibrium temperature in an annular plate whose inner boundary is kept at a temperature $u = a$ while the outer boundary is kept at temperature $u = b$. One could also interpret this solution as the equilibrium temperature of a three-dimensional domain contained between two non-coaxial cylinders held at fixed temperatures. The latter temperature will only depend upon the transverse x, y coordinates and not upon the longitudinal coordinate.

Remark: A conformal map will also preserve Neumann boundary conditions, specifying the normal derivative $\partial u / \partial \mathbf{n} = h$ on the boundary. Indeed, since a conformal map preserves angles, it maps the normal to $\partial \Omega$ to the normal to ∂D at the image point. Therefore, the transformed harmonic function $U(\xi, \eta)$ will satisfy the Neumann conditions $\partial U / \partial \mathbf{n} = H$, where H is related to h via the same equation (16.89).

Applications to Fluid Flow

Conformal mappings are particularly useful in the analysis of planar ideal fluid flow. Recall that if $\chi(z) = \varphi(x, y) + \mathrm{i}\, \psi(x, y)$ is an analytic function that represents the complex potential function for a steady state fluid flow, then we can interpret its real part $\varphi(x, y)$ as the velocity potential, while the imaginary part $\psi(x, y)$ is the harmonic conjugate stream

function. The level curves of φ are the equipotential lines, and these are orthogonal to the level curves of ψ, which are the streamlines followed by the individual fluid particles — except at stagnation points where $\chi'(z) = 0$.

Applying a conformal map $\zeta = g(z)$ leads to a transformed complex potential $\Theta(\zeta) = \Phi(\xi, \eta) + \mathrm{i}\,\Psi(\xi, \eta)$, where $\Phi(\xi, \eta)$ is the potential function and $\Psi(\xi, \eta)$ the stream function on the new domain. A key fact is that the conformal map will take isopotential lines of φ to isopotential lines of Φ and streamlines of ψ to streamlines of Ψ. Conformality implies that the orthogonality relations among isopotentials and streamlines away from stagnation points is maintained.

Let us concentrate on the case of flow past a solid object. In three dimensions, the object is assumed to have a uniform shape in the axial direction, and so we can restrict our attention to a planar fluid flow around a closed, bounded planar subset $D \subset \mathbb{R}^2 \simeq \mathbb{C}$ representing the cross-section of our cylindrical object. The (complex) velocity and potential are defined on the complementary domain $\Omega = \mathbb{C} \setminus D$ occupied by the fluid. The ideal flow assumptions of incompressibility and irrotationality are reasonably accurate if the flow is laminar, i.e., far away from turbulent. Then the velocity potential $\varphi(x, y)$ will satisfy the Laplace equation $\Delta\varphi = 0$ in the exterior domain Ω. For a solid object, we should impose the homogeneous Neumann boundary conditions

$$\frac{\partial \varphi}{\partial \mathbf{n}} = 0 \qquad \text{on the boundary} \qquad \partial\Omega = \partial D, \tag{16.96}$$

indicating that there no fluid flux into the object. We note that, according to Exercise ■, a conformal map will automatically preserve the Neumann boundary conditions.

In addition, since the flow is taking place on an unbounded domain, we need to specify the fluid motion at large distances. We shall assume our object is placed in a uniform horizontal flow, as in Figure hflow■. Thus, at large distance, the flow will not be affected by the object, and so the velocity should approximate the uniform velocity field $\mathbf{v} = (\,1, 0\,)^T$, where, for simplicity, we choose our physical units so that the asymptotic speed of the fluid is equal to 1. Equivalently, the velocity potential should satisfy

$$\varphi(x, y) \;\approx\; x, \qquad \text{when} \qquad x^2 + y^2 \gg 0.$$

Remark: An alternative physical interpretation is that the fluid is at rest, while the object moves through the fluid at unit speed 1 in a horizontal direction. For example, think of an airplane flying through the air at constant speed. If we adopt a moving coordinate system by sitting inside the plane, then the effect is as if the object is sitting still while the air is moving towards us at unit speed.

Example 16.40. The simplest example is a flat plate moving through the fluid in a horizontal direction. The plate's cross-section is a horizontal line segment, and, for simplicity, we take it to be the segment $D = [\,-1, 1\,]$ lying on the real axis. If the plate is very thin, it will have absolutely no effect on the horizontal flow of the fluid, and, indeed, the velocity potential is given by

$$\varphi(x, y) = x, \qquad\qquad x + \mathrm{i}\,y \in \Omega = \mathbb{C} \setminus [\,-1, 1\,].$$

Note that $\nabla \varphi = (\,1, 0\,)^T$, and hence this flow satisfies the Neumann boundary conditions (16.96) on the horizontal segment $D = \partial\Omega$. The corresponding complex potential is $\chi(z) = z$, with complex velocity $f(z) = \chi'(z) = 1$.

Example 16.41. Recall that the Joukowski conformal map defined by the analytic function

$$\zeta = g(z) = \frac{1}{2}\left(z + \frac{1}{z} \right) \tag{16.97}$$

squashes the unit circle $|\, z \,| = 1$ down to the real line segment $[\,-1, 1\,]$ in the ζ plane. Therefore, it will map the fluid flow outside the unit disk (the cross-section of a circular cylinder) to the fluid flow past the line segment, which, according to the previous example, has complex potential $\Theta(\zeta) = \zeta$. As a result, the complex potential for the flow past a disk is the same as the Joukowski function

$$\chi(z) = \Theta \circ g(z) = g(z) = \frac{1}{2}\left(z + \frac{1}{z} \right). \tag{16.98}$$

Except for a factor of $\frac{1}{2}$, this agrees with the flow potential we derived in Example 16.18. The difference is that, at large distances, the current potential

$$\chi(z) \;\approx\; \tfrac{1}{2}\, z \qquad \text{for} \qquad |\, z \,| \gg 1.$$

corresponds to uniform horizontal flow whose velocity $\left(\, \frac{1}{2}, 0 \,\right)^T$ is half as fast. The discrepancy between the two flows can easily be rectified by multiplying (16.98) by 2, whose only effect is to speed up the flow.

Example 16.42. Let us next consider the case of a tilted plate in a uniformly horizontal fluid flow. Thus, the cross-section is the line segment

$$z(t) = t\, e^{\,\mathrm{i}\,\theta}, \qquad -1 \le t \le 1,$$

obtained by rotating the horizontal line segment $[\,-1, 1\,]$ through an angle θ, as in Figure tilt. The goal is to construct a fluid flow past the tilted segment that is asymptotically horizontal at large distance.

The critical observation is that, while the effect of rotating a plate in a fluid flow is not so evident, we can easily rotate the disk in the flow — since it is circularly symmetric, rotations don't affect it. Thus, the rotation $w = e^{-\,\mathrm{i}\,\theta}\, z$ maps the Joukowski potential (16.98) to the complex potential

$$\Upsilon(w) = \chi(e^{\,\mathrm{i}\,\theta}\, w) = \frac{1}{2}\left(e^{\,\mathrm{i}\,\theta}\, w + \frac{e^{-\,\mathrm{i}\,\theta}}{w} \right).$$

The streamlines of the induced flow are no longer asymptotically horizontal, but rather at an angle $-\theta$. If we now apply the original Joukowski map (16.97) to the rotated flow, the circle is again squashed down to the horizontal line segment, but the flow lines continue to be at angle $-\theta$ at large distances. Thus, if we then rotate the resulting flow through an angle θ, the net effect will be to tilt the segment to the desired angle θ while rotating

the streamlines to be asymptotically horizontal. Putting the pieces together, we have the final complex potential in the form

$$\chi(z) = e^{\,\mathrm{i}\,\theta}\left(z\,\cos\theta - \,\mathrm{i}\,\sin\theta\,\sqrt{z^2 - e^{-2\,\mathrm{i}\,\theta}}\right). \tag{16.99}$$

Sample streamlines for the flow at several attack angles are plotted in Figure tilt■.

Example 16.43. As we discovered in Example 16.31, applying the Joukowski map to off-center disks will, in favorable configurations, produce airfoil-shaped objects. The fluid motion around such airfoils can thus be obtained by applying the Joukowski map to the flow past such an off-center circle.

First, an affine map $w = \alpha\,z + \beta$ will have the effect of moving the original unit disk $|\,z\,| \le 1$ to the disk $|\,w - \beta\,| \le |\,\alpha\,|$ with center β and radius $|\,\alpha\,|$. In particular, the boundary circle will continue to pass through the point $w = 1$ provided $|\,\alpha\,| = |\,1 - \beta\,|$. Moreover, as noted in Example 16.21, the angular component of α has the effect of a rotation, and so the streamlines around the new disk will, asymptotically, be at an angle $\varphi = \mathrm{ph}\,\alpha$ with the horizontal. We then apply the Joukowski transformation

$$\zeta = \frac{1}{2}\left(w + \frac{1}{w}\right) = \frac{1}{2}\left(\alpha\,z + \beta + \frac{1}{\alpha\,z + \beta}\right) \tag{16.100}$$

to map the disk to the airfoil shape. The resulting complex potential for the flow past the airfoil is obtained by substituting the inverse map

$$z = \frac{w - \beta}{\alpha} = \frac{\zeta - \beta + \sqrt{\zeta^2 - 1}}{\alpha}\,,$$

into the original potential (16.98), whereby

$$\Theta(\zeta) = \frac{1}{2}\left(\frac{\zeta - \beta + \sqrt{\zeta^2 - 1}}{\alpha} + \frac{\alpha(\zeta - \beta - \sqrt{\zeta^2 - 1}\,)}{\beta^2 + 1 - 2\,\beta\,\zeta}\right).$$

Since the streamlines have been rotated through an angle $\varphi = \mathrm{ph}\,\alpha$, we then rotate the final result back by multiplying by $e^{\,\mathrm{i}\,\varphi}$ in order to see the effect of the airfoil tiled at an angle $-\varphi$ in a horizontal flow. Sample streamlines are graphed in Figure airfoilnolift■.

We can interpret all these examples as planar cross-sections of three-dimensional fluid flows past an airplane wing oriented in the longitudinal z direction. The wing is assumed to have a uniform cross-section shape, and the flow not dependent upon the axial z coordinate. For wings that are sufficiently long and reasonable (laminar) flows, this model will be valid away from the wing tips. More complicated airfoils with varying cross-section and faster flows require a fully three-dimensional fluid model. For such problems, complex analysis is no longer applicable, and, for the most part, one must rely on numerical integration techniques. Only in recent years have computers become sufficiently powerful to compute realistic three-dimensional fluid motions — and then only in reasonable "mild" scenarios†.

† The definition of mild relies on the magnitude of the Reynolds number, [**15**].

The two-dimensional versions that have been analyzed here still provide important clues to the behavior of a three-dimensional flow, as well as useful approximations and starting points for the three-dimensional airplane wing design problem.

Unfortunately, there is a major flaw with the airfoils that we have just designed. Potential flows do not produce any lift, and hence the theory indicates that the airplane will not fly. In order to understand how lift enters into the picture, we need to study complex integration, and so we will return to this example later. In Example 16.57, we shall construct an alternative flow past an airfoil that continues to have the correct asymptotic behavior at large distances, while inducing a nonzero lift. The latter holds the secret to flight.

Poisson's Equation and the Green's Function

Although designed for solving the homogeneous Laplace equation, the method of conformal mapping can also be used to solve its inhomogeneous counterpart — the Poisson equation. As we learned in Chapter 15, to solve an inhomogeneous boundary value problem $-\Delta u = f$ on a domain Ω it suffices to solve the particular versions $-\Delta u = \delta_\zeta$ whose right hand side is a unit impulse concentrated at a point $\zeta = \xi + \mathrm{i}\,\eta \in \Omega$. The resulting solution $u(x,y) = G_\zeta(x,y) = G(x,y;\xi,\eta)$ is the Green's function for the given boundary value problem. The solution to the boundary value problem associated with a more general external forcing $f(x,y)$ is then given by a superposition principle

$$u(x,y) = \iint_\Omega G(x,y;\xi,\eta)\, f(\xi,\eta)\, d\xi\, d\eta. \tag{16.101}$$

For the planar Poisson equation, the starting point is the logarithmic potential function

$$u(x,y) = \frac{1}{2\,\pi} \log|\, z\,| = \mathrm{Re}\, \frac{1}{2\,\pi} \log z, \tag{16.102}$$

which is the solution to the Dirichlet problem

$$-\Delta u = \delta_0(x,y), \qquad (x,y) \in D, \qquad\qquad u = 0 \qquad \text{on} \qquad \partial D,$$

on the unit disk D for an impulse concentrated at the origin; see Section 15.3 for details. How do we obtain the corresponding solution when the unit impulse is concentrated at another point $\zeta = \xi + \mathrm{i}\,\eta \in D$ instead of the origin? According to Example 16.26, the linear fractional transformation

$$w = g(z) = \frac{z - \zeta}{\overline{\zeta}\, z - 1}, \qquad \text{where} \qquad |\,\zeta\,| < 1, \tag{16.103}$$

maps the unit disk to itself, moving the point $z = \zeta$ to the origin $w = g(\zeta) = 0$. The logarithmic potential $U = \dfrac{1}{2\,\pi}\, \log|\, w\,|$ will thus be mapped to the Green's function

$$G(x,y;\xi,\eta) = \frac{1}{2\,\pi}\, \log\left|\, \frac{z - \zeta}{\overline{\zeta}\, z - 1}\,\right| \tag{16.104}$$

at the point $\zeta = \xi + \mathrm{i}\,\eta$. Indeed, by the properties of conformal mapping, since U is harmonic except at the singularity $w = 0$, the function (16.104) will also be harmonic except at the image point $z = \zeta$. The fact that the mapping does not affect the delta function singularity is not hard to check; see Exercise ∎. Moreover, since the conformal map does not alter the boundary $|\,z\,| = 1$, the function (16.104) continues to satisfy the homogeneous Dirichlet boundary conditions.

Formula (16.104) reproduces the Poisson formula (15.70) for the Green's function that we derived previously using the method of images. This identification can be verified by substituting $z = r\,e^{\mathrm{i}\,\theta}$, $\zeta = \rho\,e^{\mathrm{i}\,\varphi}$, or, more simply, by noting that the numerator in the logarithmic fraction gives the potential due to a unit impulse at $z = \zeta$, while the denominator represents the image potential at $z = 1/\,\overline{\zeta}$ required to cancel out the effect of the interior potential on the boundary of the unit disk.

Now that we know the Green's function on the unit disk, we can use the methods of conformal mapping to produce the Green's function for any other simply connected domain $\Omega \subsetneq \mathbb{C}$. Let $w = g(z)$ denote the conformal map that takes the domain $z \in \Omega$ to the unit disk $w \in D$, guaranteed by the Riemann Mapping Theorem 16.33. The Green's function associated with homogeneous Dirichlet boundary conditions on Ω is explicitly given by

$$G(z;\zeta) = \frac{1}{2\,\pi}\,\log\left|\,\frac{g(z) - g(\zeta)}{\overline{g(\zeta)}\,g(z) - 1}\,\right|. \tag{16.105}$$

Example 16.44. For example, according to Example 16.25, the analytic function

$$w = \frac{z-1}{z+1}$$

maps the right half plane $x = \mathrm{Re}\,z > 0$ to the unit disk $|\,\zeta\,| < 1$. Therefore, by (16.105), the Green's function for the right half plane has the form

$$G(z;\zeta) = \frac{1}{2\,\pi}\,\log\left|\,\frac{\dfrac{z-1}{z+1} - \dfrac{\zeta-1}{\zeta+1}}{\dfrac{z-1}{z+1}\,\dfrac{\overline{\zeta}-1}{\overline{\zeta}+1} - 1}\,\right| = \frac{1}{2\,\pi}\,\log\left|\,\frac{(\overline{\zeta}+1)(z-\zeta)}{(z+1)(z-\overline{\zeta})}\,\right|. \tag{16.106}$$

One can then write the solution to the Poisson equation in a superposition as in (16.101).

16.5. Complex Integration.

All of the magic and power of calculus ultimately rests on the amazing fact that differentiation and integration are mutually inverse operations. And, just as complex functions have many remarkable differentiability properties not enjoyed by their real siblings, so complex integration theory has a extra beauty and structure beyond its more mundane real counterpart. In the remaining two sections of this chapter, we shall develop the basics of complex integration theory and discuss some of its important applications.

First, let us motivate the definition of a complex integral. As you know, the integral of a real function, $\displaystyle\int_a^b f(t)\,dt$, is usually taken along a real interval $[\,a,b\,] \subset \mathbb{R}$. In complex

function theory, integrals are taken along curves in the complex plane, and are thus intimately related to the line integrals appearing in real vector calculus. The identification of a complex number $z = x + \mathrm{i}\, y$ with a planar vector $\mathbf{x} = (\, x, y\,)^T$ will immediately connect the two concepts.

Consider a curve C in the complex plane, parametrized, as in (16.70), by $z(t) = x(t) + \mathrm{i}\, y(t)$ for $a \le t \le b$. We define the *complex integral* of a complex function $f(z)$ along the curve C to be

$$\int_C f(z)\, dz = \int_a^b f(z(t))\, \frac{dz}{dt}\, dt. \tag{16.107}$$

We shall always assume that the integrand $f(z)$ is a well-defined complex function at each point on the curve. The result of complex integration of a function along a curve is a complex number. Let us write out the integrand

$$f(z) = u(x,y) + \mathrm{i}\, v(x,y)$$

in terms of its real and imaginary parts. Also, note that

$$dz = \frac{dz}{dt}\, dt = \left(\frac{dx}{dt} + \mathrm{i}\, \frac{dy}{dt} \right) dt = dx + \mathrm{i}\, dy.$$

In this manner, we discover that the complex integral (16.107) splits up into two real line integrals

$$\int_C f(z)\, dz = \int_C (u + \mathrm{i}\, v)(dx + \mathrm{i}\, dy) = \int_C (u\, dx - v\, dy) + \mathrm{i} \int_C (v\, dx + u\, dy). \tag{16.108}$$

Example 16.45. Let us compute complex integrals

$$\int_C z^n\, dz, \tag{16.109}$$

of the monomial function $f(z) = z^n$, where n is an integer, along several different curves. We begin with the case when the integration curve C is the straight line segment along the real axis connecting the points -1 to 1, which we parametrize by $z(t) = t$ for $-1 \le t \le 1$. The defining formula (16.107) implies that the complex integral (16.109) reduces to a real integral:

$$\int_C z^n\, dz = \int_{-1}^1 t^n\, dt = \begin{cases} 0, & n = 2\,k \ge 0 \text{ is even} \\ \dfrac{2}{n+1}, & n = 2\,k+1 > 0 \text{ is odd.} \end{cases},$$

If $n \le -1$ is negative, then the singularity of the integrand at at the origin prevents the integral from converging, and so the complex integral is not defined.

Let us evaluate the same complex integral, but now along a parabolic arc P parametrized by

$$z(t) = t + \mathrm{i}\, (t^2 - 1), \qquad -1 \le t \le 1.$$

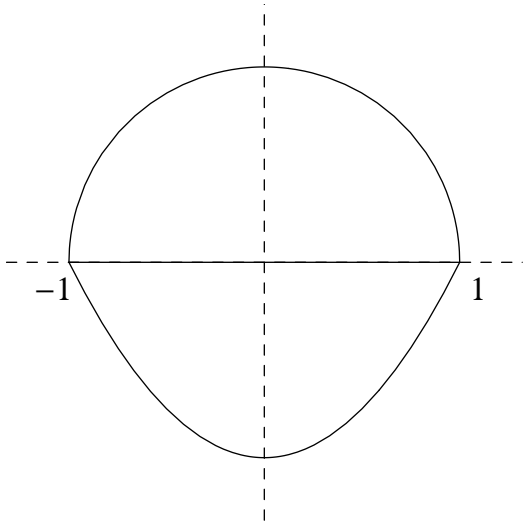


Figure 16.22. Curves for Complex Integration.

Note that, as graphed in Figure 16.22, the parabola connects the same two points. We again refer back to the basic definition (16.107) to evaluate the integral, so

$$\int_P z^n \, dz = \int_{-1}^{1} \big[\, t + \mathrm{i}\,(t^2 - 1) \,\big]^n (1 + 2\,\mathrm{i}\,t)\, dt.$$

We could, at this point, expand the resulting complex polynomial integrand, and then integrate term by term. A more elegant approach is to recognize that the integrand is an exact derivative; namely, by the chain rule

$$\frac{d}{dt}\, \frac{\big[\, t + \mathrm{i}\,(t^2 - 1) \,\big]^{n+1}}{n+1} = \big[\, t + \mathrm{i}\,(t^2 - 1) \,\big]^n (1 + 2\,\mathrm{i}\,t),$$

as long as $n \neq -1$. Therefore, we can use the Fundamental Theorem of Calculus (which works equally well for real integrals of complex-valued functions), to evaluate

$$\int_P z^n \, dz = \left. \frac{\big[\, t + \mathrm{i}\,(t^2 - 1) \,\big]^{n+1}}{n+1} \right|_{t=-1}^{1} = \begin{cases} 0, & n = 2\,k \text{ even}, \\ \dfrac{2}{n+1}, & -1 \neq n = 2\,k+1 \text{ odd}. \end{cases}$$

Thus, when $n \geq 0$ is a positive integer, we obtain the same result as before. Interestingly, in this case the complex integral is well-defined even when n is a negative integer because, unlike the real line segment, the parabolic path does *not* go through the singularity of z^n at $z = 0$. The case $n = -1$ needs to be done slightly differently. The integration of $1/z$ along the parabolic path is left as an exercise for the reader — one that requires some care. We recommend trying the exercise now, and then verifying your answer once we have become a little more familiar with basic complex integration techniques.

Finally, let us try integrating around a semi-circular arc, again with the same endpoints

-1 and 1. If we parametrize the semi-circle S^+ by $z(t) = e^{\,\mathrm{i}\,t}$, $0 \le t \le \pi$, we find

$$\int_{S^+} z^n\,dz = \int_0^\pi z^n\,\frac{dz}{dt}\,dt = \int_0^\pi e^{\,\mathrm{i}\,nt}\,\mathrm{i}\,e^{\,\mathrm{i}\,t}\,dt = \int_0^\pi \mathrm{i}\,e^{\,\mathrm{i}\,(n+1)t}\,dt$$
$$= \frac{e^{\,\mathrm{i}\,(n+1)t}}{n+1}\;\Big|_{t=0}^{\pi} = \frac{1-e^{\,\mathrm{i}\,(n+1)\pi}}{n+1} = \begin{cases} 0, & n = 2\,k \text{ even},\\ -\dfrac{2}{n+1}, & -1 \neq n = 2\,k+1 \text{ odd}. \end{cases}$$

This value is the negative of the previous cases — but this can be explained by the fact that the circular arc is oriented to go *from* 1 *to* -1 whereas the line segment and parabola both go *from* -1 *to* 1. Just as with line integrals, the direction of the curve determines the sign of the complex integral; if we reverse direction, replacing t by $-t$, we end up with the same value as the preceding two complex integrals. Moreover — again provided $n \neq -1$ — it does not matter whether we use the upper semicircle or lower semicircle to go from -1 to 1 — the result is exactly the same. However, this remark does *not* apply to the case $n = -1$. Integrating along the upper semicircle S^+ from 1 to -1 yields

$$\int_{S^+} \frac{dz}{z} = \int_0^\pi \mathrm{i}\,dt = \pi\,\mathrm{i}\,, \tag{16.110}$$

whereas integrating along the lower semicircle S^- from 1 to -1 yields the negative

$$\int_{S^-} \frac{dz}{z} = \int_0^{-\pi} \mathrm{i}\,dt = -\pi\,\mathrm{i}\,. \tag{16.111}$$

Hence, when integrating the function $1/z$, it makes a difference which direction we go around the origin.

Integrating z^n for any integer $n \neq -1$ around an entire circle gives zero — irrespective of the radius. This can be seen as follows. We parametrize a circle of radius r by $z(t) = re^{\,\mathrm{i}\,t}$ for $0 \le t \le 2\,\pi$. Then, by the same computation,

$$\oint_C z^n\,dz = \int_0^{2\,\pi} (r^n e^{\,\mathrm{i}\,nt})(r\,\mathrm{i}\,e^{\,\mathrm{i}\,t})\,dt = \int_0^{2\,\pi} \mathrm{i}\,r^{n+1} e^{\,\mathrm{i}\,(n+1)t}\,dt = \frac{r^{n+1}}{n+1}\,e^{\,\mathrm{i}\,(n+1)t}\,\Big|_{t=0}^{2\,\pi} = 0, \tag{16.112}$$

provided $n \neq -1$. Here, as in Appendix A, the circle on the integral sign serves to remind us that we are integrating around a closed curve. The case $n = -1$ remains special. Integrating once around the circle in the counter-clockwise direction yields a nonzero result

$$\oint_C \frac{dz}{z} = \int_0^{2\,\pi} \mathrm{i}\,dt = 2\,\pi\,\mathrm{i}\,. \tag{16.113}$$

Let us note that a complex integral does not depend on the particular parametrization of the curve C. It does, however, depend upon the orientation of the curve: if we traverse the curve in the reverse direction, then the complex integral changes its sign. Moreover, if we chop up the curve into two non-overlapping pieces, $C = C_1 \cup C_2$ with a common orientation, then, just as with a line integral (A.39), line integrals can be decomposed into

a sum over the pieces:

$$\int_{-C} f(z)\,dz = -\int_{C} f(z)\,dz, \qquad f(z) = \int_{C_1} f(z)\,dz \;+\; \int_{C_2} f(z)\,dz, \qquad C = C_1 \cup C_2. \tag{16.114}$$

For instance, the integral (16.113) of $1/z$ around the circle is the difference of the individual semicircular integrals (16.110), (16.111); the lower semicircular integral acquires a negative sign to switch its orientation to agree with that of the entire circle.

Note: In complex integration theory, a simple closed curve is often referred to as a *contour*, and so complex integration is sometimes referred to as *contour integration*. Unless explicitly stated, we always go around contours in the *counter-clockwise* direction.

Further experiments of this type lead us to suspect that complex integrals are usually path-independent, and hence evaluate to zero around closed contours. One must be careful, though, as the integral (16.113) makes clear. Path independence, in fact, follows from the complex version of the Fundamental Theorem of Calculus.

Theorem 16.46. *Let $f(z) = F'(z)$ be the derivative of a single-valued complex function on a domain $\Omega \subset \mathbb{C}$. Let $C \subset \Omega$ be any curve with initial point α and final point β. Then*

$$\int_C f(z)\,dz = \int_C F'(z)\,dz = F(\beta) - F(\alpha). \tag{16.115}$$

Proof: This follows immediately from the definition (16.107) and the chain rule:

$$\int_C F'(z)\,dz = \int_a^b F'(z(t))\,\frac{dz}{dt}\,dt = \int_a^b \frac{d}{dt}\,F(z(t))\,dt = F(z(b)) - F(z(a)) = F(\beta) - F(\alpha),$$

where $\alpha = z(a)$ and $\beta = z(b)$ are the endpoints of the curve. *Q.E.D.*

For example, when $n \neq -1$, the function $f(z) = z^n$ is the derivative of the single-valued function $F(z) = \dfrac{1}{n+1} z^{n+1}$. Hence

$$\int_C z^n\,dz \;=\; \frac{\beta^{n+1}}{n+1} \;-\; \frac{\alpha^{n+1}}{n+1}$$

whenever C is a curve connecting α to β. When $n < 0$, the curve is not allowed to pass through the origin $z = 0$, which is a singularity for z^n. Our earlier computations are special cases of this result.

In contrast, the function $f(z) = 1/z$ is the derivative of

$$\log z = \log |\, z \,| + \mathrm{i}\ \mathrm{ph}\, z,$$

but the complex logarithm is no longer single-valued on all of $\mathbb{C}\setminus\{0\}$, and so Theorem 16.46 cannot be applied directly. However, if our curve is contained within a simply connected subdomain that does not include the origin, $0 \notin \Omega \subset \mathbb{C}$, then we can use *any* single-valued branch of the logarithm to evaluate the integral

$$\int_C \frac{dz}{z} = \log \beta - \log \alpha,$$

where α, β are the endpoints of the curve. Since the common multiples of $2\pi\,\mathrm{i}$ cancel, the answer does not depend upon which particular branch of the logarithm is chosen, but we do need to be consistent in our choice. For example, on the upper semicircle S^+ of radius 1 going from 1 to -1,

$$\int_{S^+} \frac{dz}{z} = \log(-1) - \log 1 = \pi\,\mathrm{i}\,,$$

where we use the branch of $\log z = \log|z| + \mathrm{i}\,\mathrm{ph}\,z$ with $0 \le \mathrm{ph}\,z \le \pi$. On the other hand, if we integrate on the lower semi-circle S^- going from 1 to -1, we need to adopt a different branch, say that with $-\pi \le \mathrm{ph}\,z \le 0$. With this choice, the integral becomes

$$\int_{S^-} \frac{dz}{z} = \log(-1) - \log 1 = -\pi\,\mathrm{i}\,,$$

thus reproducing (16.110), (16.111). Pay particular attention to the different values of $\log(-1)$ in the two cases!

The most important consequence of Theorem 16.46 is that, as long as the integrand $f(z)$ has a single-valued anti-derivative, its complex integral is independent of the path connecting two points — the value only depends on the endpoints of the curve and not how one gets from point α to point β.

Theorem 16.47. *If $f(z) = F'(z)$ for $z \in \Omega$ and $C \subset \Omega$ is any closed curve, then*

$$\oint_C f(z)\,dz = 0. \tag{16.116}$$

Conversely, if (16.116) *holds for all closed curves $C \subset \Omega$ contained in the domain of definition of $f(z)$, then f admits a single-valued complex anti-derivative with $F'(z) = f(z)$.*

Proof: We have already demonstrated the first statement. As for the second, we define

$$F(z) = \int_{z_0}^{z} f(z)\,dz,$$

where $z_0 \in \Omega$ is any fixed point, and we choose any convenient curve $C \subset \Omega$ connecting[†] z_0 to z. (16.116) assures us that the value does not depend on the chosen path. The proof that this formula does define an anti-derivative of f is left as an exercise, which can be solved in the same fashion as the case of a real line integral, cf. (22.26). *Q.E.D.*

The preceding considerations suggest the following fundamental theorem, due in its general form to Cauchy. Before stating it, we introduce the convention that a complex function $f(z)$ will be called *analytic on a domain* $\Omega \subset \mathbb{C}$ provided it is analytic at every point inside Ω and, in addition, is continuous up to and including its boundary $\partial\Omega$. When Ω is bounded, its boundary $\partial\Omega$ consists of one or more simple closed curves. In general, we orient $\partial\Omega$ so that the domain is always on our left hand side. This means that the outermost boundary curve is traversed in the counter-clockwise direction, but any interior holes are take on a clockwise orientation. Our convention is depicted in Figure bdy■.

[†] This assumes Ω is a connected domain; otherwise, apply the result to its individual connected components.

Theorem 16.48. *If $f(z)$ is analytic on a bounded domain $\Omega \subset \mathbb{C}$, then*

$$\oint_{\partial\Omega} f(z)\,dz = 0. \tag{16.117}$$

Proof: If we apply Green's Theorem A.25 to the two real line integrals in (16.108), we find

$$\oint_{\partial\Omega} u\,dx - v\,dy = \iint_{\Omega} \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) = 0,$$
$$\oint_{\partial\Omega} v\,dx + u\,dy = \iint_{\Omega} \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) = 0,$$

both of which vanish by virtue of the Cauchy–Riemann equations (16.22). *Q.E.D.*

If the domain of definition of our complex function $f(z)$ is simply connected, then, by definition, the interior of any closed curve $C \subset \Omega$ is contained in Ω, and hence Cauchy's Theorem 16.48 implies the path independence of the complex integral within Ω.

Corollary 16.49. *If $f(z)$ is analytic on a simply connected domain $\Omega \subset \mathbb{C}$, then its complex integral $\displaystyle\int_C f(z)\,dz$ for $C \subset \Omega$ is independent of path. In particular,*

$$\oint_C f(z)\,dz = 0 \tag{16.118}$$

for any closed curve $C \subset \Omega$.

Remark: This result also admits a converse: a continuous function $f(z)$ that satisfies (16.118) for *all* closed curves is necessarily analytic. See [**4**] for a proof.

We will also require a slight generalization of this result.

Lemma 16.50. *If $f(z)$ is analytic in a domain that contains two simple closed curves S and C, and the entire region lying between them, then, assuming they are oriented in the same direction,*

$$\oint_C f(z)\,dz = \oint_S f(z)\,dz. \tag{16.119}$$

Proof: If C and S do not cross each other, we let Ω denote the domain contained between them, so that $\partial\Omega = C \cup S$; see Figure oints■. According to Cauchy's Theorem 16.48, $\displaystyle\oint_{\partial\Omega} f(z) = 0$. Now, our orientation convention for $\partial\Omega$ means that the outer curve, say C, is traversed in the counter-clockwise direction, while the inner curve S has the opposite, clockwise orientation. Therefore, if we assign both curves the same counter-clockwise orientation,

$$0 = \oint_{\partial\Omega} f(z) = \oint_C f(z)\,dz - \oint_S f(z)\,dz,$$

proving (16.119).

If the two curves cross, we can construct a nearby curve $K \subset \Omega$ that neither crosses, as in Figure c2■. By the preceding paragraph, each integral is equal to to that over the third curve,

$$\oint_C f(z)\,dz = \oint_K f(z)\,dz = \oint_S f(z)\,dz,$$

and formula (16.119) remains valid. *Q.E.D.*

Example 16.51. Consider the function $f(z) = z^n$ where n is an integer[†]. In (16.112), we already computed

$$\oint_C z^n\,dz = \begin{cases} 0, & n \neq -1, \\ 2\pi\,\mathrm{i}\,, & n = -1, \end{cases} \tag{16.120}$$

when C is a circle centered at $z = 0$. When $n \geq 0$, Theorem 16.46 implies that the integral of z^n is 0 over *any* closed curve in the plane. The same applies in the cases $n \leq -2$ provided the curve does not pass through the singular point $z = 0$. In particular, the integral is zero around closed curves encircling the origin, even though z^n for $n \leq -2$ has a singularity inside the curve and so Cauchy's Theorem 16.48 does not apply as stated.

The case $n = -1$ has particular significance. Here, Lemma 16.50 implies that the integral is the same as the integral around a circle — provided the curve C also goes once around the origin in a counter-clockwise direction. Thus (16.113) holds for any closed curve that goes counter-clockwise once around the origin. More generally, if the curve goes several times around the origin[‡], then

$$\oint_C \frac{dz}{z} = 2k\pi\,\mathrm{i} \tag{16.121}$$

is an integer multiple of $2\pi\,\mathrm{i}$. The integer k is called the *winding number* of the curve C, and measures the total number of times C goes around the origin. For instance, if C winds three times around 0 in a counter-clockwise fashion, then $k = 3$, while $k = -5$ indicates that the curve winds 5 times around 0 in a clockwise direction, as in Figure wind■. In particular, a winding number $k = 0$ indicates that C is not wrapped around the origin. For example, if C is viewed as a loop of string wrapped around a pole (the *pole* of $1/z$ at 0) then a winding number $k = 0$ would indicate that the string can be disentangled from the pole without cutting; nonzero winding numbers would indicate that the string is truly entangled[§].

[†] When n is fractional or irrational, the integrals are not well-defined owing to the branch point singularity at the origin.

[‡] Such a curve is definitely not simple and must necessarily cross over itself.

[§] Actually, there are more subtle three-dimensional considerations that come into play, and even strings with zero winding number cannot be removed from the pole without cutting if they are linked in some nontrivial manner, cf. [**92**].

Lemma 16.52. *If C is any simple closed curve, and a is any point not lying on C, then*

$$\oint_C \frac{dz}{z-a} = \begin{cases} 2\pi\,\mathrm{i}\,, & a \quad \text{inside } C \\ 0 & a \quad \text{outside } C. \end{cases} \tag{16.122}$$

If $a \in C$, then the integral does not converge.

Proof: Note that the integrand $f(z) = 1/(z-a)$ is analytic everywhere except at $z = a$, where it has a simple pole. If a is outside C, then Cauchy's Theorem 16.48 applies, and the integral is zero. On the other hand, if a is inside C, then Lemma 16.50 implies that the integral is equal to the integral around a circle centered at $z = a$. The latter integral can be computed directly by using the parametrization $z(t) = a + r\,e^{\,\mathrm{i}\,t}$ for $0 \le t \le 2\pi$, as in (16.113). *Q.E.D.*

Example 16.53. Let $D \subset \mathbb{C}$ be a closed and *connected* domain. Let $a, b \in D$ be two points in D. Then

$$\oint_C \left(\frac{1}{z-a} - \frac{1}{z-b} \right) dz = \oint_C \frac{dz}{z-a} - \oint_C \frac{dz}{z-b} = 0$$

for any closed curve $C \subset \Omega = \mathbb{C} \setminus D$ lying outside the domain D. This is because, by connectivity of D, either C contains both points in its interior, in which case both integrals equal $2\pi\,\mathrm{i}\,$, or C contains neither point, in which case both integrals are 0. Theorem 16.47 implies that the integrand admits a single-valued anti-derivative on the domain Ω. On the other hand, each individual term is the derivative of a multiply-valued complex logarithm. The conclusion is that, even though the individual logarithms are multiply-valued, their difference

$$F(z) = \log(z-a) - \log(z-b)$$

is a consistent, single-valued complex function on all of $\Omega = \mathbb{C} \setminus D$. There are, in fact, an infinite number of possible values, differing by integer multiples of $2\pi\,\mathrm{i}\,$. However, assigning a value at one point in Ω leads to a consistent and continuous definition on the entire domain Ω. Again, this requires that D is connected; the conclusion is *not* true, say, for the twice-punctured plane $\mathbb{C} \setminus \{a, b\}$.

We are sometimes interested in estimating the size of a complex integral. The basic inequality bounds it in terms of an arc length integral.

Proposition 16.54. *The modulus of the integral of the complex function f along a curve C is bounded by the integral of its modulus with respect to arc length:*

$$\left| \int_C f(z)\,dz \right| \;\le\; \int_C |\,f(z)\,|\,ds. \tag{16.123}$$

Proof: We begin with a simple lemma about real integrals of complex functions.

Lemma 16.55. *If $f(t)$ is a complex-valued function depending on the real variable $a \le t \le b$, then*

$$\left| \int_a^b f(t)\,dt \right| \;\le\; \int_a^b |\,f(t)\,|\,dt. \tag{16.124}$$

Proof: If $\displaystyle\int_a^b f(t)\,dt = 0$, the inequality is trivial. Otherwise, let $\theta = \operatorname{ph} \displaystyle\int_a^b f(t)\,dt$. Then, using Exercise ■,

$$\left| \int_a^b f(t)\,dt \right| = \operatorname{Re}\left[e^{-\mathrm{i}\,\theta} \int_a^b f(t)\,dt \right] = \int_a^b \operatorname{Re}\left[e^{-\mathrm{i}\,\theta} f(t) \right] dt \le \int_a^b |\, f(t)\,|\,dt,$$

which proves the lemma. *Q.E.D.*

To prove the proposition, we write out the complex integral, and use (16.124) as follows:

$$\left| \int_C f(z)\,dz \right| = \left| \int_a^b f(z(t))\,\frac{dz}{dt}\,dt \right| \le \int_a^b |\, f(z(t))\,| \left| \frac{dz}{dt} \right| dt = \int_C |\, f(z)\,|\,ds,$$

since $|\,dz\,| = |\,\dot z\,|\,dt = \sqrt{\dot x^2 + \dot y^2}\,dt = ds$ is the arc length integral element (A.30). *Q.E.D.*

Corollary 16.56. *If C has length $L = \mathcal{L}(C)$, and $f(z)$ is an analytic function such that $|\, f(z)\,| \le M$ for all points $z \in C$, then*

$$\left| \int_C f(z)\,dz \right| \le M\,L. \tag{16.125}$$

Lift and Circulation

In fluid mechanical applications, the complex integral can be assigned an important physical interpretation. As above, we consider the steady state flow of an incompressible, irrotational fluid. Let $f(z) = u(x,y) - \mathrm{i}\,v(x,y)$ denote the complex velocity corresponding to the real velocity vector $\mathbf{v} = (\,u(x,y), v(x,y)\,)^T$ at the point (x,y).

As we noted in (16.108), the integral of the complex velocity $f(z)$ along a curve C can be written as a pair of real line integrals. In the present situation,

$$\int_C f(z)\,dz = \int_C (u - \mathrm{i}\,v)(dx + \mathrm{i}\,dy) = \int_C (u\,dx + v\,dy) - \mathrm{i} \int_C (v\,dx - u\,dy). \tag{16.126}$$

According to (A.37), (A.42), the real part is the circulation integral

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C u\,dx + v\,dy, \tag{16.127}$$

while the imaginary part is minus the flux integral

$$\int_C \mathbf{v} \cdot \mathbf{n}\,ds = \int_C \mathbf{v} \wedge d\mathbf{x} = \int_C v\,dx - u\,dy, \tag{16.128}$$

for the associated steady state fluid flow!

If the complex velocity admits a single-valued complex potential

$$\chi(z) = \varphi(z) - \mathrm{i}\,\psi(z), \qquad \text{where} \qquad \chi'(z) = f(z)$$

— which is always the case if its domain of definition is simply connected — then the complex integral is independent of path, and one can use the Fundamental Theorem 16.46 to evaluate it:

$$\int_C f(z)\,dz = \chi(\beta) - \chi(\alpha) \tag{16.129}$$

for any curve C connecting α to β. Path independence of the complex integral immediately reconfirms the path independence of the flux and circulation integrals for irrotational, incompressible fluid dynamics. The real part of formula (16.129) evaluates the circulation integral

$$\int_C \mathbf{v}\cdot d\mathbf{x} = \int_C \nabla\varphi\cdot\, d\mathbf{x} = \varphi(\beta) - \varphi(\alpha), \tag{16.130}$$

as the difference in the values of the (real) potential at the endpoints α, β of the curve C. On the other hand, the imaginary part of formula (16.129) computes the flux integral

$$\int_C \mathbf{v}\wedge d\mathbf{x} = \int_C \nabla\psi\cdot\, d\mathbf{x} = \psi(\beta) - \psi(\alpha), \tag{16.131}$$

as the difference in the values of the stream function at the endpoints of the curve. Thus, the stream function acts as a "flux potential" for the flow, with the flux being independent of path. In particular, if C is a closed contour,

$$\oint_C \mathbf{v}\cdot d\mathbf{x} \;=\; 0 \;=\; \oint_C \mathbf{v}\wedge d\mathbf{x}, \tag{16.132}$$

and so there is no net circulation or flux along any closed curve in this situation.

In aerodynamics, lift is the result of the circulation of the fluid (air) around the body, [**15**, **135**]. More precisely, let $D \subset \mathbb{C}$ be a closed, bounded subset representing the cross-section of a cylindrical body, e.g., an airplane wing. The velocity vector field $\mathbf{v}$ of a steady state flow around the exterior of the body is defined on the domain $\Omega = \mathbb{C}\setminus D$. According to Blasius' Theorem, the body will experience a net lift if and only if it has nonvanishing circulation integral $\displaystyle\oint_C \mathbf{v}\cdot d\mathbf{x} \neq 0$, where C is any simple closed contour encircling the body. However, if the complex velocity admits a single-valued complex potential in Ω, then (16.132) tells us that the circulation is automatically zero, and so the body cannot experience any lift!

Example 16.57. Consider first the flow around a disk, as discussed in Examples 16.18 and 16.41. The Joukowski potential $\chi(z) = z + z^{-1}$ is a single-valued analytic function everywhere except the origin $z = 0$. Therefore, the circulation integral (16.130) around any contour encircling the disk will vanish, and hence the disk experiences no net lift. This is more or less evident from the Figure 16.11 graphing the streamlines of the flow; they are symmetric above and below the disk, and hence there cannot be any net force in the vertical direction.

Any conformal map will preserve the single-valuedness of the complex potentials, and hence will preserve the property of having zero circulation. In particular, all the flows past airfoils constructed in Example 16.43 also admit single-valued potentials, and so also have

zero circulation integral. Such an airplane will not fly, because its wings experience no lift! Of course, physical airplanes fly, and so there must be some physical assumption we are neglecting in our treatment of flow past a body. Abandoning incompressibility or irrotationality would take us outside the magical land of complex variable theory, and into the wilder regions of fully nonlinear partial differential equations of fluid mechanics. Moreover, although air is slightly compressible, water is, for all practical purposes, incompressible, and hydrofoils do experience lift when traveling through water.

The only way to introduce lift into the picture is through a (single-valued) complex velocity with a non-zero circulation integral, and this requires that its complex potential be multiply-valued. The one function that we know that has such a property is the complex logarithm

$$\lambda(z) = \log(a\,z + b), \qquad \text{whose derivative} \qquad \lambda'(z) = \frac{1}{a\,z + b}$$

is single-valued away from the singularity at $z = -b/a$. Thus, we are naturally led to introduce the family of complex potentials[†]

$$\chi_k(z) = \frac{1}{2}\left(z + \frac{1}{z}\right) - \mathrm{i}\,k\,\log z. \tag{16.133}$$

According to Exercise ■, the coefficient k must be real in order to maintain the no flux boundary conditions on the unit circle. By (16.126), the circulation is equal to the real part of the integral of the complex velocity

$$f_k(z) = \frac{d\chi_k}{dz} = \frac{1}{2} - \frac{1}{2\,z^2} - \frac{\mathrm{i}\,k}{z}\,. \tag{16.134}$$

By Cauchy's Theorem 16.48 coupled with formula (16.122), if C is a curve going once around the disk in a counter-clockwise direction, then

$$\oint_C f_k(z)\,dz = \oint_C \left(\frac{1}{2} - \frac{1}{2\,z^2} - \frac{\mathrm{i}\,k}{z}\right) dz = 2\,\pi\,k.$$

Therefore, when $\operatorname{Re} k \neq 0$, the circulation integral is non-zero, and the cylinder experiences a net lift. In Figure liftc■, the streamlines for the flow corresponding to a few representative values of k are plotted. Note the asymmetry of the streamlines that accounts for the lift experienced by the disk.

When we compose the modified lift potentials (16.133) with the Joukowski transformation (16.100), we obtain a complex potential

$$\Theta_k(\zeta) = \chi_k(z) \qquad \text{when} \qquad \zeta = \frac{1}{2}\left(w + \frac{1}{w}\right) = \frac{1}{2}\left(a\,z + \beta + \frac{1}{a\,z + \beta}\right)$$

[†] We center the logarithmic singularity at the origin in order to maintain the no flux boundary conditions on the unit circle. Moreover, Example 16.53 tells us that more than one logarithm in the potential is redundant, since the difference of any two logarithms is effectively a single-valued function, and hence contributes nothing to the circulation integral.

for flow around the corresponding airfoil — the image of the unit disk. The conformal mapping does not affect the value of the complex integrals, and hence, for any $k \neq 0$, there is a nonzero circulation around the airfoil under the modified fluid flow. This circulation is the cause of a net lift on the airfoil, and at last our airplane will fly!

However, there is now a slight embarrassment of riches, since we have now designed flows around the airfoil with an *arbitrary* value $2\pi k$ for the circulation integral, and hence having an *arbitrary* amount of lift! Which of these possible flows most closely realizes the true physical version with the correct amount of lift? In his 1902 thesis, the German mathematician Martin Kutta hypothesized that Nature chooses the constant k so as to keep the velocity of the flow at the trailing edge of the airfoil, namely $\zeta = 1$, to be finite. With some additional analysis, it turns out that this condition serves to uniquely specify k, and yields a reasonably good physical approximation to the actual lift of such an airfoil in flight, provided the tilt or attack angle of the airfoil in the flow is not too large. Further details, can be found in several references, including [**Fluid**, **93**, **135**].

16.6. Cauchy's Integral Formulae and the Calculus of Residues.

Cauchy's Integral Theorem 16.48 is a remarkably powerful result. It and its consequences underlie most important applications of complex integration. The fact that we can move the contours of complex integrals around freely — as long as we do not cross over singularities of the integrand — grants us great flexibility in their evaluation. Moreover, it leads to a method for evaluating a function *and* its derivatives through certain contour integrals.

As a consequence of Cauchy's Theorem, the value of a general complex integral around a closed contour depends only upon the nature of the singularities of the integrand that happen to lie inside the contour. This observation inspires us to develop a direct method, known as the "calculus of residues", for evaluating such integrals. The residue method effectively bypasses the Fundamental Theorem of Calculus — no antiderivatives are required! Remarkably, the method of residues can even be applied to evaluate certain types of real, definite integrals, as the final examples in this section shall demonstrate.

Cauchy's Integral Formula

The first important consequence of Cauchy's Theorem is the justly famous *Cauchy integral formulae*. It gives a formula for the value of an analytic function at a point as a certain contour integral around a closed curve encircling the point. It is worth emphasizing that Cauchy's formula is *not* a form of the Fundamental Theorem of Calculus, since we are reconstructing the function by *integration* — not its antiderivative! Cauchy's formula is a cornerstone of complex analysis. It has no real counterpart, once again underscoring the profound difference between the complex and real realms.

Theorem 16.58. *Let $\Omega \subset \mathbb{C}$ be a bounded domain with boundary $\partial\Omega$, and let $a \in \Omega$. If $f(z)$ is analytic on Ω, then*

$$f(a) = \frac{1}{2\pi\,\mathrm{i}} \oint_{\partial\Omega} \frac{f(z)}{z-a}\,dz. \tag{16.135}$$

Remark: As always, we traverse the boundary curve $\partial\Omega$ so that the domain Ω lies on our left. In most applications, Ω is simply connected, and so $\partial\Omega = C$ is a simple closed curve oriented in the counter-clockwise direction.

Proof: We first prove that the difference quotient

$$g(z) = \frac{f(z) - f(a)}{z - a}$$

is an analytic function on all of Ω. The only problematic point is at $z = a$ where the denominator vanishes. First, by the definition of complex derivative,

$$g(a) = \lim_{z \to a} \frac{f(z) - f(a)}{z - a} = f'(a)$$

exists and therefore $g(z)$ is well-defined and, in fact, continuous at $z = a$. Secondly, we can compute its derivative at $z = a$ directly from the definition:

$$g'(a) = \lim_{z \to a} \frac{g(z) - g(a)}{z - a} = \lim_{z \to a} \frac{f(z) - f(a) - f'(a)\,(z - a)}{(z - a)^2} = \tfrac{1}{2}\, f''(a),$$

where we use Taylor's Theorem C.1 (or l'Hôpital's rule) to evaluate the final limit. Since g is differentiable at $z = a$, it is an analytic function on all of Ω. Thus, we may appeal to Cauchy's Theorem 16.48, and conclude that

$$\begin{aligned} 0 = \oint_{\partial\Omega} g(z)\, dz &= \oint_{\partial\Omega} \frac{f(z) - f(a)}{z - a}\, dz = \oint_{\partial\Omega} \frac{f(z)\, dz}{z - a} - f(a) \oint_{\partial\Omega} \frac{dz}{z - a} \\ &= \oint_{\partial\Omega} \frac{f(z)\, dz}{z - a} - 2\,\pi\,\mathrm{i}\, f(a). \end{aligned}$$

The second integral was evaluated using (16.122). Rearranging terms completes the proof of the Cauchy formula. *Q.E.D.*

Remark: The proof shows that if $a \notin \overline{\Omega}$, then the Cauchy integral vanishes:

$$\frac{1}{2\,\pi\,\mathrm{i}} \oint_{\partial\Omega} \frac{f(z)}{z - a}\, dz = 0.$$

Finally, if $a \in \partial\Omega$, then the integral does not converge.

Let us see how we can apply this result to evaluate seemingly intractable complex integrals.

Example 16.59. Suppose that you are asked to evaluate the complex integral

$$\oint_C \frac{e^z\, dz}{z^2 - 2\,z - 3}$$

where C is a circle of radius 2 centered at the origin. A direct evaluation is not possible, since the integrand does not have an elementary antiderivative. However, we note that

$$\frac{e^z}{z^2 - 2\,z - 3} = \frac{e^z}{(z + 1)(z - 3)} = \frac{f(z)}{z + 1} \qquad \text{where} \qquad f(z) = \frac{e^z}{z - 3}$$

is analytic in the disk $|\, z \,| \leq 2$ since its only singularity, at $z = 3$, lies outside the contour C. Therefore, by Cauchy's formula (16.135), we immediately obtain the integral

$$\oint_C \frac{e^z\, dz}{z^2 - 2\,z - 3} = \oint_C \frac{f(z)\, dz}{z + 1} = 2\,\pi\,\mathrm{i}\, f(-1) = -\frac{\pi\,\mathrm{i}}{2\,e}\,.$$

Path independence implies that the integral has the same value on any other simple closed contour, provided it is oriented in the usual counter-clockwise direction, encircles the point $z = 1$ but not the point $z = 3$.

If the contour encloses both singularities, then we cannot apply Cauchy's formula directly. However, as we will see, Theorem 16.58 can be adapted in a direct manner to such situations. This more general result will lead us directly to the calculus of residues, to be discussed shortly.

Derivatives by Integration

The fact that we can recover values of complex functions by integration is surprising. Even more amazing is the fact that we can compute *derivatives* of complex functions by *integration*. Let us differentiate both sides of Cauchy's formula (16.135) with respect to a. The integrand in the Cauchy formula is sufficiently nice so as to allow us to bring the derivative inside the integral sign. Moreover, the derivative of the Cauchy integrand with respect to a is easily found:

$$\frac{\partial}{\partial a}\left(\frac{f(z)}{z - a}\right) = \frac{f(z)}{(z - a)^2}\,.$$

In this manner, we deduce an integral formulae for the *derivative* of an analytic function:

$$f'(a) = \frac{1}{2\,\pi\,\mathrm{i}} \oint_C \frac{f(z)}{(z - a)^2}\, dz, \tag{16.136}$$

where, as before, C is any closed curve that goes once around the point $z = a$ in a counter-clockwise direction. Further differentiation yields the general integral formulae

$$f^{(n)}(a) = \frac{n!}{2\,\pi\,\mathrm{i}} \oint_C \frac{f(z)}{(z - a)^n}\, dz \tag{16.137}$$

that expresses the n^{th} order derivative of a complex function in terms of a contour integral.

These remarkable formulae, which again have no counterpart in real function theory, can be used to prove our earlier claim that an analytic function is infinitely differentiable, and thereby complete the proof of Theorem 16.9.

Example 16.60. Let us compute the integral

$$\oint_C \frac{e^z\, dz}{z^3 - z^2 - 5\,z - 3} = \oint_C \frac{e^z\, dz}{(z + 1)^2 (z - 3)}\,,$$

around the circle of radius 2 centered at the origin. We use (16.136) with

$$f(z) = \frac{e^z}{z - 3}, \qquad\qquad f'(z) = \frac{(z - 4)\, e^z}{(z - 3)^2}\,.$$

Since $f(z)$ is analytic inside C, we conclude that

$$\oint_C \frac{e^z\,dz}{z^3 - z^2 - 5\,z - 3} = \oint_C \frac{f(z)\,dz}{(z+1)^2} = 2\,\pi\,\mathrm{i}\,f'(-1) = -\frac{5\,\pi\,\mathrm{i}}{8\,e}\,.$$

One application is the following remarkable result due to Liouville, whom we already met in Section 11.5. It says that the only bounded complex functions are the constants!

Theorem 16.61. *If $f(z)$ is defined and analytic and $|\,f(z)\,| \le M$ for all $z \in \mathbb{C}$, then $f(z)$ is constant.*

Proof: According to Cauchy's formula (16.135), for any point $a \in \mathbb{C}$,

$$f'(a) = \frac{1}{2\,\pi\,\mathrm{i}} \oint_{C_R} \frac{f(z)\,dz}{(z-a)^2},$$

where we take $C_R = \{\,|\,z - a\,| = R\,\}$ to be a circle of radius R centered at $z = a$. We then estimate the complex integral using (16.123), whence

$$|\,f'(a)\,| = \frac{1}{2\,\pi}\,\left|\,\oint_{C_R} \frac{f(z)\,dz}{(z-a)^2}\,\right| \;\le\; \frac{1}{2\,\pi} \oint_{C_R} \frac{|\,f(z)\,|}{|\,z-a\,|^2}\,ds \;\le\; \frac{1}{2\,\pi} \oint_{C_R} \frac{M}{R^2}\,ds = \frac{M}{R}\,,$$

since the length of C_R is $2\,\pi\,R$. Since $f(z)$ is analytic everywhere, we can let $R \to \infty$ and conclude that $f'(a) = 0$. But this occurs for all possible points a, and $f'(z) \equiv 0$ is everywhere zero, which suffices to prove constancy of $f(z)$. *Q.E.D.*

One immediate application is a complex analysis proof of the Fundamental Theorem of Algebra. Gauss first proved this theorem in 1799, and then gave several further proofs; see [**57**] for an extensive discussion. Although this is, in essence, a purely algebraic result, the proof given here relies in an essential way on complex analysis and complex integration.

Theorem 16.62. *Every nonconstant (complex or real) polynomial $f(z)$ has a root $z_0 \in \mathbb{C}$.*

Proof: Suppose

$$f(z) = a_n\,z^n + a_{n-1}\,z^{n-1} + \,\cdots\, + a_1\,z + a_0 \neq 0$$

for all $z \in \mathbb{C}$. Then we claim that its reciprocal

$$g(z) = \frac{1}{f(z)} = \frac{1}{a_n\,z^n + a_{n-1}\,z^{n-1} + \,\cdots\, + a_1\,z + a_0}$$

satisfies the hypotheses of Liouville's Theorem 16.61, and hence must be constant, in contradiction to our hypothesis. Therefore, $f(z)$ cannot be zero for all z, and this proves the result.

To prove the claim, first by our hypothesis that $f(z) \neq 0$, we immediately conclude that $g(z)$ is analytic for all $z \in \mathbb{C}$. Moreover, $|\,f(z)\,| \to \infty$ as $|\,z\,| \to \infty$; indeed, writing

$$|\,f(z)\,| = |\,z\,|^n\,\left|\,a_n + \frac{a_{n-1}}{z} + \,\cdots\, + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n}\,\right|,$$

the first term clearly foes to ∞ as $|\,z\,| \to \infty$, while the second term is bounded for $|\,z\,| \gg 0$. Therefore,

$$|\,g(z)\,| = \frac{1}{|\,f(z)\,|} \;\longrightarrow\; 0 \qquad \text{as} \qquad |\,z\,| \to \infty,$$

and this is enough (see Exercise ∎) to prove that $|\,g(z)\,| \le M$ is bounded for $z \in \mathbb{C}$. *Q.E.D.*

Corollary 16.63. *Every complex polynomial of degree n can be factored,*

$$f(z) = a_n\,(z - z_1)\,(z - z_2)\;\cdots\;(z - z_n)$$

where $a_1, \dots, a_n$ are the roots of $f(z)$.

Proof: The Fundamental Theorem 16.62 guarantees that there is at least one point $z_1 \in \mathbb{C}$ where $f(z_1) = 0$. Therefore, we can write

$$f(z) = (z - z_1)\,g(z)$$

where $g(z)$ is a polynomial of degree $n - 1$. The proof is completed via a straightforward induction on the degree of the polynomial. *Q.E.D.*

The Calculus of Residues

Cauchy's Theorem and Integral Formulae provide amazingly versatile tools for evaluating complicated complex integrals. Since one only needs to understand the singularities of the integrand within the domain of integration, no indefinite integration is needed to evaluate the integral! With a little more work, we are led to a general method for evaluating contour integrals, known as the *Calculus of Residues* for reasons that will soon be clear. Again, these results and methods have no counterpart in real integration theory. However, the calculus of residues can, even more remarkably, be used to evaluate a large variety of interesting definite real integrals for which no explicit indefinite integral exists.

The key idea is encapsulated in the following definition.

Definition 16.64. Let $f(z)$ be an analytic function for all z near, but not equal to a. The *residue* of $f(z)$ at the point $z = a$ is defined by the complex integral

$$\operatorname*{Res}_{z=a} f(z) = \frac{1}{2\,\pi\,\mathrm{i}} \oint_C f(z)\,dz. \tag{16.138}$$

The contour integral in (16.138) is taken once in a counter-clockwise direction around any simple, closed curve C that contains a in its interior, as illustrated in Figure residue∎. For example, C could be a small circle centered at a. We require that $f(z)$ be analytic everywhere inside C except at the point $z = a$. Lemma 16.50 implies that the value of the residue does not depend on which curve is chosen. The residue is a complex number, and tells us certain information about the singularity of $f(z)$ at $z = a$.

The simplest example is the monomial function $f(z) = c\,z^n$, where c is a complex constant and n is an integer. According to (16.112)

$$\operatorname*{Res}_{z=0} c\,z^n = \frac{1}{2\,\pi\,\mathrm{i}} \oint_C c\,z^n\,dz = \begin{cases} 0, & n \neq -1, \\ c, & n = -1. \end{cases} \tag{16.139}$$

Thus, only the exponent $n = -1$ gives a nonzero residue. The residue singles out the function $1/z$, which, not coincidentally, is the only one with a logarithmic, and multiply-valued, antiderivative.

Cauchy's Theorem 16.48, when applied to the integral in (16.138), implies that if $f(z)$ is analytic at $z = a$, then it has zero residue at a. Therefore, all the monomials, including $1/z$, have zero residue at any nonzero point:

$$\operatorname*{Res}_{z=a} c\,z^n = 0 \qquad \text{for} \qquad a \neq 0. \tag{16.140}$$

Since integration is a linear operation, the residue is a linear operator, mapping complex functions to complex numbers; thus,

$$\operatorname*{Res}_{z=a} \big[\, f(z) + g(z) \,\big] = \operatorname*{Res}_{z=a} f(z) + \operatorname*{Res}_{z=a} g(z), \qquad \operatorname*{Res}_{z=a} \big[\, c\, f(z) \,\big] = c \operatorname*{Res}_{z=a} f(z), \tag{16.141}$$

for any complex constant c. Therefore, by linearity, the residue of any finite linear combination

$$f(z) = \frac{c_{-m}}{z^m} + \frac{c_{-m+1}}{z^{m-1}} + \cdots + \frac{c_{-1}}{z} + c_0 + c_1\, z + \cdots + c_n\, z^n = \sum_{k=-m}^{n} c_k\, z^k$$

of such monomials is equal to

$$\operatorname*{Res}_{z=0} f(z) = c_{-1}.$$

Thus, the residue effectively picks out the coefficient of the term $1/z$ in such an expansion. As we shall shortly see, the same holds true for infinite series of a similar form.

The easiest nontrivial residues to compute are at poles of a function. According to (16.29), the function $f(z)$ has a *simple pole* at $z = a$ if

$$h(z) = (z - a)\, f(z) \tag{16.142}$$

is analytic at $z = a$ and $h(a) \neq 0$.

Lemma 16.65. *If $f(z) = \dfrac{h(z)}{z - a}$ has a simple pole at $z = a$, then $\operatorname*{Res}_{z=a} f(z) = h(a)$.*

Proof: We substitute the formula for $f(z)$ into the definition (16.138), and so

$$\operatorname*{Res}_{z=a} f(z) = \frac{1}{2\,\pi\,\mathrm{i}} \oint_C f(z)\, dz = \frac{1}{2\,\pi\,\mathrm{i}} \oint_C \frac{h(z)\, dz}{z - a} = h(a),$$

by Cauchy's formula (16.135). *Q.E.D.*

Example 16.66. Consider the function

$$f(z) = \frac{e^z}{z^2 - 2\,z - 3} = \frac{e^z}{(z+1)(z-3)}\,.$$

From the factorization of the denominator, we see that $f(z)$ has simple pole singularities at $z = -1$ and $z = 3$. The residues are given, respectively, by

$$\operatorname*{Res}_{z=-1} \frac{e^z}{z^2 - 2\,z - 3} = \left.\frac{e^z}{z - 3}\right|_{z=-1} = -\,\frac{1}{4\,e}\,, \qquad \operatorname*{Res}_{z=3} \frac{e^z}{z^2 - 2\,z - 3} = \left.\frac{e^z}{z + 1}\right|_{z=3} = \frac{e^3}{4}\,.$$

Since $f(z)$ is analytic everywhere else, the residue at any other point is automatically 0.

Recall that a function $g(z)$ is said to have *simple zero* at $z = a$ provided

$$g(z) = (z - a)\, k(z)$$

where $k(z)$ is analytic at $z = a$ and $k(a) = g'(a) \neq 0$. In this case, the reciprocal function

$$f(z) = \frac{1}{g(z)} = \frac{1}{(z - a)\, k(z)}$$

has a simple pole at $z = a$. The residue of the reciprocal is, by Lemma 16.65,

$$\operatorname*{Res}_{z=a} f(z) = \operatorname*{Res}_{z=a} \frac{1}{(z - a)\, k(z)} = \frac{1}{k(a)} = \frac{1}{g'(a)}\,. \tag{16.143}$$

More generally, if $f(z)$ is analytic at the point a, then the ratio $f(z)/g(z)$ has residue

$$\operatorname*{Res}_{z=a} \frac{f(z)}{g(z)} = \frac{f(a)}{g'(a)} \tag{16.144}$$

at a simple zero $z = a$ of $g(z)$.

Example 16.67. As an illustration, let us compute the residue of $\sec z = 1/\cos z$ at the point $z = \frac{1}{2}\pi$. Note that $\cos z$ has a simple zero at $z = \frac{1}{2}\pi$ since its derivative, $-\sin z$, is nonzero there. Thus, according to (16.143),

$$\operatorname*{Res}_{z=\pi/2} \sec z = \operatorname*{Res}_{z=\pi/2} \frac{1}{\cos z} = \frac{-1}{\sin \frac{1}{2}\pi} = -1.$$

The direct computation of the residue using a complex integral (16.138) is slightly harder, but instructive. For example, we may integrate $\sec z$ around a circle of radius 1 centered at $\frac{1}{2}\pi$, which we parametrize by $z(t) = \frac{1}{2}\pi + e^{\,\mathrm{i}\,t}$. According to the definition,

$$\operatorname*{Res}_{z=a} \sec z = \frac{1}{2\,\pi\,\mathrm{i}} \oint_C \frac{dz}{\cos z} = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} \frac{-2e^{\,\mathrm{i}\,t}\, dt}{e^{\,\mathrm{i}\,t} + e^{-\,\mathrm{i}\,t}}\, dt = -\,\frac{1}{\pi} \int_{-\pi}^{\pi} \frac{dt}{1 + e^{-2\,\mathrm{i}\,t}}\,.$$

We multiply the numerator and denominator in the latter integrand by $1 + e^{2\,\mathrm{i}\,t}$, and use Euler's formula (3.76) to obtain

$$\operatorname*{Res}_{z=a} \sec z = -\,\frac{1}{\pi} \int_{-\pi}^{\pi} \left[\, 1 + \,\mathrm{i}\, \frac{\sin 2\,t}{1 + \cos 2\,t} \,\right] dt = -1\,.$$

Note that the imaginary part of this integral vanishes because it is the integral of an odd function over a symmetric interval, cf. Lemma 12.11.

The Residue Theorem

Residues are the ingredients in a general method for computing contour integrals of analytic functions. The *Residue Theorem* says that the value of the integral of a complex function around a closed curve depends only on its residues at the enclosed singularities. Since the residues can be computed directly from the function, the resulting formula provides an effective mechanism for painless evaluation of complex integrals, without having to construct any sort of anti-derivative or indefinite integral. Indeed, the residue method can be employed even when the integrand does not have an anti-derivative that can be expressed in terms of elementary functions.

Theorem 16.68. *Let C be a simple, closed curve, oriented in the counter-clockwise direction. Suppose $f(z)$ is analytic everywhere inside C except for a finite number of singularities, $a_1, \dots, a_n$. Then*

$$\frac{1}{2\pi\,\mathrm{i}} \oint_C f(z)\,dz = \operatorname*{Res}_{z=a_1} f(z) + \cdots + \operatorname*{Res}_{z=a_n} f(z). \tag{16.145}$$

Proof: We draw a small circle C_i around each singularity a_i. We assume the circles all lie inside the contour C and do not cross each other, so that a_i is the only singularity contained within C_i; see Figure resC■. Definition 16.64 implies that

$$\operatorname*{Res}_{z=a_i} f(z) = \frac{1}{2\pi\,\mathrm{i}} \oint_{C_i} f(z)\,dz, \tag{16.146}$$

where the line integral is taken in the counter-clockwise direction around C_i.

Consider the domain Ω consisting of all points z which lie inside the given curve C, but outside all the small circles $C_1, \dots, C_n$; this is the shaded region in Figure resC■. By our construction, the function $f(z)$ is analytic on Ω, and hence by Cauchy's Theorem 16.48, the integral of f around the boundary $\partial\Omega$ is zero. The boundary $\partial\Omega$ must be oriented consistently, so that the domain is always lying on one's left hand side. This means that the outside contour C should be traversed in a counter-clockwise direction, whereas the inside circles C_i are in a *clockwise* direction. Therefore, the integral around the boundary of the domain Ω can be broken up into a difference

$$\begin{aligned} 0 = \frac{1}{2\pi\,\mathrm{i}} \oint_{\partial\Omega} f(z)\,dz &= \frac{1}{2\pi\,\mathrm{i}} \oint_C f(z) \;-\; \sum_{i=1}^{n} \frac{1}{2\pi\,\mathrm{i}}\, dz \oint_{C_i} f(z)\,dz \\ &= \frac{1}{2\pi\,\mathrm{i}} \oint_C f(z) \;-\; \sum_{i=1}^{n} \operatorname*{Res}_{z=a_i} f(z)\,dz. \end{aligned}$$

The minus sign converts the circle integrals to the counterclockwise orientation used in the definition (16.146) of the residues. Rearranging the final identity leads to the residue formula (16.145). *Q.E.D.*

Example 16.69. Let us use residues to evaluate the contour integral

$$\oint_C \frac{e^z}{z^2 - 2z - 3}\,dz$$

where C denotes a circle of radius r centered at the origin. According to Example 16.66, the integrand has two singularities at -1 and 3, with respective residues $-1/(4e)$ and $e^3/4$. If the radius of the circle is $r > 3$, then it goes around both singularities, and hence by the residue formula (16.145)

$$\oint_C \frac{e^z\,dz}{z^2 - 2z - 3} = 2\pi\,\mathrm{i}\,\left(-\frac{1}{4e} + \frac{e^3}{4}\right) = \frac{(e^4 - 1)\,\pi\,\mathrm{i}}{2e}\,.$$

If the circle has radius $1 < r < 3$, then it only encircles the singularity at -1, and hence

$$\oint_C \frac{e^z}{z^2 - 2z - 3}\,dz = -\frac{\pi\,\mathrm{i}}{2e}\,.$$

If $0 < r < 1$, the function has no singularities inside the circle and hence, by Cauchy's Theorem 16.48, the integral is 0. Finally, when $r = 1$ or $r = 3$, the contour passes through a singularity, and the integral does not converge.

Evaluation of Real Integrals

One important and unexpected application of the Residue Theorem 16.68 is to aid in the evaluation of certain definite real integrals. The interesting fact is that it even applies to cases in which one is unable to evaluate the corresponding indefinite integral in closed form, owing to the non-existence of an elementary anti-derivative. Nevertheless, converting the definite real integral into (part of a) complex contour integral leads to a direct evaluation via the calculus of residues that sidesteps the difficulties in finding the antiderivative. This device is indicative of a useful procedure for analyzing standard (meaning analytic) real functions by passing to their complex counterparts, which can then be tackled by the more powerful tools of complex analysis.

There are two principal types of real integral for which this technique can be applied, although numerous variations appear in more extensive treatments of the subject. First, a real trigonometric integral of the form

$$I = \int_0^{2\pi} F(\cos\theta, \sin\theta)\, d\theta \tag{16.147}$$

can often be evaluated by converting it into a complex integral around the unit circle $C = \big\{\, |\, z\,| = 1 \,\big\}$. If we set

$$z = e^{\,\mathrm{i}\,\theta} \qquad \text{so} \qquad \frac{1}{z} = e^{-\,\mathrm{i}\,\theta},$$

then

$$\cos\theta = \frac{e^{\,\mathrm{i}\,\theta} + e^{-\,\mathrm{i}\,\theta}}{2} = \frac{1}{2}\left(z + \frac{1}{z}\right), \qquad \sin\theta = \frac{e^{\,\mathrm{i}\,\theta} - e^{-\,\mathrm{i}\,\theta}}{2\,\mathrm{i}} = \frac{1}{2\,\mathrm{i}}\left(z - \frac{1}{z}\right). \tag{16.148}$$

Moreover,

$$dz = de^{\,\mathrm{i}\,\theta} = \mathrm{i}\, e^{\,\mathrm{i}\,\theta}\, d\theta = \mathrm{i}\, z\, d\theta, \qquad \text{and so} \qquad d\theta = \frac{dz}{\mathrm{i}\, z}\,. \tag{16.149}$$

Therefore, the integral (16.147) can be written in the complex form

$$I = \oint_C F\left(\frac{1}{2}\left(z + \frac{1}{z}\right), \frac{1}{2\,\mathrm{i}}\left(z - \frac{1}{z}\right)\right)\frac{dz}{\mathrm{i}\, z}\,. \tag{16.150}$$

If we know that the resulting complex integrand is well-defined and single-valued, except, possibly, for a finite number of singularities inside the unit circle, then the residue formula (16.145) tells us that the integral can be directly evaluated by adding together its residues and multiplying by $2\,\pi\,\mathrm{i}$.

Example 16.70. We compute the simple example $\displaystyle\int_0^{2\pi} \frac{d\theta}{2+\cos\theta}$. We begin by using the substitution (16.150), whence

$$\int_0^{2\pi} \frac{d\theta}{2+\cos\theta} = \oint_C \frac{dz}{\mathrm{i}\, z\left[\,2+\frac{1}{2}\left(z+\frac{1}{z}\right)\right]} = -\,\mathrm{i}\oint_C \frac{2\,dz}{z^2+4\,z+1}\,.$$

The complex integrand has singularities where its denominator vanishes:

$$z^2+4\,z+1=0, \qquad \text{so that} \qquad z=-2\pm\sqrt{3}\,.$$

Only one of these singularities, namely $-2+\sqrt{3}$ lies inside the unit circle. Therefore, applying (16.144), we find

$$-\,\mathrm{i}\oint_C \frac{2\,dz}{z^2+4\,z+1} = 2\pi \operatorname*{Res}_{z=-2+\sqrt{3}} \frac{2}{z^2+4\,z+1} = \left.\frac{4\pi}{2\,z+4}\right|_{z=-2+\sqrt{3}} = \frac{4\pi}{\sqrt{3}}\,.$$

As the student may recall from first year calculus, this particular integral can, in fact, be done directly via a trigonometric substitution. However, the computations are not particularly pleasant, and, with practice, the residue method is much simpler. Moreover, it straightforwardly applies to situations where no elementary antiderivative exists.

Example 16.71. ∎

A second type of real integral that can often be evaluated by complex residues are integrals over the entire real line, from $-\infty$ to ∞. Here the technique is a little more subtle, and we sneak up on the integral by using larger and larger closed contours that coincide with more and more of the real axis. The basic idea is contained in the following example.

Example 16.72. The problem is to evaluate the real integral

$$I = \int_0^{\infty} \frac{\cos x\; dx}{1+x^2}\,. \tag{16.151}$$

The corresponding indefinite integral cannot be evaluated in elementary terms, and so we are forced to rely on the calculus of residues. We begin by noting that the integrand is even, and hence the integral $I=\frac{1}{2}\,J$ is one half the integral

$$J = \int_{-\infty}^{\infty} \frac{\cos x\; dx}{1+x^2}$$

over the entire real line. Moreover, for x real, we can write

$$\frac{\cos x}{1+x^2} = \operatorname{Re}\frac{e^{\,\mathrm{i}\,x}}{1+x^2}, \qquad \text{and hence} \qquad J = \operatorname{Re}\int_{-\infty}^{\infty} \frac{e^{\,\mathrm{i}\,x}\,dx}{1+x^2}. \tag{16.152}$$

Let C_R be the closed contour consisting of a large semicircle of radius $R\gg 0$, which we denote by S_R, connected at the ends by the real interval $-R\le x\le R$, which is plotted

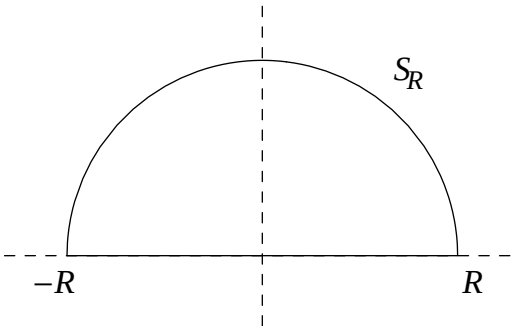


Figure 16.23. Semicircular Contour.

in Figure 16.23. The corresponding contour integral

$$\oint_{C_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2} \;=\; \int_{-R}^{R} \frac{e^{\mathrm{i}\,x}\,dx}{1+x^2} \;+\; \int_{S_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2} \tag{16.153}$$

breaks up into two pieces: the first over the real interval, and the second over the semicircle. As the radius $R \to \infty$, the semicircular contour C_R includes more and more of the real axis, and so the first integral gets closer and closer to our desired integral (16.152). If we can prove that the second, semicircular integral goes to zero, then we will be able to evaluate the integral over the real axis by contour integration, and hence by the method of residues. The fact that the semicircular integral is small is reasonable, since the integrand $(1+z^2)^{-1}e^{\mathrm{i}\,z}$ gets smaller and smaller as $|\,z\,| \to \infty$ *provided* $\operatorname{Re} z \geq 0$. A rigorous verification of this fact will appear at the end of the example.

According to the Residue Theorem 16.68, the integral (16.153) is equal to the sum of all the residues over the singularities of $f(z)$ lying inside the contour C_R. Now e^z is analytic everywhere, and so the singularities occur where the denominator vanishes, i.e., $z^2 = -1$, and so are at $z = \pm\,\mathrm{i}$. Since the semicircle lies in the upper half plane $\operatorname{Im} z > 0$, only the first singularity $z = +\,\mathrm{i}$ lies inside — provided the radius $R > 1$. To compute the residue, we use (16.143) to evaluate

$$\operatorname*{Res}_{z=\,\mathrm{i}} \frac{e^{\mathrm{i}\,z}}{1+z^2} = \left.\frac{e^{\mathrm{i}\,z}}{2\,z}\right|_{z=\,\mathrm{i}} = \frac{e^{-1}}{2\,\mathrm{i}} = \frac{1}{2\,\mathrm{i}\,e}\,.$$

Therefore, by (16.145),

$$\frac{1}{2\,\pi\,\mathrm{i}} \oint_{C_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2} = \frac{1}{2\,\mathrm{i}\,e} = \frac{\pi}{e}\,,$$

whenever $R > 1$. Thus, assuming the semicircular part of the integral does indeed become vanishingly small as $R \to \infty$, we conclude that

$$\int_{-\infty}^{\infty} \frac{e^{\mathrm{i}\,x}\,dx}{1+x^2} = \lim_{R\to\infty} \oint_{C_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2} = 2\,\pi\,\mathrm{i}\;\frac{1}{2\,\mathrm{i}\,e} = \frac{\pi}{e}\,.$$

The integral is real because its imaginary part,

$$\int_{-\infty}^{\infty} \frac{\sin x\,dx}{1+x^2} = 0,$$

is the integral of an odd function which is automatically zero. Consequently,

$$I = \int_0^\infty \frac{\cos x \, dx}{1+x^2} = \frac{1}{2} \operatorname{Re} \int_{-\infty}^\infty \frac{e^{\mathrm{i}\, x} \, dx}{1+x^2} = \frac{\pi}{2\, e} \,,$$

which is the desired result.

Finally, let us estimate the size of the semicircular integral. The integrand is bounded by

$$\left| \frac{e^{\mathrm{i}\, z}}{1+z^2} \right| \le \frac{1}{1+|\, z\,|^2} = \frac{1}{1+R^2} \qquad \text{whenever} \qquad |\, z\,| = R, \qquad \operatorname{Im} z \ge 0,$$

where we are using the fact that

$$\left| e^{\mathrm{i}\, z} \right| = e^{-y} \le 1 \qquad \text{whenever} \qquad z = x + \mathrm{i}\, y \qquad \text{with} \qquad y \ge 0.$$

According to Corollary 16.56, the size of the integral of a complex function is bounded by its maximum modulus along the curve times the length of the curve. Thus, in our case,

$$\left| \int_{S_R} \frac{e^{\mathrm{i}\, z} \, dz}{1+z^2} \right| \le \frac{1}{1+R^2} \, \mathcal{L}(S_R) = \frac{\pi\, R}{1+R^2} \le \frac{\pi}{R} \,.$$

Thus, the semicircular integral becomes vanishingly small as the radius of our semicircle goes to infinity, $R \to \infty$. This completes the justification of the method.

Example 16.73. Here the problem is to evaluate the integral

$$\int_{-\infty}^\infty \frac{dx}{1+x^4} \,. \tag{16.154}$$

The indefinite integral can, in fact, be done by partial fractions, but, as anyone who has tried this can tell you, this is not a particularly pleasant task. So, let us try using residues. Let C_R denote the same semicircular contour as in the previous example. The integrand has pole singularities where the denominator vanishes, i.e., $z^4 = -1$, and so at the four fourth roots of -1. These are

$$e^{\pi\, \mathrm{i}\, /4} = \frac{1+\mathrm{i}}{\sqrt{2}} \,, \,, \qquad e^{3\, \pi\, \mathrm{i}\, /4} = \frac{-1+\mathrm{i}}{\sqrt{2}} \,, \,, \qquad e^{5\, \pi\, \mathrm{i}\, /4} = \frac{1-\mathrm{i}}{\sqrt{2}} \,, \,, \qquad e^{7\, \pi\, \mathrm{i}\, /4} = \frac{-1-\mathrm{i}}{\sqrt{2}} \,.$$

Only the first two roots lie inside the semicircular contour C_R, provided $R > 1$. Their residues can be computed using (16.143):

$$\operatorname*{Res}_{z = e^{\pi\, \mathrm{i}\, /4}} \frac{1}{1+z^4} = \frac{1}{4\, z^3} \bigg|_{z = e^{\pi\, \mathrm{i}\, /4}} = \frac{e^{-3\, \pi\, \mathrm{i}\, /4}}{4} = \frac{-1-\mathrm{i}}{4\sqrt{2}} \,,$$

$$\operatorname*{Res}_{z = e^{3\, \pi\, \mathrm{i}\, /4}} \frac{1}{1+z^4} = \frac{1}{4\, z^3} \bigg|_{z = e^{3\, \pi\, \mathrm{i}\, /4}} = \frac{e^{-9\, \pi\, \mathrm{i}\, /4}}{4} = \frac{1-\mathrm{i}}{4\sqrt{2}} \,.$$

Therefore, by the residue formula (16.145),

$$\oint_{C_R} \frac{dz}{1+z^4} = 2\, \pi\, \mathrm{i} \, \left(\frac{-1-\mathrm{i}}{4\sqrt{2}} + \frac{1-\mathrm{i}}{4\sqrt{2}} \right) = \frac{\pi}{\sqrt{2}} \,. \tag{16.155}$$

On the other hand, we can break up the complex integral into an integral along the real axis and an integral around the semicircle:

$$\oint_{C_R} \frac{dz}{1+z^4} = \int_{-R}^{R} \frac{dx}{1+x^4} + \int_{S_R} \frac{dz}{1+z^4}.$$

The first integral goes to the desired real integral as the radius $R \to \infty$. On the other hand, on a large semicircle $|\, z \,| = R$, the integrand $1/(1+z^4)$ is small:

$$\left| \frac{1}{1+z^4} \right| \le \frac{1}{1+|\, z \,|^4} = \frac{1}{1+R^4} \qquad \text{whenever} \qquad |\, z \,| = R.$$

Thus, using Corollary 16.56, the integral around the semicircle can be bounded by

$$\left| \int_{S_R} \frac{dz}{1+z^4} \right| \le \frac{1}{1+R^4} \cdot \pi R \le \frac{\pi}{R^3} \longrightarrow 0 \qquad \text{as} \qquad R \longrightarrow \infty.$$

Thus, as $R \to \infty$, the complex integral (16.155) goes to the desired real integral (16.154), and so

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^4} = \frac{\pi}{\sqrt{2}}\,.$$

Note that the result is real and positive, as it must be.

Chapter 17

Dynamics of Planar Media

In this chapter, we continue our ascent of the dimensional ladder for linear systems. In Chapter 6, we began our journey by analyzing the equilibrium configurations of discrete systems — mass–spring chains, circuits and structures — which are governed by certain linear algebraic systems. Next, in Chapter 9, we introduced a continuous time variable to model the dynamical behavior of such discrete systems by associated systems of linear ordinary differential equations. Chapter 11 began our treatment of continuous media with the boundary value problems that describe the equilibrium configurations of one-dimensional bars, strings and beams. Dynamical motions of one-dimensional media formed the focus of Chapter 14, leading to two fundamental partial differential equations: the heat equation describing thermal diffusion, and the wave equation modeling vibrations. In Chapters 15 and 16, we focussed our attention on the boundary value problems describing equilibrium of planar bodies — plates and membranes — with primary emphasis on the all-important Laplace equation. We now turn to the analysis of the dynamical behavior of planar bodies, as governed by the two-dimensional[†] versions of the heat and wave equations. The heat equation describes diffusion of, say, heat energy or population in a homogeneous two-dimensional domain. The wave equation models small vibrations of two-dimensional membranes, e.g., drums.

Although the increase in dimension does exact a toll on our analytical prowess, we have, in fact, already mastered many of the key techniques. When applied to partial differential equation in higher dimensions, the separation of variables method often results in ordinary differential equations of a non-elementary type. Solutions are expressed in terms of certain remarkable and important non-elementary functions, including the Bessel functions in the present chapter, and the Legendre functions, spherical harmonics, and spherical Bessel functions appearing in three-dimensional problems. These so-called *special functions* do not appear in elementary calculus, but do play a starring role in more advanced applications in physics, engineering and mathematics. Most interesting special functions arise as solutions to certain second order, self-adjoint boundary value problems of Sturm–Liouville type. As such, they obey basic orthogonality relations, and thus can be used in place of the trigonometric sines and cosines that form the foundations of elementary Fourier analysis. Thus, the series solutions of higher dimensional partial differential equations lead naturally to the study of special function series. In Appendix C, we collect together the required results about the most important classes of special functions,

[†] We use the term "dimension" to refer to the number of independent space variables in the system. Time is accorded a special status, and serves to distinguish dynamics from equilibrium.

including a short presentation of the series approach for solving non-elementary ordinary differential equations.

We will also derive a multi-dimensional version of the fundamental solution, corresponding to an initial concentrated delta function force. This allows one to use a general superposition principle to solve the initial value problem. Disappointingly, conformal mappings are not particularly helpful in the dynamical universe.

Numerical methods for solving boundary value and initial value problems are, of course, essential in all but the simplest situations. The two basic methods — finite element and finite difference — have already appeared, and the only new aspect is the (substantial) complication of working in higher dimensions. Thus, in the interests of brevity, we defer the discussion of the numerical aspects of multi-dimensional partial differential equations to more advanced texts, e.g., [**nPDE**], and student projects outlined in the exercises. However, the student should be assured that, without knowledge of the qualitative features based on direct analysis and particular solutions, the design, implementation, and testing of numerical solution techniques would be severely hampered. Explicit solutions continue to play an important practical role, both as a guide for constructing numerical algorithms, as well as a convenient test of their accuracy.

17.1. Diffusion in Planar Media.

The heating of a homogeneous flat plate is modeled by the two-dimensional heat equation

$$u_t = \gamma\, \Delta u = \gamma \left(u_{xx} + u_{yy} \right), \tag{17.1}$$

where $\Delta = \partial_x^2 + \partial_y^2$ is the two-dimensional Laplacian operator. The solution $u(t, \mathbf{x}) = u(t, x, y)$ to (17.1) measures the temperature at time t at each point $\mathbf{x} = (x, y) \in \Omega$ in the domain $\Omega \subset \mathbb{R}^2$ occupied by the plate. We are assuming that there are no external heat sources on the interior of our plate, which can be arranged by covering its top and bottom with insulation. In particular, an equilibrium solution $u = u(x, y)$ does not depend on time t, so $u_t = 0$, and hence must satisfy the Laplace equation $\Delta u = 0$, which is in accordance with Chapter 15.

As in the one-dimensional version, the diffusivity coefficient $\gamma > 0$, which measures the relative speed of diffusion of heat energy throughout the medium, must be positive. Negative diffusivity results an ill-posed initial value problem, which experiences the same difficulties as its we saw in the one-dimensional backwards heat equation. The physical justification of the heat equation model will be discussed in detail shortly.

To uniquely specify the temperature $u(t, x, y)$, we must impose both initial and boundary conditions. As with the equilibrium Laplace equation, the most important are

(a) *Dirichlet boundary conditions*

$$u = h \qquad \text{on} \qquad \partial\Omega, \tag{17.2}$$

where the temperature is fixed on the boundary of the plate.

(b) *Neumann boundary conditions* that prescribe the heat flux or normal derivative

$$\frac{\partial u}{\partial n} = k \qquad \text{on} \qquad \partial\Omega, \tag{17.3}$$

with $k = 0$ corresponding to an insulated boundary.

(c) *Mixed boundary conditions*, where we impose Dirichlet conditions on part of the boundary $D \subsetneq \partial\Omega$ and Neumann conditions on the remainder $N = \partial\Omega \setminus D$. For instance, the homogeneous mixed boundary conditions

$$u = 0 \qquad \text{on} \qquad D, \qquad\qquad \frac{\partial u}{\partial n} = 0 \qquad \text{on} \qquad N, \tag{17.4}$$

correspond to insulating part of the boundary and freezing the remainder.

In all cases, the boundary data may depend upon time as well as the specific boundary point. We further specify the initial temperature of the plate

$$u(0, x, y) = f(x, y), \qquad\qquad (x, y) \in \Omega, \tag{17.5}$$

at an initial time, which for simplicity, we take as $t_0 = 0$. If the domain Ω is bounded with a boundary that is not too wild (e.g., piecewise smooth), a general theorem, [**39**], guarantees the existence of a unique solution $u(t, x, y)$ to any of these self-adjoint initial-boundary value problems for all subsequent times $t \geq 0$. Our practical goal is to both compute and understand the behavior of this solution in specific situations.

Derivation of the Diffusion Equation

The physical derivation of the two-dimensional (and three-dimensional) heat equation relies upon the same two basic thermodynamical laws that were used, in Section 14.1, to derive its one-dimensional version. The first principle is that heat energy tries to flow from hot to cold in as fast a way as possible. According to Theorem 19.39, the negative gradient $-\nabla u$ points in the direction of the steepest decrease in the temperature u at a point, and so, in an isotropic medium, heat energy will flow in that direction. Therefore, the heat flux vector $\mathbf{w}$, which measures the magnitude and direction of the flow of heat energy, should be proportional to the temperature gradient:

$$\mathbf{w}(t, x, y) = -\kappa(x, y)\, \nabla u. \tag{17.6}$$

The scalar quantity $\kappa(x, y) > 0$ measures the *thermal conductivity* of the material[†] at position $(x, y) \in \Omega$. Equation (17.6) is the multi-dimensional counterpart of *Fourier's Law of Cooling*, cf. (14.3).

The second principle is that, in the absence of external heat sources, heat can only enter a region $D \subset \Omega$ through its boundary ∂D. (Recall that the plate is insulated above and below.) Let $\varepsilon(t, x, y)$ denote the heat energy at each time and point in the domain, so that $\iint_D \varepsilon(t, x, y)\, dx\, dy$ is the heat energy contained within the region D at time t. The rate of change of heat energy is equal to the heat flux into the region through its boundary,

[†] We are assuming the material properties of the plate are not changing in time, and, moreover, are not temperature dependent. Changing the latter assumption would lead to a nonlinear diffusion equation.

which is the negative of the flux line integral (A.42), namely $\oint_{\partial D} \mathbf{w} \cdot \mathbf{n}\, ds$, where, as usual, $\mathbf{n}$ denotes the *outward* unit normal to the boundary ∂D. Therefore,

$$\frac{\partial}{\partial t} \iint_D \varepsilon(t,x,y)\, dx\, dy = - \oint_{\partial D} \mathbf{w} \cdot \mathbf{n}\; ds = - \iint_D \nabla \cdot \mathbf{w}\; dx\, dy,$$

where we apply the divergence form (A.57) of Green's Theorem to convert the flux into a double integral. We bring the time derivative inside the integral and collect the terms, whence

$$\iint_D \left(\frac{\partial \varepsilon}{\partial t} + \nabla \cdot \mathbf{w} \right) dx\, dy = 0. \tag{17.7}$$

Keep in mind that this integral formula must hold for *any* subdomain $D \subset \Omega$. Now, the only way in which an integral of a continuous function can vanish for all subdomains is if the integrand is identically zero, cf. Exercise . The net result is the basic "conservation law"

$$\frac{\partial \varepsilon}{\partial t} + \nabla \cdot \mathbf{w} = 0 \tag{17.8}$$

relating heat energy ε and heat flux $\mathbf{w}$.

As in equation (14.2), the heat energy $\varepsilon(t,x,y)$ at each time and point in the domain is proportional to the temperature,

$$\varepsilon(t,x,y) = \sigma(x,y)\, u(t,x,y), \qquad \text{where} \qquad \sigma(x,y) = \rho(x,y)\, \chi(x,y) \tag{17.9}$$

is the product of the *density* and the *heat capacity* of the material. Combining this with the Fourier Law (17.6) and the energy balance equation (17.9) leads to the general two-dimensional *diffusion equation*

$$\sigma\, \frac{\partial u}{\partial t} = \nabla \cdot \big(\kappa\, \nabla u \big). \tag{17.10}$$

In full detail, this second order partial differential equation takes the form

$$\sigma(x,y)\, \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(\kappa(x,y)\, \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(\kappa(x,y)\, \frac{\partial u}{\partial y} \right). \tag{17.11}$$

In particular, if the body is homogeneous, then both σ and κ are constant, and so general diffusion equation (17.10) reduces to the heat equation (17.1) with thermal diffusivity

$$\gamma = \frac{\kappa}{\sigma} = \frac{\kappa}{\rho\, \chi}. \tag{17.12}$$

The heat and diffusion equations are examples of *parabolic* partial differential equations, the terminology being an adaptation of that in Definition 15.1 to partial differential equations in more than two variables.

Self-Adjoint Formulation

The general diffusion equation (17.10) is in the self-adjoint form

$$u_t = - K[u] = - \nabla^* \circ \nabla u. \tag{17.13}$$

The gradient operator ∇ maps scalar fields u to vector fields $\mathbf{v} = \nabla u$. Its adjoint ∇^*, which goes in the reverse direction, is taken with respect to the weighted inner products

$$\langle\, u\,;\widetilde{u}\,\rangle = \iint_\Omega u(x,y)\,\widetilde{u}(x,y)\,\sigma(x,y)\,dx\,dy, \quad \langle\!\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle\!\rangle = \iint_\Omega \mathbf{v}(x,y)\cdot\widetilde{\mathbf{v}}(x,y)\,\kappa(x,y)\,dx\,dy, \tag{17.14}$$

between, respectively, scalar and vector fields. A straightforward integration by parts argument similar to that in Section 15.4 tells us that

$$\nabla^*\mathbf{v} = -\frac{1}{\sigma}\,\nabla\cdot(\kappa\,\mathbf{v}) = -\frac{1}{\sigma}\left(\frac{\partial(\kappa\,v_1)}{\partial x} + \frac{\partial(\kappa\,v_2)}{\partial y}\right). \tag{17.15}$$

Therefore, the right hand side of (17.13) is equal to

$$-K[u] = -\nabla^*\circ\nabla u = \frac{1}{\sigma}\,\nabla\cdot(\kappa\,\nabla u),$$

which recovers the preceding formula (17.10). As always, we need to impose suitable homogeneous boundary conditions — Dirichlet, Neumann or mixed — to ensure the validity of the integration by parts argument used to establish the adjoint formula (17.15).

In particular, to obtain the heat equation, we take σ and κ to be constant, and so (17.14) reduce, up to a constant factor, to the usual L^2 inner products between scalar and vector fields. In this case, the adjoint of the gradient is, up to a scale factor, minus the divergence: $\nabla^* = -\,\gamma\,\nabla\cdot$, where $\gamma = \kappa/\sigma$, and the general diffusion equation (17.13) reduces to the two-dimensional heat equation (17.1).

As we learned in Chapters 9 and 14, a diffusion equation (17.13) has the form of a gradient flow, decreasing the heat energy

$$E[u] = \|\,\nabla u\,\|^2 = \iint_\Omega \|\,\nabla u(x,y)\,\|^2\,\sigma(x,y)\,dx\,dy \tag{17.16}$$

as rapidly as possible. Thus, we expect its solutions to decay rapidly to thermal equilibrium, $u \to u_\star$, defined as a minimum of the energy functional.

The asymptotic equilibrium configuration will be a teim-independent solution to the heat equation, and hence satisfy the second order partial differential equation $\nabla\cdot\big(\,\kappa\,\nabla u\,\big) = 0$, or, in full detail,

$$\frac{\partial}{\partial x}\left(\,\kappa(x,y)\,\frac{\partial u_\star}{\partial x}\,\right) + \frac{\partial}{\partial y}\left(\,\kappa(x,y)\,\frac{\partial u_\star}{\partial y}\,\right) = 0, \tag{17.17}$$

subject to the prescribed boundary conditions. Note that (17.17) is a special case of the general elliptic partial differential equation (15.129) considered at the end opf Chapter 15. In particular, in the homogeneous case when the diffusivity κ is constant, the equation (17.17) reduces to Laplace's equation $\Delta u_\star = 0$, which governs the thermal equilibrium solutions for the heat equation. If the boundary conditions are homogeneous Dirchlet or mixed, then positive definiteness tells us that there is a unique equilibrium solution, namely $u_\star(x,y) = 0$, whereas in the Neumann or fully insulated case, the equilibrium will be a constant temperature distribution, the constant being the average of the initial heat energy; see Exercise ■ for details.

Remark: The heat and diffusion equations are also used to model diffusion of populations, e.g., bacteria in a petri dish or wolves in the Canadian Rockies, [**biol**]. The solution $u(t, x, y)$ represents the number of individuals near position (x, y) at time t. The diffusion is caused by random motions of the individuals. Such diffusion processes also appear in mixing of chemical reagents in solutions, with reactions introducing additional nonlinear terms that result in the broad class of *reaction–diffusion equations*, [**chem**].

17.2. Solution Techniques for Diffusion Equations.

We now discuss basic analytical (as opposed to numerical) solution techniques for the two-dimensional heat and diffusion equations of the form

$$u_t = - K[u]. \tag{17.18}$$

To start with, we shall restrict our attention to homogeneous boundary conditions. As in the one-dimensional situation of Chapter 14, the method of separation of variables is crucial. The separable solutions to any diffusion equation (17.13) are of exponential form

$$u(t, x, y) = e^{- \lambda t} v(x, y). \tag{17.19}$$

Since the linear operator $K = \nabla^* \circ \nabla$ only involves differentiation with respect to the spatial variables x, y, we find

$$\frac{\partial u}{\partial t} = - \lambda e^{- \lambda t} v(x, y), \qquad \text{while} \qquad K[u] = e^{- \lambda t} K[v].$$

Substituting back into the diffusion equation (17.18) and canceling the exponentials, we conclude that $v(x, y)$ must be an eigenfunction for the boundary value problem

$$K[v] = \lambda v. \tag{17.20}$$

The eigenfunction v is also required to satisfy the relevant homogeneous boundary conditions. In the case of the heat equation (17.1), $K[u] = -\gamma \Delta u$, and hence the eigenvalue equation (17.20) takes the form

$$\gamma \Delta v + \lambda v = 0, \qquad \text{or, in detail,} \qquad \gamma \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + \lambda v = 0. \tag{17.21}$$

This generalization of the Laplace equation is known as the *Helmholtz equation*, and was briefly discussed in Example 15.22.

The fact that K is a positive semi-definite linear operator implies that its eigenvalues are all real and non-negative. We order them in increasing size:

$$0 \leq \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots, \qquad \text{with} \qquad \lambda_n \longrightarrow \infty \qquad \text{as} \qquad n \longrightarrow \infty. \tag{17.22}$$

An eigenvalue is repeated according to the number (which is necessarily finite) of linearly independent eigenfunctions it admits. The problem has a zero eigenvalue, $\lambda_1 = 0$ if and only if the operator K is not positive definite, i.e., only in the case of pure Neumann boundary conditions. We refer the interested reader to [**39**; Chapter V] for the advanced theoretical details.

Each eigenvalue and eigenfunction pair will produce a separable solution

$$u_k(t,x,y) = e^{-\lambda_k t}\, v_k(x,y)$$

to the diffusion equation (17.18). The solutions corresponding to positive eigenvalues are exponentially decaying in time, while a zero eigenvalue, which only occurs in the positive semi-definite case, produces a constant solution. The general solution to the homogeneous boundary value problem can then be built up as a linear combination of these basic solutions, in the form of an *eigenfunction series*

$$u(t,x,y) = \sum_{k=1}^{\infty} c_k\, u_k(t,x,y) = \sum_{k=1}^{\infty} c_k\, e^{-\lambda_k t}\, v_k(x,y), \tag{17.23}$$

which is a form of generalized Fourier series. The eigenfunction coefficients c_k are prescribed by the initial conditions, which require

$$\sum_{k=1}^{\infty} c_k\, v_k(x,y) = f(x,y). \tag{17.24}$$

Thus, to solve the initial value problem, we need to expand the initial data as a series in the eigenfunctions for the Helmholtz boundary value problem.

To compute the coefficients c_k in the eigenfunction expansion (17.24), we appeal, as in the case of ordinary Fourier series, to orthogonality. Self-adjointness of the differential operator $K[u]$ implies that the corresponding eigenfunction solutions $v_1(x,y), v_2(x,y), \dots$ to (17.20) are automatically orthogonal with respect to the underlying inner product[†]

$$\langle\, v_j\,;v_k\,\rangle = 0, \qquad\qquad j \neq k.$$

The general argument establishing this result can be found in Theorem 8.20; see also Exercise ∎. As a consequence, taking the inner product of both sides of (17.24) with the eigenfunction v_k leads to the equation

$$c_k \,\|\, v_k \,\|^2 = \langle\, f\,;v_k\,\rangle \qquad \text{and hence} \qquad c_k = \frac{\langle\, f\,;v_k\,\rangle}{\|\, v_k\,\|^2}\,.$$

In this manner we recover our standard orthogonality formula (5.8) for expressing elements of a vector space in terms of an orthogonal basis. For a general diffusion equation, the orthogonality formula has the explicit form

$$c_k = \frac{\displaystyle\iint_\Omega f(x,y)\, v_k(x,y)\, \sigma(x,y)\, dx\, dy}{\displaystyle\iint_\Omega v_k(x,y)^2\, \sigma(x,y)\, dx\, dy}\,, \tag{17.25}$$

[†] If an eigenvalue is repeated, one needs to make sure that one chooses an orthogonal basis for its eigenspace.

where the weighting function $\sigma(x,y)$ was defined in (17.9). In the case of the heat equation, σ is constant and so can be canceled from both numerator and denominator, leaving the simpler formula

$$c_k = \frac{\displaystyle\iint_\Omega f(x,y)\, v_k(x,y)\, dx\, dy}{\displaystyle\iint_\Omega v_k(x,y)^2\, dx\, dy}\,. \tag{17.26}$$

Under fairly general hypotheses, it can be shown that the eigenfunctions form a *complete system*, which means that the Fourier series (17.24) will converge (at least in norm) to the function $f(x,y)$, provided it is not too bizarre. See [**39**; p. 369] for a proof of the following general theorem.

Theorem 17.1. *Let Ω be a bounded domain with piecewise smooth boundary. If $f(x,y)$ is an L^2 function on Ω, then its generalized Fourier series* (17.24) *with coefficients defined by* (17.25) *converges in norm to f. Moreover, if $f \in \mathrm{C}^2$ is twice continuously differentiable, then its generalized Fourier series converges uniformly to $f(x,y)$ for all $(x,y) \in \Omega$.*

Qualitative Properties

Before tackling simple examples where we find ourselves in a position to construct explicit formulae for the eigenfunctions and eigenvalues, let us see what the series solution (17.23) can tell us about general diffusion processes. Based on our experience with the case of a one-dimensional bar, the final conclusions will not be especially surprising. Indeed, they also apply, word for word, to diffusion processes in three-dimensional solid bodies. A reader who prefers to see explicit solution formulae may wish to skip ahead to the following section, returning here after digesting the solution formulae.

Keep in mind that we are still dealing with the solution to the homogeneous boundary value problem. The first observation is that all terms in the series solution (17.23), with the possible exception of a null eigenfunction term that appears in the semi-definite case, are tending to zero exponentially fast. Since most eigenvalues are large, all the higher order terms in the series become almost instantaneously negligible, and hence the solution can be accurately approximated by a finite sum over the first few eigenfunction modes. As time goes on, more and more of the modes can be neglected, and the solution decays to thermal equilibrium at an exponentially fast rate. The rate of convergence to thermal equilibrium is, for most initial data, governed by the smallest positive eigenvalue $\lambda_1 > 0$ for the Helmholtz boundary value problem on the domain.

In the positive definite cases of homogeneous Dirichlet or mixed boundary conditions, thermal equilibrium is $u(t,x,y) \to 0$. Thus, in these cases, the equilibrium temperature is equal to the boundary temperature — even if this temperature is only fixed on a part of the boundary. The heat dissipates away through the non-insulated part of the boundary. In the semi-definite Neumann case, corresponding to a completely insulated plate, the final thermal equilibrium temperature is constant — a multiple of the null eigenfunction

solution $u_0(t,x,y) = 1$. In this case, the general solution has the form

$$u(t,x,y) = c_0 + \sum_{k=1}^{\infty} c_k\, e^{-\lambda_k t}\, v_k(x,y), \tag{17.27}$$

where the sum is over the positive eigenmodes, $\lambda_k > 0$. Since all the summands are exponentially decaying, the equilibrium temperature $u_\star = c_0$ is the same as the constant term in the eigenfunction expansion. We evaluate this term using the orthogonality formula (17.25), and so, as $t \to \infty$,

$$u(t,x,y) \;\longrightarrow\; c_0 = \frac{\langle\, f\,;1\,\rangle}{\|\,1\,\|^2} = \frac{\displaystyle\iint_\Omega f(x,y)\,\sigma(x,y)\,dx\,dy}{\displaystyle\iint_\Omega \sigma(x,y)\,dx\,dy}\,,$$

which is a weighted average of the initial temperature over the domain. In particular, in the case of the heat equation, the weighting function σ is constant, and so the equilibrium temperature

$$u(t,x,y) \quad\longrightarrow\quad c_0 = \frac{1}{\text{area }\Omega}\iint_\Omega f(x,y)\,dx\,dy \tag{17.28}$$

equals the average or mean initial temperature distribution. In this case, the heat cannot escape through the boundary, and redistributes itself in a uniform manner over the domain.

As with its one-dimensional form, the planar heat equation has a smoothing effect on the initial temperature distribution $f(x,y)$. Assume that the eigenfunction coefficients are uniformly bounded, so $|\,c_k\,| \le M$ for some constant M. This will certainly be the case if $f(x,y)$ is piecewise continuous, but even holds for quite rough initial data, including delta functions. Then, at any time $t > 0$ after the initial instant, the coefficients $c_k\, e^{-\lambda_k t}$ in the eigenfunction series solution (17.23) are exponentially small as $k \to \infty$, which is enough to ensure smoothness[†] of the solution $u(t,x,y)$ for each $t > 0$. Therefore, a diffusion equation immediately smoothes out jumps, corners and other discontinuities in the initial data. As time progresses, the local variations in the solution become less and less, eventually (or, more accurately, asymptotically) reaching a constant equilibrium state.

For this reason, diffusion processes can be effectively applied to clean and denoise planar images. In this application, the initial data $f(x,y)$ represents the grey-scale value of the image at position (x,y), so that $0 \le f(x,y) \le 1$ with 0 representing black, and 1 representing white. As time progresses, the solution $u(t,x,y)$ represents a more and more smoothed version of the image. Although this has the effect of removing unwanted noise from the image, there is also a gradual blurring of the actual features. Thus, the "time" or "multiscale" parameter t needs to be chosen to optimally balance between the two effects — the larger t is the more noise is removed, but the more noticeable the blurring. A representative illustration appears in Figure im2. To further suppress undesirable blurring effects, recent image processing filters are based on anisotropic (and thus *nonlinear*) diffusion equations. See Sapiro, [**128**], for a survey of recent progress in this active field.

[†] For a general diffusion equation, this requires that the functions $\sigma(x,y)$ and $\kappa(x,y)$ be smooth.

Since the forwards heat equation blurs the features in an image, running it backwards in time should effectively sharpen the image. However, the one-dimensional argument presented in Section 14.1 tells us that any direct attempt to run the heat flow backwards immediately leads to difficulties, and the backwards diffusion equation is ill posed. Various strategies have been proposed to circumvent this mathematical barrier, and thereby design effective image enhancement algorithms.

Inhomogeneous Boundary Conditions and Forcing

Finally, let us briefly mention how to incorporate inhomogeneous boundary conditions and external heat sources into the problem. Consider, as a specific example, the forced heat equation

$$u_t = \gamma\, \Delta u + F(x,y), \qquad\qquad (x,y) \in \Omega, \tag{17.29}$$

where $F(x,y)$ represents an unvarying external heat source, subject to inhomogeneous Dirichlet boundary conditions

$$u = h \qquad \text{for} \qquad (x,y) \in \partial\Omega. \tag{17.30}$$

When the external forcing is fixed for all t, we expect the solution to eventually settle down to an equilibrium configuration: $u(t,x,y) \to u_star(x,y)$ as $t \to \infty$.

To determine the time-independent equilibrium temperature $u = u_\star(x,y)$, we set $u_t = 0$ in the differential equation (17.29). We immediately see that equilibrium for this problem is characterized by the solution to the Poisson equation

$$-\gamma\, \Delta u_\star = F, \qquad\qquad (x,y) \in \Omega, \tag{17.31}$$

subject to the same inhomogeneous Dirichlet boundary conditions (17.30). Positive definiteness of the Dirichlet boundary value problem implies that there is a unique equilibrium solution is , and can be characterized as the sole minimizer of the Dirichlet principle; for details see Section 15.4.

Once we have determined the equilibrium solution — usually through a numerical approximation — we set

$$\widetilde{u}(t,x,y) = u(t,x,y) - u_\star(x,y),$$

so that $\widetilde{u}$ measures the deviation of the solution u from its eventual equilibrium. By linearity $\widetilde{u}(t,x,y)$ satisfies the unforced heat equation subject to homogeneous boundary conditions:

$$\widetilde{u}_t = \gamma\, \Delta \widetilde{u}, \qquad (x,y) \in \Omega, \qquad\qquad u = 0, \qquad (x,y) \in \partial\Omega.$$

Therefore, $\widetilde{u}$ can be expanded in an eigenfunction series (17.23), and hence will decay to zero, $\widetilde{u}(t,x,y) \to 0$, at a exponentially fast rate governed by the smallest eigenvalue λ_1 of the corresponding homogeneous Helmholtz boundary value problem. Consequently, the solution to the forced, inhomogeneous problem

$$u(t,x,y) = \widetilde{u}(t,x,y) + u_\star(x,y) \;\longrightarrow\; u_\star(x,y)$$

will approach thermal equilibrium, namely $u_\star(x,y)$, at the same exponential rate as the homogeneous boundary value problem.

17.3. Explicit Solutions for the Heat Equation.

As noted earlier, explicit solution formulae are few and far between. In this section, we discuss two specific cases where the Helmholtz eigenfunctions can be found in closed form. The calculations rely on a separation of variables, which is applicable only to a rather limited restricted class of domains, which include the rectangles and disks that we discuss in detail here.

Heating of a Rectangle

A homogeneous rectangular plate

$$R = \big\{\, 0 < x < a,\ 0 < y < b \,\big\}$$

is heated to a prescribed initial temperature $u(0, x, y) = f(x, y)$ and then insulated. The sides of the plate are held at zero temperature. Our task is to determine how fast the plate returns to thermal equilibrium.

The temperature $u(t, x, y)$ evolves according to the heat equation

$$u_t = \gamma(u_{xx} + u_{yy}),$$

subject to homogeneous Dirichlet conditions

$$u(0, y) = u(a, y) = 0 = u(x, 0) = u(x, b), \qquad 0 < x < a, \quad 0 < y < b, \tag{17.32}$$

along the boundary of the rectangle. As in (17.19), the basic solutions to the heat equation are obtained from an exponential ansatz $u(t, x, y) = e^{-\lambda t}\, v(x, y)$. Substituting this expressing into the heat equation, we find that the function $v(x, y)$ solves the Helmholtz eigenvalue problem

$$\gamma(v_{xx} + v_{yy}) + \lambda\, v = 0, \qquad (x, y) \in R, \tag{17.33}$$

subject to the same homogeneous Dirichlet boundary conditions

$$v(x, y) = 0, \qquad (x, y) \in \partial R. \tag{17.34}$$

To solve the rectangular Helmholtz eigenvalue problem, we shall, as in (15.13), introduce a further separation of variables, writing

$$v(x, y) = p(x)\, q(y)$$

as the product of functions depending upon the individual Cartesian coordinates. Substituting this ansatz into the Helmholtz equation (17.33), we find

$$\gamma\, p''(x)\, q(y) + \gamma\, p(x)\, q''(y) + \lambda\, p(x)\, q(y) = 0.$$

To effect the variable separation, we collect all terms involving x on one side and all terms involving y on the other side of the equation. This is accomplished by dividing by $v = pq$ and rearranging the terms; the result is

$$\gamma\, \frac{p''}{p} = -\,\gamma\, \frac{q''}{q} - \lambda \equiv -\,\mu.$$

The left hand side of this equation only depends on x, whereas the right hand side only depends on y. As argued in Section 15.2, the only way this can occur is if the two sides equal a common *separation constant*, denoted by $-\mu$. (The minus sign is for later convenience.) In this manner, we reduce our partial differential equation to a pair of one-dimensional eigenvalue problems

$$\gamma\, p'' + \mu\, p = 0, \qquad\qquad \gamma\, q'' + (\lambda - \mu)\, q = 0,$$

each of which is subject to homogeneous Dirichlet boundary conditions

$$p(0) = p(a) = 0, \qquad\qquad q(0) = q(b) = 0.$$

To obtain a nontrivial solution to the Helmholtz equation, we seek nonzero solutions to these two supplementary eigenvalue problems. The fact that we are dealing with a rectangular domain is critical to the success of this approach.

We have already solved these particular two boundary value problems many times; see, for instance, equation (14.17). The eigenfunctions are, respectively,

$$p_m(x) = \sin \frac{m\,\pi\, x}{a}\,, \qquad m = 1, 2, 3, \ldots, \qquad\qquad q_n(y) = \sin \frac{n\,\pi\, y}{b}\,, \qquad n = 1, 2, 3, \ldots,$$

with

$$\mu = \frac{m^2\,\pi^2\,\gamma}{a^2}\,, \qquad \lambda - \mu = \frac{n^2\,\pi^2\,\gamma}{b^2}\,, \qquad \text{so that} \qquad \lambda = \frac{m^2\,\pi^2\,\gamma}{a^2} + \frac{n^2\,\pi^2\,\gamma}{b^2}\,.$$

Therefore, the separable eigenfunction solutions to the Helmholtz boundary value problem (17.32), (17.33) have the doubly trigonometric form

$$v_{m,n}(x,y) = \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}\,, \tag{17.35}$$

with corresponding eigenvalues

$$\lambda_{m,n} = \frac{m^2\,\pi^2\,\gamma}{a^2} + \frac{n^2\,\pi^2\,\gamma}{b^2} = \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\pi^2\,\gamma\,. \tag{17.36}$$

Each of these corresponds to an exponentially decaying, separable solution

$$u_{m,n}(t,x,y) = e^{-\lambda_{m,n}\,t}\, v_{m,n}(x,y) = \exp\left[-\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\pi^2\,\gamma\, t\right]\ \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b} \tag{17.37}$$

to the original heat equation.

Using the fact that the univariate sine functions form a complete system, it is not hard to prove, [**146**], that the separable eigenfunction solutions (17.37) are complete, which means that there are no non-separable eigenfunctions. As a consequence, the general solution to the initial-boundary value problem can be expressed as a linear combination

$$u(t,x,y) = \sum_{m,n=1}^{\infty} c_{m,n}\, u_{m,n}(t,x,y) = \sum_{m,n=1}^{\infty} c_{m,n}\, e^{-\lambda_{m,n}\,t}\, v_{m,n}(x,y) \tag{17.38}$$

of our eigenfunction modes. The coefficients $c_{m,n}$ are prescribed by the initial conditions, which take the form of a double Fourier sine series

$$f(x,y) = u(0,x,y) = \sum_{m,n=1}^{\infty} c_{m,n} v_{m,n}(x,y) = \sum_{m,n=1}^{\infty} c_{m,n} \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b}.$$

Self-adjointness of the Laplacian coupled with the boundary conditions implies that the eigenfunctions $v_{m,n}(x,y)$ are orthogonal with respect to the L^2 inner product on the rectangle, so

$$\langle\, v_{k,l}\,;v_{m,n}\,\rangle = \int_0^b \int_0^a v_{k,l}(x,y)\, v_{m,n}(x,y)\, dx\, dy = 0 \qquad \text{unless} \qquad k = m \quad \textit{and} \quad l = n.$$

(The skeptical reader can verify the orthogonality relations directly from the formulae for the eigenfunctions.) Thus, we can use our usual orthogonality formula (17.26) to compute the coefficients

$$c_{m,n} = \frac{\langle\, f\,;v_{m,n}\,\rangle}{\|\, v_{m,n}\,\|^2} = \frac{1}{a\,b}\int_0^b \int_0^a f(x,y) \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b}\, dx\, dy, \tag{17.39}$$

where the formula for the norms of the eigenfunctions

$$\|\, v_{m,n}\,\|^2 = \int_0^b \int_0^a v_{m,n}(x,y)^2\, dx\, dy = a\,b. \tag{17.40}$$

follows from a direct evaluation of the double integral. (Unfortunately, while the orthogonality is automatic, the computation of the norm must inevitably be done "by hand".)

The rectangle approaches thermal equilibrium at the rate equal to the smallest eigenvalue,

$$\lambda_{1,1} = \left(\frac{1}{a^2} + \frac{1}{b^2}\right)\pi^2\,\gamma, \tag{17.41}$$

which depends upon the reciprocals of the squared lengths of the sides of the rectangle and the diffusion coefficient. The larger the rectangle, or the smaller the diffusion coefficient, the smaller $\lambda_{1,1}$, and hence slower the return to thermal equilibrium. The higher modes, with m and n large, decay to zero almost instantaneously, and so the solution immediately behaves like a finite sum over a few low order modes. Assuming that $c_{1,1} \neq 0$, the slowest decaying mode in the Fourier series (17.38) is

$$c_{1,1}\, u_{1,1}(t,x,y) = c_{1,1} \exp\left[-\left(\frac{1}{a^2} + \frac{1}{b^2}\right)\pi^2\,\gamma\, t\right] \sin\frac{\pi x}{a} \sin\frac{\pi y}{b}. \tag{17.42}$$

Thus, in the long run, the temperature is of one sign, either positive or negative depending upon the sign of $c_{1,1}$, throughout the rectangle. As in the one-dimensional case, this observation is, in fact, indicative of the general phenomenon that the eigenfunction associated with the smallest positive eigenvalue of the Laplacian is of one sign throughout the domain.

A typical solution is plotted in Figure heatrect■

Heating of a Disk

Let us perform a similar analysis on the heating of a circular disk. For simplicity, we take the diffusion coefficient $\gamma = 1$. We also assume that the disk $D = \{ x^2 + y^2 \leq 1 \}$ has radius 1. We shall solve the heat equation on D subject to homogeneous Dirichlet boundary values of zero temperature at the circular edge $\partial D = C$. Thus, the full initial-boundary value problem is

$$\begin{aligned} &u_t = \Delta u, && x^2 + y^2 < 1, \qquad t > 0, \\ &u(t,x,y) = 0, && x^2 + y^2 = 1, \\ &u(0,x,y) = f(x,y), && x^2 + y^2 \leq 1. \end{aligned} \tag{17.43}$$

A simple rescaling of space and time can be used to recover the solution for an arbitrary diffusion coefficient and a disk of arbitrary radius from this particular case; see Exercise .

Since we are working in a circular domain, we instinctively pass to polar coordinates (r, θ). In view of the polar coordinate formula (15.27) for the Laplace operator, the heat equation and boundary conditions take the form

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \theta^2}\,, && 0 \leq r < 1, \qquad t > 0, \\ &\quad u(t,1,\theta) = 0, && u(0,r,\theta) = f(r,\theta), \qquad r \leq 1, \end{aligned} \tag{17.44}$$

where the solution $u(t,r,\theta)$ is defined for all $0 \leq r \leq 1$ and $t \geq 0$. Moreover,

$$u(t,r,\theta + 2\,\pi) = u(t,r,\theta)$$

must be a 2π periodic function of the angular variable to ensure that it represents a single-valued function on the entire disk.

To obtain the separable solutions

$$u(t,r,\theta) = e^{-\,\lambda\, t}\, v(r,\theta),$$

we need to solve the polar coordinate form of the Helmholtz equation

$$\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\,\frac{\partial v}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 v}{\partial \theta^2} + \lambda\, v = 0, \qquad \begin{aligned} &0 \leq r < 1, \\ &0 \leq \theta \leq 2\,\pi, \end{aligned} \tag{17.45}$$

subject to the boundary conditions

$$v(1,\theta) = 0, \qquad v(r,\theta + 2\pi) = v(r,\theta). \tag{17.46}$$

We apply a further separation of variables to the polar Helmholtz equation by writing

$$v(r,\theta) = p(r)\, q(\theta). \tag{17.47}$$

Substituting into (17.45), and then collecting together all terms involving r and all terms involving θ, we are led to the pair of ordinary differential equations

$$q'' + \mu\, q = 0, \qquad r^2\, p'' + r\, p' + (\lambda\, r^2 - \mu)\, p = 0,$$

where λ is the Helmholtz eigenvalue, and μ the separation constant. The periodicity condition (17.46) requires that $q(\theta)$ be 2π periodic, and hence

$$q(\theta) = \cos m\,\theta \qquad \text{or} \qquad \sin m\,\theta, \qquad \text{where} \qquad \mu = m^2, \tag{17.48}$$

span the eigenspace. Note that when $m = 0$, there is only one independent periodic solution, namely $q(\theta) \equiv 1$; the second solution, $q(\theta) = \theta$, does not satisfy the periodicity constraint.

Using the preceding formula for the separation constant, $\mu = m^2$, the second differential equation for $p(r)$ assumes the form

$$r^2\,\frac{d^2p}{dr^2} + r\,\frac{dp}{dr} + (\lambda\, r^2 - m^2)\,p = 0, \qquad 0 \le r \le 1. \tag{17.49}$$

Ordinarily, one requires two boundary conditions to specify a solution to such a second order boundary value problem. But our Dirichlet condition, namely $p(1) = 0$, only specifies its value at one of the endpoints: $r = 1$. The other endpoint, $r = 0$, is a *singular point* for the ordinary differential equation, because the coefficient, r^2, of the highest order derivative, p'', vanishes there. This should remind you of our solution to the Euler differential equation (15.32) when we solved the Laplace equation on the disk. As there, we only need that the solution be bounded at $r = 0$, and hence are led to require

$$|\,p(0)\,| < \infty, \qquad\qquad p(1) = 0. \tag{17.50}$$

These singular boundary conditions turn out to be sufficient to distinguish the relevant eigenfunction solutions to (17.49).

Although (17.49) arises in a variety of applications, this may be the first time that you have encountered this particular ordinary differential equation. It is not an elementary equation, and its solutions cannot be written in terms of elementary functions. Nevertheless, owing to its significance in a wide range of physical applications, its solutions have been extensively studied and are, in a sense, well-known. After some preliminary manipulations, we shall summarize the relevant properties of the solutions, leaving details and proofs to Appendix C.

To simplify the analysis, we make a preliminary rescaling of the independent variable, replacing by r by

$$z = \sqrt{\lambda}\, r.$$

Note that, by the chain rule,

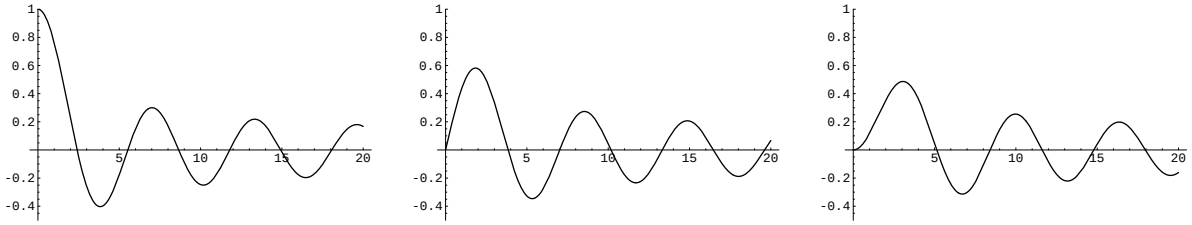
$$\frac{dp}{dr} = \sqrt{\lambda}\,\frac{dp}{dz}\,, \qquad\qquad \frac{d^2p}{dr^2} = \lambda\,\frac{d^2p}{dz^2}\,,$$

and hence

$$r\,\frac{dp}{dr} = z\,\frac{dp}{dz}\,, \qquad\qquad r^2\,\frac{d^2p}{dr^2} = z^2\,\frac{d^2p}{dz^2}\,.$$

The net effect is to eliminate the eigenvalue parameter λ (or, rather, hide it in the change of variables), so that (17.49) assumes the slightly simpler form

$$z^2\,\frac{d^2p}{dz^2} + z\,\frac{dp}{dz} + (z^2 - m^2)\,p = 0. \tag{17.51}$$



$J_0(z)$ $J_1(z)$ $J_2(z)$

Figure 17.1. Bessel Functions.

The ordinary differential equation (17.51) is known as *Bessel's equation*, named after the early 19th century astronomer Wilhelm Bessel, who used the solutions to solve a problem arising in the study of planetary orbits. The solutions to Bessel's equation are an indispensable tool in applied mathematics, physics and engineering.

The Bessel equation cannot (except in special instances) be solved in terms of elementary functions. The one thing we know for sure is that, as with any second order ordinary differential equation, there are two linearly independent solutions. However, it turns out that, up to constant multiple, only one solution remains bounded as $z \to 0$. This solution is known as the *Bessel function* of *order* m, and is denoted by $J_m(z)$. Applying the general systematic method for finding power series solutions to linear ordinary differential equations presented in Appendix C, it can be shown that the Bessel function of order m has the Taylor expansion

$$J_m(z) = \sum_{k=0}^{\infty} \frac{(-1)^k\, z^{m+2\,k}}{2^{m+2\,k}\, k!\, (m+k)!} \tag{17.52}$$
$$= \frac{z^m}{2^m\, m!}\left[\, 1 - \frac{z^2}{4\,(m+1)} + \frac{z^4}{32\,(m+1)(m+2)} - \frac{z^6}{384\,(m+1)(m+2)(m+3)} + \,\cdots\,\right]$$

at the origin $z = 0$. Verification that this series solves the Bessel equation of order m is a straightforward exercise. Moreover, a simple application of the ratio test for power series tells us that the series converges for all (complex) values of z. Indeed, the convergence is quite rapid when z is of moderate size, and so summing the series is a reasonably effective method for computing the Bessel function $J_m(z)$ — although in serious applications one adopts more sophisticated numerical techniques based on asymptotic expansions and integral formulae, [**3**, **116**]. Figure 17.1 displays graphs of the first three Bessel functions for $z > 0$. Most software packages, both symbolic and numerical, contain routines for accurately evaluating and graphing Bessel functions.

Reverting back to our original radial coordinate $r = z/\sqrt{\lambda}$, we conclude that every solution to the radial equation (17.49) which is bounded at $r = 0$ is a constant multiple

$$p(r) = J_m\big(\,\sqrt{\lambda}\, r\,\big) \tag{17.53}$$

of the rescaled Bessel function of order m. So far, we have only dealt with the boundary condition at the singular point $r = 0$. The Dirichlet condition at the other end requires

$$p(1) = J_m\big(\,\sqrt{\lambda}\,\big) = 0.$$

Therefore, in order that λ be a legitimate eigenvalue, $\sqrt{\lambda}$ must be a *root* of the m^{th} order Bessel function J_m.

Remark: We already know, owing to the positive definiteness of the Dirichlet boundary value problem, that the eigenvalues $\lambda > 0$ must be positive, so there is no problem taking the square root. Indeed, it can be proved that the Bessel functions do not have any negative roots!

The graphs of $J_m(z)$ strongly indicate, and, indeed, it can be rigorously proved, that each Bessel function oscillates between positive and negative values as z increases above 0, with slowly decreasing amplitude. As a consequence, there exists an infinite sequence of *Bessel roots*, which we number in the order in which they appear:

$$J_m(\zeta_{m,n}) = 0, \qquad \text{where} \qquad 0 < \zeta_{m,1} < \zeta_{m,2} < \zeta_{m,3} < \cdots \qquad \text{with} \quad \zeta_{m,n} \longrightarrow \infty. \tag{17.54}$$

It is worth noting that the Bessel functions are *not* periodic, and their roots are not evenly spaced.

Owing to their physical importance in a wide range of problems, the Bessel roots have been extensively tabulated in the literature, cf. [**3**, **48**]. A table of all Bessel roots that are < 12 in magnitude follows. The rows of the table are indexed by n, the root number, and the columns by m, the order of the Bessel function.

Table of Bessel Roots $\zeta_{m,n}$

	0	1	2	3	4	5	6	7
1	2.4048	3.8317	5.1356	6.3802	7.5883	8.7715	9.9361	11.0860
2	5.5201	7.0156	8.4172	9.761	11.0650	⋮	⋮	⋮
3	8.6537	10.1730	11.6200	⋮	⋮			
4	11.7920	⋮	⋮					
⋮	⋮							

Remark: According to (17.52),

$$J_m(0) = 0 \qquad \text{for} \qquad m > 0, \qquad \text{while} \qquad J_0(0) = 1.$$

However, we do not count 0 as a *bona fide* Bessel root, since it does not lead to a valid eigenfunction for the Helmholtz boundary value problem.

Summarizing our progress, the eigenvalues

$$\lambda_{m,n} = \zeta_{m,n}^2, \qquad n = 1, 2, 3, \ldots, \qquad m = 0, 1, 2, \ldots, \tag{17.55}$$

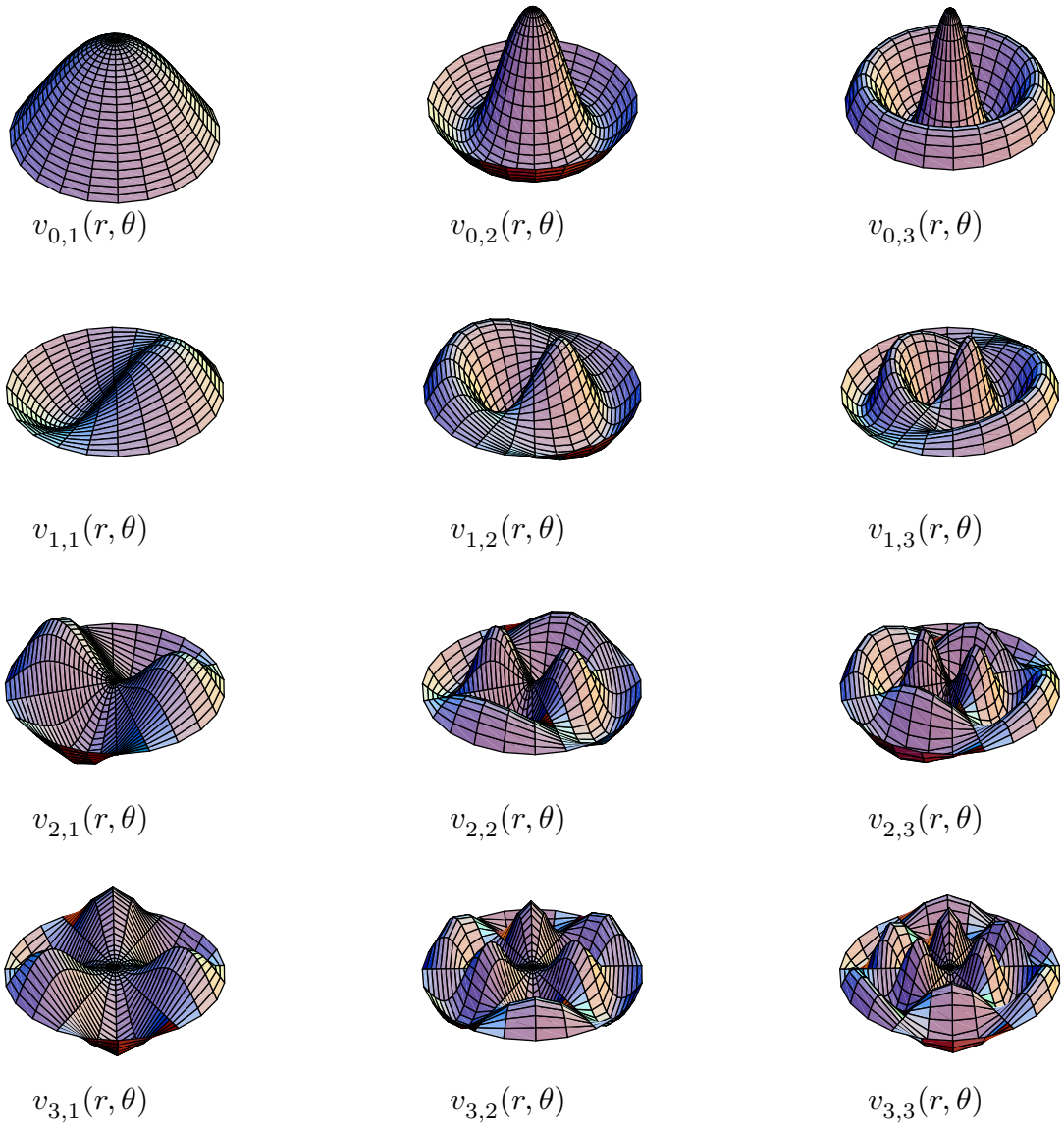


Figure 17.2. Fundamental Modes for a Disk.

of the Bessel boundary value problem (17.49), (17.50) are the squares of the roots of the Bessel function of order m. The corresponding eigenfunctions are

$$w_{m,n}(r) = J_m(\zeta_{m,n}\, r)\,, \qquad n = 1, 2, 3, \ldots, \qquad m = 0, 1, 2, \ldots, \tag{17.56}$$

defined for $0 \le r \le 1$. Combining (17.56) with the formula (17.48) for the angular components, we conclude that the separable solutions (17.47) to the polar Helmholtz boundary

value problem (17.45) are

$$\begin{aligned} v_{0,n}(r,\theta) &= J_0(\zeta_{0,n}\, r), \\ v_{m,n}(r,\theta) &= J_m(\zeta_{m,n}\, r)\, \cos m\,\theta, \\ \widehat{v}_{m,n}(r,\theta) &= J_m(\zeta_{m,n}\, r)\, \sin m\,\theta, \end{aligned} \quad \text{where} \quad \begin{array}{l} n = 1, 2, 3, \dots, \\ m = 1, 2, \dots. \end{array}$$

These solutions define the *natural modes* of vibration for a disk, and Figure 17.2 plots the first few of them. The eigenvalues $\lambda_{0,n}$ are simple, and contribute radially symmetric eigenfunctions, whereas the eigenvalues $\lambda_{m,n}$ for $m > 0$ are double, and produce two linearly independent separable eigenfunctions, with trigonometric dependence on the angular variable. As in the rectangular case, it is possible to prove that the separable eigenfunctions are *complete* — there are no other eigenfunctions — and, moreover, every (reasonable) function defined on the unit disk can be written as a generalized Fourier series in the Bessel eigenfunctions.

We have now produced the basic solutions

$$\begin{aligned} u_{0,n}(t,r) &= e^{-\,\zeta_{0,n}^2\, t}\, J_0(\zeta_{0,n}\, r), \\ u_{m,n}(t,r,\theta) &= e^{-\,\zeta_{m,n}^2\, t}\, J_m(\zeta_{m,n}\, r)\, \cos m\,\theta, \\ \widehat{u}_{m,n}(t,r,\theta) &= e^{-\,\zeta_{m,n}^2\, t}\, J_m(\zeta_{m,n}\, r)\, \sin m\,\theta, \end{aligned} \quad \begin{array}{l} n = 1, 2, 3, \dots, \\ m = 1, 2, \dots. \end{array} \tag{17.57}$$

to the Dirichlet boundary value problem for the heat equation on the unit disk. The general solution is a linear superposition, in the form of an infinite series

$$u(t,r,\theta) = \sum_{n=1}^{\infty} a_{0,n}\, u_{0,n}(t,r) + \sum_{m,n=1}^{\infty} \big[\, a_{m,n}\, u_{m,n}(t,r,\theta) + b_{m,n}\, \widehat{u}_{m,n}(t,r,\theta) \,\big].$$

As usual, the coefficients $a_{m,n}, b_{m,n}$ are determined by the initial condition, so

$$u(0,r,\theta) = \sum_{n=1}^{\infty} a_{0,n}\, v_{0,n}(r) + \sum_{m,n=1}^{\infty} \big[\, a_{m,n}\, v_{m,n}(r,\theta) + b_{m,n}\, \widehat{v}_{m,n}(r,\theta) \,\big] = f(r,\theta).$$

Thus, we must expand the initial data into a *Fourier–Bessel series*, which involves Bessel functions along with the original Fourier trigonometric functions.

According to Section 17.2, the eigenfunctions are orthogonal[†] with respect to the standard L^2 inner product

$$\langle\, u\,;v\,\rangle = \iint_D u(x,y)\, v(x,y)\, dx\, dy = \int_0^1 \int_0^{2\,\pi} f(r,\theta)\, u(r,\theta)\, v(r,\theta)\; r\, d\theta\, dr$$

on the unit disk. (Note the extra factor of r coming from the polar coordinate form (A.51) of the area element $dx\, dy = r\, dr\, d\theta$.) The norms of the Fourier–Bessel functions are given by the interesting formula

$$\| v_{m,n} \| = \| \widehat{v}_{m,n} \| = \sqrt{\pi}\; \big|\, J_m'(\zeta_{m,n}) \,\big| \tag{17.58}$$

[†] Technically, this follows from general principles except for the two eigenfunctions corresponding to the double eigenvalues, whose orthogonality must be verified by hand.

that involves the value of the *derivative* of the Bessel function at the appropriate Bessel root. A proof of this formula will be given in Appendix C; see Exercises ■, ■. A table of their numerical values follows; as above, the rows are indexed by n and the columns by m.

Norms of the Fourier–Bessel functions $\| v_{m,n} \| = \| \widehat{v}_{m,n} \|$

	0	1	2	3	4	5	6	7
1	0.9202	0.7139	0.6020	0.5287	0.4757	0.4350	0.4026	0.3759
2	0.6031	0.5319	0.4810	0.4421	0.4110	0.3854	0.3638	0.3453
3	0.4811	0.4426	0.4120	0.3869	0.3658	0.3477	0.3319	0.3180
4	0.4120	0.3870	0.3661	0.3482	0.3326	0.3189	0.3067	0.2958

Orthogonality of the eigenfunctions implies that

$$\begin{aligned} a_{m,n} &= \frac{\langle\, f\,;v_{m,n}\,\rangle}{\|\, v_{m,n}\,\|^2} = \frac{1}{\pi\, J_m'(\zeta_{m,n})^2} \int_0^1\int_0^{2\,\pi} f(r,\theta)\, J_m(\zeta_{m,n}\, r)\; r\cos m\,\theta\, d\theta\, dr, \\ b_{m,n} &= \frac{\langle\, f\,;\widehat{v}_{m,n}\,\rangle}{\|\, \widehat{v}_{m,n}\,\|^2} = \frac{1}{\pi\, J_m'(\zeta_{m,n})^2} \int_0^1\int_0^{2\,\pi} f(r,\theta)\, J_m(\zeta_{m,n}\, r)\; r\sin m\,\theta\, d\theta\, dr. \end{aligned} \tag{17.59}$$

In accordance with the general theory, each individual solution (17.57) to the heat equation decays exponentially fast, at a rate prescribed by the square of the corresponding Bessel root $\lambda_{m,n} = \zeta_{m,n}^2$. In particular, the dominant mode, meaning the one that persists the longest, is

$$u_{0,1}(t,r,\theta) = e^{-\,\zeta_{0,1}^2\, t}\, J_0(\zeta_{0,1}\, r). \tag{17.60}$$

Its decay rate $\zeta_{0,1}^2 \approx 5.783$ is the square of the first root of the Bessel function $J_0(z)$. This is the rate at which a disk whose boundary is held at zero temperature approaches thermal equilibrium. The dominant eigenfunction $v_{0,1}(r,\theta) = J_0(\zeta_{0,1}\, r) > 0$ is strictly positive within the entire disk and radially symmetric. Consequently, for most initial conditions (specifically those for which $c_{0,1} \neq 0$), the disk's temperature distribution eventually becomes entirely of one sign and radially symmetric, exponentially decaying to zero at the rate of slightly less than 6. See Figure 17.3 for a plot of a typical solution, displayed as successive times $t = 0, .04, .08, .12, .16, .2$. Note how, in accordance with the theory, the solution almost immediately acquires a radial symmetry, followed by a fairly rapid decay to thermal equilibrium.

17.4. The Fundamental Solution.

As we learned in Section 14.1, the *fundamental solution* to the heat equation measures the temperature distribution resulting from a concentrated initial heat source, e.g., a hot soldering iron applied instantaneously at one point of the plate. The physical problem is modeled mathematically by imposing a delta function as the initial condition for the heat

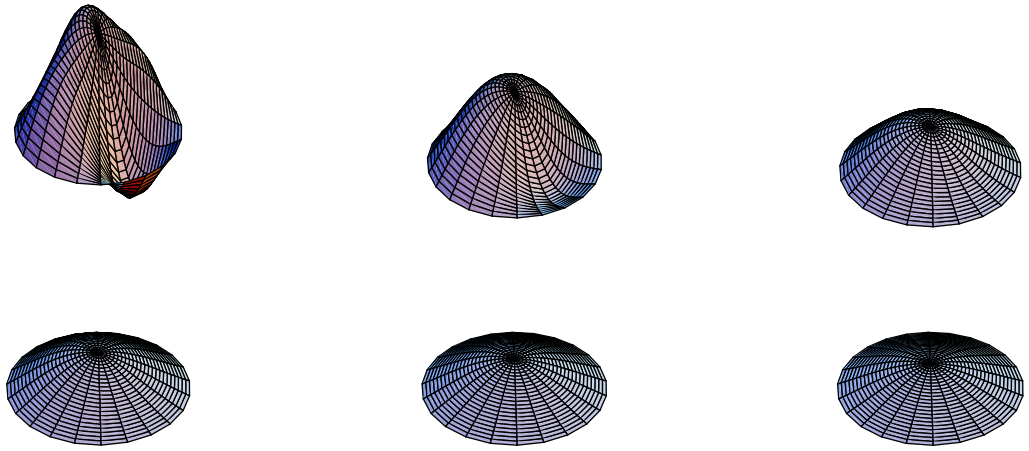


Figure 17.3. Heat Diffusion in a Disk.

equation, along with homogeneous boundary conditions of the appropriate type. Once we know the fundamental solution, we will be in a position to recover the solution for arbitrary initial data by a linear superposition principle.

As in the one-dimensional case, we shall concentrate on the most tractable case, when the domain is the entire plane: $\Omega = \mathbb{R}^2$. Our first goal will be to solve the initial value problem

$$u_t = \gamma\, \Delta u, \qquad u(0, x, y) = \delta(x - \xi, y - \eta) = \delta(x - \xi)\, \delta(y - \eta), \tag{17.61}$$

for all $t > 0$ and all $(x, y) \in \mathbb{R}^2$. The initial data is a delta function representing a concentrated unit heat source placed at position (ξ, η). The resulting solution $u = F(t, x, y; \xi, \eta)$ is the *fundamental solution* for the heat equation on all of $\mathbb{R}^2$.

The easiest route to the desired solution is the following simple lemma that uses solutions of the one-dimensional heat equation to construct solutions of the two-dimensional version.

Lemma 17.2. *If $v(t, x)$ and $w(t, x)$ are any two solutions to the one-dimensional heat equation $u_t = \gamma\, u_{xx}$, then the product*

$$u(t, x, y) = v(t, x)\, w(t, y) \tag{17.62}$$

is a solution to the two-dimensional heat equation $u_t = \gamma\, (u_{xx} + u_{yy})$.

Proof: Our assumptions imply that that $v_t = \gamma\, v_{xx}$, while $w_t = \gamma\, w_{yy}$ when we write $w(t, y)$ as a function of t and y. Therefore, when we differentiate (17.62),

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial t}\, w + v\, \frac{\partial w}{\partial t} = \gamma\, \frac{\partial^2 v}{\partial x^2}\, w + \gamma\, v\, \frac{\partial^2 w}{\partial y^2} = \gamma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right),$$

and hence $u(t, x, y)$ solves the heat equation. *Q.E.D.*

Thus, for example, if

$$v(t,x) = e^{-\gamma\omega^2 t}\,\cos\omega x, \qquad w(t,y) = e^{-\gamma\nu^2 t}\,\cos\nu y,$$

are separable solutions of the one-dimensional heat equation, then

$$u(t,x,y) = e^{-\gamma(\omega^2+\nu^2)t}\,\cos\omega x\,\cos\nu y$$

is one of the separable solutions in rectangular coordinates.

A more interesting case is to let

$$v(t,x) = \frac{1}{2\sqrt{\pi\gamma t}}\,e^{-(x-\xi)^2/4\gamma t}, \qquad w(t,y) = \frac{1}{2\sqrt{\pi\gamma t}}\,e^{-(y-\eta)^2/4\gamma t}, \tag{17.63}$$

both be the fundamental solutions (14.57) to the one-dimensional heat equation at points $x = \xi$ and $y = \eta$, respectively. Multiplying these two solutions together produces the fundamental solution for the two-dimensional problem.

Proposition 17.3. *The fundamental solution to the heat equation $u_t = \gamma\,\Delta u$ corresponding to a unit delta function placed at position $(\xi,\eta) \in \mathbb{R}^2$ at the initial time $t_0 = 0$ is*

$$F(t, x-\xi, y-\eta) = \frac{1}{4\pi\gamma t}\,e^{-[(x-\xi)^2+(y-\eta)^2]/4\gamma t}. \tag{17.64}$$

Proof: since we already know that (17.63) are solutions to the one-dimensional heat equation, Lemma 17.2 guarantees that $u(t,x,y) = v(t,x)\,w(t,y)$ as given by (17.64) solves the planar equation for $t > 0$. Moreover, at the initial time

$$u(o,x,y) = v(0,x)\,w(0,y) = \delta(x-\xi)\,\delta(y-\eta)$$

is a product of delta functions, and hence the result follows. We note that the total heat

$$\iint u(t,x,y)\,dx\,dy = \left(\int v(t,x)\,dx\right)\left(\int w(t,y)\,dy\right) = 1, \qquad t \geq 0,$$

remains constant, while

$$\lim_{t\to 0^+} u(t,x,y) \begin{cases} \infty, & (x,y) = (\xi,\eta), \\ 0, & \text{otherwise.} \end{cases}$$

has the standard delta function limit. *Q.E.D.*

As illustrated in Figure 17.4 at times $t = .01, .02, .05, .1$, the initially concentrated heat spreads out in a radially symmetric manner. The total amount of heat

$$\iint u(t,x,y)\,dx\,dy = 1, \qquad t \geq 0,$$

remains constant, but at each individual point (x,y), after a slight initial rise, the temperature decays back to zero at a rate proportional to $1/t$.

Both the planar fundamental solution and its one-dimensional have a bell-shaped Gaussian exponential profile. The one difference is the initial factor. In a one-dimensional

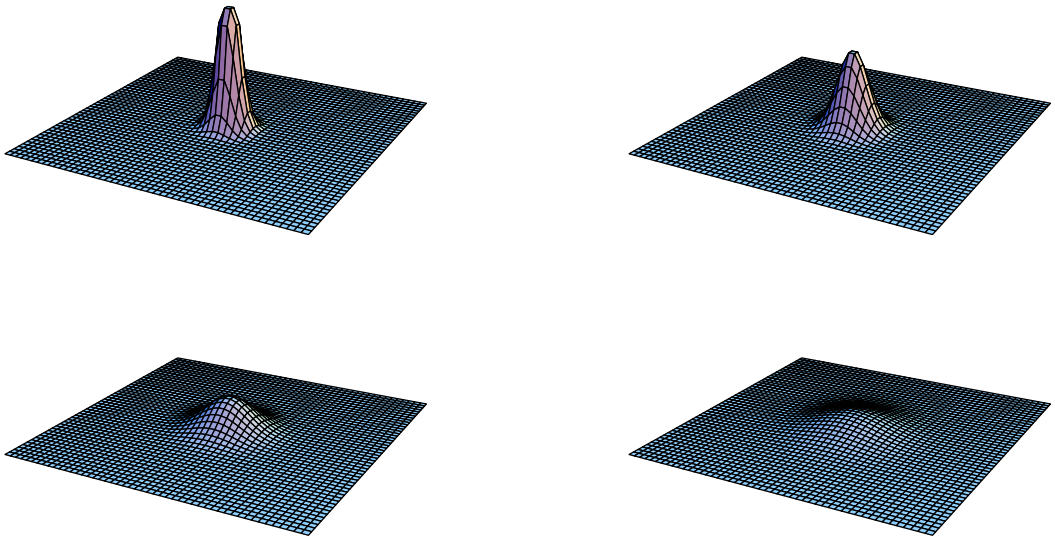


Figure 17.4. Fundamental Solution to the Planar Heat Equation.

medium, the fundamental solution decays in proportion to $1/\sqrt{t}$, whereas in two dimensions the decay is more rapid, being proportional to $1/t$. The physical explanation is that the energy is able to spread out in two independent directions, and hence diffuse away from its initial source faster. As we shall see, the decay in three-dimensional space is more rapid still, being proportional to $t^{-3/2}$ for similar reasons; see (18.97).

The principal purpose of the fundamental solution is to solve the general initial value problem. We express the initial temperature distribution as a superposition of delta function sources,

$$u(0,x,y) = f(x,y) = \iint f(\xi,\eta)\,\delta(x-\xi,y-\eta)\,d\xi\,d\eta,$$

where, at the point $(\xi,\eta) \in \mathbb{R}^2$, the source has magnitude $f(\xi,\eta)$. Linearity implies that the solution is then given by the same superposition of the associated fundamental solutions. Let us state this result as a theorem.

Theorem 17.4. *The solution to the initial value problem*

$$u_t = \gamma\,\Delta u, \qquad u(t,x,y) = f(x,y), \qquad (x,y) \in \mathbb{R}^2,$$

is given by the linear superposition formula

$$u(t,x,y) = \frac{1}{4\pi\gamma t} \iint f(\xi,\eta)\, e^{-[(x-\xi)^2+(y-\eta)^2]/4\gamma t}\,d\xi\,d\eta \tag{17.65}$$

We can interpret the solution formula (17.65) as a two-dimensional *convolution*

$$u(t,x,y) = F(t,x,y) * f(x,y) \tag{17.66}$$

of the initial data with a one-parameter family of progressively wider and shorter Gaussian filters; compare (13.113). As in the one-dimensional version, convolution with such an

integral kernel can be interpreted as a form of weighted averaging of the function, which has the effect of smoothing out and blurring the initial signal $f(x, y)$.

Example 17.5. If our initial temperature distribution is constant on a circular region, say

$$u(0, x, y) = \begin{cases} 1 & x^2 + y^2 < 1, \\ 0, & \text{otherwise}, \end{cases}$$

Then the solution can be evaluated using (17.65), as follows:

$$u(t, x, y) = \frac{1}{4\pi t} \iint_D e^{-[(x-\xi)^2+(y-\eta)^2]/4t} \, d\xi \, d\eta,$$

where the integral is over the unit disk $D = \{ \xi^2 + \eta^2 \leq 1 \}$. Let us evaluate the integral by going to polar coordinates:

$$u(t, x, y) = \frac{1}{4\pi t} \int_0^{2\pi} \int_0^1 e^{-[(x-\rho\cos\theta)^2+(y-\rho\sin\theta)^2]/4t} \, \rho \, d\rho \, d\theta.$$

Unfortunately, the final integral cannot be done in closed form in terms of elementary functions; see Exercise ■ for an expression in terms of complex Bessel functions. On the other hand, numerical evaluation of the integral is straightforward. A plot of the resulting radially symmetric solution appears in Figure h2disk■.

For more general configurations, where analytical solutions are no longer available, numerical solutions can be implemented based on a two-dimensional variant of the Crank–Nicholson scheme (14.154), relying on either finite differences or finite elements to discretize the space coordinates. We will not dwell on the details, but refer the interested reader to [**30**, **121**, **nPDE**].

17.5. The Planar Wave Equation.

The second important class of dynamical equations are those governing vibrational motions. The simplest planar system of this type is the two-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right), \tag{17.67}$$

which models the free (unforced) vibrations of a uniform two-dimensional membrane (a drum, say). Here $u(t, x, y)$ represents the displacement of the membrane at time t and position $(x, y) \in \Omega$, where $\Omega \subset \mathbb{R}^2$ is the domain representing the shape of the membrane. The constant $c^2 > 0$ encapsulates the physical properties of our membrane (density, tension, stiffness, thickness, etc.), with c being called, as in the one-dimensional version, the*wave speed*, since it turns out to be the speed at which localized signals propagate through the membrane. In this model, we are only allowing small, transverse (meaning vertical) displacements of the membrane. Large elastic vibrations lead to the nonlinear partial differential equations of elastodynamics, [**69**]. The bending vibrations of a flexible plate, which can be viewed as the two-dimensional version of a beam, are governed by a more complicated fourth order partial differential equation; see Exercise ■.

The solution $u(t, x, y)$ to the wave equation will be uniquely specified once we impose suitable boundary conditions and initial conditions. The Dirichlet conditions

$$u = h \qquad \text{on} \qquad \partial\Omega, \tag{17.68}$$

correspond to gluing our membrane to a fixed boundary or rim. On the other hand, the homogeneous Neumann conditions

$$\frac{\partial u}{\partial n} = 0 \qquad \text{on} \qquad \partial\Omega, \tag{17.69}$$

represent a free boundary where the membrane is not attached to any support. Mixed boundary conditions attach part of the boundary and leave the remaining portion free to vibrate:

$$u = h \qquad \text{on} \qquad D, \qquad\qquad \frac{\partial u}{\partial n} = 0 \qquad \text{on} \qquad N, \tag{17.70}$$

where $\partial\Omega = D \cup N$ with D and N non-overlapping. Since the wave equation is second order in time, we also need to impose two initial conditions:

$$u(0, x, y) = f(x, y), \qquad\qquad \frac{\partial u}{\partial t}\,(0, x, y) = g(x, y), \qquad\qquad (x, y) \in \Omega. \tag{17.71}$$

The first one prescribes the initial displacement of the membrane, while the second prescribes its initial velocity.

The wave equation is the simplest example of a general second order system of Newtonian form

$$\frac{\partial^2 u}{\partial t^2} = -\,K[\,u\,] = -\,\nabla^* \circ \nabla u. \tag{17.72}$$

As in (17.15), using general weighted inner products

$$\langle\, u\,;\widetilde{u}\,\rangle = \iint_\Omega u(x,y)\,\widetilde{u}(x,y)\,\rho(x,y)\,dx\,dy, \quad \langle\!\langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle\!\rangle = \iint_\Omega \mathbf{v}(x,y)\cdot\widetilde{\mathbf{v}}(x,y)\,\kappa(x,y)\,dx\,dy, \tag{17.73}$$

the adjoint to the gradient is a rescaled version of the divergence operator

$$\nabla^*\mathbf{v} = -\,\frac{1}{\rho}\,\nabla\cdot(\kappa\,\mathbf{v}).$$

Therefore, the general Newtonian system (17.72) takes the form

$$u_{tt} = -\,K[\,u\,] = \frac{1}{\rho}\,\nabla\cdot(\kappa\,\mathbf{v}),$$

or, in full detail,

$$\rho(x,y)\,\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(\,\kappa(x,y)\,\frac{\partial u}{\partial x}\,\right) + \frac{\partial}{\partial y}\left(\,\kappa(x,y)\,\frac{\partial u}{\partial y}\,\right). \tag{17.74}$$

This equation models the small transverse vibrations of a nonuniform membrane, in which $\rho(x, y) > 0$ represents the density of the membrane at the point $(x, y) \in \Omega$, while $\kappa(x, y) > 0$

represents its stiffness, in direct analogy with the one-dimensional version (14.82). In particular, if the material is homogeneous, then both ρ and κ are constant, and (17.74) reduces to the two-dimensional wave equation (17.67) with wave speed

$$c = \sqrt{\frac{\kappa}{\rho}}\,. \tag{17.75}$$

As in bars and strings, either increasing the stiffness, or decreasing the density, will cause the wave speed c to increase, and hence waves (signals) will propagate faster through the membrane.

Separation of Variables

Unfortunately, there is no explicit analytical technique comparable to the d'Alembert formula (14.121) for solving multi-dimensional wave equations. As a result, we are forced to fall back on our universal solution tool — separation of variables. Initially, the technique applies equally well to general vibration equations (17.72), and so we shall work in this context for the time being. The reader can, if desired, immediately specialize to the wave equation (17.67) itself, as explicit formulae will only be found in this case. We assume throughout that the boundary conditions — Dirichlet, Neumann, or mixed — are homogeneous; see Exercise ■ for a outline of how to handle inhomogeneous boundary conditions.

As in the one-dimensional analysis from Section 14.4, the separable solutions to the vibration equation (17.72) are found by using a trigonometric ansatz

$$u(t,x,y) = \cos\omega t\; v(x,y). \tag{17.76}$$

By linearity of K, which does not involve any t differentiation,

$$u_{tt} = -\,\omega^2\,\cos\omega t\; v(x,y), \qquad\qquad K[u] = \cos\omega t\; K[v].$$

Substituting into (17.72), and canceling out the cosine terms, we find that $v(x,y)$ must satisfy the by now familiar eigenvalue problem

$$K[v] = \omega^2\, v = \lambda\, v, \tag{17.77}$$

in which v is the eigenfunction whose eigenvalue $\lambda = \omega^2$ is equal to the square of the vibrational frequency ω. The eigenfunction $v(x,y)$ is always required to satisfy the relevant boundary conditions. Specializing to the wave equation (17.67), the eigenvalue problem (17.77) reduces to the same Helmholtz equation

$$c^2\,\Delta v + \lambda\, v = c^2\,(v_{xx} + v_{yy}) + \lambda\, v = 0 \tag{17.78}$$

that we analyzed earlier in this chapter.

As we learned, in the stable, positive definite cases — meaning either Dirichlet or mixed boundary conditions — the operator K admits an infinite sequence of positive eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \lambda_3 \le \cdots \qquad \text{with} \qquad \lambda_k \longrightarrow \infty \qquad \text{as} \qquad k \longrightarrow \infty.$$

Each eigenvalue and eigenfunction pair will produce two vibrating solutions

$$u_k(t,x,y) = \cos \omega_k t \; v_k(x,y)\,, \qquad \widetilde{u}_k(t,x,y) = \sin \omega_k t \; v_k(x,y)\,, \tag{17.79}$$

of frequency $\omega_k = \sqrt{\lambda_k}$. Note that the higher order modes vibrate faster, with progressively higher frequencies: $\omega_k \to \infty$ as $k \to \infty$.

The general solution to the initial value problem can be built up as a quasi-periodic linear combination

$$u(t,x,y) = \sum_{k=1}^{\infty} a_k \, u_k(t,x,y) + b_k \, \widetilde{u}_k(t,x,y) = \sum_{k=1}^{\infty} \big(a_k \cos \omega_k t + b_k \sin \omega_k t \big) \, v_k(x,y) \tag{17.80}$$

of the fundamental vibrational modes, in the form of an *eigenfunction series*. The eigenfunction coefficients a_k, b_k are prescribed by the initial conditions. Thus, evaluating the solution series (17.80) and its time derivative[†] at the initial time $t = 0$, we find

$$\sum_{k=1}^{\infty} a_k \, v_k(x,y) = f(x,y), \qquad \sum_{k=1}^{\infty} \omega_k \, b_k \, v_k(x,y) = g(x,y). \tag{17.81}$$

We then appeal to the orthogonality of the eigenfunctions to compute the coefficients

$$a_k = \frac{\langle \, f \, ; v_k \, \rangle}{\| \, v_k \, \|^2} = \frac{\displaystyle\iint_\Omega f \, v_k \, \rho \; dx \, dy}{\displaystyle\iint_\Omega v_k^2 \, \rho \; dx \, dy}\,, \qquad b_k = \frac{1}{\omega_k} \, \frac{\langle \, g \, ; v_k \, \rangle}{\| \, v_k \, \|^2} = \frac{\displaystyle\iint_\Omega g \, v_k \, \rho \; dx \, dy}{\omega_k \displaystyle\iint_\Omega v_k^2 \, \rho \; dx \, dy}\,, \tag{17.82}$$

via our usual formulae, (17.25). In the case of the wave equation, the density ρ is constant, and hence can be canceled from the numerator and denominator of the orthogonality formulae (17.81). As long as the initial data is reasonably well-behaved, Theorem 17.1 will justify the convergence of the resulting series solution.

The unstable, semi-definite case of pure Neumann boundary conditions, models a physical membrane that has not been attached anywhere along its boundary, and so is free to move off in a vertical direction. Here, the constant solution $v_0(x,y) \equiv 1$ is a null eigenfunction, corresponding to the zero eigenvalue $\lambda_0 = 0$. In general, each null eigenfunction provides two solutions to the vibration equation (17.72), which in the present situation are the two elementary solutions

$$u_0(t,x,y) = 1, \qquad \widetilde{u}_0(t,x,y) = t.$$

The first solution represents a membrane that has been displaced by a fixed amount in the vertical direction, while the second represents a membrane that is uniformly moving in the vertical direction with speed 1. (Think of the membrane moving in outer space unaffected

[†] We are assuming that the series converges sufficiently rapidly in order to be allowed to differentiate term by term.

by any external gravitational force.) The general solution to the vibration equation then has the series form

$$u(t,x,y) = a_0 + b_0\,t + \sum_{k=1}^{\infty} \big(a_k \cos \omega_k t + b_k \sin \omega_k t \big)\, v_k(x,y). \tag{17.83}$$

The coefficients a_k, b_k for $k > 0$ are given by the same orthogonality formulae (17.82). The only unstable, nonperiodic mode is the linearly growing term component $b_0\,t$ in (17.83). Its coefficient

$$b_0 = \frac{\langle\, g\,;1\,\rangle}{\|\,1\,\|^2} = \frac{\displaystyle\iint_\Omega g\,\rho\;dx\,dy}{\displaystyle\iint_\Omega \rho\;dx\,dy}\,,$$

is a weighted average of the initial velocity $g(x,y) = u_t(0,x,y)$ over the domain. In the case of the wave equation, the density ρ is constant, and hence

$$b_0 = \frac{1}{\text{area }\Omega}\iint_\Omega g(x,y)\,dx\,dy$$

equals the average initial velocity. If the (weighted) average initial velocity $b_0 \neq 0$ is nonzero, then the membrane will move off at an average vertical speed b_0 — while continuing to vibrate in any of the vibrational modes that have been excited by the initial displacement and/or initial velocity. Again, this is merely a two-dimensional translation of our observations of a free, vibrating bar — which in turn was the continuum version of an unsupported structure.

Remark: an interesting question is whether two differently shaped drums can have identical vibrational frequencies. Or, to state the problem in another way, can one reconstruct the shape of a drum by listening to its vibrations? The answer turns out to be "no", but for quite subtle reasons. See [**drum**] for a discussion.

17.6. Analytical Solutions of the Wave Equation.

So far, we have looked at some of the general, qualitative features of the two-dimensional vibration and wave equations. Actual analytical solutions are, of course, harder to come by, and can only be found in very special geometrical configurations. In this section, we discuss the two most important special cases — rectangular and circular membranes.

Remark: Most realistic vibration problems need to be solved numerically, by adaptations of the integration schemes outlined in Section 14.6. The spatial discretization is implemented using either finite differences, or a version of finite elements. We refer the reader to [**nPDE**] for details.

Vibration of a Rectangular Drum

Let us first consider the vibrations of a membrane in the shape of a rectangle

$$R = \big\{\, 0 < x < a,\; 0 < y < b \,\big\}$$

with side lengths a and b, whose sides are fixed to the (x,y)–plane. Thus, we seek to solve the wave equation

$$u_{tt} = c^2 \Delta u = c^2 (u_{xx} + u_{yy}), \qquad 0 < x < a, \qquad 0 < y < b, \tag{17.84}$$

subject to the initial and boundary conditions

$$\begin{aligned} u(t,0,y) = v(t,a,y) = 0 = v(t,x,0) = v(t,x,b), &\qquad 0 < x < a, \\ u(0,x,y) = f(x,y), \qquad u_t(0,x,y) = g(x,y), &\qquad 0 < y < b. \end{aligned} \tag{17.85}$$

As we saw in Section 17.3

$$c^2 (v_{xx} + v_{yy}) + \lambda\, v = 0, \qquad (x,y) \in R, \tag{17.86}$$

on a rectangle, subject to the homogeneous Dirichlet boundary conditions

$$v(0,y) = v(a,y) = 0 = v(x,0) = v(x,b), \qquad 0 < x < a, \qquad 0 < y < b, \tag{17.87}$$

are

$$v_{m,n}(x,y) = \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}\,, \qquad \text{where} \qquad \lambda_{m,n} = \pi^2\, c^2 \left(\frac{m^2}{a^2} + \frac{n^2}{b^2} \right), \tag{17.88}$$

with $m, n = 1, 2, \ldots$. The fundamental frequencies of vibration are the square roots of the eigenvalues, so

$$\omega_{m,n} = \sqrt{\lambda_{m,n}} = \pi\, c \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}\,. \tag{17.89}$$

The frequencies will depend upon the underlying geometry — meaning the side lengths — of the rectangle, as well as the wave speed, which is turn is a function of the density and stiffness of the membrane, (17.75). The higher the wave speed c, or the smaller the rectangle, the faster the vibrations. In layman's terms (17.90) quantifies the observation that smaller, stiffer drums of less dense material vibrate faster.

According to (17.79), the normal modes of vibration of our rectangle are

$$\begin{aligned} u_{m,n}(t,x,y) &= \cos \omega_{m,n}\, t\ \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}, \\ \widetilde{u}_{m,n}(t,x,y) &= \sin \omega_{m,n}\, t\ \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}. \end{aligned} \tag{17.90}$$

The general solution can be written as a double Fourier series

$$u(t,x,y) = \sum_{m,n=1}^{\infty} \big[\, a_{m,n}\, u_{m,n}(t,x,y) + b_{m,n}\, \widetilde{u}_{m,n}(t,x,y) \,\big].$$

in the normal modes. The coefficients $a_{m,n}, b_{m,n}$ are fixed by the initial displacement $u(0,x,y) = f(x,y)$ and the initial velocity $u_t(0,x,y) = g(x,y)$ as in (17.81). The orthogonality relations among the eigenfunctions imply

$$a_{m,n} = \frac{\langle\, v_{m,n}\,;f\,\rangle}{\|\, v_{m,n}\,\|^2} = \frac{1}{a\,b} \int_0^b \int_0^a f(x,y)\ \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}\, dx\, dy.$$

$$b_{m,n} = \frac{\langle\, v_{m,n}\,;g\,\rangle}{\omega_{m,n}\,\|\, v_{m,n}\,\|^2} = \frac{1}{\pi\, c \sqrt{m^2\, b^2 + n^2\, a^2}} \int_0^b \int_0^a g(x,y)\ \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}\, dx\, dy.$$

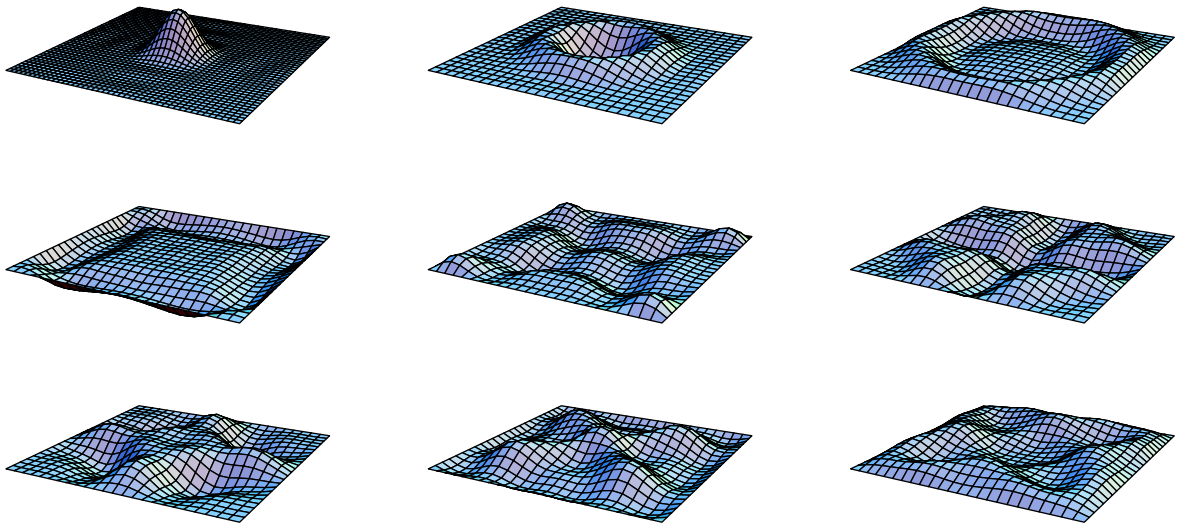


Figure 17.5. Vibrations of a Square.

Since the fundamental frequencies are not rational multiples of each other, the general solution is a genuinely quasi-periodic superposition of the various normal modes.

In Figure 17.5 we plot the solution for the initial concentrated displacement

$$u(0,x,y) = f(x,y) = e^{-100[(x-.5)^2+(y-.5)^2]}$$

starting at the center of a unit square, so $a = b = 1$. The plots are at successive times $0, .02, .04, \dots, 1.6$. Note that, unlike a one-dimensional string where a concentrated displacement remains concentrated at all subsequent times and periodically repeats, the initial displacement spreads out in a radially symmetric manner and propagates to the edges of the rectangle, where it reflects and then interacts with itself. However, owing to the quasiperiodicity of the solution, the displacement of the drum never exactly repeats itself, and the initial concentrated signal never quite reforms in the center.

Vibration of a Circular Drum

Let us next analyze the vibrations of a circular membrane with fixed Dirichlet boundary conditions. As always, we build up the solution as a quasi-periodic linear combination of the normal modes, which, by (17.79), are fixed by the eigenfunctions for the associated Helmholtz boundary value problem.

As we saw in Section 17.3, the eigenfunctions of the Helmholtz equation on a disk of radius 1, say, subject to homogeneous Dirichlet boundary conditions, are products of trigonometric and Bessel functions:

$$\begin{aligned} v_{m,n}(r,\theta) &= \cos m\theta \; J_m(\zeta_{m,n}\, r), & & m = 0,1,2,\dots, \\ \widetilde{v}_{m,n}(r,\theta) &= \sin m\theta \; J_m(\zeta_{m,n}\, r), & & n = 1,2,3,\dots . \end{aligned} \tag{17.91}$$

Here r, θ are the usual polar coordinates, while $\zeta_{m,n}$ denotes the n^{th} root of the m^{th} order Bessel function $J_m(z)$, cf. (17.54). The corresponding eigenvalue is its square, $\lambda_{m,n} = \zeta_{m,n}^2$, and hence the natural frequencies of vibration are the product of the Bessel roots times the wave speed:

$$\omega_{m,n} = c\,\sqrt{\lambda_{m,n}} \;= c\,\zeta_{m,n}. \tag{17.92}$$

A table of their values (for the case $c = 1$) can be found in the preceding section. The Bessel roots do not follow any easily discernible order or pattern, and are certainly not rational multiples of each other. Thus, the vibrations of a circular drum are also truly quasi-periodic.

The frequencies $\omega_{0,n} = c\,\zeta_{0,n}$ correspond to simple eigenvalues, with a single radially symmetric eigenfunction $J_0(\zeta_{0,n}\,r)$, while the "angular modes" $\omega_{m,n}$ with $m > 0$ are double, each possessing two linearly independent eigenfunctions (17.91). According to the general formula (17.79), each eigenfunction leads to two independent normal modes of vibration, having the explicit form

$$u_{m,n}(t,r,\theta) = \begin{Bmatrix} \cos \\ \sin \end{Bmatrix} c\,\zeta_{m,n}\,t \begin{Bmatrix} \cos \\ \sin \end{Bmatrix} m\,\theta\; J_m(\zeta_{m,n}\,r). \tag{17.93}$$

One can use either the cosine or the sine in each slot, and so the formula gives a total of four distinct normal modes associated with each Bessel function — unless $m = 0$, in which case the solutions are radially symmetric, and there are only two normal modes for each eigenvalue. The general solution is written as a series in these normal modes in the form of a Fourier–Bessel series

$$\begin{aligned} u_{m,n}(t,r,\theta) = \sum_{m,n} \Big[&\big(\, a_{m,n} \cos c\,\zeta_{m,n}\,t + b_{m,n} \sin c\,\zeta_{m,n}\,t \,\big) \cos m\,\theta \\ &+ \big(\, c_{m,n} \cos c\,\zeta_{m,n}\,t + d_{m,n} \sin c\,\zeta_{m,n}\,t \,\big) \sin m\,\theta \Big]\, J_m(\zeta_{m,n}\,r). \end{aligned} \tag{17.94}$$

The coefficients $a_{m,n}, b_{m,n}, c_{m,n}, d_{m,n}$ are determined, as usual, by the initial displacement and velocity of the membrane. In Figure vdisk■, the vibrations due to an initially concentrated displacement are displayed. Again, the motion is only quasi-periodic and never quite returns to the original configuration.

Remark: As we learned in Section 14.4, the natural frequencies of vibration a one-dimensional medium, e.g., a violin string or a column of air in a flute, are integer multiples of each other. As a consequence, the vibrations are periodic. Musically, this means that the overtones are integer multiples of the fundamental tones, and, as a result the music sounds harmonic to our ear. On the other hand, the natural frequencies of circular and rectangular drums are irrationally related, and the vibrations are only quasi-periodic. As a result, we hear a percussive sound! Thus, for some reason, our musical appreciation is psychologically attuned to the differences between rationally related/periodic and irrationally related/quasi-periodic vibrations.

Scaling and Symmetry

Both translation and scaling symmetries can be effectively employed in the analysis

of the wave equation. Let us consider the simultaneous rescaling

$$t \mapsto \alpha\, t, \qquad x \mapsto \beta\, x, \qquad y \mapsto \beta\, y, \tag{17.95}$$

of time and space, whose effect is to change the function $u(t, x, y)$ into a rescaled function

$$U(t, x, y) = u(\alpha\, t, \beta\, x, \beta\, y). \tag{17.96}$$

The chain rule relates their derivatives:

$$\frac{\partial^2 U}{\partial t^2} = \alpha^2\, \frac{\partial^2 u}{\partial t^2}\,, \qquad \frac{\partial^2 U}{\partial x^2} = \beta^2\, \frac{\partial^2 u}{\partial x^2}\,, \qquad \frac{\partial^2 U}{\partial y^2} = \beta^2\, \frac{\partial^2 u}{\partial y^2}\,.$$

Therefore, if u satisfies the wave equation

$$u_{tt} = c^2\, \Delta u,$$

then U satisfies the rescaled wave equation

$$U_{tt} = \frac{\alpha^2\, c^2}{\beta^2}\, \Delta U = \widetilde{c}^{\,2}\, \Delta U, \quad \text{where the rescaled wave speed is} \quad \widetilde{c} = \frac{\alpha\, c}{\beta}\,. \tag{17.97}$$

In particular, rescaling time by setting $\alpha = 1/c$ results in a unit wave speed $\widetilde{c} = 1$. In other words, for a given homogeneous medium, we can choose our time unit of measurement to arrange that the wave speed is equal to 1.

If we set $\alpha = \beta$, scaling space and time in the same proportion, then the wave speed does not change, $\widetilde{c} = c$, and so

$$t \mapsto \beta\, t, \qquad x \mapsto \beta\, x, \qquad y \mapsto \beta\, y, \tag{17.98}$$

defines a symmetry transformation for the wave equation. If $u(t, x, y)$ is any solution to the wave equation, then so is its rescaled version

$$U(t, x, y) = u(\beta\, t, \beta\, x, \beta\, y) \tag{17.99}$$

for any choice of (nonzero) scale parameter β. In other words, if $u(t, x, y)$ is defined on a domain Ω, then the rescaled solution $U(t, x, y)$ will be defined on the rescaled domain

$$\widetilde{\Omega} = \frac{1}{\beta}\, \Omega = \left\{ (x, y) \;\middle|\; \left(\frac{x}{\beta}\,, \frac{y}{\beta} \right) \in \Omega \right\}. \tag{17.100}$$

For example, if $\beta = \frac{1}{2}$, then the effect is to double the size of the domain. The fundamental modes for the rescaled domain have the form

$$U_n(t, x, y) = u_n(\beta\, t, \beta\, x, \beta\, y) = \cos\, \omega_n\, \beta\, t\; v_n(\beta\, x, \beta\, y),$$
$$\widetilde{U}_n(t, x, y) = \widetilde{u}_n(\beta\, t, \beta\, x, \beta\, y) = \sin\, \omega_n\, \beta\, t\; v_n(\beta\, x, \beta\, y),$$

and hence the fundamental vibrational frequencies $\widetilde{\omega}_n = \beta\, \omega_n$ are scaled by the same overall factor. Thus, when $\beta < 1$, the rescaled membrane is larger and its vibrations are slowed down by the same factor. For instance, a drum that is twice as large will vibrate twice as slowly, and hence have an octave lower overall tone. Musically, this means that all

drums of a similar shape have the same pattern of overtones, differing only in their overall pitch, which is a function of their size, tautness and density.

For example, choosing $\beta = 1/R$ will rescale the unit disk into a disk of radius R. The fundamental frequencies of the rescaled disk are

$$\widetilde{\omega}_{m,n} = \beta\, \omega_{m,n} = \frac{c}{R}\, \zeta_{m,n}, \tag{17.101}$$

where c is the wave speed and $\zeta_{m,n}$ are the Bessel roots, defined in (17.54). Consequently, the ratios $\omega_{m,n}/\omega_{m',n'}$ between vibrational frequencies are the same, independent of the size of the disk R and the wave speed c. We define the *relative vibrational frequencies*

$$\rho_{m,n} = \frac{\omega_{m,n}}{\omega_{0,1}} = \frac{\zeta_{m,n}}{\zeta_{0,1}}, \qquad \text{where} \qquad \omega_{0,1} = \frac{c\,\zeta_{0,1}}{R} \approx 2.4\, \frac{c}{R} \tag{17.102}$$

is the dominant, lowest frequency. The relative frequencies $\rho_{m,n}$ are independent of the size, stiffness or composition of the drum membrane. In the following table, we display all relative vibrational frequencies (17.102) that are < 6. As usual the columns are indexed by m and the rows by n. Once the lowest frequency $\omega_{0,1}$ has been determined — either theoretically, numerically or experimentally — all the higher overtones $\omega_{m,n} = \rho_{m,n}\, \omega_{0,1}$ are obtained by multiplication by these fixed relative frequencies of vibration.

Relative Vibrational Frequencies of a Circular Disk

	0	1	2	3	4	5	6	7	8	9	...
1	1.000	1.593	2.136	2.653	3.155	3.647	4.132	4.610	5.084	5.553	...
2	2.295	2.917	3.500	4.059	4.601	5.131	5.651	⋮	⋮	⋮	
3	3.598	4.230	4.832	5.412	5.977	⋮	⋮				
4	4.903	5.540	⋮	⋮	⋮						
⋮	⋮	⋮									

17.7. Nodal Curves.

When a membrane vibrates, the individual points move up and down in a quasi-periodic manner. As such, correlations between the motions at different points are not immediately evident. However, if the system is set to vibrate in a pure eigenmode, say

$$u_n(t,x,y) = \cos(\omega_n\, t)\, v_n(x,y),$$

then all points on the membrane move up and down at a common frequency $\omega_n = \sqrt{\lambda_n}$, which is the square root of the eigenvalue corresponding to the eigenfunction $v_n(x,y)$. The

exceptions are the points where the eigenfunction vanishes:

$$v_n(x, y) = 0. \tag{17.103}$$

Such points will not move at all. The set of all points $(x, y) \in \Omega$ that satisfy (17.103) is known as the n^{th} *nodal set* of the domain Ω. If we scatter small particles (e.g., sand or powder) over the membrane while it is performing a pure vibration, we can actually see the nodal set because the particles will, though random movement over the oscillating regions of the membrane, tend to accumulate along the stationary nodal curves.

It can be shown that, in general, the nodal set consists of a system of intersecting curves, known as the *nodal curves* of the membrane. The nodal curves partition the membrane into *nodal regions*, and intersect at critical points, $\nabla v_n = \mathbf{0}$, of the eigenfunction. Points lying in a common nodal region all vibrate in tandem, so that all the nodal region is either up or down, except, momentarily, when the *entire* membrane has zero displacement. The latter situation occurs at regular time intervals, namely whenever $\cos \omega_n t = 0$. Adjacent nodal regions, lying on the opposite sides of a nodal curve, always vibrate in opposite directions — when one side is up, the other is down, and then, as the membrane becomes momentarily flat, they simultaneously switch direction.

Example 17.6. *Circular Drums.* Since the eigenfunctions (17.91) for a disk are products of trigonometric functions in the angular variable and Bessel functions of the radius, the nodal curves for the normal modes of vibrations of a circular membrane are rays emanating from and circles centered at the origin. Thus, the nodal regions of vibration are annular sectors. Pictures of the nodal curves for the first nine fundamental modes indexed by their relative frequencies are plotted in Figure 17.6. Figure 17.2 shows a sample displacement of the membrane in each of the first twelve modes. The dominant (lowest frequency) mode is the only one that has no nodal curves; it has the form of a radially symmetric bump where the entire membrane flexes up and down. Every other mode has at least one nodal curve. For instance, the next lowest modes have frequency $\omega_{1,1}$, and are linear combinations $\alpha\, u_{1,1} + \beta\, \widetilde{u}_{1,1}$ of the two eigenfunctions. Each combination has a single diameter as a nodal curve, whose orientation depends upon the coefficients α, β. The two halves of the drum vibrate up and down in opposing directions — when the top half is up, the bottom half is down and vice versa. The next set of modes have two perpendicular diameters as nodal curves, and the four quadrants of the drum vibrate up and down in tandem, with adjacent quadrants having opposing displacements. Next is a single mode, with a circular nodal curve whose (relative) radius $\zeta_{0,2}/\zeta_{0,1} \approx 0.6276$ is the ratio of the first two roots of the order zero Bessel function; see Exercise ■ for a justification. In this case, the inner disk and outer annulus vibrate in opposing directions.

Example 17.7. *Rectangular Drums.* For a general rectangular drum, the nodal curves are relatively uninteresting. Since the normal modes (17.90) are separable products of trigonometric functions in the coordinate variables x, y, the nodal curves are regularly equi-spaced straight lines parallel to the sides of the rectangle. The internodal regions are small rectangles, all of the same size and shape, with adjacent rectangles vibrating in opposite directions.

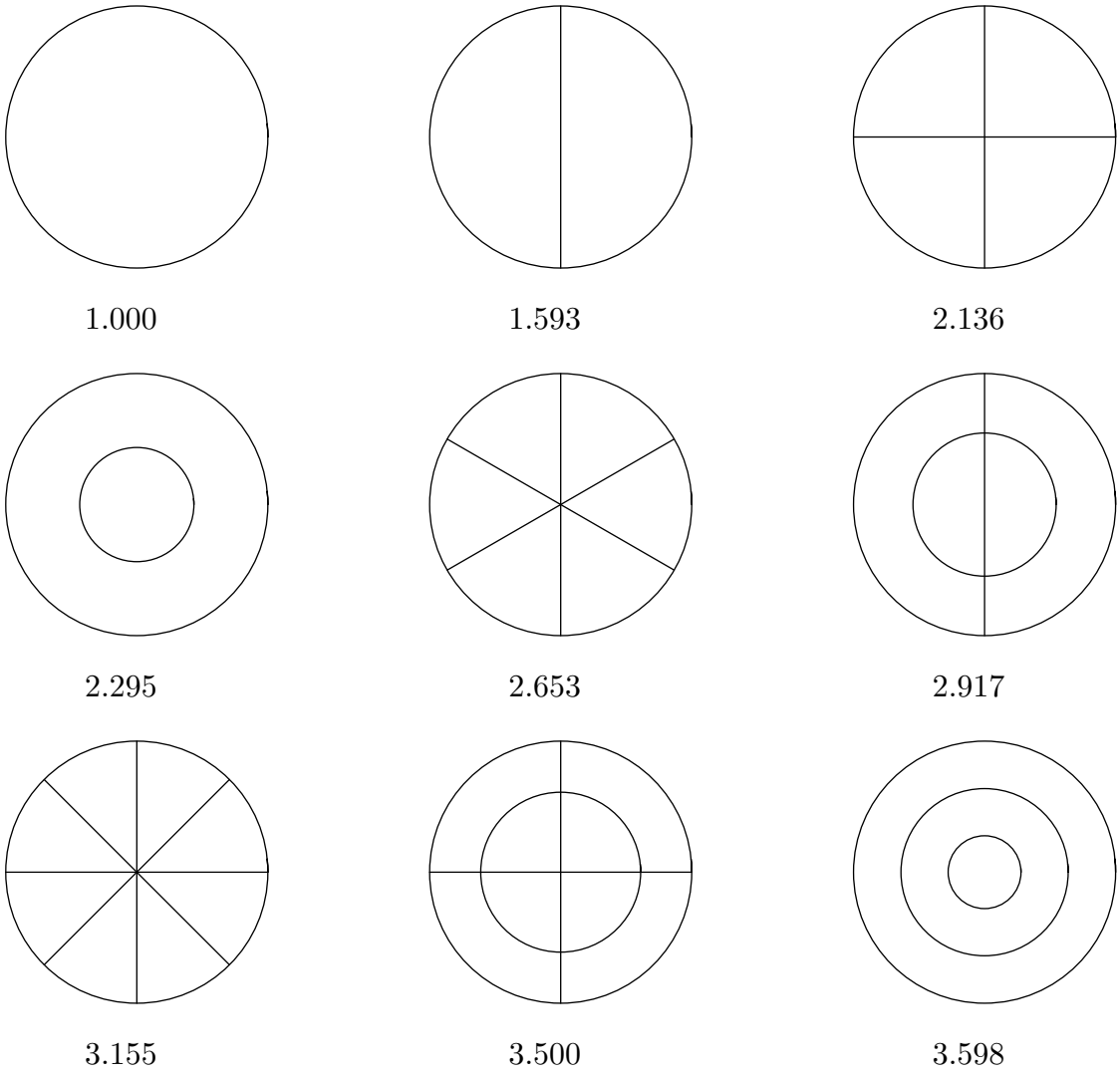


Figure 17.6. Nodal Curves and Relative Frequencies of Vibration of a Circular Membrane.

A more interesting collection of nodal curves occurs when the rectangle admits multiple eigenvalues — so-called "accidental degeneracies". If two of the eigenvalues (17.88) are equal, $\lambda_{m,n} = \lambda_{k,l}$, which occurs when

$$\frac{m^2}{a^2} + \frac{n^2}{b^2} = \frac{k^2}{a^2} + \frac{l^2}{b^2} \tag{17.104}$$

where $(m, n) \neq (k, l)$ are two distinct pairs of positive integers, then both eigenmodes vibrate with a common frequency

$$\omega = \omega_{m,n} = \omega_{k,l}.$$

As a consequence, any linear combination of the eigenmodes

$$\cos \omega\, t \left(\alpha \sin \frac{m\,\pi\, x}{a} \sin \frac{n\,\pi\, y}{b} + \beta \sin \frac{k\,\pi\, x}{a} \sin \frac{l\,\pi\, y}{b} \right), \qquad \alpha, \beta \in \mathbb{R},$$

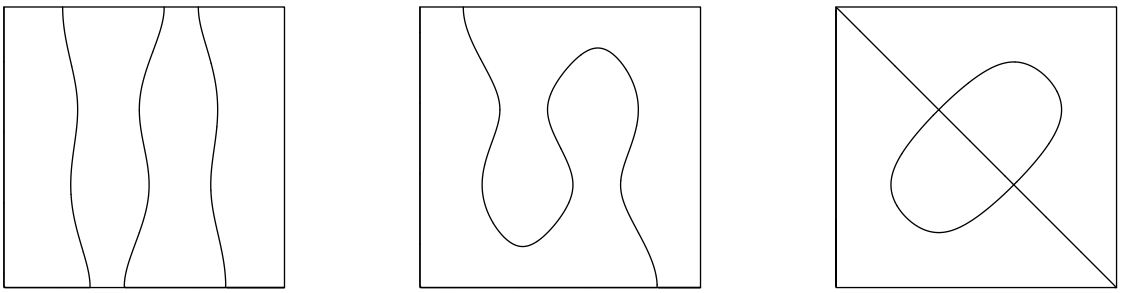


Figure 17.7. Some Nodal Curves for a Square Membrane.

will also qualify as a normal mode of vibration. The corresponding nodal curves

$$\alpha \sin \frac{m\,\pi\,x}{a} \sin \frac{n\,\pi\,y}{b} + \beta \sin \frac{k\,\pi\,x}{a} \sin \frac{l\,\pi\,y}{b} = 0 \tag{17.105}$$

have a more intriguing geometry. Their configurations can change dramatically as the relative magnitudes of α, β vary.

For example, if $R = \{\, 0 < x < 1,\ 0 < y < 1 \,\}$ is a unit square, then an accidental degeneracy, satisfying (17.104), occurs whenever

$$m^2 + n^2 = k^2 + l^2. \tag{17.106}$$

Thus, two distinct ordered pairs of positive integers (m, n) and (k, l) must have the same norm. The simplest possibility occurs whenever $m \neq n$, in which case we merely reverse the order, setting $k = n$, $l = m$. In Figure 17.7 we illustrate the nodal curves

$$\sin 4\,\pi\,x \sin \pi\,y + \beta \sin \pi\,x \sin 4\,\pi\,y = 0, \qquad \beta = .2, .5, 1,$$

corresponding to the three different linear combinations of the eigenfunctions with $m = l = 4$, $n = k = 1$. The associated vibrational frequency is $\omega_{4,1} = \pi\,c\,\sqrt{17}\,$.

Remark: Classifying the rectangles that admit such accidental degeneracies takes us into the realm of number theory, [**NumTh**]. The basic issue is to classify numbers can be written as a sum of two squared integers in more than one way, as in (17.106). Or, stated another way, find all integer points that lie on a common circle.

Chapter 18

Partial Differential Equations in Space

At last we have ascended the dimensional ladder to its ultimate rung (at least for those of us living in a three-dimensional universe): partial differential equations in physical space. Fortunately, almost everything of importance has already appeared in the one- and two-dimensional situations, and appending a third dimension is, for the most part, simply a matter of appropriately adapting the same basic constructions. Thus, separation of variables, Green's functions and fundamental solutions continue to be the weapons of choice. Unfortunately, despite the best efforts of mathematicians, the most powerful of our planar tools, conformal mapping, does *not* carry over to higher dimensions. The crux of the problem is the relative lack of conformal maps.

As before, the three primary examples are the three-dimensional Laplace equation, modeling equilibrium configurations of solid bodies, the three-dimensional heat equation, which models basic spatial diffusion processes, and the three-dimensional wave equation, governing small vibrations of solid bodies. Of course, the dimensional ladder continues to stretch onwards to general four-, five-, and even n-dimensional counterparts of these basic linear systems. However, almost all important analytical and numerical techniques already appear by the time we reach three-dimensional space, and such extensions are of interest primarily to pure mathematicians and, possibly, modern theoretical physicists.

The basic underlying solution techniques — separation of variables and Green's functions or fundamental solutions — have already appeared. In three-dimensional problems, separation of variables can be used in rectangular, cylindrical and spherical coordinates. The first two do not produce anything fundamentally new, and are therefore left to the exercises. The most important case is in spherical coordinates, and here we find new special functions known as spherical harmonics and spherical Bessel functions. These functions play important roles in a number of physical systems, including the quantum theory of atomic structure that underlies the spectroscopic and reactive properties of atoms, and hence the periodic table and, in a sense, all of modern chemistry.

The fundamental solution for the three-dimensional heat equation can be easily guessed from its one- and two-dimensional versions. The three-dimensional wave equation, surprisingly, has an explicit solution formula of d'Alembert form, albeit quite a bit more complicated. Indeed, attempts to derive such a formula for the two-dimensional version were unsuccessful, and only through a method of descent starting with the three-dimensional solution are we able to arrive at the solution to the two-dimensional wave equation. This also points out a critical difference between waves in two- and three-dimensional media. Huygens' principle states that three-dimensional waves due to a localized initial disturbance remain localized as they propagate in space; this is not true in two dimensions,

and a concentrated planar disturbance leads to a residual disturbance that never entirely disappears!

18.1. The Laplace and Poisson Equations.

We begin, as always, with systems in equilibrium. The most fundamental system is the three-dimensional Laplace equation

$$\Delta u = u_{xx} + u_{yy} + u_{zz} = 0, \tag{18.1}$$

in which $\mathbf{x} = (x, y, z)^T$ represent Cartesian coordinates in $\mathbb{R}^3$. The solutions to the Laplace equation continue to be known as *harmonic functions*. The Laplace equation models unforced equilibria; *Poisson's equation* is the inhomogeneous version

$$-\Delta u = f(x, y, z), \tag{18.2}$$

where the inhomogeneity f represents some form of external forcing.

The basic boundary value problem for the Laplace or the Poisson equation seeks a solution inside a bounded domain $\Omega \subset \mathbb{R}^3$ subject to either Dirichlet boundary conditions, prescribing the function values

$$u = h \qquad \text{on} \qquad \partial\Omega, \tag{18.3}$$

or Neumann boundary conditions prescribing its normal derivative

$$\frac{\partial u}{\partial \mathbf{n}} = k \qquad \text{on} \qquad \partial\Omega, \tag{18.4}$$

or mixed boundary conditions in which one imposes Dirichlet conditions on part of the boundary and Neumann conditions on the remainder. Keep in mind that the boundary of the domain Ω consists of one or more surfaces, which will be oriented using unit normal $\mathbf{n}$ pointing outwards, away from the domain.

The boundary value problems for the three-dimensional Laplace and Poisson equations govern a wide variety of equilibrium situations in physics. Among the areas of application, we mention:

(*a*) *Ideal fluid flow*: Here u represents the velocity potential for an incompressible, irrotational steady state fluid flow in a container, with velocity vector field $\mathbf{v} = \nabla u$. Homogeneous Neumann boundary conditions correspond to a solid boundary which fluid cannot penetrate.

(*b*) *Heat conduction*: Here u represents the temperature in a solid body. Dirichlet conditions correspond to fixing the temperature on the bounding surface(s), whereas homogeneous Neumann conditions correspond to an insulated boundary, i.e., one which does not allow any heat flux. The inhomogeneity f represents an internal heat source.

(*c*) *Elasticity*: In certain restricted situations, u represents an equilibrium deformation of a solid body, e.g., a radial deformation of a ball. Fully three-dimensional elasticity is governed by a system of partial differential equations; see Example 21.8.

(*d*) *Electrostatics*: Here u represents the electromagnetic potential in a conducting medium.

(*e*) *Gravitation*: The Newtonian gravitational potential in flat empty space is also prescribed by the Laplace equation. General relativity is a vastly more complicated system, leading to systems of nonlinear partial differential equations.

Self–Adjoint Formulation and Minimum Principles

The Laplace and Poisson equations naturally fit into our self-adjoint equilibrium framework. The construction is a straightforward adaptation of the planar version of Section 15.4. We introduce the L^2 inner products

$$\begin{aligned} \langle\, u\,;\widetilde{u}\,\rangle &= \iiint_\Omega u(x,y,z)\,\widetilde{u}(x,y,z)\,dx\,dy\,dz, \\ \langle\, \mathbf{v}\,;\widetilde{\mathbf{v}}\,\rangle &= \iiint_\Omega \mathbf{v}(x,y,z)\cdot\widetilde{\mathbf{v}}(x,y,z)\,dx\,dy\,dz, \end{aligned} \tag{18.5}$$

between scalar fields $u, \widetilde{u}$, and between vector fields $\mathbf{v}, \widetilde{\mathbf{v}}$ defined on a domain $\Omega \subset \mathbb{R}^3$. We assume that the functions in question are sufficiently nice that these inner products are well-defined; if Ω is unbounded, this requires that they decay to zero reasonably rapidly at large distances. When subject to homogeneous boundary conditions of the proper form, the adjoint of the gradient operator with respect to the L^2 inner products is minus the divergence:

$$\nabla^* = -\,\nabla\cdot\,. \tag{18.6}$$

As we have learned, the computation of the adjoint relies on an integration by parts formula. In the plane, Green's Formula (A.55) provided the basic tool. For partial differential equations in three-dimensional space, we rely on the Divergence Theorem B.36. The first step is to establish the three-dimensional analog of Green's Formula (15.78). To this end, we apply the divergence equation (B.82) to the product $u\,\mathbf{v}$ of a scalar field u and a vector field $\mathbf{v}$, leading to the identity

$$\iiint_\Omega \big(\, u\,\nabla\cdot\mathbf{v} + \nabla u\cdot\mathbf{v}\,\big)\,dx\,dy\,dz = \iiint_\Omega \nabla\cdot(u\,\mathbf{v})\,dx\,dy\,dz = \iint_{\partial\Omega} u\,(\mathbf{v}\cdot\mathbf{n})\,dS. \tag{18.7}$$

Rearranging the terms in this formula produces an integration by parts formula for volume integrals:

$$\iiint_\Omega (\nabla u\cdot\mathbf{v})\,dx\,dy\,dz = \iint_{\partial\Omega} u\,(\mathbf{v}\cdot\mathbf{n})\,dS - \iiint_\Omega u\,(\nabla\cdot\mathbf{v})\,dx\,dy\,dz. \tag{18.8}$$

Note that the gradient operator on the scalar field u has moved to become a divergence operator on the vector field $\mathbf{v}$. The boundary integral will vanish provided either

(*a*) u vanishes on Ω — these are homogeneous Dirichlet boundary conditions, or

(*b*) $\mathbf{v}\cdot\mathbf{n} = \mathbf{0}$ on $\partial\Omega$ — which leads to homogeneous Neumann boundary conditions $\nabla u = \mathbf{0}$ on $\partial\Omega$, since the gradient operator must map u to a vector field $\mathbf{v} = \nabla u$ whose normal component $\mathbf{v}\cdot\mathbf{n} == \partial u/\partial\mathbf{n}$ equals the normal derivative of u, or

(*c*) $\partial\Omega = D \cup N$ decomposes into two non-overlapping parts, and we impose Dirichlet conditions $u = 0$ on D and Neumann conditions $\mathbf{v} = \mathbf{0}$ on the remaining part N, leading to the usual mixed boundary conditions.

Thus, subject to the homogeneous boundary conditions, the integration by parts formula (18.8) takes the form

$$\langle \nabla u \,;\, \mathbf{v} \,\rangle = \langle \, u \,;\, -\nabla \cdot \mathbf{v} \,\rangle, \tag{18.9}$$

which proves (18.6). Consequently, the Laplace equation takes our well-known self-adjoint form

$$-\Delta u = -\nabla \cdot \nabla u = \nabla^* \circ \nabla u, \tag{18.10}$$

Using more general weighted inner products leads to a more general elliptic boundary value problem; see Exercise ■.

As before, (18.10) implies that the Laplacian is positive semi-definite, and positive definite provided $\ker \nabla = \{0\}$. Since, on a connected domain, only constant functions are annihilated by the gradient operator, the Dirichlet and mixed boundary conditions lead to positive definite boundary value problems, while the Neumann boundary value problem is only semi-definite. As a result, the solution to the boundary value problem can be characterized by the three-dimensional version of the Dirichlet minimization principle (15.91).

Theorem 18.1. *The solution $u(x, y, z)$ to the Poisson equation* (18.2) *subject to Dirichlet boundary conditions* (18.3) *is characterized as the unique function that minimizes the* Dirichlet integral

$$\tfrac{1}{2} \,\|\, \nabla u \,\|^2 - \langle \, u \,;\, f \,\rangle = \iiint_\Omega \, \big[\, \tfrac{1}{2} \, (u_x^2 + u_y^2 + u_z^2) - f \, u \, \big] \, dx \, dy \, dz \tag{18.11}$$

among all C^1 *functions that satisfy the prescribed boundary conditions.*

The same argument as in Section 15.4 shows that the same minimization principle applies to solution to the inhomogeneous Dirichlet boundary value problem. For mixed boundary conditions, one must append an additional boundary integral, and the solution minimize the modified Dirichlet integral

$$\iiint_\Omega \, \big[\, \tfrac{1}{2} \, (u_x^2 + u_y^2 + u_z^2) - f \, u \, \big] \, dx \, dy \, dz + \iint_N \, u \, k \, dS, \tag{18.12}$$

where $N \subset \partial\Omega$ is the Neumann part of the boundary. Details are relegated to the exercises. The minimization principle forms the foundation of the three-dimensional finite element method for constructing numerical solutions to the boundary value problem; see [**121**, **num3**] for details.

18.2. Separation of Variables.

Even in higher dimensions, separation of variables remains the workhorse of explicit solution methods for linear partial differential equations. As always, the technique is necessarily restricted to rather specific geometrical configurations. In three-dimensional space, the simplest are problems formulated on rectangular, cylindrical or spherical domains. See [**105**, **108**, **110**] for details on the more exotic types of separable coordinate systems, including ellipsoidal, toroidal, parabolic spheroidal, and so on.

The simplest domain to which the separation of variables method applies is a rectangular box, $R = \{\, 0 < x < a,\ 0 < y < b,\ 0 < z < c \,\}$ A complete separation of variables ansatz $u(x,y,z) = v(x)\, w(y)\, q(z)$ lead to a computation that is almost identical to the two-dimensional version. The details of the resulting Fourier series solution are left to the reader; see Exercise ■.

In the case when the domain is a cylinder, one passes to cylindrical coordinates to effect the separation. The solution can be written in terms of trigonometric functions and Bessel functions, with the details being outlined in Exercise ■. The most interesting case is that of a solid sphere, and this case will be developed in some detail.

Laplace's Equation in a Ball

Suppose we are given a solid spherical ball (e.g., the earth), with a specified temperature distribution on its boundary. The problem is to determine the equilibrium temperature within the ball. To simplify matters, we shall choose units in which the radius of the ball is equal to 1. Therefore, we must solve the Dirichlet boundary value problem

$$\begin{aligned} &\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, && x^2 + y^2 + z^2 < 1, \\ &u(x,y,z) = h(x,y,z), && x^2 + y^2 + z^2 = 1. \end{aligned} \tag{18.13}$$

Problems in spherical geometries usually simplify when re-expressed in terms of spherical coordinates r, φ, θ, as defined by

$$x = r \sin\varphi\, \cos\theta, \qquad y = r \sin\varphi\, \sin\theta, \qquad z = r \cos\varphi.$$

In these coordinates, the Laplace equation takes the form

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{r^2 \sin\varphi}\,\frac{\partial u}{\partial \varphi} + \frac{1}{r^2 \sin^2\varphi}\,\frac{\partial^2 u}{\partial \theta^2} \;=\; 0. \tag{18.14}$$

The derivation of this important formula is the final result of a fairly nasty chain rule computation, and is left to the reader to verify. (Set aside lots of paper and keep an eraser handy!)

To solve the spherical coordinate form of the Laplace equation, we begin by separating off the radial part of the solution, using the separation of variables ansatz

$$u(r,\varphi,\theta) = w(r)\, \chi(\varphi,\theta). \tag{18.15}$$

Substituting into (18.14), dividing the resulting equation through by the product $w\,\chi$ and placing all the terms involving r on one side of the equation yields

$$\frac{r^2}{w}\,\frac{d^2 w}{dr^2} + \frac{2r}{w}\,\frac{dw}{dr} = -\,\frac{\Delta_S[\chi]}{\chi} = \mu, \tag{18.16}$$

where

$$\Delta_S[\chi] = \frac{\partial^2 \chi}{\partial \varphi^2} + \cot\varphi\,\frac{\partial \chi}{\partial \varphi} + \frac{1}{\sin^2\varphi}\,\frac{\partial^2 \chi}{\partial \theta^2}\ . \tag{18.17}$$

The second order differential operator Δ_S, which contains only the angular components of the Laplacian operator, is of particular significance. It is known as the *spherical Laplacian*, and governs the equilibrium and dynamics of thin spherical shells, as discussed below.

Returning to the radially separated form (18.16) of the Laplace equation, our usual separation argument works. The left hand side depends only on r, while the right hand side depends only on φ, θ. This can only occur if both sides are equal to a common separation constant, denoted by μ in the equation. As a consequence, the radial component $w(r)$ satisfies the ordinary differential equation

$$r^2 w'' + 2 r w' - \mu w = 0, \tag{18.18}$$

which is of Euler type (3.76). We will put this equation aside to solve later. The angular components in (18.16) assume the form

$$\Delta_S[\chi] + \mu \chi = 0, \quad \text{or, explicitly,} \quad \frac{\partial^2 \chi}{\partial \varphi^2} + \frac{\cos \varphi}{\sin \varphi} \frac{\partial \chi}{\partial \varphi} + \frac{1}{\sin^2 \varphi} \frac{\partial^2 \chi}{\partial \theta^2} + \mu \chi = 0. \tag{18.19}$$

This second order partial differential equation constitutes the eigenvalue equation for the spherical Laplacian, and is known as the *spherical Helmholtz equation*.

To solve the spherical Helmholtz equation, we adopt a further separation of angular variables,

$$\chi(\varphi, \theta) = p(\varphi)\, q(\theta), \tag{18.20}$$

which we substitute into (18.19). Dividing the result by the product $p\, q$, and then rearranging terms, we are led to a second separated system

$$\frac{\sin^2 \varphi}{p} \frac{d^2 p}{d \varphi^2} + \frac{\cos \varphi \sin \varphi}{p} \frac{dp}{d \varphi} + \mu \sin^2 \varphi = - \frac{1}{q} \frac{d^2 q}{d \theta^2} = \nu.$$

The left hand side depends only on φ while the right hand side depends only on θ, so the two sides must equal a common separation constant, denoted by ν. The spherical Helmholtz equation then splits into a pair of ordinary differential equations

$$\sin^2 \varphi \, \frac{d^2 p}{d \varphi^2} + \cos \varphi \, \sin \varphi \, \frac{dp}{d \varphi} + (\mu \sin^2 \varphi - \nu)\, p = 0, \qquad \frac{d^2 q}{d \theta^2} + \nu \, q = 0. \tag{18.21}$$

The equation for $q(\theta)$ is easy to solve. Since the meridial angle θ varies from 0 to 2π, the function $q(\theta)$ must be a 2π periodic function. Thus, we are reduced to solving the usual periodic boundary value problem for $q(\theta)$; see, for instance, (15.30). The eigenvalue or separation constant takes on the values $\nu = m^2$, where $m = 0, 1, 2, \dots$ is an integer, and

$$q(\theta) = \cos m \, \theta \qquad \text{or} \qquad \sin m \, \theta, \qquad\qquad m = 0, 1, 2, \dots, \tag{18.22}$$

are the required eigenfunctions. Each positive $\nu = m^2 > 0$ admits two linearly independent eigenfunctions, while the $\nu = 0$ only admits the constant eigenfunction $q(\theta) \equiv 1$.

With this information, we next solve the equation for the azimuthal component $p(\varphi)$. This is not an elementary differential equation, and finding the solutions requires some work. The reasoning behind the following steps may not be immediately apparent to the

reader, since it is the result of a long, detailed study of this important differential equation by mathematicians.

First, let us eliminate the trigonometric functions. To this end, we use the change of variables

$$t = \cos\varphi, \qquad\qquad p(\varphi) = P(\cos\varphi) = P(t). \tag{18.23}$$

According to the chain rule,

$$\frac{dp}{d\varphi} = -\sin\varphi\,\frac{dP}{dt} = -\sqrt{1-t^2}\;\frac{dP}{dt}\,,$$
$$\frac{d^2p}{d\varphi^2} = \sin^2\varphi\,\frac{d^2P}{dt^2} - \cos\varphi\,\frac{dP}{dt} = (1-t^2)\,\frac{d^2P}{dt^2} - t\;\frac{dP}{dt}\,.$$

Substituting these expressions into the first equation in (18.21) and using the fact that $\nu = m^2$, we conclude that $P(t)$ must satisfy the differential equation

$$(1-t^2)^2\;\frac{d^2P}{dt^2} - 2\,t\,(1-t^2)\,\frac{dP}{dt} + \big[\,\mu\,(1-t^2) - m^2\,\big]\,P = 0. \tag{18.24}$$

Unfortunately, this differential equation is still not easy to solve, but at least its coefficients are polynomials. Equation (18.24) is known as the *Legendre differential equation* of order m, and its solutions are known as *Legendre functions*, since they were first used by Legendre to analyze the gravitational attraction of ellipsoidal bodies.

While the general solution to the Legendre equation requires a new type of special function, the solutions we are actually interested in can all be written in terms of elementary algebraic functions. First of all, since $t = \cos\varphi$, the solution only needs to be defined on the interval $-1 \le t \le 1$. The endpoints of this interval, $t = \pm 1$, correspond to the north pole, $\varphi = 0$ and the south pole, $\varphi = \pi$, of the sphere. Both endpoints are singular points for the Legendre equation since the coefficient $(1-t^2)^2$ of the leading order derivative vanishes when $t = \pm 1$. Since ultimately we need the separable solution (18.15) to be a well-defined function of x, y, z (even at points where the spherical coordinates degenerate, i.e., on the z axis), we need $p(\varphi)$ to be well-defined at $\varphi = 0, \pi$, and this requires $P(t)$ to be bounded at the singular points $t = \pm 1$. As we learned in our study of the Bessel equation, merely requiring the solution of an ordinary differential equation to be bounded at a singular point can serve as a legitimate boundary condition and serve to distinguish the relevant solutions. Requiring the solution to be bounded at both endpoints is even more restrictive:

$$|\,P(-1)\,| < \infty, \qquad\qquad |\,P(+1)\,| < \infty. \tag{18.25}$$

It turns out that this occurs only for very special values of the separation constant μ.

We will justify the following statements in Appendix C. Consider first the case $m = 0$. In this case, it turns out that the eigenfunctions, i.e., solutions to the Legendre boundary value problem (18.24), (18.25), are the *Legendre polynomials*

$$P_n(t) = \frac{1}{2^n\,n!}\;\frac{d^n}{dt^n}\;(t^2-1)^n \tag{18.26}$$

that we already encountered in Chapter 5. Equation (5.41) contains explicit formulas for the first few Legendre polynomials. Indeed, we now finally comprehend the reason for the orthogonality of the Legendre polynomials. They are the common eigenfunctions of a self-adjoint boundary value problem! Their orthogonality is a consequence of the general theorem on eigenvectors or eigenfunctions of self-adjoint linear operators, and is discussed in detail in Exercise ■.

For general $m > 0$, the eigenfunctions of the Legendre boundary value problem (18.24), (18.25) are not always polynomials. They are known as the *associated Legendre functions*, and can be constructed using the explicit formula

$$P_n^m(t) = (1-t^2)^{m/2}\,\frac{d^m}{dt^m}\,P_n(t) = \frac{(1-t^2)^{m/2}}{2^n\,n!}\,\frac{d^{n+m}}{dt^{n+m}}\,(t^2-1)^n, \qquad n = m, m+1, \dots. \tag{18.27}$$

Here is a list of the first few associated Legendre functions:

$$\begin{aligned}
&P_0^0(t) = 1, && P_1^0(t) = t,\\
&P_1^1(t) = -\sqrt{1-t^2}, && P_2^0(t) = \tfrac{3}{2}\,t^2 - \tfrac{1}{2},\\
&P_2^1(t) = -3\,t\,\sqrt{1-t^2}, && P_2^2(t) = -3\,(t^2-1),\\
&P_3^0(t) = \tfrac{5}{2}\,t^3 - \tfrac{3}{2}\,t, && P_3^1(t) = -\tfrac{3}{2}\,\sqrt{1-t^2}\,(5\,t^2-1),\\
&P_3^2(t) = -15\,(t^3-t), && P_3^3(t) = -15\,(1-t^2)^{3/2},\\
&P_4^0(t) = \tfrac{35}{8}\,t^4 - \tfrac{15}{4}\,t^2 + \tfrac{3}{8}, && P_4^1(t) = -\tfrac{5}{2}\,\sqrt{1-t^2}\,(7\,t^3-3\,t),\\
&P_4^2(t) = -\tfrac{15}{2}\,(7\,t^4 - 8\,t^2 + 1), && P_4^3(t) = -105\,t\,(1-t^2)^{3/2},\\
&P_4^4(t) = 105\,(t^4 - 2\,t^2 + 1).
\end{aligned} \tag{18.28}$$

When $m = 2k \geq n$ is even, $P_n^m(t)$ is a polynomial function, while when $m = 2k+1 \leq n$ is odd, it has an extra factor of $\sqrt{1-t^2}$ multiplying a polynomial. Keep in mind that the square root is real and positive since we are restricting our attention to the interval $-1 \leq t \leq 1$. If $m > n$ then the formula (18.27) yields zero.

Graphs of the Legendre polynomials $P_n(t) = P_n^0(t)$ can be found in Figure 5.3. In addition, Figure 18.1 displays the graphs of the associated Legendre functions $P_2^1(t), \dots, P_4^4(t)$. (The graph of $P_1^1(t)$ is omitted since it is merely a semi-circle.) Pay particular attention to the fact that the graphs have quite different vertical scales.

Theorem 18.2. *Let $m \geq 0$ be a non-negative integer. Then the eigenfunctions for the m^{th} order Legendre boundary value problem prescribed by* (18.24), (18.25) *are the associated Legendre functions $P_n^m(t)$ for $n = 0, 1, 2, \dots$. The corresponding eigenvalues are $\mu_n = n(n+1)$.*

Returning to the original azimuthal variable φ, we discover that the boundary value problem

$$\sin^2\varphi\,\frac{d^2p}{d\varphi^2} + \cos\varphi\,\sin\varphi\,\frac{dp}{d\varphi} + \mu\sin^2\varphi\cdot p - m^2\,p = 0, \qquad |\,p(0)\,|, \quad |\,p(\pi)\,| < \infty, \tag{18.29}$$

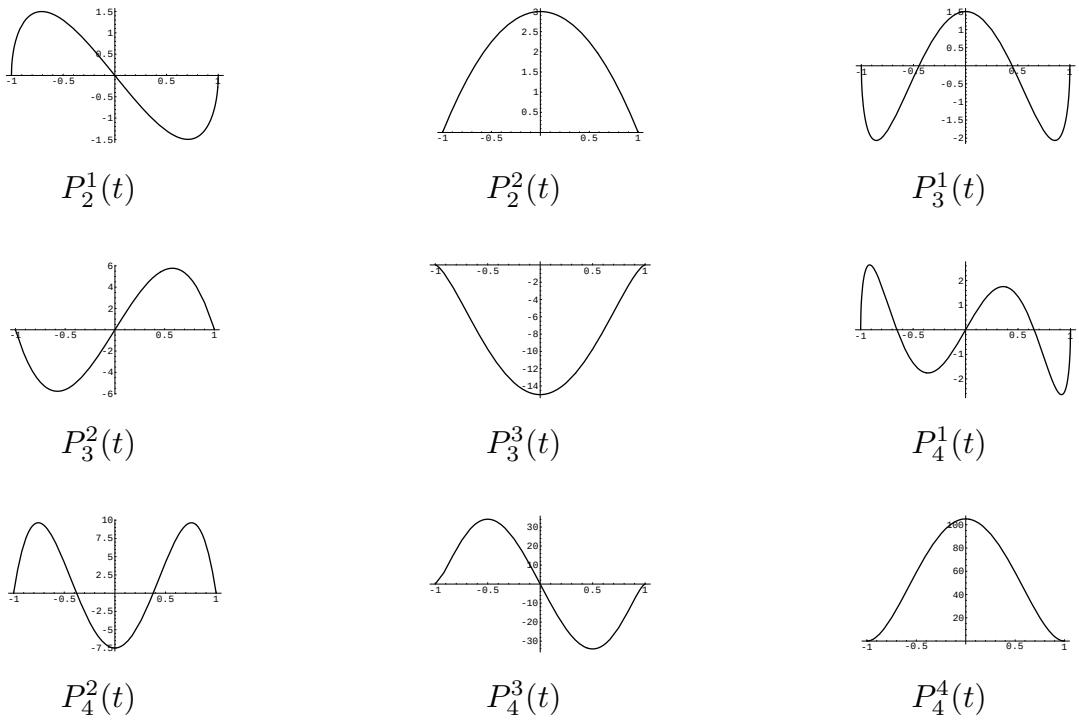


Figure 18.1. Associated Legendre Functions.

has eigenvalues and eigenfunctions

$$\mu_n = n(n+1), \qquad p_n^m(\varphi) = P_n^m(\cos\varphi), \qquad \text{for} \qquad n = m, m+1, \dots, \tag{18.30}$$

given in terms the associated Legendre functions. The n^{th} eigenvalue μ_n admits a total of $n+1$ linearly independent eigenfunctions, namely $p_n^0(\varphi), \dots, p_n^n(\varphi)$. The functions $p_n^m(\varphi)$ are, in fact, trigonometric polynomials of degree n. Here are the first few, written in Fourier form:

$$\begin{aligned}
&p_0^0(\varphi) = 1, && p_1^0(\varphi) = \cos\varphi,\\
&p_1^1(\varphi) = -\sin\varphi, && p_2^0(\varphi) = \tfrac{1}{4} + \tfrac{3}{4}\cos 2\varphi,\\
&p_2^1(\varphi) = -\tfrac{3}{2}\sin 2\varphi, && p_2^2(\varphi) = \tfrac{3}{2} - \tfrac{3}{2}\cos 2\varphi,\\
&p_3^0(\varphi) = \tfrac{3}{8}\cos\varphi + \tfrac{5}{8}\cos 3\varphi, && p_3^1(\varphi) = -\tfrac{3}{8}\sin\varphi - \tfrac{15}{8}\sin 3\varphi,\\
&p_3^2(\varphi) = \tfrac{15}{4}\cos\varphi - \tfrac{15}{4}\cos 3\varphi, && p_3^3(\varphi) = -\tfrac{45}{4}\sin\varphi + \tfrac{15}{4}\sin 3\varphi,\\
&p_4^0(\varphi) = \tfrac{9}{64} + \tfrac{5}{16}\cos 2\varphi + \tfrac{35}{64}\cos 4\varphi, && p_4^1(\varphi) = -\tfrac{5}{8}\sin 2\varphi - \tfrac{35}{16}\sin 4\varphi,\\
&p_4^2(\varphi) = \tfrac{45}{16} + \tfrac{15}{4}\cos 2\varphi - \tfrac{105}{16}\cos 4\varphi, && p_4^3(\varphi) = -\tfrac{105}{4}\sin 2\varphi + \tfrac{105}{8}\sin 4\varphi,\\
&p_4^4(\varphi) = \tfrac{315}{8} - \tfrac{105}{2}\cos 2\varphi + \tfrac{105}{8}\cos 4\varphi.
\end{aligned} \tag{18.31}$$

It is also instructive to plot the eigenfunctions in terms of the angle φ and compare with those in Figure 18.1; see Figure Lphi■.

At this stage, we have determined both angular components of our separable solutions (18.20). Multiplying the two parts together results in the spherical angle functions

$$\begin{aligned} Y_n^m(\varphi,\theta) &= \cos m\theta\ P_n^m(\cos\varphi), & & n = 0,1,2,\dots, \\ \widetilde{Y}_n^m(\varphi,\theta) &= \sin m\theta\ P_n^m(\cos\varphi), & & m = 0,1,\dots,n, \end{aligned} \tag{18.32}$$

which are known as *spherical harmonics*. The spherical harmonics $Y_n^m, \widetilde{Y}_n^m$ satisfy the spherical Helmholtz equation

$$\Delta_S\, Y_n^m + n(n+1)\, Y_n^m \;=\; 0 \;=\; \Delta_S\, \widetilde{Y}_n^m + n(n+1)\, \widetilde{Y}_n^m. \tag{18.33}$$

In other words, the spherical harmonics are the eigenfunctions for the spherical Laplacian operator, (18.17), with associated eigenvalues $\mu_n = n(n+1)$ for $n = 0,1,2,\dots$. The n^{th} eigenvalue μ_n admits a $(2n+1)$–dimensional eigenspace, spanned by the spherical harmonics

$$Y_n^0(\varphi,\theta),\quad Y_n^1(\varphi,\theta),\quad \dots,\quad Y_n^n(\varphi,\theta),\quad \widetilde{Y}_n^1(\varphi,\theta),\quad \dots,\quad \widetilde{Y}_n^n(\varphi,\theta).$$

The omitted function $\widetilde{Y}_n^0(\varphi,\theta) \equiv 0$ is trivial, and so does not contribute.

Self-adjointness of the spherical Laplacian operator implies that the spherical harmonics are orthogonal with respect to the inner product

$$\langle\, f\,;g\,\rangle = \iint_{S_1} f\, g\, dS = \int_0^{\pi}\int_0^{2\pi} f(\varphi,\theta)\, g(\varphi,\theta)\sin\varphi\ d\theta\, d\varphi, \tag{18.34}$$

where the surface area integral is over the sphere $S_1 = \{\,\|\,\mathbf{x}\,\| = 1\,\}$ of radius 1, cf. (B.40). More correctly, self-adjointness only guarantees orthogonality for the harmonics corresponding to different eigenvalues. However, by our construction, the orthogonality formula (18.34) does, in fact, hold in general. The spherical harmonic norms can be explicitly computed:

$$\|\, Y_n^0\,\|^2 = \frac{4\pi}{2n+1}\,, \qquad \|\, Y_n^m\,\|^2 = \|\, \widetilde{Y}_n^m\,\|^2 = \frac{2\pi(n+m)!}{(2n+1)(n-m)!}\,. \tag{18.35}$$

Just as with the Fourier trigonometric functions, the case $m = 0$, where the spherical harmonic $Y_n^0(\varphi)$ does not depend upon θ, is special. A proof of this formula appears in Exercise ∎.

With some further work, it can be proved that the harmonic polynomials form a complete orthogonal system of functions on the unit sphere. This means that any reasonable, e.g., piecewise C^1, function $h\colon S_1 \to \mathbb{R}$ can be expanded into a convergent *spherical Fourier series*

$$h(\varphi,\theta) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty}\left(\frac{c_{0,n}}{2}\, Y_n^0(\varphi) + \sum_{m=1}^{n}\Big[\, c_{m,n} Y_n^m(\varphi,\theta) + \widetilde{c}_{m,n}\widetilde{Y}_n^m(\varphi,\theta)\,\Big]\right) \tag{18.36}$$

in the spherical harmonics. Applying the orthogonality relations (18.34), the spherical Fourier coefficients are given by the inner products

$$c_{0,n} = \frac{2\,\langle\, f\,;Y_n^0\,\rangle}{\|\, Y_n^0\,\|^2}\,, \qquad c_{m,n} = \frac{\langle\, f\,;Y_n^m\,\rangle}{\|\, Y_n^m\,\|^2}\,, \qquad \widetilde{c}_{m,n} = \frac{\langle\, f\,;\widetilde{Y}_n^m\,\rangle}{\|\, \widetilde{Y}_n^m\,\|^2}\,, \qquad \begin{aligned} &n \ge 0, \\ &1 \le m \le n, \end{aligned}$$

or, explicitly, using the formulae (18.35) for the norms,

$$\begin{aligned} c_{m,n} &= \frac{(2n+1)(n-m)!}{2\pi\,(n+m)!} \int_0^{2\pi}\!\!\int_0^{\pi} h(\varphi,\theta)\, \cos n\theta\, P_n(\cos\varphi) \sin\varphi\; d\varphi\, d\theta, \\ \widetilde{c}_{m,n} &= \frac{(2n+1)(n-m)!}{2\pi\,(n+m)!} \int_0^{2\pi}\!\!\int_0^{\pi} h(\varphi,\theta)\, \sin n\theta\, P_n(\cos\varphi) \sin\varphi\; d\varphi\, d\theta. \end{aligned} \tag{18.37}$$

The factor $\sin\varphi$ comes from the spherical surface area formula (B.40). As with an ordinary Fourier series, the extra $\frac{1}{2}$ was introduced in the $c_{0,n}$ terms in the series (18.37) so that the formulae (18.37) are valid for all m, n. In particular, the constant term the spherical harmonic series

$$\frac{c_{0,0}}{2} = \frac{1}{4\pi} \iint_{S_1} h\, dS = \frac{1}{4\pi} \int_0^{2\pi}\!\!\int_0^{\pi} h(\varphi,\theta)\, \sin\varphi\, d\varphi\, d\theta \tag{18.38}$$

is the mean of the function f over the unit sphere.

To complete our solution to the Laplace equation on the solid ball, we still need to analyze the ordinary differential equation (18.18) for the radial component $w(r)$. Using the fact that the original separation constant is $\mu = n(n+1)$ for some non-negative integer $n \geq 0$, the radial equation (18.18) takes the form

$$r^2\, w'' + 2r\, w' - n(n+1)\, w = 0. \tag{18.39}$$

As noted earlier, this is a second order linear equation of Euler type (3.76), and can be solved by using the power ansatz $w(r) = r^\alpha$. Substituting into the equation, we find the exponent must satisfy the quadratic equation

$$\alpha^2 + \alpha - n(n+1) = 0, \qquad \text{and hence} \qquad \alpha = n \qquad \text{or} \qquad \alpha = -(n+1).$$

Therefore, the two linearly independent solutions are

$$w_1(r) = r^n \qquad \text{and} \qquad w_2(r) = r^{-n-1}. \tag{18.40}$$

Since we are only interested in solutions that remain bounded at $r = 0$ — the center of the ball — we should just retain the first solution $w(r) = r^n$ in our analysis.

At this stage, we have solved all three ordinary differential equations for the separable solutions. We combine the results (18.22), (18.32), (18.40) together to produce the spherically separable solutions (18.15) to the Laplace equation

$$\begin{aligned} H_n^m(r,\varphi,\theta) &= r^n\, Y_n^m(\varphi,\theta) = r^n\, \cos m\theta\; P_n^m(\cos\varphi), & n &= 0,1,2,\dots, \\ \widetilde{H}_n^m(r,\varphi,\theta) &= r^n\, \widetilde{Y}_n^m(\varphi,\theta) = r^n\, \sin m\theta\; P_n^m(\cos\varphi), & m &= 0,1,\dots,n, \end{aligned} \tag{18.41}$$

known as *harmonic polynomials*. As the name suggests, they are, in fact, polynomial

functions of the rectangular coordinates x, y, z. The first few harmonic polynomials are

$$\begin{aligned}
&H_0^0 = 1, && H_1^0 = z, && H_2^0 = z^2 - \tfrac{1}{2}\,x^2 - \tfrac{1}{2}\,y^2 && H_3^0 = z^3 - \tfrac{3}{2}\,x^2 z - \tfrac{3}{2}\,y^2 z \\
&&& H_1^1 = x, && H_2^1 = 3\,x\,z, && H_3^1 = 6\,x\,z^2 - \tfrac{3}{2}\,x^3 - \tfrac{3}{2}\,x\,y^2 \\
&&& \widetilde{H}_1^1 = y, && \widetilde{H}_2^1 = 3\,y\,z, && \widetilde{H}_3^1 = 6\,y\,z^2 - \tfrac{3}{2}\,x^2 y - \tfrac{3}{2}\,y^3 \\
&&&&& H_2^2 = 3\,x^2 - 3\,y^2, && H_3^2 = 15\,x^2 z - 15\,y^2 z \\
&&&&& \widetilde{H}_2^2 = 6\,x\,y, && \widetilde{H}_3^2 = 30\,x\,y\,z \\
&&&&&&& H_3^3 = 15\,x^3 - 45\,x\,y^2 \\
&&&&&&& \widetilde{H}_3^3 = 45\,x^2 y - 15\,y^3.
\end{aligned} \tag{18.42}$$

Note that H_n^m and $\widetilde{H}_n^m$ are homogeneous polynomials of degree n. Indeed, the harmonic polynomials

$$H_n^0, \quad H_n^1, \ \dots\ , \quad H_n^n, \quad \widetilde{H}_n^1, \ \dots\ \widetilde{H}_n^n$$

form a basis for the subspace of all homogeneous polynomials of degree n that solve the three-dimensional Laplace equation, which therefore has dimension $2\,n + 1$. (The unlisted $\widetilde{H}_n^0 \equiv 0$ is trivial, and so is not part of the basis.) Plotting these functions in a visually instructive manner is challenging. Since they depend upon three variables, we are in sore need of a four-dimensional viewing system to properly display and appreciate their graphs.

As we shall see, the harmonic polynomials form a complete system, and therefore the general solution to the Laplace equation on the sphere can be written as a series therein:

$$\begin{aligned}
u(r,\varphi,\theta) &= \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, H_n^0(r,\varphi) + \sum_{m=1}^{n} \Big[c_{m,n} H_n^m(r,\varphi,\theta) + \widetilde{c}_{m,n} \widetilde{H}_n^m(r,\varphi,\theta) \Big] \right) \\
&= \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, Y_n^0(\varphi) + \sum_{m=1}^{n} \Big[c_{m,n}\, r^n\, Y_n^m(\varphi,\theta) + \widetilde{c}_{m,n}\, r^n\, \widetilde{Y}_n^m(\varphi,\theta) \Big] \right).
\end{aligned} \tag{18.43}$$

To complete our solution to the boundary value problem, we substitute the harmonic polynomial series solution into the Dirichlet boundary conditions on the unit sphere $r = 1$, yielding

$$u(1,\varphi,\theta) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, Y_n^0(\varphi) + \sum_{m=1}^{n} \Big[c_{m,n} Y_n^m(\varphi,\theta) + \widetilde{c}_{m,n} \widetilde{Y}_n^m(\varphi,\theta) \Big] \right) = h(\varphi,\theta). \tag{18.44}$$

In view of the preceding remarks, the coefficients $c_{m,n}, \widetilde{c}_{m,n}$ in this harmonic polynomial series are given by the orthogonality formulae (18.37). If they are bounded — which occurs for all L^2 functions h and also certain generalized functions, including the delta function — then it is not hard to prove that the series converges everywhere, and, in fact, uniformly on any smaller ball $\| x \| = r \le r_0 < 1$.

Interestingly, if we revert to rectangular coordinates, then the spherical Fourier series

(18.43) takes the form

$$u(x,y,z) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, H_n^0(x,y,z) + \sum_{m=1}^{n} \Big[c_{m,n} H_n^m(x,y,z) + \widetilde{c}_{m,n} \widetilde{H}_n^m(x,y,z) \Big] \right). \tag{18.45}$$

The summand at order n is, in fact, a homogeneous polynomial of degree n. Therefore, the Fourier series expands the function into a power series which is, in fact, the *Taylor series expansion for the harmonic function u at the origin*! Any convergent Taylor expansion converges to an analytic function. Therefore, just like their two-dimensional siblings, harmonic functions are, in fact, analytic. According to the preceding paragraph, the radius of convergence of the spherical harmonic Fourier/Taylor series is at least one, and so $u(x,y,z)$ is analytic inside the entire ball — no matter how wild its boundary values are.

The constant term in such a Taylor series can be identified with the value of the function at the origin. On the other hand, the orthogonality formula (18.38) tells us that

$$u(0,0,0) = \frac{c_{0,0}}{2} = \frac{1}{4\,\pi} \iint_{S_1} u\, dS. \tag{18.46}$$

Therefore, we have established the three-dimensional version of the planar Theorem 15.7: the value of the harmonic function at the center of the sphere is equal to the average of its values $u = h$ on the sphere's surface.

Theorem 18.3. *If $u(\mathbf{x})$ is a harmonic function for all $\mathbf{x} \in \Omega \subseteq \mathbb{R}^3$, then u is analytic in Ω. Moreover, its value at a point $\mathbf{x}_0 \in \Omega$,*

$$u(\mathbf{x}_0) = \frac{1}{4\,\pi\, a^2} \iint_{S_a(\mathbf{x}_0)} u\; dS, \tag{18.47}$$

is equal to the average of its values on any sphere $S_a(\mathbf{x}_0) = \{\, \| \mathbf{x} - \mathbf{x}_0 \| = a \,\}$ centered at the point — provided u is harmonic on the entire enclosed ball $B_a(\mathbf{x}_0) = \{\, \| \mathbf{x} - \mathbf{x}_0 \| \le a \,\} \subset \Omega$.

Proof: It is easily checked that, under the hypothesis of the theorem,

$$U(\mathbf{x}) = u(a\,\mathbf{x} + \mathbf{x}_0)$$

is harmonic on the unit ball $\| \mathbf{x} \| \le 1$, and hence solves the boundary value problem (18.13) with boundary values $h(\mathbf{x}) = U(\mathbf{x}) = u(a\,\mathbf{x} + \mathbf{x}_0)$ on $\| \mathbf{x} \| = 1$ coming from the values of u on the sphere $S_a(\mathbf{x}_0)$. By the preceding remarks, $U(\mathbf{x})$ is analytic for $\| \mathbf{x} \| < 1$, and so $u(\mathbf{x}) = U((\mathbf{x} - \mathbf{x}_0)/a)$ is analytic inside $B_a(\mathbf{x}_0)$, and, in particular at $\mathbf{x}_0$. Since $\mathbf{x}_0$ was arbitrary, this proves analyticity of u everywhere in $E\Omega$. Moreover, according to (18.46),

$$u(\mathbf{x}_0) = U(0) = \frac{1}{4\,\pi} \iint_{S_1} U\, dS = \frac{1}{4\,\pi\, a^2} \iint_{S_a(\mathbf{x}_0)} u\; dS,$$

which proves the result. *Q.E.D.*

Arguing as in Corollary 15.8, we establish a corresponding maximum principle for harmonic functions of 3 variables.

Corollary 18.4. *A harmonic function cannot have a local maximum or minimum at any interior point of its domain of definition.*

For instance, this result implies that a body in thermal equilibrium can achieve its maximum and minimum temperature only on the boundary of the domain. In physical language, heat energy must flow away from any internal maximum and towards any internal minimum. Thus, a local maximum or minimum of temperature would preclude the body being in thermal equilibrium.

Example 18.5. In this example, we shall determine the electrostatic potential inside a hollow sphere when the upper and lower hemispheres are held at different constant potentials. This device is called a *spherical capacitor* and is realized experimentally by separating the two charged hemispheres by a thin insulating ring at the equator. A straightforward scaling argument allows us to choose our units so that the sphere has radius 1, while the potential is set equal to 1 on the upper hemisphere and 0 or grounded on the lower hemisphere. Therefore, we need to solve the Laplace equation $\Delta u = 0$ inside a solid ball $\| \mathbf{x} \| < 1$, with Dirichlet boundary conditions

$$u(x,y,z) = \begin{cases} 1, & z > 0, \\ 0, & z < 0, \end{cases} \qquad \text{on} \qquad \| \mathbf{x} \| = 1. \tag{18.48}$$

The solution will be prescribed by a harmonic polynomials series (18.45) whose coefficients are determined by the boundary values (18.48). Before making on the required computation, let us first note that since the boundary data does not depend upon the meridial angle θ, the solution $u = u(r,\varphi)$ will also be independent of θ. Therefore, we need only consider the θ-independent spherical harmonics, which are those with $m = 0$, and hence

$$u(r,\varphi) \;=\; \frac{1}{2} \sum_{n=0}^{\infty} c_n \, H_n^0(x,y,z) \;=\; \frac{1}{2} \sum_{n=0}^{\infty} c_n \, r^n \, P_n(\cos\varphi),$$

where we abbreviate $c_{0,n} = c_n$. The boundary conditions require

$$u(1,\varphi) = \frac{1}{2} \sum_{n=0}^{\infty} c_n \, P_n(\cos\varphi) = f(\varphi) = \begin{cases} 1, & 0 \le \varphi < \frac{1}{2}\,\pi, \\ 0, & \frac{1}{2}\,\pi < \varphi \le \pi. \end{cases}$$

The coefficients are given by (18.37), which, in the case $m = 0$, reduces to

$$c_n = \frac{2\,n+1}{2\,\pi} \iint_S f \, Y_n^0 \, dS = (2\,n+1) \int_0^{\pi/2} P_n(\cos\varphi) \sin\varphi \; d\varphi = (2\,n+1) \int_0^1 P_n(t)\, dt. \tag{18.49}$$

The first few are

$$c_0 = 1, \qquad c_1 = \tfrac{3}{2}, \qquad c_2 = 0, \qquad c_3 = -\tfrac{7}{8}, \qquad c_4 = 0, \qquad \dots$$

Therefore, the solution has the explicit Taylor expansion

$$\begin{aligned} u &= \tfrac{1}{2} + \tfrac{3}{4}\, r\cos\varphi - \tfrac{21}{64}\, r^3 \cos\varphi - \tfrac{35}{64}\, r^3 \cos 3\varphi + \;\cdots \\ &= \tfrac{1}{2} + \tfrac{3}{4}\, z - \tfrac{7}{8}\, z^3 - \tfrac{21}{16}\,(x^2+y^2)\, z + \;\cdots\; . \end{aligned} \tag{18.50}$$

Note in particular that the value $u(0,0,0) = \frac{1}{2}$ at the center of the sphere is the average of its boundary values, in accordance with Corollary 18.4.

Remark: The same function solves the problem of thermal equilibrium in a solid sphere with the upper hemisphere held at temperature 1° and the lower hemisphere at 0°.

Example 18.6. A closely related problem is to determine the electrostatic potential *outside* a spherical capacitor. As in the preceding example, we take our capacitor of radius 1, with electrostatic charge $+1$ on the upper hemisphere and 0 on the lower hemisphere. Here, we need to solve the Laplace equation

$$\Delta u = 0, \qquad \| \mathbf{x} \| > 1,$$

in the unbounded domain outside a solid unit ball, subject to Dirichlet boundary conditions

$$u = \begin{cases} 0, & z > 0, \\ 1, & z < 0, \end{cases} \qquad \text{on the unit sphere} \qquad \| \mathbf{x} \| = 1.$$

We expect the potential to be small at large distances $r = \| \mathbf{x} \| \gg 1$ away from the capacitor. Therefore, the non-constant harmonic polynomial solutions will not help us solve this problem, since they tend to ∞ as $\| \mathbf{x} \| \to \infty$.

However, by returning to our original separation of variables argument, we can construct a different class of solutions with the desired decay properties. When we solved the radial equation (18.39), we discarded the solution $w_2(r) = r^{-n-1}$ because it had a singularity at the origin. In the present situation, the behavior of the function at $r = 0$ is irrelevant; our current requirement is that the solution decays as $r \to \infty$, and this is now valid. Therefore, we can use the functions

$$\begin{aligned} K_n^m(x,y,z) &= r^{-2n-1} H_n^m(x,y,z) = r^{-n-1} Y_n^m(\varphi,\theta) = r^{-n-1} \cos m\theta \; P_n^m(\cos\varphi), \\ \widetilde{K}_n^m(x,y,z) &= r^{-2n-1} \widetilde{H}_n^m(x,y,z) = r^{-n-1} \widetilde{Y}_n^m(\varphi,\theta) = r^{-n-1} \sin m\theta \; P_n^m(\cos\varphi), \end{aligned} \tag{18.51}$$

for solving such exterior problems. In the present case, we only need the functions that are independent of θ, which means $m = 0$. We write the resulting solution as a series

$$u(r,\varphi) = \frac{1}{2} \sum_{n=0}^{\infty} c_n \, K_n^0(x,y,z) = \frac{1}{2} \sum_{n=1}^{\infty} c_n \, r^{-n-1} \, P_n(\cos\varphi).$$

The boundary conditions

$$u(1,\varphi) = \frac{1}{2} \sum_{n=1}^{\infty} c_n \, P_n(\cos\varphi) = f(\varphi) = \begin{cases} 1, & 0 \le \varphi < \frac{1}{2}\pi, \\ 0, & \frac{1}{2}\pi < \varphi \le \pi, \end{cases}$$

are identical with the previous example. Therefore, the coefficients are given by (18.49), leading to the series expansion

$$u = \frac{1}{2\,r} + \frac{3\cos\varphi}{4\,r^2} - \frac{21\cos\varphi + 35\cos 3\varphi}{64\,r^3} + \cdots = \frac{1}{2\,r} + \frac{3\,z}{4\,r^3} - \frac{14\,z^3 - 21\,(x^2+y^2)\,z}{16\,r^5} + \cdots, \tag{18.52}$$

where $r = \sqrt{x^2 + y^2 + z^2}$. Interestingly, at large distances, the higher order terms become negligible, and the potential looks like that associated with a point charge of magnitude $\frac{1}{2}$ — the average of the potential over the sphere — that is concentrated at the origin. This is indicative of a general fact; see Exercise ■.

18.3. The Green's Function.

We now turn to the inhomogeneous form of Laplace's equation — the Poisson equation

$$-\Delta u = f \qquad \text{for all} \qquad \mathbf{x} \in \Omega. \tag{18.53}$$

In applications, $f(\mathbf{x}) = f(x, y, z)$ represents some form of external forcing inside the solid domain. To uniquely specify the solution, we need to impose appropriate boundary conditions — Dirichlet, Neumann, or mixed. We shall mostly concentrate on the homogeneous boundary variational problem.

As we learned in Chapters 11 and 15, the solution to the Poisson equation for a general inhomogeneity $f(\mathbf{x})$ can be found by a superposition formula based on the *Green's function*, which is defined to be the particular solution corresponding to a delta function inhomogeneity that is concentrated at a single point in the domain. Thus, for each $\boldsymbol{\xi} = (\xi, \eta, \zeta) \in \Omega$, the Green's function $G(\mathbf{x}; \boldsymbol{\xi}) = G(x, y, z; \xi, \eta, \zeta)$ is the unique solution to the Poisson equation

$$-\Delta u = \delta(\mathbf{x} - \boldsymbol{\xi}) = \delta(x - \xi)\, \delta(y - \eta)\, \delta(z - \zeta) \qquad \text{for all} \qquad \mathbf{x} \in \Omega, \tag{18.54}$$

subject to the chosen homogeneous boundary conditions. The solution to the general Poisson equation (18.53) is then obtained by superposition: We write the forcing function

$$f(x, y, z) = \iiint_{\Omega} f(\xi, \eta, \zeta)\, \delta(x - \xi)\, \delta(y - \eta)\, \delta(z - \zeta)\, d\xi\, d\eta\, d\zeta$$

as a linear superposition of delta functions. By linearity, the solution

$$u(x, y, z) = \iiint_{\Omega} f(\xi, \eta, \zeta)\, G(x, y, z; \xi, \eta, \zeta)\, d\xi\, d\eta\, d\zeta \tag{18.55}$$

is then given as the same superposition of the Green's function solutions.

The Green's Function on the Entire Space

Except in a few specific instances, the explicit formula for the Green's function is difficult to find. Nevertheless, certain general, useful features can be established. The starting point is to investigate the Poisson equation (18.54) when the domain $\Omega = \mathbb{R}^3$ is all of three-dimensional space. Since the Laplacian is invariant under translations we can, without loss of generality, place our delta impulse at the origin, and solve the particular case

$$-\Delta u = \delta(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^3.$$

Since $\delta(\mathbf{x}) = 0$ for all $\mathbf{x} \neq \mathbf{0}$, the desired solution will, in fact, be a solution to the homogeneous Laplace equation

$$\Delta u = 0, \qquad \mathbf{x} \neq \mathbf{0},$$

save, possibly, for a single singularity concentrated at the origin. We impose boundary constraints by seeking a solution that goes to zero, $u \to 0$, at large distances $\| \mathbf{x} \| \to \infty$.

The Laplace equation models the equilibria of a homogeneous, isotropic medium, and so is also invariant under rotations. This indicates that, in any radially symmetric configuration, the solution $u = u(r)$ should only depend upon the distance from the origin, $r = \| \mathbf{x} \|$, and not the angular direction. Referring to the spherical coordinate form (18.14) of the Laplacian operator, if u only depends upon r, its derivatives with respect to the angular coordinates φ, θ are zero, and so $u(r)$ solves the ordinary differential equation

$$\frac{d^2u}{dr^2} + \frac{2}{r}\,\frac{du}{dr} = 0. \tag{18.56}$$

This equation is, in effect, a first order linear ordinary differential equation for $v = du/dr$ and hence is easy to solve. The solutions are of the form

$$\frac{du}{dr} = v(r) = b \log r, \qquad \text{and hence} \qquad u(r) = a + \frac{b}{r}\,,$$

where a, b are arbitrary constants. The constant solution $u(r) = a$ does not die away at large distances, nor does it have a singularity at the origin. Therefore, if our intuition is valid, the desired solution should be of the form

$$u = \frac{b}{r} = \frac{b}{\| \mathbf{x} \|} = \frac{b}{\sqrt{x^2 + y^2 + z^2}}\,. \tag{18.57}$$

Indeed, this function is harmonic — solves Laplace's equation — everywhere away from the origin, and has a singularity at $\mathbf{x} = \mathbf{0}$.

Remark: This solution is, up to constant multiple, the three-dimensional Newtonian gravitational potential due to a point mass at the origin. Its gradient

$$\mathbf{f}(\mathbf{x}) \;=\; \nabla \left(\frac{b}{\| \mathbf{x} \|} \right) \;=\; - \frac{b\,\mathbf{x}}{\| \mathbf{x} \|^3}\,.$$

defines the gravitational force vector at the point $\mathbf{x}$. When $b > 0$, the force vector $\mathbf{f}(\mathbf{x})$ points in the direction of the mass concentrated at the origin. Its magnitude

$$\| \mathbf{f} \| \;=\; \frac{b}{\| \mathbf{x} \|^2} \;=\; \frac{b}{r^2}$$

is proportional to one over the squared distance, and so satisfies the well-known inverse square law of three-dimensional Newtonian gravity.

The inverse square law also models the electrostatic forces between charged bodies. Thus, (18.57) can be interpreted as the electrostatic potential on a charged mass at position $\mathbf{x}$ due to a electric charge that is concentrated at the origin. The constant b is positive when the charges are of opposite signs, leading to an attractive force, and negative in the repulsive case of like charges.

Returning to our problem, our remaining task is to fix the multiple b such that the Laplacian of our candidate solution (18.57) has a delta function singularity at the origin; equivalently, we must find c such that

$$-\Delta r^{-1} = c\,\delta(\mathbf{x}). \tag{18.58}$$

We already know that this equation holds away from the origin, since $\delta(\mathbf{x}) = 0$ when $\mathbf{x} \neq \mathbf{0}$. To investigate near the singularity, we integrate both sides of (18.58) over a small solid ball $B_\varepsilon = \{\, r = \| \mathbf{x} \| = \varepsilon \,\}$ of radius ε:

$$-\iiint_{B_\varepsilon} \Delta r^{-1}\, dx\, dy\, dz = \iiint_{B_\varepsilon} c\,\delta(\mathbf{x})\, dx\, dy\, dz = c, \tag{18.59}$$

where we used the definition of the delta function to evaluate the right hand side. On the other hand, since $\Delta r^{-1} = \nabla \cdot \nabla r^{-1}$, we can use the divergence theorem (B.82) to evaluate the left hand integral, whence

$$\iiint_{B_\varepsilon} \Delta r^{-1}\, dx\, dy\, dz = \iiint_{B_\varepsilon} \nabla \cdot \nabla r^{-1}\, dx\, dy\, dz = \iint_{S_\varepsilon} \frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{r}\right) dS,$$

where the surface integral is over the bounding sphere $S_\varepsilon = \partial B_\varepsilon = \{\, \| \mathbf{x} \| = \varepsilon \,\}$. The normal $\mathbf{n}$ to the sphere points in the radial direction, and hence the normal derivative coincides with differentiation with respect to r. Therefore,

$$\frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{r}\right) = \frac{\partial}{\partial r}\left(\frac{1}{r}\right) = -\,\frac{1}{r^2}\,.$$

The surface integral can now be explicitly evaluated:

$$\iint_{S_\varepsilon} \frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{r}\right) dS = -\iint_{S_\varepsilon} \frac{1}{r^2}\, dS = -\iint_{S_\varepsilon} \frac{1}{\varepsilon^2}\, dS = -\,4\pi,$$

since S_ε has surface area $4\pi\varepsilon^2$. Substituting this result back into (18.59), we conclude that

$$c = 4\pi, \qquad \text{and hence} \qquad -\,\Delta r^{-1} = 4\pi\,\delta(\mathbf{x}). \tag{18.60}$$

This is our desired formula! Therefore, the Green's function for a delta function impulse at the origin is

$$G(x,y,z) = \frac{1}{4\pi\, r} = \frac{1}{4\pi\, \| \mathbf{x} \|} = \frac{1}{4\pi\, \sqrt{x^2 + y^2 + z^2}}\,. \tag{18.61}$$

If the singularity is concentrated at the point $\boldsymbol{\xi} = (\xi, \eta, \zeta)$ instead of the origin, then we merely translate the preceding solution. This leads immediately to the Green's function

$$G(\mathbf{x}; \boldsymbol{\xi}) = G(\mathbf{x} - \boldsymbol{\xi}) = \frac{1}{4\pi\, \| \mathbf{x} - \boldsymbol{\xi} \|} = \frac{1}{4\pi\, \sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}} \tag{18.62}$$

on all of space. As a consequence of the superposition formula (18.55), we have proved the following integral formula for the solutions to the Poisson equation on all of three-dimensional space.

Theorem 18.7. *A particular solution to the Poisson equation*

$$-\Delta u = f \qquad \text{for} \qquad \mathbf{x} \in \mathbb{R}^3 \tag{18.63}$$

is given by

$$u_\star(x,y,z) = \frac{1}{4\pi} \iiint_{\mathbb{R}^3} \frac{f(\xi,\eta,\zeta)\; d\xi\, d\eta\, d\zeta}{\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}}\,. \tag{18.64}$$

The general solution to the Poisson equation is

$$u(x,y,z) = u_\star(x,y,z) + w(x,y,z),$$

where $w(x,y,z)$ is an arbitrary harmonic function.

Example 18.8. In this example, we compute the gravitational (or electrostatic) potential in three-dimensional space due to a uniform solid ball, e.g., a spherical planet such as the earth. By rescaling, it suffices to consider the case when the forcing function

$$f(\mathbf{x}) = \begin{cases} 1, & \| \mathbf{x} \| < 1, \\ 0, & \| \mathbf{x} \| > 1, \end{cases}$$

is equal to 1 inside a solid ball of radius 1 and zero outside. The particular solution to the resulting Poisson equation (18.63) is given by the integral

$$u^\star(\mathbf{x}) = \frac{1}{4\pi} \iiint_{\| \boldsymbol{\xi} \| < 1} \frac{1}{\| \mathbf{x} - \boldsymbol{\xi} \|}\, d\xi\, d\eta\, d\zeta. \tag{18.65}$$

Clearly, since the forcing function is radially symmetric, the solution $u = u(r)$ is also radially symmetric. To evaluate the integral, then, we can take $\mathbf{x} = (0,0,z)$ to lie on the z axis, so that $r = \| \mathbf{x} \| = |\, z \,|$. We use cylindrical coordinates $\boldsymbol{\xi} = (\rho\cos\theta, \rho\sin\theta, \zeta)$, so that

$$\| \mathbf{x} - \boldsymbol{\xi} \| = \sqrt{\rho^2 + (z-\zeta)^2}\,.$$

See Figure Psph■. The integral in (18.65) can then be explicitly computed:

$$\frac{1}{4\pi} \int_{-1}^{1} \int_{0}^{\sqrt{1-\zeta^2}} \int_{0}^{2\pi} \frac{\rho\, d\theta\, d\rho\, d\zeta}{\sqrt{\rho^2 + (z-\zeta)^2}} \;=$$

$$= \frac{1}{2} \int_{-1}^{1} \Big(\sqrt{1 + z^2 - 2z\zeta} \;-\; |\, z - \zeta \,| \Big)\, d\zeta = \begin{cases} \dfrac{1}{3\,|\, z \,|}\,, & |\, z \,| \geq 1, \\ -\dfrac{z^2}{6} + \dfrac{1}{2}\,, & |\, z \,| \leq 1. \end{cases}$$

Therefore, by radial symmetry, the solution is

$$u(\mathbf{x}) = \begin{cases} \dfrac{1}{3r}\,, & r = \| \mathbf{x} \| \geq 1, \\ -\dfrac{r^2}{6} + \dfrac{1}{2}\,, & r = \| \mathbf{x} \| \leq 1, \end{cases} \tag{18.66}$$

plotted, as a function of $r = \| \mathbf{x} \|$ in Figure solidball■. Note that, outside the solid ball, the solution is a Newtonian potential corresponding to a point mass of magnitude $\frac{4}{3}\pi$, which is the same as the total mass of the planet. This is a well-known result in gravitation and electrostatics — the exterior potential due to a spherically symmetric mass (or electric charge) is the same as if all its mass were concentrated at its center. Thus, in outer space if you can't see a spherically symmetric planet, you can only determine its mass, not its size, by measuring the gravitational force. Interestingly, at the center of the ball, the potential is equal to $\frac{1}{2}$, not zero, which is its asymptotic value at large distances.

Bounded Domains and the Method of Images

Suppose we now wish to solve the inhomogeneous Poisson equation (18.53) on a bounded domain $\Omega \subset \mathbb{R}^3$. The spatial Green's function (18.62) is a particular solution to the underlying inhomogeneous equation

$$-\Delta u = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad \mathbf{x} \in \Omega, \tag{18.67}$$

but it does not have the proper boundary values on $\partial\Omega$. However, as we know by the principles of linearity, the general solution to any inhomogeneous linear equation has the form

$$u(\mathbf{x}) = \frac{1}{4\pi \,\| \mathbf{x} - \boldsymbol{\xi} \|} \; - \; v(\mathbf{x}), \tag{18.68}$$

where the first summand is a particular solution, which we now know, while $v(\mathbf{x})$ is an arbitrary solution to the homogeneous equation $\Delta v = 0$, i.e., an arbitrary harmonic function. The minus sign is for later convenience. The solution (18.68) satisfies the homogeneous boundary conditions provided the boundary values of $v(\mathbf{x})$ match those of the Green's function. Let us state the result in the case of the Dirichlet boundary value problem.

Theorem 18.9. *The Green's function for the homogeneous Dirichlet boundary value problem for the Poisson equation*

$$-\Delta u = f, \qquad \mathbf{x} \in \Omega, \qquad u = 0, \qquad \mathbf{x} \in \partial\Omega,$$

in a domain $\Omega \subset \mathbb{R}^3$ has the form

$$G(\mathbf{x}; \boldsymbol{\xi}) = \frac{1}{4\pi \,\| \mathbf{x} - \boldsymbol{\xi} \|} \; - \; v(\mathbf{x}; \boldsymbol{\xi}) \tag{18.69}$$

where $v(\mathbf{x}; \boldsymbol{\xi})$ is the harmonic function of $\mathbf{x}$ that satisfies

$$v(\mathbf{x}; \boldsymbol{\xi}) = \frac{1}{4\pi \,\| \mathbf{x} - \boldsymbol{\xi} \|} \qquad \text{for all} \qquad \mathbf{x} \in \partial\Omega.$$

In this manner, we have reduced the determination of the Green's function to the solution to a particular set of Laplace boundary value problems, parametrized by the point $\boldsymbol{\xi} \in \Omega$. In certain cases, the method of images will produce an explicit formula for the Green's function. As in the planar version presented in Section 15.3, the idea is to match the boundary values of the Green's function due to a delta impulse at a point inside

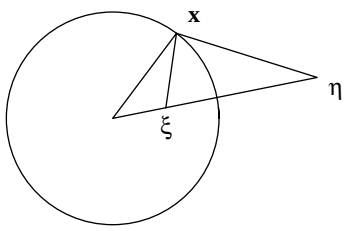


Figure 18.2. Method of Images for the Unit Sphere.

the domain with one or more Green's functions corresponding to impulses at points outside the domain — the "image points".

The case of a solid ball of radius 1 with Dirichlet boundary conditions is the easiest to handle. Indeed, the *same* geometrical construction that we used for a disk in the plane, and illustrated in Figure 18.2 applies to a solid ball in three-dimensional space. Although this is the same as Figure 15.8, we are now interpreting the picture as a three-dimensional diagram, and so the circle represents the unit sphere. We choose the image point given by *inversion*:

$$\boldsymbol{\eta} = \frac{\boldsymbol{\xi}}{\| \boldsymbol{\xi} \|^2}, \qquad \text{so that} \qquad \| \boldsymbol{\xi} \| = \frac{1}{\| \boldsymbol{\eta} \|}.$$

Applying the same similar triangles argument as in the planar case, we deduce that

$$\frac{\| \boldsymbol{\xi} \|}{\| \mathbf{x} \|} = \frac{\| \mathbf{x} \|}{\| \boldsymbol{\eta} \|} = \frac{\| \mathbf{x} - \boldsymbol{\xi} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|}, \qquad \text{and therefore} \qquad \| \mathbf{x} \| = 1.$$

As a result, the function

$$v(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{4\pi} \, \frac{\| \boldsymbol{\eta} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|} = \frac{1}{4\pi} \, \frac{\| \boldsymbol{\xi} \|}{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|}$$

has the same boundary values on the unit sphere as the free space Green's function:

$$\frac{1}{4\pi} \, \frac{\| \boldsymbol{\eta} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|} = \frac{1}{4\pi \| \mathbf{x} - \boldsymbol{\xi} \|} \qquad \text{whenever} \qquad \| \mathbf{x} \| = 1.$$

We conclude that the difference (18.69) between the two

$$G(\mathbf{x}; \boldsymbol{\xi}) = \frac{1}{4\pi} \left(\frac{1}{\| \mathbf{x} - \boldsymbol{\xi} \|} - \frac{\| \boldsymbol{\xi} \|}{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|} \right) \tag{18.70}$$

has the required properties of the Green's function: it satisfies the Laplace equation inside the unit ball except at the singularity at $\mathbf{x} = \boldsymbol{\xi}$, while $G(\mathbf{x}; \boldsymbol{\xi}) = 0$ has homogeneous Dirichlet conditions on the boundary $\| \mathbf{x} \| = 1$.

With the Green's function in hand, we can apply the general superposition formula (18.55) to arrive at a general formula for the solution to the Dirichlet boundary value problem for the Poisson equation in the unit ball.

Theorem 18.10. *The solution $u(\mathbf{x})$ to the homogeneous Dirichlet boundary value problem*

$$-\Delta u = f, \qquad \| \mathbf{x} \| < 1, \qquad\qquad u = 0, \qquad \| \mathbf{x} \| = 1$$

is given by the integral

$$u(\mathbf{x}) = \frac{1}{4\pi} \iiint_{\| \boldsymbol{\xi} \| \le 1} \left(\frac{1}{\| \mathbf{x} - \boldsymbol{\xi} \|} - \frac{\| \boldsymbol{\xi} \|}{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \mathbf{x} \|} \right) f(\boldsymbol{\xi}) \, d\xi \, d\eta \, d\zeta. \tag{18.71}$$

Example 18.11. In this example, we compute the electrostatic potential inside a sphere due to a small solid ball at its center. The outside sphere $\| \mathbf{x} \| = 1$ is assumed to be grounded, and so the potential satisfies the homogeneous Dirichlet boundary conditions there. The forcing function due to the interior charged sphere is take in the form

$$f(\mathbf{x}) = \begin{cases} 1, & \| \mathbf{x} \| < \rho, \\ 0, & \rho < \| \mathbf{x} \| < 1. \end{cases}$$

Using radial symmetry, the solution $u = u(r)$ is also radially symmetric. ■■

The Green's function can also be used to solve the inhomogeneous boundary value problem

$$-\Delta u = 0, \qquad \mathbf{x} \in \Omega, \qquad\qquad u = h, \qquad \mathbf{x} \in \partial\Omega. \tag{18.72}$$

The same argument as we applied in the two-dimensional situation works here, and the solution is

$$u(\mathbf{x}) = - \iint_{\partial\Omega} \frac{\partial G(\mathbf{x}; \boldsymbol{\xi})}{\partial \mathbf{n}} \, h(\boldsymbol{\xi}) \, dS. \tag{18.73}$$

In the case when Ω is a solid ball, this integral formula effectively sums the spherical harmonic series (18.43).

18.4. The Heat Equation in Three-Dimensional Media.

Thermal diffusion in a homogeneous solid body $\Omega \subset \mathbb{R}^3$ is governed by the three-dimensional variant of the heat equation

$$\frac{\partial u}{\partial t} = \gamma \, \Delta u = \gamma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right), \qquad\qquad (x, y, z) \in \Omega, \tag{18.74}$$

The coefficient $\gamma > 0$ measures the thermal diffusivity of the body. Positivity of the diffusivity is required in order that the heat equation be well-posed; see Section 14.1 for details. The physical derivation of the heat equation is exactly the same as the two-dimensional version (17.1), and does not need to be repeated in detail. Briefly, the temperature gradient is proportional to the heat flux vector, $\mathbf{w} = -\kappa \nabla u$, while its divergence is proportional to the rate of change of temperature, $\sigma u_t = -\nabla \cdot \mathbf{w}$. Combining these two physical laws and assuming homogeneity, whereby κ and σ are constant, produces (18.74) with $\gamma = \kappa/\sigma$.

As always, we need to impose suitable boundary conditions. These are either the Dirichlet conditions $u = h$ that specify the boundary temperature, or homogeneous Neumann conditions $\partial u/\partial \mathbf{n} = 0$ corresponding to an insulated boundary, or a mixture of the two. Given the initial temperature of the body

$$u(t_0, x, y, z) = f(x, y, z) \tag{18.75}$$

at the initial time t_0, there is a unique solution $u(t, x, y, z)$ to the initial-boundary value problem for all subsequent times $t \geq t_0$; see [**40**] for a proof.

To keep matters reasonably simple, we initially restrict our attention to the homogeneous boundary conditions. The general separation of variables method works as before. One begins by imposing an exponential ansatz $u(t, \mathbf{x}) = e^{-\lambda t}\, v(\mathbf{x})$. Substituting into the differential equation and canceling the exponentials, we deduce that v satisfies the Helmholtz eigenvalue problem

$$\gamma\, \Delta v + \lambda\, v = 0,$$

subject to the relevant boundary conditions. For Dirichlet and mixed boundary conditions, the Laplacian is a positive definite operator, and hence the eigenvalues are all strictly positive,

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots, \qquad \text{with} \qquad \lambda_n \longrightarrow \infty,$$

as $n \to \infty$. Linear superposition implies that the solution can be written as a generalized Fourier series

$$u(t, \mathbf{x}) = \sum_{n=1}^{\infty} c_n\, e^{-\lambda_n\, t}\, v_n(\mathbf{x}) \tag{18.76}$$

in the corresponding eigenfunctions $v_n(\mathbf{x})$. The coefficients c_n are uniquely prescribed by the initial condition (18.75); for $t_0 = 0$, the initial condition takes the form

$$u(0, \mathbf{x}) = \sum_{n=1}^{\infty} c_n\, v_n(\mathbf{x}) = f(\mathbf{x}). \tag{18.77}$$

Self-adjointness of the boundary value problem implies that the eigenfunctions are mutually orthogonal, and hence we can invoke the usual orthogonality formulae

$$c_n = \frac{\langle\, f\,;\, v_n\,\rangle}{\|\, v_n\,\|^2} = \frac{\displaystyle\iiint_\Omega f(\mathbf{x})\, v_n(\mathbf{x})\, dx\, dy\, dz}{\displaystyle\iiint_\Omega v_n(\mathbf{x})^2\, dx\, dy\, dz} \tag{18.78}$$

in order to compute the Fourier coefficients. Since the higher modes — the terms for $n \gg 0$ — go to zero extremely rapidly, the solution can be well approximated by the first few terms in its Fourier expansion. As a consequence, the heat equation rapidly smoothes out discontinuities and noise in the initial data, and so can be used to denoise three-dimensional and video images — although better nonlinear techniques are now available, [**128**]. The solution $u(t, \mathbf{x})$ decays exponentially fast to thermal equilibrium $u(t, \mathbf{x}) \to 0$, the same temperature as imposed on (part of) the boundary, at a rate equal to the smallest positive eigenvalue $\lambda_1 > 0$.

Unfortunately, the explicit formulae for the eigenfunctions and eigenvalues are known only for a few particular domains, [**108**]. Most explicit solution techniques for the Helmholtz boundary value problem rely on a further separation of variables. In a rectangular box, one separates into a product of functions depending upon the individual Cartesian coordinates, and the eigenfunctions are written as products of trigonometric and hyperbolic functions. See Exercise ■ for details. In a cylindrical domain, the separation is effected in cylindrical coordinates, and leads to separable solutions in terms of trigonometric and Bessel functions, as outlined in Exercise ■. The most interesting and enlightening case is a spherical domain, and we treat this particular problem in complete detail.

Heating of a Ball

Let us solve the problem of heat propagation in a solid spherical body, e.g., the earth[†]. For simplicity, we take the diffusivity $\gamma = 1$, and consider the heat equation on a solid spherical domain $B_1 = \{ \| \mathbf{x} \| < 1 \}$ of radius 1 subject to homogeneous Dirichlet boundary conditions. Once we know how to solve this particular case, an easy scaling argument outlined in Exercise ■ will allow us to find the solution for a ball of arbitrary radius and with a general diffusion coefficient.

As usual, when dealing with a spherical geometry, we adopt spherical coordinates r, φ, θ, (B.64), in terms of which the heat equation takes the form

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{r^2 \sin\varphi}\,\frac{\partial u}{\partial \varphi} + \frac{1}{r^2 \sin^2\varphi}\,\frac{\partial^2 u}{\partial \theta^2}\,, \tag{18.79}$$

where we have used our handy formula (18.14) for the Laplacian in spherical coordinates. The diffusive separation of variables ansatz $u(t, r, \varphi, \theta) = e^{-\lambda t}\, v(r, \varphi, \theta)$ requires us to analyze the Helmholtz equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{r^2 \sin\varphi}\,\frac{\partial u}{\partial \varphi} + \frac{1}{r^2 \sin^2\varphi}\,\frac{\partial^2 u}{\partial \theta^2} + \lambda\, u = 0 \tag{18.80}$$

on the unit ball $\Omega = \{ r < 1 \}$ with homogeneous Dirichlet boundary conditions. To solve the spherical coordinate form of the Helmholtz equation, we invoke a further separation of variables. To this end, we separate off the radial coordinate first by setting

$$v(r, \varphi, \theta) = w(r)\, \chi(\varphi, \theta).$$

The function χ must be 2π periodic in θ and well-defined at the poles $\varphi = 0, \pi$. Substituting this ansatz in (18.80), and separating all the r-dependent terms from the terms depending upon the angular variables φ, θ leads to a pair of differential equations; the first is an ordinary differential equation

$$r^2\, w'' + 2\, w' + (\lambda\, r^2 - \mu) w = 0, \tag{18.81}$$

[†] In this perhaps overly simplified model, we are assuming that the earth is composed of a completely homogeneous and isotropic solid material.

for the radial component $w(r)$, while the second is a familiar partial differential equation

$$\Delta_S \chi + \mu \chi = \frac{1}{\sin\varphi} \frac{\partial}{\partial \varphi} \left(\sin\varphi \frac{\partial \chi}{\partial \varphi} \right) + \frac{1}{\sin^2 \varphi} \frac{\partial^2 \chi}{\partial \theta^2} + \mu \chi = 0, \tag{18.82}$$

for its angular counterpart $\chi(\varphi, \theta)$. The operator Δ_S is the spherical Laplacian (18.17) analyzed in Section 18.2. As we learned, its eigenvalues have the form $\mu_n = n(n+1)$ for $n = 0, 1, 2, 3, \ldots$. Each eigenvalue admits $2n+1$ linearly independent eigenfunctions — the spherical harmonics $Y_n^m, \widetilde{Y}_n^m$ defined in (18.32).

The radial ordinary differential equation (18.81) can be solved by letting

$$p(r) = \sqrt{r}\, w(r).$$

We manually compute the derivatives

$$w = \frac{1}{\sqrt{r}}\, p, \qquad \frac{dw}{dr} = \frac{1}{\sqrt{r}} \frac{dp}{dr} - \frac{1}{2\, r^{3/2}}\, p, \qquad \frac{d^2 w}{dr^2} = \frac{1}{\sqrt{r}} \frac{d^2 p}{dr^2} - \frac{1}{2\, r^{3/2}} \frac{dp}{dr} + \frac{3}{4\, r^{5/2}}\, p.$$

Substituting into (18.81) with $\mu = \mu_n = n(n+1)$, and multiplying the resulting equation by $\sqrt{r}$, we discover that $p(r)$ must solve the differential equation

$$r^2 \frac{d^2 p}{dr^2} + r \frac{dp}{dr} + \left[\lambda\, r^2 - \left(n + \tfrac{1}{2} \right)^2 \right] p = 0. \tag{18.83}$$

The latter equation is identical to the rescaled Bessel equation (17.49) in which the order $m = n + \frac{1}{2}$ is a half integer, i.e., $m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$. Therefore, the solution to (18.83) that remains bounded at $r = 0$ is (up to scalar multiple) the rescaled Bessel function

$$p(r) = J_{n+1/2}\big(\sqrt{\lambda}\, r \big).$$

The corresponding solution

$$w(r) = r^{-1/2}\, J_{n+1/2}\big(\sqrt{\lambda}\, r \big) \tag{18.84}$$

to (18.81) is important enough to warrant a special name.

Definition 18.12. The *spherical Bessel function* of order $n \geq 0$ is defined by the formula

$$S_n(x) = \sqrt{\frac{\pi}{2\, x}}\; J_{n+1/2}(x) \tag{18.85}$$

involving the Bessel function of half integer order. The multiplicative factor $\sqrt{\frac{\pi}{2}}$ is included in the definition so as to avoid annoying factors of $\sqrt{\pi}$ and $\sqrt{2}$ in all subsequent formulae.

Surprisingly, unlike the Bessel functions of integer order, the spherical Bessel functions are elementary functions! According to formula (C.55), the spherical Bessel function of order 0 is

$$S_0(x) = \frac{\sin x}{x}\,. \tag{18.86}$$

The higher order spherical Bessel functions can be obtained by use of a general recurrence relation

$$S_{n+1}(x) = -\frac{dS_n}{dx} + \frac{n}{x}\, S_n(x), \tag{18.87}$$

which is a consequence of Proposition C.13. The next few are, therefore,

$$\begin{aligned} S_1(x) &= -\frac{dS_0}{dx} = -\frac{\cos x}{x} + \frac{\sin x}{x^2}\,, \\ S_2(x) &= -\frac{dS_1}{dx} + \frac{S_1}{x} = -\frac{\sin x}{x} - \frac{3\cos x}{x^2} + \frac{3\sin x}{x^3}\,, \\ S_3(x) &= -\frac{dS_2}{dx} + \frac{2\,S_1}{x}\frac{\cos x}{x} - \frac{6\sin x}{x^2} - \frac{15\cos x}{x^3} + \frac{15\sin x}{x^4}\,. \end{aligned} \tag{18.88}$$

Our radial solution (18.84) is, apart from an inessential constant multiple that we ignore, a rescaled spherical Bessel function of order n:

$$w_n(r) = S_n\big(\sqrt{\lambda}\, r\,\big)\,.$$

So far, we have not taken into account the homogeneous Dirichlet boundary condition at $r = 1$. This requires

$$w_n(1) = 0, \qquad \text{and hence} \qquad S_n\big(\sqrt{\lambda}\,\big) = 0.$$

Therefore, $\sqrt{\lambda}$ must be a root of the n^{th} order spherical Bessel function. We use the notation

$$0 < \sigma_{1,n} < \sigma_{2,n} < \sigma_{3,n} < \cdots$$

to denote the successive roots of the n^{th} order spherical Bessel function, so that

$$S_n(\sigma_{k,n}) = 0 \qquad \text{for} \qquad k = 1, 2, 3, \dots\,.$$

In particular the roots of the zero$^{\text{th}}$ order function $S_0(x) = \sin x/x$ are just the integer multiples of π, so

$$\sigma_{k,0} = k\,\pi \qquad \text{for} \qquad k = 1, 2, \ \dots\ .$$

A table of all spherical Bessel roots that are < 13 follows. The rows of the table are indexed by n, the order, while the columns are indexed by k, the root number.

Re-assembling the individual pieces, we have now demonstrated that the separable eigenfunctions of the Helmholtz equation on a solid ball of radius 1, when subject to homogeneous Dirichlet boundary conditions, are products of spherical Bessel functions and spherical harmonics,

$$v_{k,m,n}(r,\varphi,\theta) = S_n(\sigma_{k,n}\, r)\, Y_n^m(\varphi,\theta), \qquad \widetilde{v}_{k,m,n}(r,\varphi,\theta) = S_n(\sigma_{k,n}\, r)\, \widetilde{Y}_n^m(\varphi,\theta). \tag{18.89}$$

The corresponding eigenvalues

$$\lambda_{k,n} = \sigma_{k,n}^2, \qquad n = 0, 1, 2, \dots, \qquad k = 1, 2, 3, \dots\,, \tag{18.90}$$

Table of Spherical Bessel Roots $\sigma_{k,n}$

	0	1	2	3	...
1	3.1416	6.2832	9.4248	12.5664	...
2	4.4934	7.7253	10.9041	⋮	
3	5.7635	9.0950	12.3229	⋮	
4	8.1826	11.7049	⋮		
5	9.3558	12.9665	⋮		
6	10.5128	⋮			
7	11.6570	⋮			
8	12.7908	⋮			
⋮	⋮				

are given by the squared spherical Bessel roots. Since there are $2n+1$ independent spherical harmonics of order n, each eigenvalue $\lambda_{k,n}$ admits $2n+1$ linearly independent eigenfunctions, namely $v_{k,0,n}, \dots, v_{k,n,n}, \widetilde{v}_{k,1,n}, \dots, v_{k,n,n}$. (We omit the trivial case $\widetilde{v}_{k,0,n} \equiv 0$.) In particular, the radially symmetric solutions are the eigenfunctions for $m = n = 0$, namely

$$v_k(r) = v_{k,0,0}(r) = S_0(\sigma_{k,0}\, r) = \frac{\sin k\pi\, r}{k\pi\, r}, \qquad\qquad k = 1, 2, \dots . \tag{18.91}$$

It can be shown that the separable solutions (18.89) form a complete system of eigenfunctions, [**39**].

We have thus completely determined the basic separable solutions to the heat equation on a solid unit ball subject to homogeneous Dirichlet boundary conditions. They are products of exponential functions of time, spherical Bessel functions of the radius and the spherical harmonics:

$$\begin{aligned} u_{k,m,n}(t, r, \varphi, \theta) &= e^{-\sigma_{k,n}^2\, t}\; S_n(\sigma_{k,n}\, r)\, Y_n^m(\varphi, \theta), \\ \widetilde{u}_{k,m,n}(t, r, \varphi, \theta) &= e^{-\sigma_{k,n}^2\, t}\; S_n(\sigma_{k,n}\, r)\, \widetilde{Y}_n^m(\varphi, \theta). \end{aligned} \tag{18.92}$$

The general solution can be written as an infinite "Fourier–Bessel–spherical harmonic" series in these fundamental modes

$$\begin{aligned} &u(t, r, \theta, \varphi) = \\ &= \sum_{n=0}^{\infty} \sum_{k=1}^{\infty} e^{-\sigma_{k,n}^2\, t}\; S_n(\sigma_{k,n}\, r) \left(\frac{c_{0,n}}{2}\, Y_n^0(\varphi, \theta) + \sum_{m=1}^{n} \Big[c_{m,n} Y_n^m(\varphi, \theta) + \widetilde{c}_{m,n} \widetilde{Y}_n^m(\varphi, \theta) \Big] \right). \end{aligned} \tag{18.93}$$

The series coefficients are uniquely prescribed by the initial data, based on the usual orthogonality relations among the eigenfunctions. Detailed formulae are relegated to the

exercises. In particular, the slowest decaying mode is the spherically symmetric function

$$u_{k,0,0}(t,r) = \frac{e^{-\pi^2 t}\, \sin \pi r}{r} \tag{18.94}$$

corresponding to the smallest eigenvalue $\lambda_{1,0} = \pi^2$. Therefore, the overall decay rate to thermal equilibrium of a unit sphere is at a rate equal to $\pi^2 \approx 9.8696$, or, to a very rough approximation, 10.

The Fundamental Solution to the Heat Equation

For the heat equation (as well as more general diffusion equations), the fundamental solution measures the response of the body to a concentrated unit heat source. Thus, given a point $\boldsymbol{\xi} = (\xi, \eta, \zeta) \in \Omega$ in the domain occupied by the body, the fundamental solution

$$u(t, \mathbf{x}) = F(t, \mathbf{x}; \boldsymbol{\xi}) = F(t, x, y, z; \xi, \eta, \zeta)$$

solves the initial-boundary value problem

$$u_t = \Delta u, \qquad u(0, \mathbf{x}) = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad \text{for} \qquad \mathbf{x} \in \Omega, \quad t > 0, \tag{18.95}$$

subject to the homogeneous boundary conditions of the required form — which can be either Dirichlet, Neumann or mixed.

In general, there is no explicit formula for the fundamental solution, although in certain domains one can construct a (generalized) Fourier series in the associated eigenfunctions. The one case amenable to a complete analysis is when the heat is distributed over all of three-dimensional space, so $\Omega = \mathbb{R}^3$. To this end, we recall that Lemma 17.2 showed how to construct solutions of the two-dimensional heat equation as products of one-dimensional solutions. In a similar manner, if $v(t,x)$, $w(t,x)$ and $q(t,x)$ are any three solutions to $u_t = \gamma\, u_{xx}$, then the product

$$u(t,x,y) = v(t,x)\, w(t,y)\, q(t,z) \tag{18.96}$$

is a solution to the three-dimensional heat equation $u_t = \gamma\, (u_{xx} + u_{yy} + u_{zz})$. In particular, choosing

$$v(t,x) = \frac{1}{2\sqrt{\pi\, \gamma\, t}}\, e^{-(x-\xi)^2/4\gamma t}, \qquad w(t,y) = \frac{1}{2\sqrt{\pi\, \gamma\, t}}\, e^{-(y-\eta)^2/4\gamma t},$$
$$q(t,z) = \frac{1}{2\sqrt{\pi\, \gamma\, t}}\, e^{-(z-\zeta)^2/4\gamma t},$$

to all be one-dimensional fundamental solutions, we are immediately led to the three-dimensional fundamental solution in the form of a three-dimensional Gaussian kernel.

Theorem 18.13. *The fundamental solution*

$$F(t, \mathbf{x}; \boldsymbol{\xi}) = F(t, \mathbf{x} - \boldsymbol{\xi}) = \frac{e^{-\| \mathbf{x} - \boldsymbol{\xi} \|^2/4\gamma t}}{8\, (\pi\, \gamma\, t)^{3/2}} \tag{18.97}$$

solves the three-dimensional heat equation $u_t = \gamma\, \Delta u$ on $\mathbb{R}^3$ with an initial temperature equal to a delta function concentrated at the point $\mathbf{x} = \boldsymbol{\xi}$.

Thus, the initially concentrated heat energy immediately begins to spread out in a radially symmetric manner, with a minuscule, but nonzero effect felt at arbitrarily large distances away from the initial concentration. At each individual point $\mathbf{x} \in \mathbb{R}^3$, after an initial warm-up, the temperature decays back to zero at a rate proportional to $t^{-3/2}$ — even more rapidly than in two dimensions because, intuitively, there are more directions for the heat energy to disperse.

To solve the general initial value problem with the initial temperature $u(0, x, y, z) = f(x, y, z)$ distributed over all of space, we first write

$$f(x, y, z) = \iiint f(\boldsymbol{\xi})\, \delta(\mathbf{x} - \boldsymbol{\xi})\, d\xi\, d\eta\, d\zeta$$

as a linear superposition of delta functions. By linearity, the solution to the initial value problem is given by the corresponding superposition

$$u(t, \mathbf{x}) = \frac{1}{8\,(\pi\,\gamma\,t)^{3/2}} \iiint f(\boldsymbol{\xi})\, e^{-\,\|\,\mathbf{x}-\boldsymbol{\xi}\,\|^2/4\,\gamma\,t}\, d\xi\, d\eta\, d\zeta. \tag{18.98}$$

of the fundamental solutions. Since the fundamental solution has exponential decay as $\|\,\mathbf{x}\,\| \to \infty$, the superposition formula is valid even for initial temperature distributions which are moderately increasing at large distances. We remark that the integral (18.98) has the form of a three-dimensional convolution

$$u(t, \mathbf{x}) = F(t, \mathbf{x}) * f(\mathbf{x}) = \iiint f(\boldsymbol{\xi})\, F(t, \mathbf{x} - \boldsymbol{\xi})\, d\xi\, d\eta\, d\zeta \tag{18.99}$$

of the initial data with a one-parameter family of increasingly spread out Gaussian filters. Thus, convolution with a Gaussian kernel has the same smoothing effect on functions.

Example 18.14. ∎

More general situations must be solved by numerical integration and approximation.

18.5. The Wave Equation in Three-Dimensional Media.

Certain classes of vibrations of a uniform solid body are governed by the three-dimensional wave equation

$$u_{tt} = c^2 \Delta u = c^2 (u_{xx} + u_{yy} + u_{zz}). \tag{18.100}$$

The solution $u(t, \mathbf{x}) = u(t, x, y, z)$ represents a scalar-valued displacement of the body at time t and position $\mathbf{x} = (x, y, z) \in \Omega \subset \mathbb{R}^3$. For example, $u(t, \mathbf{x})$ might represent the radial displacement of the body. One imposes suitable boundary conditions, e.g., Dirichlet, Neumann or mixed, on $\partial\Omega$, along with a pair of initial conditions

$$u(0, \mathbf{x}) = f(\mathbf{x}), \qquad \frac{\partial u}{\partial t}\,(0, \mathbf{x}) = g(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{18.101}$$

that specify the initial displacement and initial velocity of the body. As long as the initial and boundary data are reasonably nice, there exists a unique solution to the initial-boundary value problem for all $-\infty < t < \infty$. Thus, in contrast to the heat equation, one can follow solutions to the wave equation backwards in time; see also Exercise ∎.

Remark: Since the solution $u(t, \mathbf{x})$ to the wave equation is scalar-valued, it cannot measure the full range of possible three-dimensional motions of a solid body. The more complicated dynamical systems governing the elastic motions of solids are discussed in Exercise ■.

Remark: The wave equation also governs the propagation of electromagnetic waves, such as light, radio, X-rays, etc., in a homogeneous medium, including (in the absence of gravitational effects) empty space. Each individual component of the electric and magnetic vector fields $\mathbf{E}, \mathbf{B}$ satisfy the wave equation, in which c denotes the velocity of light; see Exercise ■ for details.

We initially concentrate on the homogeneous boundary value problem. The fundamental vibrational modes are found by imposing our usual trigonometric ansatz

$$u(t, x, y, z) = \cos \omega\, t\; v(x, y, z).$$

Substituting into the wave equation (18.100), we discover (yet again) that $v(x, y, z)$ must be an eigenfunction solving the associated Helmholtz eigenvalue problem

$$\Delta v + \lambda\, v = 0, \qquad \text{where} \qquad \lambda = \frac{\omega^2}{c^2}\,, \tag{18.102}$$

along with the relevant boundary conditions. In the positive definite cases, i.e., Dirichlet and mixed boundary conditions, the eigenvalues $\lambda_k = \omega_k^2/c^2 > 0$ are all positive. Each eigenfunction $v_k(x, y, z)$ yields two vibrational solutions

$$u_k(t, x, y, z) = \cos \omega_k\, t\; v_k(x, y, z), \qquad \widetilde{u}_k(t, x, y, z) = \sin \omega_k\, t\; v_k(x, y, z),$$

of frequency $\omega_k = c\sqrt{\lambda_k}$ equal to the square root of the corresponding eigenvalue. The general solution is a quasi-periodic linear combination

$$u(t, x, y, z) = \sum_{k=1}^{\infty} \big(a_k \cos \omega_k t + b_k \sin \omega_k t \,\big)\, v_k(x, y, z) \tag{18.103}$$

of these fundamental vibrational modes. The coefficients a_k, b_k are uniquely prescribed by the initial conditions (18.101). Thus,

$$u(0, x, y, z) = \sum_{k=1}^{\infty} a_k\, v_k(x, y, z) = f(x, y, z),$$
$$\frac{\partial u}{\partial t}\,(0, x, y, z) = \sum_{k=1}^{\infty} \omega_k\, b_k\, v_k(x, y, z) = g(x, y, z).$$

The explicit formulas follow immediately from the mutual orthogonality of the eigenfunctions:

$$a_k = \frac{\langle\, f\,;\, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iiint_\Omega f\, v_k\; dx\, dy\, dz}{\displaystyle\iiint_\Omega v_k^2\; dx\, dy\, dz}\,, \qquad b_k = \frac{1}{\omega_k}\, \frac{\langle\, g\,;\, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iiint_\Omega g\, v_k\, dx\, dy\, dz}{\omega_k \displaystyle\iiint_\Omega v_k^2\; dx\, dy\, dz}\,. \tag{18.104}$$

In the positive semi-definite Neumann boundary value problem, there is an additional zero eigenvalue $\lambda_0 = 0$ corresponding to the constant null eigenfunction $v_0(x, y, z) \equiv 1$. This results in two additional terms in the eigenfunction expansion — a constant term

$$a_0 = \frac{1}{\operatorname{vol}\Omega} \iiint_\Omega f(x, y, z)\, dx\, dy\, dz$$

that equals the average initial displacement, and an unstable mode $b_0\, t$ that grows linearly in time, whose speed

$$b_0 = \frac{1}{\operatorname{vol}\Omega} \iiint_\Omega g(x, y, z)\, dx\, dy\, dz$$

is the average of the initial velocity over the entire body. The unstable mode will be excited if and only if there is a non-zero net initial velocity, $b_0 \neq 0$.

Most of the basic solution techniques we learned in the two-dimensional case apply here, and we will not dwell on the details. The case of a rectangular box is a particularly straightforward application of the method of separation of variables, and is outlined in the exercises. A similar analysis, now in cylindrical coordinates, can be applied to the case of a vibrating cylinder. The most interesting case is that of a solid spherical ball, which is the subject of the next subsection.

Vibrations of a Ball

Let us focus on the radial vibrations of a solid ball, as modeled by the three-dimensional wave equation (18.100). The solution $u(t, \mathbf{x})$ represents the radial displacement of the particle that is situated at position $\mathbf{x}$ when the ball is at rest.

For simplicity, we look at the Dirichlet boundary value problem on a ball of radius 1. The normal modes of vibration are governed by the Helmholtz equation (18.102) on $B_1 = \{ \| \mathbf{x} \| < 1 \}$ subject to homogeneous Dirichlet boundary conditions. According to (18.89), the eigenfunctions are

$$\begin{aligned} v_{k,m,n}(r, \varphi, \theta) &= S_n(\sigma_{k,n}\, r)\, Y_n^m(\varphi, \theta), \\ \widetilde{v}_{k,m,n}(r, \varphi, \theta) &= S_n(\sigma_{k,n}\, r)\, \widetilde{Y}_n^m(\varphi, \theta), \end{aligned} \qquad \begin{aligned} &k = 1, 2, 3, \dots\ , \\ &m = 0, 1, 2, \dots\ , \\ &n = 0, \dots, m. \end{aligned} \tag{18.105}$$

Here S_n denotes the n^{th} order spherical Bessel function (18.85), $\sigma_{k,n}$ is its k^{th} root, while $Y_n^m, \widetilde{Y}_n^m$ are the spherical harmonics (18.32). Each eigenvalue

$$\lambda_{k,n} = \sigma_{k,n}^2, \qquad n = 0, 1, 2, \dots, \qquad k = 1, 2, 3, \dots,$$

corresponds to $2n + 1$ independent eigenfunctions, namely

$$v_{k,0,n}(r, \varphi, \theta),\ v_{k,1,n}(r, \varphi, \theta),\ \dots\ v_{k,n,n}(r, \varphi, \theta),\ \widetilde{v}_{k,1,n}(r, \varphi, \theta),\ \dots\ \widetilde{v}_{k,n,n}(r, \varphi, \theta),$$

where we discard the trivial case $\widetilde{v}_{k,0,n}(r, \varphi, \theta) \equiv 0$. As a consequence, the fundamental vibrational frequencies of a solid ball

$$\omega_{k,n} = c\, \sqrt{\lambda_{k,n}}\ = c\, \sigma_{k,n}, \qquad n = 0, 1, 2, \dots, \qquad k = 1, 2, 3, \dots, \tag{18.106}$$

are equal to the spherical Bessel roots $\sigma_{k,n}$ multiplied by the wave speed. There are a total of $2(2n+1)$ independent vibrational modes associated with each distinct frequency (18.106), namely

$$\begin{aligned}
u_{k,m,n}(t,r,\varphi,\theta) &= \cos(c\,\sigma_{k,n}\,t)\ S_n(\sigma_{k,n}\,r)\,Y_n^m(\varphi,\theta),\\
\widehat{u}_{k,m,n}(t,r,\varphi,\theta) &= \sin(c\,\sigma_{k,n}\,t)\ S_n(\sigma_{k,n}\,r)\,Y_n^m(\varphi,\theta),\\
\widetilde{u}_{k,m,n}(t,r,\varphi,\theta) &= \cos(c\,\sigma_{k,n}\,t)\ S_n(\sigma_{k,n}\,r)\,\widetilde{Y}_n^m(\varphi,\theta),\\
\widehat{\widetilde{u}}_{k,m,n}(t,r,\varphi,\theta) &= \sin(c\,\sigma_{k,n}\,t)\ S_n(\sigma_{k,n}\,r)\,\widetilde{Y}_n^m(\varphi,\theta).
\end{aligned} \qquad \begin{array}{l} k = 1,2,3,\dots\,,\\ m = 0,1,2,\dots\,,\\ n = 0,\dots,m. \end{array} \tag{18.107}$$

In particular, the radially symmetric modes of vibration have, according to (18.86), the elementary form

$$\begin{aligned}
u_{k,0,0}(r,\varphi,\theta) &= \cos c\,k\,\pi\,t\ S_0(k\,\pi\,r) = \cos c\,k\,\pi\,t\ \frac{\sin k\,\pi\,r}{r}\,,\\
\widehat{u}_{k,0,0}(r,\varphi,\theta) &= \sin c\,k\,\pi\,t\ S_0(k\,\pi\,r) = \sin c\,k\,\pi\,t\ \frac{\sin k\,\pi\,r}{r}\,,
\end{aligned} \qquad k = 1,2,3,\dots\,. \tag{18.108}$$

Their vibrational frequencies, $\omega_{k,0} = c\,k\,\pi$, are integral multiples of the lowest frequency $\omega_{0,1} = \pi$. Therefore, interestingly, if you only excite the radially symmetric modes, the ball would vibrate periodically motion.

More generally, adopting the same scaling argument as in (17.101), we conclude that the fundamental frequencies for a solid ball of radius R and wave speed c are given by $\omega_{k,n} = c\,\sigma_{k,n}/R$. The relative vibrational frequencies

$$\frac{\omega_{k,n}}{\omega_{1,0}} = \frac{\sigma_{k,n}}{\sigma_{1,0}} = \frac{\sigma_{k,n}}{\pi} \tag{18.109}$$

are independent of the size of the ball R or the wave speed c. In the accompanying table, we display all relative vibrational frequencies that are less than 4 in magnitude. The rows are indexed by n, the order of the spherical harmonic, while the columns are indexed by k, the root number.

The purely radial modes of vibration (18.108) have individual frequencies

$$\omega_{k,0} = \frac{k\,\pi\,c}{R}\,, \qquad \text{so} \qquad \frac{\omega_{k,n}}{\omega_{1,0}} = k,$$

and appear in the first row of the table. The lowest frequency is $\omega_{1,0} = \pi\,c/R$, corresponding to a vibration with period $2\,\pi/\omega_{1,0} = 2\,R/c$. In particular, for the earth, the radius $R \approx 6,000$ km and the wave speed in rock is, on average, $c \approx 5$ km/sec, so that the fundamental mode of vibration has period $2\,R/c \approx 2400$ seconds, or 40 minutes. Vibrations of the earth are also known as *seismic waves* and, of course, earthquakes are their most severe manifestation. Therefore understanding the modes of vibration is an issue of critical importance in geophysics and civil engineering, including the design of structures, buildings and bridges and the avoidance of resonant frequencies.

Of course, we have suppressed almost all interesting terrestrial geology in this very crude approximation, which has been based on the assumption that the earth is a uniform

Relative Spherical Bessel Roots $\sigma_{k,n}/\sigma_{1,0}$

	0	1	2	3	...
1	1.0000	2.0000	3.0000	4.0000	...
2	1.4303	2.4590	3.4709	⋮	
3	1.8346	2.8950	3.9225	⋮	
4	2.2243	3.3159	⋮		
5	2.6046	3.7258	⋮		
6	2.9780	⋮			
7	3.3463	⋮			
8	3.7105	⋮			
⋮	⋮				

body, vibrating only in its radial direction. A more realistic modeling of the vibrations of the earth requires an understanding of the basic partial differential equations of linear and nonlinear elasticity, [**69**]. Nonuniformities in the earth lead to scattering of the resulting vibrational waves. These in turn are used to understand the geological structures underneath the ground. For instance, setting off and then measuring small scale seismic vibrations is the primary means of determining its underlying structure, with oil and mineral exploration being a particularly important application. We refer the interested reader to [**6**] for a comprehensive introduction to mathematical seismology.

Remark: The number of spherical harmonics governs the energy levels or orbital shells occupied by electrons in an atom. In chemistry, the electron levels are indexed by order n of the spherical harmonic, and traditionally labeled by a letter in the sequence $p, s, d, f, \dots$. Thus, the order $n = 0$ spherical harmonics correspond to the p shells; the 3 harmonics of order $n = 1$ are the s shells, and so on. Since electrons are allowed to have one of two possible spins, the Pauli exclusion principle tells us that each energy shell can be occupied by at most two electrons. Thus, the number of electrons that can reside in the n^{th} energy level of an atom is $2(2n+1)$, the same as the number of linearly independent solutions to the wave equation associated with a given energy level. The configuration of energy shells and electrons in atoms are responsible for the periodic table. Thus, hydrogen has a single electron in the p shell. Helium has two electrons in the p shell. Lithium has 3 electrons, with two of them filling the first p shell and the third in the second p shell. Neon has 10 electrons filling the two p and first three s shells. And so on. The chemical properties of the elements are, to a very large extent, determined by the placement of the electrons within the different shells. See [**Chem**] for further details.

Example 18.15. The radial vibrations of a hollow spherical shell (e.g., an elastic

balloon) are governed by the differential equation

$$u_{tt} = c^2\, \Delta_S[u] = c^2 \left(\frac{\partial^2 u}{\partial \varphi^2} + \cot \varphi\, \frac{\partial u}{\partial \varphi} + \frac{1}{\sin^2 \varphi}\, \frac{\partial^2 u}{\partial \theta^2} \right), \tag{18.110}$$

where Δ_S denotes the spherical Laplacian (18.17). The radial displacement $u(t, \varphi, \theta)$ of a point on the sphere only depends on time t and the angular coordinates φ, θ. The solution $u(t, \varphi, \theta)$ is required to be 2π periodic in the meridial angle θ and bounded at the poles $\varphi = 0, \pi$.

According to (18.32), the n^{th} eigenvalue $\lambda_n = n(n+1)$ of the spherical Laplacian leads to $2n + 1$ linearly independent spherical harmonic eigenfunctions

$$Y_n^0(\varphi, \theta), \quad Y_n^1(\varphi, \theta), \quad \dots, \quad Y_n^n(\varphi, \theta), \quad \widetilde{Y}_n^1(\varphi, \theta), \quad \dots, \quad \widetilde{Y}_n^n(\varphi, \theta).$$

As a consequence, the fundamental frequencies of vibration for a spherical shell are

$$\omega_n = c\, \sqrt{\lambda_n} \, = c\, \sqrt{n(n+1)}\,, \qquad\qquad n = 0, 1, 2, \dots\,. \tag{18.111}$$

The vibrational solutions are quasi-periodic combinations of the fundamental modes

$$\begin{aligned} &\cos \sqrt{n(n+1)}\; t\; Y_n^m(\varphi, \theta), \qquad\qquad &&\sin \sqrt{n(n+1)}\; t\; Y_n^m(\varphi, \theta), \\ &\cos \sqrt{n(n+1)}\; t\; \widetilde{Y}_n^m(\varphi, \theta), \qquad\qquad &&\sin \sqrt{n(n+1)}\; t\; \widetilde{Y}_n^m(\varphi, \theta), \end{aligned} \tag{18.112}$$

involving the spherical harmonics. The smallest positive eigenvalue is $\lambda_1 = 2$, yielding a lowest tone of frequency $\omega_1 = c\sqrt{2}$. The higher order frequencies are irrational multiples of the lowest order one, $\omega_1 = c$, and hence a spherical bell sounds percussive to our ears.

The spherical Laplacian operator is only positive semi-definite, since the lowest mode has eigenvalue $\lambda_0 = 0$, which corresponds to the constant null eigenfunction $v_0(\varphi, \theta) = Y_0^0(\varphi, \theta) \equiv 1$. Therefore, the wave equation admits an unstable mode $b_{0,0}\, t$, corresponding to a uniform radial inflation. The coefficient

$$b_{0,0} = \frac{3}{4\pi} \iint_{S_1} \frac{\partial u}{\partial t}\, (0, \varphi, \theta)\, dS$$

represents the sphere's average initial velocity. The existence of such an unstable mode is an artifact of the simplified linear model we are using, that fails to account for nonlinearly elastic effects that serve to constrain the inflation of a spherical balloon.

18.6. Spherical Waves and Huygens' Principle.

The fundamental solution to the wave equation measures the effect of applying an instantaneous concentrated unit impulse at a single point. Two physical examples to keep in mind are the light waves propagating from a sudden concentrated blast, e.g., a stellar supernova or a lightning bolt, and the sound waves from an explosion or thunderclap, propagating in air at a much slower speed.

In a uniform isotropic medium, e.g., empty space, the initial blast leads to a spherically expanding wave, moving away at the speed of light or sound in all directions. Using translation invariance, we can assume that the source is at the origin, and so the solution

$u(t, \mathbf{x})$ should only depend on the distance $r = \| \mathbf{x} \|$ from the source. We change over to spherical coordinates and look for a solution $u = u(t, r)$ to the three-dimensional wave equation with no angular dependence. Substituting the formula (18.14) for the spherical Laplacian and setting the angular derivatives to 0, we are led to the partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} \right) \tag{18.113}$$

that governs the propagation of spherically symmetric waves. It turns out, surprisingly, that we can solve this partial differential equation. The secret is to multiply both sides of the equation by r. The resulting equation can be written in the form

$$\frac{\partial^2 (}{\partial r^2}\, u) t = r \, \frac{\partial^2 u}{\partial t^2} = c^2 \left(r \, \frac{\partial^2 u}{\partial r^2} + 2 \, \frac{\partial u}{\partial r} \right) = c^2 \, \frac{\partial^2}{\partial r^2} \, (r \, u),$$

and so (18.113) reduces to

$$\frac{\partial^2 w}{\partial t^2} = c^2 \, \frac{\partial^2 w}{\partial r^2} \,, \qquad \text{where} \qquad w(t, r) = r \, u(t, r). \tag{18.114}$$

Therefore the function $w(t, r)$ is a solution to the one-dimensional wave equation!

According to Theorem 14.8, the general solution to (18.114) has the d'Alembert form

$$w(t, r) = p(r - c\, t) + q(r + c\, t),$$

where $p(\xi)$ and $q(\eta)$ are *arbitrary* functions of a single characteristic variable. Reverting back to $u = w/r$, we conclude that the spherically symmetric solutions to the three-dimensional wave equation are all of the form

$$u(t, r) \;=\; \frac{p(r - c\, t)}{r} \;+\; \frac{q(r + c\, t)}{r} \,. \tag{18.115}$$

The first term

$$u(t, r) = \frac{p(r - c\, t)}{r} \tag{18.116}$$

in the solution (18.115) represents a wave moving at speed c in the direction of increasing r — away from the origin. It describes the effect of a variable light source concentrated at the origin. Think, for instance, of a pulsating quasar in interstellar space. To highlight this interpretation, let us look at the basic case when $p(s) = \delta(s - a)$ be a delta function at $s = a$; more general such solutions can then be assembled by linear superposition. The solution

$$u(t, r) = \frac{\delta(r - c\, t - a)}{r} = \frac{\delta\big(\, r - c(t - t_0)\,\big)}{r} \,, \qquad \text{where} \qquad t_0 = -\, \frac{a}{c} \,. \tag{18.117}$$

will represent a concentrated spherical wave. At the instant $t = t_0$, the light is entirely concentrated at the origin $r = 0$. The light impulse then moves away from the origin at speed c in all directions. At each later time $t > t_0$, the initially concentrated light source is now spread out over the surface of a sphere of radius $r = c\,(t - t_0)$. The intensity of

the signal at each point on the sphere, however, has decreased by a factor $1/r$, and so, the farther from the source, the weaker the signal. An observer sitting at a fixed point away from the source will only see an instantaneous flash of light as the spherical wave passes by. A similar phenomenon holds for sound waves — the sound of the explosion will only last momentarily. Thunder and lightning are the most familiar examples of this everyday phenomenon. On the other hand, for $t < t_0$, the impulse is concentrated at a negative radius $r = c\,(t - t_0) < 0$. To interpret this, note that, for a given value of the spherical angles $\varphi\theta$, the point

$$x = r\,\sin\varphi\,\cos\theta, \qquad y = r\,\sin\varphi\,\sin\theta, \qquad z = r\,\cos\varphi,$$

for $r < 0$ lies on the antipodal point of the sphere of radius $|\,r\,|$, so that replacing r by $-r$ has the same effect as changing $\mathbf{x}$ to $-\mathbf{x}$. Thus, the solution (18.117) represents a concentrated spherically symmetric light wave arriving from the edges of the universe at speed c, that strengthens in intensity as it collapses into the origin at $t = t_0$. After collapse, it immediately reappears in expanding form.

The second solution in the d'Alembert formula (18.115) has, in fact, exactly the same physical form. Indeed, if we set

$$\widehat{r} = -r, \qquad p(\xi) = -\,q(-\xi), \qquad \text{then} \qquad \frac{q(r + c\,t)}{r} = \frac{p(\widehat{r} - c\,t)}{\widehat{r}}\,.$$

Therefore, to represent the general radially symmetric solution to the three-dimensional wave equation, we only need use one of these constituents, and thus only need to consider solutions of the form (18.117) from now on.

In order to utilize such spherical wave solutions, we need to understand the nature of their originating singularity. For simplicity, we set $a = 0$ in (18.117) and concentrate on the particular solution

$$u(t, r) = \frac{\delta(r - c\,t)}{r}\,, \tag{18.118}$$

which has a singularity at the origin $r = 0$ when $t = 0$. We need to pin down precisely which sort of distribution this solution represents. Invoking the limiting definition of a distribution is tricky, and it will be easier to use the dual definition as a linear functional. Thus, at a fixed time $t \geq 0$, we must evaluate the inner product

$$\langle\, u\,;f\,\rangle = \iiint\, u(t, x, y, z)\, f(x, y, z)\, dx\, dy\, dz$$

of the solution with a smooth test function $f(\mathbf{x}) = f(x, y, z)$. We convert to spherical coordinates using the change of variables formula (B.66), whereby

$$\begin{aligned} \langle\, u\,;f\,\rangle &= \int_0^{\infty}\int_0^{2\,\pi}\int_0^{\pi} \frac{\delta(r - c\,t)}{r}\, f(r, \varphi, \theta)\, r^2 \sin\varphi\; d\varphi\, d\theta\, dr \\ &= c\,t\, \int_0^{2\,\pi}\int_0^{\pi} f(c\,t, \varphi, \theta)\, \sin\varphi\; d\varphi\, d\theta. \end{aligned} \tag{18.119}$$

Therefore, $\langle\, u\,;f\,\rangle = 4\pi\, c\, t\ \mathrm{M}^{\mathbf{0}}_{c\,t}\,[\,f\,]$, where

$$\mathrm{M}^{\mathbf{0}}_{c\,t}\,[\,f\,] = \frac{1}{4\pi\, c^2\, t^2}\iint_{S_{c\,t}} f\ dS = \frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi} f(c\,t,\varphi,\theta)\,\sin\varphi\ d\varphi\, d\theta \tag{18.120}$$

is the mean or average value of the function f on the sphere $S_{c\,t} = \|\,\mathbf{x}\,\| = c\,t$ of radius $r = c\,t$ centered at the origin $\mathbf{0}$. In particular, in the limit as $t \to 0$, the mean over the sphere of radius $r = 0$ is equal to the value of the function at the origin:

$$\mathrm{M}^{\mathbf{0}}_{0}\,[\,f\,] = f(\mathbf{0}), \tag{18.121}$$

and so, at $t = 0$, the formula implies that $\langle\, u\,;f\,\rangle = 0$ for *all* functions f. Consequently, $u(0,r) \equiv 0$ represents a trivial zero initial displacement.

How, then, can the solution be nonzero? Clearly, this must be the result of a nonzero initial velocity. Thus, we differentiate (18.119) with respect to t, whereby

$$\begin{aligned}\left\langle \frac{\partial u}{\partial t}\,;\, f\,\right\rangle &= \frac{\partial}{\partial t}\,\langle\, u\,;f\,\rangle\\ &= c\int_0^{2\pi}\int_0^{\pi} f(c\,t,\varphi,\theta)\,\sin\varphi\ d\varphi\, d\theta + c^2\, t\int_0^{2\pi}\int_0^{\pi}\frac{\partial f}{\partial r}\,(c\,t,\varphi,\theta)\,\sin\varphi\ d\varphi\, d\theta\\ &= 4\pi\, c\,\mathrm{M}^{\mathbf{0}}_{c\,t}\,[\,f\,] + 4\pi\, c^2\, t\,\mathrm{M}^{\mathbf{0}}_{c\,t}\left[\,\frac{\partial f}{\partial r}\,\right].\end{aligned} \tag{18.122}$$

The result is a linear combination of the mean of f and of its radial derivative f_r over the sphere of radius $c\,t$. In particular, at $t = 0$, using (18.121),

$$\langle\, u_t\,;f\,\rangle\,\big|_{t=0} = 4\pi\, c\,\mathrm{M}^{\mathbf{0}}_{0}\,[\,f\,] = 4\pi\, c\, f(\mathbf{0}),$$

We conclude that, at $t = 0$, the initial velocity

$$u_t(0,r) = 4\pi\, c\,\delta(\mathbf{x})$$

is a multiple of a delta function at the origin! Dividing through by $4\pi\, c$, we conclude that the spherical expanding wave

$$u(t,r) = \frac{\delta(r - c\,t)}{4\pi\, c\, r} \tag{18.123}$$

is the solution to the initial value problem

$$u(0,\mathbf{x}) \equiv 0, \qquad \frac{\partial u}{\partial t}\,(0,\mathbf{x}) = \delta(\mathbf{x}),$$

corresponding to an initial unit velocity impulse concentrated at the origin. This solution can be viewed as the three-dimensional version of striking a piano string with a hammer.

More generally, if our unit impulse is concentrated at the point $\boldsymbol{\xi}$, we invoke the translational symmetry of the wave equation to conclude that the function

$$G(t,\mathbf{x};\boldsymbol{\xi}) = \frac{\delta\big(\,\|\,\mathbf{x} - \boldsymbol{\xi}\,\| - c\,t\,\big)}{4\pi\, c\,\|\,\mathbf{x} - \boldsymbol{\xi}\,\|}\,, \qquad t \ge 0, \tag{18.124}$$

is the solution to the wave equation that satisfies the initial conditions

$$G(0, \mathbf{x}; \boldsymbol{\xi}) = 0, \qquad \frac{\partial G}{\partial t}\,(0, \mathbf{x}; \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}). \tag{18.125}$$

By linearity, any superposition of these spherical waves will also be a solution to the wave equation. Thus, for the initial conditions

$$u(0, x, y, z) = 0, \qquad \frac{\partial u}{\partial t}\,(0, x, y, z) = g(x, y, z), \tag{18.126}$$

representing a zero initial displacement, we write the initial velocity

$$g(\mathbf{x}) = \iiint g(\boldsymbol{\xi})\, \delta(\mathbf{x} - \boldsymbol{\xi})\, dx\, dy\, dz$$

as a superposition of delta functions, and immediately conclude that the relevant solution is the identical superposition of spherical waves

$$u(t, \mathbf{x}) = \frac{1}{4\pi c} \iiint g(\boldsymbol{\xi})\, \frac{\delta\big(\,\|\,\mathbf{x} - \boldsymbol{\xi}\,\| - c\,t\,\big)}{\|\,\mathbf{x} - \boldsymbol{\xi}\,\|}\, d\xi\, d\eta\, d\zeta = \frac{1}{4\pi c^2\, t} \iint_{\|\,\boldsymbol{\xi} - \mathbf{x}\,\| = c\,t} g(\boldsymbol{\xi})\, dS. \tag{18.127}$$

Therefore the value of the solution at a point $\mathbf{x}$ and time $t \geq 0$ is equal to

$$u(t, \mathbf{x}) = t\, \mathrm{M}^{\mathbf{x}}_{c\,t}\,[\,g\,], \tag{18.128}$$

namely t times the mean of the initial velocity function g over a sphere of radius $r = c\,t$ centered at the point $\mathbf{x}$.

Example 18.16. Let us set the wave speed $c = 1$ for simplicity. Suppose that the initial velocity

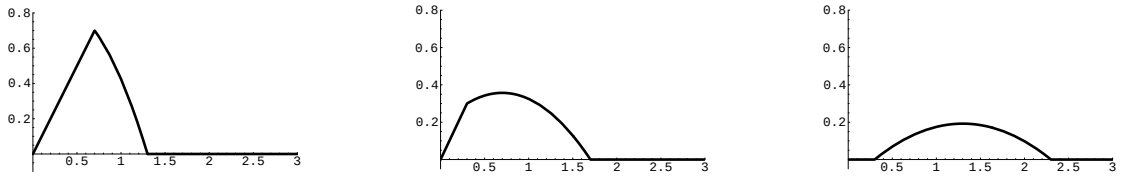
$$g(\mathbf{x}) = \begin{cases} 1, & \|\,\mathbf{x}\,\| < 1, \\ 0, & \|\,\mathbf{x}\,\| > 1 \end{cases}$$

is 1 within the unit ball B_1 centered at the origin, and 0 outside the ball. According to the formula (18.127) for the solution at a point $\mathbf{x}$ and time $t \geq 0$, we need to compute the average value of g over a sphere $S^{\mathbf{x}}_t$ of radius t centered at $\mathbf{x}$. Since $g = 0$ outside the unit sphere, its average will be equal to the surface area of that part of the sphere that is contained inside the unit ball, i.e., $S^{\mathbf{x}}_t \cap B_1$, divided by the total surface area of $S^{\mathbf{x}}_t$, namely $4\pi\, t^2$. The two spheres will intersect if either

$$(a)\ \ r > 1 \text{ and } r - 1 < t < r + 1, \qquad \text{or} \qquad (b)\ \ r < 1 \text{ and } 1 - r < t < 1 + r.$$

If $t > 1 + r$ or $t < r - 1$ and $r > 1$, then the sphere of radius t lies entirely outside the unit ball, and so the mean is 0; if $t < 1 - r$ and $r < 1$, then the sphere lies entirely within the unit ball and the mean is 1. Otherwise, referring to Figure bs■, and using Exercise ■, we see that the area of the spherical cap $S^{\mathbf{x}}_t \cap B_1$ is, by the Law of Cosines,

$$2\pi\, t^2(1 - \cos\alpha) = 2\pi\, t^2 \left(1 - \frac{1 - r^2 - t^2}{2\,r\,t}\right) = \frac{\pi\, t}{2\,r}\,[\,1 - (t - r)^2\,],$$



$r = .3$ $r = .7$ $r = 1.3$

Figure 18.3. Time Plot of the Solution to Wave Equation at Three Fixed Positions.

where $r = \| \mathbf{x} \|$ and $\varphi = \alpha$ denotes the azimuthal angle describing the circle of intersection between the two spheres. Therefore,

$$\mathrm{M}^{\mathbf{x}}_{ct}[\, g\,] = \begin{cases} 1, & 0 \le t \le 1 - r, \\ \dfrac{1 - (t-r)^2}{4\, r\, t}, & |\, r - 1\,| \le t \le 1 + r, \\ 0, & 0 \le t \le r - 1 \quad \text{or} \quad t \ge 1 + r. \end{cases} \tag{18.129}$$

The solution (18.128) is obtained by multiplying by t, and hence for $t \ge 0$,

$$u(t, \mathbf{x}) = \begin{cases} t, & 0 \le t \le 1 - \| \mathbf{x} \|, \\ \dfrac{1 - \big(t - \| \mathbf{x} \|\big)^2}{4 \| \mathbf{x} \|}, & |\, 1 - \| \mathbf{x} \| \,| \le t \le 1 + \| \mathbf{x} \|, \\ 0, & 0 \le t \le \| \mathbf{x} \| - 1 \quad \text{or} \quad t \ge 1 + \| \mathbf{x} \|. \end{cases} \tag{18.130}$$

As illustrated in Figure 18.3, an observer sitting inside the sphere at a distance $r < 1$ away from the origin will experience a linearly increasing light intensity followed by a parabolic decrease to 0 intensity, where it remains from then on. If the observer is closer to the edge than the center, the parabolic portion will continue to increase for a while before eventually tapering off. On the other hand, an observer sitting outside the sphere will experience, after an initially dark period, a parabolic increase to a maximal intensity and then symmetrical decrease, returning to dark after a total time laps of 2. We also show a plot of u at a function of $r = \| \mathbf{x} \|$ for various times in Figure wbr. Note that the light stays brightest in a sphere of gradually decreasing radius. At time $t = 1$ there remains a cusp, after which the solution is bright inside the domain lying between two concentric spheres of respective radii $t - 1$ and $t + 1$.

The solution described by formula (18.127) only handles initial velocities. How do we construct a solution corresponding to a nonzero initial displacement? Surprisingly, the answer is differentiation. The key observation is that if $u(t, \mathbf{x})$ is any (sufficiently smooth) solution to the wave equation, so is its time derivative

$$v(t, \mathbf{x}) = \frac{\partial u}{\partial t}\,(t, \mathbf{x}).$$

This follows at once from differentiating both sides of the wave equation with respect to t and using the equality of mixed partial derivatives. Physically, this implies that the velocity of a wave obeys the same evolutionary principle as the wave itself, which is a manifestation

of the linearity and time-independence (autonomy) of the equation. Suppose u has initial conditions

$$u(0, \mathbf{x}) = f(\mathbf{x}), \qquad u_t(0, \mathbf{x}) = g(\mathbf{x}).$$

What are the initial conditions for its derivative $v = u_t$? Clearly, its initial displacement $v(0, \mathbf{x}) = u_t(0, \mathbf{x}) = g(\mathbf{x})$ equals the initial velocity of u. As for its initial velocity, we have

$$\frac{\partial v}{\partial t} = \frac{\partial^2 u}{\partial t^2} = c^2 \, \Delta u$$

because we are assuming that u solves the wave equation. Thus, at the initial time

$$\frac{\partial v}{\partial t}\,(0, \mathbf{x}) = c^2 \, \Delta u(0, \mathbf{x}) = c^2 \, \Delta f(\mathbf{x})$$

equals c^2 times the Laplacian of the initial displacement[†]. In particular, if u satisfies the initial conditions

$$u(0, \mathbf{x}) = 0, \qquad u_t(0, \mathbf{x}) = g(\mathbf{x}), \tag{18.131}$$

then $v = u_t$ satisfies the initial conditions

$$v(0, \mathbf{x}) = g(\mathbf{x}), \qquad v_t(0, \mathbf{x}) = 0. \tag{18.132}$$

Thus, paradoxically, to solve the initial displacement problem we differentiate the initial velocity solution (18.127) with respect to t, and hence

$$v(t, \mathbf{x}) = \frac{\partial u}{\partial t}\,(t, \mathbf{x}) = \frac{\partial}{\partial t}\left(t \, \mathrm{M}^{\mathbf{x}}_{ct}\,[\, g\,] \right) = \mathrm{M}^{\mathbf{x}}_{ct}\,[\, g\,] + c\, t \, \mathrm{M}^{\mathbf{x}}_{ct}\left[\frac{\partial g}{\partial \mathbf{n}} \right], \tag{18.133}$$

using our computation in (18.122). Therefore, $v(t, \mathbf{x})$ is a linear combination of the mean of the function g and the mean of its normal or radial derivative $\partial g/\partial \mathbf{n}$, taken over a sphere of radius $c\, t$ centered at the point $\mathbf{x}$. In particular, to obtain the solution corresponding to a concentrated initial displacement,

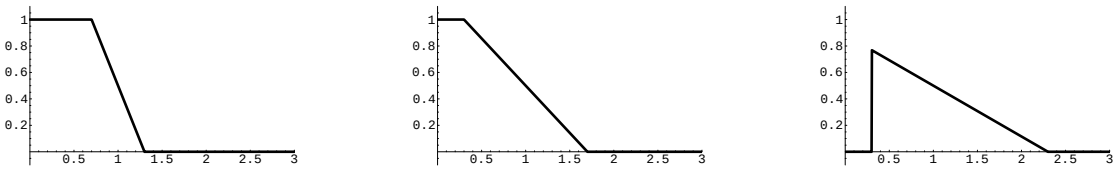
$$F(0, \mathbf{x}; \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad \frac{\partial F}{\partial t}\,(0, \mathbf{x}; \boldsymbol{\xi}) = 0, \tag{18.134}$$

we differentiate the solution (18.124), so

$$F(t, \mathbf{x}; \boldsymbol{\xi}) = \frac{\partial G}{\partial t}\,(t, \mathbf{x}; \boldsymbol{\xi}) = - \, \frac{\delta'\big(\| \, \mathbf{x} - \boldsymbol{\xi} \, \| - c\, t \, \big)}{4 \, \pi \, \| \, \boldsymbol{\xi} - \mathbf{x} \, \|}\,, \tag{18.135}$$

which represents a spherically expanding doublet, cf. Figure 11.10. Thus, interestingly, a concentrated initial displacement spawns a spherical doublet or derived delta wave, whereas a concentrated initial velocity spawns a singlet delta wave.

[†] A similar device is used to initiate the numerical solution method for the wave equation; see Section 14.6.



$r = .3$ $r = .7$ $r = 1.3$

Figure 18.4. Time Plot of Solutions to Wave Equation at Three Fixed Position.

Example 18.17. Let $c = 1$ for simplicity. Consider the initial displacement

$$u(0, \mathbf{x}) = f(\mathbf{x}) = \begin{cases} 1, & \| \mathbf{x} \| < 1, \\ 0, & \| \mathbf{x} \| > 1 \end{cases}$$

along with zero initial velocity, corresponding to an instantaneously illuminated solid glass ball. To obtain the solution, we try differentiating (18.130) with respect to t, leading to

$$u(t, \mathbf{x}) = \begin{cases} 1, & 0 \le t < 1 - \| \mathbf{x} \|, \\ \dfrac{\| \mathbf{x} \| - t}{2 \| \mathbf{x} \|}, & | \, 1 - \| \mathbf{x} \| \, | \le t \le 1 + \| \mathbf{x} \|, \\ 0, & 0 \le t < \| \mathbf{x} \| - 1 \quad \text{or} \quad t > 1 + \| \mathbf{x} \|. \end{cases} \tag{18.136}$$

As illustrated in Figure 18.4, an observer sitting inside the ball at radius $r < 1$ will begin by experiencing a constant intensity, followed by a sudden jump, then linear decrease, and finally a jump back to quiescent, while an observer sitting outside, with $r > 1$, will experience, after an initially dark period, a sudden jump in the light intensity, followed by a linear decrease to darkness. ∎ The size of the jump depends upon the distance from the ball.

By linearity, we can combine the two solutions (18.128), (18.133) together, and have thus established a d'Alembert-type solution formula for the wave equation in three-dimensional space.

Theorem 18.18. *The solution to the initial value problem*

$$u_{tt} = c^2 \, \Delta u, \qquad u(0, \mathbf{x}) = f(\mathbf{x}), \qquad \frac{\partial u}{\partial t}(0, \mathbf{x}) = g(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^3, \tag{18.137}$$

for the wave equation in three-dimensional space is given by

$$u(t, \mathbf{x}) = \mathrm{M}^{\mathbf{x}}_{c\,t}[\, f \,] + c\, t \, \mathrm{M}^{\mathbf{x}}_{c\,t}\left[\frac{\partial f}{\partial \mathbf{n}} \right] \; + \; t \, \mathrm{M}^{\mathbf{x}}_{c\,t}[\, g \,], \tag{18.138}$$

where $\mathrm{M}^{\mathbf{x}}_{c\,t}[\, f \,]$ *denotes the average value of the function* f *over a sphere of radius* $c\,t$ *centered at position* $\mathbf{x}$.

Observe that the value of the solution (18.138) at a point $\mathbf{x}$ and time t only depends upon the values of the initial displacements and velocities at a distance $c\,t$ away. Physically, this means that the light that we see at a given time t arrived from points at a distance exactly $d = c\,t$ away at time $t = 0$. In particular, a sharp, localized initial signal — whether

initial displacement or initial velocity — that is concentrated near a point produces a sharp, localized response concentrated on a sphere surrounding the point at all subsequent times. In our three-dimensional universe, we only witness the light from an explosion for a brief moment, after which if there is no subsequent light source, the view returns to darkness. Similarly, a sharp sound remains sharply concentrated, with diminishing magnitude, as it propagates through space. This phenomenon was first highlighted the seventeenth century Dutch scientist Christiaan Huygens and is known as *Huygens' Principle* in his honor. Remarkably, as we will show next, Huygens' Principle does not hold in a two dimensional universe! In the plane, concentrated impulses will be spread out as time progresses.

The Method of Descent

So far, we have explicitly determined the response of the wave equation to an initial displacement and initial velocity in one- and three-dimensional space. The two-dimensional case

$$u_{tt} = c^2 \Delta u = c^2 (u_{xx} + u_{yy}). \tag{18.139}$$

is, counter-intuitively, more complicated! For instance, looking for a radially symmetric solution $u(t, r)$ leads to the partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right) \tag{18.140}$$

which, unlike its three-dimensional cousin (18.113), cannot be so easily integrated.

However, our solution to the three-dimensional problem can be easily adapted to construct a solution using the so-called *method of descent*. Any solution $u(t, x, y)$ to the two-dimensional wave equation (18.139) can be viewed as a solution to the three-dimensional wave equation (18.100) that does not depend upon the vertical z coordinate, whence $\partial u/\partial z = 0$. Clearly, if the initial data does not depend on z, then the resulting solution $u(t, x, y)$ will also be independent of z.

Consider first the solution formula (18.127) corresponding to initial conditions

$$u(0, x, y) = 0, \qquad \frac{\partial u}{\partial t}\,(0, x, y) = g(x, y). \tag{18.141}$$

of zero initial displacement, but nonzero initial velocity. We rewrite the formula in the form of a surface integral over the sphere $S_{c\,t} = \big\{ \,\| \boldsymbol{\xi} \,\| = c\,t \,\big\}$ centered at the origin:

$$u(t, \mathbf{x}) = \frac{1}{4\,\pi\,c^2\,t} \iint_{S_{c\,t}} g(\boldsymbol{\xi})\, dS = \frac{1}{4\,\pi\,c^2\,t} \iint_{\| \boldsymbol{\xi} \| = c\,t} g(\mathbf{x} + \boldsymbol{\xi})\, dS. \tag{18.142}$$

Imposing the condition that $g(x, y)$ does not depend upon the z coordinate, we see that the integrals over the upper and lower hemispheres

$$S^+_{c\,t} = \big\{ \,\| \boldsymbol{\xi} \,\| = c\,t, \;\; \zeta \geq 0 \,\big\}, \qquad\qquad S^-_{c\,t} = \big\{ \,\| \boldsymbol{\xi} \,\| = c\,t, \;\; \zeta \leq 0 \,\big\},$$

are identical. As in (B.45), to evaluate the upper hemispherical integral, we parametrize the upper hemisphere as the graph $\zeta = \sqrt{(c\,t)^2 - \xi^2 - \eta^2}$ over the disk $D_{c\,t} =$

$\{\, \xi^2 + \eta^2 \le c^2\, t^2 \,\}$, and so

$$u(t,x,y) = \frac{1}{2\,\pi\, c^2\, t} \iint_{S^+_{c\,t}} g(\mathbf{x} + \boldsymbol{\xi})\, dS = \frac{1}{2\,\pi\, c} \iint_{D_{c\,t}} \frac{g(x+\xi, y+\eta)}{\sqrt{(c\,t)^2 - \xi^2 - \eta^2}}\, d\xi\, d\eta, \tag{18.143}$$

which solves the initial value problem (18.141). In particular, if we take the initial velocity $g(x,y) = \delta(x-\xi)\,\delta(y-\eta)$ to be a concentrated impulse, then the resulting solution is

$$G(t,x,y;\xi,\eta) = \begin{cases} 0, & (x-\xi)^2 + (y-\eta) > c\,t, \\ \dfrac{1}{2\,\pi\, c\sqrt{c^2\, t^2 - (x-\xi)^2 - (y-\eta)^2}}, & (x-\xi)^2 + (y-\eta) < c\,t. \end{cases} \tag{18.144}$$

Thus, given a concentrated impulse in the velocity at time $t = 0$, an observer sitting at position $\mathbf{x}$ will first experience a concentrated light wave at time $t = \| \mathbf{x} - \boldsymbol{\xi} \|/c$. However, in contrast to the three-dimensional solution, the observer will continue to experience a non-zero signal after the initial disturbance has passed, with decreasing magnitude proportional to $1/(c\,t)$; see the first graph in Figure nhp2■. Thus, although the initial condition is concentrated, in contrast to the three-dimensional case, the resulting solution is not. In a two-dimensional universe, Huygens' principle is *not* valid. A two-dimensional creature would experience not only a initial effect of any sound or light wave but also an "afterglow" with slowly diminishing magnitude. It would be like living in a permanent echo chamber, and so understanding and acting upon sensory phenomena would more challenging in a two-dimensional universe. In general, Huygens' principle is only valid in odd-dimensional spaces; see also [**17**] for recent advances in the classification of partial differential equations that admit a Huygens' principle.

Similarly, the solution to the initial displacement conditions

$$u(0,x,y) = f(x,y), \qquad \frac{\partial u}{\partial t}\,(0,x,y) = 0, \tag{18.145}$$

can be obtained by differentiation with respect to t. Thus,

$$u(t,x,y) = \frac{\partial}{\partial t}\left(\frac{1}{2\,\pi\, c} \iint_{D_{c\,t}} \frac{f(x+\xi, y+\eta)}{\sqrt{(c\,t)^2 - \xi^2 - \eta^2}}\, d\xi\, d\eta \right) \tag{18.146}$$

is the desired solution. The general solution is a linear combination of the two types of solutions (18.143), (18.146). Note that the solution at a point $\mathbf{x}$ at time t depends upon the initial displacement and velocity on the entire disk of radius $r\,t$ centered at the point, and not just on the points a distance $c\,t$ away.

Remark: Since the solutions to the two-dimensional wave equation can be interpreted as three-dimensional solutions with no z dependence, a concentrated delta impulse in the two-dimensional wave equation would correspond to a concentrated line impulse in three dimensions. If light starts propagating from the line at $t = 0$, after the initial signal reaches us, we will continue to receive light from points that are progressively farther away along the line, which accounts for the two-dimensional afterglow.

Chapter 19

Nonlinear Systems

Nonlinearity is ubiquitous in physical phenomena. Fluid mechanics, elasticity, relativity, chemical reactions, combustion, ecology, biomechanics, and many, many others are all governed by inherently nonlinear equations. (The one notable exception is quantum mechanics, which is a fundamentally linear theory. More recent attempts at grand unification of all fundamental physical theories, such as string theory and conformal field theory, do venture into the nonlinear realm.) For this reason, an increasingly large fraction of modern mathematical research is devoted to the analysis of nonlinear systems. The advent of powerful computers has finally placed nonlinearity within our grasp, and thereby fomented a revolution in our understanding and development of nonlinear mathematics. Indeed, many of the most important modern analytical techniques drew their inspiration from early computer forays into the uncharted nonlinear wilderness.

Why, then, have we spent the overwhelming majority of this text developing purely linear mathematics? The facile answer, of course, is that nonlinear systems are vastly more difficult to analyze. In the nonlinear regime, many basic questions remain unanswered; existence and uniqueness of solutions are not guaranteed; explicit formulae are difficult to come by; linear superposition is no longer available; numerical approximations are not always sufficiently accurate; etc., etc. But, a more intelligent answer is that, without a proper understanding of linear phenomena and linear mathematics, one has no foundation upon which to erect a nonlinear analysis. Therefore, in an introductory text on applied mathematics, we are forced to develop in detail the proper linear foundations to aid us when we confront the nonlinear beast.

Moreover, many important physical systems are "weakly nonlinear", in the sense that, while nonlinear effects do play an essential role, the linear terms dominate the system, and so, to a first approximation, the system is close to linear. As a result, the underlying nonlinear phenomena can be understood by suitably perturbing their linear approximations. Historically, while certain nonlinear problems date back to Newton (for example the n body problem arising in celestial mechanics and planetary motion), significant progress in understanding weak nonlinearities only began after computers became sufficiently powerful tools. The truly nonlinear regime is, even today, only sporadically modeled and even less well understood. Despite dramatic advances in both hardware and mathematical algorithms, many nonlinear systems, for instance Einsteinian gravitation, still remain beyond the capabilities of today's computers and algorithms.

Space limitations imply that we can only provide a brief overview of some of the key ideas and phenomena that arise when venturing into the nonlinear realm. This chapter is devoted to the study of nonlinear functions and equations. In the remaining chapters, we

shall ascend the nonlinear "dimensional ladder", passing from equilibrium to dynamics and from discrete to continuous, mimicking our linear ascent that guided the logical progression in the preceding chapters of the text.

We begin with an analysis of the iteration of nonlinear functions. Building on our experience with iteration of linear systems, we will discover that functional iteration, when it converges, provides a powerful mechanism for solving equations and optimization. When it fails to converge, even very simple nonlinear iterations can lead to remarkably complex, chaotic behavior. The second section is devoted to basic solution techniques for nonlinear systems, and includes the bisection method, iterative methods, and the powerful Newton method. The third section is devoted to optimization, i.e., the minimization of nonlinear functions on finite-dimensional spaces. As we know, the equilibrium configurations of discrete mechanical systems are minimizers of the potential energy in the system. The locations where the gradient of the function vanishes are the critical points, and include the local minima and maxima as well as non-optimizing saddle points. Nondegenerate critical points are classified by a second derivative test based on Hessian matrix. These results from multivariable calculus will be developed in a form that readily generalizes to minimization problems on infinite-dimensional function space, to be presented in Chapter 21. Numerical optimization procedures rely on iterative procedures, and we present those connected with a gradient descent approach.

19.1. Iteration of Functions.

Iteration, or repeated application of a function, plays an essential role in the modern theories of dynamical systems. Iteration can be regarded as a discrete dynamical system, in which the continuous time variable has been "quantized". Even iterating a very simple quadratic function leads to an amazing variety of phenomena, including convergence, period doubling, and chaos. Discrete dynamical systems arise not just in mathematics, but also underlie the theory of growth and decay of biological populations, predator-prey models, spread of communicable diseases such as AIDS, and host of other natural phenomena. Moreover, many numerical solution methods — for systems of algebraic equations, ordinary differential equations, partial differential equations and so on — rely in essence on an iterative method, and so the basic results on function iteration play a key role in the analysis of convergence and efficiency of such numerical techniques.

In general, an iterative system of the form

$$\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)}), \tag{19.1}$$

is also known as a *discrete dynamical system*. A solution is a discrete collection of points $\mathbf{u}^{(k)}$ in which the index $k = 0, 1, 2, 3, \dots$ takes on non-negative integer values. One might also consider negative integral values $k = -1, -2, \dots$ of the index, but we will not. The superscripts on $\mathbf{u}^{(k)}$ refer to the iteration number, and do *not* denote derivatives. The index k may be viewed as the discrete "time" for the system, indicating the number of days, years, seconds, etc.

The function[†] $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ is usually assumed to be continuous. Later on we shall also

[†] Complex iteration is based on a complex-valued function $\mathbf{g}\colon \mathbb{C}^n \to \mathbb{C}^n$.

require that $\mathbf{g}$ be reasonably smooth, meaning that it has at least one or two continuous partial derivatives everywhere. Chapter 10 dealt with the case when $\mathbf{g}(\mathbf{u}) = A\,\mathbf{u}$ is a linear function, necessarily given by multiplication by an $n \times n$ matrix A. In this chapter, we allow nonlinear functions into the picture.

Once we specify an initial condition, say

$$\mathbf{u}^{(0)} = \mathbf{c}, \tag{19.2}$$

for the initial iterate, then the solution is easy to compute mechanically:

$$\mathbf{u}^{(1)} = \mathbf{g}(\mathbf{u}^{(0)}) = \mathbf{g}(\mathbf{c}), \quad \mathbf{u}^{(2)} = \mathbf{g}(\mathbf{u}^{(1)}) = \mathbf{g}(\mathbf{g}(\mathbf{c})), \quad \mathbf{u}^{(3)} = \mathbf{g}(\mathbf{u}^{(2)}) = \mathbf{g}(\mathbf{g}(\mathbf{g}(\mathbf{c}))), \quad \dots$$

and so on. Therefore, unlike continuous dynamical systems, existence and uniqueness of solutions is immediate. As long as each successive iterate $\mathbf{u}^{(k)}$ lies in the domain of definition of $\mathbf{g}$ one merely repeats the process to produce the solution,

$$\mathbf{u}^{(k)} = \mathbf{g} \circ \mathbf{g} \circ \cdots \circ \mathbf{g}(\mathbf{c}), \qquad\qquad k = 0, 1, 2, \dots, \tag{19.3}$$

which is obtained by composing the function $\mathbf{g}$ with itself a total of k times. In other words, the solution to a discrete dynamical system corresponds to repeatedly pushing the $\mathbf{g}$ key on your calculator. For example, repeatedly hitting the `sin` key corresponds to a solution to the system $u^{(k+1)} = \sin u^{(k)}$. For simplicity, we shall tacitly assume that the function $\mathbf{g}$ is defined on all of $\mathbb{R}^n$. Otherwise, we must always be careful that the successive iterates $\mathbf{u}^{(k)}$ never leave the domain of definition of $\mathbf{g}$, which would cause the iteration to break down.

While the solution to a discrete dynamical system is essentially trivial, understanding its behavior is definitely not. Sometimes the solution converges to a particular value — the key requirement for numerical solution methods. Sometimes it goes off to ∞, or, more precisely, $\| \mathbf{u}^{(k)} \| \to \infty$. Sometimes the solution repeats itself after a while. And sometimes it behaves in a random, chaotic manner — all depending on the function $\mathbf{g}$ and, at times, the initial condition $\mathbf{c}$. Although any of these cases may appear and play a role in applications, we shall mostly concentrate upon understanding the case of convergence of the iterates.

Definition 19.1. A *fixed point* or *equilibrium solution* for a discrete dynamical system (19.1) is a vector $\mathbf{u}^\star \in \mathbb{R}^n$ such that

$$\mathbf{g}(\mathbf{u}^\star) = \mathbf{u}^\star. \tag{19.4}$$

We easily see that every fixed point provides a constant solution, namely $\mathbf{u}^{(k)} \equiv \mathbf{u}^\star$, to the discrete dynamical system. Moreover, solutions that converge always converge to a fixed point.

Proposition 19.2. *If a solution to a discrete dynamical system converges,*

$$\lim_{k \to \infty} \mathbf{u}^{(k)} = \mathbf{u}^\star,$$

then the limit $\mathbf{u}^\star$ is a fixed point of the system.

Proof: This is a simple consequence of the continuity of $\mathbf{g}$. We have

$$\mathbf{u}^\star = \lim_{k\to\infty} \mathbf{u}^{(k+1)} = \lim_{k\to\infty} \mathbf{g}(\mathbf{u}^{(k)}) = \mathbf{g}\left(\lim_{k\to\infty} \mathbf{u}^{(k)}\right) = \mathbf{g}(\mathbf{u}^\star),$$

the last two equalities following from the continuity of $\mathbf{g}$. *Q.E.D.*

Of course, not every solution to a discrete dynamical system will necessarily converge, but Proposition 19.2 says that if it does, then it must converge to a fixed point. Thus, the goal is to understand when a solution converges, and, if so, to which fixed point — if there is more than one. (In the linear case, only the actual convergence is a significant issues since most linear systems admit exactly one fixed point, namely $\mathbf{u}^\star = \mathbf{0}$.) Fixed points are roughly divided into three classes: *asymptotically stable*, with the property that all nearby solutions converge to it, *stable*, with the property that all nearby solutions stay nearby, and *unstable*, almost all of whose nearby solutions diverge away from the fixed point. Thus, from a practical standpoint, convergence of the iterates of a discrete dynamical system requires asymptotic stability of the fixed point.

Scalar Functions

As always, the first step is to thoroughly understand the scalar case, and so we begin with a discrete dynamical system

$$u^{(k+1)} = g(u^{(k)}), \qquad u^{(0)} = c, \tag{19.5}$$

in which $g\colon \mathbb{R} \to \mathbb{R}$ is a continuous, scalar-valued function. As noted above, we will assume, for simplicity, that g is defined everywhere, and so the iterates $u^{(0)}, u^{(1)}, u^{(2)}, \dots$ are all well-defined.

The linear case $g(u) = a\,u$ was treated in Section 10.1, following (10.2). The simplest "nonlinear" case is that of an affine function

$$g(u) = a\,u + b, \tag{19.6}$$

leading to an *affine discrete dynamical system*

$$u^{(k+1)} = a\,u^{(k)} + b. \tag{19.7}$$

The only fixed point is the solution to

$$u^\star = g(u^\star) = a\,u^\star + b, \qquad \text{namely,} \qquad u^\star = \frac{b}{1-a}\,. \tag{19.8}$$

The formula for $u^\star$ requires that $a \neq 1$, and, indeed, the case $a = 1$ has no fixed point, as the reader can easily confirm; see Exercise ■. Since we already know the value of $u^\star$, we can easily analyze the difference

$$e^{(k)} = u^{(k)} - u^\star, \tag{19.9}$$

between the iterate $u^{(k)}$ and the fixed point. The smaller $e^{(k)}$ is, the closer $u^{(k)}$ is to the desired fixed point. In many applications, the iterate $u^{(k)}$ is viewed as an approximation

to the fixed point $u^\star$, and so $e^{(k)}$ is interpreted as the *error* in the k^{th} iterate. Subtracting the fixed point equation (19.8) from the iteration equation (19.7), we find

$$u^{(k+1)} - u^\star = a\,(u^{(k)} - u^\star).$$

Therefore the errors $e^{(k)}$ satisfy a *linear iteration*

$$e^{(k+1)} = a\,e^{(k)}, \qquad \text{and hence} \qquad e^{(k)} = a^k\,e^{(0)}. \tag{19.10}$$

Therefore, as we already demonstrated in Section 10.1, the solutions to this scalar linear iteration converge,

$$e^{(k)} \longrightarrow 0 \qquad \text{and hence} \qquad u^{(k)} \longrightarrow u^\star, \qquad \text{if and only if} \qquad |\,a\,| < 1.$$

This is the criterion for *asymptotic stability* of the fixed point, or, equivalently, convergence of the affine iterative system (19.7). The magnitude of $|\,a\,| < 1$ determines the rate of convergence, and the closer it is to 0, the faster the iterates approach to the fixed point.

Example 19.3. Suppose $g(u) = \frac{1}{4}\,u + 2$, and so we consider the iterative scheme

$$u^{(k+1)} = \tfrac{1}{4}\,u^{(k)} + 2.$$

Starting with the initial condition $u^{(0)} = 0$, the ensuing values are

k	1	2	3	4	5	6	7	8
$u^{(k)}$	2.0	2.5	2.625	2.6562	2.6641	2.6660	2.6665	2.6666

Thus, after 8 iterations, the iterates have converged to the fixed point $u^\star = \frac{8}{3}$ to 4 decimal places. The rate of convergence is $\frac{1}{4}$, and indeed

$$|\,u^{(k)} - u^\star\,| = \left(\tfrac{1}{4}\right)^k (u^{(0)} - u^\star) = \tfrac{8}{3}\,\left(\tfrac{1}{4}\right)^k \longrightarrow 0 \qquad \text{as} \qquad k \longrightarrow \infty.$$

Let us now turn to the fully nonlinear case. In general, near a given point, any (smooth) nonlinear function can be approximated by its tangent line, which is an affine function; see Figure tl1. Therefore, if we are close to a fixed point $u^\star$, then we might expect the behavior of the nonlinear system will behave very much like iteration of its affine approximation. And, indeed, this intuition turns out to be essentially correct. This result forms our first concrete example of *linearization*, in which the analysis of a nonlinear system is based on its linear (or, more correctly, affine) approximation.

The explicit formula for the tangent line to $g(u)$ near the fixed point $u = u^\star$ is

$$g(u) \;\approx\; g(u^\star) + g'(u^\star)(u - u^\star) \;=\; a\,u + b, \tag{19.11}$$

where

$$a = g'(u^\star), \qquad b = g(u^\star) - g'(u^\star)\,u^\star = \big(1 - g'(u^\star)\big)\,u^\star.$$

Note that $u^\star = b/(1-a)$ remains a fixed point for the affine approximation. According to the preceding discussion, the convergence of the iterates for the affine approximation is governed by the size of the coefficient $a = g'(u^\star)$. This observation inspires the key stability criterion for fixed points of scalar iterative systems.

Theorem 19.4. *Suppose $g(u)$ is a continuously differentiable scalar function. Suppose $u^\star = g(u^\star)$ is a fixed point. If $|\, g'(u^\star)\,| < 1$, then $u^\star$ is a stable fixed point, and hence any sequence of iterates $u^{(k)}$ which starts out sufficiently close to $u^\star$ will converge to $u^\star$. On the other hand, if $|\, g'(u^\star)\,| > 1$, then $u^\star$ is an unstable fixed point, and the only iterates which converge to it are those that land exactly on it, i.e., $u^{(k)} = u^\star$ for some $k \geq 0$.*

Proof: The goal is to prove that the errors $e^{(k)} = u^{(k)} - u^\star$ between the k^{th} iterate and the true fixed point tend to 0 as $k \to \infty$. To this end, we try to estimate $e^{(k+1)}$ in terms of $e^{(k)}$. According to (19.5) and the Mean Value Theorem C.3 from calculus,

$$e^{(k+1)} = u^{(k+1)} - u^\star = g(u^{(k)}) - g(u^\star) = g'(v)\,(u^{(k)} - u^\star) = g'(v)\,e^{(k)}, \tag{19.12}$$

for some v lying between $u^{(k)}$ and $u^\star$. By continuity, if $|\, g'(u^\star)\,| < 1$ at the fixed point, then we can choose $0 < \rho < 1$ such that

$$|\, g'(v)\,| \leq \rho < 1 \qquad \text{whenever} \qquad |\, v - u^\star\,| < \delta \tag{19.13}$$

holds in a (perhaps small) interval surrounding the fixed point. If $|\, e^{(k)}\,| = |\, u^{(k)} - u^\star\,| < \delta$, then the point v in (19.12) satisfies (19.13). Therefore,

$$|\, u^{(k+1)} - u^\star\,| \;\leq\; \rho\,|\, u^{(k)} - u^\star\,|, \qquad \text{and hence} \qquad |\, e^{(k+1)}\,| \;\leq\; \rho\,|\, e^{(k)}\,|. \tag{19.14}$$

In particular, since $\rho < 1$, if $|\, u^{(k)} - u^\star\,| < \delta$, then $|\, u^{(k+1)} - u^\star\,| < \delta$, and hence the subsequent iterate $u^{(k+1)}$ also lies in the interval where (19.13) holds. Iterating, we conclude that the errors satisfy

$$e^{(k)} \;\leq\; \rho^k\, e^{(0)}, \qquad \text{and hence} \qquad e^{(k)} = |\, u^{(k)} - u^\star\,| \;\longrightarrow\; 0 \quad \text{as} \quad k \to \infty, \tag{19.15}$$

which completes the proof of the theorem in the stable case. The proof in unstable case is left as Exercise ■ for the reader. *Q.E.D.*

Remark: The borderline cases $g'(u^\star) = \pm 1$ are *not* covered by the theorem. For a linear system, these cases are stable, but not asymptotically stable. For nonlinear systems, such borderline situations require more detailed knowledge of the nonlinear terms in order to resolve the status — stable or unstable — of the fixed point. Despite their importance in certain applications, we will not try to analyze such borderline cases any further here. From now on, we will only deal with asymptotically stable fixed points, and, for brevity, usually omit the adjective "asymptotically".

Example 19.5. Given constants ϵ, m, the trigonometric equation

$$u = m + \epsilon\, \sin u \tag{19.16}$$

is known as *Kepler's equation*. It arises in the study of planetary motion, with $|\, \epsilon\,| < 1$ representing the *eccentricity* of an elliptical planetary orbit and m its *mean anomaly*; see Figure Kepler■. The desired solution u is the *eccentric anomaly*, and governs the motion of the planet around the ellipse. Details can be found in [**76**; p. 119].

The solutions to Kepler's equation are the fixed points of the discrete dynamical system based on the function $g(u) = m + \epsilon \sin u$. Note that

$$|\, g'(u)\,| = |\, \epsilon\, \cos u\,| = |\, \epsilon\,| < 1, \tag{19.17}$$

which automatically implies that the as yet unknown fixed point is stable. Indeed, Exercise ∎ implies that condition (19.17) is enough to prove the existence of a unique stable fixed point. In the particular case $m = \epsilon = \frac{1}{2}$, the result of iterating $u^{(k+1)} = \frac{1}{2} + \frac{1}{2}\sin u^{(k)}$ starting with $u^{(0)} = 0$ is

k	1	2	3	4	5	6	7	8	9
$u^{(k)}$	0.5	0.7397	0.8370	0.8713	0.8826	0.8862	0.8873	0.8877	0.8878

After 13 iterations, we have converged sufficiently close to the solution (fixed point) $u^\star = 0.887862$ to have computed its value to 7 decimal places.

Remark: Inspection of the proof of Theorem 19.4 reveals that we never really used the differentiability of g, except to verify the inequality

$$| g(u) - g(v) | \le \rho\, | u - v | \qquad \text{for some fixed } \rho. \tag{19.18}$$

A function that satisfies (19.18) for all u nearby a given point v is called *Lipschitz continuous*, in honor of the 19th century German mathematician Rudolf Lipschitz. The Mean Value Theorem C.3 implies that any continuously differentiable function $g \in \mathrm{C}^1$ is automatically Lipschitz continuous, but there are nondifferentiable examples. The simplest is the absolute value function $g(u) = | u |$, which is Lipschitz continuous, since

$$| g(u) - g(v) | \;=\; \Big|\; | u | - | v | \;\Big| \;\le\; | u - v | \qquad \text{for any} \qquad u, v \in \mathbb{R},$$

but is not differentiable at $u = 0$. On the other hand, as its name indicates, Lipschitz continuity does imply continuity. Thus, stability of the fixed point follows from the weaker hypothesis that $g(u)$ is Lipschitz continuous at $u^\star$ with *Lipschitz constant* $\rho < 1$.

Example 19.6. The simplest truly nonlinear example is a quadratic polynomial. The most important case is the so-called *logistic map*

$$g(u) = \lambda\, u(1 - u), \tag{19.19}$$

where $\lambda \neq 0$ is a fixed non-zero parameter. (The case $\lambda = 0$ is completely trivial. Why?) In fact, an elementary change of variables can make any quadratic iterative system into one involving a logistic map; see Exercise ∎.

The fixed points of the logistic map are the solutions to the quadratic equation

$$u = \lambda\, u(1 - u), \qquad \text{or} \qquad \lambda\, u^2 - \lambda\, u + 1 = 0.$$

Using the quadratic formula, we conclude that $g(u)$ has two fixed points:

$$u_1^\star = 0, \qquad u_2^\star = 1 - \frac{1}{\lambda}\,.$$

Let us apply Theorem 19.4 to determine their stability. The derivative is

$$g'(u) = \lambda - 2\,\lambda\, u, \qquad \text{and so} \qquad g'(u_1^\star) = \lambda, \qquad g'(u_2^\star) = 2 - \lambda.$$

Therefore, if $| \lambda | < 1$, the first fixed point is stable, while if $1 < \lambda < 3$, the second fixed point is stable. For $\lambda < -1$ or $\lambda > 3$ neither fixed point is stable, and we expect the iterates to not converge at all.

Numerical experiments with this example show that it is the source of an amazingly diverse range of behavior, depending upon the value of the parameter λ. In the following table ■ , we display the results of iteration starting with initial point $u^{(0)} = 1$. As expected from Theorem 19.4, the iterates converge to one of the fixed points in the range $-1 < \lambda < 3$, except when $\lambda = 1$. For λ a little bit larger than $\lambda_1 = 3$, the iterates do not converge to a fixed point; an example appears in the table ■ . But it does not take long for them to settle down and switch back and forth between two particular values. This behavior indicates that there is a (stable) *period* 2 *orbit* for the discrete dynamical system, in accordance with the following definition.

Definition 19.7. A *period k orbit* of a discrete dynamical system is a solution that satisfies $u^{(n+k)} = u^{(n)}$ for all $n = 0, 1, 2, \ldots$. The (*minimal*) *period* is the smallest positive value of k for which this condition holds.

Thus, a fixed point

$$u^{(0)} = u^{(1)} = u^{(2)} = \cdots$$

is a period 1 orbit. A period 2 orbit satisfies

$$u^{(0)} = u^{(2)} = u^{(4)} = \cdots \qquad \text{and} \qquad u^{(1)} = u^{(3)} = u^{(5)} = \cdots ,$$

but $u^{(0)} \neq u^{(1)}$, as otherwise the minimal period would be 1. Similarly, a period 3 orbit has

$$u^{(0)} = u^{(3)} = u^{(6)} = \cdots , \qquad u^{(1)} = u^{(4)} = u^{(7)} = \cdots , \qquad u^{(2)} = u^{(5)} = u^{(8)} = \cdots ,$$

with $u^{(0)}, u^{(1)}, u^{(2)}$ distinct. Stability implies that nearby iterates converge to this periodic solution.

For the logistic map, the period 2 orbit persists until $\lambda = \lambda_2 \approx 3.4495$, after which the iterates alternate between four values — a period 4 orbit. This again changes at $\lambda = \lambda_3 \approx 3.5441$, after which the iterates end up alternating between eight values. In fact, there is an increasing sequence of values

$$3 = \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \cdots ,$$

where, for any $\lambda_n < \lambda \leq \lambda_{n+1}$, the iterates eventually follow a period 2^n orbit. Thus, as λ passes through each value λ_n the period of the orbit doubles from 2^n to $2 \cdot 2^n = 2^{n+1}$, and the discrete dynamical system experiences a *bifurcation*. The bifurcation values λ_n lie closer and closer together, piling up on an eventual limit $\lambda_\star = \lim_{n \to \infty} \lambda_n \approx 3.5699$, at which point the period has become infinitely large. The entire phenomena is known as a *period doubling cascade*. Interestingly, the ratios of the distances between successive bifurcation points approaches a well-defined limit,

$$\frac{\lambda_{n+2} - \lambda_{n+1}}{\lambda_{n+1} - \lambda_n} \quad \longrightarrow \quad 4.6692\ldots , \tag{19.20}$$

known as *Feigenbaum's constant*. In the 1970's, the American physicist Mitchell Feigenbaum, [**53**], discovered that this period doubling cascade appears in a broad range of

discrete dynamical systems. Even more remarkably, in all cases, the corresponding ratios of distances between bifurcation points has the *same* limiting value. This was subsequently proved by Oscar Lanford in 1982, [**101**].

After λ passes the limiting value $\lambda_\star$, "all hell breaks loose". The iterates become completely chaotic†, moving at random over the interval $[0, 1]$. But this is not the end of the story. Embedded within this chaotic regime are certain small ranges of λ where the system settles down to a stable orbit, whose period is not necessarily a power of 2. In fact, there exist values of λ for which the iterates settle down to a stable orbit of period m for *any* positive integer m. For instance, as λ increases past $\lambda_{3\star} \approx 3.83$, a period 3 orbit appears for a while; then it experiences a succession of period doubling cascade of period 6, 12, 24, ... orbits, each persisting on a shorter and shorter interval of parameter values, until chaos breaks out yet again. There is a well-prescribed order in which the periodic cases appear, and each period m is followed by a very closely spaced sequence of period doubling bifurcations, of periods $2^n m$ for $n = 1, 2, 3, \ldots$, after which the iterates revert to completely chaotic behavior until the next periodic case emerges. The ratios of distances between bifurcation points have the same Feigenbaum limit (19.20). Finally, these periodic and chaotic windows all pile up on the ultimate parameter value $\lambda^\star_\star = 4$. And then, when $\lambda > 4$, all the iterates go off to ∞, and the system ceases to be interesting.

The reader is encouraged to write a simple computer program and perform some numerical experiments. In particular, Figure log■ shows the asymptotic behavior of the iterates for values of the parameter in the interesting range $2 < \lambda < 4$. The horizontal axis is λ, and the marked points show the ultimate fate of the iteration for the given value of λ. For instance, the single curve lying above low values of λ represents a fixed point; this bifurcates into a pair of curves representing a stable period 2 orbit, which then bifurcates into 4 curves representing a period 4 orbit, and so on. Chaotic behavior is indicated by a somewhat random pattern of points lying above the value of λ. To plot this figure, we ran the iteration $u^{(n)}$ for $0 \le n \le 100$, and then discarded the first 50 points, plotting the next 50 iterates $u^{(51)}, \ldots, u^{(100)}$. Investigation of the fine detailed structure of the logistic map requires yet more iterations with increased accuracy. In addition one should discard more of the initial iterates so as to give the system enough time to settle down to a stable periodic orbit or continue in a chaotic manner.

Remark: So far, we have only looked at real scalar iterative systems. Complex discrete dynamical systems display yet more remarkable and fascinating behavior. The complex version of the logistic iteration equation leads to the justly famous Mandelbrot set, [**102**], with its stunning, psychedelic fractal structure, [**120**].

The rich range of phenomena in evidence even in such extremely simple nonlinear iterative systems is astounding. While intimations of this first appeared in the late nineteenth century research of the influential French mathematician Henri Poincaré, serious investigations were delayed until the advent of the computer era, which precipitated an explosion of research activity in the area of dynamical systems. Similar period doubling

† The term "chaotic" does have a precise mathematical definition, but the reader can take it more figuratively for the purposes of this elementary introduction.

cascades and chaos can be found in a broad range of nonlinear systems, [**7**], and are often encountered in physical applications, [**107**]. A modern explanation of fluid turbulence is that it is a (very complicated) form of chaos.

Quadratic Convergence

Let us now return to the more mundane case when the iterates converge to a stable fixed point of the discrete dynamical system. In applications, we are interested in computing a precise[†] numerical value for the fixed point, and hence the speed of convergence of the iterates is of crucial importance.

According to Theorem 19.4, the convergence rate of an iterative system is essentially governed by the magnitude of the derivative $|\, g'(u^\star)\,|$ at the fixed point. The basic inequality (19.14) for the errors $e^{(k)} = u^{(k)} - u^\star$, namely

$$|\, e^{(k+1)}\,| \;\le\; \rho\,|\, e^{(k)}\,|,$$

is known as a *linear convergence estimate*. It means that the error decreases by a factor of at least ρ at each step. If the k^{th} iterate $u^{(k)}$ approximates the fixed point $u^\star$ correctly to m decimal places, so $|\, e^{(k)}\,| < .5 \times 10^{-m}$, then the $(k+1)^{\text{st}}$ iterate satisfies

$$|\, e^{(k+1)}\,| < .5 \times 10^{-m}\,\rho = .5 \times 10^{-m + \log_{10}\rho}.$$

More generally, for any $j > 0$,

$$|\, e^{(k+j)}\,| < .5 \times 10^{-m}\,\rho^j = .5 \times 10^{-m + j\,\log_{10}\rho},$$

which means that the $(k+j)^{\text{th}}$ iterate $u^{(k+j)}$ has at least[‡]

$$m - j\,\log_{10}\rho = m + j\,\log_{10}\rho^{-1}$$

correct decimal places. For instance, if $\rho = .1$ then each new iterate produces one new decimal place of accuracy (at least), while if $\rho = .9$ then it typically takes $22 \approx -1/\log_{10} .9$ iterates to produce just one additional accurate digit!

As a consequence, there is a huge advantage — particularly in the application of iterative methods to the numerical solution of equations — to arranging that $|\, g'(u^\star)\,|$ be as small as possible. The fastest convergence rate of all will occur when $g'(u^\star) = 0$. Now the constant ρ in (19.14) can be taken to be arbitrarily small, although the smaller ρ is, the smaller the interval $|\, v - u^\star\,| < \delta$ on which (19.14) applies, and so the closer one must be to the fixed point. Be that as it may, once the iterates start converging, they will get closer and closer to the fixed point, and so the rate of convergence will speed up accordingly. In fact, for such functions, the rate of convergence is not just slightly, but dramatically faster than linear.

[†] The degree of precision is to be specified by the user and the application.

[‡] Note that since $\rho < 1$, the logarithm $\log_{10}\rho^{-1} = -\log_{10}\rho > 0$ is positive.

Theorem 19.8. *Let $g(u) \in \mathrm{C}^2$. Suppose $u^\star = g(u^\star)$ is a fixed point such that $g'(u^\star) = 0$. Then, for all iterates $u^{(k)}$ sufficiently close to $u^\star$, the errors $e^{(k)} = u^{(k)} - u^\star$ satisfy the* quadratic convergence estimate

$$| e^{(k+1)} | \;\leq\; \sigma \, | e^{(k)} |^2 \tag{19.21}$$

for some constant $\sigma > 0$.

Proof: Just as in the proof of the linear convergence estimate (19.14), the proof relies on approximating the function by a simpler function near the fixed point. For linear convergence, an affine approximation sufficed, but in this case we require a higher order, quadratic approximation. Instead of the mean value formula (19.12), we now use the first order Taylor expansion (C.6) of g near $u^\star$:

$$g(u) = g(u^\star) + g'(u^\star)\,(u - u^\star) + \tfrac{1}{2}\, g''(w)\,(u - u^\star)^2, \tag{19.22}$$

where the error term depends on an (unknown) point w that lies between u and $u^\star$. At a fixed point, the constant term is $g(u^\star) = u^\star$. Furthermore, under our hypothesis $g'(u^\star) = 0$, and so the Taylor expansion (19.22) reduces to

$$g(u) - u^\star = \tfrac{1}{2}\, g''(w)\,(u - u^\star)^2.$$

Therefore,

$$| g(u) - u^\star | \leq \sigma \, | u - u^\star |^2, \tag{19.23}$$

where σ is chosen so that

$$\tfrac{1}{2}\, | g''(w) | \leq \sigma \tag{19.24}$$

for all w sufficiently close to $u^\star$. Therefore, the magnitude of σ is governed by the size of the *second derivative* of the iterative function $g(u)$ near the fixed point. We apply (19.23) to estimate the error

$$| e^{(k+1)} | = | u^{(k+1)} - u^\star | = | g(u^{(k)}) - g(u^\star) | \;\leq\; \sigma \, | u^{(k)} - u^\star |^2 = \sigma \, | e^{(k)} |^2,$$

which establishes the quadratic convergence estimate (19.21). *Q.E.D.*

Let us see how the quadratic estimate (19.21) speeds up the convergence rate. Following our earlier argument, suppose $u^{(k)}$ is correct to m decimal places, so

$$| e^{(k)} | < .5 \times 10^{-m}.$$

Then (19.21) implies that

$$| e^{(k+1)} | < .5 \times (10^{-m})^2 \, \sigma = .5 \times 10^{-2m + \log_{10} \sigma},$$

and so $u^{(k+1)}$ has $2m - \log_{10} \sigma$ accurate decimal places. If $\sigma \approx g''(u^\star)$ is of moderate size, we essentially *double* the number of accurate decimal places in just a single iterate! A second iteration will double the number of accurate digits yet again. Thus, the convergence of a quadratic iteration scheme is *extremely* rapid, and, barring round-off errors, one can produce any desired number of digits of accuracy in a very short time. For example, if we start with an initial guess that is accurate in the first decimal digit, then a linear iteration with $\rho = .1$ will require 49 iterations to obtain 50 decimal place accuracy, whereas a quadratic iteration (with $\sigma = 1$) will only require 6 iterations to obtain $2^6 = 64$ decimal places of accuracy!

Example 19.9. Consider the function

$$g(u) = \frac{4u^3 + 2u - 1}{3u^2 + 1}.$$

There is a unique fixed point $u^\star = g(u^\star)$ which is the solution to the cubic equation

$$u^3 + u - 1 = 0.$$

Note that

$$g'(u) = \frac{6u^4 + 6u^2 - 6u}{(3u^2+1)^2} = \frac{6u\,(u^3 + u - 1)}{(3u^2+1)^2},$$

and hence $g'(u^\star)$ vanishes at the fixed point. Theorem 19.8 implies that the iterations should exhibit quadratic convergence to the root. Indeed, we find, starting with $u^{(0)} = 0$, the following values. Not the dramatically faster convergence, especially when contrasted with the linearly convergent scheme based on ∎

For a general discrete dynamical system, the appearance of a quadratically convergent fixed point is a matter of luck. The construction of general purpose quadratically convergent iterative methods for solving equations will be the focus of the following Section 19.2.

Vector–Valued Iteration

Extending the preceding analysis to vector-valued iterative systems is not especially difficult. We will build on our experience with linear iterative systems, and so the reader should review the basic concepts and results from Chapter 10 before proceeding to the nonlinear cases presented here.

We begin by fixing a norm $\| \cdot \|$ on $\mathbb{R}^n$. Since we will also be computing the associated matrix norm $\| A \|$, as defined in Theorem 10.17, it may be computationally more convenient to adopt either the 1 or the ∞ norms rather than the standard Euclidean norm. As far as the theory goes, however, the precise choice of norm is unimportant.

We begin by defining the vector-valued counterpart of the basic linear convergence condition (19.18).

Definition 19.10. A function $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ is *Lipschitz continuous* at a point $\mathbf{a} \in \mathbb{R}^n$ if there exists a constant $\rho \geq 0$, known as the *Lipschitz constant*, such that

$$\| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{a}) \| \;\leq\; \rho \, \| \mathbf{u} - \mathbf{a} \| \tag{19.25}$$

for all $\mathbf{u}$ sufficiently close to $\mathbf{a}$, i.e., $\| \mathbf{u} - \mathbf{a} \| < \delta$ for some fixed $\delta > 0$.

Example 19.11. Consider the function $\mathbf{g}(\mathbf{u}) = \begin{pmatrix} |\, u - v \,| \\ \max\{\, |\, u \,|, |\, v \,| \,\} \end{pmatrix}$, defined for $\mathbf{u} = (\, u, v \,)^T \in \mathbb{R}^2$. Although $\mathbf{g}$ is not differentiable, it does satisfy the Lipschitz estimate (19.25) for the 1 norm $\| \mathbf{u} \|_1 = |\, u \,| + |\, v \,|$. Indeed,

$$\begin{aligned} \| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{a}) \| \;&\leq\; \Big|\; |\, u - v \,| - |\, a - b \,| \;\Big| \;+\; \Big|\; \max\{\, |\, u \,|, |\, v \,| \,\} - \max\{\, |\, a \,|, |\, b \,| \,\} \;\Big| \\ &\leq\; 2 \big(\; |\, u - a \,| + |\, v - b \,| \;\big) = 2 \, \| \mathbf{u} - \mathbf{a} \|_1. \end{aligned}$$

Thus, (19.25) holds with uniform Lipschitz constant $\rho = 2$.

Remark: The notion of "Lipschitz continuity" appears to depend on the underlying choice of matrix norm. However, the fact that all norms on a finite-dimensional vector space are essentially equivalent — see Theorem 3.19 — implies that this concept is, in fact, independent of the choice of norm. However, one should keep in mind that the value of the Lipschitz constant ρ is norm-dependent.

The Lipschitz inequality (19.25) provides an immediate proof of the basic convergence theorem for iteration of a discrete dynamical system (19.1). Recall that a fixed point is called *asymptotically stable* if $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ for every initial condition $\mathbf{u}^{(0)} = \mathbf{c}$ sufficiently close to $\mathbf{u}^\star$.

Theorem 19.12. *If $\mathbf{u}^\star = \mathbf{g}(\mathbf{u}^\star)$ is a fixed point for the discrete dynamical system (19.1) and $\mathbf{g}$ is Lipschitz continuous at $\mathbf{u}^\star$ with Lipschitz constant $\rho < 1$, then $\mathbf{u}^\star$ is an asymptotically stable fixed point.*

Proof: The proof is a copy of the last part of the proof of Theorem 19.4. We write

$$\| \mathbf{u}^{(k+1)} - \mathbf{u}^\star \| = \| \mathbf{g}(\mathbf{u}^{(k)}) - \mathbf{g}(\mathbf{u}^\star) \| \le \rho \| \mathbf{u}^{(k)} - \mathbf{u}^\star \|,$$

using the assumed Lipschitz estimate (19.25). Iterating this basic inequality immediately demonstrates that

$$\| \mathbf{u}^{(k)} - \mathbf{u}^\star \| \ \le \ \rho^k \| \mathbf{u}^{(0)} - \mathbf{u}^\star \| \qquad \text{for} \qquad k = 0, 1, 2, 3, \dots .$$

Since $\rho < 1$, the right hand side tends to 0 as $k \to \infty$, and hence $\mathbf{u}^{(k)} \to \mathbf{u}^\star$. *Q.E.D.*

For more complicated functions, the direct verification of the Lipschitz inequality (19.25) is not particularly easy. However, as in the scalar case, any continuously differentiable function is automatically Lipschitz continuous.

Theorem 19.13. *If $\mathbf{g}(\mathbf{u}) \in \mathrm{C}^1$ has continuous first order partial derivatives for all $\mathbf{u}$ sufficiently close to $\mathbf{u}^\star$, then $\mathbf{g}$ is Lipschitz continuous at $\mathbf{u}^\star$.*

Proof: According to the first order Taylor expansion (C.10) of a vector-valued function at a point $\mathbf{u}^\star$ takes the form

$$\mathbf{g}(\mathbf{u}) = \mathbf{g}(\mathbf{u}^\star) + \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) + R(\mathbf{u} - \mathbf{u}^\star). \tag{19.26}$$

Here

$$\mathbf{g}'(\mathbf{u}) = \begin{pmatrix} \dfrac{\partial g_1}{\partial u_1} & \dfrac{\partial g_1}{\partial u_2} & \cdots & \dfrac{\partial g_1}{\partial u_n} \\ \dfrac{\partial g_2}{\partial u_1} & \dfrac{\partial g_2}{\partial u_2} & \cdots & \dfrac{\partial g_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial g_n}{\partial u_1} & \dfrac{\partial g_n}{\partial u_2} & \cdots & \dfrac{\partial g_n}{\partial u_n} \end{pmatrix}, \tag{19.27}$$

is the $n \times n$ *Jacobian matrix* of the vector-valued function $\mathbf{g}$ whose entries are the partial derivatives of its individual components. The remainder term in (19.26) satisfies[†]

$$\| R(\mathbf{v}) \| \le \sigma \| \mathbf{v} \|^2 \qquad \text{whenever} \qquad \| \mathbf{v} \| \le \varepsilon,$$

for some positive constant $\sigma > 0$. If the corresponding matrix norm of the Jacobian matrix at $\mathbf{u}^\star$ satisfies

$$\| \mathbf{g}'(\mathbf{u}^\star) \| = \rho^\star,$$

then, by the triangle inequality and the definition (10.23) of matrix norm,

$$\begin{aligned} \| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{u}^\star) \| &= \| \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) + R(\mathbf{u} - \mathbf{u}^\star) \| \le \| \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) \| + \| R(\mathbf{u} - \mathbf{u}^\star) \| \\ &\le \| \mathbf{g}'(\mathbf{u}^\star) \|\, \| \mathbf{u} - \mathbf{u}^\star \| + \sigma \| \mathbf{u} - \mathbf{u}^\star \|^2 \le (\rho^\star + \varepsilon\,\sigma) \| \mathbf{u} - \mathbf{u}^\star \|, \end{aligned} \tag{19.28}$$

whenever $\| \mathbf{u} - \mathbf{u}^\star \| \le \varepsilon$. This proves that $\mathbf{g}$ is Lipschitz continuous at $\mathbf{u}^\star$ with Lipschitz constant $\rho = \rho^\star + \varepsilon\,\sigma$. Note that, by choosing ε small enough, we can ensure that the Lipschitz constant ρ is arbitrarily close to the matrix norm $\rho^\star$. *Q.E.D.*

For a continuously differentiable function, then, asymptotic stability is a consequence of the size, or, more correctly, the *spectral radius* of the Jacobian matrix at the fixed point.

Theorem 19.14. *Suppose* $\mathbf{g}(\mathbf{u}) \in \mathrm{C}^2$. *If* $\mathbf{u}^\star = \mathbf{g}(\mathbf{u}^\star)$ *is a fixed point such that* $\mathbf{g}'(\mathbf{u}^\star)$ *is a convergent matrix, then* $\mathbf{u}^\star$ *is asymptotically stable. The rate of convergence of the iterative scheme* $\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)})$ *to* $\mathbf{u}^\star$ *is governed by the spectral radius of* $\mathbf{g}'(\mathbf{u}^\star)$.

Proof: If $\mathbf{g}'(\mathbf{u}^\star)$ is convergent, and hence has spectral radius strictly less than 1, then Corollary 10.29 assures us that there exists a matrix norm such that

$$\| \mathbf{g}'(\mathbf{u}^\star) \| = \rho^\star < 1. \tag{19.29}$$

Defining σ as in the proof of Theorem 19.13, we then choose $\varepsilon > 0$ so that

$$\rho = \rho^\star + \varepsilon\,\sigma < 1.$$

Then (19.28) implies that

$$\| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{u}^\star) \| \le \rho \| \mathbf{u} - \mathbf{u}^\star \|, \qquad \text{provided} \qquad \| \mathbf{u} - \mathbf{u}^\star \| < \varepsilon.$$

As before, this suffices to proves convergence of the iterates to $\mathbf{u}^\star$. *Q.E.D.*

Example 19.15. ∎

Theorem 19.14 tells us that initial values $\mathbf{u}^{(0)}$ that are sufficiently near a stable fixed point $\mathbf{u}^\star$ are guaranteed to converge to it. In the linear case, closeness of the initial data to the fixed point was not, in fact, an issue; all stable fixed points are, in fact, globally stable. For nonlinear iteration, it is of critical importance, and one does not typically expect iteration starting with far away initial data to converge to the desired fixed point.

[†] We can use any convenient norm on $\mathbb{R}^n$.

An interesting (and difficult) problem is to determine the so-called *basin of attraction* of a stable fixed point, defined as the set of all initial data that ends up converging to it. As in the elementary logistic map (19.19), initial values that lie outside a basin of attraction can lead to divergent iterates, periodic orbits, or even exhibit chaotic behavior. The full range of possible phenomena is a subject of contemporary research in dynamical systems theory and in numerical analysis, [**7**].

The smaller the spectral radius or matrix norm of the Jacobian matrix at the fixed point, the faster the iterates converge to it. As in the scalar case, quadratic convergence will occur when the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star) = \mathrm{O}$ is the zero matrix[†], i.e., *all* first order partial derivatives of the components of $\mathbf{g}$ vanish at the fixed point. The quadratic convergence estimate

$$\| \mathbf{u}^{(k+1)} - \mathbf{u}^\star \| \;\le\; \sigma \, \| \mathbf{u}^{(k)} - \mathbf{u}^\star \|^2 \tag{19.30}$$

is a consequence of the second order Taylor expansion at the fixed point. Details of the proof are left as an exercise.

Example 19.16. ∎

In general, the existence of a fixed point of an iterative system is *not* automatic. One way is to observe the iterates starting with suitably selected initial data; if they converge, then Proposition 19.2 assures us that their limit is a fixed point. There is one important class of maps for which we have a theoretical justification, not only of the existence, but also the uniqueness of a fixed point.

Definition 19.17. A map $\mathbf{g}\colon \Omega \to \Omega$ is called a *contraction mapping* if it has Lipschitz constant $\rho < 1$ at all points in Ω.

Therefore, applying a contraction mapping reduces the distance between points. As a result, a contraction mapping shrinks the size of its domain; see Figure contract∎. As a result, as the iterations proceed, the domain gets smaller and smaller and the iterates become trapped. If the original domain is closed and bounded, then it is forced to shrink down to a single point, which is the unique fixed point of the iterative system.

The simplest example of a contraction mapping is the scaling map $\mathbf{g}(\mathbf{u}) = \rho\,\mathbf{u}$ with $0 < \rho < 1$. Starting with the unit ball $B_1 = \big\{ \| \mathbf{u} \| \le 1 \big\}$, at the k^{th} iteration the points have been mapped into a contracted sphere of radius ρ^k. As $k \to \infty$ these contracted domains become smaller and smaller, converging in the limit to the unique fixed point $\mathbf{u}^\star = \mathbf{0}$. A precise statement of the *Contraction Mapping Theorem* follows; see [**map**] for the proof.

Theorem 19.18. *If $\mathbf{g}\colon \Omega \to \Omega$ is a contraction mapping defined on a closed bounded domain $\Omega \subset \mathbb{R}^n$ then $\mathbf{g}$ admits a unique fixed point $\mathbf{u}^\star \in \Omega$. Moreover, starting with any initial point $\mathbf{u}^{(0)} \in \Omega$, the iterates necessarily converge to the fixed point $\mathbf{u}^{(k)} \to \mathbf{u}^\star$.*

More sophisticated, powerful fixed point theorems require advanced knowledge of algebraic topology and will not be developed in this text. See [**fixed**] for details.

[†] Having zero spectral radius is not sufficient for quadratic convergence; see Exercise ∎.

19.2. Solution of Equations and Systems.

The solution of nonlinear equations and systems of equations is, of course, a problem of utmost importance in mathematics and its manifold applications. In the general situation, we are given a collection of m functions depending upon n variables, and we are interested in finding all solutions $\mathbf{u} = (\, u_1, u_2, \dots, u_n \,)^T$ to the system

$$f_1(u_1, \dots, u_n) = 0, \qquad \dots \qquad f_m(u_1, \dots, u_n) = 0. \tag{19.31}$$

In practice, as in the linear case, we are primarily interested in the case when the number of equations is equal to the number of unknowns, $m = n$, as one can only expect both existence and uniqueness of solutions in such situations. This point will be discussed in further detail below.

There is no universal direct solution method for nonlinear equations and systems comparable to Gaussian elimination. As a result, numerical solution techniques rely almost exclusively on iterative algorithms. In this section, we shall present the principal methods for numerically approximating the solution(s) to a system. We shall only discuss general purpose algorithms. Specialized methods for particular classes of equations, e.g., methods designed for solving polynomial equations, can be found in numerical analysis texts, e.g., [**30**, **121**]. Of course, the most important specialized methods — those designed for solving linear systems — will continue to play a critical role, even in the nonlinear regime.

The Bisection Method

We begin, as always, with the scalar case. Thus, we are given a real-valued function $f\colon \mathbb{R} \to \mathbb{R}$, and seek its *roots*, i.e., the real[†] solution(s) to the scalar equation

$$f(u) = 0. \tag{19.32}$$

Here are some prototypical examples:

(*a*) Find the roots of the quintic polynomial equation

$$u^5 + u + 1 = 0. \tag{19.33}$$

Graphing the left hand side of the equation, as in Figure 19.1, convinces us that there is just one real root, lying somewhere between -1 and $-.5$. While there are explicit algebraic formulas for the roots of quadratic, cubic, and quartic polynomials, a famous theorem[‡] due to the Norwegian mathematician Nils Henrik Abel in the early 1800's states that there is *no* such formula for generic fifth order polynomial equations.

(*b*) As noted in Example 19.5, the trigonometric Kepler equation

$$u - \epsilon \sin u = m$$

[†] Complex roots to complex equations will be discussed later.

[‡] A modern proof of this fact relies on Galois theory, [**62**].

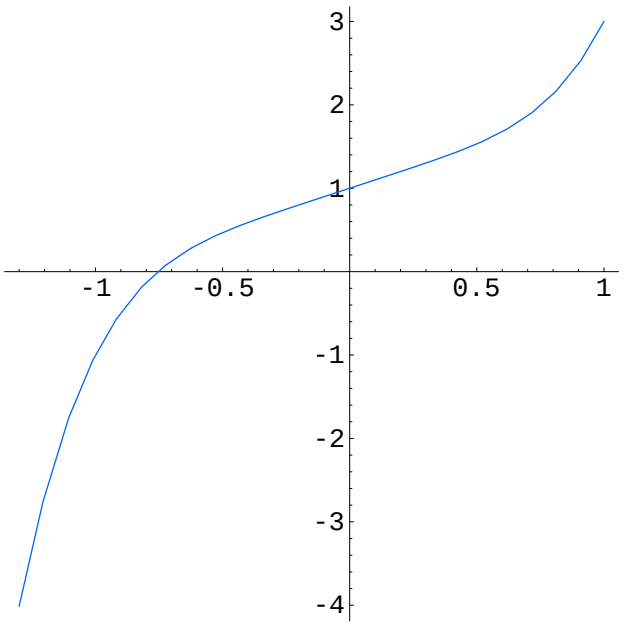


Figure 19.1. Graph of $u^5 + u + 1$.

arises in the study of planetary motion. Here ϵ, m are fixed constants, and we seek a corresponding solution u. We have already looked at one iterative solution method for this equation.

(*c*) Chemistry ■

The most primitive method for solving scalar equations, and *the only one that is guaranteed to work in all cases*, is the bisection algorithm. While it has an iterative flavor, it cannot be properly classed as a method governed by functional iteration as defined in the preceding section, and so must be studied directly in its own right.

The starting point is the Intermediate Value Theorem, which we state in simplified form, without proof. See Figure 19.2 for an illustration, and [**9**] for a proof.

Theorem 19.19. *Let $f(u)$ be a continuous scalar function. Suppose we can find two points $a < b$ where the values of $f(a)$ and $f(b)$ take* opposite *signs, so either $f(a) < 0$ and $f(b) > 0$, or $f(a) > 0$ and $f(b) < 0$. Then there exists at least one point $a < u^\star < b$ where $f(u^\star) = 0$.*

The hypothesis can be compactly written as $f(a)\, f(b) < 0$. Note that if $f(a) = 0$ or $f(b) = 0$, then finding a root is trivial. If $f(a)$ and $f(b)$ have the same sign, then there may or may not be a root in between. Figure 19.3 plots the functions $u^2 + 1$, u^2 and $u^2 - 1$, on the interval $-2 \le u \le 2$. The first has two simple roots; the second has a single double root, while the third has no root. Also, continuity of the function on the entire interval $[a, b]$ is an essential hypothesis. For example, the function $f(u) = 1/u$ satisfies $f(-1) = -1$ and $f(1) = 1$, but there is no root to the equation $1/u = 0$.

Note carefully that the Theorem 19.19 does *not* say there is a unique root between a and b. There may be many roots, or even, in pathological examples, infinitely many. All the theorem guarantees is that there is at least one root.

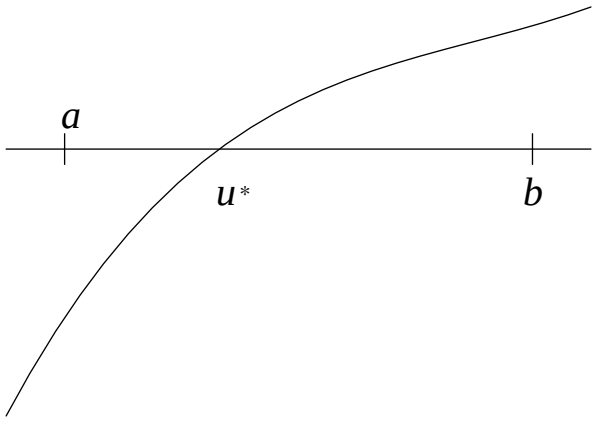


Figure 19.2. Intermediate Value Theorem.

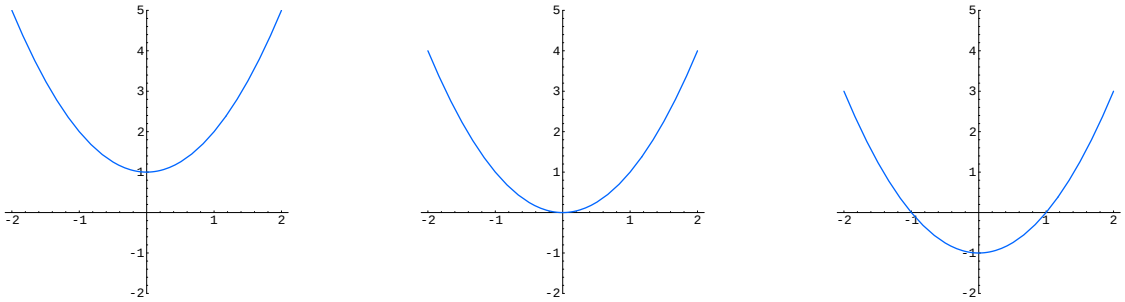


Figure 19.3. Roots of Functions.

Once we are assured that a root exists, bisection amounts to a "divide and conquer" strategy. Starting with the endpoints, the goal is to locate a root $a < u^\star < b$ between them. Lacking any additional evidence, a good strategy would be to try the midpoint $c = \frac{1}{2}(a+b)$ as a first guess for the root. If, by some miracle, $f(c) = 0$, then we are done, since we have found a solution! Otherwise (and typically) we look at the sign of $f(c)$. There are two possibilities. If $f(a)$ and $f(c)$ are of opposite signs, then the Intermediate Value Theorem tells us that there is a root $u^\star$ lying between $a < u^\star < c$. Otherwise, $f(c)$ and $f(b)$ must have opposite signs, and so there is a root $c < u^\star < b$. In either event, we apply the same method to the interval in which we are assured a root lies, and repeat the procedure. Each iteration halves the length of the interval, and chooses the half in which a root must be. (There may, of course, be a root in the other half, but we cannot be sure of this, and so discard it from further consideration.) The root we home in on lies trapped in intervals of smaller and smaller width, and so convergence of the method is guaranteed. Figure bisect■ illustrates the steps in a particular example.

k	$u^{(k)}$	$v^{(k)}$	$w^{(k)} = \frac{1}{2}(u^{(k)} + v^{(k)})$	$f(w^{(k)})$
0	1	2	1.5	.75
1	1	1.5	1.25	−.1875
2	1.25	1.5	1.375	.2656
3	1.25	1.375	1.3125	.0352
4	1.25	1.3125	1.2813	−.0771
5	1.2813	1.3125	1.2969	−.0212
6	1.2969	1.3125	1.3047	.0069
7	1.2969	1.3047	1.3008	−.0072
8	1.3008	1.3047	1.3027	−.0002
9	1.3027	1.3047	1.3037	.0034
10	1.3027	1.3037	1.3032	.0016
11	1.3027	1.3032	1.3030	.0007
12	1.3027	1.3030	1.3029	.0003
13	1.3027	1.3029	1.3028	.0001
14	1.3027	1.3028	1.3028	−.0000

Example 19.20. The roots of the quadratic equation

$$f(u) = u^2 + u - 3 = 0$$

are given by the quadratic formula

$$u_1^\star = \frac{-1 + \sqrt{13}}{2} \approx 1.302775\ldots, \qquad u_2^\star = \frac{-1 - \sqrt{13}}{2} \approx -2.302775\ldots.$$

Let us see how one might approximate them by applying the Bisection Algorithm. We start the procedure by choosing the points $a = u^{(0)} = 1$, $b = v^{(0)} = 2$, noting that $f(1) = -1$ and $f(2) = 3$ have opposite signs and hence we are guaranteed that there is at least one root between 1 and 2. In the first step we look at the midpoint of the interval $[1, 2]$, which is 1.5, and evaluate $f(1.5) = .75$. Since $f(1) = -1$ and $f(1.5) = .75$ have opposite signs, we know that there is a root lying between 1 and 1.5. Thus, we use $u^{(1)} = 1$ and $v^{(1)} = 1.5$ as the endpoints of the next interval, and continue. The next midpoint is at 1.25, where $f(1.25) = -.1875$ has the opposite sign to $f(1.5) = .75$, and so a root lies between $u^{(2)} = 1.25$ and $v^{(2)} = 1.5$. The process is then iterated as long as desired — or, more practically, as long as your computer's precision does not become an issue.

The accompanying table displays the result of the algorithm, rounded off to four decimal digits. Thus, after 14 iterations the Bisection Algorithm has computed the positive root $u_1^\star$ correctly to 4 decimal places. A similar bisection starting with the interval from $u^{(1)} = -3$ to $v^{(1)} = -2$ will produce the negative root.

The formal implementation of the algorithm is governed by the following program. The endpoints of the k^{th} interval are denoted by $u^{(k)}$ and $v^{(k)}$. The midpoint is $w^{(k)} = \frac{1}{2}\big(u^{(k)} + v^{(k)}\big)$, and the main decision is whether $w^{(k)}$ should be the right or left hand endpoint of the new interval. The integer n, governing the number of iterations, will be prescribed in accordance with how close we wish to approximate the solution $u^\star$.

The algorithm produces two sequences of approximations such that $u^{(k)} < u^\star < v^{(k)}$ lies between them. Both converge monotonically to the root, one from below and the other from above:

$$a = u^{(0)} \le u^{(1)} \le u^{(2)} \le \cdots \le u^{(k)} \longrightarrow u^\star \longleftarrow v^{(k)} \le \cdots \le v^{(2)} \le v^{(1)} \le v^{(0)} = b.$$

In other words, the solution $u^\star$ is trapped inside a sequence of intervals $[\,u^{(k)}, v^{(k)}\,]$ of progressively shorter and shorter length. Since we cut the interval in half at each step of the algorithm, the length of the interval $[\,u^{(k)}, v^{(k)}\,]$ is exactly half that of $[\,u^{(k-1)}, v^{(k-1)}\,]$, and so

$$v^{(k)} - u^{(k)} = \tfrac{1}{2}\,(v^{(k-1)} - u^{(k-1)}).$$

Iterating this formula, we conclude that

$$v^{(n)} - u^{(n)} = \big(\tfrac{1}{2}\big)^n\,(v^{(0)} - u^{(0)}) = \big(\tfrac{1}{2}\big)^n\,(b-a)\,.$$

The final approximation

$$w^{(n)} = \tfrac{1}{2}(u^{(n)} + v^{(n)})$$

lies in middle of its interval, and hence must be within a distance

$$|\,w^{(n)} - u^\star\,| \le \tfrac{1}{2}\,(v^{(n)} - u^{(n)}) = \big(\tfrac{1}{2}\big)^{n+1}\,(b-a)$$

of the root. Consequently, if we want to approximate the root within a prescribed tolerance ε, we should choose the number of iterations n so that

$$\big(\tfrac{1}{2}\big)^{n+1}\,(b-a) < \varepsilon, \qquad \text{or} \qquad n > \log_2 \frac{b-a}{\varepsilon} - 1\,. \tag{19.34}$$

The Bisection Method

```
start
    if f(a) f(b) < 0 set u^(0) = a, v^(0) = b
    for k = 0 to n − 1
        set w^(k) = ½(u^(k) + v^(k))
        if f(w^(k)) = 0, stop; print u* = w^(k)
        if f(w^(k)) f(u^(k)) < 0, set u^(k+1) = w^(k), v^(k+1) = v^(k)
            else set u^(k+1) = u^(k), v^(k+1) = w^(k)
    next k
    print u* = w^(n) = ½(u^(n) + v^(n))
end
```

Theorem 19.21. *If $f(u)$ is a continuous function, with $f(a)\,f(b) < 0$, then the bisection algorithm starting with $u^{(0)} = a, v^{(0)} = b$ will converge to a solution to $f(u) = 0$ lying between a and b. After n steps, the midpoint $w^{(n)} = \frac{1}{2}\,(u^{(n)} + v^{(n)})$ will be within a tolerance of $\varepsilon = 2^{-n-1}(b - a)$ of the solution.*

For example, in the case of the quadratic equation in Example 19.20, after 14 iterations, we have approximated the positive root to within

$$\varepsilon = \left(\tfrac{1}{2}\right)^{15}(2 - 1) \;\approx\; 3.052 \times 10^{-5},$$

reconfirming our observation that we have accurately computed the first four decimal places of the root. If we need 10 decimal places, we set our tolerance to $\varepsilon = 10^{-11}$, and so, according to (19.34), must perform $n = 36 > 35.54 \approx \log_2 10^{11} - 1$ successive bisections.

Example 19.22. As noted at the beginning of this section, the quintic equation

$$f(u) = u^5 + u + 1 = 0$$

has one real root, whose value can be readily computed by bisection. We start the algorithm with the initial points $u^{(0)} = -1$, $v^{(0)} = -.5$, noting that $f(-1) = -1 < 0$ while $f(0) = 1 > 0$ are of opposite signs. In order to compute the root to 6 decimal places, we set $\varepsilon = 10^{-7}$ in (19.34), and so need to perform $n = 23 > 22.25 \approx \log_2 10^7 - 1$ bisections. Indeed, the algorithm produces the approximation $u^\star \approx -0.754878$ to the root, and the displayed digits are guaranteed to be accurate.

Fixed Point Methods

The Bisection method converges in all cases — provided it can be properly started by locating two points where the function takes opposite signs. This may be tricky if the function has two very closely spaced roots and is, say, negative only for a very small

interval between them, and may be impossible for multiple roots, e.g., the root $u^\star = 0$ of the quadratic function $f(u) = u^2$. When applicable, its convergence rate is completely predictable, but not especially fast. Worse, it has no immediately apparent extension to systems of equations, since there is *no* counterpart to the Intermediate Value Theorem for vector-valued functions.

Most other methods for solving equations rely on some form of fixed point iteration. Thus, we seek to replace the system of equations (19.32) with a fixed point system

$$\mathbf{u} = \mathbf{g}(\mathbf{u}). \tag{19.35}$$

The key requirements are

(*a*) The solution $\mathbf{u}^\star$ to (19.32) is also a fixed point for equation (19.35), and

(*b*) $\mathbf{u}^\star$ is, in fact a stable fixed point, so the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star)$ is a convergent matrix, or, slightly more restrictively, $\| \mathbf{g}'(\mathbf{u}^\star) \| < 1$ for a given matrix norm.

If both requirements are satisfied, then, *provided we choose the initial iterate* $\mathbf{u}^{(0)} = \mathbf{c}$ *sufficiently close to* $\mathbf{u}^\star$, the iterates $\mathbf{u}^{(k)}$ will converge to the desired solution $\mathbf{u}^\star$ as $k \to \infty$. Thus, the key to the practical use of functional iteration for solving equations is the proper design of an iterative system, coupled with a reasonably good initial guess for the solution.

Example 19.23. To solve the cubic equation

$$f(u) = u^3 - u - 1 = 0 \tag{19.36}$$

we note that $f(1) = -1$ while $f(2) = 5$, and so there is a root between 1 and 2. Indeed, the bisection algorithm gives the approximate value $u^\star \approx 1.3247$ after 17 iterations.

Let us try to find the same root by fixed point iteration. As a first, naïve, guess, we rewrite the cubic equation in fixed point form

$$u = 1 - u^3 = \widetilde{g}(u).$$

Starting with the initial guess $u^{(0)} = 1.5$, the successive iterates are given by

$$u^{(k+1)} = \widetilde{g}(u^{(k)}) = 1 - (u^{(k)})^3, \qquad k = 0, 1, 2, \dots .$$

However, their values

$$\begin{aligned} &u^{(0)} = 1.5, && u^{(1)} = -2.375, && u^{(2)} = 14.3965, \\ &u^{(3)} = -2,983, && u^{(4)} = 2.654 \times 10^{10}, && u^{(5)} = -1.869 \times 10^{31}, \quad \dots \end{aligned}$$

rapidly become unbounded and fail to converge. This could have been predicted by the convergence criterion in Theorem 19.4. Indeed, $\widetilde{g}'(u) = -3\,u^2$ and so $|\,\widetilde{g}'(u)\,| > 3$ for all $1 \leq u$, including the root $u^\star$. This means that $u^\star$ is an unstable fixed point, and we cannot expect the iterates to converge to it.

On the other hand, we can rewrite the equation (19.36) in the alternative iterative form

$$u = \sqrt[3]{1 + u} \; = g(u).$$

In this case

$$0 \;\leq\; g'(u) \;=\; \frac{1}{3(1+u)^{2/3}} \;\leq\; \frac{1}{3} \qquad \text{for} \qquad u > 0.$$

Thus, the stability condition (19.13) is satisfied, and we anticipate convergence at a rate of at least $\frac{1}{3}$. (The bisection method converges more slowly, at rate $\frac{1}{2}$.) Indeed, the first few iterates $u^{(k+1)} = \sqrt[3]{1 + u^{(k)}}$ are

$$1.5, \qquad 1.3571, \qquad 1.33086, \qquad 1.32588, \qquad 1.32494, \qquad 1.32476, \qquad 1.32473,$$

and we have converged to the root, correct to four decimal places, in only 6 iterations.

Newton's Method

As we learned in Section 19.1, the speed of convergence of an iterative method based on a scalar function $g(u)$ is governed by the magnitude of its derivative, $|\, g'(u^\star)\,|$, at the fixed point. Thus, to design an iterative method to solve an equation $f(u) = 0$, we need

(a) a function $g(u)$ whose fixed points $u^\star$ coincide with the solutions,

(b) whose derivative at the fixed point is as small as possible.

In particular, if we can arrange that $g'(u^\star) = 0$, then, instead of a relatively slow linear convergence rate, the numerical solution method will satisfy the dramatically faster quadratic convergence estimate of Theorem 19.8, with all its consequent advantages.

Now, the first condition requires that $g(u) = u$ whenever $f(u) = 0$. A little thought will convince you that the iterative function should take the form

$$g(u) = u - \lambda(u)\, f(u), \tag{19.37}$$

where $\lambda(u)$ is a reasonably nice function. If $f(u^\star) = 0$, then clearly $u^\star = g(u^\star)$, and so $u^\star$ is a fixed point. The converse holds provided $\lambda(u) \neq 0$ is never zero.

For a quadratically convergent method, the second requirement is that the derivative of $g(u)$ be zero at the fixed point solutions. We compute

$$g'(u) = 1 - \lambda'(u)\, f(u) - \lambda(u)\, f'(u).$$

Thus, $g'(u^\star) = 0$ at a solution to $f(u^\star) = 0$ if and only if

$$0 = 1 - \lambda'(u^\star)\, f(u^\star) - \lambda(u^\star)\, f'(u^\star) = 1 - \lambda(u^\star)\, f'(u^\star).$$

Consequently, we should require that

$$\lambda(u^\star) = \frac{1}{f'(u^\star)} \tag{19.38}$$

to ensure a quadratically convergent iterative scheme. This assumes that $f'(u^\star) \neq 0$, which means that $u^\star$ is a *simple root* of f. We leave aside multiple roots, which require a different argument and method, to be outlined in Exercise ■.

Of course, there are many functions $\lambda(u)$ that satisfy (19.38), since we only need to specify its value at a single point. The problem is that we do not know $u^\star$ — after all this is what we are trying to compute — and so cannot compute the value of the derivative of f there. However, we can circumvent this apparent difficulty by a simple device: we impose equation (19.38) at all points,

$$\lambda(u) = \frac{1}{f'(u)}\,, \tag{19.39}$$

which certainly guarantees that it hold at the solution $u^\star$. The result is the function

$$g(u) = u - \frac{f(u)}{f'(u)}, \tag{19.40}$$

that yields the iteration scheme known as *Newton's method*. It dates back to Isaac Newton, the founder of the calculus, and, to this day, remains *the* most important general purpose algorithm for solving equations. Newton's method starts with an initial guess $u^{(0)}$ to be supplied by the user, and then successively computes

$$u^{(k+1)} = u^{(k)} - \frac{f(u^{(k)})}{f'(u^{(k)})}. \tag{19.41}$$

Provided the initial guess is sufficiently close, the iterates $u^{(k)}$ are guaranteed to converge to the (simple) root $u^\star$ of f.

Theorem 19.24. *Suppose $f(u) \in \mathrm{C}^2$ is twice continuously differentiable. Let $u^\star$ be a solution to the equation $f(u) = 0$ such that $f'(u^\star) \neq 0$. Given an initial guess $u^{(0)}$ sufficiently close to $u^\star$, the Newton iteration scheme* (19.41) *converges at a quadratic rate to the solution $u^\star$.*

Proof: By continuity, if $f'(u^\star) \neq 0$, then $f'(u) \neq 0$, and hence the Newton iterative function (19.40) is well defined and continuously differentiable for all u sufficiently close to $u^\star$. Since $g'(u) = f(u)\, f''(u)/f'(u)^2$, we have $g'(u^\star) = 0$, as promised by our construction. Hence, the result is an immediate consequence of Theorem 19.8. *Q.E.D.*

Example 19.25. Consider the cubic equation

$$f(u) = u^3 - u - 1 = 0,$$

that we already solved in Example 19.23. The function used in the Newton iteration is

$$g(u) = u - \frac{f(u)}{f'(u)} = u - \frac{u^3 - u - 1}{3\,u^2 - 1},$$

which is well-defined as long as $u \neq \pm \frac{1}{\sqrt{3}}$. We will try to avoid these singular points. The iterative procedure

$$u^{(k+1)} = g(u^{(k)}) = u^{(k)} - \frac{(u^{(k)})^3 - u^{(k)} - 1}{3\,(u^{(k)})^2 - 1}$$

with initial guess $u^{(0)} = 1.5$ produces the following values:

$$1.5, \qquad 1.34783, \qquad 1.32520, \qquad 1.32472,$$

which gives the root correctly to 5 decimal places after only three iterations. The quadratic convergence of Newton's method implies that, roughly, each new iterate doubles the number of correct decimal places. Thus, if we need to compute the root accurately to 40

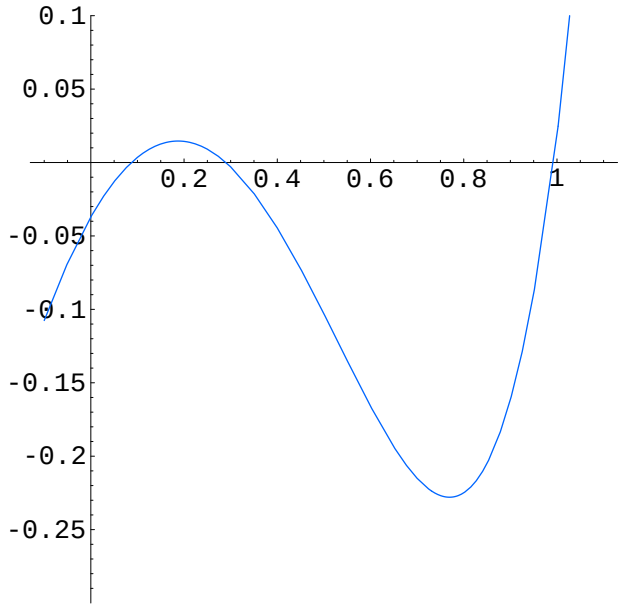


Figure 19.4. The function $f(u) = u^3 - \frac{3}{2}\,u^2 + \frac{5}{9}\,u - \frac{1}{27}$.

decimal places†, it would only require 3 further iterations! this underscores the tremendous advantage that the Newton algorithm offers over competing methods such as bisection or naïve iteration.

Example 19.26. The cubic polynomial equation

$$f(u) = u^3 - \tfrac{3}{2}\,u^2 + \tfrac{5}{9}\,u - \tfrac{1}{27} = 0$$

has

$$f(0) = -\tfrac{1}{27}\,, \qquad f\big(\tfrac{1}{3}\big) = \tfrac{1}{54}\,, \qquad f\big(\tfrac{2}{3}\big) = -\tfrac{1}{27}\,, \qquad f(1) = \tfrac{1}{54}\,.$$

The Intermediate Value Theorem 19.19 guarantees that there are three roots on the interval $[0,1]$: one between 0 and $\frac{1}{3}$, the second between $\frac{1}{3}$ and $\frac{2}{3}$, and the third between $\frac{2}{3}$ and 1. The graph in Figure 19.4 reconfirms this observation. Since we are dealing with a cubic polynomial, there are no other roots.

It takes sixteen iterations of the bisection algorithm starting with the three subintervals $\big[\,0\,,\tfrac{1}{3}\,\big]$, $\big[\,\tfrac{1}{3}\,,\tfrac{2}{3}\,\big]$ and $\big[\,\tfrac{2}{3}\,,1\,\big]$ to produce the roots to six decimal places:

$$u_1^\star \;\approx\; .085119, \qquad u_2^\star \;\approx\; .451805, \qquad u_3^\star \;\approx\; .963076.$$

Incidentally, if we start with the interval $[0,1]$ and apply bisection, we converge (perhaps surprisingly) to the largest root $u_3^\star$ in 17 iterations.

Fixed point iteration based on the formulation

$$u = g(u) = -\,u^3 + \tfrac{3}{2}\,u^2 + \tfrac{4}{9}\,u + \tfrac{1}{27}$$

† This assumes we are working in a sufficiently high precision arithmetic so as to avoid round-off errors.

can be used to find the first and third roots, but not the second root. For instance, starting with $u^{(0)} = 0$ produces $u_1^\star$ to 5 decimal places after 23 iterations, whereas starting with $u^{(0)} = 1$ produces $u_3^\star$ to 5 decimal places after 14 iterations. The reason we cannot produce $u_2^\star$ is due to the magnitude of the derivative

$$g'(u) = -3\,u^2 + 3\,u + \tfrac{4}{9}$$

at the roots, which is

$$g'(u_1^\star) \;\approx\; 0.678065, \qquad g'(u_2^\star) \;\approx\; 1.18748, \qquad g'(u_3^\star) \;\approx\; 0.551126.$$

Thus, $u_1^\star$ and $u_3^\star$ are stable fixed points, but $u_2^\star$ is unstable. However, because $g'(u_1^\star)$ and $g'(u_3^\star)$ are both bigger than .5, this iterative algorithm converges *slower* than ordinary bisection!

Finally, Newton's method is based upon iteration of the function

$$g(u) = u - \frac{f(u)}{f'(u)} = u - \frac{u^3 - \frac{3}{2}\,u^2 + \frac{5}{9}\,u - \frac{1}{27}}{3\,u^2 - 3\,u + \frac{5}{9}}\,.$$

Starting with an initial guess of $u^{(0)} = 0$, the method computes $u_1^\star$ to 5 decimal places after only 4 iterations; starting with $u^{(0)} = .5$, it produces $u_2^\star$ after 2 iterations; while starting with $u^{(0)} = 1$ produces $u_3^\star$ after 3 iterations — a dramatic speed up over the other two methods.

Newton's method has a very pretty graphical interpretation, that helps us understand what is going on and why it converges so fast. Given the equation $f(u) = 0$, suppose we know an approximate value $u = u^{(k)}$ for a solution. Nearby $u^{(k)}$, we can approximate the nonlinear function $f(u)$ by its tangent line at the given point $u^{(k)}$, which has the equation

$$y = f(u^{(k)}) + f'(u^{(k)})(u - u^{(k)}). \tag{19.42}$$

As long as the tangent line is not horizontal — which requires $f'(u^{(k)}) \neq 0$ — it crosses the axis at the abscissa

$$u^{(k+1)} = u^{(k)} - \frac{f(u^{(k)})}{f'(u^{(k)})}\,,$$

which represents a new, and, presumably more accurate, approximation to the desired root. The procedure is illustrated pictorially in Figure Newton■. Note that the passage from $u^{(k)}$ to $u^{(k+1)}$ is exactly the Newton iteration step (19.41). In this manner, Newton's method can be viewed as successive approximation of the function by its tangent line and then using the root of the resulting affine function as the next approximation to the root of the function.

Given sufficiently accurate initial guesses, Newton's method will then rapidly produce accurate values for the simple roots to the equation in question. In practice, barring special structure in the problem, Newton's method is the root-finding algorithm of choice. The one caveat is that we need to come up with a reasonably close initial guess to the root we are seeking. Otherwise, there is no guarantee that it will converge at all, although if the Newton iterations do converge, we know that the limiting value is a root of our equation.

The behavior of Newton's method as we change parameters and vary the initial guess is very similar to the logistic map, and includes period doubling bifurcations and chaotic behavior. The reader is invited to experiment with simple examples, some of which are provided in Exercise ■. For further details, see [**120**].

Example 19.27. For fixed values of the eccentricity ϵ, Kepler's equation

$$u - \epsilon \sin u = m \tag{19.43}$$

can be viewed as a implicit equation defining the eccentric anomaly u as a function of the mean anomaly m. To solve the equation by Newton's method, we introduce the iterative function

$$g(u) \;=\; u \;-\; \frac{u - \epsilon \sin u - m}{1 - \epsilon \cos u}\,.$$

Notice that when $|\,\epsilon\,| < 1$, the denominator never vanishes and so the iteration remains well-defined everywhere. Starting with an initial guess $u^{(0)}$, we are assured that the method will quickly converge to the solution.

Fixing the eccentricity ϵ, we can employ a *continuation method* to determine how the solution $u^\star = h(m)$ depends upon the mean anomaly m. Namely, we start at $m = m_0 = 0$ with the obvious solution $u^\star = h(0) = 0$. Then, to compute the solution at successive closely spaced values $0 < m_1 < m_2 < m_3 < \cdots$, we use the previously computed value as an initial guess $u^{(0)} = h(m_k)$ for the value of the solution at the next mesh point m_{k+1}, and run the Newton scheme until it converges to the value $u^\star = h(m_{k+1})$. As long as m_{k+1} is reasonably close to m_k, Newton's method will converge to the solution quite quickly.

The continuation method will quickly produce the values of u at the sample points m_k. Intermediate values can either be determined by an interpolation scheme, e.g., a cubic spline fit of the data, or by running the Newton scheme using the closest known value as an initial condition. A plot for the value $\epsilon = .5$ appears in Figure kepler■.

Systems of Equations

Let us now turn our attention to systems of equations. We shall only consider the case when there are the same number of equations as unknowns:

$$f_1(u_1, \ldots, u_n) = 0, \qquad \ldots \qquad f_n(u_1, \ldots, u_n) = 0. \tag{19.44}$$

We shall write the system (19.44) in vector form

$$\mathbf{f}(\mathbf{u}) = \mathbf{0}, \tag{19.45}$$

where $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued function of n variables. Also, we do not necessarily require that $\mathbf{f}$ be defined on all of $\mathbb{R}^n$, although this does simplify the exposition to a certain degree.

We shall only consider solutions that are isolated, meaning separated from all the others. More formally:

Definition 19.28. A solution $\mathbf{u}^\star$ to a system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ is called *isolated* if there exists $\delta > 0$ such that $\mathbf{f}(\mathbf{u}) \neq \mathbf{0}$ for all $\mathbf{u}$ satisfying $0 < \| \mathbf{u} - \mathbf{u}^\star \| < \delta$.

Example 19.29. Consider the planar equation

$$x^2 + y^2 = (x^2 + y^2)^2.$$

Rewriting the equation in polar coordinates as

$$r = r^2 \qquad \text{or} \qquad r(r-1) = 0,$$

we immediately see that the solutions consist of the origin $x = y = 0$ and all points on the unit circle $r^2 = x^2 + y^2 = 1$. Only the origin is an isolated solution.

Typically, the solutions to a system of n equations in n unknowns are isolated, although this is not always the case. For example, if A is a singular $n \times n$ matrix, then the solutions to $A\,\mathbf{u} = \mathbf{0}$ consist of a nontrivial subspace of $\mathbb{R}^n$ and so are not isolated. Nonlinear systems with non-isolated solutions can similarly be viewed as having some form of degeneracy. In general, the computation of non-isolated solutions, e.g., solving the implicit equations for a curve or surface, is a much more difficult problem, and we will not attempt to discuss these issues in this introductory presentation. However, our continuation approach to the Kepler equation in Example 19.27 gives a hint as to how one might proceed in such situations.

In the case of a single scalar equation, the simple roots are the most amenable to practical computation. In higher dimensions, the role of the derivative of the function is played by the Jacobian matrix (19.27), and this motivates the following definition.

Definition 19.30. A solution $\mathbf{u}^\star$ to a system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ is called *nonsingular* if the associated Jacobian matrix is nonsingular there: $\det \mathbf{f}'(\mathbf{u}^\star) \neq 0$.

Note that the Jacobian matrix is square if and only if the system has the same number of equations as unknowns, and so this is a requirement for a solution to be nonsingular. Moreover, the Inverse Function Theorem, [**9**, **126**], from multivariable calculus implies that a nonsingular solution is necessarily isolated.

Theorem 19.31. *If $\mathbf{u}^\star$ is a nonsingular solution to the system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$, then $\mathbf{u}^\star$ is an isolated solution.*

As with simple roots of scalar equations, nonsingular solutions of systems are the most amenable to practical computation. Non-isolated solutions, as well as isolated solutions with singular Jacobian matrices, are much more difficult to compute, and very few useful solution algorithms exist in such degenerate situations.

Now, let us turn to numerical solution techniques. The first remark is that, unlike the scalar case, proving existence of a solution to a system of equations is often a difficult problem. There is no counterpart to the Intermediate Value Theorem 19.19 for vector-valued functions; it is easy to construct examples of vector-valued functions, whose entries take on both positive and negative values, but for which there are no solutions to the system (19.45); see Exercise ■ for one simple example. For this reason, there is no decent analog of the Bisection method for systems of equations.

On the other hand, Newton's method can be straightforwardly adapted to compute nonsingular solutions to systems of equations, and forms *the* most widely used method for

this purpose. The derivation proceeds in very similar manner to the scalar case. First, we replace the system (19.45) by a fixed point system

$$\mathbf{u} = \mathbf{g}(\mathbf{u}) \tag{19.46}$$

having the same solutions. By direct analogy with (19.37), any (reasonable) fixed point method will take the form

$$\mathbf{g}(\mathbf{u}) = \mathbf{u} - L(\mathbf{u})\,\mathbf{f}(\mathbf{u}), \tag{19.47}$$

where $L(\mathbf{u})$ is an $n \times n$ matrix-valued function. Clearly, if $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ then $\mathbf{g}(\mathbf{u}) = \mathbf{u}$; conversely, if $\mathbf{g}(\mathbf{u}) = \mathbf{u}$, then $L(\mathbf{u})\,\mathbf{f}(\mathbf{u}) = \mathbf{0}$. If we further require that the matrix $L(\mathbf{u})$ be nonsingular, i.e., $\det L(\mathbf{u}) \neq 0$, then every fixed point of the iterator (19.47) will be a solution to the system (19.45) and vice versa.

According to Theorem 19.14, the speed of convergence (if any) of the iterative method

$$\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)}) \tag{19.48}$$

is governed by the spectral radius or matrix norm of the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star)$ at the fixed point. In particular, if

$$\mathbf{g}'(\mathbf{u}^\star) = \mathrm{O} \tag{19.49}$$

is the zero matrix, then the method is quadratically convergent. Computing the derivative using the matrix version of the Leibniz rule for the derivative of a matrix product, cf. Exercise ■, we find

$$\mathbf{g}'(\mathbf{u}^\star) = \mathrm{I} - L(\mathbf{u}^\star)\,\mathbf{f}'(\mathbf{u}^\star), \tag{19.50}$$

where I is the $n \times n$ identity matrix; see Exercise ■ for details. (Fortunately, all the terms that involve derivatives of the entries of $L(\mathbf{u})$ go away since $\mathbf{f}(\mathbf{u}^\star) = \mathbf{0}$ by assumption.) Therefore, the quadratic convergence criterion (19.49) holds if and only if

$$L(\mathbf{u}^\star)\,\mathbf{f}'(\mathbf{u}^\star) = \mathrm{I}\,, \qquad \text{and hence} \qquad L(\mathbf{u}^\star) = \mathbf{f}'(\mathbf{u}^\star)^{-1} \tag{19.51}$$

should be the inverse of the Jacobian matrix of $\mathbf{f}$ at the solution, which, fortuitously, was already assumed to be nonsingular.

As in the scalar case, we don't know the solution $\mathbf{u}^\star$, but we can arrange that condition (19.51) holds by setting

$$L(\mathbf{u}) = \mathbf{f}'(\mathbf{u})^{-1}$$

everywhere — or at least everywhere that $\mathbf{f}$ has a nonsingular Jacobian matrix. The resulting fixed point system

$$\mathbf{u} = \mathbf{g}(\mathbf{u}) = \mathbf{u} - \mathbf{f}'(\mathbf{u})^{-1}\,\mathbf{f}(\mathbf{u}), \tag{19.52}$$

leads to the quadratically convergent *Newton iteration scheme*

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \mathbf{f}'(\mathbf{u}^{(k)})^{-1}\,\mathbf{f}(\mathbf{u}^{(k)}). \tag{19.53}$$

All it requires is that we guess an initial value $\mathbf{u}^{(0)}$ that is sufficiently close to the desired solution $\mathbf{u}^\star$. We are then guaranteed that the iterates $\mathbf{u}^{(k)}$ converge quadratically fast to $\mathbf{u}^\star$.

Theorem 19.32. *Let $\mathbf{u}^\star$ be a nonsingular solution to the system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$. Then, provided $\mathbf{u}^{(0)}$ is sufficiently close to $\mathbf{u}^\star$, the Newton iteration scheme* (19.53) *converges at a quadratic rate to the solution: $\mathbf{u}^{(k)} \to \mathbf{u}^\star$.*

Example 19.33. Consider the pair of simultaneous cubic equations

$$f_1(u,v) = u^3 - 3\,u\,v^2 - 1 = 0, \qquad f_2(u,v) = 3\,u^2\,v - v^3 = 0. \tag{19.54}$$

It is not difficult to prove that there are three solutions:

$$\mathbf{u}_1^\star = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2^\star = \begin{pmatrix} -.5 \\ .866025\ldots \end{pmatrix}, \qquad \mathbf{u}_3^\star = \begin{pmatrix} -.5 \\ -.866025\ldots \end{pmatrix}.$$

The Newton scheme relies on the Jacobian matrix

$$\mathbf{f}'(\mathbf{u}) = \begin{pmatrix} 3\,u^2 - 3\,v^2 & -6\,u\,v \\ 6\,u\,v & 3\,u^2 - 3\,v^2 \end{pmatrix}.$$

Since $\det \mathbf{f}'(\mathbf{u}) = 9(u^2 + v^2)$ is non-zero except at the origin, all three solutions are non-singular, and hence, for a sufficiently close initial value, Newton's method will converge. We compute the inverse Jacobian matrix explicitly:

$$\mathbf{f}'(\mathbf{u})^{-1} = \frac{1}{9(u^2+v^2)} \begin{pmatrix} 3\,u^2 - 3\,v^2 & 6\,u\,v \\ -6\,u\,v & 3\,u^2 - 3\,v^2 \end{pmatrix}.$$

Hence, in this particular example, the Newton iterator (19.52) is

$$\mathbf{g}(\mathbf{u}) = \begin{pmatrix} u \\ v \end{pmatrix} - \frac{1}{9(u^2+v^2)} \begin{pmatrix} 3\,u^2 - 3\,v^2 & 6\,u\,v \\ -6\,u\,v & 3\,u^2 - 3\,v^2 \end{pmatrix} \begin{pmatrix} u^3 - 3\,u\,v^2 - 1 \\ 3\,u^2\,v - v^3 \end{pmatrix}.$$

Implementing (19.53). Starting with ■ we converge to ■

Remark: The alert reader may notice that in this example, we are in fact merely computing the cube roots of unity, i.e., equations (19.54) are the real and imaginary parts of the complex equation $z^3 = 1$ when $z = u + \mathrm{i}\,v$. A complete map of the basins of attraction converging to the three different roots has a remarkably complicated, fractal-like structure, as illustrated in Figure Newt3■.

Example 19.34. A robot arm consists of two rigid rods that are joined end-to-end to a fixed point in the plane, which we take as the origin $\mathbf{0}$. The arms are free to rotate, and the problem is to configure them so that the robots hand ends up at the prescribed position $\mathbf{a} = (\,a, b\,)^T$. The first rod has length ℓ and makes an angle α with the horizontal, so its end is at position $\mathbf{v}_1 = (\,\ell\cos\alpha, \ell\sin\alpha\,)^T$. The second rod has length m and makes an angle β with the horizontal, and so is represented by the vector $\mathbf{v}_2 = (\,m\cos\beta, m\sin\beta\,)^T$. The hand at the end of the second arm is at position $\mathbf{v}_1 + \mathbf{v}_2$, and the problem is to find values for the angles α, β so that $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{a}$. To this end, we need to solve the system of equations

$$\ell\cos\alpha + m\cos\beta = a, \qquad \ell\sin\alpha + m\sin\beta = b. \tag{19.55}$$

k	$\alpha^{(k)}$	$\beta^{(k)}$	$x^{(k)}$	$y^{(k)}$	$z^{(k)}$	$w^{(k)}$
0	0.0000	1.5708	2.0000	0.0000	2.0000	1.0000
1	0.0000	2.5708	2.0000	0.0000	1.1585	0.5403
2	0.3533	2.8642	1.8765	0.6920	0.9147	0.9658
3	0.2917	2.7084	1.9155	0.5751	1.0079	0.9948
4	0.2987	2.7176	1.9114	0.5886	1.0000	1.0000
5	0.2987	2.7176	1.9114	0.5886	1.0000	1.0000

To compute the solution, we shall apply Newton's method. First, we compute the Jacobian matrix of the system with respect to α, β, which is

$$\mathbf{f}'(\alpha, \beta) = \begin{pmatrix} -\ell \sin\alpha & -m \sin\beta \\ \ell \cos\alpha & m \cos\beta \end{pmatrix}.$$

As a result, the Newton iteration equation (19.53) has the explicit form

$$\begin{pmatrix} \alpha^{(k+1)} \\ \beta^{(k+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{pmatrix} - \\ - \frac{1}{\ell m \sin(\beta^{(k)} - \alpha^{(k)})} \begin{pmatrix} -\ell \cos\alpha^{(k)} & m \sin\beta^{(k)} \\ -\ell \cos\alpha^{(k)} & m \sin\beta^{(k)} \end{pmatrix} \begin{pmatrix} \ell \cos\alpha^{(k)} + m \cos\beta^{(k)} - a \\ \ell \sin\alpha^{(k)} + m \sin\beta^{(k)} - b \end{pmatrix}.$$

when running the iteration, one must be careful to avoid points at which $\alpha^{(k)} - \beta^{(k)} = 0$ or π, i.e., where the robot arm has straightened out.

As an example, let us assume that the rods have lengths $\ell = 2$, $m = 1$, and the desired location of the hand is at $\mathbf{a} = (\,1, 1\,)^T$. We start with an initial guess of $\alpha^{(0)} = 0$, $\beta^{(0)} = \frac{1}{2}\pi$, so the first rod lies along the x–axis and the second is perpendicular. The first few Newton iterates are given in the accompanying table.The first column gives the iterate number k. The second and third columns indicate the angles $\alpha^{(k)}$, $\beta^{(k)}$ of the rods. The fourth and fifth give the position $\big(\,x^{(k)}, y^{(k)}\,\big)^T$ of the joint or elbow, while the final two indicate the position $\big(\,z^{(k)}, w^{(k)}\,\big)^T$ of the robot's hand.

Thus, the robot has rapidly converged to one of the two possible configurations. Convergence is dependent upon the initial configuration, and the iterates do not always settle down. For instance, if $\| \mathbf{a} \| > \ell + m$, there is no possible solution, since the arms are too short for the hand to reach to desired location; thus, no choice of initial conditions will lead to a convergent scheme and the robot arm flaps around in a chaotic manner.

Now that we have gained some experience with Newton's method for systems of equations, some supplementary remarks are in order. As we learned back in Chapter 1, except perhaps in very low-dimensional situations, one should not invert a matrix directly, but rather use Gaussian elimination, or, in favorable situations, a linear iterative scheme, e.g., Jacobi, Gauss–Seidel or SOR, to solve a linear system. So it is better to write the Newton

equation (19.53) in unsolved, implicit form

$$\mathbf{f}'(\mathbf{u}^{(k)})\,\mathbf{v}^{(k+1)} = -\,\mathbf{f}(\mathbf{u}^{(k)}), \qquad \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{v}^{(k)}. \tag{19.56}$$

Given the iterate $\mathbf{u}^{(k)}$, we first compute the Jacobian matrix $\mathbf{f}'(\mathbf{u}^{(k)})$, and then use our preferred linear systems solver to find $\mathbf{v}^{(k)}$. Adding $\mathbf{u}^{(k)}$ to the result immediately yields the updated approximation $\mathbf{u}^{(k+1)}$ to the solution.

Therefore, the main bottleneck in the implementation of the Newton scheme, particularly for large systems, is solving the linear system (19.56). The coefficient matrix $\mathbf{f}'(\mathbf{u}^{(k)})$ must be recomputed at each step of the iteration, and hence knowing the solution to the k^{th} linear system does not help us solve the next one in the sequence. Having to re-implement a complete Gaussian elimination at every step will tend to slow down the algorithm, particularly in high dimensional situations involving many equations in many unknowns.

One simple dodge for speeding up the computation is to note that, once we start converging, $\mathbf{u}^{(k)}$ will be very close to $\mathbf{u}^{(k-1)}$ and so we will probably not go far wrong by using $\mathbf{f}'(\mathbf{u}^{(k-1)})$ in place of the updated Jacobian matrix $\mathbf{f}'(\mathbf{u}^{(k)})$. Since we have already solved the linear system with coefficient matrix $\mathbf{f}'(\mathbf{u}^{(k-1)})$, we know its $L\,U$ factorization, and hence can use forward and back substitution to quickly solve the modified system

$$\mathbf{f}'(\mathbf{u}^{(k-1)})\,\mathbf{v}^{(k+1)} = -\,\mathbf{f}(\mathbf{u}^{(k)}), \qquad \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{v}^{(k)}. \tag{19.57}$$

If $\mathbf{u}^{(k+1)}$ is still close to $\mathbf{u}^{(k-1)}$, we can continue to use $\mathbf{f}'(\mathbf{u}^{(k-1)})$ as the coefficient matrix when proceeding on to the next iterate $\mathbf{u}^{(k+2)}$. We continue until there has been a notable change in the iterates, at which stage we can revert to solving the correct, unmodified linear system (19.56) by Gaussian elimination. In this version of the algorithm, we update the coefficient matrix every few iterations, particularly if the value of the approximations has significantly changed. This device may dramatically reduce the total amount of computation required to approximate the solution to a prescribed accuracy. The down side is that this *quasi-Newton scheme* is only linearly convergent, and so does not home in on the root as fast as the unmodified implementation. The user needs to balance the trade-off between speed of convergence versus amount of time needed to solve the linear system at each step in the process.

19.3. Optimization.

We have already remarked on the importance of quadratic minimization principles to characterize the equilibrium solutions of a variety of linear systems. In nonlinear mathematics, optimization loses none of its centrality, and the wealth of practical applications has spawned an entire subdiscipline of applied mathematics. Physical systems naturally seek to minimize the potential energy function, and so determination of the possible equilibrium configurations requires solving a nonlinear minimization principle. Engineering design is guided by a variety of optimization constraints, such as performance, safety, cost and marketability. Non-quadratic minimization principles also arise in the fitting of data by more general schemes beyond the simple linear least squares approximation method

discussed in Section 4.3. Additional applications arise in economics and financial mathematics — one often wishes to minimize costs or maximize profits — in manufacturing, in biological and ecological systems, in pattern recognition and signal processing, and in statistics.

The Objective Function

Throughout this section, the function $F(\mathbf{u}) = F(u_1, \dots, u_n)$ to be minimized — the energy, cost, entropy, performance, etc. — will be called the *objective function*. As such, it depends upon one or more variables $\mathbf{u} = (\, u_1, u_2, \dots, u_n \,)^T$ that belong to a prescribed subset $\Omega \subset \mathbb{R}^n$.

Definition 19.35. A point $\mathbf{u}^\star \in \Omega$ is a *global minimum* of the objective function on the domain Ω if

$$F(\mathbf{u}^\star) \le F(\mathbf{u}) \qquad \text{for all} \qquad \mathbf{u} \in \Omega. \tag{19.58}$$

The minimum is called *strict* if

$$F(\mathbf{u}^\star) < F(\mathbf{u}) \qquad \text{for} \qquad \mathbf{u}^\star \neq \mathbf{u} \in \Omega. \tag{19.59}$$

The point is called a *local minimum* if the inequality holds just for points $\mathbf{u}$ nearby $\mathbf{u}^\star$, i.e., satisfying $\| \mathbf{u} - \mathbf{u}^\star \| < \delta$ for some $\delta > 0$. Thus, strict local minima are *isolated*.

The definition of a *maximum* — local or global — is the same, but with the reversed inequality: $F(\mathbf{u}^\star) \ge F(\mathbf{u})$ or, in the strict case, $F(\mathbf{u}^\star) > F(\mathbf{u})$. Alternatively, a maximum of $F(\mathbf{u})$ is the same as a minimum of the negative $-F(\mathbf{u})$. Therefore, every result that applies to minimization of a function can easily be translated into a result on maximization, which allows us to concentrate exclusively on the minimization problem without any loss of generality. We will use *extremum*[†] as a shorthand term for either a maximum or a minimum.

Remark: As we already noted in Section 4.1, *any* system of equations can be readily converted into a minimization principle. Thus, given a system (19.45), we consider the function[‡]

$$F(\mathbf{u}) = \| \mathbf{f}(\mathbf{u}) \|^2 = f_1(u_1, \dots, u_n)^2 + \, \cdots \, + f_n(u_1, \dots, u_n)^2. \tag{19.60}$$

By the basic properties of the norm, the minimum value is $F(\mathbf{u}) = 0$, and this is achieved if and only if $\mathbf{f}(\mathbf{u}) = \mathbf{0}$, i.e., at a solution to the system.

In contrast to the much more complicated existence question for systems of equations, there is an general theorem that guarantees the existence of minima (and, hence, maxima) for a very broad class of optimization problems.

Theorem 19.36. *If $F \colon \Omega \to \mathbb{R}$ is continuous, and $\Omega \subset \mathbb{R}^n$ is closed and bounded, then F has at least one global minimum $\mathbf{u}^\star \in \Omega$.*

[†] Curiously, the term "optimum" is not used.

[‡] We use the standard Euclidean norm, but any other norm would work equally well here.

See [**125**, **126**] for a proof. Although Theorem 19.36 assures us of the existence of a global minimum of any continuous function on a bounded domain, it does not guarantee uniqueness, nor does it indicate how to go about finding it. Just as with the solution of nonlinear systems of equations, it is quite rare that one can find exact formulae for the minima of non-quadratic functions. Our goal, then, is to formulate practical algorithms that can accurately compute the minima of general nonlinear functions. A naïalgorithm, but one that is often successfully applied in practical problems, [**121**, **opt**], is to select a reasonably dense set of sample points $u^{(k)}$ in the domain and compare the values of $f(u^{(k)})$. If the points are sufficiently densely distributed and the function is not too wild, this will give a good approximation to the minimum. The algorithm can be speeded up by using sophisticated methods of selecting the sample points.

As the student no doubt remembers, there are two different possible types of minima. An *interior minimum* occurs at an interior point of the domain of definition of the function, whereas a *boundary minimum* occurs on its boundary $\partial\Omega$. Interior local minima are easier to find, and, to keep the presentation simple, we shall focus our efforts on them.

Let us review the basic procedure for optimizing scalar functions that you learned in calculus.

Example 19.37. Let us optimize the scalar function

$$f(u) = 8\,u^3 + 5\,u^2 - 6\,u$$

on the domain $-1 \le u \le 1$. As you learned in first year calculus, the first step to finding the minimum is to look at the *critical points* where the derivative vanishes:

$$f'(u) = 24\,u^2 + 10\,u - 6 = 0, \qquad \text{and hence} \qquad u = \tfrac{1}{3},\ -\tfrac{3}{4}.$$

To ascertain the local nature of the two critical points, we apply the second derivative test. Since $f''(u) = 48\,u + 10$, we have

$$f''\left(\tfrac{1}{3}\right) = 26 > 0, \qquad \text{whereas} \qquad f''\left(-\tfrac{3}{4}\right) = -26 < 0,$$

and we conclude that $\frac{1}{3}$ is a local minimum, while $\frac{3}{4}$ is a local maximum.

To find the global minimum and maximum on the interval $[-1, 1]$, we must also take into account the boundary points ± 1. Comparing the function values at the four points,

$$f(-1) = 3, \qquad f\left(\tfrac{1}{3}\right) = -\tfrac{31}{27} \approx -1.148, \qquad f\left(-\tfrac{3}{4}\right) = \tfrac{63}{16} = 3.9375, \qquad f(1) = 7,$$

we see that $\frac{1}{3}$ is the global minimum, whereas 1 is the global maximum. This is borne out by the graph of the function in Figure 19.5.

The Gradient

As the student learns in multi-variable calculus, the (interior) extrema — minima and maxima — of a smooth function $F(\mathbf{u}) = F(u_1, \dots, u_n)$ are necessarily *critical points*, meaning places where the gradient of F vanishes. The gradient of a function is, of course, the vector whose entries are its first order partial derivatives:

$$\nabla F(\mathbf{u}) = \left(\frac{\partial F}{\partial u_1}, \ \dots \ , \frac{\partial F}{\partial u_n} \right)^T. \tag{19.61}$$

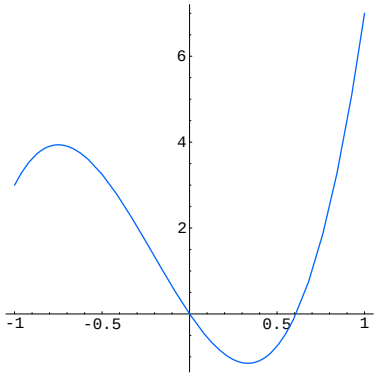


Figure 19.5. The function $8u^3 + 5u^2 - 6u$.

Let us, in preparation for more general minimization problems over infinite-dimensional function spaces, reformulate the definition of the gradient in a more intrinsic manner. An important but subtle point is that the gradient operator, in fact, relies upon the introduction of an inner product on the underlying vector space. The "standard version" (19.61) is based upon on the Euclidean inner product on $\mathbb{R}^n$. Altering the inner product will change the formula for the gradient!

Definition 19.38. Let V be an inner product space. Given a function $F: \Omega \to \mathbb{R}$ defined on an open domain $\Omega \subset V$, its *gradient* at a point $\mathbf{u} \in \Omega$ is the vector $\nabla F(\mathbf{u}) \in V$ that satisfies

$$\langle\, \nabla F(\mathbf{u})\,;\mathbf{v}\,\rangle = \frac{d}{dt}\, F(\mathbf{u} + t\,\mathbf{v})\,\Big|_{t=0} \qquad \text{for all} \qquad \mathbf{v} \in V. \tag{19.62}$$

The left hand side of (19.62) is known as the *directional derivative* of F with respect to $\mathbf{v} \in V$, typically denoted by $\partial F/\partial \mathbf{v}$.

In the Euclidean case, when $F(\mathbf{u}) = F(u_1, \dots, u_n)$ is a function of n variables, defined for $\mathbf{u} \in \mathbb{R}^n$, we can use the chain rule to compute

$$\begin{aligned} \frac{d}{dt}\, F(\mathbf{u} + t\,\mathbf{v}) &= \frac{d}{dt}\, F(u_1 + t\,v_1, \dots, u_n + t\,v_n) \\ &= v_1\, \frac{\partial F}{\partial u_1}(\mathbf{u} + t\,\mathbf{v}) \;+\; \cdots \;+\; v_n\, \frac{\partial F}{\partial u_n}(\mathbf{u} + t\,\mathbf{v}). \end{aligned} \tag{19.63}$$

Setting $t = 0$, the right hand side of (19.62) reduces to

$$\frac{d}{dt}\, F(\mathbf{u} + t\,\mathbf{v})\,\Big|_{t=0} = v_1\, \frac{\partial F}{\partial u_1}\,(\mathbf{u}) \;+\; \cdots \;+\; v_n\, \frac{\partial F}{\partial u_n}\,(\mathbf{u}) = \mathbf{v} \cdot \nabla F(\mathbf{u}) = \nabla F(\mathbf{u}) \cdot \mathbf{v}.$$

Therefore, the directional derivative equals the Euclidean dot product between the usual gradient of the function (19.61) and the direction vector $\mathbf{v}$.

A function $F(\mathbf{u})$ is *continuously differentiable* if and only if its gradient $\nabla F(\mathbf{u})$ is a continuously varying function of $\mathbf{u}$. This is equivalent to the requirement that the first order partial derivatives $\partial F/\partial u_i$ are all continuous. As usual, we use $\mathrm{C}^1(\Omega)$ to denote the vector space of all continuously differentiable scalar-valued functions defined on a domain

$\Omega \subset \mathbb{R}^n$. From now on, all objective functions are assumed to be continuously differentiable on their domain of definition.

Remark: In this chapter, we will only deal with the standard Euclidean dot product and hence the usual gradient (19.61). However, all results can be readily translated into more general situations, e.g., weighted inner products. Details are outlined in Exercise ■.

More generally, if $\mathbf{u}(t)$ represents a parametrized curve contained within the domain of definition of $F(\mathbf{u})$, then the instantaneous rate of change in the scalar quantity F as we move along the curve is given by

$$\frac{d}{dt} F(\mathbf{u}(t)) = \left\langle \nabla F(\mathbf{u}) \,;\, \frac{d\mathbf{u}}{dt} \right\rangle , \tag{19.64}$$

which is the directional derivative of F with respect to the velocity or tangent vector $\mathbf{v} = \dot{\mathbf{u}}$ to the curve. For instance, our rate of ascent or descent as we travel through the mountains is given by the dot product of our velocity vector with the gradient of the elevation function. This leads us to one important interpretation of the gradient vector.

Theorem 19.39. *The gradient ∇F of a scalar function $F(\mathbf{u})$ points in the direction of its steepest increase. The negative gradient, $-\nabla F$, which points in the opposite direction, indicates the direction of steepest decrease.*

For example, if $F(u, v)$ represents the elevation of a mountain range at position (u, v) on a map, then ∇F tells us the direction that is steepest uphill, while $-\nabla F$ points directly downhill — the direction water will flow. Similarly, if $F(u, v, w)$ represents the temperature of a solid body, then ∇F tells us the direction in which it is getting the hottest. Heat energy (like water) will flow in the opposite, coldest direction, namely that of the negative gradient vector $-\nabla F$.

You need to be careful in how you interpret Theorem 19.39. Clearly, the faster you move along a curve, the faster the function $F(\mathbf{u})$ will vary, and one needs to take this into account when comparing the rates of change along different curves. The easiest way to normalize is to assume that the tangent vector $\mathbf{a} = \dot{\mathbf{u}}$ has norm 1, so $\| \mathbf{a} \| = 1$ and we are passing through the point $\mathbf{u}$ with unit speed. Once this is done, Theorem 19.39 is an immediate consequence of the Cauchy–Schwarz inequality (3.16). Indeed,

$$\left| \frac{\partial F}{\partial \mathbf{a}} \right| = | \mathbf{a} \cdot \nabla F | \leq \| \mathbf{a} \| \, \| \nabla F \| = \| \nabla F \|, \qquad \text{when} \qquad \| \mathbf{a} \| = 1,$$

with equality if and only if $\mathbf{a} = c \nabla F$ points in the same direction as the gradient. Therefore, the maximum rate of change is when $\mathbf{a} = \nabla F / \| \nabla F \|$ is the unit vector in the direction of the gradient, while the minimum is achieved when $\mathbf{a} = -\nabla F / \| \nabla F \|$ points in the opposite direction. As a result, Theorem 19.39 tells us how to move if we wish to minimize a scalar function as rapidly as possible.

Theorem 19.40. *A curve $\mathbf{u}(t)$ will realize the steepest decrease in the scalar field $F(\mathbf{u})$ if and only if it satisfies the gradient flow equation*

$$\dot{\mathbf{u}} = -\nabla F(\mathbf{u}). \tag{19.65}$$

Critical Points

Let us now prove that the gradient vanishes at any local minimum of the function. The most important thing about this proof is that it only relies on the intrinsic definition of gradient, and therefore applies to any function on any inner product space. Moreover, even though the gradient can change if we alter the underlying inner product, the condition that it vanishes at a local extremum does not.

Definition 19.41. A point $\mathbf{u}^\star$ is called a *critical point* of the objective function $F(\mathbf{u})$ if

$$\nabla F(\mathbf{u}^\star) = \mathbf{0}. \tag{19.66}$$

Theorem 19.42. *If $\mathbf{u}^\star \in \Omega$ is a local (interior) minimum of $F(\mathbf{u})$, then $\nabla F(\mathbf{u}^\star) = \mathbf{0}$, and so $\mathbf{u}^\star$ is a critical point.*

Proof: Let $\mathbf{v} \in \mathbb{R}^n$ be any vector. Consider the function

$$g(t) = F(\mathbf{u}^\star + t\,\mathbf{v}) = F(u_1^\star + t\,v_1, \dots, u_n^\star + t\,v_n), \tag{19.67}$$

where $t \in \mathbb{R}$ is sufficiently small to ensure that $\mathbf{u}^\star + t\,\mathbf{v} \in \Omega$ remains inside the domain of F. Thus, g measures the values of F along a straight line passing through $\mathbf{u}^\star$ in the direction[†] prescribed by $\mathbf{v}$. Since $\mathbf{u}^\star$ is a local minimum,

$$F(\mathbf{u}^\star) \le F(\mathbf{u}^\star + t\,\mathbf{v}), \qquad \text{and hence} \qquad g(0) \le g(t)$$

for all t sufficiently close to zero. In other words, $g(t)$, as a function of the single variable t, has a local minimum at $t = 0$. By the basic calculus result on minima of functions of one variable, the derivative of $g(t)$ must vanish at $t = 0$. Therefore, by the definition (19.62) of gradient,

$$0 = g'(0) = \frac{d}{dt}\, F(\mathbf{u}^\star + t\,\mathbf{v}) \,\Big|_{t=0} = \langle\, \nabla F(\mathbf{u}^\star)\,;\mathbf{v}\,\rangle.$$

We conclude that the gradient vector $\nabla F(\mathbf{u}^\star)$ at the critical point must be orthogonal to every vector $\mathbf{v} \in \mathbb{R}^n$. The only vector that is orthogonal to every vector in an inner product space is the zero vector, and hence $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. *Q.E.D.*

Remark: As we learned, the gradient vector ∇F points in the direction of the steepest increase in the function, while its negative, $-\nabla F(\mathbf{u})$, points in the direction of steepest decrease. At a minimum of the function, all directions are increasing, and so there *is* no direction of steepest decrease. The only way that the gradient can avoid this little dilemma is for it to vanish, which provides an intuitive explanation of why minima (and maxima) must be critical points.

[†] If $\mathbf{v} = \mathbf{0}$, then the line degenerates to a point, but the ensuing argument remains (trivially) valid.

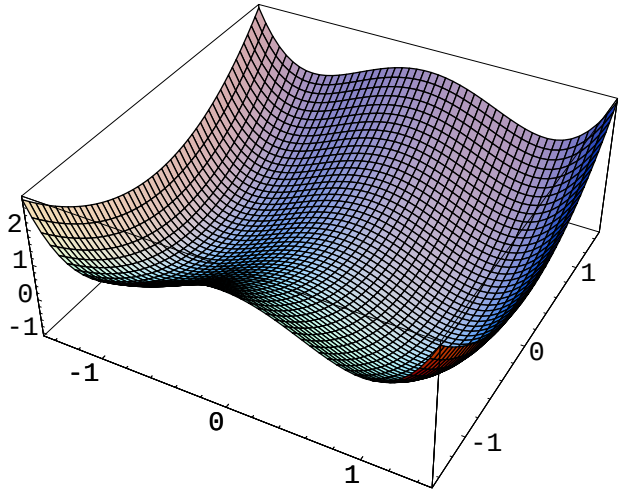


Figure 19.6. The Function $u^4 - 2u^2 + v^2$.

Thus, provided the objective function is continuously differentiable, every interior minimum, both local and global, is necessarily a critical point. The converse is not true; critical points can be maxima; they can also be saddle points or of some degenerate form. The basic analytical method[‡] for determining the (interior) minima of a given function is to first find all its critical points by solving the system of equations (19.66). Each critical point then needs to be more closely examined — as it could be either a minimum, or a maximum, or neither.

Example 19.43. Consider the function

$$F(u, v) = u^4 - 2u^2 + v^2,$$

which is defined and continuously differentiable on all of $\mathbb{R}^2$. Since $\nabla F = \big(\, 4u^3 - 4u, 2v \,\big)^T$, its critical points are obtained by solving the system of equations

$$4u^3 - 4u = 0, \qquad 2v = 0.$$

The solutions to the first equation are $u = 0, \pm 1$, while the second equation requires $v = 0$. Therefore, F has three critical points:

$$\mathbf{u}_1^\star = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2^\star = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_3^\star = \begin{pmatrix} -1 \\ 0 \end{pmatrix}. \tag{19.68}$$

Inspecting the graph in Figure 19.6, we suspect that the first critical point $\mathbf{u}_1^\star$ is a saddle point, whereas the other two are both global minima for the function, with the same value $F(\mathbf{u}_2^\star) = F(\mathbf{u}_3^\star) = -1$. This will be confirmed once we learn how to rigorously distinguish critical points.

[‡] Numerical methods are discussed below.

If $F(\mathbf{u})$ is defined on a closed subdomain $\Omega \subset \mathbb{R}^n$, then its minima may also occur at boundary points $\mathbf{u} \in \partial\Omega$, and there is no requirement that the gradient vanish at such boundary minima. The analytical determination of boundary extrema relies on the method of Lagrange multipliers, and we refer the interested reader to [**9**, **38**]. If the domain is unbounded, one must also worry about the asymptotic behavior of the function for large $\mathbf{u}$. In order to keep our presentation simple, we shall relegate these more involved issues to a more advanced text.

The student should also pay attention to the distinction between local minima and global minima. Both are critical points. In the absence of theoretical justification, the only practical way to determine whether or not a minimum is global is to find all the different local minima and see which one gives the smallest value. In many examples arising in applications, when $F(\mathbf{u})$ is often an energy function, one knows that the function is bounded from below, and hence, from general principles, that a global minimum exists, even when the domain is unbounded.

The Second Derivative Test

The status of critical point — minimum, maximum, or neither — can often be resolved by analyzing the second derivative of the objective function at the critical point. Let us first review the one variable second derivative test from first year calculus.

Proposition 19.44. *Let $g(t) \in \mathrm{C}^2$ be a scalar function, and suppose $t^\star$ a critical point, so $g'(t^\star) = 0$. If $t^\star$ is a local minimum, then $g''(t^\star) \geq 0$. Conversely, if $g''(t^\star) > 0$, then $t^\star$ is a strict local minimum. Similarly, $g''(t^\star) \leq 0$ is required at a local maximum, while $g''(t^\star) < 0$ implies that $t^\star$ is a strict local maximum.*

The proof of this result relies on the quadratic Taylor approximation

$$g(t) \;\approx\; g(t^\star) + \tfrac{1}{2}\,(t - t^\star)^2\, g''(t^\star)$$

near the critical point, (C.7), where we use the fact that $g'(t^\star) = 0$ and so the linear terms in the Taylor polynomial vanish. If $g''(t^\star) \neq 0$, then the quadratic Taylor polynomial has a minimum or maximum at $t^\star$ according to the sign of the second derivative. In the borderline case, when $g''(t^\star) = 0$, the second derivative test is inconclusive, and the point could be either maximum, minimum, saddle point, or degenerate. One must then look at the higher order terms in the Taylor expansion to resolve the issue; see Exercise ■.

In multi-variate calculus, the "second derivative" is represented by the $n \times n$ *Hessian matrix*

$$\nabla^2 F(\mathbf{u}) = \begin{pmatrix} \dfrac{\partial^2 F}{\partial u_1^2} & \dfrac{\partial^2 F}{\partial u_1\, \partial u_2} & \cdots & \dfrac{\partial^2 F}{\partial u_1\, \partial u_n} \\ \dfrac{\partial^2 F}{\partial u_2\, \partial u_1} & \dfrac{\partial^2 F}{\partial u_2^2} & \cdots & \dfrac{\partial^2 F}{\partial u_2\, \partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 F}{\partial u_n\, \partial u_1} & \dfrac{\partial^2 F}{\partial u_n\, \partial u_2} & \cdots & \dfrac{\partial^2 F}{\partial u_n^2} \end{pmatrix}, \tag{19.69}$$

named after the early eighteenth German mathematician Ludwig Otto Hesse. The entries of the Hessian are the second order partial derivatives of the objective function. If $F(\mathbf{u}) \in \mathrm{C}^2$ has continuous second order partial derivatives, then its Hessian matrix is symmetric, $\nabla^2 F(\mathbf{u}) = \nabla^2 F(\mathbf{u})^T$, which is a restatement of the fact that its mixed partial derivatives are equal: $\partial^2 F/\partial u_i\, \partial u_j = \partial^2 F/\partial u_j\, \partial u_i$, cf. [**9**, **38**]. For the applicability of the second derivative test, this is an essential ingredient.

The second derivative test for a local minimum of scalar function relies on the positivity of its second derivative. For a function of several variables, the corresponding condition is that the Hessian matrix be positive definite, as in Definition 3.22. More specifically:

Theorem 19.45. *Let $F(\mathbf{u}) = F(u_1, \dots, u_n) \in \mathrm{C}^2(\Omega)$ be a real-valued, twice continuously differentiable function defined on an open domain $\Omega \subset \mathbb{R}^n$. If $\mathbf{u}^\star \in \Omega$ is a (local, interior) minimum for F, then it is necessarily a critical point, so $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. Moreover, the Hessian matrix* (19.69) *must be positive semi-definite at the minimum, so $\nabla^2 F(\mathbf{u}^\star) \geq 0$. Conversely, if $\mathbf{u}^\star$ is a critical point with positive definite Hessian matrix $\nabla^2 F(\mathbf{u}^\star) > 0$, then $\mathbf{u}^\star$ is a strict local minimum of F.*

Proof: We return to the proof of Theorem 19.42. Given a local minimum $\mathbf{u}^\star$, the scalar function $g(t) = F(\mathbf{u}^\star + t\,\mathbf{v})$ in (19.67) has a local minimum at $t = 0$. As noted above, basic calculus tells us that its derivatives at $t = 0$ must satisfy

$$g'(0) = 0, \qquad g''(0) \geq 0. \tag{19.70}$$

The first condition leads to the critical point equation $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. A straightforward chain rule calculation produces the formula

$$g''(0) = \sum_{i,j=1}^{n} \frac{\partial^2 F}{\partial u_i\, \partial u_j}(\mathbf{u}^\star)\, v_i\, v_j = \mathbf{v}^T\, \nabla^2 F(\mathbf{u}^\star)\, \mathbf{v}.$$

As a result, the second condition in (19.70) requires that

$$\mathbf{v}^T\, \nabla^2 F(\mathbf{u}^\star)\, \mathbf{v} \geq 0.$$

Since this condition is required for every direction $\mathbf{v} \in \mathbb{R}^n$, the Hessian matrix $\nabla^2 F(\mathbf{u}^\star) \geq 0$ satisfies the criterion for positive semi-definiteness, proving the first part of the theorem.

Conversely, if the Hessian $\nabla^2 F(\mathbf{u}^\star) > 0$ is positive definite, then,

$$g''(0) = \mathbf{v}^T\, \nabla^2 F(\mathbf{u}^\star)\, \mathbf{v} > 0 \qquad \text{for all} \qquad \mathbf{v} \neq \mathbf{0},$$

and so $t = 0$ is a strict local minimum for $g(t)$. Since this occurs for every $\mathbf{v} \in V$, this implies[†] $F(\mathbf{u}^\star) < F(\mathbf{u})$ for all $\mathbf{u}$ near $\mathbf{u}^\star$ and so $\mathbf{u}^\star$ is a strict local minimum. *Q.E.D.*

[†] We are ignoring some technical details that need cleaning up for a completely rigorous proof, which relies on the multivariable Taylor expansion of $F(\mathbf{u})$. See Appendix C.

A maximum requires a negative semi-definite Hessian matrix. If, moreover, the Hessian at the critical point is negative definite, then the critical point is a strict local maximum. If the Hessian matrix is indefinite, then the critical point is a saddle point, and neither minimum nor maximum. In the borderline case — when the Hessian is only positive or negative semi-definite at the critical point, then the second derivative test is inconclusive. Resolving the nature of the critical point requires more detailed knowledge of the objective function, e.g., its higher order derivatives.

Example 19.46. As a first, elementary example, consider the quadratic function

$$F(u,v) = u^2 - 2\,u\,v + 3\,v^2.$$

To minimize F, we begin by computing its gradient $\nabla F = \begin{pmatrix} 2\,u - 2\,v \\ -2\,u + 6\,v \end{pmatrix}$. Solving the pair of equations $\nabla F = \mathbf{0}$, namely

$$2\,u - 2\,v = 0, \qquad\qquad -2\,u + 6\,v = 0,$$

we see that the only critical point is the origin $u = v = 0$. To test whether the origin is a maximum or minimum, we further compute the Hessian matrix

$$H = \nabla^2 F(u,v) = \begin{pmatrix} F_{uu} & F_{uv} \\ F_{uv} & F_{vv} \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 6 \end{pmatrix}.$$

Using the methods of Section 3.5, we easily prove that the Hessian matrix is positive definite. Therefore, by Theorem 19.45, $\mathbf{u}^\star = \mathbf{0}$ is a strict local minimum of F.

Indeed, we recognize $F(u,v)$ to be, in fact, a homogeneous positive definite quadratic form, which can be written in the form

$$F(u,v) = \mathbf{u}^T K\,\mathbf{u}, \qquad \text{where} \qquad K = \begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix} = \tfrac{1}{2}\,H, \qquad \mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}.$$

Positive definiteness of the coefficient matrix K implies that $F(u,v) > 0$ for all $\mathbf{u} = (\,u,v\,)^T \neq \mathbf{0}$, and hence $\mathbf{0}$ is, in fact, a global minimum.

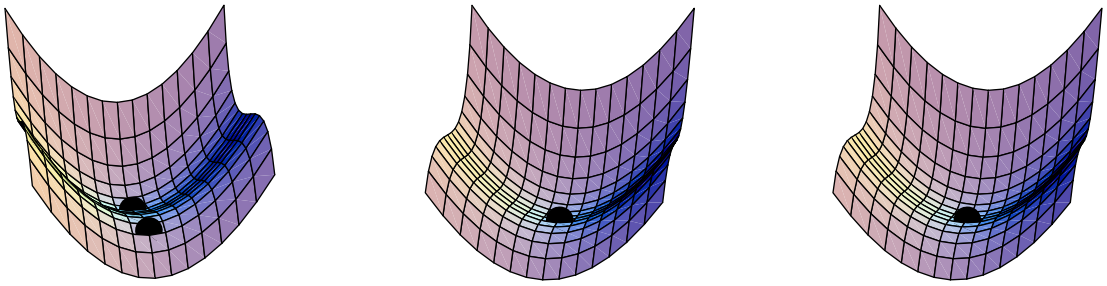
In general, any quadratic function $Q(\mathbf{u}) = Q(u_1,\dots,u_n)$ can be written in the form

$$Q(\mathbf{u}) = \mathbf{u}^T K\,\mathbf{u} - 2\,\mathbf{b}^T\mathbf{u} + c = \sum_{i,j=1}^{m} k_{ij}\,u_i\,u_j - 2\sum_{i=1}^{n} b_i\,u_i + c, \tag{19.71}$$

where $K = K^T$ is a symmetric $n \times n$ matrix, $\mathbf{b} \in \mathbb{R}^n$ is a fixed vector, and $c \in \mathbb{R}$ is a scalar. A straightforward computation produces the formula for its gradient and Hessian matrix:

$$\nabla Q(\mathbf{u}) = 2\,K\,\mathbf{u} - 2\,\mathbf{b}, \qquad\qquad \nabla^2 Q(\mathbf{u}) = 2\,K. \tag{19.72}$$

As a result, the critical points of the quadratic function are the solutions to the linear system $K\,\mathbf{u} = \mathbf{b}$. If K is nonsingular, there is a unique critical point $\mathbf{u}^\star$, which is a strict local minimum if and only if $K > 0$ is positive definite. In fact, Theorem 4.1 tells us that, in the positive definite case, $\mathbf{u}^\star$ is a strict *global* minimum for $Q(\mathbf{u})$. Thus, the algebraic



$u^2 + v^2 - v^3$ $u^2 + v^4$ $u^2 + v^3$

Figure 19.7. Critical Points.

approach of Chapter 4 provides additional, global information that cannot be gleaned directly from the local, multivariable calculus Theorem 19.45. But algebra is only able to handle quadratic minimization problems with ease. The analytical classification of minima and maxima of more complicated objective functions necessarily relies the gradient and Hessian criteria of Theorem 19.45.

Example 19.47. The function

$$F(u,v) = u^2 + v^2 - v^3 \qquad \text{has gradient} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u \\ 2\,v - 3\,v^2 \end{pmatrix}.$$

There are two solutions to the critical point equation $\nabla F = \mathbf{0}$: $\mathbf{u}_1^\star = (\,0,0\,)^T$ and $\mathbf{u}_2^\star = \left(\,0, \frac{2}{3}\,\right)^T$. The Hessian matrix of the objective function is

$$\nabla^2 F(u,v) = \begin{pmatrix} 2 & 0 \\ 0 & 2 - 6\,v \end{pmatrix}.$$

At the first critical point, the Hessian $\nabla^2 F(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ is positive definite. Therefore, the origin is a strict local minimum. On the other hand, $\nabla^2 F\left(\,0, \frac{2}{3}\,\right) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$ is indefinite, and hence $\mathbf{u}_2^\star = \left(\,0, \frac{2}{3}\,\right)^T$ a saddle point. The function is graphed in Figure 19.7, with the critical points indicated by the small solid balls. The origin is, in fact, only a local minimum, since $F(0,0) = 0$, whereas $F(0,v) < 0$ for all $v > 1$. Thus, there is no global minimum or maximum on $\mathbb{R}^2$.

Next, consider the function

$$F(u,v) = u^2 + v^4, \qquad \text{with gradient} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u \\ 4\,v^3 \end{pmatrix}.$$

The only critical point is the origin $u = v = 0$. The origin is a strict global minimum because $F(u,v) > 0 = F(0,0)$ for all $(u,v) \neq (\,0,0\,)^T$. However, its Hessian matrix

$$\nabla^2 F(u,v) = \begin{pmatrix} 2 & 0 \\ 0 & 12\,v^2 \end{pmatrix}$$

is only positive semi-definite at the origin, $\nabla^2 F(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$, and the second derivative test is inconclusive.

On the other hand, the origin $u = v = 0$ is also the only critical point for the function

$$F(u,v) = u^2 + v^3 \qquad \text{with} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u \\ 3\,v^2 \end{pmatrix}.$$

The Hessian matrix is

$$\nabla^2 F(u,v) = \begin{pmatrix} 2 & 0 \\ 0 & 6\,v \end{pmatrix}, \qquad \text{and so} \qquad \nabla^2 F(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

is the same positive semi-definite matrix at the critical point. However, in this case $(0,0)$ is not a local minimum; indeed

$$F(0,v) < 0 = F(0,0) \qquad \text{whenever} \qquad v < 0,$$

and so there exist points arbitrarily close to the origin where F takes on smaller values. As illustrated in Figure 19.7, the origin is, in fact, a degenerate saddle point.

Finally, the function

$$F(u,v) = u^2 - 2\,u\,v + v^2 \qquad \text{has gradient} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u - 2\,v \\ -2\,u + 2\,v \end{pmatrix},$$

and so every point $u = v$ is a critical point. The Hessian matrix

$$\nabla^2 F(u,v) = \begin{pmatrix} F_{uu} & F_{uv} \\ F_{uv} & F_{vv} \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

is positive semi-definite everywhere. Since $F(u,u) = 0$, while $F(u,v) = (u-v)^2 > 0$ when $u \neq v$, each of these critical points is a non-isolated local minimum, but not a strict local minimum. Thus, comparing the three preceding examples, we see that a semi-definite Hessian cannot completely distinguish critical points.

Finally, the reader should always remember that first and second derivative tests only determine the local behavior of the function near the critical point. They cannot be used to determine whether or not we are at a global minimum. This requires some additional analysis, and, often, a fair amount of ingenuity.

Minimization of Scalar Functions

In practical optimization, one typically bypasses the preliminary characterization of minima as critical points, and instead implements a direct iterative procedure that constructs a sequence of successively better approximations. As the computation progresses, the approximations are adjusted so that the objective function is made smaller and smaller, which, we hope, will ensure that we are converging to some form of minimum.

As always, to understand the issues involved, it is essential to consider the simplest scalar situation. Thus, we are given the problem of minimizing a scalar function $F(u)$ on a bounded interval $a \leq u \leq b$. The minimum value can either be at an endpoint or an interior minimum. Let us first state a result that plays a similar role to the Intermediate Value Theorem 19.19 that formed the basis of the bisection method for finding roots.

Lemma 19.48. *Suppose that $F(u)$ is defined and continuous for all $a \leq u \leq b$. Suppose that we can find a point $a < c < b$ such that $F(c) < F(a)$ and $F(c) < F(b)$. Then $F(u)$ has a minimum at some point $a < u^\star < b$.*

The proof is an easy consequence of Theorem 19.36. Therefore, if we find three points $a < c < b$ satisfying the conditions of the lemma, we are assured of the existence of a local minimum for the function between the two endpoints. Once this is done, we can design an algorithm to home in on the minimum $u^\star$. We choose another point, say d between a and c and evaluate $F(d)$. If $F(d) < F(c)$, then $F(d) < F(a)$ also, and so the points $a < d < c$ satisfy the hypotheses of Lemma 19.48. Otherwise, if $F(d) > F(c)$ then the points $d < c < b$ satisfy the hypotheses of the lemma. In either case, a local minimum has been narrowed down to a smaller interval, either $[a, c]$ or $[d, b]$. In the unlikely even that $F(d) = F(c)$, one can try another point instead — unless the objective function is constant, one will eventually find a suitable value of d. Iterating the method will produce a sequence of progressively smaller and smaller intervals in which the minimum is trapped, and, just like the bisection method, the endpoints of the intervals get closer and closer to $u^\star$.

The one question is how to choose the point d. We described the algorithm when it was selected to lie between a and c, but one could equally well try a point between c and b. To speed up the algorithm, it makes sense to place d in the larger of the two subintervals $[a, c]$ and $[c, b]$. One could try placing d in the midpoint of the interval, but a more inspired choice is to place it at position ■ The result is the *Golden Section Method*, and is outlined in the accompanying program. At each stage, the length of the interval has been reduced by a factor of $\frac{1}{2}\big(\sqrt{5} - 1\big) \approx .61803$. Thus, the convergence rate is linear, and a bit slower than the bisection algorithm.

Another strategy is to use an interpolating polynomial through the three points on the graph of $F(u)$ and use the minimum value of that polynomial as the next approximation to the minimum. According to Exercise ■, the minimizing value is at

$$d = \frac{m\,s - n\,t}{s - t}\,,$$

where

$$s = \frac{F(c) - F(a)}{c - a}\,, \qquad t = \frac{F(b) - F(c)}{b - c}\,, \qquad m = \frac{a + c}{2}\,, \qquad n = \frac{c + b}{2}\,.$$

As long as $a < c < b$ satisfy the hypothesis of Lemma 19.48, we are assured that the quadratic interpolant has a minimum (and not a maximum!), and that the minimum remains between the endpoints of the interval. If the length of the interval is small, the minimum value should be a good approximation to the minimizer $u^\star$ of $F(u)$ itself. Once d is determined, the algorithm proceeds as before. In this case, convergence is not quite guaranteed, or, in unfavorable situations, could be much slower than in the preceding method. One can even try using the method when the function values do not satisfy the hypothesis of Lemma 19.48, although now the new point d will not necessarily lie between a and b. Worse, the quadratic interpolant may have a maximum at d, and one ends up going in the wrong direction, which can even happen in the minimizing case due to the

discrepancy between it and the objective function $F(u)$. Thus, this case must be handled with more caution, and convergence of the scheme is much more fraught with danger.

A final idea is to focus not on the objective function $F(u)$ but rather its derivative $f(u) = F'(u)$. The critical points of F are the roots of $f(u) = 0$, and so one can use one of the solution methods, e.g., bisection or Newton's method, to find the critical points. Of course, one must then take care that the critical point $u^\star$ is indeed a minimum, as it could equally well be a maximum of the original objective function. (It will probably not be a saddle point, as these do not correspond to simple roots of $f(u)$.) But this can be checked by looking at the sign of $F''(u^\star) = f'(u^\star)$ at the root; indeed, if we use Newton's method we will be computing the derivative at each stage of the algorithm, and can stop looking if the derivative is of the wrong sign.

Gradient Descent

Now, let us turn our attention to multi-dimensional optimization problems. We are seeking to minimize a (smooth) scalar objective function $F(\mathbf{u}) = F(u_1, \dots, u_n)$. According to Theorem 19.39, at any given point $\mathbf{u}$ in the domain of definition of F, the negative gradient vector $-\nabla F(\mathbf{u})$, if nonzero, points in the direction of the steepest decrease in F. Thus, to minimize F, an evident strategy is to "walk downhill", and, to be efficient, walk downhill as fast as possible, namely in the direction $-\nabla F(\mathbf{u})$. After walking in this direction for a little while, we recompute the gradient, and this tells us the new direction to head downhill. With luck, we will eventually end up at the bottom of the valley, i.e., at a (local) minimum value of the objective function.

This simple idea forms the basis of the *gradient descent* method for minimizing the objective function $F(\mathbf{u})$. In a numerical implementation, we start the iterative procedure with an initial guess $\mathbf{u}^{(0)}$, and let $\mathbf{u}^{(k)}$ denote the k^{th} approximation to the minimum $\mathbf{u}^\star$. To compute the next approximation, we move away from $\mathbf{u}^{(k)}$ in the direction of the negative gradient, and hence

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - t_k \, \nabla F(\mathbf{u}^{(k)}) \tag{19.73}$$

for some positive scalar $t_k > 0$ that indicates how far we move in the negative gradient direction. We are free to adjust t_k so as to optimize our descent path, and this is the key to the success of the method.

If $\nabla F(\mathbf{u}^{(k)}) \neq \mathbf{0}$, then, at least when $t_k > 0$ is sufficiently small,

$$F(\mathbf{u}^{(k+1)}) < F(\mathbf{u}^{(k)}), \tag{19.74}$$

and so $\mathbf{u}^{(k+1)}$ is, presumably, a better approximation to the desired minimum. Clearly, we cannot choose t_k too large or we run the risk of overshooting the minimum and reversing the inequality (19.74). Think of walking downhill in the Swiss Alps. If you walk too far in a straight line, which is what happens as t_k increases, then you might very well miss the valley and end up higher than you began — not a good strategy for descending to the bottom! On the other hand, if we choose t_k too small, taking very tiny steps, then the method will converge to the minimum much too slowly.

How should we choose an optimal value for the factor t_k? Keep in mind that the goal is to minimize $F(\mathbf{u})$. Thus, a good strategy would be to set t_k equal to the value of $t > 0$ that minimizes the scalar objective function

$$g(t) = F\big(\, \mathbf{u}^{(k)} - t\, \nabla F(\mathbf{u}^{(k)}) \,\big) \tag{19.75}$$

obtained by restricting $F(\mathbf{u})$ to the ray emanating from $\mathbf{u}^{(k)}$ that lies in the negative gradient direction. Physically, this corresponds to setting off in a straight line in the direction of steepest decrease in our altitude, and continuing on in this direction until we cannot go down any further. Barring luck, we will not have reached the actual bottom of the valley, but must then readjust our direction and continue on down the hill in a series of straight line paths, each connecting $\mathbf{u}^{(k)}$ to $\mathbf{u}^{(k+1)}$.

In practice, one can rarely compute the minimizing value $t^\star$ of $g(t)$ exactly. Instead, we use one of the scalar minimization algorithms presented in the previous subsection. Note that we only need to look for a minimum among positive values of $t > 0$, since our choice of the negative gradient direction assures us that, at least for t sufficiently small and positive, $g(t) < g(0)$.

Example 19.49. ∎

Conjugate Gradients

The one complication with the basic gradient descent method is that it may take a long time to reach the minimum. This is a danger if the scalar factors t_k are small, and we end up taking very tiny steps in each round of the iteration. This occurs if we are looking for a minimum in a long narrow valley, as illustrated in Figure valley∎. The initial step takes us into the valley, but then we spend a long time meandering back and forth along the valley floor before we come close to the true minimum.

One method to avoid such difficulties, and speed up the convergence rate of the scheme, is to use the method of *conjugate directions*, modeled on the quadratic minimization procedure discussed in in Section 16.2.

Chapter 20

Nonlinear Ordinary Differential Equations

Most physical processes are modeled by differential equations. First order ordinary differential equations, also known as dynamical systems, arise in a wide range of applications, including population dynamics, mechanical systems, planetary motion, ecology, chemical diffusion, etc., etc. See [**19**, **72**, **ODES**] for additional material and applications.

The goal of this chapter is to study and solve initial value problems for nonlinear systems of ordinary differential equations. Of course, very few nonlinear systems can be solved explicitly, and so one must typically rely on a suitable numerical scheme in order to approximate the solution. However, numerical schemes do not always give accurate results. Without some basic theoretical understanding of the nature of solutions, equilibrium points, and their stability, one would not be able to understand when numerical solutions (even those provided by standard well-used packages) are to be trusted. Moreover, when testing a numerical scheme, it helps to have already assembled a repertoire of nonlinear problems in which one already knows one or more explicit analytic solutions. Further tests and theoretical results can be based on first integrals (also known as conservation laws) or, more generally, Lyapunov functions. Although we have only space to touch on these topics briefly, but, we hope, this will whet the reader's appetite for delving into this subject in more depth. The references [**19**, **46**, **72**, **80**, **85**] can be profitably consulted.

Our overriding emphasis will be on those properties of solutions that have physical relevance. Finding a solution to a differential equation is not be so important if that solution never appears in the physical model represented by the system, or is only realized in exceptional circumstances. Thus, equilibrium solutions, which correspond to configurations in which the physical system does not move, only occur in everyday a physics if they are stable. An unstable equilibrium will not appear in practice, since slight perturbations in the system or its physical surroundings will immediately dislodge the system far away from equilibrium.

Finally, we present a few of the most basic numerical solution techniques for ordinary differential equations. We begin with the most basic Euler scheme, and work up to the Runge–Kutta fourth order method, which is one of the most popular methods for everyday applications.

20.1. First Order Systems of Ordinary Differential Equations.

In this section, we introduce the basic object of study — initial value problems for first order systems of ordinary differential equations. We have already dealt with the linear case in Chapter 9, and so our emphasis will be on nonlinear phenomena. Our emphasis

will be on first order systems. Indeed, as we shall see, there is a simple reformulation that converts higher order equations and systems into equivalent first order systems, and so we do not lose any generality by restricting our attention to the first order situation.

Scalar Ordinary Differential Equations

As always, to study a new problem, it is essential to begin with the simplest case. Consider the scalar, first order ordinary differential equation

$$\frac{du}{dt} = F(t,u). \tag{20.1}$$

In many applications, the independent variable t represents time, and the unknown function $u(t)$ is some dynamical physical quantity. Under appropriate conditions on the right hand side (to be formalized in the following section), the solution $u(t)$ is uniquely specified by its value at a single time,

$$u(t_0) = u_0. \tag{20.2}$$

The combination (20.1), (20.2) is referred to as an *initial value problem*, and our goal is to devise both analytical and numerical solution methods.

The simplest class is the *autonomous* differential equations, meaning that the right hand side does not explicitly depend upon the time variable:

$$\frac{du}{dt} = F(u). \tag{20.3}$$

Autonomous scalar equations can be solved by a direct integration. We first divide both sides by $F(u)$, whereby $\dfrac{1}{F(u)}\,\dfrac{du}{dt} = 1$, and then integrate with respect to t; the result is

$$\int \frac{1}{F(u)}\,\frac{du}{dt}\,dt = \int dt = t + k,$$

where k is a constant of integration. The left hand integral can be evaluated by the change of variables that replaces t by u, with $du = (du/dt)\,dt$, and so

$$\int \frac{1}{F(u)}\,\frac{du}{dt}\,dt = \int \frac{du}{F(u)} = G(u),$$

where $G(u)$ represents any convenient anti-derivative[†] of the function $1/F(u)$. Thus, the solution can be written in implicit form

$$G(u) = t + k. \tag{20.4}$$

A more direct version of the method is to formally rewrite the differential equation (20.3) in the "separated form"

$$\frac{du}{F(u)} = dt, \qquad \text{and then integrate both sides} \qquad G(u) = \int \frac{du}{F(u)} = \int dt = t + k.$$

[†] Technically, a second constant of integration should appear here, but this can be absorbed into the previous constant k, and so is unnecessary.

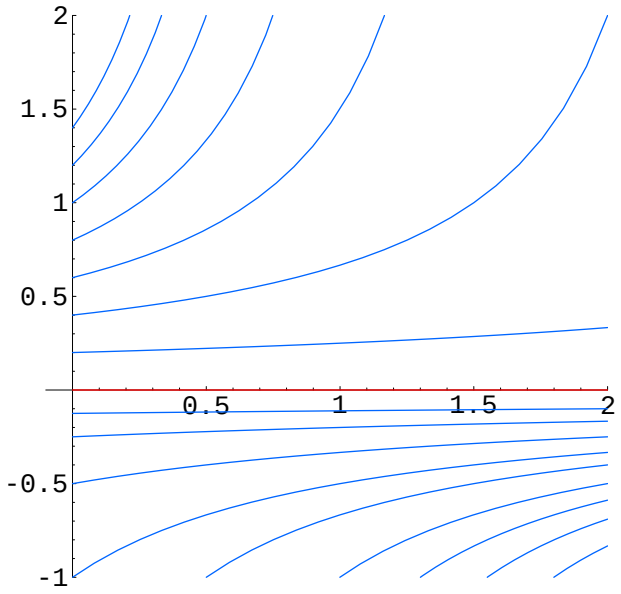


Figure 20.1. Solutions to $\dot{u} = u^2$.

If we are able to solve the implicit equation (20.4), we may thereby obtain the explicit solution

$$u(t) = H(t + k) \tag{20.5}$$

in terms of the inverse function $H = G^{-1}$. Finally, to satisfy the initial condition (20.2), we set $t = t_0$ in the implicit solution formula (20.4), whereby $G(u_0) = t_0 + k$. Therefore, the solution to our initial value problem is

$$G(u) - G(u_0) = t - t_0, \qquad \text{or, in explicit form,} \qquad u(t) = H\big(t - t_0 + G(u_0)\big). \tag{20.6}$$

Before completing the analysis of the solution method, let us consider an elementary example.

Example 20.1. Consider the autonomous initial value problem

$$\frac{du}{dt} = u^2, \qquad u(t_0) = u_0.$$

To solve the differential equation, we rewrite it in the separated form

$$\frac{du}{u^2} = dt, \qquad \text{and then integrate both sides:} \qquad -\frac{1}{u} = \int \frac{du}{u^2} = t + k.$$

Solving for u, we deduce the general solution formula

$$u = -\frac{1}{t + k}\,.$$

To specify the integration constant k, we evaluate u at the initial time t_0; this implies

$$u_0 = -\frac{1}{t_0 + k}\,, \qquad \text{so that} \qquad k = -\frac{1}{u_0} - t_0.$$

Therefore, the solution to the initial value problem is

$$u = \frac{u_0}{1 - u_0(t - t_0)}\,. \tag{20.7}$$

Figure 20.1 shows the graphs of some typical solutions.

Observe that as t approaches the critical value $t_\star = t_0 + 1/u_0$, the solution blows up: $u(t) \to \infty$. The blow-up time depends upon the initial data — the larger u_0 is, the sooner the solution goes off to infinity. If the initial data is negative, the solution is well-defined for all $t > t_0$, but the solution has a singularity in the past. The only solution that exists for all positive and negative time is the constant solution $u(t) \equiv 0$, corresponding to the initial condition $u_0 = 0$.

In general, the constant or *equilibrium solutions* to an autonomous ordinary differential equation play a particularly important role. If $u(t) \equiv u^\star$ is a constant solution, then $du/dt \equiv 0$, and hence the differential equation (20.3) implies that $F(u^\star) = 0$. Therefore, the equilibrium solutions coincide with the *roots* of the function $F(u)$. In point of fact, the derivation of our formula for the solution (20.4) assumed that we were *not* at an equilibrium point where $F(u) = 0$. In the preceding example, our final solution formula (20.7) to the implicit equation happened to include the equilibrium solution $u(t) \equiv 0$, but this was a lucky accident, and one must typically take more care that such solutions do not elude us when applying the integration method.

Example 20.2. Although a population of people, animals, or bacteria consists of individuals, the aggregate behavior can often be effectively modeled by a continuous dynamical system. As first proposed by the English economist Thomas Malthus, the population of a species grows, roughly, in proportion to its size. Thus, the number of individuals $N(t)$ in a species at time t satisfies a first order differential equation of the form

$$\frac{dN}{dt} = \rho\, N, \tag{20.8}$$

where the proportionality factor ρ measures the rate of growth, namely the difference between the birth rate and the death rate. Thus, if births exceed deaths, $\rho > 0$, and the population increases, whereas if $\rho < 0$, more individuals are dying and the population shrinks.

In the very simplest model, the growth rate ρ is assumed to be independent of the population size, and we have the simple linear ordinary differential equation (8.1) that we solved at the beginning of Chapter 8. The solutions satisfy the exponential or Malthusian growth law $N(t) = N_0\, e^{\rho\, t}$, where $N_0 = N(0)$ is the initial population size. Thus, if $\rho > 0$ the population grows without limit, while if $\rho < 0$ the population dies out , $N(t) \to 0$, at an exponentially fast rate. The Malthus population model gives a reasonably accurate description of the behavior of an isolated population in an environment with unlimited resources.

In a more realistic model, the growth rate will depend upon the size of the population as well as external environmental factors. For example, in an environment with limited resources, small populations will grow, whereas an excessively large population will have

insufficient resources to survive, and so its growth rate will be negative. In other words, the growth rate $\rho(N) > 0$ when $N < N_\star$, while $\rho(N) < 0$ when $N > N_\star$, where the *carrying capacity* $N_\star > 0$ is a number that depends upon the resource availability. The simplest class of functions that satifies these two inequalities are of the form $\rho(N) = \lambda(N_\star - N)$, where $\lambda > 0$ is a positive constant. Subsituting this expression for the growth rate into (20.8) leads to the nonlinear population model

$$\frac{dN}{dt} = \lambda N (N_\star - N). \tag{20.9}$$

In deriving this model, we assumed that the environment is not changing over time; a dynamical environment would lead to a more complicated non-autonomous differential equation.

Before analyzing the solutions to the nonlinear population model, let us make a change of variables, and set $u(t) = N(t)/N_\star$, so that u represents the size of the population in proportion to the *carrying capacity* $N_\star$ of the environment. Then $u(t)$ satisfies the so-called *logistic differential equation*

$$\frac{du}{dt} = \lambda u (1 - u), \qquad\qquad u(0) = u_0, \tag{20.10}$$

where we assign the initial time to be $t_0 = 0$ for simplicity. This differential equation is the continuous counterpart of the logistic map (19.19). However, unlike its discrete cousin, the logistic differential equation is quite sedate, and its solutions easily understood. First, there are two equilibrium solutions: $u(t) \equiv 0$ and $u(t) \equiv 1$, obtained by setting the right hand side of the equation equal to zero. The first represents a nonexistent population with no individuals and hence no reproduction. The second equilibrium solution corresponds to a population $N(t) \equiv N_\star$ that is at the ideal size for the environment, and so deaths exactly balance births. In all other situations, the population size will vary over time.

To integrate the logistic differential equation, we proceed as above, first writing it in the separated form $\dfrac{du}{u(1-u)} = \lambda\, dt$. Integrating both sides, and using partial fractions,

$$\lambda t + k = \int \frac{du}{u(1-u)} = \int \left(\frac{du}{u} + \frac{du}{1-u} \right) = \log \left| \frac{u}{1-u} \right| ,$$

where k is a constant of integration. Therefore

$$\frac{u}{1-u} = c e^{\lambda t}, \qquad \text{where} \qquad c = \pm e^k.$$

Solving for u, we deduce the solution

$$u(t) = \frac{c e^{\lambda t}}{1 + c e^{\lambda t}} \,. \tag{20.11}$$

The constant of integration is fixed by the initial condition. Solving the algebraic equation

$$u_0 = u(0) = \frac{c}{1+c} \qquad \text{yields} \qquad c = \frac{u_0}{1 - u_0} \,.$$

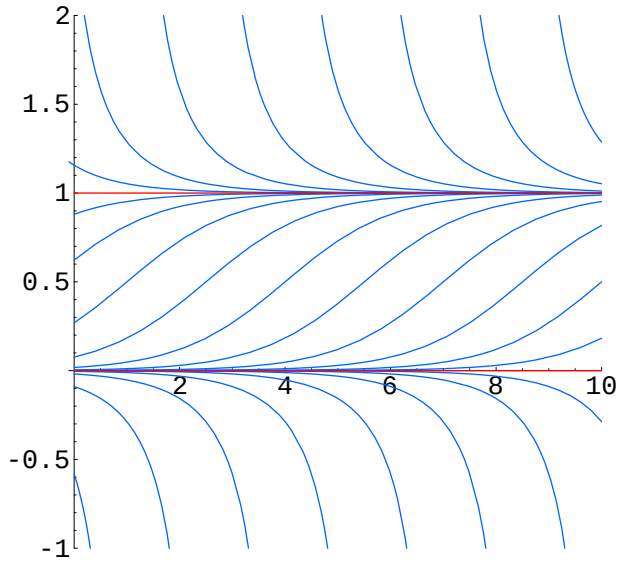


Figure 20.2. Solutions to $u' = u(1-u)$.

Substituting the result back into the solution formula (20.11) and simplifying, we find

$$u(t) = \frac{u_0\, e^{\lambda t}}{1 - u_0 + u_0\, e^{\lambda t}}\,. \tag{20.12}$$

The solutions are illustrated in Figure 20.2. Interestingly, the equilibrium solutions are not covered by the integration method, but do appear in the final solution formula, corresponding to initial data $u_0 = 0$ and $u_0 = 1$ respectively.

When using the logistic equation to model population dynamics, the initial data is assumed to be positive, $u_0 > 0$. As time $t \to \infty$, the solution (20.12) tends to the equilibrium value $u(t) \to 1$. For small initial values $u_0 \ll 1$ the solution initially grows at an exponential rate λ, corresponding to a population with unlimited resources. However, as the population increases, the gradual lack of resources tends to slow down the growth rate, and eventually the population saturates at the equilibrium value. On the other hand, if $u_0 > 1$, the population is too large to be sustained by the resources, and so dies off until it reaches the same saturation value. If $u_0 = 0$, then the solution remains at equilibrium $u(t) \equiv 0$. Finally, when $u_0 < 0$, the solution only exists for a finite amount of time, with $u(t) \to -\infty$ as $t \to t_\star = \log(u_0/(u_0 - 1))$. Of course, this final case does not correspond to a physical situation since we cannot have a negative population!

The separation of variables method used to solve autonomous equations can be straightforwardly extended to a special class of non-autonomous equations. A *separable* ordinary differential equation has the form

$$\frac{du}{dt} = a(t)\, F(u), \tag{20.13}$$

in which the right hand side is the product of a function of t only times a function of u.

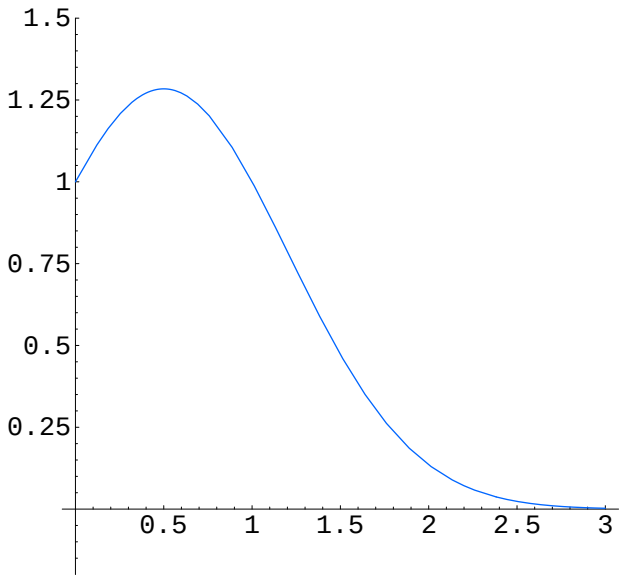


Figure 20.3. Solution to the Initial Value Problem $\dot{u} = (1 - 2t)\, u$, $u(0) = 1$.

To solve the equation, we rewrite it in the separated form

$$\frac{du}{F(u)} = a(t)\, dt.$$

Integrating both sides leads to the solution in implicit form

$$G(u) = \int \frac{du}{F(u)} = \int a(t)\, dt = A(t) + k. \tag{20.14}$$

The integration constant k is then fixed by the initial condition.

Example 20.3. Let us solve the initial value problem

$$\frac{du}{dt} = (1 - 2t)\, u, \qquad u(0) = 1. \tag{20.15}$$

We begin by writing the differential equation in separated form $\dfrac{du}{u} = (1 - 2t)\, dt$ Integrating both sides leads to

$$\log u = \int \frac{du}{u} = \int (1 - 2t)\, dt = t - t^2 + k.$$

We can readily solve for

$$u = c \exp(t - t^2),$$

where $c = \pm e^k$. This consitutes the general solution to the differential equation, and happens to include the equilibrium solution $u(t) \equiv 0$. The initial condition requires that $c = 1$, and hence $u(t) = e^{t - t^2}$ is the unique solution to the initial value problem. The solution is graphed in Figure 20.3.

First Order Systems

A first order system of ordinary differential equation has the general form

$$\frac{du_1}{dt} = F_1(t, u_1, \dots, u_n), \qquad \cdots \qquad \frac{du_n}{dt} = F_n(t, u_1, \dots, u_n). \tag{20.16}$$

The unknowns $u_1(t), \dots, u_n(t)$ are scalar functions of the real variable t, which usually represents time. We shall write the system more compactly in vector form

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}), \tag{20.17}$$

so that $\mathbf{F}\colon \Omega \to \mathbb{R}^n$ is a vector-valued function of $n+1$ variables defined on an open domain $\Omega \subset \mathbb{R}^{n+1}$. By a *solution* to the differential equation, we mean a vector-valued function $\mathbf{u}(t)$ that is defined and continuously differentiable on an interval $a < t < b$, and, moreover, satisfies the differential equation on its interval of definition. The solution $\mathbf{u}(t)$ serves to parametrize a curve $C \subset \mathbb{R}^n$, called the solution *trajectory* or *orbit*.

In this chapter, we shall be concerned with initial value problems for first order systems of ordinary differential equations. The general initial conditions are

$$u_1(t_0) = a_1, \qquad u_2(t_0) = a_2, \qquad \cdots \qquad u_n(t_0) = a_n, \tag{20.18}$$

or, in vectorial form, $\mathbf{u}(t_0) = \mathbf{a}$. Here t_0 is a prescribed initial time, while the vector $\mathbf{a} = (\, a_1, a_2, \dots, a_n \,)^T$ prescribes the initial position of the desired solution. In favorable situations, the initial conditions serve to uniquely specify a solution to the differential equations — at least for a short time. As in the scalar case, the general issues of existence and uniquenss of solutions must be properly addressed.

A system of differential equations is called *autonomous* if the right hand side does not explicitly depend upon the time t, and so has the form

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}). \tag{20.19}$$

One important class of autonomous first order systems are the steady state fluid flows in two and three dimensions. In this case, $\mathbf{F}(\mathbf{u})$ represents the fluid velocity vector field at the position $\mathbf{u}$. A solution $\mathbf{u}(t)$ represents the trajectory of an individual fluid particle. The differential equation tells us that the fluid velocity at each point of its trajectory matches the prescribed vector field. Appendices A and B contain additional details.

An *equilibrium solution* to an autonomous system is defined to be a constant solution: $\mathbf{u}(t) \equiv \mathbf{u}^\star$ for all t. Since the solution is constant, its derivative must vanish, $d\mathbf{u}/dt \equiv \mathbf{0}$. Hence, every equilibrium solution corresponds to a *root* or solution to the system of algebraic equations

$$\mathbf{F}(\mathbf{u}^\star) = \mathbf{0} \tag{20.20}$$

prescribed by the vanishing of the right hand side of the system.

Example 20.4. A *predator-prey system* is a simplified ecological model of two species: the predators which feed on the prey. For example, the predators might be lions in the Serengeti and the prey zebra. We let $u(t)$ represent the number of prey, and $v(t)$ the number of predators at time t. Both species obey a population growth model of the form (20.8), and so the dynamical equations have the general form

$$\frac{du}{dt} = \rho\, u, \qquad\qquad \frac{dv}{dt} = \sigma\, v,$$

where the individual growth rates ρ, σ may depend upon the other species. The more prey, i.e., the larger u is, the faster the predators reproduce, while a lack of prey will cause them to die off. On the other hand, the more predators, the faster the prey are consumed and the slower their net rate of growth. If we assume that the environment has unlimited resources, then the simplest model that incorporates these assumptions is the *Lotka–Volterra system*

$$\frac{du}{dt} = \alpha\, u - \delta\, u\, v, \qquad\qquad \frac{dv}{dt} = -\beta\, v + \gamma\, u\, v, \tag{20.21}$$

corresponding to growth rates $\rho = \alpha - \delta\, v$, $\sigma = -\beta + \gamma\, u$. The parameters $\alpha, \beta, \gamma, \delta > 0$ are all positive, and their precise values will depend upon the species involved and how they interact; they are determined by field data, along, perhaps, with educated guesses. In particular, α represents the unrestrained growth rate of the prey in the absence of predators, while $-\beta$ represents the rate that the predators die off in the absence of food. The nonlinear terms model the interaction of the two species. The initial conditions $u(t_0) = u_0$, $v(t_0) = v_0$ represent the initial populations of the two species.

We will discuss the integration of the predator-prey system (20.21) in Section 20.3. Here, let us content ourselves with determining the possible equilibria. Setting the right hand sides of the system to zero, leads to the algebraic system

$$0 = \alpha\, u - \delta\, u\, v = u(\alpha - \delta\, v), \qquad\qquad 0 = -\beta\, v + \gamma\, u\, v = v(-\beta + \gamma\, u).$$

There are two distinct equilibria, namely

$$u_1^\star = v_1^\star = 0, \qquad\qquad u_2^\star = \beta/\gamma, \quad v_2^\star = \alpha/\delta.$$

The first is the uninteresting equilibrium where there are no animals — no predators and no prey. The second is a nontrivial solution in which both populations maintain a steady value, in which the birth rate of the prey is precisely sufficient to continuously feed the predators. Is this a feasible solution? Or, to state the question more mathematically, is this a stable equilibrium? We shall develop the tools to answer this question below.

Higher Order Systems

Many physical systems are modeled by nonlinear systems of differential equations depending upon higher order derivatives of the unknowns. But there is an easy trick that will reduce any higher order ordinary differential equation or system to an equivalent first order system. "Equivalence" means that there is a one-to-one correspondence between

solutions. This implies that in practice, it suffices to analyze first order systems. There is no need to develop a separate theory for higher order ordinary differential equations. Moreover, almost all numerical solution algorithms are designed for first order systems, and so to numerically integrate a higher order system, one also must place it into equivalent first order form.

We have already encountered the basic idea in our discussion of the phase plane approach to second order scalar equations. Given a second order equation

$$\frac{d^2u}{dt^2} = F\left(t, u, \frac{du}{dt}\right), \tag{20.22}$$

we set $v = \dfrac{du}{dt}$. Since $\dfrac{dv}{dt} = \dfrac{d^2u}{dt^2}$, the functions u, v satisfy the first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = F(t, u, v). \tag{20.23}$$

Conversely, it is easy to check that if $(\, u(t), v(t)\,)^T$ is any solution to the system, then its first component defines a solution to the scalar equation. The basic initial conditions $u(t_0) = u_0, v(t_0) = v_0$ for the first order system translate into a pair of initial conditions $u(t_0) = u_0, \dot{u}(t_0) = v_0$ specifying the value of the solution and its first order derivative for the second order equation.

Example 20.5. The forced *van der Pol equation*

$$\frac{d^2u}{dt^2} + \alpha\,(u^2 - 1)\,\frac{du}{dt} + u = f(t) \tag{20.24}$$

arises in the modeling of an electrical circuit with a triode whose resistance changes with the current, [**EE**], certain chemical reactions, and wind-induced motions of structures. To convert it into an equivalent first order system, we set $v = du/dt$, whence

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = f(t) - \,\alpha\,(u^2 - 1)v - u. \tag{20.25}$$

Similarly, given a third order equation

$$\frac{d^3u}{dt^3} = F\left(t, u, \frac{du}{dt}\,, \frac{d^2u}{dt^2}\right), \qquad \text{we set} \qquad v = \frac{du}{dt}\,, \qquad w = \frac{dv}{dt} = \frac{d^2u}{dt^2}\,.$$

The variables u, v, w satisfy the equivalent first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = w, \qquad \frac{dw}{dt} = F(t, u, v, w).$$

The general method of construction should now be clear.

Example 20.6. The Newtonian equations for a mass m moving in a potential force field are a second order system of the form $m\,d^2\mathbf{u}/dt^2 = \nabla F(\mathbf{u})$ in which $\mathbf{u}(t)$ represents the position of the mass and $F(\mathbf{u}) = F(u, v, w)$ the potential function. In components,

$$m\,\frac{d^2u}{dt^2} = \frac{\partial F}{\partial u}\,, \qquad m\,\frac{d^2v}{dt^2} = \frac{\partial F}{\partial v}\,, \qquad m\,\frac{d^2w}{dt^2} = \frac{\partial F}{\partial w}\,. \tag{20.26}$$

For example, a planet moving in the sun's gravitational field satisfies the system with the gravitational potential

$$F(\mathbf{u}) = \frac{\alpha}{\| \mathbf{u} \|} = \frac{\alpha}{\sqrt{u^2 + v^2 + w^2}}, \qquad \nabla F(\mathbf{u}) = - \frac{\alpha \, \mathbf{u}}{\| \mathbf{u} \|^3} = \frac{\alpha}{(u^2 + v^2 + w^2)^{3/2}} \begin{pmatrix} u \\ v \\ w \end{pmatrix}.$$

To convert the second order Newtonian equations into a first order system, we let $\mathbf{v} = \dot{\mathbf{u}}$ be the velocity vector, with components $p = du/dt$, $q = dv/dt$, $r = dw/dt$, and so

$$\begin{aligned} &\frac{du}{dt} = p, && \frac{dv}{dt} = q, && \frac{dw}{dt} = r, \\ &\frac{dp}{dt} = \frac{1}{m} \, \frac{\partial F}{\partial u}\,(u,v,w), && \frac{dq}{dt} = \frac{1}{m} \, \frac{\partial F}{\partial v}\,(u,v,w), && \frac{dr}{dt} = \frac{1}{m} \, \frac{\partial F}{\partial w}\,(u,v,w). \end{aligned} \tag{20.27}$$

Example 20.7. There is a simple trick for changing any non-autonomous system into an autonomous system involving one additional variable. Namely, one introduces an extra coordinate $u_0 = t$ to represent the time, which satisfies the elementary differential equation $du_0/dt = 1$ with initial condition $u_0(t_0) = t_0$. Thus, the original system (20.16) can be written in an equivalent autonomous form

$$\frac{du_0}{dt} = 1, \qquad \frac{du_1}{dt} = F_1(u_0, u_1, \dots, u_n), \qquad \cdots \qquad \frac{du_n}{dt} = F_n(u_0, u_1, \dots, u_n). \tag{20.28}$$

For example, an autonomous form of the forced van der Pol system is

$$\frac{du_0}{dt} = 1, \qquad \frac{du_1}{dt} = u_2, \qquad \frac{du_2}{dt} = f(u_0) - \alpha\,(u_1^2 - 1)u_2 - u_1. \tag{20.29}$$

20.2. Existence, Uniqueness, and Continuous Dependence.

There is no general analytical method that will solve all differential equations. Indeed, even relatively simple first order, scalar, non-autonomous ordinary differential equations cannot be solved in closed form. One example is the particular *Riccati equation*

$$\frac{du}{dt} = u^2 + t \tag{20.30}$$

whose solution cannot be written in terms of elementary functions, although there is a solution formula that relies on Airy functions, cf. Exercise ■. The *Abel equation*

$$\frac{du}{dt} = u^3 + t \tag{20.31}$$

fares even worse, since its general solution cannot be written in terms of known special functions. Understanding when a given differential equation can be solved in terms of elementary functions or known special functions is an active area of contemporary research. In this vein, we cannot resist mentioning that the most important class of exact solution techniques for differential equations are those based on symmetry methods. An

introduction can be found in the first author's graduate level monograph [**117**]; see also [**32**, **84**].

Existence

Before worrying about how to solve a differential equation, either analytically, qualitatively, or numerically, it makes sense to resolve the basic issues of existence and uniqueness. First, does a solution exist? If, not, it makes no sense trying to find one. Second, is the solution uniquely determined by the initial conditions? Otherwise, the differential equation does not have much relevance in physical applications since we cannot use it as a predictive tool.

Unlike partial differential equations, which must be treated on a case-by-case basis, there are completely satisfactory general results that answer the existence and uniqueness questions for initial value problems for systems of ordinary differential equations. (However, boundary value problems are more delicate.) We will not take time to discuss the proofs of these fundamental results, which can be found in most advanced textbooks on ordinary differential equations, including [**19**, **72**, **80**, **85**].

Let us begin by stating the fundamental existence theorem.

Theorem 20.8. *Let* $\mathbf{F}(t, \mathbf{u})$ *be a continuous function*[†]*, then the initial value problem*

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}), \qquad \mathbf{u}(t_0) = \mathbf{a}, \tag{20.32}$$

has a solution $\mathbf{u} = \mathbf{f}(t)$ *defined for nearby times* $| t - t_0 | < \delta$ *for some* $\delta > 0$.

The existence theorem guarantees that the solution exists — at least for times sufficiently close to the initial instant t_0. This may be the most that can be said, although in many systems the maximal interval $\alpha < t < \beta$ of existence of the solution might be much larger — possibly infinite $-\infty < t < \infty$. The interval of existence typically depends upon both the equation and the particular initial data. For instance, in the elementary Example 20.1, the solutions to the initial value problem only exist up until time $1/u_0$, and so the larger the initial data, the shorter the time of existence. It is worth noting that this phenomenon did not appear in the linear regime, where, barring singularities that appear in the equation itself, solutions to a linear ordinary differential equation are guaranteed to exist for all time.

In practice, we will always extend solutions to their maximal interval of existence. If there is a point beyond which the solution cannot be extended, then either the solution $\| \mathbf{u}(t) \| \to \infty$ becomes unbounded in a finite time, or, if the right hand side $F(t, \mathbf{u})$ is only defined on a subset $\Omega \subset \mathbb{R}^{n+1}$, then the solution reaches the boundary $\partial\Omega$ in a finite time. Thus, a solution to an ordinary differential equation cannot suddeenly vanish into thin air. A proof of this fact can be found in the above-mentioned references.

[†] If $\mathbf{F}(t, \mathbf{u})$ is only defined on a domain $\Omega \subset \mathbb{R}^{n+1}$, then we must assume that the initial conditions $(t_0, \mathbf{a}) \in \Omega$ belong to the domain of definition.

Remark: The existence theorem can be readily adapted to apply to higher order systems of ordinary differential equations through our trick for converting a higher order system into a first order system by introducing additional variables. The appropriate initial conditions are induced from those of the equivalent first order system, as in the second order example (20.22) discussed above.

Uniqueness

As important as existence is the question of uniqueness. Does the initial value problem have more than one solution? If so, then we cannot use the differential equation to predict the behavior of the solution from its initial conditions, and so such a problem is probably worthless for applications. While continuity of the function $\mathbf{F}(t, \mathbf{u})$ is enough to guarantee that a solution exists, it is not quite enough to ensure uniqueness of the solution to the initial value problem. The difficulty can be appreciated by looking at an elementary example.

Example 20.9. Consider the nonlinear initial value problem

$$\frac{du}{dt} = \frac{5}{3}\, u^{2/5}, \qquad u(0) = 0. \tag{20.33}$$

Since the right hand side $F(u) = \frac{5}{3}\, u^{2/5}$ is a continuous function, Theorem 20.8 assures us of the existence of a solution. This autonomous scalar equation can be easily solved by separation of variables:

$$\int \frac{3}{5}\, \frac{du}{u^{2/5}} = u^{3/5} = t + c, \qquad \text{and so} \qquad u = (t + c)^{5/3}.$$

Substituting into the initial condition implies that $c = 0$, and hence $u(t) = t^{5/3}$ is a solution to the initial value problem.

On the other hand, since the right hand side vanishes at $u = 0$, the constant function $u(t) \equiv 0$ is an equilibrium solution to the differential equation. (Here is an example where the separation of variables method fails to recover the equilibrium solution.) Moreover, the equilibrium solution also has the initial value $u(0) = 0$. Therefore, we have constructed two different solutions to the initial value problem (20.33). Uniqueness is *not* valid! Worse yet, there are, in fact, an *infinite* number of solutions to the initial value problem. For any $a > 0$, the function

$$u(t) = \begin{cases} 0, & 0 \le t \le a, \\ (t - a)^{5/3}, & t \ge a, \end{cases} \tag{20.34}$$

is differentiable everywhere, even at $t = a$. Moreover, it satisfies both the differential equation and the initial condition. Several of the solutions are plotted in Figure 20.4.

In conclusion, to ensure uniqueness of solutions, we need to impose a stronger restriction than mere continuity on the differential equation. The proof of the following basic uniqueness theorem can be found in the above references.

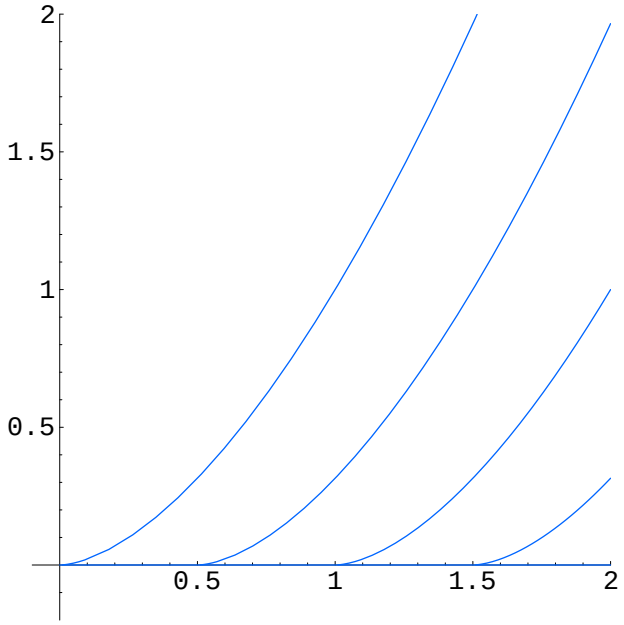


Figure 20.4. Solutions to the Differential Equation $\dot{u} = \frac{5}{3}\, u^{2/5}$.

Theorem 20.10. *If* $\mathbf{F}(t, \mathbf{u}) \in \mathrm{C}^1$ *is continuously differentiable, then there exists one and only one solution*† *to the initial value problem* (20.32).

Thus, the difficulty with the differential equation (20.33) is that the function $F(u) = \frac{5}{3}\, u^{2/5}$, although continuous everywhere, is not differentiable at $u = 0$, and hence the uniqueness theorem does not apply. On the other hand, $F(u)$ is continuously differentiable away from $u = 0$, and so any nonzero initial condition $u(t_0) = u_0 \neq 0$ will produce a unique solution — for as long as it remains away from the problematic value $u = 0$.

Remark: While having continuous partial derivatives is sufficient to guarantee uniqueness, this condition can, in fact, be slightly weakened. It suffices to require that $\mathbf{F}(t, \mathbf{u})$ is continuous as a function of t and satisfies the *Lipschitz condition*

$$\| \mathbf{F}(t, \mathbf{v}) - \mathbf{F}(t, \mathbf{u}) \| \leq C(t) \, \| \mathbf{v} - \mathbf{u} \| \tag{20.35}$$

for all $t, \mathbf{u}, \mathbf{v}$ and some positive $C(t) > 0$. See (19.18) above and the subsequent discussion for more details on Lipschitz continuity.

Blanket Hypothesis: From now on, all differential equations must satisfy the uniqueness criterion that their right hand side is continuously differentiable, or, at least, satisfies the Lipschitz inequality (20.35).

One important consequence of the uniqueness theorem is that a solution $\mathbf{u}(t)$ to an autonomous system of ordinary differential equations is either in equilibrium, not varying in time, so $\dot{\mathbf{u}} \equiv \mathbf{0}$, or is moving at all times where defined, i.e., $\dot{\mathbf{u}} \neq \mathbf{0}$ everywhere. In other

† As always, we extend all solutions to their maximal interval of existence.

words, it is mathematically impossible for a solution to reach an equilibrium position in a finite amount of time — although it may well approach equilibrium in an asymptotic fashion as $t \to \infty$. Physically, this result has the interesting and counterintuitive consequence that a system never actually attains an equilibrium position! Even at very large times, there is always some very slight residual motion. In practice, though, once the solution gets sufficiently close to equilibrium, we are unable to detect the motion, and the physical system has, for all practical purposes, reached its stationary equilibrium configuration. And, of course, the inherent motion of the atoms and molecules not included in the simplified model would hide the infinitesimal residual effects of the original solution.

Proposition 20.11. *If $\mathbf{u}(t)$ is any solution to an autonomous ordinary differential equation such that $\mathbf{u}(t_\star) = \mathbf{u}^\star$ at some time $t_\star$, then $\mathbf{u}(t) \equiv \mathbf{u}^\star$ is the equilibrium solution.*

Proof: We regard $\mathbf{u}(t_\star) = \mathbf{u}^\star$ as initial data for the given solution $\mathbf{u}(t)$ at the initial time $t_\star$. Since $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$, the constant function $\mathbf{u}^\star(t) \equiv \mathbf{u}^\star$ is a solution of the differential equation that satisfies the same initial conditions. Therefore, by our blanket uniqueness hypothesis, the solution in question has to agree with it. *Q.E.D.*

Without uniqueness, the result is false. For example, the function $u(t) = (t - t_\star)^{5/3}$ is a solution to the scalar ordinary differential equation (20.33) that reaches equilibrium, $u^\star = 0$, in a finite time $t = t_\star$.

Although a solution cannot reach equilibrium in a finite time, it can certainly have a well-defined limiting value. It can be proved that such a limit point is necessarily an equilibrium solution. A proof of this result can be found in the above-mentioned references.

Theorem 20.12. *If $\mathbf{u}(t)$ is any solution to an autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ such that $\lim_{t \to \infty} \mathbf{u}(t) = \mathbf{u}^\star$, then $\mathbf{u}^\star$ is an equilibrium solution.*

The same conclusion holds if we run time backwards: if $\lim_{t \to -\infty} \mathbf{u}(t) = \mathbf{u}^\star$, then $\mathbf{u}^\star$ is also an equilibrium point. Of course, limiting equilibrium points are but one of a variety of possible long term behaviors of solutions to ordinary differential equations. Solutions can also become unbounded, can approach periodic orbits, known as *limit cycles*, or even turn out to be completely chaotic, depending upon the nature of the system and the initial conditions.

Continuous Dependence

In a physical applications, it is rare, if not infeasible, to be able to prescribe the initial conditions exactly. Rather, experimental and physical errors will only allow us to say that the initial conditions are approximately equal to those in our mathematical model. Thus, we need to be sure that a small error in our initial measurements do not produce a large effect in the solution. A similar argument can be made for any physical parameters, e.g., masses, charges, frictional coefficients, etc., that appear in the differential equation itself. A slight change in the parameters should not have a dramatic effect on the solution.

Mathematically, what we are after is a criterion of *continuous dependence* of solutions upon both initial data and parameters. Fortunately, the desired result holds without any additional assumptions on the differential equation, other than requiring that the parameters appear continuously. We state both results in a single theorem.

Theorem 20.13. *Consider an initial value problem problem*

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}, \boldsymbol{\mu}), \qquad \mathbf{u}(t_0) = \mathbf{a}(\boldsymbol{\mu}), \tag{20.36}$$

in which the differential equation and/or the initial conditions depend continuously upon one or more parameters $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$. *Then the unique*[†] *solution* $\mathbf{u}(t, \boldsymbol{\mu})$ *depends continuously upon the parameters.*

Example 20.14. Let us look at a perturbed version of the initial value problem

$$\frac{du}{dt} = \alpha\, u^2, \qquad u(0) = u_0 + \mu,$$

that we considered in Example 20.1. We regard μ as a small perturbation of our original initial data u_0, and α as a variable parameter in the equation. The solution is

$$u(t, \mu) = \frac{u_0 + \mu}{1 - \alpha\,(u_0 + \mu)\, t}\,.$$

Note that, where defined, this is a continuous function of both parameters μ, α. Thus, a small change in the initial data, or in the equation, produces a small change in the solution — at least for times near the initial time.

Continuous dependence does not preclude nearby solutions from eventually becoming far apart. Indeed, the blow-up time $t_\star = 1/\big[\,\alpha\,(u_0 + \mu)\,\big]$ for a solution depends upon both the initial data and the parameter in the equation. Thus, as we approach blow up, solutions that started out very close to each other will get arbitrarily far apart; see Figure 20.1 for an illustration.

In light of this example, the continuous dependence of solutions upon parameters does *not* prevent solutions to the ordinary differential equation from being chaotic and/or having "sensitive dependence" on initial conditions. A very tiny change in the initial conditions has a negligible short term effect upon the solution, but over longer time intervals the differences between the two solutions can be dramatic. Further development of these ideas can be found in [**7**, **44**] and elsewhere.

20.3. Stability.

Once a solution to a system of ordinary differential equations has settled down, its limiting value is an equilibrium solution. However, not all equilibria appear in this way. The only steady state solutions that one directly observes in a physical system are the stable equilibria. Unstable equilibria are unsustainable in any realistic situation, and will disappear when subjected to even the tiniest perturbation, e.g., a breath of air, or outside traffic jarring the experimental apparatus. Thus, finding the equilibrium solutions to a system of ordinary differential equations is only half the battle; one must then understand their stability properties in order to characterize those that can be realized in normal

[†] We continue to impose our blanket uniqueness hypothesis.

physical circumstances. We shall exclusively work with autonomous systems $\dfrac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u})$ in our presentation. We assume throughout that the right hand side $\mathbf{F}(\mathbf{u})$ is continuously differentiable, so as to ensure the uniqueness of solutions to the initial value problem.

As we noted in Theorem 20.12, if a solution tends to a single point as $t \to \infty$, then that point must be an equilibrium solution. If *every* solution that starts out near a given equilibrium solution tends to it, the equilibrium is called *asymptotically stable*. If the solutions that start out nearby stay nearby, then the equilibrium is *stable*. More formally:

Definition 20.15. An equilibrium solution $\mathbf{u}^\star$ to an autonomous system of first order ordinary differential equations is called

(i) *stable* if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that if $\| \mathbf{u}_0 - \mathbf{u}^\star \| < \delta$, then the solution $\mathbf{u}(t)$ with initial conditions $\mathbf{u}(0) = \mathbf{u}_0$ satisfies $\| \mathbf{u}(t) - \mathbf{u}^\star \| < \varepsilon$ for all $t \geq t_0$.

(ii) *asymptotically stable* if $\mathbf{u}^\star$ is stable and, in addition, there exists $\delta_0 > 0$ such that if $\| \mathbf{u}_0 - \mathbf{u}^\star \| < \delta_0$, then $\mathbf{u}(t) \to \mathbf{u}^\star$ as $t \to \infty$.

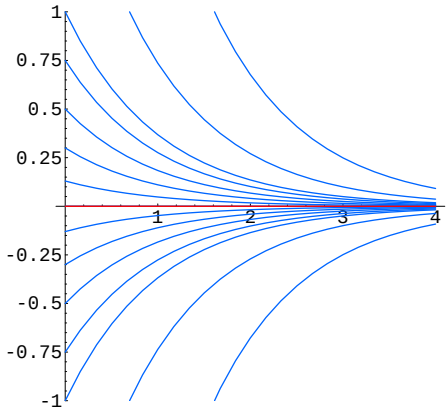
Example 20.16. As we saw, the logistic differential equation (20.10) has two equilibrium solutions, corresponding to the two solutions to the equation $\lambda\, u(1-u) = 0$. The first equilibrium solution $u_1^\star = 0$ is unstable, since all nearby solutions go away from it at an exponentially fast rate. On the other hand, the other equilibrium solution $u_2^\star = 1$ is asymptotically stable, since any solution with initial condition $0 < u_0$ tends to it at an exponentially fast rate. The solution graphs in Figure 20.1 illustrate the behavior of the solutions.

Example 20.17. Consider an autonomous (meaning constant coefficient) homogeneous linear planar system

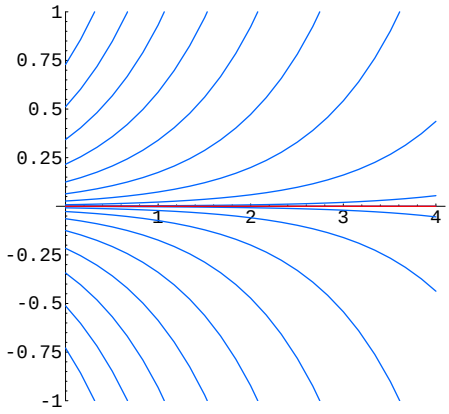
$$\frac{du}{dt} = a\,u + b\,v, \qquad \frac{dv}{dt} = c\,u + d\,v,$$

with coefficient matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The origin $u^\star = v^\star = 0$ is an evident equilibrium, solution, and, moreover, is the only equilibrium provided $\ker A = \{\mathbf{0}\}$. According to the results in Section 9.3, the stability of the origin equilibrium depends upon the eigenvalues of the coefficient matrix A. The origin is (globally) asymptotically stable if and only if both eigenvalues are real and negative. The origin is stable, but not asymptotically stable if and only if both eigenvalues are purely imaginary, or if 0 is a double eigenvalue and so $A = \mathrm{O}$. In all other cases, the origin is unstable. Below we will see how this simple linear analysis has direct bearing on the stability problem for nonlinear planar systems.

In Chapter 9, we established general criteria, based on the eigenvalues of the coefficient matrix, that guarantee the stability of equilibrium solutions to constant coefficient (i.e., autonomous) linear systems of ordinary differential equations. As we shall see, except in borderline situations, the same stability criteria carry over to equilibrium solutions of nonlinear ordinary differential equations. In analogy with the discrete case discussed in Section 19.1, we approximate the nonlinear system near an equilibrium point by its linearization.



Stable Equilibrium



Unstable Equilibrium

Figure 20.5.

Stability of Scalar Differential Equations

The stability analysis for scalar ordinary differential equations

$$\frac{du}{dt} = F(u) \tag{20.37}$$

is particularly easy. We continue to impose our blanket hypothesis, ensuring uniqueness, that F is continuously differentiable, or, at the very least, is a Lipschitz function of u.

The first observation is that all non-equilibrium solutions $u(t)$ are strictly monotone functions, meaning they are either always increasing or always decreasing. Indeed, when $F(u) > 0$, then, according to the equation, the derivative $\dot{u} > 0$, and hence $u(t)$ is increasing at such a point. Vice versa, solutions are decreasing at any point where $F(u) < 0$. By continuity, any non-monotone solution would have to pass through an equilibrium value where $F(u^\star) = 0$, which would be in violation of Proposition 20.11. This proves the claim.

As a consequence of monotonicity, there are only three possible behaviors for a non-equilibrium solution: $u(t)$ either

(*a*) becomes unbounded at some finite time: $u(t) \to \infty$ or $-\infty$ as $t \to t^\star$, or

(*b*) exists for all $t \geq t_0$, but become unbounded as $t \to \infty$, or

(*c*) exists for all $t \geq t_0$ and has a limiting value, $u(t) \to u^\star$ as $t \to \infty$, which, by Theorem 20.12 must be an equilibrium point.

Let us look more carefully at the last eventuality. Let $u^\star$ be an equilibrium, so $F(u^\star) = 0$. Suppose that $F(u) > 0$ for all u lying slightly below $u^\star$. Any solution that starts out below, but sufficiently close to $u^\star$ must be increasing. Moreover, $u(t) < u^\star$ for all t since the solution cannot pass through the equilibrium point. Therefore, $u(t)$ must be a solution of type (*c*). It must have limiting value $u^\star$, since by assumption, this is the only equilibrium solution it can increase to. Therefore, in this situation, the equilibrium point $u^\star$ is *asymptotically stable from below*; solutions that start out slightly below return to it in the limit. On the other hand, if $F(u) < 0$ for all u slightly below $u^\star$, then any solution that start out in this regime will be monotonically decreasing, and so move away from the equilibrium point, which is thus *unstable from below*.

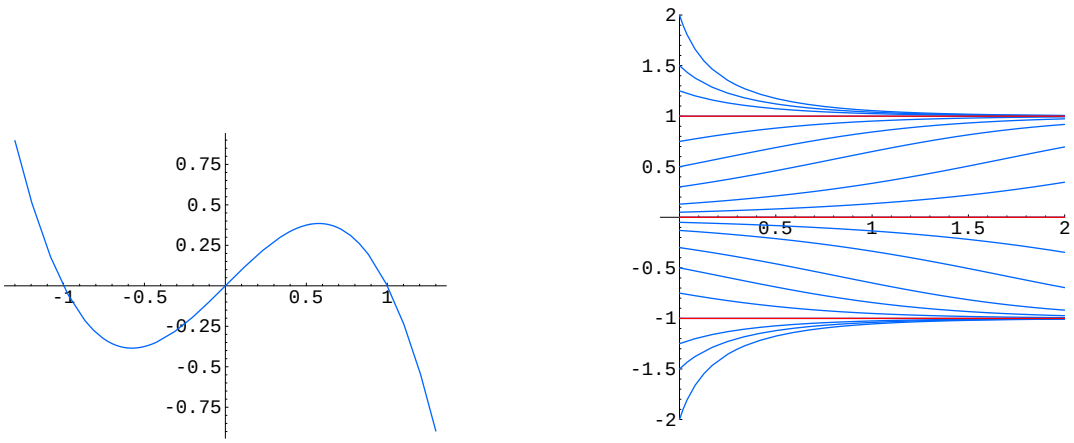


Figure 20.6. Stability of $\dot{u} = u - u^3$.

By the same reasoning, if $F(u) < 0$ for u slightly above $u^\star$, then such solutions will be monotonically decreasing, bounded from below by $u^\star$, and hence have no choice but to tend to $u^\star$ in the limit. Under this condition, the equilibrium point is *asymptotically stable from above*. The reverse inequality $F(u) > 0$ corresponds to solutions that increase away from $u^\star$, which is then *unstable from above*. Combining the two stable cases produces the basic condition for asymptotic stability of scalar ordinary differential equations.

Theorem 20.18. *A equilibrium point $u^\star$ of an autonomous scalar differential equation is asymptotically stable if and only if $F(u) > 0$ for $u^\star - \delta < u < u^\star$ and $F(u) < 0$ for $u^\star < u < u^\star + \delta$, for some $\delta > 0$.*

In other words, if $F(u)$ switches sign from positive to negative as one goes through the equilibrium point from left to right, then the equilibrium is stable. If the inequalities are reversed, and $F(u)$ goes from negative to positive, then the equilibrium point is unstable An equilibrium point where $F(u)$ is of one sign on both sides, e.g., the point $u^\star = 0$ for $F(u) = u^2$, is stable from one side, and unstable from the other; see Exercise ■ for details.

Example 20.19. Consider the differential equation

$$\frac{du}{dt} = u - u^3. \tag{20.38}$$

Solving the algebraic equation $F(u) = u - u^3 = 0$, we find that the equation has three equilibria: $u_1^\star = -1$, $u_2^\star = 0$, $u_3^\star = +1$, The graph of the function $F(u) = u - u^3$ switches from positive to negative at the first equilibrium point, which proves its stability. Similarly, the graph goes back from positive to negative at $u_2^\star = 0$, proving the instability of the second equilibrium. Finally, the third equilibrium is stable because $F(u)$ changes from negative to positive there.

With this information coupled with monotonicity, we can completely characterize the behavior of all solutions to the system. Any solution with negative initial condition $u_0 < 0$ will end up, asymptotically, at the first equilibrium, $u(t) \to -1$ as $t \to \infty$. Indeed, if $u_0 < -1$, then $u(t)$ is monotonically increasing to -1, while if $-1 < u_0 < 0$, the solution is decreasing towards -1. On the other hand, if $u_0 > 0$ the solution ends up at the last

equilibrium, $u(t) \to +1$; those with $0 < u_0 < 1$ are monotonically increasing, while those with $1 < u_0$ are decreasing. The only solution that does not end up at either -1 or $+1$ as $t \to \infty$ is the unstable equilibrium solution $u(t) \equiv 0$. The solutions are plotted in Figure 20.6; note the curves converge to the stable solutions ± 1 and diverge from the unstable solution 0 as $t \to \infty$.

Thus, the sign of the function $F(u)$ nearby an equilibrium determines its stability. In most instances, this can be checked by looking at the derivative of the function at the equilibrium. If $F'(u^\star) < 0$, then we are in the stable situation, whereas if $F'(u^\star) > 0$, then we are unstable on both sides.

Theorem 20.20. *Let $u^\star$ be a equilibrium point for a scalar ordinary differential equation $\dot{u} = F(u)$. If $F'(u^\star) < 0$, then $u^\star$ is asymptotically stable. If $F'(u^\star) > 0$, then $u^\star$ is unstable.*

Thus, in the preceding example, $F'(u) = 1 - 3u^2$, and we compute its value at the equilibria: $F'(-1) = -2 < 0$, $F'(0) = 1 > 0$, $F'(1) = -2 < 0$. The signs reconfirm our conclusion that ± 1 are stable equilibria, while 0 is unstable.

Theorem 20.20 is not quite as powerful as the direct test in Theorem 20.18, but does have the advantage of being a bit easier to use, and, more significantly, generalizing to systems of ordinary differential equations. In the borderline case when $F'(u^\star) = 0$, the derivative test is inconclusive, and requires further analysis to resolve the stability of the equilibrium in question. For example, the equations $du/dt = u^3$ and $du/dt = -u^3$ both satisfy $F'(0) = 0$ at the equilibrium point $u^\star = 0$. But, using the criterion of Theorem 20.18, we conclude that the former has an unstable equilibrium, while the latter is stable.

Linearization and Stability

In higher dimensional situations, we can no longer rely on monotonicity properties of solutions, and a more sophisticated approach to the stability of equilibrium solutions is required. The key idea is already contained in the second characterization of stable equilibria in Theorem 20.20. The derivative $F'(u)$ determines the slope of its tangent line, which is the linear approximation to the function $F(u)$ at the equilibrium point. In Chapter 9, we established the basic stability criteria for the linearized differential equation. In most situations, linear stability or instability carries over to the corresponding nonlinear system.

Let us revisit the scalar case $\dot{u} = F(u)$ from this point of view. *Linearization* of a scalar function at a point means to replace it by its tangent line approximation

$$F(u) \;\approx\; F(u^\star) + F'(u^\star)(u - u^\star) \tag{20.39}$$

If $u^\star$ is an equilibrium point, then $F(u^\star) = 0$ and the first term disappears. Therefore, we expect that, near the equilibrium point, the solutions to the nonlinear ordinary differential equation (20.37) should be well approximated by its linearization

$$\frac{du}{dt} = F'(u^\star)(u - u^\star).$$

Let us rewrite this equation in terms of the deviation $v(t) = u(t) - u^\star$ of the solution from equilibrium. Since $u^\star$ is fixed, $dv/dt = du/dt$, and so the linearized equation takes the elementary form

$$\frac{dv}{dt} = a\, v, \qquad \text{where} \qquad a = F'(u^\star) \tag{20.40}$$

is the value of the derivative at the equilibrium point. Note that the original equilibrium point $u = u^\star$ corresponds to the zero equilibrium point $v = v^\star = 0$ of the linearized equation. We already know that the linear differential equation (20.40) has an asymptotically stable equilibrium at $v^\star = 0$ if and only if $a = F'(u^\star) < 0$, while for $a = F'(u^\star) > 0$ the origin is unstable. In this manner, the linearized stability criteria reproduce those we found in Theorem 20.20.

The same linearization technique can be applied to analyze the stability of an equilibrium solution $\mathbf{u}^\star$ to a first order autonomous system $\dfrac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u})$. We approximate the function $\mathbf{F}(\mathbf{u})$ at an equilibrium point where $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$ by its first order Taylor expansion

$$\mathbf{F}(\mathbf{u}) \;\approx\; \mathbf{F}(\mathbf{u}^\star) + \mathbf{F}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) \;=\; \mathbf{F}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star), \tag{20.41}$$

where $\mathbf{F}'(\mathbf{u}^\star)$ denotes its $n \times n$ Jacobian matrix (19.27) at the equilibrium point. Thus, for nearby solutions, we expect that the deviation from equilibrium $\mathbf{v}(t) = \mathbf{u}(t) - \mathbf{u}^\star$ will be governed by the linearization

$$\frac{d\mathbf{v}}{dt} = A\,\mathbf{v}, \qquad \text{where} \qquad A = \mathbf{F}'(\mathbf{u}^\star). \tag{20.42}$$

According to Theorem 9.17, the linearized system (20.42) has an asymptotically stable zero solution if and only if all the eigenvalues of the coefficient matrix $A = \mathbf{F}'(\mathbf{u}^\star)$ have negative real part. On the other hand, if one or more of the eigenvalues has positive real part, then the zero solution is unstable. It can be proved, [**72**, **80**], that these conditions are also sufficient for asymptotic stability and instability in the nonlinear case.

Theorem 20.21. *Let $\mathbf{u}^\star$ be an equilibrium point for the first order ordinary differential equation $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If all of the eigenvalues of the Jacobian matrix $\mathbf{F}'(\mathbf{u}^\star)$ have negative real part,* $\operatorname{Re} \lambda_j < 0$, *then $\mathbf{u}^\star$ is asymptotically stable. If, on the other hand, $\mathbf{F}'(\mathbf{u}^\star)$ has one or more eigenvalues with positive real part, then $\mathbf{u}^\star$ is an unstable equilibrium.*

The borderline case occurs when one or more of the eigenvalues is either 0 or purely imaginary, i.e., $\operatorname{Re} \lambda_j = 0$, while all other eigenvalues have negative real part. In these cases, the linearization test is inconclusive, and we need more detailed information (which may not be easy to come by) on how the nonlinear terms might affect any borderline eigenvalues lying on the imaginary axis. Their effect may be to nudge the eigenvalue into the left half plane, stabilizing the solutions, or into the right half plane, destabilizing them.

Example 20.22. The second order ordinary differential equation

$$m\, \frac{d^2\theta}{dt^2} + \mu\, \frac{d\theta}{dt} + \kappa \sin\theta = 0 \tag{20.43}$$

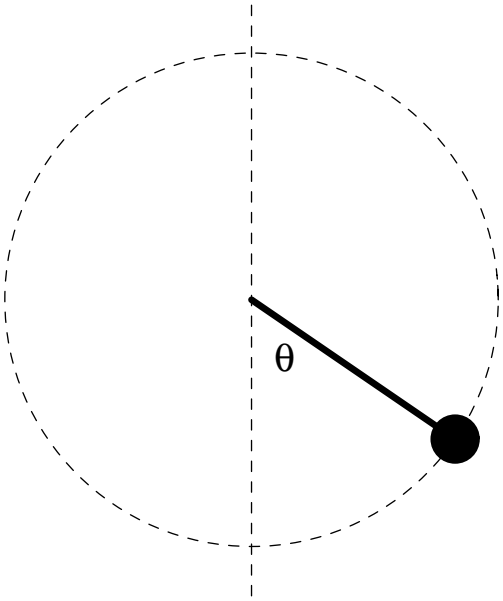


Figure 20.7. The Pendulum.

describes the damped oscillations of a rigid pendulum that rotates on a pivot under a gravitational force. The unknown function $\theta(t)$ measures the angle of the pendulum from the vertical, as illustrated in Figure 20.7. The constant $m > 0$ is the mass of the pendulum bob, $\mu > 0$ the coefficient of friction, assumed here to be strictly positive, and $\kappa > 0$ the restoring gravitational force.

In order to study the equilibrium solutions and their stability, we begin by converting this equation into a first order system by setting $u(t) = \theta(t)$, $v(t) = \dfrac{d\theta}{dt}$, and so

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\alpha \sin u - \beta v, \qquad \text{where} \qquad \alpha = \frac{\kappa}{m}, \qquad \beta = \frac{\mu}{m}, \tag{20.44}$$

are both positive constants. The equilibria occur where the right hand sides of the first order system (20.44) simultaneously vanish:

$$v = 0, \qquad -\alpha \sin u - \beta v = 0, \qquad \text{and hence} \qquad u = 0, \ \pm\pi, \ \pm 2\pi, \ \dots \ .$$

Thus, the system has infinitely many equilibrium points $\mathbf{u}_k^\star = (k\pi, 0)$ for $k = 0, \pm 1, \pm 2, \dots$.

The equilibrium point $\mathbf{u}_0^\star = (0, 0)$ corresponds to $\theta = 0$, $\dot\theta = 0$, which means that the pendulum is at rest at the bottom of its arc. Our physical intuition leads us to expect this to describe a stable configuration, as the frictional effects will eventually damp out small motions of the pendulum. The next equilibrium $\mathbf{u}_1^\star = (\pi, 0)$ corresponds to $\theta = \pi$, $\dot\theta = 0$, which means that the pendulum stays motionless at the top of its arc. Theoretically, this is a possible equilibrium configuration, but highly unlikely to be observed in practice and thus should be unstable. Now, since $u = \theta$ is an angular variable, equilibria whose u values differ by an integer multiple of 2π define the same physical configuration, and hence should have identical stability properties. Therefore, the remaining equilibria $\mathbf{u}_k^\star$ physically

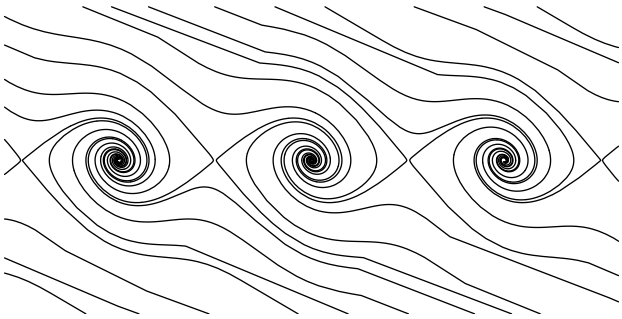


Figure 20.8. The Underdamped Pendulum.

correspond to one or the other of these two possible equilibrium positions; when $k = 2\,j$ is even, the pendulum is at the bottom, while when $k = 2\,j + 1$ is odd, the pendulum is at the top.

Let us now confirm our intuition using the linearization stability criterion of Theorem 20.21. The right hand side of the system, namely

$$\mathbf{F}(u,v) = \begin{pmatrix} v \\ -\alpha \sin u - \beta\, v \end{pmatrix}, \qquad \text{has Jacobian matrix} \qquad \mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ -\alpha \cos u & -\beta \end{pmatrix}.$$

At the bottom equilibrium $\mathbf{u}_0^\star = (0,0)$, the Jacobian matrix

$$\mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ -\alpha & -\beta \end{pmatrix} \qquad \text{has eigenvalues} \qquad \lambda = \frac{-\beta \pm \sqrt{\beta^2 - 4\,\alpha}}{2}\,.$$

Under our assumption that $\alpha, \beta > 0$, the eigenvalues both have negative real part, and hence the origin is a stable equilibrium. If $\beta^2 < 4\,\alpha$ — the *underdamped* case — the eigenvalues are complex, and hence, in the terminology of Section 9.3, the origin is a *stable focus*. In the phase plane, the solutions spiral in to the focus, which corresponds to a pendulum with damped oscillations of decreasing magnitude. On the other hand, if $\beta^2 > 4\,\alpha$, then the system is *overdamped*. Both eigenvalues are negative, and the origin is a *stable node*. In this case, the solutions decay exponentially fast. Physically, this would be like a pendulum moving in a vat of molasses. In both cases, the phase portrait of the nonlinear motion near the equilibrium position closely matches the linearized problem. The same analysis holds at all even multiples of π — which really represent the same bottom equilibrium point.

On the other hand, at the top equilibrium $\mathbf{u}_1^\star = (\pi, 0)$, the Jacobian matrix

$$\mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ \alpha & -\beta \end{pmatrix} \qquad \text{has eigenvalues} \qquad \lambda = \frac{-\beta \pm \sqrt{\beta^2 + 4\,\alpha}}{2}\,.$$

In this case, one of the eigenvalues is real and positive while the other is negative. The linearized system has an unstable saddle point, and hence the nonlinear system is also unstable at this equilibrium point. Any tiny perturbation of an upright pendulum will dislodge it, sending it into an oscillatory motion which eventually ends up at the stable bottom equilibrium.

The complete phase portrait of an underdamped pendulum appears in Figure 20.8. Note that, as advertised, almost all solutions end up spiraling into the stable equilibria.

Solutions with a large initial velocity will spin several times around the center, but eventually the cumulative effects of frictional forces win out and the pendulum ends up in a damped oscillatory mode. The unstable equilibria have the same basic saddle shape as their linear counterparts. Each gives rise to two special solutions in which the pendulum spins around a few times, and, in the $t \to \infty$ limit, ends up upright at the unstable equilibrium position. However, this solution is practically impossible to achieve in a physical environment as any tiny perturbation — e.g., a breath of air — will cause the pendulum to sightly deviate and then end up decaying into the usual damped oscillatory motion at the bottom.

A deeper analysis shows that equilibria whose eigenvalues do not lie on the imaginary axis, so $\operatorname{Re} \lambda_j \neq 0$ for all j, are *structurally stable*. This means that not only are the stability properties dictated by the linearized approximations, but, nearby the equilibrium point, solutions to the nonlinear system are slight perturbations of those of the corresponding linearized system. For instance, stable foci of the linearized system correspond to stable foci of the nonlinear counterpart, while unstable saddle points remain saddle points, although the saddle rays are slightly curved as they depart from the equilibrium. In other words, the structural stability of linear systems, as discussed at the end of Section 9.3 also carries over to the nonlinear regime near an equilibrium. This result is known as the *Center Manifold Theorem*, and a complete statement and proof can be found, for instance, in [**72**, **80**].

Example 20.23. Consider the unforced van der Pol system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -(u^2 - 1)\, v - u.$$

that we derived in Example 20.5. The only equilibrium point is at the origin $u = v = 0$. Computing the Jacobian matrix of the right hand side,

$$\mathbf{F}'(u, v) = \begin{pmatrix} 0 & 1 \\ 2\,u\,v - 1 & 1 \end{pmatrix}, \qquad \text{hence} \qquad \mathbf{F}'(0, 0) = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

The eigenvalues of $\mathbf{F}'(0, 0)$ are $\frac{1}{2}(1 \pm \mathrm{i}\sqrt{3}\,)$, and correspond to an unstable focus of the linearized system near the equilibrium point. Therefore, the origin is an unstable equilibrium for nonlinear van der Pol system. Solutions starting out near $\mathbf{0}$ spiral away.

On the other hand, it can be shown that solutions that are sufficiently far away spiral in towards the center. So what happens to the solutions? As illustrated in the phase plane portrait Figure 20.9, all of the solutions spiral towards a *stable* periodic orbit, known as a *limit cycle* for the system. Any initial data will eventually end up following the periodic orbit as it circles around the origin. Proof of the existence of a limit cycle relies on the more sophisticated *Poincaré–Bendixson Theory* for planar autonomous systems, which can be found in [**72**].

Example 20.24. The nonlinear system

$$\frac{du}{dt} = u\,(v - 1), \qquad \frac{dv}{dt} = 4 - u^2 - v^2,$$

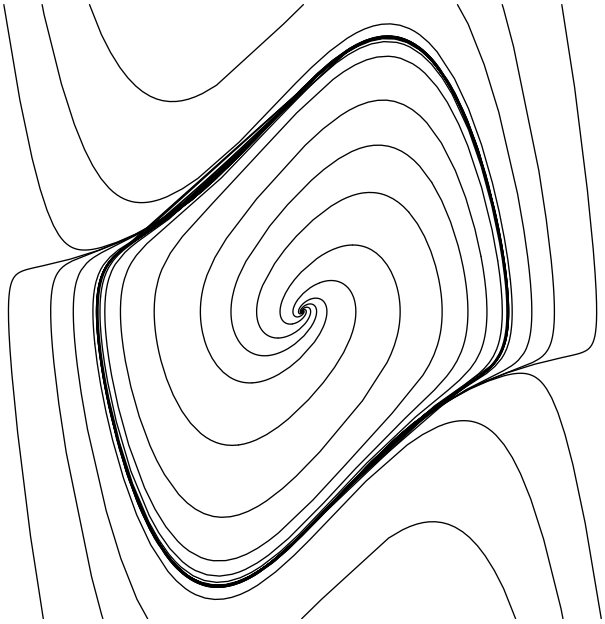


Figure 20.9. Phase Portrait of the van der Pol System.

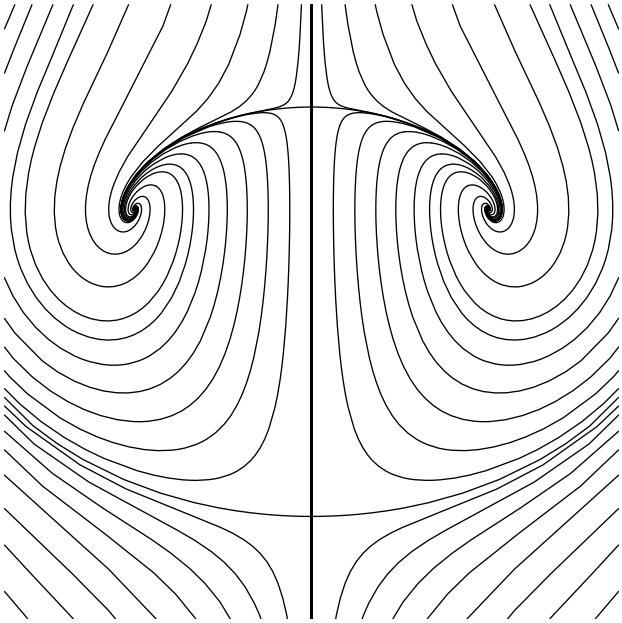


Figure 20.10. Phase Portrait for $\dot{u} = u\,(v-1),\ \dot{v} = 4 - u^2 - v^2$..

has four equilibria: $(0, \pm 2)$ and $(\pm\sqrt{3}\,, 1)$. The Jacobian matrix for the system is

$$\mathbf{F}'(u,v) = \begin{pmatrix} v-1 & u \\ -2\,u & -2\,v \end{pmatrix}.$$

A table of the eigenvalues at the equilibrium points and their stability follows: A complete phase portrait can be found in Figure 20.10

Equilibrium Point	Jacobian matrix	Eigenvalues	Stability
$(0,2)$	$\begin{pmatrix} 1 & 0 \\ 0 & -4 \end{pmatrix}$	$1,\ -4$	unstable saddle
$(0,-2)$	$\begin{pmatrix} -3 & 0 \\ 0 & 6 \end{pmatrix}$	$-3, 6$	unstable saddle
$(\sqrt{3}\,,1)$	$\begin{pmatrix} 0 & -\sqrt{3} \\ 2\sqrt{3} & -2 \end{pmatrix}$	$-1 \pm \mathrm{i}\sqrt{5}$	stable focus
$(-\sqrt{3}\,,1)$	$\begin{pmatrix} 0 & -\sqrt{3} \\ 2\sqrt{3} & -2 \end{pmatrix}$	$-1 \pm \mathrm{i}\sqrt{5}$	stable focus

Conservative Systems

When modeling a physical system that includes some form of damping — due to friction, viscosity, or dissipation — the linearization test for stability of equilibria will usually suffice. However, when dealing with conservative systems, when there is no damping and so energy is preserved, the test is usually inconclusive, and one must rely on alternative stability criteria. In many instances, one can exploit the conservation of energy for this purpose. We return to our general philosophy that minimizers of an energy function should be stable (but not necessarily asymptotically stable) equilibria.

To say that the energy that is conserved means that it is constant on solutions. Such a quantity is known as a *first integral* or *conservation law* for the system of ordinary differential equations. Additional examples include conservation of mass and conservation of linear and angular momentum. Let us state the general definition.

Definition 20.25. A *first integral* of an autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ is a real-valued function $I(\mathbf{u})$ which is constant on solutions.

The constant value of the first integral will depend upon the solution, and is fixed by its value at the initial time t_0. In other words, a first integral must satisfy

$$I(\mathbf{u}(t)) = I(\mathbf{u}(t_0)) \tag{20.45}$$

whenever $\mathbf{u}(t)$ is a solution to the differential equation. Or, to rephrase this condition in another, equivalent manner, every solution to the dynamical system is constrained to move along a single level set $\{\, I(\mathbf{u}) = c \,\}$ of the first integral I. Any constant function, $I(\mathbf{u}) \equiv c$, is trivially a first integral, but it carries no information whatsoever about the solutions, and so is uninteresting. We will call any autonomous system that possesses a nontrivial first integral $I(\mathbf{u})$ a *conservative system*.

How do we find first integrals? In applications, one often appeals to physical principles such as conservation of energy, momentum, or mass. Mathematically, the most convenient

way to check whether a function is constant is to verify that its derivative is identically zero. Thus, differentiating (20.45) with respect to t and making use of the chain rule leads to the basic condition

$$0 = \frac{d}{dt}\, I(\mathbf{u}(t)) = \nabla I(\mathbf{u}(t)) \cdot \frac{d\mathbf{u}}{dt} = \nabla I(\mathbf{u}(t)) \cdot \mathbf{F}(\mathbf{u}(t)). \tag{20.46}$$

The final expression is the directional derivative of $I(\mathbf{u})$ with respect to the vector field $\mathbf{v} = \mathbf{F}(\mathbf{u})$ that specifies the differential equation, cf. (19.62). Writing out (20.46) in detail, we find that a first integral $I(u_1, \dots, u_n)$ must satisfy a first order linear partial differential equation

$$F_1(u_1, \dots, u_n)\, \frac{\partial I}{\partial u_1} + \, \cdots \, + F_n(u_1, \dots, u_n)\, \frac{\partial I}{\partial u_n} = 0. \tag{20.47}$$

As such, it looks harder to solve[†] than the original ordinary differential equation! Usually, one is forced to rely on either physical intuition, intelligent guesswork, or, as a last resort, luck to find first integrals. A deeper fact, due to the pioneering twentieth century mathematician Emmy Noether, cf. [**113**, **117**], is that first integrals and conservation laws are the result of underlying symmetry properties of the differential equation. Like many nonlinear methods, it research.

Let us specialize to the case of a planar autonomous system

$$\frac{du}{dt} = F(u, v), \qquad \frac{dv}{dt} = G(u, v). \tag{20.48}$$

According to (20.47), a first integral $I(u, v)$ of this system must satisfy the linear partial differential equation

$$F(u, v)\, \frac{\partial I}{\partial u} \; + \; G(u, v)\, \frac{\partial I}{\partial v} = 0. \tag{20.49}$$

This nonlinear first order partial differential equation can be solved as follows[‡]. Consider the auxiliary first order scalar ordinary differential equation[§]

$$\frac{dv}{du} = \frac{G(u, v)}{F(u, v)} \tag{20.50}$$

for $v = h(u)$. Note that (20.50) can be formally obtained by dividing the second equation in the original system (20.48) by the first, and then canceling the time differentials dt. Suppose we can write the general solution to the scalar equation (20.50) in the implicit form

$$I(u, v) = c, \tag{20.51}$$

[†] In fact, the general solution method of such partial differential equations, [**117**], relies on the integration of ordinary differential equations. But then we are back to where we started!

[‡] See Section 22.1 for an alternative perspective on such partial differential equations.

[§] We assume that $F(u, v) \neq 0$. Otherwise, $I(u) = u$ is itself a first integral, and the system reduces to a scalar equation for v; see Exercise ■.

where c is a constant of integration. We claim that the function $I(u,v)$ is a first integral of the original system (20.48). Indeed, differentiating (20.51) with respect to u and using the chain rule, we find

$$0 = \frac{d}{du}\, I(u,v) = \frac{\partial I}{\partial u} + \frac{dv}{du}\,\frac{\partial I}{\partial v} = \frac{\partial I}{\partial u} + \frac{G(u,v)}{F(u,v)}\,\frac{\partial I}{\partial v}\ .$$

Clearing the denominator, we conclude that $I(u,v)$ solves the partial differential equation (20.49), which justifies our claim.

Example 20.26. As an elementary example, consider the linear equation

$$\frac{du}{dt} = -\,v, \qquad \frac{dv}{dt} = u. \tag{20.52}$$

To construct a first integral, we construct auxiliary equation (20.50), which is

$$\frac{dv}{du} = -\,\frac{u}{v}\,.$$

This first order ordinary differential equation can be solved by separating variables:

$$v\,dv = -\,u\,du, \qquad \text{and hence} \qquad \tfrac{1}{2}\,u^2 + \tfrac{1}{2}\,v^2 = c,$$

where c is the constant of integration. Therefore, by the preceding result,

$$I(u,v) = \tfrac{1}{2}\,u^2 + \tfrac{1}{2}\,v^2$$

is a first integral. The level sets of $I(u,v)$ are the concentric circles centered at the origin, and we recover the fact that the solutions of (20.52) go around the circles. The origin is a stable equilibrium — a center.

This simple example hints at the importance of first integrals in stability theory. The following key result confirms our general philosophy that energy minimizers, or, more generally, minimizers of first integrals, are necessarily stable equilibria.

Theorem 20.27. *Let $I(\mathbf{u})$ be a first integral for the autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If $\mathbf{u}^\star$ is a strict local minimum of I, then $\mathbf{u}^\star$ is a stable equilibrium point for the system.*

Proof: We first prove that $\mathbf{u}^\star$ is an equilibrium. Indeed, the solution $\mathbf{u}(t)$ with initial condition $\mathbf{u}(t_0) = \mathbf{u}^\star$ must maintain the value of $I(\mathbf{u}(t)) = I(\mathbf{u}^\star)$. But, by definition of a strict local minimum, there are no points near $\mathbf{u}^\star$ that have the same value of I, and hence, by continuity, the solution cannot leave the point $\mathbf{u}^\star$.

To prove stability, we set

$$M(r) = \max\{\, I(\mathbf{u}) \mid \|\,\mathbf{u} - \mathbf{u}^\star\,\| \le r\,\}\,, \qquad m(r) = \min\{\, I(\mathbf{u}) \mid \|\,\mathbf{u} - \mathbf{u}^\star\,\| = r\,\}\,.$$

Thus $M(r)$ is the maximum value of the integral over a ball[†] of radius r centered at the minimum, while $m(r)$ is the minimum over its boundary sphere of radius r. Since I is

[†] We write as if the norm is the Euclidean norm, but any other norm will work equally well for this proof.

continuous, so are m and M. Since $\mathbf{u}^\star$ is a local minimum, $M(r) \geq m(r) > I(\mathbf{u}^\star)$ for any $0 < r < \varepsilon$ sufficiently small.

For each such $\varepsilon > 0$, we can choose a $\delta > 0$ such that $M(\delta) < m(\varepsilon)$. Then, if $\mathbf{u}(t_0) = \mathbf{u}_0$ satisfies $\| \mathbf{u}_0 - \mathbf{u}^\star \| \leq \delta$, then $I(\mathbf{u}_0) \leq M(\delta)$. But $I(\mathbf{u}(t))$ is fixed, and so the resulting solution $\mathbf{u}(t)$ cannot cross the sphere of radius ε, since all points $\mathbf{v}$ on the sphere have a strictly larger value of $I(\mathbf{v}) \geq m(\varepsilon) > M(\delta) \geq I(\mathbf{u}_0)$. Therefore, $\| \mathbf{u}(t) - \mathbf{u}^\star \| < \varepsilon$, and hence we have fulfilled the stability criteria of Definition 20.15. *Q.E.D.*

Remark: If we reverse the inequalities, then the proof of Theorem 20.27 will also apply to strict local maximum of the first integral I, and so they are also stable equilibria. Or, to phrase it another way, maxima of $I(\mathbf{u})$ are minima of its negative $-I(\mathbf{u})$, which is also a first integral. Saddle points, however, are rarely stable. While at first sight, this may appear to contradict our intuition, the fact is that energy functions typically do not have maxima. Indeed, the energy is the sum of kinetic and potential contributions. While potential energy can admit maxima, e.g., the pendulum at the top of its arc, these are only saddle points for the full energy function, since the kinetic energy term can always be increased by moving a bit faster.

Example 20.28. Consider the specific predator-prey system

$$\frac{du}{dt} = 2\,u - u\,v, \qquad \frac{dv}{dt} = -9\,v + 3\,u\,v,$$

modeling populations of, say, lions and zebra, which is a special case of (20.21). According to Example 20.4, there are two possible equilibria:

$$u_1^\star = v_1^\star = 0, \qquad u_2^\star = 3, \quad v_2^\star = 2.$$

Let us try to determine their stability by the linearization criterion. The Jacobian matrix for the system is $\mathbf{F}'(u,v) = \begin{pmatrix} 2-v & -u \\ 3\,v & 3\,u - 9 \end{pmatrix}$. At the first, trivial equilibrium,

$$\mathbf{F}'(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -9 \end{pmatrix}, \qquad \text{with eigenvalues } \ 2 \ \text{ and } \ -9.$$

Since there is one positive and one negative eigenvalue, the origin is an unstable saddle point. On the other hand, at the nonzero equilibrium, the Jacobian matrix

$$\mathbf{F}'(3,2) = \begin{pmatrix} 0 & -3 \\ 6 & 0 \end{pmatrix}, \qquad \text{has purely imaginary eigenvalues} \qquad \pm 3\sqrt{2}\ \mathrm{i}\,.$$

Since they are purely imaginary, the linearized system has a stable center. But as we are in a borderline situation, Theorem 20.21 cannot be applied, and the linearization stability test is inconclusive.

It turns out that the predator-prey model has a first integral, and so represents a conservative system. In view of (20.50), we need to solve the auxiliary equation.

$$\frac{dv}{du} = \frac{-9\,v + 3\,u\,v}{2\,u - u\,v} = \frac{-9/u + 3}{2/v - 1}\,.$$

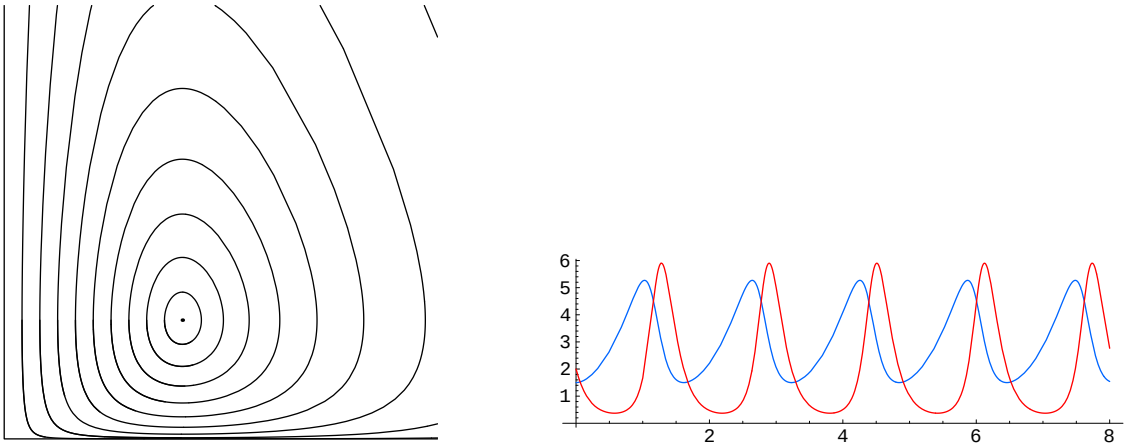


Figure 20.11. Phase Portrait and Graph of the Predator-Prey System.

Fortunately, this is a separable first order ordinary differential equation. Integrating,

$$2\,\log v - v = \int \left(\frac{2}{v} - 1\right) dv = \int \left(-\frac{9}{u} + 3\right) du = -9\log u + 3\,u + c,$$

where c is the constant of integration. Writing the solution in the form (20.51), we conclude that

$$I(u,v) = 9\,\log u - 3\,u + 2\,\log v - v = c,$$

is a first integral of the system. The solutions to the system must stay on the level sets of I. Note that

$$\nabla I(u,v) = \begin{pmatrix} 9/u - 3 \\ 2/v - 1 \end{pmatrix}, \qquad \text{and hence} \qquad \nabla I(3,2) = \mathbf{0},$$

which shows that the second equilibrium is a critical point. (The zero equilibrium is a singularity of I.) Moreover, the Hessian matrix at the critical point $\nabla^2 I(3,2) = \begin{pmatrix} -3 & 0 \\ 0 & -1 \end{pmatrix}$ is negative definite, and hence $\mathbf{u}_2^\star = (\,3,2\,)^T$ is a strict local maximum of the integral $I(u,v)$. Thus, Theorem 20.27 (rephrased for maxima) proves that the equilibrium point is a stable center.

The first integral serves to completely characterize the qualitative behavior of the system. In the physically relevant region, i.e., the upper right quadrant $Q = \{\,u > 0,\ v > 0\,\}$ where both populations are positive, all of the level sets of the first integral are closed curves encircling the equilibrium point $\mathbf{u}_2^\star = (\,3,2\,)^T$. The solutions move along these closed curves, and hence form a family of periodic solutions, illustrated in Figure 20.11. Thus, in such an idealized ecological model, for any initial conditions starting with some zebra and lions, i.e., $u(t_0), v(t_0) > 0$, the populations will maintain a balance over the long term, but each varies periodically between maximum and minimum values. Observe also that the maximum and minimum values of the two populations are not achieved simultaneously. Starting with a small number of predators, the number of prey will initially increase. The predators then have more food, and so also increase in numbers. At a certain

critical point, the predators are sufficiently numerous as to kill prey faster than they can reproduce. At this point, the prey population has reached its maximum, and begins to decline. But it takes a while for the predator population to feel the effect, and so it continues to increase. However, eventually the increasingly rapid decline in the number of prey begins to affect the predators, which subsequent achieve their maximum number. After this, both populations are in decline. Eventually, enough predators have died off so as to relieve the pressure on the prey, whose population bottoms out, and then slowly begins to rebound. Later, the number of predators also reaches a minimum, at which point the entire growth and decay cycle starts over again. Such periodic phenomena are observed, roughly, in many natural ecological systems.

The period of the population cycle depends upon how far away from the stable equilibrium it lies. Near equilibrium, the solutions are close to those of the linearized system which, in view of the eigenvalues, are periodic, with frequency $3\sqrt{2}$, and period $\sqrt{2}\,\pi/3$. However, solutions that are far away from equilibrium have much longer periods, and so greater imbalances between lions and zebras leads to longer periods, and more radically varying numbers of the two populations. Understanding the mechanisms behind these population cycles is becoming increasingly important in the proper management of natural resources, [**biol**].

Example 20.29. In our next example, we look at the undamped oscillations of a pendulum. When we set the friction coefficient $\mu = 0$, the nonlinear second order ordinary differential equation (20.43) reduces to

$$m\,\frac{d^2\theta}{dt^2} + \kappa \sin\theta = 0. \tag{20.53}$$

As before, we convert this into a first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\,\alpha\,\sin u, \qquad \text{where} \qquad u(t) = \theta(t), \quad v(t) = \frac{d\theta}{dt}\,, \quad \alpha = \frac{\kappa}{m}\,. \tag{20.54}$$

The equilibria, $\mathbf{u}^\star_k = (n\,\pi, 0)$ for $n = 0, \pm 1, \pm 2, \dots$, are the same as in the damped case, i.e., the pendulum is either at the top (n even) or the bottom (n odd) of the circle.

Let us try the linearized stability test. In this case, the Jacobian matrix of (20.54) is $\mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ -\,\alpha\,\cos u & 0 \end{pmatrix}$. At the top equilibria $\mathbf{u}^\star_{2\,k+1} = \big(\,(2\,k+1)\,\pi, 0\,\big)^T$,

$$\mathbf{F}'\big(\,(2\,k+1)\,\pi, 0\,\big) = \begin{pmatrix} 0 & 1 \\ \alpha & 0 \end{pmatrix} \qquad \text{has real eigenvalues} \qquad \pm\,\sqrt{\alpha}\,,$$

and hence these equilibria are unstable saddle points, just as in the damped version. On the other hand, at the bottom equilibria $\mathbf{u}^\star_{2\,k} = \big(\,2\,k\,\pi, 0\,\big)^T$, the Jacobian matrix

$$\mathbf{F}'(2\,k\,\pi, 0) = \begin{pmatrix} 0 & 1 \\ -\,\alpha & 0 \end{pmatrix}, \qquad \text{has purely imaginary eigenvalues} \qquad \pm\,\mathrm{i}\,\sqrt{\alpha}\,.$$

Without the benefit of damping, the linearized stability test is inconclusive, and the stability of the bottom equilibria remains in doubt.

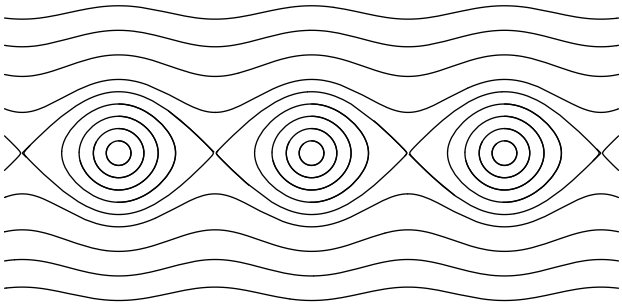


Figure 20.12. The Undamped Pendulum.

Since we are dealing with a conservative system, the total energy of the pendulum, namely

$$E(u,v) = \tfrac{1}{2} m v^2 + \kappa (1 - \cos u) = \frac{m}{2} \left(\frac{d\theta}{dt} \right)^2 + \kappa (1 - \cos \theta) \tag{20.55}$$

should provide us with a first integral. Note that E is a sum of two terms, which represent, respectively, the kinetic energy due to the motion, and the potential energy[†] due to the height of the pendulum bob. To verify that $E(u,v)$ is indeed a first integral, we compute

$$\frac{dE}{dt} = m v \, \frac{dv}{dt} + \kappa \sin u \, \frac{du}{dt} = - m v \, \alpha \sin u + \kappa v \sin u = 0, \qquad \text{since} \qquad \alpha = \frac{\kappa}{m} .$$

Therefore, E is indeed constant on solutions, reconfirming the physical basis of the model.

The phase plane solutions to the pendulum equation move along the level sets of the energy function $E(u,v)$, which are plotted in Figure 20.12. Its critical points are the equilibria; these are where

$$\nabla E(\mathbf{u}) = \begin{pmatrix} \kappa \sin u \\ m v \end{pmatrix} = \mathbf{0}, \qquad \text{and hence} \qquad \mathbf{u} = \mathbf{u}^\star_n = \begin{pmatrix} n \pi \\ 0 \end{pmatrix}$$

for some integer n. To characterize the critical points, we appeal to the second derivative test, and so evaluate the Hessian $\nabla^2 E(u,v) = \begin{pmatrix} \kappa \cos u & 0 \\ 0 & m \end{pmatrix}$. At the bottom equilibria $\mathbf{u}^\star_{2k}$, the Hessian $\nabla^2 E(2k\pi, 0) = \begin{pmatrix} \kappa & 0 \\ 0 & m \end{pmatrix}$ is positive definite, since κ and m are positive constants. Therefore, the bottom equilibria are strict local minima of the energy, and so Theorem 20.27 guarantees their stability. The upper equilibrium points $\mathbf{u}^\star_{2k+1}$ are saddle points for the energy function since their Hessian $\nabla^2 E\big((2k+1)\pi, 0 \big) = \begin{pmatrix} -\kappa & 0 \\ 0 & m \end{pmatrix}$ is indefinite. Indeed, the phase portrait of the nonlinear pendulum nearby the unstable equilibria looks like a perturbed version of a linear saddle point.

[†] In a physical system, the potential energy is only defined up to an additive constant. Here we have fixed the zero energy level to be at the bottom of the pendulum's arc.

Each stable equilibrium is surrounded by a family of closed elliptically-shaped level curves, and hence forms a center. Each closed level curve corresponds to a periodic solution of the system, in which the pendulum moves back and forth. Near the equilibrium, the period is close to that of the linearized system, namely $2\pi/\sqrt{\alpha}$ as predicted by the eigenvalues. This fact underlies the use of pendulum-based clocks in time keeping, first recognized by Galileo. Grandfather clocks keep accurate time because the amplitude of the oscillations of their pendula are small. However, as we move further away from the equilibrium point, the solutions with very large amplitude oscillations, in which the pendulum becomes nearly vertical, have much longer periods.

The limiting case of the periodic solutions is of particular interest. The pair of curves connecting two distinct unstable equilibria are known as the *homoclinic orbits*, and play an essential role in the more advanced analysis of the pendulum under perturbations. Physically, a homoclinic orbit corresponds to a pendulum that starts out just shy of vertical, goes through exactly one full rotation, and eventually (as $t \to \infty$) ends up vertical again.

Finally, the level sets lying above and below the "cat's-eyes" formed by the periodic orbits are known as the *running orbits*. Since $u = \theta$ is a 2π periodic angular variable, the running orbits $\big(\, u(t), v(t) \,\big)^T = (\theta(t), \dot\theta(t))^T$, in fact, also correspond to periodic physical motions, in which the pendulum rotates around and around its pivot point. Since energy is conserved, the rotations persist forever. The larger the total energy $E(u, v)$, the farther away from the u–axis the level set, and the faster the pendulum spins.

In summary, the nature of a solution to the pendulum equation is almost entirely characterized by its energy:

$E = 0,$	stable equilibria,
$0 < E < 2\kappa,$	periodic oscillating orbits,
$E = 2\kappa,$	unstable equilibria and homoclinic orbits,
$E > 2\kappa,$	running orbits.

Example 20.30. The equations governing the rotation of a rigid body around a fixed point are known as the *Euler equations* of rigid body mechanics, [**65**], in honor of the prolific eighteenth century Swiss mathematician Leonhard Euler. According to Exercise ■, the eigenvectors of the positive definite inertia tensor of the body prescribe the three mutually orthogonal principal axes of rotation. The corresponding eigenvalues $0 < I_1, I_2, I_3$ are the principal *moments of inertia* of the body. Let $u_1(t), u_2(t), u_3(t)$ denote the angular momenta of the body around its three principal axes. In the absence of external forces, the dynamical system governing a rotating body takes the symmetric form

$$\frac{du_1}{dt} = \frac{I_2 - I_3}{I_2\, I_3}\, u_2\, u_3, \qquad \frac{du_2}{dt} = \frac{I_3 - I_1}{I_1\, I_3}\, u_1\, u_3, \qquad \frac{du_3}{dt} = \frac{I_1 - I_2}{I_1\, I_2}\, u_1\, u_2. \tag{20.56}$$

This system models, for example, the dynamics of a satellite spinning in its orbit around the earth. The solution will prescribe the rotations of the satellite around its center of mass, but not the overall motion of the center of mass as the satellite orbits the earth.

Let us assume that the body has three different moments of inertia, which we place in increasing order $0 < I_1 < I_2 < I_3$. The equilibria of the system are where the right

hand sides simultaneously vanish, which require that either $u_2 = u_3 = 0$ or $u_1 = u_3 = 0$ or $u_1 = u_2 = 0$. In other words, every point on the three coordinate axes is an equilibrium configuration! Since the variables represent angular momenta, the equilibria correspond to the body spinning around one of its principal axes at a fixed angular velocity. Let us analyze the stability of these equilibrium configurations. The linearization test fails completely — as it must do whenever dealing with a non-isolated equilibrium point. But the Euler equations turn out to have two independent first integrals:

$$E(\mathbf{u}) = \frac{1}{2}\left(\frac{u_1^2}{I_1} + \frac{u_2^2}{I_2} + \frac{u_3^2}{I_3}\right), \qquad A(\mathbf{u}) = \frac{1}{2}\left(u_1^2 + u_2^2 + u_3^2\right). \tag{20.57}$$

The first is the total kinetic energy, while the second is the total angular momentum. The proof that $dE/dt = 0 = dA/dt$ for any solution is left to the reader.

Since both E and A are constant, the solutions to the system are constrained to move along a common level set $C = \{E = e, A = a\}$. Thus, the solution curves are given by intersecting the sphere $S = \{A = a\}$ of radius $\sqrt{2\,a}$ with the ellipsoid where $L = \{E = e\}$. In Figure rigid■, we have graphed the intersection curves $C = S \cap L$ of a fixed sphere with a family of ellipsoids corresponding to different values of the kinetic energy. The six equilibria on the sphere are at its intersections with the coordinate axes. Those on the x and z axes are surrounded by closed periodic orbits, and hence are stable equilibria; indeed, they are, respectively, local minima and maxima of the energy when restricted to the sphere. On the other hand, the two equilibria on the y axis have the form of unstable saddle points. We conclude that a body that spins around its principal axes corresponding to the smallest or the largest moments of inertia is stable, whereas one that spins around the axis corresponding to the intermediate moment of inertia is unstable. This mathematical deduction can be demonstrated physically by flipping a solid rectangular object, e.g., this book, up into the air. It is easy to arrange it to spin around its long axis or its short axis in a stable manner, but it will balk at attempts to make it rotate around its middle axis!

Lyapunov's Method

Systems that incorporate damping and/or frictional effects do not typically have first integrals. From a physical standpoint, the damping will cause the total energy of the system to be a decreasing function of time. Eventually, the systems returns to an equilibrium position, and the extra energy has been dissipated away. However, this physical law has implications for the behavior of solutions. In particular, it can also be used to prove stability, even in cases when the linearization stability test is inconclusive. The nineteenth century Russian mathematician Alexander Lyapunov was the first to pinpoint the importance of such functions.

Definition 20.31. A function $L(\mathbf{u})$ is known as a *Lyapunov function* for the first order system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ if it satisfies

$$\frac{d}{dt}\,L(\mathbf{u}(t)) \;\le\; 0 \qquad \text{for all solutions} \qquad \mathbf{u}(t). \tag{20.58}$$

It is worth pointing out that one can verify the Lyapunov inequality (20.58) without actually having to solve the system. Namely, by the same chain rule computation as we used to establish the first integral criterion (20.46), we find $\dfrac{d}{dt}\, L(\mathbf{u}) = \nabla L(\mathbf{u}) \cdot \mathbf{F}(\mathbf{u})$, and hence $L(\mathbf{u})$ is a Lyapunov function if and only if

$$\nabla L(\mathbf{u}) \cdot \mathbf{F}(\mathbf{u}) \;\le\; 0 \qquad \text{for all} \qquad \mathbf{u}.$$

However, unlike first integrals which can, at least in principle, be systematically constructed by solving a first order partial differential equation, finding Lyapunov functions is more of an art form, usually relying on physical intuition or inspired guesswork.

The Lyapunov inequality (20.58) implies that a Lyapunov function must be decreasing,

$$L(\mathbf{u}(t)) \;\le\; L(\mathbf{u}(t_0)) \qquad \text{for all} \qquad t > t_0,$$

when evaluated on any solution to the system. The proof of Theorem 20.27 can be readily adapted to prove stability of a system with a Lyapunov function. Details can be found in [**72**, **80**].

Theorem 20.32. *If $L(\mathbf{u})$ is a Lyapunov function for the autonomous system of ordinary differential equations $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ and $\mathbf{u}^\star$ is a strict local minimum of L, then $\mathbf{u}^\star$ is a stable equilibrium point for the system. If the Lyapunov inequality* (20.58) *is strict for all nearby solutions (except, of course, the equilibrium itself), then the minimum $\mathbf{u}^\star$ is, in fact, asymptotically stable.*

In a damped mechanical system, the energy is decreasing, and so plays the role of a Lyapunov function. Unlike first integrals, maxima of Lyapunov functions are *not* stable.

Example 20.33. Return to the planar system

$$\frac{du}{dt} = v, \qquad\qquad \frac{dv}{dt} = -\,\alpha\, \sin u - \beta\, v,$$

describing the damped oscillations of a pendulum, as in (20.44). Physically, we expect that the damping will cause a continual decrease in the total energy in the system, which, by (20.55) is

$$E = \tfrac{1}{2}\, m\, v^2 + \kappa\, (1 - \cos u).$$

We compute its time derivative, when $u(t), v(t)$ is a solution to the damped system. Recalling that $\alpha = \kappa/m$, $\beta = \mu/m$, we find

$$\frac{dE}{dt} = m\, v\, \frac{dv}{dt} + \kappa \sin u\, \frac{du}{dt} = m\, v\, (-\,\alpha\, \sin u - \beta\, v) + \kappa\, v \sin u = -\,\mu\, v^2 \le 0,$$

since we are assuming that the frictional coefficient $\mu > 0$. Therefore, the energy satisfies the Lyapunov stability criterion, and hence Theorem 20.32 re-establishes the stability of the energy minima $u = 2\,k\,\pi$, $v = 0$, where the damped pendulum is at the bottom of the arc.

20.4. Numerical Solution Methods.

Since we are not able to solve the vast majority of differential equations in explicit, analytic form, the design of suitable numerical algorithms for accurately approximating the solutions is an essential component of the applied mathematician's toolbox. The ubiquity of differential equations in all areas of applications has inspired the tremendous research effort devoted to the development of numerical solution methods, some dating back to the beginnings of the calculus. Nowadays, many excellent computer packages are available for numerically solving ordinary differential equations. All give reliable and accurate results for a broad range of systems, at least for solutions over moderately long time periods. However, all of these packages and the underlying methods have their limitations, and it is essential that one be able to to recognize when the methods are working as advertised, and when they are giving spurious results! Here is where the theory, particularly the classification of equilibria and their stability properties, as well as first integrals and Lyapunov functions, can play an essential role. Explicit solutions, when known, can also be used as test cases for tracking the reliability and accuracy of a chosen numerical scheme.

In this section, we concentrate on numerical methods for initial value problems. We shall develop and analyze a few of the most basic single step schemes, culminating in the very popular fourth order Runge–Kutta method. This should only serve as a very basic introduction to the subject, and many other useful methods can be found in more specialized texts, [**node**]. Some equations are more difficult to accurately approximate than others, and a variety of more specialized methods are employed when confronted with a recalcitrant system.

Euler's Method

The key issues already appear when confronting the simplest first order ordinary differential equation

$$\frac{du}{dt} = F(t,u), \qquad\qquad u(t_0) = u_0. \tag{20.59}$$

To keep matters simple, we will concentrate on the scalar case; however, the methods are all phrased in a manner that allows them to be readily adapted to first order systems — just replace the scalar functions $u(t)$ and $F(t,u)$ by vectors $\mathbf{u}$ and $\mathbf{F}(t,\mathbf{u})$ throughout. (The time t, of course, remains a scalar.) Higher order ordinary differential equations are usually handled by first converting them into an equivalent first order system, as discussed in Section 20.1, and then applying the numerical methods thereunto. (An exception is the finite element methods for boundary value problems, introduced in Section 11.6, which works directly on the higher order equation.)

We begin with the very simplest method for solving the initial value problem (20.59). The method is named after Euler — although Newton and contemporaries were well aware of such a simple technique. Euler's method is rarely used because much more efficient and accurate techniques can be implemented with minimal additional work. Nevertheless, the method lies at the core of the entire subject, and must be thoroughly understood before progressing on to the more sophisticated algorithms that are used in real-life computations.

Starting at the initial point t_0, we introduce *mesh points*

$$t_0 < t_1 < t_2 < t_3 < \cdots ,$$

continuing on until we reach a desired final time $t_n = t_\star$. The mesh points t_k should be fairly closely spaced. In our presentation, we will always adopt a uniform *step size*, and so

$$h = t_{k+1} - t_k > 0, \tag{20.60}$$

does not depend on k and is assumed to be relatively small. This assumption serves to simplify the analysis, and does not significantly affect the underlying ideas. For a uniform step size, the k^{th} mesh point is at $t_k = t_0 + k\,h$. More sophisticated *adaptive* methods, in which the step size is adjusted in order to maintain accuracy of the numerical solution, can be found in more specialized texts, e.g., [**node**].

A numerical algorithm will recursively compute approximations $u_k \approx u(t_k)$ to the sampled values of the solution $u(t)$ at the chosen mesh points. Our goal is to make the *error* $E_k = u_k - u(t_k)$ in the approximation at each time t_k as small as possible. If required, the values of the solution $u(t)$ between mesh points may be computed by a subsequent interpolation procedure, e.g., based upon cubic splines.

Euler's method begins with the standard first order Taylor approximation to the solution. Thus, we approximate $u(t)$ near the mesh point t_k by its tangent line

$$u(t) \;\approx\; u(t_k) + (t - t_k)\,\frac{du}{dt}\,(t_k) = u(t_k) + (t - t_k)\,F(t_k, u(t_k)),$$

where we replace the derivative du/dt of the solution by the right hand side of the governing differential equation (20.59). In particular, the approximate value of the solution at the subsequent mesh point is

$$u(t_{k+1}) \;\approx\; u(t_k) + (t_{k+1} - t_k)\,F(t_k, u(t_k)). \tag{20.61}$$

This simple idea forms the basis of Euler's method.

Since in practice we only know the approximation u_k to the value of $u(t_k)$ at the current mesh point, we are forced to replace $u(t_k)$ by its approximation u_k. We thereby convert (20.61) into the iterative scheme

$$u_{k+1} = u_k + (t_{k+1} - t_k)\,F(t_k, u_k). \tag{20.62}$$

In particular, when using a uniform step size (20.60), *Euler's method* takes the simple form

$$u_{k+1} = u_k + h\,F(t_k, u_k). \tag{20.63}$$

As sketched in Figure 20.13, the method starts off approximating the solution reasonably well, but gradually loses accuracy as the errors accumulate.

Euler's method is the simplest example of a *one-step* numerical scheme for integrating an ordinary differential equation. The term "one-step" refers to the fact that the value for the succeeding approximation, $u_{k+1} \approx u(t_{k+1})$, depends only upon the current value, $u_k \approx u(t_k)$, which is one mesh point or step in back. To understand how Euler's method works in practice, we begin by looking at a problem we know how to solve. As usual, the

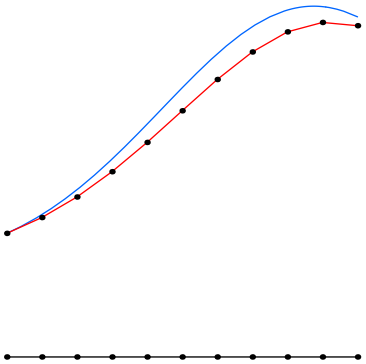


Figure 20.13. Euler's Method.

best way to test a numerical solution method is to begin by trying it on a problem with a known solution, since then we can determine exactly how large the resulting approximation error is.

Example 20.34. The simplest "nontrivial" ordinary differential equation is

$$\frac{du}{dt} = u, \qquad u(0) = 1.$$

The solution to the initial value problem is, of course, the exponential function $u(t) = e^t$. Since $F(t, u) = u$, Euler's method (20.63) with a fixed step size $h > 0$ takes the form

$$u_{k+1} = u_k + h\, u_k = (1 + h)\, u_k.$$

This linear iterative equation is easy to solve:

$$u_k = (1 + h)^k u_0 = (1 + h)^k,$$

which is our proposed approximation to the solution $u(t_k) = e^{t_k}$ at the mesh point $t_k = k\, h$. Therefore, by adopting the Euler scheme to solve the differential equation, we are effectively approximating the exponential function

$$e^{t_k} = e^{k\, h} \;\approx\; (1 + h)^k$$

by a power. When we replace the mesh time $t_k = k\, h$ by t, we recover, in the limit, a well-known calculus formula:

$$e^t = \lim_{h \to 0} (1 + h)^{t/h} = \lim_{k \to \infty} \left(1 + \frac{t}{k} \right)^k.$$

The student familiar with the theory of compound interest, [**int**], will recognize this particular approximation. As the time interval of compounding, h, gets smaller and smaller, the amount in the savings account approaches an exponential. Pay attention to the fact that the smaller the step size, the more steps, and hence the more work required to reach a given time. Thus, for time $t = 1$ we need $k = 10$ steps of size $h = .1$, but $k = 1000$ steps of size $h = .001$.

How good is the resulting approximation? The *error*

$$E(t_k) = E_k = u_k - e^{t_k}$$

measures the difference between the true solution and its numerical approximation at time $t = t_k = k\,h$. Let us tabulate the error at the particular times $t = 1, 2$ and 3 for various values of the step size h. The actual solution values are

$$e^1 = e = 2.718281828\ldots, \qquad e^2 = 7.389056096\ldots, \qquad e^3 = 20.085536912\ldots.$$

In this case, the approximate solution always underestimates the true solution.

h	$E(1)$	$E(2)$	$E(3)$
.1	$-.125$	$-.662$	-2.636
.01	$-.0134$	$-.0730$	$-.297$
.001	$-.00135$	$-.00738$	$-.0301$
.0001	$-.000136$	$-.000739$	$-.00301$
.00001	$-.0000136$	$-.0000739$	$-.000301$

Some key observations:

(*i*) The further t is away from the initial point $t_0 = 0$, the larger the magnitude of the error for a given step size.

(*ii*) On the other hand, the smaller the step size, the smaller the error. The trade-off is that more computational effort[†] is required to produce the numerical approximation.

(*iii*) The error is more or less in proportion to the step size. Decreasing the step size by a factor of $\frac{1}{10}$ decreases the error by a similar amount, but simultaneously increases the amount of computation by a factor of 10.

The final observation is indicative of the fact that the Euler method is of *first order*, which means that the error depends *linearly*[‡] on the step size h. More specifically, at a fixed time t, the error is bounded by

$$| E(t) | = | u_k - u(t) | \le C(t)\, h, \qquad \text{when} \qquad t = t_k = k\, h, \tag{20.64}$$

for some positive $C(t) > 0$ that depends upon the time and the particular solution, but not on the step size.

[†] In this case, there happens to be an explicit formula for the numerical solution which can be used to bypass the iterations. However, in almost any other situation, one cannot compute the approximation u_k without having first determined the intermediate values $u_0, \ldots, u_{k-1}$.

[‡] See the discussion of the order of iterative methods in Section 19.1 for motivation.

$h = .1$ $h = .01$

Figure 20.14. Euler's Method for $\dot{u} = \left(1 - \frac{4}{3}t\right) u$.

Example 20.35. The solution to the initial value problem

$$\frac{du}{dt} = \left(1 - \tfrac{4}{3}t\right) u, \qquad u(0) = 1. \tag{20.65}$$

was found in Example 20.3 by the method of separation of variables: $u(t) = \exp\left(t - \frac{2}{3}t^2\right)$. Euler's method leads to the iterative numerical scheme

$$u_{k+1} = u_k + h\left(1 - \tfrac{4}{3}t_k\right) u_k, \qquad u_0 = 1,$$

to approximate the solution. The following table lists the errors $E(t_k) = u_k - u(t_k)$ between the values computed by the Euler scheme and the actual solution values

$$u(1) = 1.395612425\ldots, \qquad u(2) = 0.513417119\ldots, \qquad u(3) = 0.049787068\ldots,$$

for several different step sizes.

h	$E(1)$	$E(2)$	$E(3)$
0.1000	0.07461761	0.03357536	−0.00845267
0.0100	0.00749258	0.00324416	−0.00075619
0.0010	0.00074947	0.00032338	−0.00007477
0.0001	0.00007495	0.00003233	−0.00000747

As in the previous example, each decrease in step size by a factor of 10 leads to one additional decimal digit of accuracy in the computed solution. In Figure 20.14 we compare the graphs of the actual and numerical solutions for step sizes $h = .1$ and .01.

Taylor Methods

In general, the order of a numerical solution method governs both the accuracy of its approximations and the speed at which they converge to the true solution. Although the Euler method is simple and easy to implement, it is only a first order method, and therefore of rather limited utility for efficiently computing accurate approximations to the

solution to an initial value problem. Thus, there is a great need to devise simple numerical methods that have a much higher order of accuracy.

Our derivation of the Euler method was based on a first order Taylor approximation to the solution. An evident way to design a higher order method is to employ a higher order Taylor approximation. The Taylor series expansion for the solution $u(t)$ at the succeeding mesh point $t_{k+1} = t_k + h$ has the form

$$u(t_{k+1}) = u(t_k + h) = u(t_k) + h\,\frac{du}{dt}(t_k) + \frac{h^2}{2}\,\frac{d^2u}{dt^2}(t_k) + \,\cdots\,. \tag{20.66}$$

As we just saw, we can evaluate the first derivative term $\dfrac{du}{dt} = F(t,u)$ through use of the underlying differential equation. The second derivative term can be found by differentiating with respect to t. Invoking the chain rule,

$$\begin{aligned}\frac{d^2u}{dt^2} = \frac{d}{dt}\,\frac{du}{dt} = \frac{d}{dt}\,F(t,u(t)) &= \frac{\partial F}{\partial t}(t,u) + \frac{\partial F}{\partial u}(t,u)\,\frac{du}{dt}\\ &= \frac{\partial F}{\partial t}(t,u) + \frac{\partial F}{\partial u}(t,u)\,F(t,u) \equiv F^{(2)}(t,u).\end{aligned} \tag{20.67}$$

This operation is sometimes known as the *total derivative*, indicating that that we must treat the second variable u as a function of t. Substituting the resulting formula into (20.66) and truncating at order h^2 leads to the *second order Taylor method*

$$\begin{aligned}u_{k+1} &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k)\\ &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,\left(\frac{\partial F}{\partial t}(t_k,u_k) + \frac{\partial F}{\partial u}(t_k,u_k)\,F(t_k,u_k)\right),\end{aligned} \tag{20.68}$$

in which, as before, we replace the solution value $u(t_k)$ by its computed approximation u_k. The resulting method is of second order, meaning that the error function satisfies the quadratic error estimate

$$|\,E(t)\,| = |\,u_k - u(t)\,| \le C(t)\,h^2 \qquad \text{when} \qquad t = t_k = k\,h. \tag{20.69}$$

Example 20.36. Let us explicitly formulate the second order Taylor method for the initial value problem (20.65). Here

$$\begin{aligned}\frac{du}{dt} &= F(t,u) = \left(1 - \tfrac{4}{3}\,t\right)u,\\ \frac{d^2u}{dt^2} &= \frac{d}{dt}\,F(t,u) = -\tfrac{4}{3}\,u + \left(1 - \tfrac{4}{3}\,t\right)\,\frac{du}{dt} = -\tfrac{4}{3}\,u + \left(1 - \tfrac{4}{3}\,t\right)^2 u,\end{aligned}$$

and so (20.68) becomes

$$u_{k+1} = u_k + h\left(1 - \tfrac{4}{3}\,t_k\right)u_k + \frac{h^2}{2}\left[-\tfrac{4}{3}\,u_k + \left(1 - \tfrac{4}{3}\,t_k\right)^2 u_k\right], \qquad u_0 = 1.$$

The following table lists the errors between the values computed by the second order Taylor scheme and the actual solution values, as given in Example 20.35.

h	$E(1)$	$E(2)$	$E(3)$
0.100	0.00276995	−0.00133328	0.00027753
0.010	0.00002680	−0.00001216	0.00000252
0.001	0.00000027	−0.00000012	0.00000002

In accordance with the quadratic error estimate (20.69), a decrease in the step size by a factor of $\frac{1}{10}$ leads in an increase in accuracy of the solution by a factor $\frac{1}{100}$, i.e., an increase in 2 significant decimal places in the numerical approximation of the solution.

Higher order Taylor methods are obtained by including further terms in the expansion (20.66). For example, to derive a third order Taylor method, we include the third order term where we evaluate the third derivative by differentiating (20.67), and so

$$\begin{aligned} \frac{d^3u}{dt^3} &= \frac{d}{dt}\,\frac{d^2u}{dt^2} = \frac{d}{dt}\,F^{(2)}(t,u) = \frac{\partial F^{(2)}}{\partial t} \;+\; \frac{\partial F^{(2)}}{\partial u}\,\frac{du}{dt} = \frac{\partial F^{(2)}}{\partial t} \;+\; F\,\frac{\partial F^{(2)}}{\partial u} \\ &= \frac{\partial^2 F}{\partial t^2} + 2\,F\,\frac{\partial^2 F}{\partial t\,\partial u} + F^2\,\frac{\partial^2 F}{\partial u^2} + \frac{\partial F}{\partial t}\,\frac{\partial F}{\partial u} + F\left(\frac{\partial F}{\partial u}\right)^2 \equiv F^{(3)}(t,u). \end{aligned} \tag{20.70}$$

The resulting third order Taylor method is

$$u_{k+1} = u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k) + \frac{h^3}{6}\,F^{(3)}(t_k,u_k), \tag{20.71}$$

where the last two summand are given by (20.67), (20.70), respectively. The higher order expressions are even worse, and a good symbolic manipulation system is almost essential. (Although, in the past, mathematicians were able to perform these sorts of computations by hand!)

Whereas higher order Taylor methods are easy to motivate, they are rarely used in practice. There are two principal difficulties:

(*a*) Owing to their dependence upon the partial derivatives of $F(t,u)$, they require the right hand side of the differential equation to be rather smooth.

(*b*) Even worse, the explicit formulae become exceedingly complicated, even for relatively simple functions $F(t,u)$. Efficient evaluation of the multiplicity of terms in the Taylor approximation becomes a significant concern.

As a result, mathematicians soone abandoned the Taylor series approach, and began to look elsewhere for high order, efficient integration methods.

Error Analysis

Before continuing our investigations, we need to engage in a more serious discussion of the error in a numerical scheme. A general *one-step* numerical method can be written in the form

$$u_{k+1} = G(h,t_k,u_k), \tag{20.72}$$

where G is a prescribed function of the current value u_k, the point t_k itself, and the step size $h = t_{k+1} - t_k$, which, for illustrative purposes, we assume to be fixed. We leave the discussion of multi-step methods, in which G could also depend upon the earlier values $u_{k-1}, u_{k-2}, \dots$, to more advanced texts, e.g., [**node**].

In any numerical integration scheme there are, in general, three sources of error.

(*i*) The first is the *local error* committed in the current step of the algorithm. Even if we had managed to compute a completely accurate value of the solution $u_k = u(t_k)$ at time t_k, the numerical approximation scheme (20.72) is not exact, and will therefore introduce an error into the next computed value $u_{k+1} \approx u(t_{k+1})$.

(*ii*) The second source of error is due to the error that is already present in the current approximation $u_k \approx u(t_k)$. The local errors tend to accumulate as we continue to integrate the differential equation, and the net result is the *global error* in the scheme. The global error is what we actually observe in practice.

(*iii*) Finally, if the initial condition $u_0 \approx u(t_0)$ is not computed accurately, this *initial error* will also make a contribution. For example, if $u(t_0) = \pi$, then we introduce some initial error by using a decimal approximation, say $\pi \approx 3.14159$.

The third error source is relatively unimportant, and will be ignored in our discussion, i.e., we will assume $u_0 = u(t_0)$ is exact. Then the global error will be an accumulation of successive local errors, and so we must first understand the local error in detail.

To measure the local error in going from t_k to t_{k+1}, we compare the exact solution value $u(t_{k+1})$ with its numerical approximation (20.72) under the assumption that the current computed value is correct: $u_k = u(t_k)$. Of course, in practice this is never the case, and so the local error is an artificial quantity. Be that as it may, in most circumstances the local error is (*a*) easy to estimate, and, (*b*) provides a very good guide to the global accuracy of the numerical scheme. To estimate the local error, we assume that the step size h is small and approximate the solution $u(t)$ by its Taylor expansion

$$\begin{aligned} u(t_{k+1}) &= u(t_k) + h\,\frac{du}{dt}(t_k) + \frac{h^2}{2}\,\frac{d^2u}{dt^2}(t_k) + \frac{h^3}{6}\,\frac{d^3u}{dt^3}(t_k) + \,\cdots \\ &= u_k + h\,F(t_k, u_k) + \frac{h^2}{2}\,F^{(2)}(t_k, u_k) + \frac{h^3}{6}\,F^{(3)}(t_k, u_k) + \,\cdots\,, \end{aligned} \tag{20.73}$$

where we have used (20.67), (20.70), etc., to evaluate the derivative terms, and then our assumption to replace $u(t_k)$ by u_k. On the other hand, a direct Taylor expansion, in h, of the numerical scheme produces

$$u_{k+1} = G(h, t_k, u_k) = G(0, t_k, u_k) + h\,\frac{\partial G}{\partial h}\,(0, t_k, u_k) + \frac{h^2}{2}\,\frac{\partial^2 G}{\partial h^2}\,(0, t_k, u_k) + \,\cdots\,. \tag{20.74}$$

The local error is obtained by comparing these two Taylor expansions.

Definition 20.37. A numerical integration method is of *order* n if the Taylor expansions (20.73), (20.74) of the exact and numerical solutions agree up to order h^n.

For example, the Euler method

$$u_{k+1} = G(h, t_k, u_k) = u_k + h\,F(t_k, u_k),$$

is already in the form of a Taylor expansion — there are no terms involving h^2 and higher powers. Comparing with the exact expansion (20.73), we see that the constant and order h terms are the same, but the order h^2 terms differ. Thus, according to the definition, the Euler method is a first order method. Similarly, the Taylor method (20.68) is a second order method, because it was explicitly designed to match the constant, h and h^2 terms in the Taylor expansion of the solution (20.73). For a general Taylor method of order n, one chooses $G(h, t_k, u_k)$ to be exactly the order n Taylor polynomial.

Under fairly general hypotheses, it can be proved that if the method has order n as measured by the local error, then the *global error* is bounded by a multiple of h^n. In other words, if the initial condition $u_0 = u(t_0)$ is accurate, then the computed value u_k differs from the solution at time t_k by an amount

$$| u_k - u(t_k) | \leq M h^n \tag{20.75}$$

where the constant $M > 0$ may depend on the time t_k and the solution $u(t)$. The error bound justifies our numerical observations. For a method of order n, decreasing the step size by a factor of $\frac{1}{10}$ will decrease the error in the solution by a factor of at least 10^{-n}, and so, roughly, we expect to pick up an additional n digits of accuracy in the solution value — at least up until the point that round-off errors begin to play a role in the computation. This rule of thumb needs to be taken with a small grain of salt; nevertheless, it is amply borne out in our test examples. Readers interested in a complete error analysis of numerical integration schemes should consult a more specialized text, e.g., [**node**].

The bottom line is the higher its order, the more accurate the numerical scheme, and hence the larger the step size that can be used to produce the solution to a desired accuracy. If the total amount of computation has also decreased, then the high order method is to be preferred over a simpler, lower order method. Our goal now is to find another route to the design of higher order methods that avoids the complications inherent in a direct Taylor expansion. More specifically, we seek suitably compact combinations of function values that reproduce the Taylor expansion of the solution (20.73) to high order.

An Equivalent Integral Equation

The secret to the design of effective higher order numerical algorithms is to replace the differential equation by an equivalent integral equation. By way of motivation, recall that, in general, differentiation is a badly behaved process; a reasonable function can have an unreasonable derivative. On the other hand, integration ameliorates; even quite nasty functions have relatively well-behaved integrals. For the same reason, accurate numerical integration is relatively easy, whereas numerical differentiation should be avoided if possible. While we have not dealt directly with integral equations in this text, the subject has been extensively developed, [**39**], and has many important applications.

Conversion of the initial value problem (20.59) to an integral equation is straightforward. We integrate both sides of the differential equation from the initial point t_0 to a variable time t. The Fundamental Theorem of Calculus is used to explicitly evaluate the left hand integral:

$$u(t) - u(t_0) = \int_{t_0}^{t} \dot{u}(s)\, ds = \int_{t_0}^{t} F(s, u(s))\, ds.$$

Rearranging terms, we arrive at the key result.

Theorem 20.38. *There is a one-to-one correspondence between solutions to the initial value problem*

$$\frac{du}{dt} = F(t,u), \qquad\qquad u(t_0) = u_0,$$

and solutions to the integral equation

$$u(t) = u(t_0) + \int_{t_0}^{t} F(s,u(s))\,ds. \tag{20.76}$$

Proof: We already showed that the solution $u(t)$ to the initial value problem satisfies the integral equation (20.76). Conversely, suppose that $u(t)$ solves the integral equation. The Fundamental Theorem of Calculus tells us that the right hand side of (20.76) has derivative $\dfrac{du}{dt} = F(t,u(t))$ equal to the integrand. Moreover, at $t = t_0$, the integral has the same upper and lower limits, and so vanishes, which implies that $u(t) = u(t_0) = u_0$ has the correct initial conditions. *Q.E.D.*

Remark: Unlike the differential equation, the integral equation (20.76) requires no additional initial condition — it is automatically built into the equation. The proofs of the fundamental existence and uniqueness Theorems 20.8 and 20.10 for ordinary differential equations are, in fact, based on the integral reformulation of the initial value problem; see [**72**, **80**] for details.

The integral equation reformulation is equally valid for systems of first order ordinary differential equations. As noted above, the functions $\mathbf{u}(t)$ and $\mathbf{F}(t,\mathbf{u}(t))$ become vector valued. Integrating a vector-valued function is accomplished by integrating its individual components. Details are left to the reader.

Implicit and Predictor–Corrector Methods

From this point onwards, we shall abandon the original initial value problem, and turn our attention to trying to numerically solve the equivalent integral equation (20.76). Let us rewrite the equation, starting at the mesh point t_k instead of t_0, and integrating until time $t = t_{k+1}$. The result is the basic integral formula

$$u(t_{k+1}) = u(t_k) + \int_{t_k}^{t_{k+1}} F(s,u(s))\,ds \tag{20.77}$$

that (implicitly) computes the value of the solution at the subsequent mesh point. Comparing this formula with the Euler method

$$u_{k+1} = u_k + h\,F(t_k,u_k), \qquad \text{where} \qquad h = t_{k+1} - t_k,$$

and assuming for the moment that $u_k = u(t_k)$ is exact, we discover that we are merely approximating the integral by

$$\int_{t_k}^{t_{k+1}} F(s,u(s))\,ds \;\approx\; h\,F(t_k,u(t_k)). \tag{20.78}$$

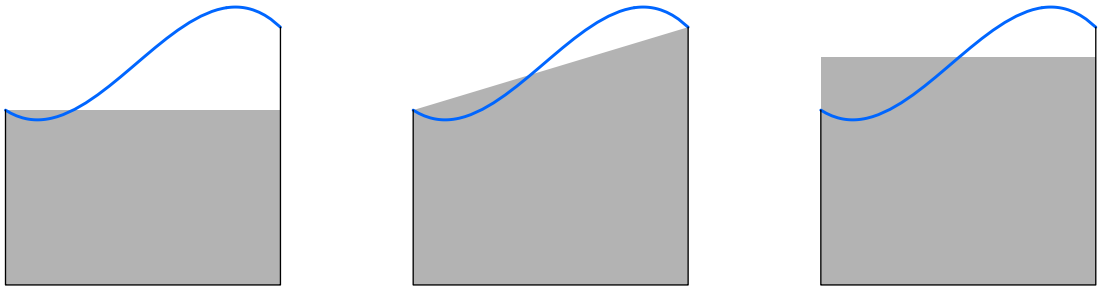


Figure 20.15. Numerical Integration Methods.

Formula (20.78) is the left-endpoint rule for numerical integration, that approximates the area under the curve $g(t) = F(t, u(t))$ between $t_k \le t \le t_{k+1}$ by the area of a rectangle whose height $g(t_k) = F(t_k, u(t_k)) \approx F(t_k, u_k)$ is prescribed by the left-hand endpoint of the graph. As indicated in Figure 20.15, this is a reasonable, but not especially accurate method of numerical integration.

In first year calculus, you no doubt encountered much better methods of approximating the integral of a function. One of these is the *trapezoid rule*, which approximates the integral of the function $g(t)$ by the area of a trapezoid obtained by connecting the two points $(t_k, g(t_k))$ and $(t_{k+1}, g(t_{k+1}))$ on the graph of g by a straight line, as in Figure 20.15. Let us therefore try replacing (20.78) by the more accurate trapezoidal approximation

$$\int_{t_k}^{t_{k+1}} F(s, u(s))\, ds \;\approx\; \tfrac{1}{2}\, h \,\big[\, F(t_k, u(t_k)) + F(t_{k+1}, u(t_{k+1})) \,\big]. \tag{20.79}$$

Substituting this approximation into the integral formula (20.77), and replacing the solution values $u(t_k), u(t_{k+1})$ by their numerical approximations, leads to the (hopefully) more accurate numerical scheme

$$u_{k+1} = u_k + \tfrac{1}{2}\, h \,\big[\, F(t_k, u_k) + F(t_{k+1}, u_{k+1}) \,\big], \tag{20.80}$$

known as the *Trapezoid method*. The trapezoid method is an *implicit scheme*, since the updated value u_{k+1} appears on both sides of the equation, and hence is only defined implicitly.

Example 20.39. Consider the differential equation $\dot u = \big(1 - \frac{4}{3}\, t\big)\, u$ studied in Examples 20.35 and 20.36. The trapezoid rule with a fixed step size h takes the form

$$u_{k+1} = u_k + \tfrac{1}{2}\, h \,\big[\, \big(1 - \tfrac{4}{3}\, t_k\big)\, u_k + \big(1 - \tfrac{4}{3}\, t_{k+1}\big)\, u_{k+1} \,\big].$$

In this case, we can explicit solve for the updated solution value, leading to the recursive formula

$$u_{k+1} = \frac{1 + \frac{1}{2}\, h\, \big(1 - \frac{4}{3}\, t_k\big)}{1 - \frac{1}{2}\, h\, \big(1 - \frac{4}{3}\, t_{k+1}\big)}\; u_k = \frac{1 + \frac{1}{2}\, h - \frac{2}{3}\, h\, t_k}{1 - \frac{1}{2}\, h + \frac{2}{3}\, h\, (t_k + h)}\; u_k. \tag{20.81}$$

Implementing this scheme for three different step sizes gives the following errors between the computed solution and true solution at times $t = 1, 2, 3$.

h	$E(1)$	$E(2)$	$E(3)$
0.100	−0.00133315	0.00060372	−0.00012486
0.010	−0.00001335	0.00000602	−0.00000124
0.001	−0.00000013	0.00000006	−0.00000001

The numerical data is in full accord with the fact that method is of second order. For each reduction in step size by $\frac{1}{10}$, the accuracy in the solution increases by, roughly, a factor of $\frac{1}{100}$; that is, the numerical solution acquires two additional accurate decimal digits.

The main problem with the trapezoid scheme (and any other implicit scheme) is immediately apparent. The updated approximate value for the solution u_{k+1} appears on both sides of the equation (20.80). Only for very simple functions $F(t,u)$ can one expect to solve (20.80) explicitly for u_{k+1} in terms of the known quantities t_k, u_k and $t_{k+1} = t_k + h$. The alternative is to employ a numerical equation solver such as the bisection algorithm or Newton's method to compute u_{k+1}. In the case of Newton's method, one would use the current approximation u_k as a first guess for the new approximation u_{k+1} — similar to the continuation method discussed in Example 19.27. The resulting scheme takes a little bit of work to program, but can be effective in certain situations.

An alternative, less complicated strategy is based on the following far-reaching idea. We already know a half-way decent approximation to the solution value u_{k+1} — namely that provided by the more primitive Euler scheme

$$\widetilde{u}_{k+1} = u_k + h\, F(t_k, u_k). \tag{20.82}$$

Let's use this estimated value in place of u_{k+1} on the right hand side of the implicit equation (20.80). The result

$$\begin{aligned} u_{k+1} &= u_k + \tfrac{1}{2}\, h \,\big[\, F(t_k, u_k) + F(t_k + h, \widetilde{u}_{k+1}) \,\big] \\ &= u_k + \tfrac{1}{2}\, h \,\big[\, F(t_k, u_k) + F\big(\, t_k + h, u_k + h\, F(t_k, u_k)\,\big)\,\big]. \end{aligned} \tag{20.83}$$

is known as the *improved Euler method*. It is a completely explicit method since there is no need to solve any equation for the updated value u_{k+1}.

Example 20.40. For our favorite equation $\dot{u} = \big(1 - \frac{4}{3}\,t\,\big)\,u$, the improved Euler method begins with the Euler approximation $\widetilde{u}_{k+1} = u_k + h\,\big(1 - \frac{4}{3}\,t_k\,\big)\,u_k$, and then replaces it by the improved value

$$\begin{aligned} u_{k+1} &= u_k + \tfrac{1}{2}\, h \,\big[\, \big(1 - \tfrac{4}{3}\,t_k\,\big)\,u_k + \big(1 - \tfrac{4}{3}\,t_{k+1}\,\big)\,\widetilde{u}_{k+1}\,\big] \\ &= u_k + \tfrac{1}{2}\, h \,\big[\, \big(1 - \tfrac{4}{3}\,t_k\,\big)\,u_k + \big(1 - \tfrac{4}{3}\,(t_k + h)\,\big)\,\big(u_k + h\,\big(1 - \tfrac{4}{3}\,t_k\,\big)\,u_k\,\big)\,\big] \\ &= \Big[\, \big(1 - \tfrac{2}{3}\,h^2\,\big)\,\big[\,1 + h\,\big(1 - \tfrac{4}{3}\,t_k\,\big)\,\big] + \tfrac{1}{2}\,h^2\,\big(1 - \tfrac{4}{3}\,t_k\,\big)^2\,\Big]\,u_k. \end{aligned}$$

Implementing this scheme leads to the following errors in the numerical solution at the indicated times. The improved Euler method performs comparably to the fully implicit scheme (20.81), and significantly better than the original Euler method.

h	$E(1)$	$E(2)$	$E(3)$
0.100	−0.00070230	0.00097842	0.00147748
0.010	−0.00000459	0.00001068	0.00001264
0.001	−0.00000004	0.00000011	0.00000012

The improved Euler method is the simplest of a large family of so-called *predictor–corrector algorithms*. In general, one begins a relatively crude method — in this case the Euler method — to *predict* a first approximation $\widetilde{u}_{k+1}$ to the desired solution value u_{k+1}. One then employs a more sophisticated, typically implicit, method to *correct* the original prediction, by replacing the required update u_{k+1} on the right hand side of the implicit scheme by the less accurate prediction $\widetilde{u}_{k+1}$. The resulting explicit, corrected value u_{k+1} will, provided the method has been designed with due care, be an improved approximation to the true solution.

The numerical evidence in Example 20.40 indicates that the improved Euler scheme is a second order method. To verify this experimental prediction, we expand the right hand side of (20.83) in a Taylor series in h, and then compare, term by term, with the solution expansion (20.73). First,

$$F\big(t_k + h, u_k + h\,F(t_k, u_k)\big) = F + h\big(F_t + F\,F_u\big) + \tfrac{1}{2}\,h^2\big(F_{tt} + 2\,F\,F_{tu} + F^2\,F_{uu}\big) + \cdots,$$

where all the terms on the right hand side are evaluated at t_k, u_k. Substituting into (20.83), we find

$$u_{k+1} = u_k + h\,F + \tfrac{1}{2}\,h^2\big(F_t + F\,F_u\big) + \tfrac{1}{4}\,h^2\big(F_{tt} + 2\,F\,F_{tu} + F^2\,F_{uu}\big) + \cdots. \tag{20.84}$$

The Taylor expansions (20.73), (20.84) agree in their order $1, h$ and h^2 terms, but differ at order h^3. This confirms that the improved Euler method is of second order.

We can design a range of numerical solution schemes by implementing alternative numerical approximations to the basic integral equation (20.77). For example, the midpoint rule approximates the integral of the function $g(t)$ by the area of the rectangle whose height is the value of the function at the midpoint:

$$\int_{t_k}^{t_{k+1}} g(s)\,ds \;\approx\; h\,g\big(t_k + \tfrac{1}{2}\,h\big), \qquad \text{where} \qquad h = t_{k+1} - t_k. \tag{20.85}$$

See Figure 20.15 for an illustration. The midpoint rule is known to have the same order of accuracy as the trapezoid method. Substituting into (20.77) leads to the approximation

$$u_{k+1} = u_k + \int_{t_k}^{t_{k+1}} F(s, u(s))\,ds \;\approx\; u_k + h\,F\big(t_k + \tfrac{1}{2}\,h, u\big(t_k + \tfrac{1}{2}\,h\big)\big).$$

Of course, we don't know the value of the solution $u\big(t_k + \frac{1}{2}\,h\big)$ at the midpoint, but can predict it through a straightforward adaptation of the basic Euler approximation:

$u\left(t_k + \frac{1}{2}\,h\right) \approx u_k + \frac{1}{2}\,h\,F(t_k, u_k)$. The result is the *midpoint method*

$$u_{k+1} = u_k + h\,F\left(t_k + \tfrac{1}{2}\,h, u_k + \tfrac{1}{2}\,h\,F(t_k, u_k)\right). \tag{20.86}$$

A comparison of the terms in the Taylor expansions of (20.73), (20.86) reveals that the midpoint method is also of second order.

Runge–Kutta Methods

The improved Euler and midpoint methods are the most elementary incarnations of a general class of numerical schemes for ordinary differential equations that were first systematically studied by the German mathematicians Carle Runge and Martin Kutta in the late nineteenth century. The Runge–Kutta methods are by far the most popular and powerful general-purpose numerical methods for integrating ordinary differential equations. While not appropriate in all possible situations, Runge–Kutta schemes are surprisingly adaptable and perform quite efficiently and accurately in a wide variety of systems, and, barring complications, tend to be the method of choice in most basic applications. They and their variants comprise the engine that powers most computer software for solving general initial value problems for systems of ordinary differential equations.

A general *Runge–Kutta method* takes the form

$$u_{k+1} = u_k + h\,\sum_{i=1}^{m} c_i\,F(t_{i,k}, u_{i,k}), \tag{20.87}$$

where m, the number of summands, is referred to as the number of *terms* in the method. Each $t_{i,k}$lying in the k^{th} mesh interval, and so $t_k \le t_{i,k} \le t_{k+1}$. The second argument $u_{i,k}$ should be viewed as an approximation to the solution at the point $t_{i,k}$, so $u_{i,k} \approx u(t_{i,k})$, and is computed by a simpler Runge–Kutta method of the same general format. Such methods are very flexible. One is free to choose the coefficients c_i, the times $t_{i,k}$, as well as the intermediate approximations $u_{i,k}$. As always, the *order* of the method is is indicated by the power of h to which the Taylor expansions of the numerical method (20.87) and the actual solution (20.73) agree. Clearly, the more terms we include in the Runge–Kutta formula (20.87), the more free parameters available to match terms in the solution's Taylor series, and so the higher the potential order of the method. Thus, the goal is to arrange said parameters so that the method has a high order of accuracy, while, simultaneously, avoiding unduly complicated, and hence computationally costly, formulae.

Both the improved Euler and midpoint methods are particular cases of a class of two term Runge–Kutta methods of the form

$$u_{k+1} = u_k + h\left[\,a\,F(t_k, u_k) + b\,F\big(t_k + \lambda h, u_k + \lambda\,h\,F(t_k, u_k)\big)\,\right], \tag{20.88}$$

based on the current mesh point $t_{k,1} = t_k$ and one intermediate point $t_{k,2} = t_k + \lambda h$, so $0 \le \lambda \le 1$. We use the basic Euler method to approximate the solution value $u_{k,2} = u_k + \lambda\,h\,F(t_k, u_k)$ at the intermediate value $t_{k,2}$. The improved Euler method uses $a = b = \frac{1}{2}$, and $\lambda = 1$, while the midpoint method corresponds to $a = 0, b = 1$ and $\lambda = \frac{1}{2}$. The

possible values of a, b and λ are to be determined by matching the Taylor expansion

$$\begin{aligned} u_{k+1} &= u_k + h\,\big[\, a\,F(t_k,u_k) + b\,F\big(t_k + \lambda\,h, u_k + \lambda\,h\,F(t_k,u_k)\big)\,\big] \\ &= u_k + h\,(a+b)\,F(t_k,u_k) + h^2\,b\,\lambda\,\left[\,\frac{\partial F}{\partial t}\,(t_k,u_k) + F(t_k,u_k)\,\frac{\partial F}{\partial u}\,(t_k,u_k)\right] + \,\cdots\,. \end{aligned}$$

(in powers of h) of the right hand side of (20.88) with the Taylor expansion (20.73) of the solution $u(t_{k+1}) = u(t_k + h)$ to as high an order as possible. The constant terms, u_k, are the same. For the order h and order h^2 terms to agree, we must have, respectively,

$$a + b = 1, \qquad\qquad b\,\lambda = \tfrac{1}{2}.$$

Therefore, setting $a = 1 - \mu$, $b = \mu$, and $\lambda = 1/(2\,\mu)$, where μ is arbitrary[†], leads to the family of two term, second order Runge–Kutta methods of the form

$$u_{k+1} = u_k + h\,\left[\,(1-\mu)\,F(t_k,u_k) + \mu\,F\left(t_k + \frac{h}{2\,\mu}, u_k + \frac{h}{2\,\mu}\,F(t_k,u_k)\right)\right]. \tag{20.89}$$

The case $\mu = \frac{1}{2}$ corresponds to the improved Euler method (20.83), while $\mu = 1$ gives the midpoint method (20.86). Unfortunately, none of these methods are able to match all of the third order terms in the Taylor expansion for the solution, and so we are left with a one-parameter family of two step Runge–Kutta methods, all of second order, that include the improved Euler and midpoint rules as particular instances. The cases when $\frac{1}{2} \le \mu \le 1$ all perform more or less comparably, and there is no special reason to prefer one over the other.

Thus, to construct a third order Runge–Kutta method, we need to take at least $m \ge 3$ terms in (20.87). A rather complicated symbolic computation will produce a range of valid schemes; the results can be found in [**nODE**]. Finding relatively simple, but high order Runge–Kutta methods is a rather tedious process, and we leave a complete discussion of the available options to a more advanced treatment. In practical applications, a particularly simple fourth order, four term method has become the most popular. The method, often abbreviated as RK4, takes the form

$$u_{k+1} = u_k + \frac{h}{6}\,\big[\,F(t_k,u_k) + 2\,F(t_{2,k},u_{2,k}) + 2\,F(t_{3,k},u_{3,k}) + F(t_{4,k},u_{4,k})\,\big], \tag{20.90}$$

where the times and function values are successively computed according to the following procedure:

$$\begin{aligned} t_{2,k} &= t_k + \tfrac{1}{2}\,h, & u_{2,k} &= u_k + \tfrac{1}{2}\,h\,F(t_k,u_k), \\ t_{3,k} &= t_k + \tfrac{1}{2}\,h, & u_{3,k} &= u_k + \tfrac{1}{2}\,h\,F(t_{2,k},u_{2,k}), \\ t_{4,k} &= t_k + h, & u_{4,k} &= u_k + h\,F(t_{3,k},u_{3,k}). \end{aligned} \tag{20.91}$$

The four term RK4 scheme (20.90), (20.91) is, in fact, a fourth order method. This is confirmed by demonstrating that the Taylor series expansion of the right hand side of

[†] Although we should restrict $\mu \ge \frac{1}{2}$ in order that $0 \le \lambda \le 1$.

(20.90) in powers of h matches all of the terms in the true Taylor series (20.73) up to and including those of order h^4, and hence the local truncation error is of order h^5. This is not a computation for the faint-hearted — bring lots of paper and erasers, or, better yet, a good computer algebra package! The RK4 scheme but one instance of a large family of possible fourth order, four term Runge–Kutta methods, but is by far the most popular owing to its relative simplicity.

Example 20.41. For our favorite equation $\dot{u} = \left(1 - \frac{4}{3}t\right) u$, the RK4 method leads to the following errors at the indicated times.

h	$E(1)$	$E(2)$	$E(3)$
0.100	-1.944×10^{-7}	1.086×10^{-6}	4.592×10^{-6}
0.010	-1.508×10^{-11}	1.093×10^{-10}	3.851×10^{-10}
0.001	-1.332×10^{-15}	-4.741×10^{-14}	1.932×10^{-14}

The results are phenomenally good — much better than any of the other methods. Each decrease in the step size by a factor of $\frac{1}{10}$ leads to 4 more decimal digits of accuracy, in accordance with it being a fourth order method.

Actually, it is not entirely fair to compare the accuracy of the methods at the same step size. Each iteration of the RK4 method requires four evaluations of the function $F(t, u)$, and hence takes the same computational effort as four Euler iterations, or, equivalently, two improved Euler iterations. Thus, the more revealing comparison would be between RK4 at step size h, Euler at step size $\frac{1}{4}h$, and improved Euler at step size $\frac{1}{2}h$, as these involve roughly the same amount of computational effort. The resulting errors $E(1)$ at time $t = 1$ are listed in the following table:

h	Euler	Improved Euler	Midpoint	Runge–Kutta 4
.1	1.872×10^{-2}	-1.424×10^{-4}	2.687×10^{-4}	-1.944×10^{-7}
.01	1.874×10^{-3}	-1.112×10^{-6}	2.788×10^{-6}	-1.508×10^{-11}
.001	1.87×10^{-4}	-1.080×10^{-8}	2.799×10^{-8}	-1.332×10^{-15}

The Runge–Kutta method clearly outperforms its rivals. At a step size of .1, it is almost as accurate as the improved Euler and midpoint methods with step size .0005, and hence 200 times the amount of computation, while the Euler method would require a step size of approximately $.24 \times 10^{-6}$, and would be $4,000,000$ times as slow as Runge–Kutta! With a step size of .001, RK4 computes a solution value that is near the limits imposed by machine accuracy using single precision arithmetic. The high performance level and accuracy of RK4 immediately explains its popularity for a broad range of applications.

Example 20.42. As noted above, by replacing the function values u_k by vectors $\mathbf{u}_k$, one can immediately apply the RK4 method to integrate initial value problems for first order systems of ordinary differential equations. Consider, by way of example, the Lotka–Volterra system

$$\frac{du}{dt} = 2\,u - u\,v, \qquad \frac{dv}{dt} = -9\,v + 3\,u\,v, \tag{20.92}$$

analyzed in Example 20.28. To find a numerical solution, we write $\mathbf{u} = (\,u, v\,)^T$ for the solution vector, and $\mathbf{F}(\mathbf{u}) = (\,2\,u - u\,v, -9\,v + 3\,u\,v\,)^T$. Thus, the Euler method for this system for step size h uses $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + h\,\mathbf{F}(\mathbf{u}^{(k)})$, or, explicitly, as a nonlinear iterative system

$$u^{(k+1)} = u^{(k)} + h\,(2\,u^{(k)} - u^{(k)}\,v^{(k)}), \qquad v^{(k+1)} = v^{(k)} + h\,(-9\,v^{(k)} + 3u^{(k)}\,v^{(k)}).$$

The improved Euler and Runge–Kutta schemes are implemented in a similar fashion. Phase plane pictures of the three numerical algorithms starting with initial conditions $u^{(0)} =$ ■ , $v^{(0)} =$ appear in Figure 9.3. Recall that the solution is supposed to travel periodically around a closed curve, given by the level set

$$I(u, v) = 9\,\log u - 3\,u + 2\,\log v - v = ■$$

of the first integral. The Euler method spirals away from the periodic solution, while ■ . Since we do not have an analytic formula† for the solution, we cannot measure the precise error in the methods. However, the first integral is supposed to remain constant on the solution trajectories, and so one means of monitoring the accuracy of the solution is by the variation in the numerical values of $I(u^{(k)}, v^{(k)})$. These are graphed in ■

In practical implementations, it is important to know whether the numerical solution is accurate or not. Since the accuracy of the computation is dependent upon the step size h, one should adjust h so as to maintain a preassigned level of accuracy. The result is an *adaptive* method, in which the step size is allowed to change during the course of the algorithm, in response to some measurement of overall error in the computation. Inaccurate solutions values would require a suitable reduction in step size. On the other hand, if the solution is more accurate than the application requires, one could increase h in order to reduce the total amount of computational effort.

How might one decide when a method is giving inaccurate results, since one presumably does not know the true solution and so has nothing to directly test the numerical approximation against? A useful idea is to integrate the differential equation using two different methods, usually of different orders of accuracy, and comparing the results. If the two solution values are reasonably close, then one is usually safe in assuming that the methods are both giving accurate results, while in the event that they differ beyond some preassigned tolerance, then one needs to re-evaluate the step size. Several methods

† However, compare Exercise ■ ■ .

are used in practical situation, the most popular of which is known as the Runge–Kutta–Fehlberg method, which combines a fourth and a fifth order Runge–Kutta scheme. Details can be found in more advanced treatments of the subject, e.g., [**nODE**].

Stiff Differential Equations

While the Runge–Kutta fourth order method with a sufficiently small step size will successively integrate a broad range of differential equations — at least over not unduly long time intervals — it does occasionally experience unexpected difficulties. While we have not developed sufficiently sophisticated analytical tools to conduct a thorough analysis, it will be instructive to look at why a breakdown might occur in a simpler context.

Example 20.43. The simple linear initial value problem

$$\frac{du}{dt} = -250\,u, \qquad u(0) = 1, \tag{20.93}$$

is an instructive and sobering example. The explicit solution is easy; it is a very rapidly decreasing exponential

$$u(t) = e^{-250\,t} \qquad \text{with} \qquad u(1) \;\approx\; 2.69 \times 10^{-109}.$$

The following table gives the result of approximating the solution $u(1)$ at $t = 1$ using three of our numerical integration schemes for several step sizes:

h	Euler	Improved Euler	RK4
.1	6.34×10^{13}	3.99×10^{24}	2.81×10^{41}
.01	4.07×10^{17}	1.22×10^{21}	1.53×10^{-19}
.001	1.15×10^{-125}	6.17×10^{-108}	2.69×10^{-109}

The results are not misprints! When the step size is .1, the computed solution values are perplexingly large, and appear to represent an exponentially growing solution — the complete opposite of the rapidly decaying true solution. Reducing the step size beyond a critical threshold suddenly transforms the numerical solution to an exponentially decaying function. Only the fourth order RK4 method with step size $h = .001$ — and hence a total of $1,000$ steps — does a reasonable job at approximating the correct value of the solution at $t = 1$.

The reader may well ask what is going on? The solution couldn't be simpler — why is it so difficult to compute it? To illustrate the basic issue, let us analyze how the Euler method handles such differential equations. Consider the initial value problem

$$\frac{du}{dt} = \lambda\,u, \qquad u(0) = 1, \tag{20.94}$$

with an exponential solution

$$u(t) = e^{\lambda\,t}.$$

As in Example 20.34, the Euler method with step size h relies on the iterative scheme

$$u_{k+1} = (1 + \lambda\, h)\, u_k, \qquad\qquad u_0 = 1,$$

with solution

$$u_k = (1 + \lambda\, h)^k. \tag{20.95}$$

If $\lambda > 0$, the exact solution, $e^{\lambda\, t}$, is exponentially growing. Since $1 + \lambda\, h > 1$, the numerical iterates are also growing, albeit at a somewhat slower rate. In this case, there is no inherent surprise with the numerical approximation procedure — in the short run it gives fairly accurate results, but eventually trails behind the exponentially growing solution. On the other hand, if $\lambda < 0$, then the exact solution $e^{\lambda\, t}$ is exponentially decaying and positive. But now, if $\lambda\, h < -2$, then $1 + \lambda\, h < -1$, and the iterates (20.95) grow exponentially fast in magnitude, with alternating signs. In this case, the numerical solution is nowhere close to the true solution, which explains the previously observed pathological behavior. If $-1 < 1 + \lambda\, h < 0$, the numerical solutions decay in magnitude, but continue to alternate between positive definite and negative values. Thus, to correctly model the qualitative features of the solution and obtain a numerically respectable approximation, we need to choose the step size h so as to ensure that $) < 1 + \lambda h$, and hence $h < -1/\lambda$ when $\lambda < 0$. For the given value $\lambda = -250$, then, we need to choose $h < \frac{1}{250} = .004$.

Thus, the numerical methods for ordinary differential equations exhibit a form of conditional stability, cf. Section 14.6. Paradoxically, the larger negative λ is — and hence the faster the solution tends to a trivial zero equilibrium — the *more* difficult and expensive the numerical integration. The ordinary differential equation (20.93) is the simplest example of what is known as a *stiff differential equation*. In general, an equation or system is stiff if it has one or more very rapidly decaying solutions. In the case of linear systems $\dot{\mathbf{u}} = A\,\mathbf{u}$, stiffness occurs whenever the coefficient matrix A has an eigenvalue with a large negative real part: Re $\lambda \ll 0$. It only takes one such solution to render the equation stiff, and ruin the numerical computation of even the well behaved solutions! Curiously, the component of the actual solution corresponding to such large negative eigenvalues is almost irrelevant, as it becomes almost instanteously tiny. However, the presence of such an eigenvalue continues to render the numerical solution to the system very difficult, even to the point of exhausting any available computing resources. Stiff equations require more sophisticated numerical procedures to integrate, and we refer the reader to [**numODE**, **HairerWanner2**] for details.

Chapter 21

The Calculus of Variations

We have already had ample encounters with Nature's propensity to optimize. Minimization principles form one of the most powerful tools for formulating mathematical models governing the equilibrium configurations of physical systems. Moreover, the design of numerical integration schemes such as the powerful finite element method are also founded upon a minimization paradigm. This chapter is devoted to the mathematical analysis of minimization principles on infinite-dimensional function spaces — a subject known as the "calculus of variations", for reasons that will be explained as soon as we present the basic ideas. Solutions to minimization problems in the calculus of variations lead to boundary value problems for ordinary and partial differential equations. Numerical solutions are primarily based upon a nonlinear version of the finite element method. The methods developed to handle such problems prove to be fundamental in many areas of mathematics, physics, engineering, and other applications.

The history of the calculus of variations is tightly interwoven with the history of calculus, and has merited the attention of a remarkable range of mathematicians, beginning with Newton, then developed as a field of mathematics in its own right by the Bernoulli family. The first major developments appeared in the work of Euler, Lagrange and Laplace. In the nineteenth century, Hamilton, Dirichlet and Hilbert are but a few of the outstanding contributors. In modern times, the calculus of variations has continued to occupy center stage in research, including major theoretical advances, along with wide-ranging applications in physics, engineering and all branches of mathematics. In this chapter, we will only have time to scratch the surface of the vast area of classical and contemporary research.

Minimization problems amenable to the methods of the calculus of variations serve to characterize the equilibrium configurations of almost all continuous physical systems, ranging through elasticity, solid and fluid mechanics, electro-magnetism, gravitation, quantum mechanics, and many, many others. Many geometrical systems, such as minimal surfaces, can be conveniently formulated as optimization problems. Moreover, numerical approximations to the equilibrium solutions of such boundary value problems are based on a nonlinear finite element approach that reduced the infinite-dimensional minimization problem to a finite-dimensional problem, to which we can apply the optimization techniques learned in Section 19.3.

We have already treated the simplest problems in the calculus of variations. As we learned in Chapters 11 and 15, minimization of a quadratic functional requires solving an associated boundary value problem for a linear differential equation. Just as the vanishing of the gradient of a function of several variables singles out the critical points, among which are the minima, both local and global, so a similar "functional gradient" will distinguish

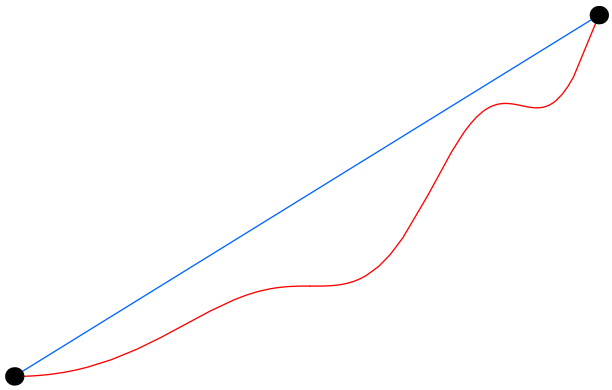


Figure 21.1. The Shortest Path is a Straight Line.

the candidate functions that might be minimizers of the functional. The finite-dimensional gradient leads to a system of algebraic equations; the functional gradient leads to a boundary value problem for a nonlinear ordinary or partial differential equation. Thus, the passage from finite to infinite dimensional nonlinear systems mirrors the transition from linear algebraic systems to boundary value problems.

21.1. Examples of Variational Problems.

The best way to introduce the subject is to introduce some concrete examples of both mathematical and practical importance. These particular minimization problems played a key role in the historical development of the calculus of variations. And they still serve as an excellent motivation for learning its basic constructions.

Minimal Curves and Geodesics

The *minimal curve problem* is to find the shortest path connecting two points. In its simplest manifestation, we are given two distinct points

$$\mathbf{a} = (a, \alpha) \qquad \text{and} \qquad \mathbf{b} = (b, \beta) \qquad \text{in the plane} \quad \mathbb{R}^2. \tag{21.1}$$

Our goal is to find the curve of shortest length connecting them. "Obviously", as you learn in childhood, the shortest path between two points is a straight line; see Figure 21.1. Mathematically, then, the minimizing curve we are after should be given as the graph of the particular affine function[†]

$$y = c\,x + d = \frac{\beta - \alpha}{b - a}\,(x - a) + \alpha \tag{21.2}$$

passing through the two points. However, this commonly accepted "fact" that (21.2) is the solution to the minimization problem is, upon closer inspection, perhaps not so immediately obvious from a rigorous mathematical standpoint.

[†] We assume that $a \neq b$, i.e., the points $\mathbf{a}, \mathbf{b}$ do not lie on a common vertical line.

Let us see how we might properly formulate the minimal curve problem. Let us assume that the minimal curve is given as the graph of a smooth function $y = u(x)$. Then, according to (A.27), the length of the curve is given by the standard arc length integral

$$J[u] = \int_a^b \sqrt{1 + (u')^2}\, dx, \tag{21.3}$$

where we abbreviate $u' = du/dx$. The function is required to satisfy the boundary conditions

$$u(a) = \alpha, \qquad u(b) = \beta, \tag{21.4}$$

in order that its graph pass through the two prescribed points (21.1). The minimal curve problem requires us to find the function $y = u(x)$ that minimizes the arc length functional (21.3) among all reasonable functions satisfying the prescribed boundary conditions. The student should pause to reflect on whether it is mathematically obvious that the affine function (21.2) is the one that minimizes the arc length integral (21.3) subject to the given boundary conditions. One of the motivating tasks of the calculus of variations, then, is to rigorously prove that our childhood intuition is indeed correct.

Indeed, the word "reasonable" *is* important. For the arc length functional to be defined, the function $u(x)$ should be at least piecewise C^1, i.e., continuous with a piecewise continuous derivative. If we allow discontinuous functions, then the straight line (21.2) does not, in most cases, give the minimizer; see Exercise ■. Moreover, continuous functions which are not piecewise C^1 may not have a well-defined length. The more seriously one thinks about these issues, the less evident the solution becomes. But, rest assured that the "obvious" solution (21.2) does indeed turn out to be the true minimizer. However, a fully rigorous mathematical proof of this fact requires a proper development of the calculus of variations machinery.

A closely related problem arises in optics. The general principle, first formulated by the seventeenth century French mathematician Pierre de Fermat, is that when a light ray moves through an optical medium, e.g., a vacuum, it travels along a path that will minimize the travel time. As always, Nature seeks the most economical solution! Let $c(x, y)$ denote the speed of light at each point in the medium[†]. Speed is equal to the time derivative of distance traveled, namely, the arc length (21.3) of the curve $y = u(x)$ traced by the light ray. Thus,

$$c(x, u(x)) = \frac{ds}{dt} = \sqrt{1 + u'(x)^2}\ \frac{dx}{dt}\ .$$

Integrating from start to finish, we conclude that the total travel time of the light ray is equal to

$$T[u] = \int_0^T dt = \int_a^b \frac{dt}{dx}\, dx = \int_a^b \frac{\sqrt{1 + u'(x)^2}}{c(x, u(x))}\, dx. \tag{21.5}$$

Fermat's Principle states that, to get from one point to another, the light ray follows the curve $y = u(x)$ that minimizes this functional. If the medium is homogeneous, then

[†] For simplicity, we only consider the two-dimensional case here.

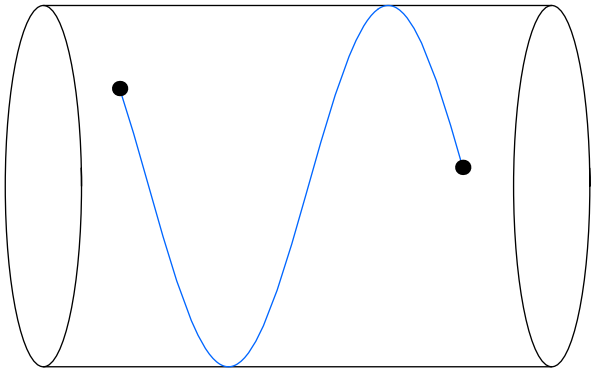


Figure 21.2. Geodesics on a Cylinder.

$c(x,y) \equiv c$ is constant, and $T[u]$ equals a multiple of the arc length functional, whose minimizers are the "obvious" straight lines. In an inhomogeneous medium, the path taken by the light ray is no longer evident, and we are in need of a systematic method for solving the minimization problem. All of the known laws of optics and lens design, governing focusing, refraction, etc., all follow as consequences of the minimization principle, [**optics**].

Another problem of a similar ilk is to construct the *geodesics* on a curved surface, meaning the curves of minimal length. In other words, given two points $\mathbf{a}, \mathbf{b}$ on a surface $S \subset \mathbb{R}^3$, we seek the curve $C \subset S$ that joins them and has the minimal possible length. For example, if S is a circular cylinder, then the geodesic curves turn out to be straight lines parallel to the center line, circles orthogonal to the center line, and spiral helices; see Figure 21.2 for an illustration. Similarly, the geodesics on a sphere are arcs of great circles; these include the circumpolar paths followed by airplanes around the globe. However, both of these claims are in need of rigorous justification.

In order to mathematically formulate the geodesic problem, we suppose, for simplicity, that our surface $S \subset \mathbb{R}^3$ is realized as the graph[†] of a function $z = F(x,y)$. We seek the geodesic curve $C \subset S$ that joins the given points

$$\mathbf{a} = (a, \alpha, F(a,\alpha)), \qquad \text{and} \qquad \mathbf{b} = (b, \beta, F(b,\beta)), \qquad \text{on the surface} \qquad S.$$

Let us assume that C can be parametrized by the x coordinate, in the form

$$y = u(x), \qquad\qquad z = F(x, u(x)).$$

In particular, this requires $a \neq b$. The length of the curve is given by the usual arc length integral (B.17), and so we must minimize the functional

$$\begin{aligned} J[u] &= \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2 + \left(\frac{dz}{dx}\right)^2}\, dx \\ &= \int_a^b \sqrt{1 + \left(\frac{du}{dx}\right)^2 + \left(\frac{\partial F}{\partial x}\,(x,u(x)) + \frac{\partial F}{\partial u}\,(x,u(x))\,\frac{du}{dx}\right)^2}\, dx, \end{aligned}$$

[†] Cylinders are not graphs, but can be placed within this framework by passing to cylindrical coordinates. Similarly, spherical surfaces are best treated in spherical coordinates.

subject to the boundary conditions

$$u(a) = \alpha, \qquad u(b) = \beta.$$

For example, the geodesics on the paraboloid

$$z = \tfrac{1}{2}\, x^2 + \tfrac{1}{2}\, y^2 \tag{21.6}$$

can be found by minimizing the functional

$$J[\,u\,] = \int_a^b \sqrt{1 + (u')^2 + (x + u\,u')^2}\, dx \tag{21.7}$$

subject to prescribed boundary conditions.

Minimal Surfaces

The minimal surface problem is a natural generalization of the minimal curve problem. In its simplest manifestation, we are given a simple closed curve $C \subset \mathbb{R}^3$. The problem is to find the surface S of least total area among all those whose boundary $\partial S = C$ coincides with the given curve. Therefore, we seek to minimize the surface area integral

$$\text{area } S = \iint_S \, dS$$

over all possible surfaces $S \subset \mathbb{R}^3$ with the prescribed boundary curve $\partial S = C$. Such an area–minimizing surface is known as a *minimal surface* for short.

Physically, if we take a wire in the shape of the curve C and dip it into soapy water, then the surface tension forces in the resulting soap film will force it to minimize surface area, and hence be a minimal surface[†]. For example, if the curve is a closed plane curve, e.g., a circle, then the minimal surface will just be the planar region enclosed by the curve. But, if the curve C twists into the third dimension, then the shape of the minimizer is by no means evident. Soap films and bubbles have been the source of much fascination, physical, æsthetical and mathematical, over the centuries. The least area problem is also known as *Plateau's Problem*, after the nineteenth century French physicist Joseph Plateau, who conducted systematic experiments. A satisfactory solution to the simplest version of the minimal surface problem was only achieved in the mid twentieth century, [**109**, **112**]. Problems arising in engineering design, architecture, and biology, such as foams, membranes and drug delivery methods, make this problem of continued contemporary importance and an active area of research.

Let us mathematically formulate the search for a minimal surface as a problem in the calculus of variations. For simplicity, we shall assume that the bounding curve C projects down to a simple closed curve $\Gamma = \partial\Omega$ that bounds an open domain $\Omega \subset \mathbb{R}^2$ in the (x, y) plane, as in Figure minsurf■. The space curve $C \subset \mathbb{R}^3$ is then given by $z = g(x, y)$

[†] More correctly, the soap film will realize a local but not necessarily global minimum for the surface area functional.

for $(x, y) \in \partial\Omega$. For reasonable curves C, we expect that the minimal surface S will be described as the graph of a function $z = u(x, y)$ parametrized by $(x, y) \in \Omega$. The surface area of such a graph is given by the double integral

$$J[u] = \iint_\Omega \sqrt{1 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2}\, dx\, dy; \tag{21.8}$$

see (B.39). To find the minimal surface, then, we seek the function $z = u(x, y)$ that minimizes the surface area integral (21.8) when subject to the Dirichlet boundary conditions

$$u(x, y) = g(x, y) \qquad \text{for} \qquad (x, y) \in \partial\Omega \tag{21.9}$$

that prescribe the boundary curve C. As we shall see, the solutions to this minimization problem satisfy a certain nonlinear second order partial differential equation, given in (21.50) below.

A simple version of the minimal surface problem, that still contains many interesting features, is to find minimal surfaces of revolution. Recall that a *surface of revolution* is obtained by revolving a plane curve about an axis, which, for definiteness, we take to be the x axis. Thus, given two points $\mathbf{a} = (a, \alpha), \mathbf{b} = (b, \beta) \in \mathbb{R}^2$, our goal is to find the curve $y = u(x)$ joining them such that the surface of revolution obtained by revolving the curve around the x-axis has the least surface area. According to Exercise ■, the area of such a surface of revolution is given by

$$J[u] = \int_a^b 2\pi\, |\, u\, |\, \sqrt{1 + (u')^2}\; dx. \tag{21.10}$$

We seek a minimizer of this integral among all functions $u(x)$ that satisfy the boundary conditions $u(a) = \alpha,\ \ u(b) = \beta$.

$$u(a) = \alpha, \qquad\qquad u(b) = \beta.$$

The minimal surface of revolution can be physically realized by stretching a soap film between two wire circles, of radius α and β, placed a distance $b - a$ apart. Symmetry considerations will require the minimizing surface to be rotationally symmetric. Interestingly, the revolutionary surface area functional (21.10) is exactly the same as the optical functional (21.5) when the light speed at a point is inversely proportional to its distance from the horizontal axis, namely $c(x, y) = 1/2\,\pi\, |\, y\, |$.

21.2. The Simplest Variational Problem.

Even the preceding, rather limited collection of examples of variational problems should already convince the reader of the practical utility of the calculus of variations. Let us now discuss the most basic analytical techniques for solving such minimization problems. We will exclusively deal with the classical approach, leaving more modern direct methods — the function space equivalent of the gradient descent method — to a more in–depth treatment of the subject, [**cvar**].

Let us concentrate on the simplest class of variational problems, in which the unknown is a continuously differentiable scalar function, and the functional to be minimized depends

upon at most its first derivative. The basic minimization problem, then, is to determine the function $y = u(x) \in \mathrm{C}^1[a,b]$ that minimizes the *objective functional*

$$J[u] = \int_a^b L(x,u,u')\,dx \tag{21.11}$$

subject to certain prescribed boundary conditions. The integrand $L(x,u,p)$ is known as the *Lagrangian* for the variational problem, in honor of Joseph–Louis Lagrange, who was one of the founders of the subject. We usually assume that $L(x,u,p)$ is a reasonably smooth function of all three of its (scalar) arguments x, u and p, which represents the derivative u'. For example, the arc length functional (21.3) has Lagrangian function $L(x,u,p) = \sqrt{1+p^2}$, whereas in the surface of revolution problem (21.10), we have $L(x,u,p) = 2\,\pi\,|\,u\,|\sqrt{1+p^2}$. (In the latter case, the points where $u = 0$ are slightly problematic, since L is not continuously differentiable there.)

In order to uniquely specify a minimizing function, we must impose suitable boundary conditions. All of the usual suspects — Dirichlet (fixed), Neumann (free), as well as mixed and periodic boundary conditions — that arose in Chapter 11 are also of interest here. In the interests of brevity, we shall concentrate on the Dirichlet boundary conditions

$$u(a) = \alpha, \qquad\qquad u(b) = \beta, \tag{21.12}$$

as these are the most common in physical problems, although some of the exercises will investigate other types.

The First Variation and the Euler–Lagrange Equation

According to Section 19.3, the (local) minimizers of a (sufficiently nice) function defined on a finite-dimensional vector space are initially characterized as critical points, where the gradient of the function vanishes. An analogous construction applies in the infinite-dimensional context treated by the calculus of variations. Every minimizer $u_\star$ of a sufficiently nice functional $J[u]$ is a "critical function", meaning that the functional gradient $\nabla J[u_\star] = 0$ vanishes at that function. Indeed, the justification of this result that was outlined in Section 19.3 continues to apply here; see, in particular, the proof of Theorem 19.42. Of course, not every critical point turns out to be a minimum. In nondegenerate situations, the classification of critical points into local minima, maxima, or saddle points, relies on the second derivative test. The functional version of the second derivative test — the second variation — is the topic of Section 21.3.

Thus, our first order of business is to learn how to compute the gradient of a functional that is defined on an infinite-dimensional function space. Adapting the general Definition 19.38 of the gradient of a function defined on an inner product space, the gradient $\nabla J[u]$ of the functional (21.11) should be defined by the same basic formula

$$\langle\, \nabla J[u]\,;v\,\rangle = \frac{d}{dt}\,J[u + t\,v]\,\Big|_{t=0}. \tag{21.13}$$

Here $v(x)$ is a function — the "direction" in which the derivative is computed. Classically, v is known as a "variation" in the function u, sometimes written $v = \delta u$, whence the term

"calculus of variations". The gradient operator on functionals is often referred to as the *variational derivative*. The inner product used in (21.13) is taken (again for simplicity) to be the standard L^2 inner product

$$\langle f ; g \rangle = \int_a^b f(x)\, g(x)\, dx \tag{21.14}$$

on function space.

Now, starting with (21.11), we have

$$J[u + t\, v] = \int_a^b L(x, u + t\, v, u' + t\, v')\, dx. \tag{21.15}$$

We need to compute the derivative of the integral with respect to t. Assuming smoothness of the integrand allows us to bring the derivative inside the integral and so, by the chain rule,

$$\begin{aligned} \frac{d}{dt}\, J[u + t\, v] &= \int_a^b \frac{d}{dt}\, L(x, u + t\, v, u' + t\, v')\, dx \\ &= \int_a^b \left[v\, \frac{\partial L}{\partial u}\, (x, u + t\, v, u' + t\, v') + v'\, \frac{\partial L}{\partial p}\, (x, u + t\, v, u' + t\, v') \right] dx. \end{aligned}$$

Therefore, setting $t = 0$ to evaluate (21.13), we find

$$\langle \nabla J[u] ; v \rangle = \int_a^b \left[v\, \frac{\partial L}{\partial u}\, (x, u, u') + v'\, \frac{\partial L}{\partial p}\, (x, u, u') \right] dx. \tag{21.16}$$

The resulting integral often referred to as the *first variation* of the functional $J[u]$. The condition $\langle \nabla J[u] ; v \rangle = 0$ for a minimizer is known as the *weak form* of the variational principle.

To obtain the strong form, the right hand side of (21.16) needs to be written as an inner product,

$$\langle \nabla J[u] ; v \rangle = \int_a^b \nabla J[u]\, v\, dx = \int_a^b h\, v\, dx$$

between some function $h(x) = \nabla J[u]$ and the variation v. The first term has this form, but the derivative v' appearing in the second term is problematic. However, as the reader of Chapter 11 already knows, the secret behind removing derivatives in an integral formula is integration by parts. If we set

$$\frac{\partial L}{\partial p}\, (x, u(x), u'(x)) \equiv r(x),$$

we can re-express the offending term as

$$\int_a^b r(x)\, v'(x)\, dx = \big[\, r(b)\, v(b) - r(a)\, v(a) \,\big] - \int_a^b r'(x)\, v(x)\, dx, \tag{21.17}$$

where — again by the chain rule —

$$r'(x) = \frac{d}{dx}\left(\frac{\partial L}{\partial p}\,(x,u,u')\right) = \frac{\partial^2 L}{\partial x\,\partial p}\,(x,u,u') + u'\,\frac{\partial^2 L}{\partial u\,\partial p}\,(x,u,u') + u''\,\frac{\partial^2 L}{\partial p^2}\,(x,u,u')\,. \tag{21.18}$$

So far we have not imposed any conditions on our variation $v(x)$. We are comparing the values of $J[u]$ only among the functions that satisfy the prescribed boundary conditions, namely

$$u(a) = \alpha, \qquad\qquad u(b) = \beta.$$

Therefore, we must make sure that the varied function $\widehat{u}(x) = u(x) + t\,v(x)$ remains within this space of functions, and so it must satisfy the same boundary conditions $\widehat{u}(a) = \alpha$, $\widehat{u}(b) = \beta$. But $u(x)$ already satisfies the boundary conditions, and so the variation $v(x)$ must satisfy the corresponding homogeneous boundary conditions

$$v(a) = 0, \qquad\qquad v(b) = 0. \tag{21.19}$$

As a result, both boundary terms in our integration by parts formula (21.17) vanish, and we can write (21.16) as

$$\langle\, \nabla J[u]\,;v\,\rangle = \int_a^b \nabla J[u]\,v\;dx = \int_a^b v\,\left[\,\frac{\partial L}{\partial u}\,(x,u,u') - \frac{d}{dx}\left(\frac{\partial L}{\partial p}\,(x,u,u')\right)\right]\,dx.$$

We conclude that

$$\nabla J[u] = \frac{\partial L}{\partial u}\,(x,u,u') - \frac{d}{dx}\left(\frac{\partial L}{\partial p}\,(x,u,u')\right). \tag{21.20}$$

This is our explicit formula for the functional gradient or variational derivative of the functional (21.11) with Lagrangian $L(x,u,p)$. Note that the gradient $\nabla J[u]$ of a functional is a *function*.

The *critical functions* $u(x)$ — which include all local minimizers — are, by definition, where the functional gradient vanishes: $\nabla J[u] = 0$. Thus, $u(x)$ must satisfy

$$\frac{\partial L}{\partial u}\,(x,u,u') - \frac{d}{dx}\,\frac{\partial L}{\partial p}\,(x,u,u') = 0. \tag{21.21}$$

In view of (21.18), we see that (21.21) is, in fact, a second order ordinary differential equation,

$$E(x,u,u',u'') = \frac{\partial L}{\partial u}\,(x,u,u') - \frac{\partial^2 L}{\partial x\,\partial p}\,(x,u,u') - u'\,\frac{\partial^2 L}{\partial u\,\partial p}\,(x,u,u') - u''\,\frac{\partial^2 L}{\partial p^2}\,(x,u,u') = 0,$$

known as the *Euler–Lagrange equation* associated with the variational problem (21.11). Any solution to the Euler–Lagrange equation that is subject to the assumed boundary conditions forms a critical point for the functional, and hence is a potential candidate for the desired minimizing function. And, in many cases, the Euler–Lagrange equation suffices to characterize the desired minimizer without further ado.

Theorem 21.1. *Suppose the Lagrangian function is at least twice continuously differentiable*: $L(x, u, p) \in \mathrm{C}^2$. *Then any* C^2 *minimizer* $u(x)$ *to the corresponding functional* $J[u] = \displaystyle\int_a^b L(x, u, u')\, dx$ *must satisfy the associated Euler–Lagrange equation* (21.21).

Let us now investigate what the Euler–Lagrange equation tells us about the examples of variational problems presented at the beginning of this section. One word of warning: there do exist seemingly reasonable functionals whose minimizers are not, in fact, C^2, and hence do not solve the Euler–Lagrange equation; see [**14**] for examples. Fortunately, in the problems we usually consider, such pathologies do not appear.

Curves of Shortest Length

Consider the problem of finding the curve of shortest length connecting two points $\mathbf{a} = (a, \alpha), \mathbf{b} = (b, \beta) \in \mathbb{R}^2$ in the plane. As we saw in Section 21.2, this requires minimizing the arc length integral

$$J[u] = \int_a^b \sqrt{1 + (u')^2}\, dx \qquad \text{with Lagrangian} \qquad L(x, u, p) = \sqrt{1 + p^2}\,.$$

Since

$$\frac{\partial L}{\partial u} = 0, \qquad \frac{\partial L}{\partial p} = \frac{p}{\sqrt{1 + p^2}}\,,$$

the Euler–Lagrange equation (21.21) in this case takes the form

$$0 \;=\; -\,\frac{d}{dx}\,\frac{u'}{\sqrt{1 + (u')^2}} \;=\; -\,\frac{u''}{(1 + (u')^2)^{3/2}}.$$

Since the denominator does not vanish, the Euler–Lagrange equation reduces to the simplest second order ordinary differential equation

$$u'' = 0. \tag{21.22}$$

All solutions to the Euler–Lagrange equation are affine functions, $u = c\,x + d$, whose graphs are straight lines. Since our solution must also satisfy the boundary conditions $\alpha = u(a)$, $\beta = u(b)$, the only critical function, and hence the sole candidate to be a minimizer, is the unique straight line

$$y = \frac{\beta - \alpha}{b - a}\,(x - a) + \alpha \tag{21.23}$$

passing through the two points. Thus, the Euler–Lagrange equation helps to reconfirm our intuition that straight lines minimize distance.

Be that as it may, the fact that a function satisfies the Euler–Lagrange equation and the boundary conditions merely gives it the status of a candidate for minimizing the variational problem. By the same token, a critical function is also a candidate for *maximizing* the variational problem, too. The nature of the critical functions can only be distinguished by a second derivative test, which requires further work. Of course, for the present problem, we "know" that a straight line cannot maximize distance, and must be

the minimizer. Nevertheless, the reader should have a little nagging doubt that we have completely solved the minimum distance problem ...

Minimal Surface of Revolution

Consider next the problem of finding the curve connecting two points having a surface of revolution of minimal surface area. For simplicity, we assume that the curve is given by the graph of a *non-negative* function $y = u(x) \geq 0$. According to (21.10), the required curve will minimize the functional

$$J[u] = \int_a^b u \sqrt{1 + (u')^2}\, dx, \qquad \text{with Lagrangian} \qquad L(x, u, p) = u \sqrt{1 + p^2}\,, \tag{21.24}$$

where we have dropped an irrelevant factor of 2π and used our positivity assumption to omit the absolute value on u in the integrand. Since

$$\frac{\partial L}{\partial u} = \sqrt{1 + p^2}\,, \qquad\qquad \frac{\partial L}{\partial p} = \frac{u\, p}{\sqrt{1 + p^2}}\,,$$

the Euler–Lagrange equation (21.21) is

$$\sqrt{1 + (u')^2} \;-\; \frac{d}{dx} \frac{u\, u'}{\sqrt{1 + (u')^2}} \;=\; \frac{1 + (u')^2 - u\, u''}{(1 + (u')^2)^{3/2}} = 0. \tag{21.25}$$

Therefore, to find the critical functions, we need to solve a nonlinear second order ordinary differential equation — and not one in a familiar form.

Fortunately, there is a little trick[†] we can use to find the solution. If we multiply by u', then we can rewrite the result as an exact derivative

$$u' \left(\frac{1 + (u')^2 - u\, u''}{(1 + (u')^2)^{3/2}} \right) \;=\; \frac{d}{dx} \frac{u}{\sqrt{1 + (u')^2}} = 0.$$

We conclude that

$$\frac{u}{\sqrt{1 + (u')^2}} = c, \tag{21.26}$$

where c is a constant of integration. The left hand side of (21.26), being constant on the entire solution, is a *first integral* for the differential equation, cf. Definition 20.25. The resulting equation is an implicit form of an autonomous first order differential equation. Solving for

$$\frac{du}{dx} = u' = \frac{\sqrt{u^2 - c^2}}{c}$$

leads to an autonomous first order ordinary differential equation, which we can immediately solve:

$$\int \frac{c\, du}{\sqrt{u^2 - c^2}} = x + \delta,$$

[†] Actually, as with many tricks, this is really an indication that something profound is going on. Noether's Theorem, a result of fundamental importance in modern physics that relates symmetries and conservation laws, [**64**, **117**], underlies the integration method. See also Exercise ■.

where δ is a constant of integration. According to Exercise ■, the most useful form of the integral is in terms of the inverse to the hyperbolic function $\cosh z = \frac{1}{2}(e^z + e^{-z})$, whereby

$$\cosh^{-1} \frac{u}{c} = x + \delta, \qquad \text{and hence} \qquad u = c \cosh\left(\frac{x + \delta}{c} \right). \tag{21.27}$$

In this manner, we have produced the general solution to the Euler–Lagrange equation (21.25). Any solution that also satisfies the boundary conditions provides a critical function for the surface area functional (21.24), and hence is a candidate for the minimizer.

The curve prescribed by the graph of a hyperbolic cosine function (21.27) is known as a *catenary*. It is *not* a parabola, even though to the untrained eye it looks similar. Interestingly, the catenary is the same profile as a hanging chain. Owing to their minimizing properties, catenaries are quite common in engineering design — for instance the cables in a suspension bridge such as the Golden Gate Bridge are catenaries, as is the arch in St. Louis.

So far, we have not taken into account the boundary conditions $u(a) = \alpha$ and $u(b) = \beta$. It turns out that there are three distinct possibilities, depending upon the configuration of the boundary points:

(a) There is precisely one value of the two integration constants c, δ that satisfies the two boundary conditions. In this case, it can be proved that this catenary is the unique curve that minimizes the area of its associated surface of revolution.

(b) There are two different possible values of c, δ that satisfy the boundary conditions. In this case, one of these is the minimizer, and the other is a spurious solution — one that corresponds to a saddle point for the functional.

(c) There are *no* values of c, δ that allow (21.27) to satisfy the two boundary conditions. This occurs when the two boundary points $\mathbf{a}, \mathbf{b}$ are relatively far apart. In this configuration, the physical soap film spanning the two circular wires breaks apart into two circular disks, and this defines the minimizer for the problem, i.e., there is no surface of revolution that has a smaller surface area than the two disks. (In the first two cases, this is not valid; the minimizing catenary has a smaller surface area than the two disks.) However, the "function"[†] that minimizes this configuration consists of two vertical lines from $\mathbf{a}$ and $\mathbf{b}$ to the x axis along with the portion of the axis lying between them. We can approximate this function by a sequence of genuine functions that give progressively smaller and smaller values to the surface area functional (21.10), but the actual minimum is not attained among the class of (smooth) functions.

Thus, even in such a reasonably simple example, a number of the subtle complications arising in the calculus of variations can already be seen. Lack of space precludes a more detailed development of these ideas here, and we refer the interested reader to more specialized books devoted to the calculus of variations, including [**39**, **64**].

[†] Here "function" must be taken in a very broad sense, as this situation does not even correspond to a generalized function!

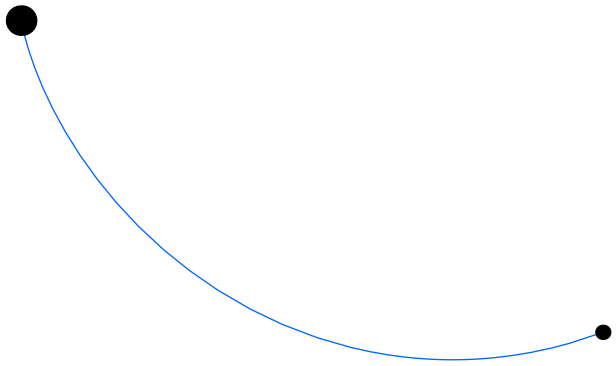


Figure 21.3. The Brachistrochrone Problem.

The Brachistochrone Problem

The most famous classical variational problem is the so-called *brachistochrone problem*. The word "brachistochrone" means "minimal time" in Latin. An experimenter lets a bead slide down a wire that connects two prescribed points. The goal is to shape the wire in such a way that, starting from rest, the bead slides from one end to the other in minimal time. Naïve guesses for the wire's shape, including a straight line, a parabola, and a circular arc, are wrong. One can do better through a careful analysis of the associated variational problem. The brachistochrone problem was originally posed by Johann Bernoulli in 1696, and served as an inspiration for much of the subsequent development of the subject.

We take the starting point of the bead at the origin: $\mathbf{a} = (0, 0)$. The wire will bend downwards, and to avoid annoying minus signs in the subsequent formulae, we take the vertical y axis to point downwards, so the wire has the shape given by the graph of $y = u(x) > 0$. The end point $\mathbf{b} = (b, \beta)$ is assumed to lie below and to the right, and so $b > 0$ and $\beta > 0$; the set-up is sketched in Figure 21.3. The first step is to find the formula for the transit time of the bead sliding along the wire. Arguing as in our derivation of the optics functional (21.5), if v denotes the speed of descent of the bead, then the total travel time is

$$T[u] = \int_0^b \frac{\sqrt{1 + (u')^2}}{v}\, dx. \tag{21.28}$$

We shall use conservation of energy to determine a formula for the speed v as a function of the position along the wire.

The kinetic energy of the bead is $\frac{1}{2} m v^2$, where m is its mass and $v \geq 0$ its speed of descent. On the other hand, due to our sign convention, the potential energy of the bead when it is at height y is $-mgy$, where m is its mass and g the gravitational force, and we take the initial height $y = 0$ as the zero potential energy level. The bead is initially at rest, with 0 kinetic energy and 0 potential energy. Assuming that frictional forces are negligible, conservation of energy implies that

$$0 = \tfrac{1}{2} m v^2 - mgy.$$

We can solve this equation to determine the bead's speed as a function of its height:

$$v = \sqrt{2gy}\,. \tag{21.29}$$

Substituting this expression into (21.28), we conclude that the shape $y = u(x)$ of the wire is obtained by minimizing the functional

$$T[u] = \int_0^b \sqrt{\frac{1 + (u')^2}{2\,g\,u}}\, dx. \tag{21.30}$$

The associated Lagrangian is

$$L(x, u, p) = \sqrt{\frac{1 + p^2}{u}}\,,$$

where we omit an irrelevant factor of $2g$ (or adopt physical units in which $g = \frac{1}{2}$). We compute

$$\frac{\partial L}{\partial u} \;=\; -\,\frac{\sqrt{1 + p^2}}{2\,u^{3/2}}\,, \qquad\qquad \frac{\partial L}{\partial p} \;=\; \frac{p}{\sqrt{u\,(1 + p^2)}}\,.$$

Therefore, the Euler–Lagrange equation for the brachistochrone functional (21.30) is

$$-\,\frac{\sqrt{1 + (u')^2}}{2\,u^{3/2}} \;-\; \frac{d}{dx}\,\frac{u'}{\sqrt{u\,(1 + (u')^2)}} \;=\; -\,\frac{2\,u\,u'' + (u')^2 + 1}{2\,\sqrt{u\,(1 + (u')^2)}} \;=\; 0,$$

and is equivalent to the nonlinear second order ordinary differential equation

$$2\,u\,u'' + (u')^2 + 1 = 0.$$

Rather than try to solve this differential equation directly, we note that the Lagrangian does not depend upon x, and therefore we can use the result of Exercise ■ that states that the Hamiltonian

$$H(x, u, p) = L - p\,\frac{\partial L}{\partial p} = \frac{1}{\sqrt{u(1 + p^2)}}$$

is a first integral, and hence

$$\frac{1}{\sqrt{u(1 + (u')^2)}} = k, \qquad \text{which we rewrite as} \qquad u(1 + (u')^2) = c,$$

where $c = 1/k^2$ is a constant. Solving for the derivative u' results in the first order autonomous ordinary differential equation

$$\frac{du}{dx} = \sqrt{\frac{c - u}{u}}\,.$$

This equation can be explicitly solved by separation of variables, and so, integrating from the initial point $x = u = 0$,

$$\int_0 \sqrt{\frac{u}{c - u}}\; du = x.$$

The integration can be done by use of a trigonometric substitution, namely

$$u = \tfrac{1}{2}\,c\,(1 - \cos r), \qquad \text{whereby} \qquad x = c \int_0 (1 - \cos r), r\, dr = \tfrac{1}{2}\,c(r - \sin r). \tag{21.31}$$

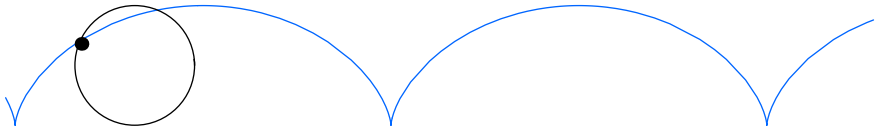


Figure 21.4. A Cycloid.

The resulting pair of equations (21.31) serve to parametrize a curve $(x(r), u(r))$ known as a *cycloid*. According to Exercise ■, a cycloid can be visualized as the curve that is traced by a point sitting on the edge of a rolling wheel. Thus, all solutions to the Euler–Lagrange equation are the cycloids, described in parametric form by (21.31). Any cycloid which satisfies the boundary conditions supplies us with a critical function, and hence a candidate for the solution to the brachistochrone minimization problem.

With a little more work, it can be proved that there is precisely one value of the integration constant c that satisfies the two boundary conditions, and, moreover, that this particular cycloid minimizes the brachistochrone functional. An example of a cycloid is plotted in Figure 21.4. Interestingly, in certain configurations, namely if $\beta < 2\,b/\pi$, the cycloid that solves the brachistrochrone problem dips below the lower endpoint **b**.

21.3. The Second Variation.

The solutions to the Euler–Lagrange boundary value problem are the critical functions for the variational principle, meaning that they cause the functional gradient to vanish. In the finite-dimensional theory, being a critical point is only a necessary condition for minimality. One must impose additional conditions, based on the second derivative of the objective function at the critical point, in order to guarantee that it is a minimum and not a maximum or saddle point. Similarly, in the calculus of variations, the solutions to the Euler–Lagrange equation may also include (local) maxima, as well as other non-extremal critical functions. To distinguish between the different possible solutions, we need to formulate a second derivative test for the objective functional on an infinite-dimensional function space. In the calculus of variations, the second derivative of a functional is known as its *second variation*, the Euler–Lagrange expression being also known as the *first variation*.

In the finite-dimensional version, the second derivative test was based on the positive definiteness of the Hessian matrix. The justification relied on a second order Taylor expansion of the objective function at the critical point. Thus, in an analogous fashion, we expand the objective functional $J[\,u\,]$ near the critical function. Consider the scalar function $g(t) = J[\,u + t\,v\,]$, where the function $v(x)$ represents a variation. The second order Taylor expansion of $g(t)$ takes the form

$$g(t) = J[\,u + t\,v\,] = J[\,u\,] + t\,K[\,u;v\,] + \tfrac{1}{2}\,t^2\,Q[\,u;v\,] + \,\cdots\,.$$

The first order terms are linear in the variation v, and given by an inner product

$$g'(0) = K[\,u;v\,] = \langle\, \nabla J[\,u\,]\,;v\,\rangle$$

between the variation and the functional gradient. In particular, if $u = u^\star$ is a critical function, then the first order terms vanish,

$$K[u^\star; v] = \langle \nabla J[u^\star]\,; v \,\rangle = 0$$

for all allowable variations v, meaning those that satisfy the homogeneous boundary conditions. Therefore, the nature of the critical function $u^\star$ — minimum, maximum, or neither — is, in most cases, determined by the second derivative terms

$$g''(0) = Q[u^\star; v].$$

As in the finite-dimensional Theorem 19.45, if u is a minimizer, then $Q[u; v] \geq 0$. Conversely, if $Q[u; v] > 0$ for $v \neq \mathbb{0}$, i.e., the second derivative terms satisfy a condition of positive definiteness, then u will be a strict local minimizer. This forms the crux of the second derivative test.

Let us explicitly evaluate the second derivative terms for the simplest variational problem (21.11). We need to expand the scalar function

$$g(t) = J[u + t\,v] = \int_a^b L(x, u + t\,v, u' + t\,v')\,dx$$

in a Taylor series around $t = 0$. The linear terms in t were already found in (21.16), and so we need to compute the quadratic terms:

$$Q[u; v] = g''(0) = \int_a^b \big[\, A\,v^2 + 2\,B\,v\,v' + C\,(v')^2 \,\big]\,dx, \tag{21.32}$$

where the coefficient functions

$$A(x) = \frac{\partial^2 L}{\partial u^2}\,(x, u, u')\,, \qquad B(x) = \frac{\partial^2 L}{\partial u\,\partial p}\,(x, u, u')\,, \qquad C(x) = \frac{\partial^2 L}{\partial p^2}\,(x, u, u')\,, \tag{21.33}$$

are found by evaluating certain second order derivatives of the Lagrangian at the critical function $u(x)$. The quadratic functional (21.32) is known as the *second variation* of the original functional $J[u]$, and plays the role of the Hessian matrix for functionals. In contrast to the first variation, it is not possible to eliminate all of the derivatives on v in the quadratic functional (21.32) through integration by parts. This causes significant complications for the analysis.

To formulate conditions that the critical function be a minimizer for the functional, we need to determine when such a quadratic functional is positive definite, meaning that $Q[u; v] > 0$ for all non-zero allowable variations $v(x) \neq \mathbb{0}$. Clearly, if the integrand is positive definite at each point, so

$$A(x)\,v^2 + 2\,B(x)\,v\,v' + C(x)\,(v')^2 > 0 \qquad \text{for all} \qquad a \leq x \leq b, \qquad (v, v') \neq \mathbf{0}, \tag{21.34}$$

then the second variation $Q[u; v]$ is also positive definite.

Example 21.2. For the arc length minimization functional (21.3), the Lagrangian is $L(x, u, p) = \sqrt{1 + p^2}$. To analyze the second variation, we first compute

$$\frac{\partial^2 L}{\partial u^2} = 0, \qquad \frac{\partial^2 L}{\partial u\, \partial p} = 0, \qquad \frac{\partial^2 L}{\partial p^2} = \frac{1}{(1 + p^2)^{3/2}}\,.$$

For the critical straight line function $u = u^\star$ given in (21.23), we evaluate at $p = u' = (\beta - \alpha)/(b - a)$, and so

$$A(x) = \frac{\partial^2 L}{\partial u^2} = 0, \qquad B(x) = \frac{\partial^2 L}{\partial u\, \partial p} = 0, \qquad C(x) = \frac{\partial^2 L}{\partial p^2} = k \equiv \frac{(b-a)^3}{\big[\,(b-a)^2 + (\beta - \alpha)^2\,\big]^{3/2}}\,.$$

Therefore, the second variation functional (21.32) is

$$Q[\,u^\star; v\,] = \int_a^b k\,(v')^2\,dx,$$

where $k > 0$ is a positive constant. Thus, $Q[\,u^\star; v\,] = 0$ vanishes if and only if v is a constant function. But the variation v is required to satisfy the homogeneous boundary conditions $v(a) = v(b) = 0$, and hence the functional is positive definite for all allowable nonzero variations. Therefore, we can finally conclude that the straight line is, indeed, a (local) minimizer for the arc length functional. We have at last justified our intuition that the shortest distance between two points is a straight line!

In general, as the following example points out, the pointwise positivity condition (21.34) is overly restrictive.

Example 21.3. Consider the quadratic functional

$$Q[\,v\,] = \int_0^1 \big[\,(v')^2 - v^2\,\big]\,dx. \tag{21.35}$$

The claim is that $Q[\,v\,] > 0$ is positive definite for all nonzero v / ⊕subject to homogeneous Dirichlet boundary conditions $v(0) = 0 = v(1)$. This result is not trivial! Indeed, the boundary conditions play an essential role, since choosing $v(x) \equiv c$ to be any constant function will produce a negative value for the functional: $Q[\,v\,] = -\,c^2$.

To prove the claim, consider the quadratic functional

$$\widetilde{Q}[\,v\,] = \int_0^1 (v' + v \tan x)^2\,dx \;\geq\; 0,$$

which is clearly positive semi-definite since the integrand is everywhere ≥ 0; moreover, the integral vanishes if and only if v satisfies the first order linear ordinary differential equation

$$v' + v \tan x = 0, \qquad \text{for all} \qquad 0 \leq x \leq 1.$$

The only solution that also satisfies boundary condition $v(0) = 0$ is the trivial one $v \equiv 0$. We conclude that $\widetilde{Q}[\,v\,] = 0$ if and only if $v \equiv 0$, and hence $\widetilde{Q} > 0$ is a positive definite quadratic functional.

Let us expand the latter functional,

$$\begin{aligned}\widetilde{Q}[v] &= \int_0^1 \big[\,(v')^2 + 2\,v\,v'\tan x + v^2\tan^2 x\,\big]\,dx\\ &= \int_0^1 \big[\,(v')^2 - v^2\,(\tan x)' + v^2\tan^2 x\,\big]\,dx = \int_0^1 \big[\,(v')^2 - v^2\,\big]\,dx = Q[v].\end{aligned}$$

In the second equality, we integrated the middle term by parts, using $(v^2)' = 2\,v\,v'$, and noting that the boundary terms vanish. Since $\widetilde{Q}[v]$ is positive definite, so is $Q[v]$, justifying the previous claim.

To see how subtle this result is, consider the almost identical quadratic functional

$$\widehat{Q}[v] = \int_0^4 \big[\,(v')^2 - v^2\,\big]\,dx. \tag{21.36}$$

The only difference is in the upper limit to the integral. A quick computation shows that the function $v(x) = x(4-x)$ satisfies the homogeneous Dirichlet boundary conditions $v(0) = 0 = v(4)$, but

$$\widehat{Q}[v] = \int_0^4 \big[\,(4-2\,x)^2 - x^2(4-x)^2\,\big]\,dx = -\tfrac{128}{5} < 0.$$

Therefore, $\widehat{Q}[v]$ is *not* positive definite. Our preceding analysis does not apply because the function $\tan x$ becomes singular at $x = \frac{1}{2}\pi$, and so the auxiliary integral $\displaystyle\int_0^4 (v' + v\tan x)^2\,dx$ does not converge.

The complete analysis of positive definiteness of quadratic functionals is quite subtle. The strange appearance of $\tan x$ in this particular example turns out to be an important clue! In the interests of brevity, let us just state without proof a fundamental theorem, and refer the interested reader to [**64**] for full details.

Theorem 21.4. *Let $A(x), B(x), C(x) \in \mathrm{C}^0[a,b]$ be continuous functions. The quadratic functional*

$$Q[v] = \int_a^b \big[\,A\,v^2 + 2\,B\,v\,v' + C\,(v')^2\,\big]\,dx$$

is positive definite, so $Q[v] > 0$ for all $v \neq 0$ satisfying the homogeneous Dirichlet boundary conditions, provided

(a) *$C(x) > 0$ for all $a \le x \le b$, and*

(b) *For any $a < c \le b$, the only solution to the associated linear Euler–Lagrange boundary value problem*

$$-(C\,w')' + (A - B')\,w = 0, \qquad w(a) = 0 = w(c), \tag{21.37}$$

is the trivial function $w(x) \equiv 0$.

Remark: A value c for which (21.37) has a nontrivial solution is known as a *conjugate point* to a. Thus, condition (b) can be restated that the variational problem has no conjugate points in the interval $[\,a,b\,]$.

Example 21.5. The quadratic functional

$$Q[\,v\,] = \int_0^b \big[\,(v')^2 - v^2\,\big]\,dx \tag{21.38}$$

has Euler–Lagrange equation

$$-w'' - w = 0.$$

The solutions $w(x) = k\sin x$ satisfy the boundary condition $w(0) = 0$. The first conjugate point occurs at $c = \pi$ where $w(\pi) = 0$. Therefore, Theorem 21.4 implies that the quadratic functional (21.38) is positive definite *provided* the upper integration limit $b < \pi$. This explains why the first quadratic functional (21.35) is positive definite, since there are no conjugate points on the interval $[\,0,1\,]$, while the second (21.36) is *not* because the first conjugate point π lies on the interval $[\,0,4\,]$.

In the case when the quadratic functional arises as the second variation of a functional (21.11), then the coefficient functions A, B, C are given in terms of the Lagrangian $L(x,u,p)$ by formulae (21.33). In this case, the first condition in Theorem 21.4 requires

$$\frac{\partial^2 L}{\partial p^2}\,(x,u,u') > 0 \tag{21.39}$$

for the minimizer $u(x)$. This is known as the *Legendre condition*. The second, *conjugate point condition* requires that the so-called *linear variational equation*

$$-\frac{d}{dx}\left(\frac{\partial^2 L}{\partial p^2}\,(x,u,u')\,\frac{dw}{dx}\right) + \left(\frac{\partial^2 L}{\partial u^2}\,(x,u,u') - \frac{d}{dx}\,\frac{\partial^2 L}{\partial u\,\partial p}\,(x,u,u')\right) w = 0 \tag{21.40}$$

has no nontrivial solutions $w(x) \not\equiv 0$ that satisfy $w(a) = 0$ and $w(c) = 0$ for $a < c \le b$.

21.4. Multi-dimensional Variational Problems.

The calculus of variations encompasses a very broad range of mathematical applications. The methods of variational analysis can be applied to an enormous variety of physical systems, in which the equilibrium configurations minimize a suitable functional — typically, the potential energy of the system. The minimizing configurations are among the critical points of the functional where its functional gradient vanishes. Following similar computational procedures as in the simple one-dimensional version, we find that the critical functions are characterized as solutions to a system of partial differential equations, called the Euler–Lagrange equations associated with the variational principle. Each solution to the boundary value problem specified by the Euler–Lagrange equations is, thus, a candidate minimizer for the variational problem. In many applications, the Euler–Lagrange equations suffice to single out the desired physical solutions, and one does not continue on to the considerably more difficult second variation.

Implementation of the variational calculus for functionals in higher dimensions will be illustrated by looking at a specific example — a first order variational problem involving a single scalar function of two variables. Thus, we consider a functional in the form

$$J[u] = \iint_\Omega L(x, y, u, u_x, u_y)\, dx\, dy, \tag{21.41}$$

of a double integral over a prescribed domain $\Omega \subset \mathbb{R}^2$. The *Lagrangian* $L(x, y, u, p, q)$ is assumed to be a sufficiently smooth function of its five arguments. Our goal is to find the function(s) $u = f(x, y)$ that minimize the given functional among all sufficiently smooth functions that satisfy a set of prescribed boundary conditions on $\partial\Omega$. The most important are our usual Dirichlet, Neumann and mixed boundary conditions. For simplicity, we concentrate on the Dirichlet boundary value problem

$$u(x, y) = g(x, y) \qquad \text{for} \qquad (x, y) \in \partial\Omega. \tag{21.42}$$

The First Variation

The basic necessary condition for an extremum (minimum or maximum) is obtained in precisely the same manner as in the one-dimensional framework. Consider the function

$$g(t) \equiv J[u + t\, v] = \iint_\Omega L(x, y, u + t\, v, u_x + t\, v_x, u_y + t\, v_y)\, dx\, dy$$

for $t \in \mathbb{R}$. The variation $v(x, y)$ is assumed to satisfy homogeneous Dirichlet boundary conditions

$$v(x, y) = 0 \qquad \text{for} \qquad (x, y) \in \partial\Omega, \tag{21.43}$$

to ensure that $u + t\, v$ satisfies the same boundary conditions (21.42) as u itself. Under these conditions, if u is a minimizer, then the scalar function $g(t)$ will have a minimum at $t = 0$, and hence $g'(0) = 0$. When computing $g'(t)$, we assume that the functions involved are sufficiently smooth so as to allow us to bring the derivative d/dt inside the integral and then apply the chain rule. At $t = 0$, the result is

$$g'(0) = \frac{d}{dt}\, J[u + t\, v]\,\Big|_{t=0} = \iint_\Omega \left(v\, \frac{\partial L}{\partial u} + v_x\, \frac{\partial L}{\partial p} + v_y\, \frac{\partial L}{\partial q} \right) dx\, dy, \tag{21.44}$$

where the derivatives of L are evaluated at x, y, u, u_x, u_y. To identify the functional gradient, we need to rewrite this integral in the form of an inner product

$$g'(0) = \langle\, \nabla J[u]\,;\, v\,\rangle = \iint_\Omega h(x, y)\, v(x, y)\, dx\, dy, \qquad \text{where} \qquad h = \nabla J[u].$$

As before, we need to remove the offending derivatives from v. In two dimensions, the requisite integration by parts formula

$$\iint_\Omega \frac{\partial v}{\partial x}\, w_1 + \frac{\partial v}{\partial y}\, w_2\, dx\, dy = \oint_{\partial\Omega} v\,(-w_2\, dx + w_1\, dy) - \iint_\Omega v \left(\frac{\partial w_1}{\partial x} + \frac{\partial w_2}{\partial y} \right) dx\, dy, \tag{21.45}$$

in which w_1, w_2 are arbitrary smooth functions, appears in (15.79). Setting $w_1 = \partial L/\partial p, w_2 = \partial L/\partial q$, we find

$$\iint_\Omega \left(v_x \frac{\partial L}{\partial p} + v_y \frac{\partial L}{\partial q} \right) dx\, dy = - \iint_\Omega v \left[\frac{\partial}{\partial x} \left(\frac{\partial L}{\partial p} \right) + \frac{\partial}{\partial y} \left(\frac{\partial L}{\partial q} \right) \right] dx\, dy,$$

where the boundary integral vanishes when $v(x, y)$ satisfies the homogeneous Dirichlet boundary conditions (21.43) that we impose on the allowable variations. Substituting this result back into (21.44), we conclude that

$$g'(0) = \iint_\Omega v \left[\frac{\partial L}{\partial u} - \frac{\partial}{\partial x} \left(\frac{\partial L}{\partial p} \right) - \frac{\partial}{\partial y} \left(\frac{\partial L}{\partial q} \right) \right] dx\, dy = 0. \tag{21.46}$$

The quantity in brackets is the desired first variation or functional gradient:

$$\nabla J[u] = \frac{\partial L}{\partial u} - \frac{\partial}{\partial x} \left(\frac{\partial L}{\partial p} \right) - \frac{\partial}{\partial y} \left(\frac{\partial L}{\partial q} \right),$$

which must vanish at a critical function. We conclude that the minimizer $u(x, y)$ must satisfy the *Euler–Lagrange equation*

$$\frac{\partial L}{\partial u}(x, y, u, u_x, u_y) \; - \; \frac{\partial}{\partial x} \left(\frac{\partial L}{\partial p}(x, y, u, u_x, u_y) \right) \; - \; \frac{\partial}{\partial y} \left(\frac{\partial L}{\partial q}(x, y, u, u_x, u_y) \right) \; = \; 0. \tag{21.47}$$

Once we explicitly evaluate the derivatives, the net result is a second order partial differential equation

$$L_u - L_{xp} - L_{yq} - u_x L_{up} - u_y L_{uq} - u_{xx} L_{pp} - 2\, u_{xy} L_{pq} - u_{yy} L_{qq}, \tag{21.48}$$

where we use subscripts to indicate derivatives of both u and L, the latter being evaluated at x, y, u, u_x, u_y. Solutions to the Euler–Lagrange equation are critical functions for the variational problem, and hence include any local and global minimizers. Determination of which solutions are geniune minima requires a further analysis of the positivity properties of the second variation, which is beyond the scope of our introductory treatment. Indeed, a complete analysis of the positive definiteness of the second variation of multi-dimensional variational problems is very complicated, and still awaits a completely satisfactory resolution!

Example 21.6. As a first elementary example, consider the Dirichlet minimization problem

$$J[u] = \iint_\Omega \tfrac{1}{2} \left(u_x^2 + u_y^2 \right) dx\, dy \tag{21.49}$$

that we first encountered in our analysis of the solutions to the Laplace equation (15.91). In this case, the associated Lagrangian is

$$L = \tfrac{1}{2}(p^2 + q^2), \qquad \text{with} \qquad \frac{\partial L}{\partial u} = 0, \qquad \frac{\partial L}{\partial p} = p = u_x, \qquad \frac{\partial L}{\partial q} = q = u_y.$$

Therefore, the Euler–Lagrange equation (21.47) becomes

$$-\frac{\partial}{\partial x}\,(u_x) - \frac{\partial}{\partial y}\,(u_y) = -\,u_{xx} - u_{yy} = -\,\Delta u = 0,$$

which is the two-dimensional Laplace equation. Subject to the boundary conditions, the solutions, i.e., the harmonic functions, are the critical functions for the Dirichlet variational principle. This reconfirms the Dirichlet characterization of harmonic functions as minimizers of the variational principle, as stated in Theorem 15.13. However, the calculus of variations approach, as developed so far, leads to a much weaker result since it only singles out the harmonic functions as *candidates* for minimizing the Dirichlet integral; they could just as easily be maximing functions or saddle points. In the quadratic case, the direct algebraic approach is, when applicable, the more powerful, since it assures us that the solutions to the Laplace equation really do minimize the integral among the space of functions satisfying the appropriate boundary conditions. However, the direct method is restricted to quadratic variational problems, whose Euler–Lagrange equations are linear partial differential equations. In nonlinear cases, one really does need to utilize the full power of the variational machinery.

Example 21.7. Let us derive the Euler–Lagrange equation for the minimal surface problem. From (21.8), the surface area integral

$$J[u] = \iint_\Omega \sqrt{1 + u_x^2 + u_y^2}\; dx\, dy \qquad \text{has Lagrangian} \qquad L = \sqrt{1 + p^2 + q^2}\,.$$

Note that

$$\frac{\partial L}{\partial u} = 0, \qquad \frac{\partial L}{\partial p} = \frac{p}{\sqrt{1 + p^2 + q^2}}\,, \qquad \frac{\partial L}{\partial q} = \frac{q}{\sqrt{1 + p^2 + q^2}}\,.$$

Therefore, replacing $p \to u_x$ and $q \to u_y$ and then evaluating the derivatives, the Euler–Lagrange equation (21.47) becomes

$$-\frac{\partial}{\partial x}\,\frac{u_x}{\sqrt{1 + u_x^2 + u_y^2}} - \frac{\partial}{\partial y}\,\frac{u_y}{\sqrt{1 + u_x^2 + u_y^2}} = \frac{-\,(1 + u_y^2)u_{xx} + 2u_x u_y u_{xy} - (1 + u_x^2)u_{yy}}{(1 + u_x^2 + u_y^2)^{3/2}} = 0.$$

Thus, a surface described by the graph of a function $u = f(x, y)$ is a candidate for minimizing surface area provided it satisfies the *minimal surface equation*

$$(1 + u_y^2)\, u_{xx} - 2\, u_x\, u_y\, u_{xy} + (1 + u_x^2)\, u_{yy} = 0. \tag{21.50}$$

Thus, we are confronted with a complicated, nonlinear, second order partial differential equation, which has been the focus of some of the most sophisticated and deep analysis over the preceding two centuries, with significant progress on understanding its solution only within the past 70 years. We have not developed the sophisticated analytical and numerical techniques that are required to have anything of substance to say about its solutions here, and will refer the interested reader to the advanced texts [**109**, **112**].

Example 21.8. The small deformations of an elastic body $\Omega \subset \mathbb{R}^n$ are described by the *displacement* field, $\mathbf{u}\colon \Omega \to \mathbb{R}^n$. Each material point $\mathbf{x} \in \Omega$ in the undeformed body will move to a new position $\mathbf{x} + \mathbf{u}(\mathbf{x})$ in the deformed body $\widetilde{\Omega} = \{\, \mathbf{x} + \mathbf{u}(\mathbf{x}) \,|\, \mathbf{x} \in \Omega \,\}$. The one-dimensional case governs bars, beams and rods, two-dimensional bodies include thin plates and shells, while $n = 3$ for fully three-dimensional solid bodies. See [**8**, **69**] for details and physical derivations.

For small deformations, we can use a linear theory to approximate the much more complicated equations of nonlinear elasticity. The simplest case is that of an isotropic, homogeneous planar body $\Omega \subset \mathbb{R}^2$, i.e., a thin plate. The equilibrium mechanics are described by the deformation function $\mathbf{u}(\mathbf{x}) = (\, u(x,y), v(x,y)\,)^T$. A detailed physical analysis of the constitutive assumptions leads to the following minimization principle

$$\begin{aligned} J[\,u,v\,] &= \iint_\Omega \left[\, \tfrac{1}{2}\,\mu \,\|\, \nabla \mathbf{u} \,\|^2 + \tfrac{1}{2}\,(\lambda+\mu)(\nabla\cdot\mathbf{u})^2 \,\right] dx\,dy \\ &= \iint_\Omega \left[\, \left(\, \tfrac{1}{2}\,\lambda + \mu \,\right)(u_x^2 + v_y^2) + \tfrac{1}{2}\,\mu\,(u_y^2 + v_x^2) + (\lambda+\mu)\,u_x\,v_y \,\right] dx\,dy. \end{aligned} \tag{21.51}$$

The parameters λ, μ are known as the *Lamé moduli* of the material, and govern its intrinsic elastic properties. They are measured by performing suitable experiments on a sample of the material. Physically, (21.51) represents the stored (or potential) energy in the body under the prescribed displacement. Nature, as always, seeks the displacement that will minimize the total energy.

To compute the Euler–Lagrange equations, we consider the functional variation $g(t) = J[\,u + t\,f, v + t\,g\,]$, in which the individual variations f, g are arbitrary functions subject only to the given homogeneous boundary conditions. If u, v minimize J, then $g(t)$ has a minimum at $t = 0$, and so we are led to compute

$$g'(0) = \langle\, \nabla J \,;\, \mathbf{f} \,\rangle = \iint_\Omega \,(f\,\nabla_u J + g\,\nabla_v J)\,dx\,dy,$$

which we write as an inner product (using the standard L^2 inner product between vector fields) between the variation $\mathbf{f}$ and the functional gradient $\nabla J = (\,\nabla_u J, \nabla_v J\,)^T$. For the particular functional (21.51), we find

$$g'(0) = \iint_\Omega \left[\, \left(\,\lambda + 2\,\mu\,\right)(u_x\,f_x + v_y\,g_y) + \mu\,(u_y\,f_y + v_x\,g_x) + (\lambda+\mu)\,(u_x\,g_y + v_y\,f_x) \,\right] dx\,dy.$$

We use the integration by parts formula (21.45) to remove the derivatives from the variations f, g. Discarding the boundary integrals, which are used to prescribe the allowable boundary conditions, we find

$$g'(0) = -\iint_\Omega \left(\begin{array}{r} \left[\, (\lambda+2\,\mu)\,u_{xx} + \mu\,u_{yy} + (\lambda+\mu)\,v_{xy} \,\right] f + \\ + \left[\, (\lambda+\mu)\,u_{xy} + \mu\,v_{xx} + (\lambda+2\,\mu)\,v_{yy} \,\right] g \end{array} \right) dx\,dy.$$

The two terms in brackets give the two components of the functional gradient. Setting them equal to zero, we derive the second order linear system of Euler–Lagrange equations

$$(\lambda+2\,\mu)\,u_{xx} + \mu\,u_{yy} + (\lambda+\mu)\,v_{xy} = 0, \qquad (\lambda+\mu)\,u_{xy} + \mu\,v_{xx} + (\lambda+2\,\mu)\,v_{yy} = 0, \tag{21.52}$$

known as *Navier's equations*, which can be compactly written as

$$\mu\,\Delta\mathbf{u} + (\mu + \lambda)\,\nabla(\nabla \cdot \mathbf{u}) = \mathbf{0} \tag{21.53}$$

for the displacement vector $\mathbf{u} = (\,u, v\,)^T$. The solutions to are the critical displacements that, under appropraite boundary conditions, minimize the potential energy functional.

Since we are dealing with a quadratic functional, a more detailed algebraic analysis will demonstrate that the solutions to Navier's equations are the minimizers for the variational principle (21.51). Although only valid in a limited range of physical and kinematical conditions, the solutions to the planar Navier's equations and its three-dimensional counterpart are successfully used to model a wide class of elastic materials.

Chapter 22

Nonlinear Partial Differential Equations

The last topic to be touched on in this book is the vast and active contemporary research area of nonlinear partial differential equations. Leaving aside quantum mechanics, which remains a purely linear theory, most real-world physical systems, including gas dynamics, fluid mechanics, elasticity, relativity, biology, thermodynamics, and so on, are modeled by *nonlinear* partial differential equations. Attempts to survey even a tiny fraction of such a all-encompassing range of phenomena, methods, results, and mathematical developments, are necessarily doomed to failure. So we will concentrate on a handful of prototypical, but very important examples, arising in the study of nonlinear waves and heat conduction. Specific topics include shock waves, blow up, similarity solutions, and solitons. We will only be able to consider nonlinear partial differential equations modeling dynamical behavior in one (spce) dimension. The much more compliocated nonlinear systems that govern our three-dimensional dynamical universe remain on the cutting edge of contemporary research activity.

Historically, i.e., before the advent of high powered computers, relatively little was known about the extraordinary range of behavior exhibited by nonlinear partial differential equations. Most of the most basic phenomena that now drive modern-day research, such as solitons, chaos, stability, blow-up, singularities, asymptotics, etc., remained undetected, or only dimly outlined. The last fifty years has witnessed a remarkable blossoming of our understanding, due in large part to the advent of large scale computing and significant advances in numerical methods for integating nonlinear systems. Numerical experimentation suddenly exposed many unexpected phenomena, including chaos and solitons, to the light of day. New analytical methods, new mathematical theories, and new computational algorithms have precipitated this revolution in our understanding and study of nonlinear systems, an activity that continues to grow in intensity and breadth. Each leap in computing power and theoretical advances has led to yet deeper understanding of nonlinear phenomena, but also points out how far we have yet to go. To make sense of this bewildering variety of methods, equations, and results, it is essential build upon a firm foundation of, first of all, linear systems theory, and secondly, nonlinear algebraic and ordinary differential equations.

We arrange our presentation according to the order of the underlying differential equation. First order nonlinear partial differential equations govern nonlinear waves and vibrations. Such nonlinear wave motions arise in gas dynamics, water waves, elastodynamics, chemical reactions, flood waves in rivers, chromatography, traffic flow, and a range of biological and ecological systems. One of the most important nonlinear phenomena, with no linear counterpart, is the break down of solutions in finite time, resulting in the forma-

tion of discontinuous shock waves. A striking example is the supersonic boom produced by an airplane that breaks the sound barrier. As in the linear wave equation, the signals propagate along the characteristics, but in the nonlinear case the characteristics can cross each other, indicating the onset of a shock wave.

Second order partial differential equations govern nonlinear diffusion processes, including heat flow and population dynamics. The simplest and most important equation, known as Burgers' equation, can, surprisingly, be linearized by transforming it to the heat equation. This accident provides an essential glimpse into the world of nonlinear diffusion processes. As we discover, as the diffusion or viscosity tends to zero, the solutions to burgers' equation tend to the shock waves solutions to the first order dispersionless limiting equation.

Third order partial differential equations arise in the study of dispersive wave motion, including water waves, plasma waves and others. We first treat the linear dispersive model, contrasting it with the hyperbolic models we encountered earlier in this book. The distinction between group and wave velocity — seen when waves propagate over water — is exposed. Finally, we introduce the remarkable Korteweg–deVries equation, which serves as a model for nonlinear water waves. Despite being nonlinear, it supports stable localized traveling wave solutions, known as *solitons*, that even maintain their shape under collisions. The Korteweg–deVries equation is an example of an integrable system, since it can be solved by an associated linear problem.

22.1. Nonlinear Waves and Shocks.

Before attempting to tackle any nonlinear partial differential equations, we should carefully review the solution to the simplest linear first order partial differential equation — the one-way or *unidirectional wave equation*

$$u_t + c\, u_x = 0. \tag{22.1}$$

First, assume that the *wave velocity* c is constant. According to Proposition 14.7, a solution $u(t, x)$ to this partial differential equation is constant along the characteristic lines of slope

$$\frac{dx}{dt} = c, \qquad \text{namely} \qquad x - ct = \text{constant} \tag{22.2}$$

As a consequence, the solutions are all of the form

$$u = p(x - ct)$$

where $p(\xi)$ is an arbitrary function of the *characteristic variable* $\xi = x - ct$. To a stationary observer, the solution is a wave of unchanging form moving at velocity c. The case $c > 0$ corresponds to a wave that translates to the right, as illustrated in Figure 22.1.

Slightly more complicated, but still linear, is the wave equation

$$u_t + c(x)\, u_x = 0, \tag{22.3}$$

where the variable wave velocity $c(x)$ depends upon the position of the wave. This equation models unidirectional waves propagating through a non-uniform, but static medium. Generalizing the constant coefficient construction (22.2), we define the *characteristic curves*

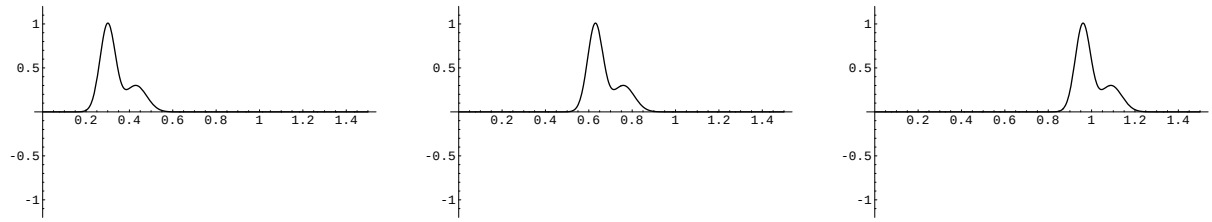


Figure 22.1. Traveling Wave of Constant Form.

for the wave equation (22.3) to be the solutions to the autonomous ordinary differential equation

$$\frac{dx}{dt} = c(x). \tag{22.4}$$

Thus, unlike the constant velocity version, the characteristics are no longer straight lines. Nevertheless, the preceding observation retains its validity.

Proposition 22.1. *The solutions to the linear wave equation* (22.3) *are constant on the characteristic curves.*

Proof: Let $x(t)$ be a characteristic curve, i.e., a solution to (22.4), parametrized by the time t. The value of a solution $u(t,x)$ of the wave equation at the point $(t,x(t))$ on the given characteristic curve is $h(t) = u(t,x(t))$. Our goal is to prove that $h(t)$ is a constant function of t, and, as usual, this is done by proving that its derivative is identically zero. To differentiate $h(t)$, we invoke the chain rule:

$$\frac{dh}{dt} = \frac{d}{dt}\,u(t,x(t)) = \frac{\partial u}{\partial t}(t,x(t)) + \frac{dx}{dt}\,\frac{\partial u}{\partial x}(t,x(t)) = \frac{\partial u}{\partial t}(t,x(t)) + c(x(t))\,\frac{\partial u}{\partial x}(t,x(t)) = 0.$$

We replaced dx/dt by $c(x)$ since we are assuming that $x(t)$ is a characteristic curve, and hence satisfies (22.4). The final combination of derivatives is zero whenever u solves the wave equation (22.1). Therefore, $h(t) = u(t,x(t))$ is constant. *Q.E.D.*

Since the characteristic curve differential equation (22.4) is autonomous, it can be immediately solved:

$$h(x) = \int \frac{dx}{c(x)} = t + \delta, \tag{22.5}$$

where δ is the constant of integration. Therefore, the characteristic curves are defined by the formula $x = g(t+\delta)$, where $g = h^{-1}$ is the inverse function.

Any function which is constant along the curves defined by (22.5) must be a function of the *characteristic variable*[†] $\xi = h(x) - t$. As a consequence, Proposition 22.1 implies that we can write the solution to the wave equation in the form

$$u(t,x) = p(h(x) - t), \tag{22.6}$$

[†] The present definition of characteristic variable has changed slightly from the constant velocity case.

where $p(\xi)$ is an arbitrary function of the characteristic variable. It is easy to verify directly that (22.6) that, provided $h(x)$ is defined by (22.5), $u(t,x)$ solves the partial differential equation (22.3) for *any* choice of function $p(\xi)$.

To find the solution that satsifies the given initial conditions

$$u(0,x) = f(x) \tag{22.7}$$

we merely substitute the solution (22.6), leading to an implicit equation

$$p(h(x)) = f(x), \qquad \text{and hence} \qquad p(\xi) = f \circ h^{-1}(\xi) = f[g(\xi)].$$

Graphically, the solution must be constant along each characteristic curve. Therefore, to find the value of the solution $u(t,x)$ at a given point, we look at the characteristic curve passing through (t,x). If this curve intersects the x axis at the point $(0,y)$, then $u(t,x) = u(0,y) = f(y)$. The construction is illustrated in Figure ccx.

Example 22.2. Consider the equation

$$\frac{\partial u}{\partial t} + \frac{1}{x^2+1}\,\frac{\partial u}{\partial x} = 0. \tag{22.8}$$

According to (22.4), the characteristic curves are the solutions to the first order ordinary differential equation

$$\frac{dx}{dt} = \frac{1}{x^2+1}\,.$$

Integrating, we find

$$\int (x^2+1)\,dx = \tfrac{1}{3}\,x^3 + x = t + \delta,$$

and the resulting characteristic curves are plotted in Figure wcxx.

The general solution to the equation takes the form

$$u = p\left(\tfrac{1}{3}\,x^3 + x - t\right),$$

where $p(\xi)$ is an arbitrary function of the characteristic variable $\xi = \frac{1}{3}\,x^3 + x - t$. A typical solution, corresponding to initial data $u(t,0) =$ is plotted in Figure wcxx. The fact that the characteristic curves are not straight means that, although the wave remains constant along each individual curve, a stationary observer will witness a dynamically changing profile as the wave moves along. Waves speed up as they arrive at the origin, and then slow down once they pass by. As a result, we observe the wave spreading out as it approaches the origin, and then contracting as it moves off to the right.

Example 22.3. Consider the equation

$$u_t - x\,u_x = 0. \tag{22.9}$$

The characteristic curves are the solutions to

$$\frac{dx}{dt} = -x, \qquad \text{and so} \qquad x\,e^t = c, \tag{22.10}$$

where c is the constant of integration. The solution takes the form

$$u = p(x\,e^t), \tag{22.11}$$

where $p(\xi)$ is an arbitrary function. Therefore, for initial data

$$u(0,x) = f(x) \qquad \text{the solution is} \qquad u = f(x\,e^t).$$

For example, the solution

$$u(t,x) = \frac{1}{(x\,e^t)^2 + 1} = \frac{e^{-2\,t}}{x^2 + e^{-2\,t}}$$

corresponding to initial data $u(t,0) = f(x) = (x^2+1)^{-1}$ is plotted in Figure wavecx■. Note that since the characteristic curves all converge on the t axis, the solution becomes more and more concentrated at the origin.

A Nonlinear Wave Equation

One of the simplest possible nonlinear partial differential equations is the *nonlinear wave equation*

$$u_t + u\,u_x = 0. \tag{22.12}$$

first systematically studied by Riemann†. Since it appears in so many applications, this equation goes under a variety of names in the literature, including the Riemann, inviscid Burgers', dispersonless Korteweg–deVries, and so on.

The equation (22.12) has the form of a unidirectional wave equation $u_t + c\,u_x = 0$ in which the wave velocity $c = u$ depends, not on the position x, but rather on the magnitude of the wave. Larger waves move faster, and overtake smaller waves. Waves of depression, where $u < 0$, move in the reverse direction.

Fortunately, the method of characteristics that was developed for linear wave equations also works in the present nonlinear situation and leads to a complete solution to the equation. Mimicking our previous construction, (22.4), let us define the *characteristic curves* of the nonlinear wave equation (22.12) by the formula

$$\frac{dx}{dt} = u(t,x). \tag{22.13}$$

In this case, the characteristics depend upon the solution, and so it appears that we will be not able to specify the characteristics until we know the solution $u(t,x)$. Be that as it may, the solution $u(t,x)$ remains constant along its characteristic curves, and this observation will allow us to pin both down.

† In addition to his contributions to complex analysis, partial differential equations and number theory, Riemann also was the inventor of Riemannian geometry, which proved absolutely essential for Einstein's theory of general relativity some 70 years later!

First, to prove this claim, assume that $x = x(t)$ parametrizes a characteristic curve. We need to show that $h(t) = u(t, x(t))$ is constant along the curve. As before, we differentiate. Using the chain rule and (22.13), we deduce that

$$\frac{dh}{dt} = \frac{d}{dt}\, u(t, x(t)) = \frac{\partial u}{\partial t}\,(t, x(t)) + \frac{dx}{dt}\,\frac{\partial u}{\partial x}(t, x(t)) = \frac{\partial u}{\partial t}\,(t, x(t)) + u(t, x(t))\,\frac{\partial u}{\partial x}(t, x(t)) = 0.$$

The final expression vanishes because u is assumed to solve the wave equation (22.12) at all values of (t, x). Since the derivative of $h(t) = u(t, x(t))$ is zero, this quantity must be a constant, as stated.

Now, since the solution $u(t, x(t))$ is constant on the characteristic curve, the right hand side of its defining equation (22.13) is a constant. Therefore, the derivative dx/dt is constant, and the characteristic curve is a *straight line*! Consequently, each characteristic curve

$$x = u\,t + \delta,$$

is a straight line of slope u, which we call the *characteristic slope* of the line. The value of the solution on each characteristic line is its characteristic slope. The larger u is, the steeper the characteristic line, and the faster the wave moves.

The characteristic variable $\xi = x - t\,u$ depends upon the solution, which can now be written in implicit form

$$u = f(x - t\,u), \tag{22.14}$$

where $f(\xi)$ is an arbitrary function of the characteristic variable. For example, if $f\xi) = \alpha\,\xi + \beta$ is an affine function, then

$$u = \alpha(x - t\,u) + \beta, \qquad \text{and hence} \qquad u(t, x) = \frac{\alpha\,x + \beta}{1 + \alpha\,t}\,. \tag{22.15}$$

If $\alpha > 0$, this represents a straight line solution that gradually flattens out as $t \to \infty$. On the other hand, if $\alpha < 0$, the line rapidly steepens to vertical as $t \to t_\star = -1/\alpha$ when the solution blows up.

To construct a solution $u(t, x)$ to the initial value problem

$$u(0, x) = f(x) \tag{22.16}$$

we note that, at $t = 0$, the implicit solution formula formula (22.14) reduces to $u(0, x) = f(x)$, and hence the function f coincides with the initial data! However, because (22.14) defines $u(t, x)$ implicitly, it is not clear

(*a*) whether it can be solved to give a well-defined value for the solution $u(t, x)$, and,

(*b*) if so, what are the solution's qualitative features and dynamical behavior.

A more instructive and revealing strategy is based on the following geometrical construction. Through each point $(y, 0)$ on the x axis, draw the characteristic line $x = t\,f(y) + y$ whose slope $f(y)$ equals the value of the initial data at that point. According to the preceding argument, the solution will have the same value, $u = f(y)$, on the entire characteristic line $x = t\,f(y) + y$. For example, if $f(y) = y$, then $u(t, x) = y$ whenever $x = t\,y + y$. Eliminating y, we recover the previous solution $u(t, x) = x/(t + 1)$.

Figure 22.2. Rarefaction Wave.

Now, the problem with this construction is immediately apparent from the illustrative Figure Rsol. Any characteristic lines which are not parallel must cross each other. The value of the solution is supposed to be equal to the slope of the characteristic line, and so at the point of crossing, the solution is supposed to have two *different* values, one corresponding to each line. Something is clearly amiss, and we need to understand the construction and the resulting solution in more depth.

There are three basic scenarios. The first, trivial case is when all the characteristic lines are parallel and so the difficulty does not arise. In this case, the characteristic lines have the same slope, say c, which means that $u = c$ has the same value on each one. Therefore, $u \equiv c$ is a trivial constant solution.

The next simplest case occurs when the initial data $f(x)$ is everywhere increasing, so $f(x) \leq f(y)$ whenever $x \leq y$, which is assured if the derivative $f'(x) \geq 0$ is never negative. In this case, as in Figure chli, the characteristic lines emanting from the x axis fan out into the right half plane, and so never cross each other at any $t \geq 0$. Each point (t, x) for $t \geq 0$ lies on a unique characteristic line, and the value of the solution at (t, x) is equal to the slope of the line. Consequently, the solution is well-defined at all future times. Physically, such a solultion represents a wave of *rarefaction*, which gradually spreads out as time progresses. A typical example is plotted in Figure 22.2, corresponding to initial data $u(0, x) = \tan^{-1} x + \frac{\pi}{2}$.

The more interesting case is when $f'(x) < 0$. Now the characteristic lines starting at $t = 0$ cross at some future time. If a point (t, x) lies on two or more characteristic lines of different slopes, the value of the solution $u(t, x)$, which should equal the characteristic slope, is no longer uniquely determined. Although one might be tempted to deal with such multiply-valued solutions in a purely mathematical framework, from a physical standpoint this is unacceptable. The solution u is supposed to represent a physical quantity, e.g., density, velocity, pressure, etc., and must therefore have a unique value at each point. The mathematical model has broken down, and fails to agree with the physical process.

Before confronting this difficulty, let us first, theoreticallly, understand what happens when we continue the solution as a multiply-valued function. To be specific, consider the initial data

$$u(0, x) = \tfrac{\pi}{2} - \tan^{-1} x, \tag{22.17}$$

plotted in the first figure in Figure shu. In the companion picture we plot the characteristic lines for this particular initial data. Initially, the characteristic lines do not cross, and the solution is a well-defined single-valued functions. However, there is a critical time $t = t_\star > 0$ when the first two lines cross each other at a particular point $(t_\star, x_\star)$. After the

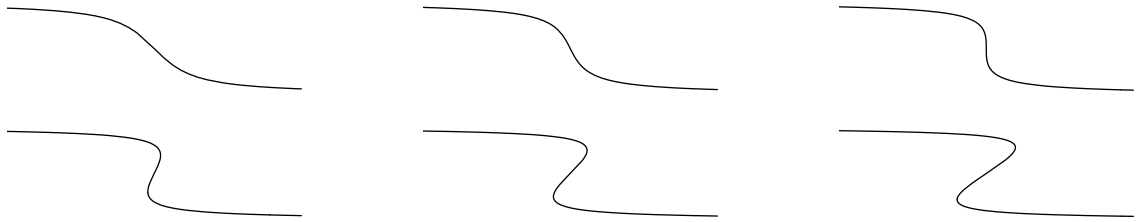


Figure 22.3. Multiply Valued Solution.

critical time, the (t, x) plane contains a wedge-shaped region, each point of which lies on the intersection of three different characteristic lines with different slopes; at such points, the solution will achieve three different values. Outside the wedge, the points only belong to a single characteristic line, and the solution remains single valued there. (The boundary of the wedge is special, consisting of points where only two characteristic lines cross.)

To understand what is going on, look now at the sequence of pictures of the solution at successive times in Figure 22.3. Since the initial data is positive, $f(x) > 0$, all the characteristic slopes are positive. As a consequence, all the points on the solution curve will move to the right, at a speed equal to their height. Since the initial data is decreasing, points to the left will move faster than those to the right, and eventually overtake them. Thus, as time progresses, the solution gradually steepens. At the critical time $t_\star$ when the first two characteristic lines cross, the tangent to the solution curve $u(t_\star, x)$ has become vertical at the point $x_s tar$, and so $u_x(t, x_\star) \to \infty$ as $t \to t_\star$. After the critical time, the solution graph $u(t, x)$ for fixed $t > t_\star$ remains a continuous curve in the (x, u) plane, but no longer represents the graph of a single-valued function. The overlapping lobes correspond to points (t, x) lying in the aforementioned wedge.

The critical time can be determined from the implicit solution formula (22.14). Indeed, if we differentiate with respect to x, we find

$$\frac{\partial u}{\partial x} = \frac{\partial}{\partial x} f(x - t\,u) = f'(\xi) \left(1 - t\, \frac{\partial u}{\partial x} \right), \qquad \text{where} \qquad \xi = x - t\,u$$

is the characteristic variable, which is constant along the characteristic lines. Solving,

$$\frac{\partial u}{\partial x} = \frac{f'(\xi)}{1 + t\, f'(\xi)}\,.$$

Therefore, the slope

$$\frac{\partial u}{\partial x} \;\longrightarrow\; \infty \qquad \text{as} \qquad t \;\longrightarrow\; -\,\frac{1}{f'(\xi)}\,.$$

In other words, if the initial data has negative slope at position x, so $f'(x) < 0$, then the solution along the characteristic line emanating from the point $(x, 0)$ will break down at the time $-1/f'(x)$. As a consequence, the first of these critical times is at

$$t_\star = \min \left\{ -\,\frac{1}{f'(x)} \;\middle|\; f'(x) < 0 \right\}. \tag{22.18}$$

For instance, for the particular initial configuration (22.17) represented by the pictures,

$$f(x) = \frac{\pi}{2} - \tan^{-1} x, \qquad f'(x) = -\frac{1}{1+x^2},$$

and so the critical time is

$$t_\star = \min(1 + x^2) = 1.$$

As noted above, the triply-valued matheamtical solution is physically untenable. So what happens after the critical time $t_\star$? One needs to choose which of the three possible values should be used at each given point in the triply-valued wedge.

The mathematics is incapable of providing us with the answer, and we must reconsider the physical system that we are modeling.

The mathematics by itself incapable of telling us how to continue with this solution past the critical time at which the characteristics begin to cross. We therefore need to return to the physics underlying the partial differential equation, and ask what sort of phenomenon we are trying to model. The most instructive is to view the equation as a simple model of compressible fluid flow in a single space variable, e.g., gas in a pipe. If we push a piston down the end of a long pipe then the gas will move ahead of the piston and compress. If we push the piston too fast, the gas will compress near the piston. However, if the piston moves too rapidly the gas piles up on top of itself and a shock wave forms.

The physical assumption that underlies the specification of where the shock wave appears is known as an *entropy condition*. The simplest version, which applies to many physical systems, is an equal area rule. Draw the vertical shock line where the areas of the two lobes in the multiply valued solution are equal, as in Figure ea.

Note that this implies irreversibility of the solutions to the nonlinear wave equation. One cannot simply run time backwards and expect the shock to disappear. However, this is a different issue than the irreversibility of the heat equation, which was due to its ill-posedness in backwards time. One can run the nonlinear wave equation backwards, but this would result, typically, in the formation of a different collection of shocks.

Example 22.4. An interesting case is when the initial data has the form of a step function with a single jump discontinuity:

$$u(0,x) = f(x) = a + b\,\sigma(x) = \begin{cases} a, & x < 0, \\ b, & x > 0 \end{cases}. \tag{22.19}$$

If $a > b > 0$, then the initial data is in the form of a shock. If we use the mathematical solution by continuing along the characteristic lines, the solution at time t is multiply-valued in the region $bt < x < at$ where it assumes both values a and b as illustrated in Figure sws. If we use the equal area rule, we draw the shock line halfway along, at $x = \frac{1}{2}(a+b)t$. Therefore, the shock moves with speed $\frac{1}{2}(a+b)$ equal to one half the magnitude of the jump (and the value of the step function at the jump according to the Fourier convention). Behind the shock the solution has value a and in front the smaller value b. A graph of the characteristic lines appears in Figure swsch.

By way of contrast, if $0 < a < b$, then the characteristic lines diverge from the shock point, and the mathematical solution is not well-defined in the wedge-shaped region

$a\,t < x < b\,t$. We must decide how to connect the two regions where the solution is defined. Physical reasoning points to using an affine or striaght line to connect the two parts of the solution. Indeed, a simple modification of the solution (22.15) yields the function

$$u(t,x) = \frac{x}{t}\,,$$

which not only solves the differential equation, but also has the required values $u(t, a\,t) = a$, and $u(t, b\,t) = b$ at the edge of the wedge. Therefore, the desired solution is the *rarefaction wave*

$$u(t,x) = \begin{cases} a, & x \le a\,t, \\ x/t, & a\,t \le x \le b\,t, \\ b, & x \ge b\,t \end{cases},$$

which is graphed in Figure swsr.

These two simple solutions epiomize the basic phenomenon modeled by our nonlinear wave equation — rarefaction wave where the solution is spreading out, that correspond to regions where $f'(x) > 0$, and waves of compression, where $f'(x) < 0$, where the solution contracta and eventually lead to shocks. Anyone caught in a traffic jam recognized the compresssion waves, where the traffic is bunched together and almost stationery, while the interspersed rarefaction waves represent freely moving traffic. (An intelligent drive wil recognize the rarefaction waves moving through the jam and use them to switch lanes!) The observed, and frustrating trafffic jam phenomenon is a direct result of the nonlinear wave model for traffic flow.

An entropy condition such as the equal area rule allows us to progress beyond the formation of a simple shock. As other characteristic lines cross, additional shocks form. The shocks themselves continue propagate, often at different velocities. When a fast-moving shock catches up with a slow moving shock, one must decide how to merge the shocks together to retain a physically consistent solution. At this point, the mathematical details have become too complicated for us to pursue in any more detail, and we refer the reader to [**147**] for a detailed discussion, along with applications to equations of gas dynamics, flood waves in rivers, motion of glaciers, chomotography, traffic flow and many other physical systems.

22.2. Nonlinear Diffusion.

First order partial differential equations, beginning with the simple scalar equation (22.12), and progressing through the equations of gas dynamics on to the full-blown Euler equations of fluid mechanics, model conservative wave motion. Such models fail to account for frictional and viscous effects.

When a shock wave forms, there is a breakdown in the mathematical solution to the equation. But the physical processes continue. This indicates that our assumptions governing the physical situation modeled by the partial differential equation are not complete, and are neglecting certain significant physical effects. In the case of gas dynamics, the nonlinear wave equation (22.12) does not build in any damping effects due to viscosity in the fluid. Dissipative or frictional or viscous effects are, as we know, governed by second

order differential operators. The simplest is the linear heat equation which models a broad range of dissipative phenomena, but fails to take into account nonlinear physical effects.

As in the linear heat equation, dissipative effects such as friction and viscosity are governed by second order elliptic differential operators, and hence intorduce second order terms into the wave model. In this section, we study the very simplest model that includes both nonlinar wave motion and dissipation, known as Burgers' equation.

Burgers' Equation

The simplest nonlinear diffusion equation is known as† *Burgers' equation*, and takes the form

$$u_t + \beta\, u\, u_x = \gamma\, u_{xx}. \tag{22.20}$$

The term $\gamma\, u_{xx}$ represents linear diffusion, as in the heat equation. The diffusion coefficient $\gamma > 0$ must be positive in order that the equation be well-posed. The second, nonlinear term represents a simple advection. In the inviscid limit, as the diffusion goes to zero, $\gamma \to 0$, Burgers' equation reduces to the nonlinear wave equation (22.12), which, as a result, is often referred to as the *inviscid Burgers' equation* in the literature. One can also interpret the linear term as modeling viscosity, and so Burgers' equation represents a very simplified version of the equations of viscous fluid mechanics. The higher the viscosity γ, the faster the diffusion.

We will only consider solutions $u(t,x)$ which are globally defined for all $-\infty < x < \infty$ and for times $t > t_0$ after an initial time which, for simplicity, we take as $t_0 = 0$. As with both the heat equation and the nonlinear wave equation (22.12), the solution will be specified by its initial values

$$u(0,x) = f(x), \qquad -\infty < x < \infty. \tag{22.21}$$

The initial data $f(x)$ is assumed to be reasonably smooth, say C^1, and bounded as $|\, x\,| \to \infty$.

Small solutions of Burgers' equation, $|\, u(t,x)\,| \ll 1$, will tend to act like solutions to the heat equation, since the nonlinear terms will be neglible. On the other hand, for large solutions $|\, u(t,x)\,| \gg 1$, the nonlinear terms will dominant the equation, and we expect the solution to behave like the nonlinear waves we analyzed in Section 22.1. Thus, the question naturally arises: do the solutions of Burgers' equation experience shocks, or does the diffusion have a sufficient effect to smooth out any potential discontinuities. As we will see, the latter scenario is correct. Assuming $\gamma > 0$, it can be proved, [**147**], that the initial value problem (22.20), (22.21) for Burgers' equation has a unique solution $u(t,x)$ that is smooth and exists for all positive time $t > 0$. The diffusion, no matter how small, is sufficient to prevent the formation of any shocks or other discontinuities in the solution.

A typical simple solution is plotted in Figure Burgers. The initial data is the same as in the shock wave solution for the inviscid version that we plotted in Figure swa. We

† Note that the apostrophe goes after the "s" since the equation is named after the applied mathematician J. Burgers, [**Burgers**]. It was first studied in this context by Bateman, [**16**], although it does appear in older pure mathematical texts.

take the diffusion coefficient to be small: $\gamma = .01$, and the nonlinearity $\beta = 1$. As you can see, the wave initially steepens just like its inviscid counterpart. However, at a certain point, the diffusion prevents the wave from becoming vertical and then moving into the shock regime. Instead, once the initial steepening is finished, the wave takes the form of a very sharp, but nevertheless smooth, transition, looking like a slightly smoothed-out form of the equal area shock wave solution that we found in Section 22.1.

Indeed, the profound fact is that, as the diffusion $\gamma \to 0$ becomes very small, the solutions to Burgers' equation (22.20) converge, in the inviscid limit, to the shock wave solution to (22.12) constructed by the equal area rule. This observation is in accordance with our physical intuition, that all physical systems retain a very small dissipative component, that serves to smooth out discontinuities that appear in the theoretical model that fails to take the dissipation/viscosity/damping/etc. into account. It also has very important theoretical consequences. The way to characterize the discontinuous solutions to the inviscid nonlinear wave equation is as the limit, as the viscosity goes to zero, of classical solutions to the Burgers' equation. This is known as *viscosity solution* method. If the viscous terms are as in Burgers' equation, the resulting viscosity solutions are consistent with the equal area rule for drawing the shocks. More generally, this method allows one to continue the solutions into the regimes where multiple shocks merge and interact.

The Hopf–Cole Transformation

While the Burgers' equation is a fully nonlinear partial differential equation, there is a remarkable nonlinear transformation that converts it into the linear heat equation. This result first appears in an exercise in the nineteenth century differential equations textbook by Forsyth, [**59**; vol. 6, p. 102]. Its modern rediscovery by Eberhard Hopf, [**83**], and Julian Cole, [**35**], was a milestone in the study of nonlinear partial differential equations.

One simple-minded way to convert a linear equation into a more complicated nonlinear equation is to make a nonlinear changes of variables. The resulting nonlinear equation is said to be *linearizable* since it can be linearized by inverting the change of variables. Recognizing when a nonlinear equation can, in fact, be linearized by a sujitable change of variables is a challenging problem, and the subject of much contemporary research, [**E**]. In practice, "nonlinearizing" a linear equation by a randomly chosen changes of variables rarely leads to a nonlinear equation of any interest. However, sometimes there is a luck accident, and the linearizing change of variables can make a profound impact on our understanding of the nonlinear version.

Our starting point is the linear heat equation

$$v_t = \gamma\, v_{xx}. \tag{22.22}$$

Among all the possible nonlinear changes of dependent variable†, one of the simplest that might spring to mind is an exponential function. Consider the nonlinear change of variables

$$v = e^{\sigma\, \varphi}, \tag{22.23}$$

† Nonlinear changes of the independent variables t, x alone will only lead to a linear partial differential equation, albeit with nonconstant coefficients; see Exercise ■.

where $\varphi(t,x)$ is a new function. The change of variables is valid provided $v(t,x) > 0$ is a *positive* solution to the heat equation. Fortunately, this is not hard to arrange: if the initial data $v(0,x) > 0$ is strictly positive, then the solution is positive for all $t > 0$. Physically, if the temperature in a fully insulated bar starts out everywhere above freezing, in the absence of external heat sources, it can never dip below freezing at any later time.

To find the differential equation satisfied by the new function φ, we compute the relations among their derivatives using the chain rule:

$$v_t = \sigma\,\varphi_t\,e^{\sigma\,\varphi}, \qquad v_x = \sigma\,\varphi_x\,e^{\sigma\,\varphi}, \qquad v_{xx} = \big(\,\sigma\,\varphi_{xx} + \sigma^2\,\varphi_x^2\,\big)\,e^{\sigma\,\varphi}.$$

We substitute the first and last formulae into the heat equation (22.22) and canceling a common exponential factor. We conclude that $\varphi(t,x)$ satisfies the nonlinear partial differential equation

$$\varphi_t = \gamma\,\varphi_{xx} + \gamma\,\sigma\,\varphi_x^2, \tag{22.24}$$

known as the *potential Burgers' equation*, for reasons that will soon become apparent.

The second step in the procedure is to differentiate the potential Burgers' equation with respect to x; the result is

$$\varphi_{tx} = \gamma\,\varphi_{xxx} + 2\,\gamma\,\sigma\,\varphi_x\,\varphi_{xx}. \tag{22.25}$$

If we now set

$$\frac{\partial \varphi}{\partial x} = u, \tag{22.26}$$

then the resulting partial differential equation is a form of Burgers' equation (22.20):

$$u_t = \gamma\,u_{xx} + 2\,\gamma\,\sigma\,u\,u_x.$$

Indeed, if we define $\beta = -2\,\gamma\,\sigma$, then the two equations coincide. Let us summarize the resulting *Hopf–Cole transformation*.

Theorem 22.5. *If $v(t,x) > 0$ is any positive solution to the linear heat equation $v_t = \gamma\,v_{xx}$, then*

$$u(t,x) = \frac{\partial}{\partial x}\left(-\frac{\gamma}{2\,\beta}\,\log v(t,x)\right) = -\frac{\gamma}{2\,\beta}\,\frac{v_x}{v}\,, \tag{22.27}$$

solves Burgers' equation $u_t + \beta\,u\,u_x = \gamma\,u_{xx}$.

Do all solutions to Burgers' equation arise in this way? In order to decide, we run the derivation of the transformation in reverse. The intermediate function $\varphi(t,x)$ in (22.26) can be interpreted as a *potential function* for the solution $u(t,x)$, which can be physically interpreted as a fluid velocity given as the one-dimensional gradient of its potential. The potential is only determined up to a constant, or, more accurately, up to a function of t, and so $\widetilde{\varphi}(t,x) = \varphi(t,x) + h(t)$ is an equally valid potential. Substituting the potential relation (22.26) into Burgers' equation (22.20) leads to (22.25)

$$\varphi_{tx} + \beta\,\varphi_x\,\varphi_{xx} = \gamma\,\varphi_{xxx}.$$

Integrating both sides with respect to x produces

$$\varphi_t = \gamma\,\varphi_{xx} + \gamma\,\sigma\,\varphi_x^2 + h(t),$$

for some integration "constant" $h(t)$. Replacing $\varphi = \widetilde{\varphi} + H(t)$ where $H' = h$, we find that the alternative potential $\widetilde{\varphi}$ does satsify the potential Burgers' equation, and thus comes from a positive solution to the heat equation by the exponential changes of variables. Thus, the answer is yes: every solution to Burgers' equation comes from a positive solution to the heat equation via the Hopf–Cole transformation.

Example 22.6. For example, the separable solution

$$v(t,x) = \alpha + \beta\, e^{-\omega^2 t} \cos \omega\, x$$

to the heat equation leads to the solution

$$u(t,x) = \frac{\gamma}{2\,\beta}\, \frac{\beta\,\omega\, e^{-\omega^2 t} \sin \omega\, x}{\alpha + \beta\, e^{-\omega^2 t} \cos \omega\, x}$$

to Burgers' equation; a typical examplke is plotted in Figure Bcos■. We must require that $\alpha > |\,\beta\,|$ in order that $v(t,x) > 0$ be a positive solution to the heat equation for $t \geq 0$; otherwise the solution to Burgers' equation would have singularities at the roots of u. This particular solution primarily feels the effects of the diffusivity, and goes to zero exponentially fast.

To solve the initial value problem (22.20), (22.21) we note that the initial conditions transform, via (22.27), to

$$v(0,x) = h(x) = \exp \varphi(0,x) = \exp \int_0^x dy f(x), \tag{22.28}$$

where 0 can be replaced by any convenient starting point for the integral, e.g., $-\infty$. According to the solution formula (14.59) (adapted to general diffusivity; see Exercise ■), the solution to the initial value problem (22.22), (22.28) for the heat equation can be written as a convolution with the fundamental solution:

$$v(t,x) = \frac{1}{2\sqrt{\pi\,\gamma\,t}} \int_{-\infty}^{\infty} e^{-(x-y)^2/(4\,\gamma\,t)}\, h(y)\, dy.$$

Therefore, the solution to the Burgers' initial value problem (22.20), (22.21) is

$$u(t,x) = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{x-y}{t}\, e^{-G(t,x,y)/2\,\gamma}\, dy}{\displaystyle\int_{-\infty}^{\infty} e^{-G(t,x,y)/2\,\gamma}\, dy} \qquad \text{where} \qquad G(t,x,y) = \int_0^y f(z)\, dz + \frac{(x-y)^2}{2\,\gamma}\,. \tag{22.29}$$

Example 22.7. to demonstrate the smoothing effect of the diffusion terms, let us take the initial data

$$u(0,x) = \begin{cases} a, & x < 0, \\ b, & x > 0, \end{cases} \tag{22.30}$$

in the form of a step function. We assume that $a > b$, which would correspond to a shock wave in the inviscid limit. (The reader is asked to analyze the case $a < b$ which corresponds to a rarefaction wave.) The solution takes the form

$$u(t,x) = b + \frac{a-b}{1 + w(t,x)\, \exp \dfrac{a-b}{2\gamma}(x - ct)}$$

where

$$c = \frac{a+b}{2}, \qquad w(t,x) \;=\; \frac{\dfrac{\sqrt{\pi}}{2} \;-\; \operatorname{erf} \dfrac{bt - x}{\sqrt{4\pi\gamma\, t}}}{\dfrac{\sqrt{\pi}}{2} \;-\; \operatorname{erf} \dfrac{x - at}{\sqrt{4\pi\gamma\, t}}}$$

where $\operatorname{erf} z$ denotes the error function (14.61). The solution is plotted in Figure Bshock■. Note that the sharp transition region for the schock has been smoothed out. The larger the diffusion coefficient in relation to the initial solution heights a, b, the more significant the smoothing effect.

Example 22.8. Consider the case when the initial data $u(0,x) = a\,\delta(x)$ is a concentrated delta function impulse at the origin. In the solution formula (22.29), starting the integral for $G(t,x,y)$ at 0 is problematic, but as noted earlier, any other starting point will lead to a valid formula. Thus, we take

$$G(t,x,y) = \int_{-\infty}^{y} a\,\delta(z)\,dz + \frac{(x-y)^2}{2\gamma} = \begin{cases} \dfrac{(x-y)^2}{2\gamma}, & y < 0, \\ a + \dfrac{(x-y)^2}{2\gamma}, & y > 0. \end{cases}$$

Substituting this into (22.29), we can evaluate the upper integral in elementary terms, while the lower integral involves the error function (14.61):

$$u(t,x) = \sqrt{\frac{\gamma}{c}}\; \frac{(e^{a/2\gamma} - 1)\, e^{-x^2/(4\gamma t)}}{\frac{1}{2}\sqrt{\pi}\,(e^{a/2\gamma} + 1) + (e^{a/2\gamma} - 1)\, \operatorname{erf}(x/\sqrt{4\pi\gamma\, t}\,)}\,. \tag{22.31}$$

A graph of this solution appears in Figure hump■. As you can see, the delta function diffuses out, but, unlike the heat equation, the wave does not remain symmetric owing to the advection terms in the equation. The effect is to steepen in front as it propagates. We note that (22.31) is in the form of a similarity solution

$$u(t,x) = \sqrt{\frac{\gamma}{c}}\; F\left(\frac{x}{\sqrt{4\pi\gamma\, t}}\right),$$

which could perhaps have been predicted from the scaling invariance of the initial data.

If $a \ll 1$ is small, then the nonlinear terms in burgers' equation are negligible, and the solution is very close to the fundamental source solution to the heat equation. On the other hand, for large $a \gg 1$, one would expect the advection terms to dominate, and the onmly effect of diffusion being a smoothing at any abrupt discontinuity. Indeed, for large a the leading edge of the solution is in the form of a shock wave. As $a \to \infty$, the solution converges to the similarity solution

$$u(t,x) \longrightarrow \begin{cases} \dfrac{x}{t}\,, & 0 \le x \le \sqrt{2\,a\,t}\,, \\ 0, & \text{otherwise.} \end{cases}$$

of the inviscid wave equation (22.12).

22.3. Dispersion and Solitons.

Finally, we look at a remarkable third order evolution equation that originally arose in the modeling of surface water waves, that serves to introduce yet further phenomena, both linear and nonlinear. The third order derivative models dispersion, in which waves of different frequencies move at different speeds. Coupled with the same nonlinearity as in the inviscid and viscous Burgers' (22.12), (22.20), the result is one of the most remarkable equaitons in all of mathematics, with far-reaching implicaitons, njot only in fluid mechanics and applications, but even in complex function theory, physics, etc., etc.

Linear Dispersion

The simplest linear partial differential equation of a type that we have not yet considered is the third order equation

$$u_t = u_{xxx} \tag{22.32}$$

It is the third member of the hierarchy of simple evolution equations that starts with the simple ordinary differential equation $u_t = u$, then proceeds to the unidirectional wave equation $u_t = u_x$, and then the heat equation $u_t = u_{xx}$. Each member of the hierarchy has its own range of phenomenology. The third order case is a simple model for linear dispersive waves.

We shall only look at the equation on the entire line, so $x \in \mathbb{R}$, and so can ignore additional complications caused by boundary conditions. The solution to the equation is uniquely specified by initial data

$$u(0,x) = f(x), \qquad\qquad -\infty < x < \infty.$$

See [**X**] for a proof.

Let us apply the Fourier transform to solve the equation. Using separation of variables Substitute

$$u(t,x) = e^{\,\mathrm{i}\,\omega\,t + \mathrm{i}\,k\,x}$$

where ω is the frequency and k is called the *wave number*. We find $\omega = k^3\ 2\pi\blacksquare$ is the dispersion relation. Therefore, the solution is given by superposition as a Fourier integral

$$u(t,x) = \int_{-\infty}^{\infty} e^{\,\mathrm{i}\,k^3\,t + \mathrm{i}\,k\,x}\,\widehat{f}(k)\,dk$$

In particular, the solution with a concentrated initial disturbance

$$u(0,x) = \delta(x) \qquad \text{is} \qquad u(t,x) = \operatorname{Ai}\Big(\frac{x}{t^{1/3}}\Big)$$

in terms of the Airy function. See Figure ee3 for a graph.

Fundamwental solution and superposition

Although energy is conserved, unlike the heat and diffusion equations, the dispersion of waves means that the solution dies out.

group velocity and wave velocity.

The Korteweg–deVries Equation

The simplest wave equation that combines dispersion with nonlinearity is the celebrated *Korteweg–deVries equation*

$$u_t + u_{xxx} + u\,u_x = 0. \tag{22.33}$$

The equation was first derived by the French applied mathematician Boussinesq, [**22**; eq. (30), p. 77], [**23**; eqs. (283, 291)], in 1872 as a model for surface water waves. It was rediscovered by the Dutch mathematicians Korteweg and de Vries, [**96**], over two decades later. More recently, in the early 1960's, Kruskal and Zabusky, [**155**], rederived it as a continuum limit of a model of nonlinear mass-spring chains studied by Fermi, Pasta and Ulam, [**54**]. Their numerical experiments on the eqation opened the door to the undertanding of its many remarkable properties. It has a critical balance between nonlinear effects and dispersion, leading to integrability.

The most important special solutions to the Korteweg–deVries equation are the traveling waves. We assume that

$$u = v(\xi) = v(x - ct)$$

to be a wave of permanent form, translating to the right with speed c, thatis, a solution to $u_t + c\,u_x = 0$. Note that

$$\frac{\partial u}{\partial t} = -\,c\,v'(\xi), \qquad \frac{\partial u}{\partial x} = v'(\xi), \qquad \frac{\partial^3 u}{\partial x^3} = v'''(\xi).$$

Therefore, $v(\xi)$ satsifies the third order nonlinear ordinary differential equation

$$v''' + v\,v' - c\,v = 0. \tag{22.34}$$

Moreover, we impose boundary conditions

$$\lim_{\xi \to \pm\infty} v(\xi) = \lim_{\xi \to \pm\infty} v'(\xi) = \lim_{\xi \to \pm\infty} v''(\xi) = 0. \tag{22.35}$$

This equation can be integrated. First, note that it can be written as a derivative:

$$\frac{d}{d\xi}\big[\,v'' + \tfrac{1}{2}\,v^2 - c\,v\,\big] = 0, \qquad \text{and hence} \qquad v'' + \tfrac{1}{2}\,v^2 - c\,v = a,$$

where a is a constant of integration. However, the boundary conditions as $\pm\infty$ imply that $a = 0$. Multiplying the latter equation by v' allows us to integrate a second time

$$\frac{d}{d\xi}\left[\,\tfrac{1}{2}(v')^2 + \tfrac{1}{6}\,v^3 - \tfrac{1}{2}\,c\,v^2\,\right] = v'\left[\,v'' + \tfrac{1}{2}\,v^2 - c\,v\,\right] = 0.$$

Integrating both sides of the equation,

$$\tfrac{1}{2}(v')^2 + \tfrac{1}{6}\,v^3 - \tfrac{1}{2}\,c\,v^2 = b,$$

where b is a second constant of integration, which, again by the boundary conditions (22.35), is also $b = 0$.

We also assume that the wave is localized, meaning that u and its derivatives tend to 0 as $|\,x\,| \to \infty$. Therefore, $v(\xi)$ satisfies the first order autonomous ordinary differential equation

$$\frac{dv}{d\xi} = v\sqrt{c - \tfrac{1}{3}\,v}\,.$$

We integrate by the usual method:

$$\int \frac{dv}{v\sqrt{c - \frac{1}{3}\,v}} = \xi + \delta.$$

The solution has the form

$$v(\xi) = 3\,c\,\operatorname{sech}^2\left[\,\tfrac{1}{2}\sqrt{c}\,\xi + \delta\,\right],$$

where

$$\operatorname{sech} y = \frac{1}{\cosh y} = \frac{2}{e^y + e^{-y}}\,,$$

is the *hyperbolic secant function*. Hence, the localized traveling wave solutions of the Korteweg–deVries equation equation are of the form

$$u(t,x) = 3\,c\,\operatorname{sech}^2\left[\,\tfrac{1}{2}\sqrt{c}\,(x - c\,t) + \delta\,\right], \tag{22.36}$$

where $c > 0$ and δ are arbitrary constants. The parameter c measures the velocity of the wave. It also masures its amplitude, since the maximum value of $u(t,x)$ is $3\,c$ since $\operatorname{sech} y$ has a maximum value of 1 at $y = 0$. Therefore, the taller the wave, the faster it moves. See Figure soliton■ for a graph.

The solution (22.36) is known as a *solitary wave solution* since it represents a localized wave that travels unchanged in shape. Such waves were first observed by the British engineer J. Scott Russell, [**131**], who tells the remarkable incident of chasing such a wave generated by the sudden motion of a barge along an Edinburgh canal on horseback for several miles. The mathematician Airy claimed that such waves could not exist, but he based his analysis upon a linearized theory. Boussinesq's establishment of the surface wve model demonstrated that such localized disturbances can result from nonlinear effects in the system.

Remark: In the Korteweg–deVries equation model, one can find arbitrarily tall soliton solutions. In physical water waves, if the wave is too tall it will break. Indeed, it can be rigorously proved that the full water wave equations admit solitary wave solutions, but there is a wave of greatest height, beyond which a wave will tend to break. The solitary water waves are not geniune solitons, since there is a small, but measureable, effect when two waves collide.

These nonlinear traveling wave solutions were discovered by Kruskal and Zabusky, [**155**], to have remarkable properties. For this reason they have been given a special new name — *soliton*. Ordinarily, combining two solutions to a nonlinear equation can be quite unpredictable, and one might expect any number of scenarios to occur. If you start with initial conditions representing a taller wave to the left of a shorter wave, the solution of the Korteweg–deVries equation runs as follows. The taller wave moves faster, and so catches up the shorter wave. They then have a very complicated nonlinear interaction, as expected. But, remarkably, after a while they emerge from the interaction unscathed. the smaller wave is now in back and the larger one in front. After this, they proceed along their way, with the smaller one lagging behind the high speed tall wave. the only effect of their encounter is a phase shift, meaning a change in the value of the phase parameter δ in each wave. See Figure solitons■. After the interaction, the position of the soliton if it had traveled unhindered by the other is shown in a dotted line. Thus, they behave like colliding paricles, which is the genesis of the word "soliton".

A similar phenomenon holds for several such soliton solutions. After some time where the various waves interact, they finally emerge with the largest soliton in front, and then in order to the smallest one in back, all progressing at their own speed, and so gradually drawing apart.

Moreover, starting with an *arbitrary* initial disturbance

$$u(0, x) = f(x)$$

it can be proved that after some time, the solution disintegrates into a finite number of solitons of different heights, moving off to the right, plus a small dispersive tail moving to the left that rapidly disappears. Propving this remarkable result is beyond the scope of this book. It relies on the method of *inverse scattering*, that connects the Korteweg–deVries equation with a linear eigenvalue problem of fundamental importance in one-dimensional quantum mechanics. The solitons correspond to the bound states of a quantum potential. We refer the interested reader to the introductory text [**50**] and the more advanced monograph [**1**] for details.

Chaos and integrability are the two great themes in modern nonlinear applied mathematics, and the student is well-advised to pursue both.

There is a remarkable transformation, known as the inverse scattering transform, which is a form of nonlinear Fourier transform, that can be used to solve the Korteweg–deVries equation. Its fascinating properties continue to be of great current research interest to this day.

22.4. Conclusion and Bon Voyage.

These are your first wee steps in a vast new realm. We are unable to discuss nonlinear

partial differential equations arising in fluid mechanics, in elasticity, in relativity, in differential geometry, n computer vision, in mathematical biology. We bid the reader adieu and farewell.

Appendix A

Vector Calculus in Two Dimensions

so far, we have concentrated on problems of one-dimensional media — bars, beams and strings. In order to study the partial differential equations describing the equilibria and dynamics of planar media, we need to review the basics of vector calculus in the two dimensions. We begin with a discussion of plane curves and domains. Many physical quantities, including force and velocity, are determined by vector fields, and we review the basic concepts. The key differential operators in planar vector calculus are the gradient and divergence operations, along with the Jacobian matrix for maps from $\mathbb{R}^2$ to itself. There are three basic types of line integrals: integrals with respect to arc length, for computing lengths of curves, masses of wires, center of mass, etc., ordinary line integrals of vector fields for computing work and fluid circulation, and flux line integrals for computing flux of fluids and forces. Next, we review the basics of double integrals of scalar functions over plane domains. Line and double integrals are connected by the justly famous Green's theorem, which is the two-dimensional version of the fundamental theorem of calculus. The integration by parts argument required to characterize the adjoint of a partial differential operator rests on the closely allied Green's formula.

Space limitations require us to go through this material fairly rapidly, and we assume that you already gained sufficient familiarity with most of these concepts in a sophomore-level multi-variable calculus course. More details, and full justifications of these results can be found in many of the standard vector calculus texts, including [**9**].

A.1. Plane Curves.

We begin our review by collecting together the basic facts concerning geometry of plane curves. A *curve* $C \subset \mathbb{R}^2$ is parametrized by a pair of continuous functions

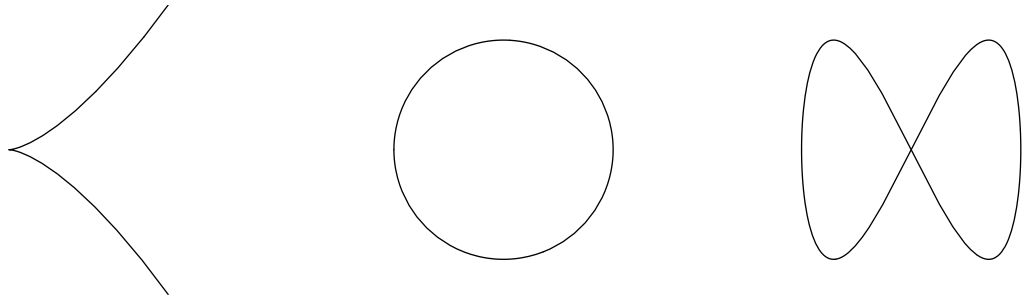
$$\mathbf{x}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \in \mathbb{R}^2, \tag{A.1}$$

where the scalar *parameter* t varies over an (open or closed) interval $I \subset \mathbb{R}$. When it exists, the *tangent vector* to the curve at the point $\mathbf{x}$ is described by the derivative,

$$\frac{d\mathbf{x}}{dt} = \dot{\mathbf{x}} = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix}. \tag{A.2}$$

We shall often use Newton's dot notation to abbreviate derivatives with respect to the parameter t.

Physically, we can think of a curve as the trajectory described by a particle moving in the plane. The parameter t is identified with the time, and so $\mathbf{x}(t)$ gives the position of the



Cusped Curve Circle Figure Eight

Figure A.1. Planar Curves.

particle at time t. The tangent vector $\dot{\mathbf{x}}(t)$ measures the velocity of the particle at time t; its magnitude[†] $\| \dot{\mathbf{x}} \| = \sqrt{\dot{x}^2 + \dot{y}^2}$ is the speed, while its orientation (assuming the velocity is nonzero) indicates the instantaneous direction of motion of the particle as it moves along the curve. Thus, by the *orientation* of a curve, we mean the direction of motion or parametrization, as indicated by the tangent vector. Reversing the orientation amounts to moving backwards along the curve, with the individual tangent vectors pointing in the opposite direction.

The curve parametrized by $\mathbf{x}(t)$ is called *smooth* provided its tangent vector is continuous and everywhere *nonzero*: $\dot{\mathbf{x}} \neq \mathbf{0}$. This is because curves with vanishing derivative may have corners or cusps; a simple example is the first curve plotted in Figure A.1, which has parametrization

$$\mathbf{x}(t) = \begin{pmatrix} t^2 \\ t^3 \end{pmatrix}, \qquad \dot{\mathbf{x}}(t) = \begin{pmatrix} 2t \\ 3t^2 \end{pmatrix},$$

and has a cusp at the origin when $t = 0$ and $\dot{\mathbf{x}}(0) = 0$. Physically, a particle trajectory remains smooth as long as the speed of the particle is never zero, which effectively prevents the particle from instantaneously changing its direction of motion. A closed curve is *smooth* if, in addition to satisfying $\dot{\mathbf{x}}(t) \neq \mathbf{0}$ at all points $a \leq t \leq b$, the tangents at the endpoints match up: $\dot{\mathbf{x}}(a) = \dot{\mathbf{x}}(b)$. A curve is called *piecewise smooth* if its derivative is piecewise continuous and nonzero everywhere. The corners in a piecewise smooth curve have well-defined right and left tangents. For example, polygons, such as triangles and rectangles, are piecewise smooth curves. In this book, all curves are assumed to be at least piecewise smooth.

A curve is *simple* if it has no self-intersections: $\mathbf{x}(t) \neq \mathbf{x}(s)$ whenever $t \neq s$. Physically, this means that the particle is never in the same position twice. A curve is *closed* if $\mathbf{x}(t)$ is defined for $a \leq t \leq b$ and its endpoints coincide: $\mathbf{x}(a) = \mathbf{x}(b)$, so that the particle ends up where it began. For example, the unit circle

$$\mathbf{x}(t) = (\cos t, \sin t)^T \qquad \text{for} \qquad 0 \leq t \leq 2\pi,$$

[†] Throughout this chapter, we always use the standard Euclidean inner product and norm. With some care, all of the concepts can be adapted to other choices of inner product. In differential geometry and relativity, one even allows the inner product and norm to vary from point to point, [**49**].

is closed and simple†, while the curve

$$\mathbf{x}(t) = (\cos t, \sin 2t)^T \qquad \text{for} \qquad 0 \le t \le 2\pi,$$

is not simple since it describes a figure eight that intersects itself at the origin. Both curves are illustrated in Figure A.1.

Assuming the tangent vector $\dot{\mathbf{x}}(t) \neq \mathbf{0}$, then the *normal* vector to the curve at the point $\mathbf{x}(t)$ is the orthogonal or perpendicular vector

$$\dot{\mathbf{x}}^{\perp} = \begin{pmatrix} \dot{y} \\ -\dot{x} \end{pmatrix} \tag{A.3}$$

of the same length $\| \dot{\mathbf{x}}^{\perp} \| = \| \dot{\mathbf{x}} \|$. Actually, there are two such normal vectors, the other being the negative $-\dot{\mathbf{x}}^{\perp}$. We will always make the "right-handed" choice (A.3) of normal, meaning that as we traverse the curve, the normal always points to our right. If a simple closed curve C is oriented so that it is traversed in a counterclockwise direction — the standard mathematical orientation — then (A.3) describes the outwards-pointing normal. If we reverse the orientation of the curve, then both the tangent vector and normal vector change directions; thus (A.3) would give the inwards-pointing normal for a simple closed curve traversed in the clockwise direction.

The same curve C can be parametrized in many different ways. In physical terms, a particle can move along a prescribed trajectory at a variety of different speeds, and these correspond to different ways of parametrizing the curve. Conversion from one parametrization $\mathbf{x}(t)$ to another $\widetilde{\mathbf{x}}(\tau)$ is effected by a *change of parameter*, which is a smooth, invertible function $t = g(\tau)$; the reparametrized curve is then $\widetilde{\mathbf{x}}(\tau) = \mathbf{x}(g(\tau))$. We require that $dt/d\tau = g'(\tau) > 0$ everywhere. This ensures that each t corresponds to a unique value of τ, and, moreover, the curve remains smooth and is traversed in the same overall direction under the reparametrization. On the other hand, if $g'(\tau) < 0$ everywhere, then the orientation of the curve is reversed under the reparametrization. We shall use the notation $-C$ to indicate the curve having the same shape as C, but with the reversed orientation.

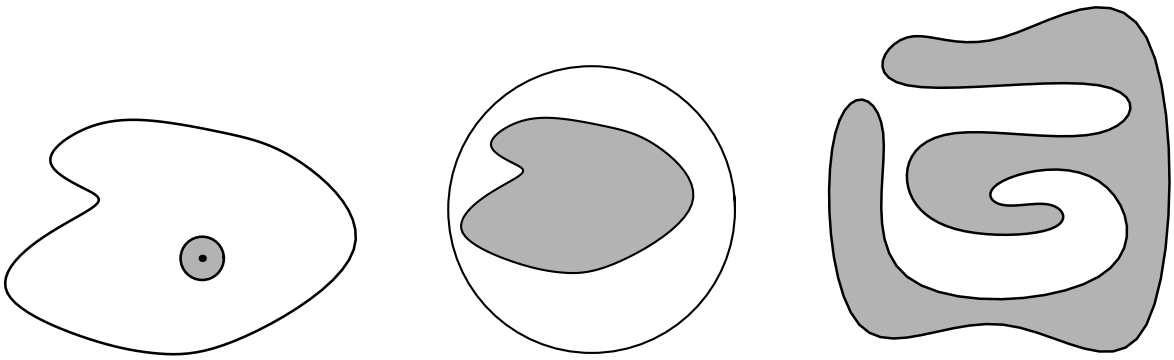
Example A.1. The function $\mathbf{x}(t) = (\cos t, \sin t)^T$ for $0 < t < \pi$ parametrizes a semi-circle of radius 1 centered at the origin. If we set† $\tau = -\cot t$ then we obtain the less evident parametrization

$$\widetilde{\mathbf{x}}(\tau) = \left(\frac{1}{\sqrt{1+\tau^2}}\,, \; -\frac{\tau}{\sqrt{1+\tau^2}} \right)^T \qquad \text{for} \qquad -\infty < \tau < \infty$$

of the *same* semi-circle, in the *same* direction. In the familiar parametrization, the velocity vector has unit length, $\| \dot{\mathbf{x}} \| \equiv 1$, and so the particle moves around the semicircle in the counterclockwise direction with unit speed. In the second parametrization, the particle

† For a closed curve to be simple, we require $\mathbf{x}(t) \neq \mathbf{x}(s)$ whenever $t \neq s$ *except* at the ends, where $\mathbf{x}(a) = \mathbf{x}(b)$ is required for the ends to close up.

† The minus sign is to ensure that $d\tau/dt > 0$.



Interior Point Bounded Domain A Simple Closed Curve

Figure A.2. Topology of Planar Domains.

slows down near the endpoints, and, in fact, takes an infinite amount of time to traverse the semicircle from right to left.

A.2. Planar Domains.

A plate or other two-dimensional body occupies a region in the plane, known as a *domain*. The simplest example is an open circular disk

$$D_r(\mathbf{a}) = \big\{\, \mathbf{x} \in \mathbb{R}^2 \;\big|\; \|\, \mathbf{x} - \mathbf{a} \,\| < r \,\big\} \tag{A.4}$$

of radius r centered at a point $\mathbf{a} \in \mathbb{R}^2$. In order to properly formulate the mathematical tools needed to understand boundary value problems and dynamical equations for such bodies, we first need to review basic terminology from point set topology of planar sets. Many of the concepts carry over as stated to subsets of any higher dimensional Euclidean space $\mathbb{R}^n$.

Let $\Omega \subset \mathbb{R}^2$ be any subset. A point $\mathbf{a} \in \Omega$ is called an *interior point* if some small disk centered at $\mathbf{a}$ is entirely contained within the set: $D_\varepsilon(\mathbf{a}) \subset \Omega$ for some $\varepsilon > 0$; see Figure A.2. The set Ω is *open* if every point is an interior point. A set K is *closed* if and only if its complement $\Omega = \mathbb{R}^2 \setminus K = \{\mathbf{x} \;/\; \mathbf{x} \notin K\}$ is open.

Example A.2. If $f(x,y)$ is any continuous real-valued function, then the subset $\{\, f(x,y) > 0 \,\}$ where f is strictly positive is open, while the subset $\{\, f(x,y) \geq 0 \,\}$ where f is non-negative is closed. One can, of course, replace 0 by any other constant, and also reverse the direction of the inequalities, without affecting the conclusions.

In particular, the set

$$D_r = \{\, x^2 + y^2 < r^2 \,\} \tag{A.5}$$

consisting of all points of (Euclidean) norm strictly less than r, defines an *open disk* of radius r centered at the origin. On the other hand,

$$K_r = \{\, x^2 + y^2 \leq r^2 \,\} \tag{A.6}$$

is the *closed disk* of radius r, which includes the bounding circle

$$C_r = \{\, x^2 + y^2 = r^2 \,\}. \tag{A.7}$$

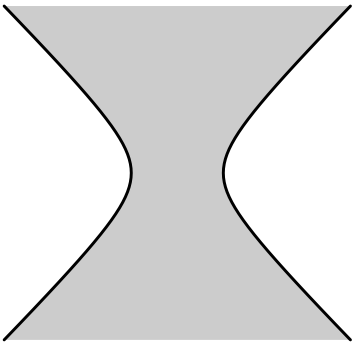


Figure A.3. Open Sets Defined by a Hyperbola.

A point $\mathbf{x}^\star$ is a *limit point* of a set Ω if there exists a sequence of points $\mathbf{x}^{(n)} \in \Omega$ converging to it, so that[†] $\mathbf{x}^{(n)} \to \mathbf{x}^\star$ as $n \to \infty$. Every point $\mathbf{x} \in \Omega$ is a limit point (just take all $\mathbf{x}^{(n)} = \mathbf{x}$) but the converse is not necessarily valid. For example, the points on the circle (A.7) are all limit points for the open disk (A.5). The *closure* of a set Ω, written $\overline{\Omega}$, is defined as the set of all limit points of Ω. In particular, a set K is closed if and only if it contains all its limit points, and so $K = \overline{K}$. The *boundary* $\partial\Omega$ of a subset Ω consists of all limit points which are not interior points. If Ω is open, then its closure is the disjoint union of the set and its boundary $\overline{\Omega} = \Omega \cup \partial\Omega$. Thus, the closure of the open disk D_r is the closed disk $\overline{D}_r = D_r \cup C_r$; the circle $C_r = \partial D_r = \partial \overline{D}_r$ forms their common boundary.

An open subset that can be written as the union, $\Omega = \Omega_1 \cup \Omega_2$, of two disjoint, nonempty, open subsets, so $\Omega_1 \cap \Omega_2 = \varnothing$, is called *disconnected*. For example, the open set

$$\Omega = \{\, x^2 - y^2 > 1 \,\} \tag{A.8}$$

is disconnected, consisting of two disjoint "sectors" bounded by the two branches of the hyperbola $x^2 - y^2 = 1$; see Figure A.3. On the other hand, the complementary open set

$$\widehat{\Omega} = \{\, x^2 - y^2 < 1 \,\} \tag{A.9}$$

is *connected*, and consists of all points between the two hyperbolas.

A subset is called *bounded* if it is contained inside a (possibly large) disk, i.e., $\Omega \subset D_r$ for some $r > 0$, as in the second picture in Figure A.2. Thus, both the closed and the open disks (A.5), (A.6) are bounded, whereas the two hyperbolic sectors (A.8), (A.9) are both unbounded.

The class of subsets for which the boundary value problems for the partial differential equations of equilibrium mechanics are properly prescribed can now be defined.

Definition A.3. A *planar domain* is a connected, open subset $\Omega \subset \mathbb{R}^2$ whose boundary $\partial\Omega$ consists of one or more piecewise smooth, simple curves, such that Ω lies entirely on one side of each of its boundary curve(s).

[†] See Section 12.5 for more details on convergence.

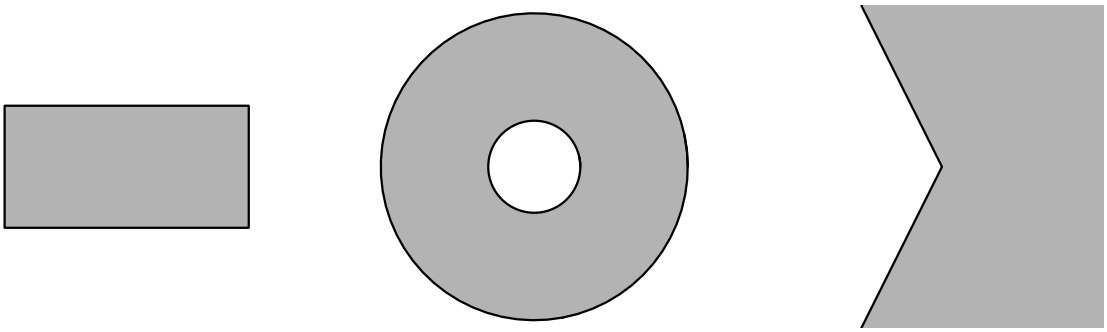


Figure A.4. Planar Domains.

The last condition is to avoid dealing with pathologies. For example, the subset $\Omega \setminus C$ obtained by cutting out a curve C from the interior of an open set Ω would not be an allowable domain.

Example A.4. The open rectangle $R = \{\, a < x < b, c < y < d \,\}$ is an open, connected and bounded domain. Its boundary is a piecewise smooth curve, since there are corners where the tangent does not change continuously.

The *annulus*

$$r^2 < x^2 + y^2 < R^2, \qquad \text{for fixed} \qquad 0 < r < R, \tag{A.10}$$

is an open, connected, bounded domain whose boundary consists of two disjoint concentric circles. The degenerate case of a *punctured disk*, when $r = 0$, is *not* a domain since its boundary consists of a circle and a single point — the origin.

Another well-studied example is the wedge-shaped domain $W = \{\, \alpha < \theta < \beta \,\}$ consisting of all points whose angular coordinate $\theta = \tan^{-1} y/x$ lies between two prescribed values. If $0 < \beta - \alpha < 2\,\pi$, then the wedge is a domain whose boundary consists of two connected rays. However, if $\beta = \alpha + 2\,\pi$, then the wedge is obtained by cutting the plane along a single ray at angle α. The latter case does not comply with our definition of a domain since the wedge now lies on both sides of its boundary ray.

Any connected domain is automatically *pathwise connected* meaning that any two points can be connected by (i.e., are the endpoints of) a curve lying entirely within the domain. If the domain is bounded, which is the most important case for boundary value problems, then its boundary consists of one or more piecewise smooth, simple, closed curves. A bounded domain Ω is called *simply connected* if it has just one such boundary curve; this means that Ω is connected and has no holes, and so its boundary $\partial\Omega = C$ is a simple closed curve that contains Ω in its interior. For instance, an open disk and a rectangle are both simply connected, whereas an annulus is not.

The Jordan Curve Theorem states the intuitively obvious, but actually quite deep, result that any simple closed curve divides the plane $\mathbb{R}^2$ into two disjoint, connected, open domains — its *interior*, which is bounded and simply connected, and its *exterior*, which is unbounded and not simply connected. This result is illustrated in the final figure in Figure A.2; the interior of the indicated simple closed curve is shaded in gray while the

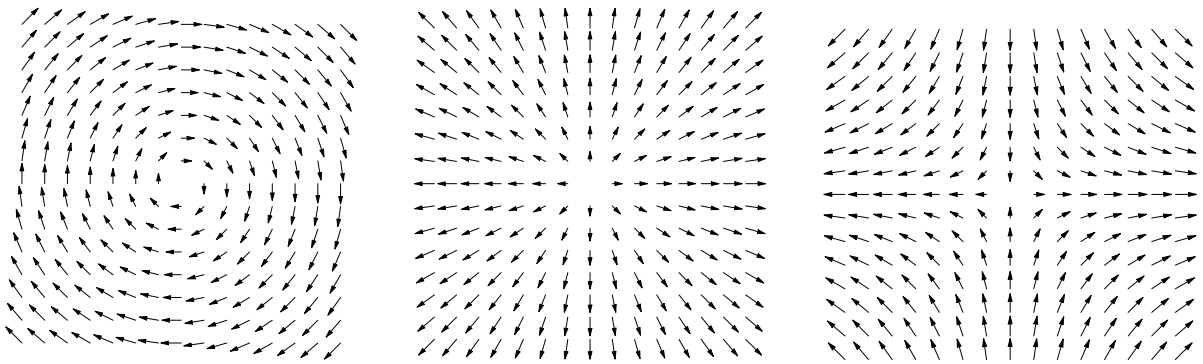


Figure A.5. Vector Fields.

exterior is in white. Note that the each subdomain lies entirely on one side of the curve, which forms their common boundary.

The following result is often used to characterize the simple connectivity of more general planar subsets, including unbounded domains.

Lemma A.5. *A planar domain $\Omega \subset \mathbb{R}^2$ is* simply connected *if it is connected and, moreover, if the interior of any simple closed curve $C \subset \Omega$ is also contained in Ω.*

For example, an annulus (A.10) is not simply connected because the interior of a circle going around the hole is not entirely contained within the annulus. On the other hand, the unbounded domain (A.9) lying between two branches of a hyperbola is simply connected, even though its boundary consists of two disjoint, unbounded curves.

A.3. Vector Fields.

A vector-valued function $\mathbf{v}(x,y) = \begin{pmatrix} v_1(x,y) \\ v_2(x,y) \end{pmatrix}$ is known as a (planar) *vector field*. A vector field assigns a vector $\mathbf{v}(x,y)$ to each point $(\,x,y\,)^T$ in its domain of definition, and hence defines a (in general nonlinear) function $\mathbf{v}\colon \Omega \to \mathbb{R}^2$. The vector field can be conveniently depicted by drawing an arrow representing the vector $\mathbf{v} = \mathbf{v}(x,y)$ starting at its point of definition $(\,x,y\,)^T$. See Figure A.5 for some representative sketches.

Example A.6. Vector fields arise very naturally in physics and engineering applications from physical forces: gravitational, electrostatic, centrifugal, etc. A *force field* $\mathbf{f}(x,y) = (\,f_1(x,y), f_2(x,y)\,)^T$ describes the direction and magnitude of the force experienced by a particle at position (x,y). In a planar universe, the gravitational force field exerted by a point mass concentrated at the origin has, according to Newtonian gravitational theory, magnitude proportional to[†] $1/r$, where $r = \| \mathbf{x} \|$ is the distance to the origin, and is directed towards the origin. Thus, the vector field describing gravitational force has

[†] In three-dimensional Newtonian gravity, $1/r$ is replaced by $1/r^2$.

the form

$$\mathbf{f} = -\gamma \frac{\mathbf{x}}{\| \mathbf{x} \|} = \left(\frac{-\gamma x}{\sqrt{x^2 + y^2}}, \frac{-\gamma y}{\sqrt{x^2 + y^2}} \right)^T, \tag{A.11}$$

where $\gamma > 0$ denotes the constant of proportionality, namely the product of the two masses times the universal gravitational constant. The same force law applies to the attraction, $\gamma > 0$, and repulsion, $\gamma < 0$, of electrically charged particles.

Newton's Laws of planetary motion produce the second order system of differential equations

$$m \frac{d^2\mathbf{x}}{dt^2} = \mathbf{f}.$$

The solutions $\mathbf{x}(t)$ describe the trajectories of planets subject to a central gravitational force, e.g., the sun. They also govern the motion of electrically charged particles under a central electric charge, e.g., classical (i.e., not quantum) electrons revolving around a central nucleus. In three-dimensional Newtonian mechanics, planets move along conic sections — ellipses in the case of planets, and parabolas and hyperbolas in the case of non-recurrent objects like some comets. Interestingly (and not as well-known), the corresponding two-dimensional theory is not as neatly described — the typical orbit of a planet around a planar sun does not form a simple closed curve, [**45**]!

Example A.7. Another important example is the velocity vector field $\mathbf{v}$ of a steady-state fluid flow. The vector $\mathbf{v}(x, y)$ measures the instantaneous velocity of the fluid particles (molecules or atoms) as they pass through the point (x, y). "Steady-state" means that the velocity at a point (x, y) does not vary in time — even though the individual fluid particles are in motion. If a fluid particle moves along the curve $\mathbf{x}(t) = (x(t), y(t))^T$, then its velocity at time t is the derivative $\mathbf{v} = \dot{\mathbf{x}}$ of its position with respect to t. Thus, for a time-independent velocity vector field $\mathbf{v}(x, y) = (\, v_1(x, y), v_2(x, y)\,)^T$, the fluid particles will move in accordance with an autonomous, first order system of ordinary differential equations

$$\frac{dx}{dt} = v_1(x, y), \qquad \frac{dy}{dt} = v_2(x, y). \tag{A.12}$$

According to the basic theory† of systems of ordinary differential equations , an individual particle's motion $\mathbf{x}(t)$ will be uniquely determined solely by its initial position $\mathbf{x}(0) = \mathbf{x}_0$. In fluid mechanics, the trajectories of particles are known as the *streamlines* of the flow. The velocity vector $\mathbf{v}$ is everywhere tangent to the streamlines. When the flow is steady, the streamlines do not change in time. Individual fluid particles experience the same motion as they successively pass through a given point in the domain occupied by the fluid.

As a specific example, consider the vector field

$$\mathbf{v}(x, y) = \begin{pmatrix} -\omega y \\ \omega x \end{pmatrix}, \tag{A.13}$$

† See Section 20.2 for details.

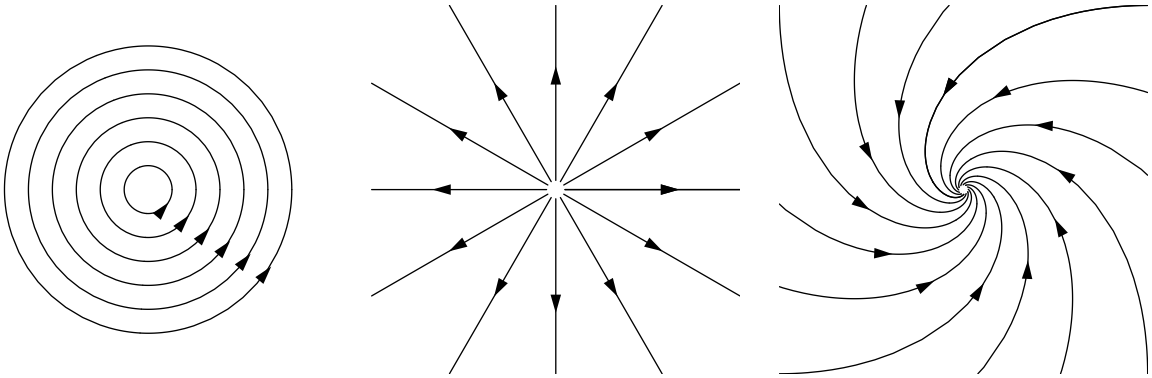


Figure A.6. Steady State Fluid Flows.

for fixed $\omega > 0$, which is plotted in the first figure in Figure A.5. The corresponding fluid trajectories are found by solving the associated first order system of ordinary differential equations

$$\dot{x} = -\omega\, y, \qquad \dot{y} = \omega\, x,$$

with initial conditions $x(0) = x_0$, $y(0) = y_0$. This is a linear system, and can be solved by the eigenvalue and eigenvector techniques presented in Chapter 9. The resulting flow

$$x(t) = x_0 \cos \omega t - y_0 \sin \omega t, \qquad y(t) = x_0 \sin \omega t + y_0 \cos \omega t,$$

corresponds to a fluid that is uniformly rotating around the origin. The streamlines are concentric circles, and the fluid particles rotate around the circles in a counterclockwise direction with angular velocity ω, as illustrated in Figure A.6. Note that the fluid velocity $\mathbf{v}$ is everywhere tangent to the circles. The origin is a stagnation point, since the velocity field $\mathbf{v} = \mathbf{0}$ vanishes there, and the particle at the origin does not move.

As another example, the radial vector field

$$\mathbf{v}(x, y) = \alpha \mathbf{x} = \begin{pmatrix} \alpha\, x \\ \alpha\, y \end{pmatrix} \tag{A.14}$$

corresponds to a fluid source, $\alpha > 0$, or sink, $\alpha < 0$, at the origin, and is plotted in the second figure in Figure A.5. The solution to the first order system of ordinary differential equations $\dot{\mathbf{x}} = \alpha \mathbf{x}$ with initial conditions $\mathbf{x}(0) = \mathbf{x}_0$ gives the radial flow $\mathbf{x}(t) = e^{\alpha\, t}\, \mathbf{x}_0$. The streamlines are the rays emanating from the origin, and the motion is outwards (source) or inwards (sink) depending on the sign of α. As in the rotational flow, the origin is a stagnation point.

Combining the radial and circular flow vector fields,

$$\mathbf{v}(x, y) = \begin{pmatrix} \alpha\, x - \omega\, y \\ \omega\, x + \alpha\, y \end{pmatrix} \tag{A.15}$$

leads to a swirling source or sink — think of the water draining out of your bathtub. Again, the flow is found by integrating a linear system of ordinary differential equations

$$\dot{x} = \alpha\, x - \omega\, y, \qquad \dot{y} = \omega\, x + \alpha\, y.$$

Solving as in Chapter 9, we find that the fluid particles follow the spiral streamlines

$$x(t) = e^{\alpha t}\big(\, x_0 \cos \omega t - y_0 \sin \omega t \,\big), \qquad\qquad y(t) = e^{\alpha t}\big(\, x_0 \sin \omega t + y_0 \cos \omega t \,\big),$$

again illustrated in Figure A.6.

Remark: All of the phase portraits for linear systems of first order ordinary differential equations in two variables presented in Section 9.3 can be reinterpreted as streamline plots for steady state fluid flows. Additional, nonlinear examples, along with numerical solution techniques, can be found in Chapter 20.

Remark: Of course, physical fluid motion occurs in three-dimensional space. However, any planar flow can also be viewed as a particular type of three-dimensional fluid motion that does not depend upon the vertical coordinate. The motion on every horizontal plane is the same, and so the planar flow represents a cross-section of the full three-dimensional motion. For example, slicing a steady flow past a vertical cylinder by a transverse horizontal plane results in a planar flow around a circle; see Figure fcyl■.

A.4. Gradient and Curl.

In the same vein, a scalar-valued function $u(x, y)$ is often referred to as a *scalar field*, since it assigns a scalar to each point $(\, x, y \,)^T$ in its domain of definition. Typical physical examples of scalar fields include temperature, deflection of a membrane, height of a topographic map, density of a plate, and so on.

The *gradient* operator ∇ maps a scalar field $u(x, y)$ to the vector field

$$\nabla u = \operatorname{grad} u = \begin{pmatrix} \partial u/\partial x \\ \partial u/\partial y \end{pmatrix} \tag{A.16}$$

consisting of its two first order partial derivatives. The scalar field u is often referred to as a *potential function* for its gradient vector field ∇u. For example, the gradient of the potential function $u(x, y) = x^2 + y^2$ is the radial vector field $\nabla u = (\, 2x, 2y \,)^T$. Similarly, the gradient of the logarithmic potential function

$$u(x, y) = -\gamma \, \log r = -\tfrac{1}{2}\, \gamma \, \log(x^2 + y^2)$$

is the gravitational force (A.11) exerted by a point mass concentrated at the origin. Additional physical examples include the velocity potential of certain fluid velocity vector fields and the electromagnetic potential whose gradient describes the electromagnetic force field.

Not every vector field admits a potential because not every vector field lies in the range of the gradient operator ∇. Indeed, if $u(x, y)$ has continuous second order partial derivatives, and

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \mathbf{v} = \nabla u = \begin{pmatrix} u_x \\ u_y \end{pmatrix},$$

then, by the equality of mixed partials,

$$\frac{\partial v_1}{\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial y}\right) = \frac{\partial v_2}{\partial x}\,.$$

The resulting equation

$$\frac{\partial v_1}{\partial y} = \frac{\partial v_2}{\partial x} \tag{A.17}$$

constitutes one of the necessary conditions that a vector field must satisfy in order to be a gradient. Thus, for example, the rotational vector field (A.13) does not satisfy (A.17), and hence is *not* a gradient. There is *no* potential function for such circulating flows.

The difference between the two terms in (A.17) is known as the *curl* of the planar vector field $\mathbf{v} = (v_1, v_2)$, and denoted by[†]

$$\nabla \wedge \mathbf{v} = \operatorname{curl} \mathbf{v} = \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}\,. \tag{A.18}$$

Notice that the curl of a planar vector field is a scalar field. (In contrast, in three dimensions, the curl of a vector field is a vector field — see (B.76).) Thus, a necessary condition for a vector field to be a gradient is that its curl vanish identically: $\nabla \wedge \mathbf{v} \equiv 0$.

Even if the vector field has zero curl, it still may not be a gradient. Interestingly, the general criterion depends only upon the topology of the domain of definition, as clarified in the following theorem.

Theorem A.8. *Let* $\mathbf{v}$ *be a smooth vector field defined on a domain* $\Omega \subset \mathbb{R}^2$. *If* $\mathbf{v} = \nabla u$ *for some scalar function* u, *then* $\nabla \wedge \mathbf{v} \equiv 0$. *If* Ω *is simply connected, then the converse holds*: *if* $\nabla \wedge \mathbf{v} \equiv 0$ *then* $\mathbf{v} = \nabla u$ *for some potential function* u *defined on* Ω.

As we shall see, this result is a direct consequence of Green's Theorem A.25.

Example A.9. The vector field

$$\mathbf{v} = \left(\frac{-y}{x^2 + y^2}\,,\ \frac{x}{x^2 + y^2} \right)^T \tag{A.19}$$

satisfies $\nabla \wedge \mathbf{v} \equiv 0$. However, there is no potential function defined for all $(x, y) \neq (0, 0)$ such that $\nabla u = \mathbf{v}$. As the reader can check, the angular coordinate

$$u = \theta = \tan^{-1} \frac{y}{x} \tag{A.20}$$

satisfies $\nabla \theta = \mathbf{v}$, but θ is not well-defined on the entire domain since it experiences a jump discontinuity of magnitude 2π as we go around the origin. Indeed, $\Omega = \{\mathbf{x} \neq \mathbf{0}\}$ is *not* simply connected, and so Theorem A.8 does not apply. On the other hand, if we restrict $\mathbf{v}$ to any simply connected subdomain $\widehat{\Omega} \subset \Omega$ that does not encircle the origin, then the angular coordinate (A.20) can be unambiguously and smoothly defined on $\widehat{\Omega}$, and does serve as a single-valued potential function for $\mathbf{v}$; see Exercise ∎

[†] In this text, we adopt the more modern wedge notation $\wedge$ for what is often denoted by a cross $\times$.

In fluid mechanics, the curl of a vector field measures the local circulation in the associated steady state fluid flow. If we place a small paddle wheel in the fluid, then its rate of spinning will be in proportion to $\nabla \wedge \mathbf{v}$. (An explanation of this fact will appear below.) The fluid flow is called *irrotational* if its velocity vector field has zero curl, and hence, assuming Ω is simply connected, is a gradient: $\mathbf{v} = \nabla u$. In this case, the paddle wheel will not spin. The scalar function $u(x, y)$ is known as the *velocity potential* for the fluid motion. Similarly, a force field that is given by a gradient $\mathbf{f} = \nabla \varphi$ is called a *conservative force field*, and the function φ defines the force potential.

Suppose $u(\mathbf{x}) = u(x, y)$ is a scalar field. Given a parametrized curve $\mathbf{x}(t) = \big(x(t), y(t) \big)^T$, the composition $f(t) = u(\mathbf{x}(t)) = u(x(t), y(t))$ indicates the behavior as we move along the curve. For example, if $u(x, y)$ represents the elevation of a mountain range at position (x, y), and $\mathbf{x}(t)$ represents our position at time t, then $f(t) = u(\mathbf{x}(t))$ is our altitude at time t. Similarly, if $u(x, y)$ represents the temperature at (x, y), then $f(t) = u(\mathbf{x}(t))$ measures our temperature at time t.

The rate of change of the composite function is found through the chain rule

$$\frac{df}{dt} = \frac{d}{dt}\, u(x(t), y(t)) = \frac{\partial u}{\partial x}\, \frac{dx}{dt} \;+\; \frac{\partial u}{\partial y}\, \frac{dy}{dt} = \nabla u \cdot \dot{\mathbf{x}}, \tag{A.21}$$

and hence equals the dot product between the gradient $\nabla u(\mathbf{x}(t))$ and the tangent vector $\dot{\mathbf{x}}(t)$ to the curve at the point $\mathbf{x}(t)$. For instance, our rate of ascent or descent as we travel through the mountains is given by the dot product of our velocity vector with the gradient of the elevation function. The dot product between the gradient and a fixed vector $\mathbf{a} = \big(a, b \big)^T$ is known as the *directional derivative* of the scalar field $u(x, y)$ in the direction $\mathbf{a}$, and denoted by

$$\frac{\partial u}{\partial \mathbf{a}} = \mathbf{a} \cdot \nabla u = a\, u_x + b\, u_y. \tag{A.22}$$

Thus, the rate of change of u along a curve $\mathbf{x}(t)$ is given by its directional derivative $\partial u / \partial \dot{\mathbf{x}} = \nabla u \cdot \dot{\mathbf{x}}$, as in (A.21), in the tangent direction. This leads us to one important interpretation of the gradient vector.

Proposition A.10. *The gradient ∇u of a scalar field points in the direction of steepest increase of u. The negative gradient, $-\nabla u$, which points in the opposite direction, indicates the direction of steepest decrease of u.*

For example, if $u(x, y)$ represents the elevation of a mountain range at position (x, y) on a map, then ∇u tells us the direction that is steepest uphill, while $-\nabla u$ points directly downhill — the direction water will flow. Similarly, if $u(x, y)$ represents the temperature of a two-dimensional body, then ∇u tells us the direction in which it gets hottest the fastest. Heat energy (like water) will flow in the opposite direction, namely in the direction of the vector $-\nabla u$. This basic fact underlies the derivation of the multi-dimensional heat and diffusion equations.

You need to be careful in how you interpret Theorem 19.39. Clearly, the faster you move along a curve, the faster the function $u(x, y)$ will vary, and one needs to take this into account when comparing the rates of change along different curves. The easiest way to normalize is to assume that the tangent vector $\mathbf{a} = \dot{\mathbf{x}}$ has norm 1, so $\| \mathbf{a} \| = 1$ and we

are going through $\mathbf{x}$ with unit speed. Once this is done, Theorem 19.39 is an immediate consequence of the Cauchy–Schwarz inequality (3.16). Indeed,

$$\left| \frac{\partial u}{\partial \mathbf{a}} \right| = | \mathbf{a} \cdot \nabla u | \leq \| \mathbf{a} \| \, \| \nabla u \| = \| \nabla u \|, \qquad \text{when} \qquad \| \mathbf{a} \| = 1,$$

with equality if and only if $\mathbf{a} = c \, \nabla u$ points in the same direction as the gradient. Therefore, the maximum rate of change is when $\mathbf{a} = \nabla u / \| \nabla u \|$ is the unit vector in the direction of the gradient, while the minimum is achieved when $\mathbf{a} = - \nabla u / \| \nabla u \|$ points in the opposite direction. As a result, Theorem 19.39 tells us how to move if we wish to minimize a scalar function as rapidly as possible.

Theorem A.11. *A curve $\mathbf{x}(t)$ will realize the steepest decrease in the scalar field $u(\mathbf{x})$ if and only if it satisfies the gradient flow equation*

$$\dot{\mathbf{x}} = - \nabla u, \qquad \text{or} \qquad \begin{aligned} \frac{dx}{dt} &= - \frac{\partial u}{\partial x}(x,y), \\ \frac{dy}{dt} &= - \frac{\partial u}{\partial y}(x,y). \end{aligned} \tag{A.23}$$

The only points at which the gradient does not tell us about the directions of increase/decrease are the *critical points*, which are, by definition, points where the gradient vanishes: $\nabla u = \mathbf{0}$. These include local maxima or minima of the function, i.e., mountain peaks or bottoms of valleys, as well as other types of critical points like saddle points that represent mountain passes. In such cases, we must look at the second or higher order derivatives to tell the directions of increase/decrease; see Section 19.3 for details.

Remark: Theorem A.11 forms the basis of gradient descent methods for numerically approximating the maxima and minima of functions. One begins with a guess (x_0, y_0) for the minimum and then follows the gradient flow in to the minimum by numerically integrating the system of ordinary differential equations (A.23). This idea will be developed in detail in Chapter 19.

Example A.12. Consider the function $u(x,y) = x^2 + 2\,y^2$. Its gradient vector field is $\nabla u = (\, 2\,x, 4\,y \,)^T$, and hence the gradient flow equations (A.23) take the form

$$\dot{x} = -2\,x, \qquad \dot{y} = -4\,y.$$

The solution that starts out at initial position $(\, x_0, y_0 \,)^T$ is

$$\mathbf{x}(t) = x_0 \, e^{-2\,t}, \qquad y(t) = y_0 \, e^{-4\,t}. \tag{A.24}$$

Note that the origin is a stable fixed point for this linear dynamical system, and the solutions $\mathbf{x}(t) \to \mathbf{0}$ converge exponentially fast to the minimum of the function $u(x,y)$. If we start out not on either of the coordinate axes, so $x_0 \neq 0$ and $y_0 \neq 0$, then the trajectory (A.24) is a semi-parabola of the form $y = c x^2$, where $c = y_0 / x_0^2$; see the first picture in Figure A.7 ■ . These curves, along with the four coordinate semi-axes, are the paths to follow to get to the minimum $\mathbf{0}$ the fastest.

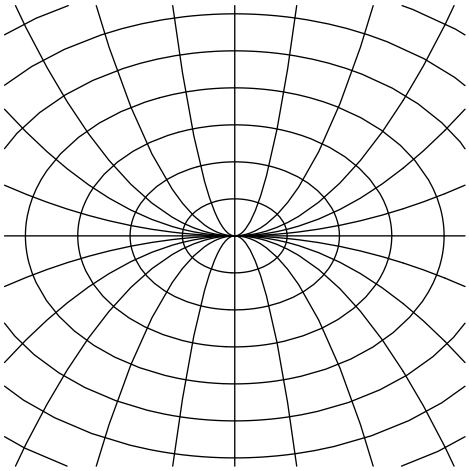


Figure A.7. Orthogonal System of Ellipses and Parabolas.

Level Sets

Let $u(x, y)$ be a scalar field. The curves defined by the implicit equation

$$u(x, y) = c \tag{A.25}$$

holding the function $u(x, y)$ constant are known as its *level sets*. For instance, if $u(x, y)$ represents the elevation of a mountain range, then its level sets are the usual contour lines on a topographic map. The Implicit Function Theorem tells us that, away from critical points, the level sets of a planar function are simple, though not necessarily closed, curves.

Theorem A.13. *If $u(x, y)$ has continuous partial derivatives, and, at a point, $\nabla u(x_0, y_0) \neq \mathbf{0}$, then the level set passing through the point $(\, x_0, y_0 \,)^T$ is a smooth curve near the point in question.*

Critical points, where $\nabla u = \mathbf{0}$, are either isolated points, or points of intersection of level sets. For example the level sets of the function $u(x, y) = 3x^2 - 2x^3 + y^2$ are plotted in Figure ls■. The function has critical points at $(\, 0, 0 \,)^T$ and $(\, 0, 1 \,)^T$. The former is a local minimum, and forms an isolated level point, while the latter is a saddle point, and is the point of intersection of the level curves $\{\, u = 1 \,\}$.

If we parametrize an individual level set by $\mathbf{x}(t) = (\, x(t), y(t) \,)^T$, then (A.25) tells us that the composite function $u(x(t), y(t)) = c$ is constant along the curve and hence its derivative

$$\frac{d}{dt}\, u(x(t), y(t)) = \nabla u \cdot \dot{\mathbf{x}} = 0$$

vanishes. We conclude that the tangent vector $\dot{\mathbf{x}}$ to the level set is *orthogonal* to the gradient direction ∇u at each point. In this manner, we have established the following additional important interpretation of the gradient, which is illustrated in Figure A.8.

Theorem A.14. *The gradient ∇u of a scalar field u is everywhere orthogonal to its level sets $\{\, u = c \,\}$.*

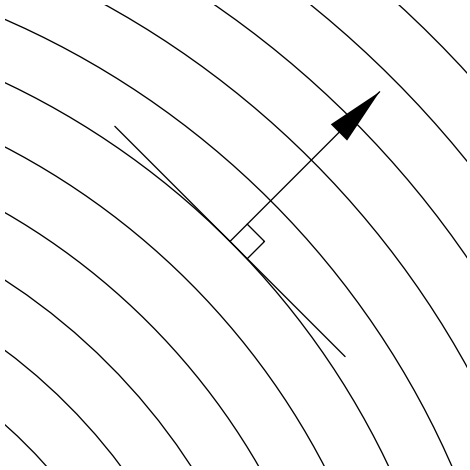


Figure A.8. Level Sets and Gradient.

Comparing Theorems A.11 and A.14, we conclude that the curves of steepest descent are always orthogonal (perpendicular) to the level sets of the function. Thus, if we want to hike uphill the fastest, we should keep our direction of travel always perpendicular to the contour lines. Similarly, if $u(x, y)$ represents temperature in a planar body at position (x, y) then the level sets are the curves of constant temperature, known as the *isotherms*. Heat energy will flow in the negative gradient direction, and hence orthogonally to the isotherms.

Example A.15. Consider again the function $u(x, y) = x^2 + 2y^2$ from Example A.12. Its level sets $u(x, y) = x^2 + 2y^2 = c$ form a system of concentric ellipses centered at the origin, illustrated in the second picture in Figure A.7. Theorem A.14 implies that the parabolic trajectories (A.24) followed by the solutions to the gradient flow equations form an orthogonal system of curves to the ellipses. This is evident in the third picture in Figure A.7, showing that the ellipses and parabolas intersect everywhere at right angles.

A.5. Integrals on Curves.

As you know, integrals of scalar functions, $\displaystyle\int_a^b f(t)\,dt$, are taken along real intervals $[a, b] \subset \mathbb{R}$. In higher dimensional calculus, there are a variety of possible types of integrals. The closest in spirit to one-dimensional integration are "line[†] integrals", in which one integrates along a curve. In planar calculus, line integrals come in three flavors. The most basic are the integrals of scalar functions with respect to arc length. Such integrals are used to compute lengths of curves, and masses of one-dimensional objects like strings and wires. The second and third varieties are used to integrate a vector field along a curve. Integrating the tangential component of the vector field is used, for instance, to compute work and measure circulation along the curve. The last type integrates the normal component of the

[†] A more accurate term would be "curve integral", but the terminology is standard and will not be altered in this text.

vector field along the curve, and represents flux (fluid, heat, electromagnetic, etc.) along the curve.

Arc Length

The *length* of the plane curve $\mathbf{x}(t)$ over the parameter range $a \le t \le b$ is computed by integrating the (Euclidean) norm[†] of its tangent vector:

$$\mathcal{L}(C) = \int_a^b \left\| \frac{d\mathbf{x}}{dt} \right\| \, dt = \int_a^b \sqrt{\dot{x}^2 + \dot{y}^2} \, dt. \tag{A.26}$$

The formula is justified by taking the limit of sums of lengths of small approximating line segments, [**9**]. For example, if the curve is given as the graph of a function $y = f(x)$ for $a \le x \le b$, then its length is computed by the familiar calculus formula

$$\mathcal{L}(C) = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \, dx. \tag{A.27}$$

It is important to verify that the length of a curve does not depend upon any particular parametrization (or even direction of traversal) of the curve.

Example A.16. The length of a circle $\mathbf{x}(t) = \begin{pmatrix} r\cos t \\ r \sin t \end{pmatrix}$, $0 \le t \le 2\pi$, of radius r is given by

$$\mathcal{L}(C) = \int_0^{2\pi} \left\| \frac{d\mathbf{x}}{dt} \right\| \, dt = \int_0^{2\pi} r \, dt = 2\pi r,$$

verifying the well-known formula for its circumference. On the other hand, the curve

$$\mathbf{x}(t) = \begin{pmatrix} a\cos t \\ b \sin t \end{pmatrix}, \qquad 0 \le t \le 2\pi, \tag{A.28}$$

parametrizes an ellipse with semi-axes a, b. Its arc length is given by the integral

$$s = \int_0^{2\pi} \sqrt{a^2 \sin^2 t + b^2 \cos^2 t} \, dt.$$

Unfortunately, this integral cannot be expressed in terms of elementary functions. It is, in fact, an *elliptic integral*, [**48**], so named for this very reason!

A curve is said to be parametrized by *arc length*, written $\mathbf{x}(s) = (\, x(s), y(s) \,)^T$, if one traverses it with constant, unit speed, which means that

$$\left\| \frac{d\mathbf{x}}{ds} \right\| = 1 \tag{A.29}$$

[†] Alternative norms lead to alternative notions of curve length, of importance in the study of curved spaces in differential geometry. In Einstein's theory of relativity, one allows the norm to vary from point to point, and hence length will vary over space.

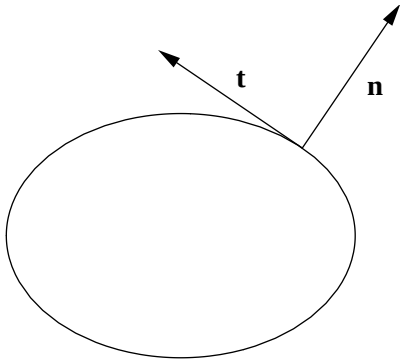


Figure A.9. The Moving Frame for an Ellipse.

at all points. In other words, the length of that part of the curve between arc length parameter values $s = s_0$ and $s = s_1$ is exactly equal to $s_1 - s_0$. To convert from a more general parameter t to arc length $s = \sigma(t)$, we must compute

$$s = \sigma(t) = \int_a^t \left\| \frac{d\mathbf{x}}{dt} \right\| dt \qquad \text{and so} \qquad ds = \| \dot{\mathbf{x}} \| \, dt = \sqrt{\dot{x}^2 + \dot{y}^2}\, dt. \tag{A.30}$$

The *unit tangent* to the curve at each point is obtained by differentiating with respect to the arc length parameter:

$$\mathbf{t} = \frac{d\mathbf{x}}{ds} = \frac{\dot{\mathbf{x}}}{\| \dot{\mathbf{x}} \|} = \left(\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right), \qquad \text{so that} \qquad \| \mathbf{t} \| = 1. \tag{A.31}$$

(As always, we require $\dot{\mathbf{x}} \neq \mathbf{0}$.) The *unit normal* to the curve is orthogonal to the unit tangent,

$$\mathbf{n} = \mathbf{t}^{\perp} = \left(\frac{dy}{ds}, -\frac{dx}{ds} \right) = \left(\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \frac{-\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \right), \qquad \text{so that} \qquad \begin{array}{l} \| \mathbf{n} \| = 1, \\ \mathbf{n} \cdot \mathbf{t} = 0. \end{array} \tag{A.32}$$

At each point on the curve, the vectors $\mathbf{t}, \mathbf{n}$ form an orthonormal basis of $\mathbb{R}^2$ known as the *moving frame* along the curve. For example, for the ellipse (A.28) with semi-axes a, b, the unit tangent and normal are given by

$$\mathbf{t} = \frac{1}{a^2 + b^2} \begin{pmatrix} -a \sin t \\ b \cos t \end{pmatrix}, \qquad \mathbf{n} = \frac{1}{a^2 + b^2} \begin{pmatrix} b \cos t \\ a \sin t \end{pmatrix},$$

and graphed in Figure A.9. Actually, a curve has two unit normals at each point — one points to our right side and the other to our left side as we move along the curve. The normal $\mathbf{n}$ in (A.32) is the right-handed normal, and is the traditional one to choose; the opposite, left-handed normal is its negative $-\mathbf{n}$. If we traverse a simple closed curve in a counterclockwise direction, then the right-handed normal $\mathbf{n}$ is the unit *outward* normal, pointing to the curve's exterior.

Arc Length Integrals

We now explain how to integrate scalar functions along curves. Suppose first that C is a (piecewise) smooth curve that is parametrized by arc length, $\mathbf{x}(s) = (x(s), y(s))$ for $0 \le s \le \ell$, where ℓ is the total length of C. If $u(\mathbf{x}) = u(x, y)$ is any scalar field, we define its *arc length integral* along the curve C to be

$$\int_C u\, ds = \int_0^\ell u(x(s), y(s))\, ds. \tag{A.33}$$

For example, if $\rho(x, y)$ represents the density at position (x, y) of a wire bent in the shape of a curve C, then the arc length integral $\displaystyle\int_C \rho(x, y)\, ds$ computes the total mass of the wire. In particular, the length of the curve is (tautologously) given by

$$\mathcal{L}(C) = \int_C ds = \int_0^\ell ds = \ell. \tag{A.34}$$

If we use an alternative parametrization $\mathbf{x}(t)$, with $a \le t \le b$, then the arc length integral is computed using the change of parameter formula (A.30), and so

$$\int_C u\, ds = \int_a^b u(\mathbf{x}(t)) \left\| \frac{d\mathbf{x}}{dt} \right\| dt = \int_a^b u(x(t), y(t)) \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt. \tag{A.35}$$

Changing the orientation of the curve does *not* alter the value of this type of line integral. Moreover, if we break up the curve into two nonoverlapping pieces, then the arc length integral decomposes into a sum:

$$\int_C u\, ds = \int_{-C} u\, ds, \qquad \int_C u\, ds = \int_{C_1} u\, ds \;+\; \int_{C_2} u\, ds, \qquad C = C_1 \cup C_2. \tag{A.36}$$

Example A.17. A circular wire radius 1 has density proportional to the distance of the point from the x axis. The mass of the wire is computed by the arc length integral

$$\oint_C |\, y\,|\, ds = \int_0^{2\pi} |\sin t\,|\, dt = 4.$$

The arc length integral was evaluated using the parametrization $\mathbf{x}(t) = (\, a\cos t, a \sin t\,)^T$ for $0 \le t \le 2\pi$, whereby $ds = \| \dot{\mathbf{x}} \|\, dt = dt$.

Line Integrals of Vector Fields

There are two intrinsic ways of integrating a vector field along a curve. In the first version, we integrate its tangential component $\mathbf{v} \cdot \mathbf{t}$, where $\mathbf{t} = d\mathbf{x}/ds$ is the unit tangent vector, with respect to arc length.

Definition A.18. The *line integral* of a vector field $\mathbf{v}$ along a parametrized curve $\mathbf{x}(t)$ is given by

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C v_1(x,y)\,dx + v_2(x,y)\,dy = \int_C \mathbf{v} \cdot \mathbf{t}\; ds. \tag{A.37}$$

To evaluate the line integral, we parametrize the curve by $\mathbf{x}(t)$ for $a \le t \le b$, and then

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_a^b \mathbf{v}(\mathbf{x}(t)) \, \frac{d\mathbf{x}}{dt}\, dt = \int_a^b \left[v_1(x(t),y(t)) \, \frac{dx}{dt} + v_2(x(t),y(t)) \, \frac{dy}{dt} \right] dt. \tag{A.38}$$

This result follows from the formulae (A.30), (A.31) for the arc length and unit tangent vector. In general, line integrals are independent of how the curve is parametrized — as long as it is traversed in the same direction. Reversing the direction of parameterization, i.e., changing the orientation of the curve, changes the sign of the line integral — because it reverses the direction of the unit tangent. As before, line integrals can be decomposed into sums over components:

$$\int_{-C} \mathbf{v} \cdot d\mathbf{x} = -\int_C \mathbf{v} \cdot d\mathbf{x}, \qquad \int_C \mathbf{v} \cdot d\mathbf{x} = \int_{C_1} \mathbf{v} \cdot d\mathbf{x} \; + \; \int_{C_2} \mathbf{v} \cdot d\mathbf{x}, \qquad C = C_1 \cup C_2. \tag{A.39}$$

In the second formula, one must take care to orient the two parts C_1, C_2 in the same direction as C.

Example A.19. Let C denote the circle of radius r centered at the origin. We will compute the line integral of the rotational vector field (A.19), namely

$$\oint_C \mathbf{v} \cdot d\mathbf{x} = \oint_C \frac{y\,dx - x\,dy}{x^2 + y^2}\,.$$

The circle on the integral sign serves to remind us that we are integrating around a closed curve. We parameterize the circle by

$$x(t) = r\cos t, \qquad y(t) = r \sin t, \qquad 0 \le t \le 2\,\pi.$$

Applying the basic line integral formula (A.38), we find

$$\oint_C \frac{y\,dx - x\,dy}{x^2 + y^2} = \int_0^{2\,\pi} \frac{-\,r^2 \sin^2 t - r^2 \cos^2 t}{r^2}\, dt = -\,2\,\pi,$$

independent of the circle's radius. Note that the parametrization goes around the circle once in the counterclockwise direction. If we go around once in the clockwise direction, e.g., by using the parametrization $\mathbf{x}(t) = (r\sin t, r\cos t)$, then the resulting line integral equals $+\,2\,\pi$.

If $\mathbf{v}$ represents the velocity vector field of a steady state fluid, then the line integral (A.37) represents the *circulation* of the fluid around the curve. Indeed, $\mathbf{v} \cdot \mathbf{t}$ is proportional to the force exerted by the fluid in the direction of the curve, and so the circulation integral measures the average of the tangential fluid forces around the curve. Thus, for example,

the rotational vector field (A.19) has a net circulation of -2π around any circle centered at the origin. The minus sign tells us that the fluid is circulating in the clockwise direction — opposite to the direction in which we went around the circle.

A fluid flow is *irrotational* if the circulation is zero for all closed curves. An irrotational flow will not cause a paddle wheel to rotate — there will be just as much fluid pushing in one direction as in the opposite, and the net tangential forces will cancel each other out. The connection between circulation and the curl of the velocity vector field will be made evident shortly.

If the vector field happens to be the gradient of a scalar field, then we can readily evaluate its line integral.

Theorem A.20. *If $\mathbf{v} = \nabla u$ is a gradient vector field, then its line integral*

$$\int_C \nabla u \cdot d\mathbf{x} = u(\mathbf{b}) - u(\mathbf{a}) \tag{A.40}$$

equals the difference in values between the potential function at the endpoints $\mathbf{a} = \mathbf{x}(a)$ and $\mathbf{b} = \mathbf{x}(b)$ of the curve C.

Thus, the line integral of a gradient field is *independent of path*; its value does not depend on how you get from point $\mathbf{a}$ to point $\mathbf{b}$. In particular, if C is a closed curve, then

$$\oint_C \nabla u \cdot d\mathbf{x} = 0,$$

since the endpoints coincide: $\mathbf{a} = \mathbf{b}$. In fact, independence of path is both necessary and sufficient for the vector field to be a gradient.

Theorem A.21. *Let $\mathbf{v}$ be a vector field defined on a domain Ω. Then the following are equivalent:*

(a) *The line integral $\displaystyle\int_C \mathbf{v} \cdot d\mathbf{x}$ is independent of path.*

(b) $\displaystyle\oint_C \mathbf{v} \cdot d\mathbf{x} = 0$ *for every closed curve C.*

(c) $\mathbf{v} = \nabla u$ *is the gradient of some potential function defined on Ω.*

In such cases, a potential function can be computed by integrating the vector field

$$u(\mathbf{x}) = \int_{\mathbf{a}}^{\mathbf{x}} \mathbf{v} \cdot d\mathbf{x}. \tag{A.41}$$

Here $\mathbf{a}$ is any fixed point (which defines the zero potential level), and we evaluate the line integral over any curve that connects $\mathbf{a}$ to $\mathbf{x}$; path-independence says that it does not matter which curve we use to get from $\mathbf{a}$ to $\mathbf{x}$. The proof that $\nabla u = \mathbf{v}$ is left as an exercise.

Example A.22. The line integral

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C (x^2 - 3\,y)\,dx + (2 - 3\,x)\,dy$$

of the vector field $\mathbf{v} = \big(\, x^2 - 3\,y, 2 - 3\,x \,\big)^T$ is independent of path. Indeed, parametrizing a curve C by $(x(t), y(t))$, $a \le t \le b$, leads to

$$\int_C (x^2 - 3\,y)\,dx + (2 - 3\,x)\,dy = \int_a^b \left[(x^2 - 3\,y)\,\frac{dx}{dt} + (2 - 3\,x)\,\frac{dy}{dt} \right] dt$$
$$= \int_a^b \frac{d}{dt}\big(\, x^3 - 3\,x\,y + 2\,y \,\big)\,dt = \big(\, x^3 - 3\,x\,y + 2\,y \,\big)\,\Big|_{t=a}^{b}.$$

The result only depends on the endpoints $\mathbf{a} = (\, x(a), y(a) \,)^T$, $\mathbf{b} = (\, x(b), y(b) \,)^T$, and not on the detailed shape of the curve. Integrating from $\mathbf{a} = \mathbf{0}$ to $\mathbf{b} = (x, y)$ produces the potential function

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

As guaranteed by (A.41), $\nabla u = \mathbf{v}$.

On the other hand, the line integral

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C (x^3 - 2\,y)\,dx + x^2\,dy$$

of the vector field $\mathbf{v} = \big(\, x^3 - 2\,y, x^2 \,\big)^T$ is not path-independent, and so $\mathbf{v}$ does not admit a potential function. Indeed, integrating from $(0, 0)$ to $(1, 1)$ along the straight line segment $\{\, (t, t) \,|\, 0 \le t \le 1 \,\}$, produces

$$\int_C (x^3 - 2\,y)\,dx + x^2\,dy = \int_0^1 \big(\, t^3 - 2\,t + t^2 \,\big)\,dt = -\tfrac{5}{12}.$$

On the other hand, integrating along the parabola $\{\, (t, t^2) \,|\, 0 \le t \le 1 \,\}$, yields a different value

$$\int_C (x^3 - 2\,y)\,dx + x^2\,dy = \int_0^1 \big(\, t^3 - 2\,t^2 + 2\,t^3 \,\big)\,dt = \tfrac{1}{12}.$$

If $\mathbf{v}$ represents a force field, then the line integral (A.37) represents the amount of *work* required to move along the given curve. Work is defined as force, or, more correctly, the tangential component of the force in the direction of motion, times distance. The line integral effectively totals up the infinitesimal contributions, the sum total representing the total amount of work expended in moving along the curve. Note that the work is independent of the parametrization of the curve. In other words (and, perhaps, counter-intuitively[†]), the amount of work expended doesn't depend upon how fast you move along the curve.

According to Theorem A.21, the work does not depend on the route you use to get from one point to the other if and only if the force field admits a potential function: $\mathbf{v} = \nabla u$. Then, by (A.40), the work is just the difference in potential at the two points. In

[†] The reason this doesn't agree with our intuition about work is that we are not taking frictional effects into account, and these are typically velocity-dependent.

particular, for a gradient vector field there is no net work required to go around a closed path.

Flux

The second type of line integral is found by integrating the normal component of the vector field along the curve:

$$\int_C \mathbf{v} \cdot \mathbf{n} \, ds. \tag{A.42}$$

Using the formula (A.32) for the unit normal, we find that the inner product can be rewritten in the alternative form

$$\mathbf{v} \cdot \mathbf{n} = v_1 \, \frac{dy}{ds} - v_2 \, \frac{dx}{ds} = \mathbf{v}^\perp \cdot \mathbf{t} \, ,$$

where $\mathbf{t} = d\mathbf{x}/ds$ is the unit tangent, while

$$\mathbf{v}^\perp = (\, -v_2, v_1 \,)^T \tag{A.43}$$

is a vector field that is everywhere orthogonal to the velocity vector field $\mathbf{v} = (\, v_1, v_2 \,)^T$. Thus, the normal line integral (A.42) can be rewritten as a tangential line integral

$$\int_C \mathbf{v} \cdot \mathbf{n} \, ds = \int_C v_1 \, dy - v_2 \, dx = \int_C \mathbf{v} \wedge d\mathbf{x} = \int_C \mathbf{v}^\perp \cdot \, d\mathbf{x} = \int_C \mathbf{v}^\perp \cdot \mathbf{t} \, ds. \tag{A.44}$$

If $\mathbf{v}$ represents the velocity vector field for a steady-state fluid flow, then the inner product $\mathbf{v} \cdot \mathbf{n}$ with the unit normal measures the *flux* of the fluid flow across the curve at the given point. The flux is positive if the fluid is moving in the normal direction $\mathbf{n}$ and negative if it is moving in the opposite direction. If the vector field admits a potential, $\mathbf{v} = \nabla u$, then the flux

$$\mathbf{v} \cdot \mathbf{n} = \nabla u \cdot \mathbf{n} = \frac{\partial u}{\partial \mathbf{n}} \tag{A.45}$$

equals its *normal derivative*, i.e., the directional derivative of the potential function u in the normal direction to the curve. The line integral $\displaystyle\int_C \mathbf{v} \cdot \mathbf{n} \, ds$ sums up the individual fluxes, and so represents the total flux across the curve, meaning the total volume of fluid that passes across the curve per unit time — in the direction assigned by the unit normal $\mathbf{n}$. In particular, if C is a simple closed curve and $\mathbf{n}$ is the outward normal, then the flux integral (A.42) measures the net outflow of fluid across C; if negative, it represents an inflow. The total flux is zero if and only if the total amount of fluid contained within the curve does not change. Thus, in the absence of sources or sinks, an incompressible fluid, such as water, will have zero net flux around any closed curve since the total amount of fluid within any given region cannot change.

Example A.23. For the radial vector field

$$\mathbf{v} = \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \qquad \text{we have} \qquad \mathbf{v}^\perp = \begin{pmatrix} -y \\ x \end{pmatrix}.$$

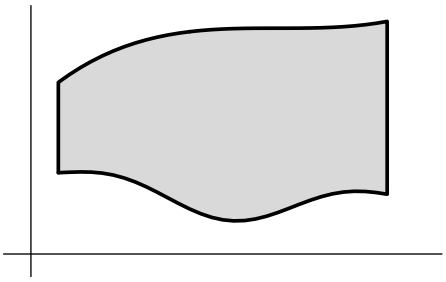


Figure A.10. Double Integration Domain.

As we saw in Example A.7, $\mathbf{v}$ represents the fluid flow due to a source at the origin. Thus, the resulting fluid flux across a circle C of radius r is computed using the line integral

$$\oint_C \mathbf{v} \cdot \mathbf{n}\, ds = \oint_C x\, dy - y\, dx = \int_0^{2\pi} r^2 \sin^2 t + r^2 \cos^2 t\, dt = 2\pi\, r^2.$$

Therefore, the source fluid flow has a net outflow of $2\pi\, r^2$ units across a circle of radius r. This is not an incompressible flow!

A.6. Double Integrals.

We assume that the student is familiar with the foundations of multiple integration, and merely review a few of the highlights in this section. Given a scalar function $u(x, y)$ defined on a domain Ω, its *double integral*

$$\iint_\Omega u(x, y)\, dx\, dy = \iint_\Omega u(\mathbf{x})\, d\mathbf{x} \tag{A.46}$$

is equal to the volume of the solid lying underneath the graph of u over Ω. If $u(x, y)$ represents the density at position $(x, y)^T$ in a plate having the shape of the domain Ω, then the double integral (A.46) measures the total mass of the plate. In particular,

$$\text{area } \Omega = \iint_\Omega dx\, dy$$

is equal to the *area* of the domain Ω.

In the particular case when

$$\Omega = \big\{\, \varphi(x) < y < \psi(x), \quad a < x < b \,\big\} \tag{A.47}$$

is given as the region lying between the graphs of two functions, as in Figure A.10, then we can evaluate the double integral by repeated scalar integration,

$$\iint_\Omega u(x, y)\, dx\, dy = \int_a^b \left(\int_{\varphi(x)}^{\psi(x)} u(x, y)\, dy \right) dx, \tag{A.48}$$

in the two coordinate directions. Fubini's Theorem states that one can equally well evaluate the integral in the reverse order

$$\iint_\Omega u(x,y)\,dx\,dy = \int_c^d \left(\int_{\alpha(y)}^{\beta(y)} u(x,y)\,dx \right) dy \tag{A.49}$$

in the case

$$\Omega = \big\{\, \alpha(y) < x < \beta(y), \quad c < y < d \,\big\} \tag{A.50}$$

lies between the graphs of two functions of y.

Example A.24. Compute the volume of the solid lying under the positive part of the paraboloid $z = 1 - x^2 - y^2$. Note that $z > 0$ if and only if $x^2 + y^2 < 1$, and hence we should evaluate the double integral

$$\iint_\Omega (1 - x^2 - y^2)\,dx\,dy$$

over the unit disk $\Omega = \big\{\, x^2 + y^2 < 1 \,\big\}$. We may represent the disk in the form (A.47), so that

$$\Omega = \big\{ -\sqrt{1-x^2} < y < \sqrt{1-x^2}, \quad -1 < x < 1 \,\big\}.$$

Therefore, we evaluate the volume by repeated integration

$$\begin{aligned} \iint_\Omega [\,1 - x^2 - y^2\,]\,dx\,dy &= \int_{-1}^1 \left[\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (1 - x^2 - y^2)\,dy \right] dx \\ &= \int_{-1}^1 \left[\,(y - x^2\,y - \tfrac{1}{3}\,y^3)\,\Big|_{y=-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \right] dx = \int_{-1}^1 \tfrac{4}{3}(1-x^2)^{3/2}\,dx = \tfrac{1}{2}\,\pi. \end{aligned}$$

The final integral is most easily effected via a trigonometric substitution.

Alternatively, and much easier, one can use polar coordinates to evaluate the integral. The unit disk takes the form $D = \{\, 0 \le r < 1, 0 \le \theta < 2\,\pi \,\}$, and so

$$\iint_D (1 - x^2 - y^2)\,dx\,dy = \iint_D (1 - r^2)\,r\,dr\,d\theta = \int_0^1 \left(\int_0^{2\,\pi} (r - r^3)\,d\theta \right) dr = \tfrac{1}{2}\,\pi.$$

We are using the standard formula

$$dx\,dy = r\,dr\,d\theta \tag{A.51}$$

for the area element in polar coordinates, [**9**].

The polar integration formula (A.51) is a consequence of the general change of variables formula for double integrals. If

$$x = x(s,t), \qquad\qquad y = y(s,t),$$

is an invertible change of variables that maps $(\,s,t\,)^T \in D$ to $(\,x,y\,)^T \in \Omega$, then

$$\iint_\Omega u(x,y)\,dx\,dy = \iint_D stU(s,t) \left| \frac{\partial(x,y)}{\partial(s,t)} \right|. \tag{A.52}$$

Here $U(s,t) = u(x(s,t), y(s,t))$ denotes the function when rewritten in the new variables, while

$$\frac{\partial(x,y)}{\partial(s,t)} = \det \begin{pmatrix} x_s & x_t \\ y_s & y_t \end{pmatrix} = \frac{\partial x}{\partial s}\,\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\,\frac{\partial y}{\partial s} \tag{A.53}$$

is the *Jacobian determinant* of the functions x, y with respect to the variables s, t, which measures the local change in area under the map.

In the event that the domain of integration is more complicated than either (A.47) or (A.50), then one performs "surgery" by chopping up the domain

$$\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_k$$

into smaller pieces. The pieces Ω_i are not allowed to overlap, and so have at most their boundary curves in common. The double integral

$$\iint_\Omega u(x,y)\,dx\,dy = \iint_{\Omega_1} u(x,y)\,dx\,dy + \;\cdots\; + \iint_{\Omega_k} u(x,y)\,dx\,dy \tag{A.54}$$

can then be evaluated as a sum of the double integrals over the individual pieces.

A.7. Green's Theorem.

For double integrals, the role of the Fundamental Theorem of Calculus is played by *Green's Theorem*. The Fundamental Theorem relates an integral over an interval $I = [a,b]$ to an evaluation at the boundary $\partial I = \{a, b\}$, which consists of the two endpoints of the interval. In a similar manner, Green's Theorem relates certain double integrals over a planar domain Ω to line integrals around its boundary curve(s) $\partial\Omega$.

Theorem A.25. *Let $\mathbf{v}(\mathbf{x})$ be a smooth vector field defined on a bounded domain $\Omega \subset \mathbb{R}^2$. Then the line integral of $\mathbf{v}$ around the boundary $\partial\Omega$ equals the double integral of the curl of $\mathbf{v}$ over the domain. This result can be written in either of the equivalent forms*

$$\iint_\Omega \nabla \wedge \mathbf{v}\,d\mathbf{x} = \oint_{\partial\Omega} \mathbf{v}\cdot\,d\mathbf{x}, \qquad \iint_\Omega \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}\right) dx\,dy = \oint_{\partial\Omega} v_1\,dx + v_2\,dy\,. \tag{A.55}$$

An outline of the proof appears in Exercise ■. Green's Theorem was first formulated in 1828 by the English mathematician and miller George Green, and, contemporaneously, by the Russian mathematician Mikhail Ostrogradski.

Example A.26. Let us apply Green's Theorem A.25 to the particular vector field $\mathbf{v} = (\,0, x\,)^T$. Since $\nabla \wedge \mathbf{v} \equiv 1$, we find

$$\oint_{\partial\Omega} x\,dy = \iint_\Omega dx\,dy = \text{area } \Omega.$$

This means that we can compute the area of a planar domain by computing the indicated line integral around its boundary! For example, to compute the area of a disk D_r of radius r, we parametrize its bounding circle C_r by $(\,r\cos t, r\sin t\,)^T$ for $0 \le t \le 2\pi$, and compute

$$\text{area } D_r = \oint_{C_r} x\,dy = \int_0^{2\pi} r^2\cos^2 t\,dt = \pi\,r^2.$$

If we interpret $\mathbf{v}$ as the velocity vector field associated with a steady state fluid flow, then the right hand side of formula (A.55) represents the circulation of the fluid around the boundary of the domain Ω. Green's Theorem implies that the double integral of the curl of the velocity vector must equal this circulation line integral.

According to (mean2), if we divide the double integral in (A.55) by the area of the domain,

$$\frac{1}{\text{area } \Omega} \iint_{\Omega} \nabla \wedge \mathbf{v}\, d\mathbf{x} = \mathrm{M}_{\Omega}\,[\,\nabla \wedge \mathbf{v}\,],$$

we obtain the mean of the curl $\nabla \wedge \mathbf{v}$ of the vector field over the domain. In particular, if the domain Ω is very small, then $\nabla \wedge \mathbf{v}$ does not vary much, and so its value at any point in the domain is more or less equal to the mean. On the other hand, the right hand side of (A.55) represents the circulation around the boundary $\partial\Omega$. Thus, we conclude that the curl $\nabla \wedge \mathbf{v}$ of the velocity vector field represents an "infinitesimal circulation" at the point it is evaluated. In particular, the fluid is irrotational, with no net circulation around any curve, if and only if $\nabla \wedge \mathbf{v} \equiv 0$ everywhere. Under the assumption that its domain of definition is simply connected, Theorem A.21 tell us that this is equivalent to the existence of a velocity potential u with $\nabla u = \mathbf{v}$.

We can also apply Green's Theorem A.25 to flux line integrals of the form (A.42). Using the identification (A.44) followed by (A.55), we find that

$$\oint_{\partial\Omega} \mathbf{v} \cdot \mathbf{n}\, ds = \oint_{\partial\Omega} \mathbf{v}^{\perp} \cdot\, d\mathbf{x} = \iint_{\Omega} \nabla \wedge \mathbf{v}^{\perp}\, dx\, dy.$$

However, note that the curl of the orthogonal vector field (A.43), namely

$$\nabla \wedge \mathbf{v}^{\perp} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} = \nabla \cdot \mathbf{v}, \tag{A.56}$$

coincides with the *divergence* of the original velocity field. Combining these together, we have proved the divergence or flux form of Green's Theorem:

$$\iint_{\Omega} \nabla \cdot \mathbf{v}\, dx\, dy = \oint_{\partial\Omega} \mathbf{v} \cdot \mathbf{n}\, ds. \tag{A.57}$$

As before, Ω is a bounded domain, and $\mathbf{n}$ is the unit outward normal to its boundary $\partial\Omega$.

In the fluid flow interpretation, the right hand side of (A.57) represents the net fluid flux out of the region Ω. Thus, the double integral of the divergence of the flow vector must equal this net change in area. Thus, in the absence of sources or sinks, the divergence of the velocity vector field, $\nabla \cdot \mathbf{v}$ will represent the local change in area of the fluid at each point. In particular, if the fluid is incompressible if and only if $\nabla \cdot \mathbf{v} \equiv 0$ everywhere.

An ideal fluid flow is both incompressible, $\nabla \cdot \mathbf{v} = 0$, and irrotational, $\nabla \wedge \mathbf{v} = \mathbf{0}$. Assuming its domain is simply connected, we introduce velocity potential $u(x, y)$, so $\nabla u = \mathbf{v}$. Therefore,

$$0 = \nabla \cdot \mathbf{v} = \nabla \cdot \nabla u = u_{xx} + u_{yy}. \tag{A.58}$$

Therefore, the velocity potential for an incompressible, irrotational fluid flow is a harmonic function, i.e., it satisfies the Laplace equation! Water waves are typically modeled in this manner, and so many problems in fluid mechanics rely on the solution to Laplace's equation.

Appendix B

Vector Calculus in Three Dimensions

Before continuing on to the analysis of partial differential equations in three space dimensions, we should first review the fundamentals of three-dimensional vector calculus. The student is expected to have already encountered most of these topics in an introductory multi-variable calculus course. We shall be dealing with calculus on curves, surfaces and solid bodies in three-dimensional space. The three methods of integration — line, surface and volume (triple) integrals — and the fundamental vector differential operators — gradient, curl and divergence — are intimately related. The differential operators and integrals underlie the multivariate versions of the fundamental theorem of calculus, known as Stokes' Theorem and the Divergence Theorem.

All of these topics will be reviewed in rapid succession, with most details being relegated to the exercises. A more detailed development can be found in any reasonable multi-variable calculus text, including [**9**, **38**, **58**].

B.1. Dot and Cross Product.

We begin by reviewing the basic algebraic operations between vectors in three-dimensional space $\mathbb{R}^3$. We shall continue to use column vector notation

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = (v_1, v_2, v_3)^T \in \mathbb{R}^3.$$

The standard basis vectors of $\mathbb{R}^3$ are

$$\mathbf{e}_1 = \mathbf{i} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{e}_2 = \mathbf{j} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{e}_3 = \mathbf{k} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{B.1}$$

We prefer the former notation, as it easily generalizes to n-dimensional space. Any vector

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = v_1 \, \mathbf{e}_1 + v_2 \, \mathbf{e}_2 + v_3 \, \mathbf{e}_3$$

is a linear combination of the basis vectors. The coefficients he v_1, v_2, v_3 are the coordinates of the vector with respect to the standard basis.

Space comes equipped with an orientation — either right- or left-handed. One cannot alter[†] the orientation by physical motion, although looking in a mirror — or, mathematically, performing a reflection — reverses the orientation. The standard basis vectors are graphed with a right-hand orientation, as in Figure rhr■. When you point with your right hand, $\mathbf{e}_1$ lies in the direction of your index finger, $\mathbf{e}_2$ lies in the direction of your middle finger, and $\mathbf{e}_3$ is in the direction of your thumb. In general, a set of three linearly independent vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is said to have a *right-handed orientation* if they have the same orientation as the standard basis. It is not difficult to prove that this is the case if and only if the determinant of the 3×3 matrix whose columns are the given vectors is positive: $\det(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3) > 0$. Interchanging the order of the vectors may switch their orientation; for example if $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are right-handed, then $\mathbf{v}_2, \mathbf{v}_1, \mathbf{v}_3$ is left-handed.

We have already made extensive use of the Euclidean *dot product*

$$\mathbf{v} \cdot \mathbf{w} = v_1 w_1 + v_2 w_2 + v_3 w_3, \qquad \text{where} \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}, \tag{B.2}$$

along with the Euclidean norm

$$\| \mathbf{v} \| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + v_3^2}\,. \tag{B.3}$$

As in Definition 3.1, the dot product is bilinear, symmetric: $\mathbf{v} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{v}$, and positive. The Cauchy–Schwarz inequality

$$| \mathbf{v} \cdot \mathbf{w} | \;\leq\; \| \mathbf{v} \| \, \| \mathbf{w} \|. \tag{B.4}$$

implies that the dot product can be used to measure the angle θ between the two vectors $\mathbf{v}$ and $\mathbf{w}$:

$$\mathbf{v} \cdot \mathbf{w} = \| \mathbf{v} \| \, \| \mathbf{w} \| \, \cos\theta. \tag{B.5}$$

(See also (3.15).)

Remark: In this chapter, we will only use the Euclidean dot product and its associated norm. Adapting the constructions to more general norms and inner products is an interesting exercise, but will not concern us here.

Also of great importance — but particular to three-dimensional space — is the *cross product* between vectors. While the dot product produces a scalar, the three-dimensional cross product produces a vector, defined by the formula

$$\mathbf{v} \wedge \mathbf{w} = \begin{pmatrix} v_2 w_3 - v_3 w_2 \\ v_3 w_1 - v_1 w_3 \\ v_1 w_2 - v_2 w_1 \end{pmatrix} \qquad \text{where} \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}, \tag{B.6}$$

[†] This assumes that space is identified with the three-dimensional Euclidean space $\mathbb{R}^3$, or, more generally, an oriented three-dimensional manifold, [**dg**].

We have chosen to employ the more modern wedge notation rather the more traditional cross symbol, $\mathbf{v} \times \mathbf{w}$, for this quantity. The cross product formula is most easily memorized as a formal 3×3 determinant

$$\mathbf{v} \wedge \mathbf{w} = \det \begin{pmatrix} v_1 & w_1 & \mathbf{e}_1 \\ v_2 & w_2 & \mathbf{e}_2 \\ v_3 & w_3 & \mathbf{e}_3 \end{pmatrix} = (v_2 w_3 - v_3 w_2)\, \mathbf{e}_1 + (v_3 w_1 - v_1 w_3)\, \mathbf{e}_2 + (v_1 w_2 - v_2 w_1)\, \mathbf{e}_3, \tag{B.7}$$

involving the standard basis vectors (B.1). We note that, like the dot product, the cross product is a bilinear function, meaning that

$$(c\,\mathbf{u} + d\,\mathbf{v}) \wedge \mathbf{w} = c\,(\mathbf{u} \wedge \mathbf{w}) + d\,(\mathbf{v} \wedge \mathbf{w}), \qquad \mathbf{u} \wedge (c\,\mathbf{v} + d\,\mathbf{w}) = c\,(\mathbf{u} \wedge \mathbf{v}) + d\,(\mathbf{u} \wedge \mathbf{w}), \tag{B.8}$$

for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^3$ and any scalars $c, d \in \mathbb{R}$. On the other hand, unlike the dot product, the cross product is an *anti-symmetric* quantity

$$\mathbf{v} \wedge \mathbf{w} = -\,\mathbf{w} \wedge \mathbf{v}, \tag{B.9}$$

which changes its sign when the two vectors are interchanged. In particular, the cross product of a vector with itself is automatically zero:

$$\mathbf{v} \wedge \mathbf{v} = \mathbf{0}.$$

Geometrically, the cross product vector $\mathbf{u} = \mathbf{v} \wedge \mathbf{w}$ is orthogonal to the two vectors $\mathbf{v}$ and $\mathbf{w}$:

$$\mathbf{v} \cdot (\mathbf{v} \wedge \mathbf{w}) = \mathbf{0} = \mathbf{w} \cdot (\mathbf{v} \wedge \mathbf{w}).$$

Thus, when $\mathbf{v}$ and $\mathbf{w}$ are linearly independent, their cross product $\mathbf{u} = \mathbf{v} \wedge \mathbf{w} \neq \mathbf{0}$ defines a normal direction to the plane spanned by $\mathbf{v}$ and $\mathbf{w}$. The direction of the cross product is fixed by the requirement that $\mathbf{v}, \mathbf{w}, \mathbf{u} = \mathbf{v} \wedge \mathbf{w}$ form a right-handed triple. The length of the cross product vector is equal to the area of the parallelogram defined by the two vectors, which is

$$\| \mathbf{v} \wedge \mathbf{w} \| = \| \mathbf{v} \|\, \| \mathbf{w} \|\, |\sin \theta\,| \tag{B.10}$$

where θ is than angle between the two vectors, as in Figure para. Consequently, the cross product vector is zero, $\mathbf{v} \wedge \mathbf{w} = \mathbf{0}$, if and only if the two vectors are collinear (linearly dependent) and hence only span a line.

The *scalar triple product* $\mathbf{u} \cdot (\mathbf{v} \wedge \mathbf{w})$ between three vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ is defined as the dot product between the first vector with the cross product of the second and third vectors. The parenthesis is often omitted because there is only one way to make sense of $\mathbf{u} \cdot \mathbf{v} \wedge \mathbf{w}$. Combining (B.2), (B.7), shows that one can compute the triple product by the determinantal formula

$$\mathbf{u} \cdot \mathbf{v} \wedge \mathbf{w} = \det \begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix}. \tag{B.11}$$

By the properties of the determinant, permuting the order of the vectors merely changes the sign of the triple product:

$$\mathbf{u} \cdot \mathbf{v} \wedge \mathbf{w} = -\,\mathbf{v} \cdot \mathbf{u} \wedge \mathbf{w} = +\,\mathbf{v} \cdot \mathbf{w} \wedge \mathbf{u} = \cdots.$$

The triple product vanishes, $\mathbf{u} \cdot \mathbf{v} \wedge \mathbf{w} = 0$, if and only if the three vectors are linearly dependent, i.e., coplanar or collinear. The triple product is positive, $\mathbf{u} \cdot \mathbf{v} \wedge \mathbf{w} > 0$ if and only if the three vectors form a right-handed basis. Its magnitude $|\, \mathbf{u} \cdot \mathbf{v} \wedge \mathbf{w} \,|$ measures the volume of the parallelepiped spanned by the three vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$, as in Figure ppp.

B.2. Curves.

A *space curve* $C \subset \mathbb{R}^3$ is parametrized by a vector-valued function

$$\mathbf{x}(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} \in \mathbb{R}^3, \qquad a \leq t \leq b, \tag{B.12}$$

that depends upon a single parameter t that varies over some interval. We shall always assume that $\mathbf{x}(t)$ is continuously differentiable. The curve is *smooth* provided its *tangent vector* is continuous and everywhere *nonzero*:

$$\frac{d\mathbf{x}}{dt} = \dot{\mathbf{x}} = \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} \neq \mathbf{0}. \tag{B.13}$$

As in the planar situation, the smoothness condition (B.13) precludes the formulation of corners, cusps or other singularities in the curve.

Physically, we can think of a curve as the trajectory described by a particle moving in space. At each time t, the tangent vector $\dot{\mathbf{x}}(t)$ represents the instantaneous velocity of the particle. Thus, as long as the particle moves with nonzero speed, $\| \dot{\mathbf{x}} \| = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} > 0$, its trajectory is necessarily a smooth curve.

Example B.1. A charged particle in a constant magnetic field moves along the curve

$$\mathbf{x}(t) = \begin{pmatrix} \rho \cos t \\ \rho \sin t \\ c\, t \end{pmatrix}, \tag{B.14}$$

where $c > 0$ and $\rho > 0$ are positive constants. The curve describes a circular *helix* of *radius* ρ spiraling up the z axis. The parameter c determines the *pitch* of the helix, indicating how tightly its coils are wound; the smaller c is, the closer the winding. See Figure helix for an illustration. DNA is, remarkably, formed in the shape of a (bent and twisted) double helix. The tangent to the helix at a point $\mathbf{x}(t)$ is the vector

$$\dot{\mathbf{x}}(t) = \begin{pmatrix} -\rho \sin t \\ \rho \cos t \\ c \end{pmatrix}.$$

Note that the speed of the particle,

$$\| \dot{\mathbf{x}} \| = \sqrt{\rho^2 \sin^2 t + \rho^2 \cos^2 t + c^2} = \sqrt{\rho^2 + c^2}\,, \tag{B.15}$$

remains constant, although the velocity vector $\dot{\mathbf{x}}$ twists around.

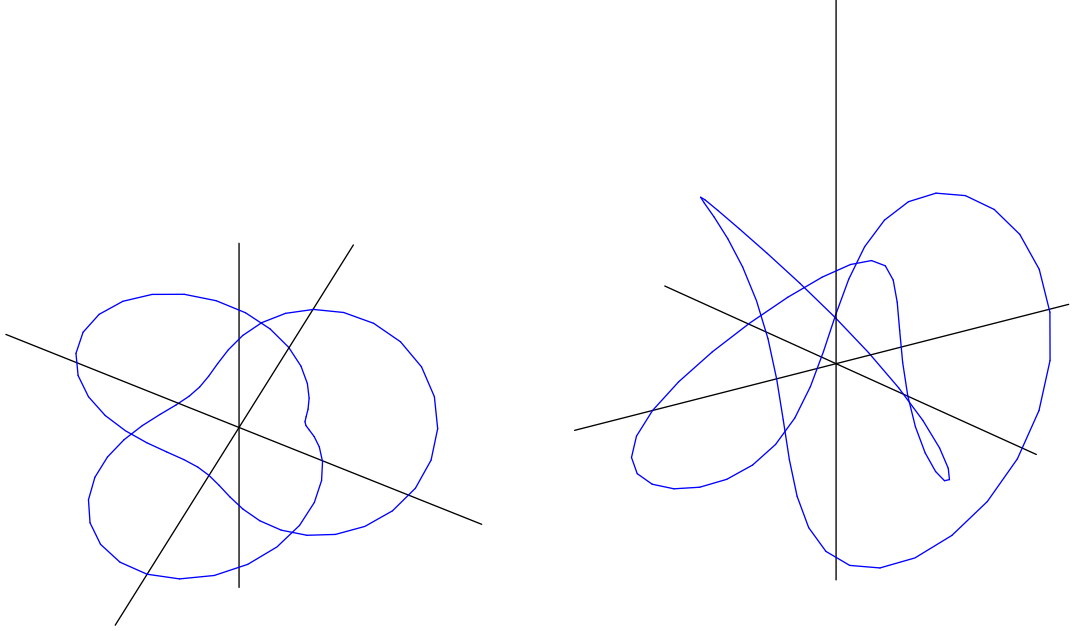


Figure B.1. Two Views of a Trefoil Knot.

Most of the terminology introduced in Chapter A for planar curves carries over to space curves without significant alteration. In particular, a curve is *simple* if it never crosses itself, and *closed* if its ends meet, $\mathbf{x}(a) = \mathbf{x}(b)$. In the plane, simple closed curves are all topologically equivalent, meaning one can be continuously deformed to the other. In space, this is no longer true. Closed curves can be knotted, and thus have nontrivial topology.

Example B.2. The curve

$$\mathbf{x}(t) = \begin{pmatrix} (2 + \cos 3\,t) \cos 2\,t \\ (2 + \cos 3\,t) \sin 2\,t \\ \sin 3\,t \end{pmatrix} \qquad \text{for} \qquad 0 \leq t \leq 2\,\pi, \tag{B.16}$$

describes a closed curve that is in the shape of a trefoil knot, as depicted in Figure B.1. The trefoil is a genuine knot, meaning it cannot be deformed into an unknotted circle without cutting and retying. (However, a rigorous proof of this fact is not easy.) The trefoil is the simplest of the "toroidal knots", investigated in more detail in Exercise ■.

The study and classification of knots is a subject of great historical importance. Indeed, they were first considered from a mathematical viewpoint in the nineteenth century, when the English applied mathematician William Thompson (later Lord Kelvin), [**Kelvin**], proposed a theory of atoms based on knots! In recent years, knot theory has witnessed a tremendous revival, owing to its great relevance to modern day mathematics and physics. We refer the interested reader to the advanced text [**92**] for details.

B.3. Line Integrals.

In Section A.5, we encountered three different types of line integrals along plane curves. Two of these — integrals with respect to arc length, (A.35), and circulation integrals, (A.37) — are directly applicable to space curves. On the other hand, for three-dimensional flows, the analog of the flux line integral (A.42) is a surface integral, and will be discussed later in the chapter.

Arc Length

The *length* of the space curve $\mathbf{x}(t)$ over the parameter range $a \leq t \leq b$ is computed by integrating the norm of its tangent vector:

$$\mathcal{L}(C) = \int_a^b \left\| \frac{d\mathbf{x}}{dt} \right\| dt = \int_a^b \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\; dt\,. \tag{B.17}$$

It is not hard to show that the length of the curve is independent of the parametrization — as it should be.

Starting at the endpoint $\mathbf{x}(a)$, the *arc length parameter* s is given by

$$s = \int_a^t \left\| \frac{d\mathbf{x}}{dt} \right\| dt \qquad \text{and so} \qquad ds = \| \dot{\mathbf{x}} \|\, dt = \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\; dt. \tag{B.18}$$

The arc length s measures the distance along the curve starting from the initial point $\mathbf{x}(a)$. Thus, the length of the part of the curve between $s = \alpha$ and $s = \beta$ is exactly $\beta - \alpha$. It is often convenient to reparametrize the curve by its arc length, $\mathbf{x}(s)$. This has the same effect as moving along the curve at unit speed, since, by the chain rule,

$$\frac{d\mathbf{x}}{ds} = \frac{dx}{dt}\,\frac{dt}{ds} = \frac{\dot{\mathbf{x}}}{\| \dot{\mathbf{x}} \|}\,, \qquad \text{so that} \qquad \left\| \frac{d\mathbf{x}}{ds} \right\| = 1.$$

Therefore $d\mathbf{x}/ds$ is the unit tangent vector pointing in the direction of motion along the curve.

Example B.3. The length of one turn of a helix (B.14) is, using (B.15),

$$\mathcal{L}(C) = \int_0^{2\pi} \left\| \frac{d\mathbf{x}}{dt} \right\| dt = \int_0^{2\pi} \sqrt{\rho^2 + c^2}\; dt = 2\,\pi\,\sqrt{\rho^2 + c^2}\,.$$

The arc length parameter, measured from the point $\mathbf{x}(0) = (\,r, 0, 0\,)^T$ is merely a rescaling,

$$s = \int_0^t \sqrt{\rho^2 + c^2}\; dt = \sqrt{\rho^2 + c^2}\; t,$$

of the original parameter t. When the helix is parametrized by arc length,

$$\mathbf{x}(s) = \left(\rho \cos \frac{s}{\sqrt{\rho^2 + c^2}}\,,\ \rho \sin \frac{s}{\sqrt{\rho^2 + c^2}}\,,\ \frac{c\,s}{\sqrt{\rho^2 + c^2}} \right)^T,$$

we move along it with unit speed. It now takes time $s = 2\,\pi\,\sqrt{\rho^2 + c^2}$ to complete on turn of the helix.

Example B.4. To compute the length of the trefoil knot (B.16), we begin by computing the tangent vector

$$\frac{d\mathbf{x}}{dt} = \begin{pmatrix} -2\,(2 + \cos 3t) \sin 2t - 3\,\sin 3t\,\cos 2t \\ 2\,(2 + \cos 3t) \cos 2t - 3\,\sin 3t\,\sin 2t \\ 3\,\cos 3t \end{pmatrix}.$$

After some algebra involving trigonometric identities, we find

$$\| \dot{\mathbf{x}} \| = \sqrt{27 + 16 \cos 3t + 2 \cos 6t} \, ,$$

which is never 0. Unfortunately, the resulting arc length integral

$$\int_0^{2\pi} \| \dot{\mathbf{x}} \| \, dt = \int_0^{2\pi} \sqrt{27 + 16 \cos 3t + 2 \cos 6t} \; dt$$

cannot be completed in elementary terms. Numerical integration can be used to find the approximate value 31.8986 for the length of the knot.

The *arc length integral* of a scalar field $u(\mathbf{x}) = u(x, y, z)$ along a curve C is

$$\int_C u \, ds = \int_0^{\ell} u(\mathbf{x}(s)) \, ds = \int_0^{\ell} u(x(s), y(s), z(s)) \, ds, \tag{B.19}$$

where ℓ is the total length of the curve. For example, if $\rho(x, y, z)$ represents the density at position $\mathbf{x} = (x, y, z)$ of a wire bent in the shape of the curve C, then $\displaystyle\int_C \rho \, ds$ represents the total mass of the wire. In particular, the integral

$$\int_C ds = \int_0^{\ell} ds = \ell$$

recovers the length of the curve.

If it is not convenient to work directly with the arc length parametrization, we can still compute the arc length integral in terms of the original parametrization $\mathbf{x}(t)$ for $a \leq t \leq b$. Using the change of parameter formula (B.18), we find

$$\int_C u \, ds = \int_a^b u(\mathbf{x}(t)) \, \| \dot{\mathbf{x}} \| \, dt = \int_a^b u(x(t), y(t), z(t)) \, \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} \, dt. \tag{B.20}$$

Example B.5. The density of a wire that is wound in the shape of a helix is proportional to its height. Let us compute the mass of one full turn of the helical wire. Thus, the density is given by $\rho(x, y, z) = a \, z$, where a is the constant of proportionality, and we are assuming $z \geq 0$. Substituting into (B.20), the total mass of the wire is

$$\mathcal{L}(C) = \int_C a \, z \, ds = \int_0^{2\pi} a \, c \, t \, \sqrt{r^2 + c^2} \, dt = 2 \, \pi^2 \, a \, c \, \sqrt{r^2 + c^2} \, .$$

Line Integrals of Vector Fields

As in the two-dimensional situation (A.37), the *line integral* of a vector field $\mathbf{v}$ along a parametrized curve $\mathbf{x}(t)$ is obtained by integration of its tangential component with respect to the arc length. The tangential component of $\mathbf{v}$ is given by

$$\mathbf{v} \cdot \mathbf{t}, \qquad \text{where} \qquad \mathbf{t} = \frac{d\mathbf{x}}{ds}$$

is the unit tangent vector to the curve. Thus, the line integral of $\mathbf{v}$ is written as

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C v_1(x,y,z)\,dx + v_2(x,y,z)\,dy + v_3(x,y,z)\,dz = \int_C \mathbf{v} \cdot \mathbf{t}\; ds. \qquad (B.21)$$

We can evaluate the line integral in terms of an arbitrary parametrization of the curve by the general formula

$$\begin{aligned} \int_C \mathbf{v} \cdot d\mathbf{x} &= \int_a^b \mathbf{v}(\mathbf{x}(t)) \cdot \frac{d\mathbf{x}}{dt}\; dt \qquad (B.22)\\ &= \int_a^b \left[v_1(x(t),y(t),z(t))\,\frac{dx}{dt} + v_2(x(t),y(t),z(t))\,\frac{dy}{dt} + v_3(x(t),y(t),z(t))\,\frac{dz}{dt} \right] dt. \end{aligned}$$

Line integrals in three dimensions enjoy all of the properties of their two-dimensional siblings: Reversing the direction of parameterization along the curve changes the sign; also, the integral can be decomposed into sums over components:

$$\int_{-C} \mathbf{v} \cdot d\mathbf{x} = -\int_C \mathbf{v} \cdot d\mathbf{x}, \qquad \int_C \mathbf{v} \cdot d\mathbf{x} = \int_{C_1} \mathbf{v} \cdot d\mathbf{x} \;+\; \int_{C_2} \mathbf{v} \cdot d\mathbf{x}, \qquad C = C_1 \cup C_2. \qquad (B.23)$$

If $\mathbf{f}(\mathbf{x})$ represents a force field, e.g., gravity, electromagnetic force, etc., then its line integral $\displaystyle\int_C \mathbf{f} \cdot d\mathbf{x}$ represents the *work* done by moving along the curve. As in two dimensions, work is independent of the parametrization of the curve, i.e., the particle's speed of traversal.

Example B.6. Our goal is to move a mass through the force field $\mathbf{f} = (\,y, -x, 1\,)^T$ starting from the initial point $(\,1,0,1\,)^T$ and moving vertically to the final point $(\,1,0,2\,\pi\,)^T$. *Question*: does it require more work to move in a straight line $\mathbf{x}(t) = (\,1,0,t\,)^T$ or along the spiral helix $\mathbf{x}(t) = (\,\cos t, \sin t, t\,)^T$, where, in both cases, $0 \le t \le 2\,\pi$? The work line integral has the form

$$\int_C \mathbf{f} \cdot d\mathbf{x} = \int_C y\,dx - x\,dy + dz = \int_0^{2\,\pi} \left[y\,\frac{dx}{dt} - x\,\frac{dy}{dt} + \frac{dz}{dt} \right] dt.$$

Along the straight line, the amount of work is

$$\int_C \mathbf{f} \cdot d\mathbf{x} = \int_0^{2\,\pi} dt = 2\,\pi.$$

As for the spiral helix,

$$\int_C \mathbf{f} \cdot d\mathbf{x} = \int_0^{2\,\pi} \left[-\sin^2 t - \cos^2 t + 1 \right] dt = 0.$$

Thus, although we travel a more roundabout route, it takes no work to move along the helix!

The reason for the second result is that the force vector field $\mathbf{f}$ is everywhere orthogonal to the tangent to the curve: $\mathbf{f} \cdot \mathbf{t} = 0$, and so there is no tangential force exerted upon the motion. In such cases, the work line integral

$$\int_C \mathbf{f} \cdot d\mathbf{x} = \int_C \mathbf{f} \cdot \mathbf{t} \, ds = 0$$

automatically vanishes. In other words, it takes no work whatsoever to move in any direction which is orthogonal to the given force vector.

B.4. Surfaces.

Curves are one-dimensional, and so can be traced out by a single parameter. Surfaces are two-dimensional, and hence require two distinct parameters. Thus, a *surface* $S \subset \mathbb{R}^3$ is parametrized by a vector-valued function

$$\mathbf{x}(p,q) = \big(x(p,q), y(p,q), z(p,q) \big)^T \tag{B.24}$$

that depends on two variables. As the parameters $(p,q) \in \Omega$ range over a prescribed plane domain $\Omega \subset \mathbb{R}^2$, the locus of points $\mathbf{x}(p,q)$ traces out the surface in space. See Figure surf■ for an illustration. The parameters are often thought of as defining a system of *local coordinates* on the curved surface.

We shall always assume that the surface is *simple*, meaning that it does not intersect itself, so $\mathbf{x}(p,q) = \mathbf{x}(\widetilde{p}, \widetilde{q})$ if and only if $p = \widetilde{p}$ and $q = \widetilde{q}$. In practice, this condition can be quite hard to check! The boundary

$$\partial S = \{ \, \mathbf{x}(p,q) \mid \ (p,q) \in \partial\Omega \, \} \tag{B.25}$$

of a simple surface consists of one or more simple curves, as in Figure surf■. If the underlying parameter domain Ω is bounded and simply connected, then $\partial\Omega$ is a simple closed plane curve, and so ∂S is also a simple closed curve.

Example B.7. The simplest instance of a surface is a *graph* of a function. The parameters are the x, y coordinates, and the surface coincides with the portion of the graph of the function $z = u(x,y)$ that lies over a fixed domain $(x,y) \in \Omega \subset \mathbb{R}^2$, as illustrated in Figure gsurf■. Thus, a graphical surface has the parametric form

$$\mathbf{x}(p,q) = \big(p, q, u(p,q) \big)^T, \qquad\qquad (p,q) \in \Omega.$$

Thus, the parametrization identifies $x = p$ and $y = q$, while $z = u(p,q) = u(x,y)$ represents the height of the surface above the point $(x,y) \in \Omega$.

For example, the upper *hemisphere* S_r^+ of radius r centered at the origin can be parametrized as a graph

$$z = \sqrt{r^2 - x^2 - y^2}, \qquad\qquad x^2 + y^2 < r^2, \tag{B.26}$$

sitting over the disk $D_r = \{ \, x^2 + y^2 < r^2 \, \}$ of radius r. The boundary of the hemisphere is the image of the circle $C_r = \partial D_r = \{ \, x^2 + y^2 = r^2 \, \}$ of radius r, and is itself a circle of radius r sitting in the x, y plane: $\partial S_r^+ = \{ \, x^2 + y^2 = r, z = 0 \, \}$.

Remark: One can interpret the Dirichlet problem (15.1), (15.4) for the two-dimensional Laplace equation as the problem of finding a surface S that is the graph of a harmonic function with a prescribed boundary $\partial S = \{\, z = h(x,y) \text{ for } (x,y) \in \partial\Omega \,\}$.

Example B.8. A *sphere* S_r of radius r can be explicitly parametrized by two angular variables φ, θ in the form

$$\mathbf{x}(\varphi,\theta) = (r\sin\varphi\cos\theta, r\sin\varphi\sin\theta, r\cos\varphi), \qquad 0 \le \theta < 2\pi, \quad 0 \le \varphi \le \pi. \tag{B.27}$$

The reader can easily check that $\|\,\mathbf{x}\,\|^2 = r^2$, as it should be. As illustrated in Figure sangle■, θ measures the *meridial angle* or *longitude*, while φ measures the *azimuthal angle* or *latitude*. Thus, the upper hemisphere S_r^+ is obtained by restricting the azimuthal parameter to the range $0 \le \varphi \le \frac{1}{2}\pi$. Each parameter value φ, θ corresponds to a unique point on the sphere, *except* when $\varphi = 0$ or π. All points $(\theta, 0)$ are mapped to the north pole $(\,0,0,r\,)$, while all points (θ,π) are mapped to the south pole $(\,0,0,-r\,)$. Away from the poles, the spherical angles provide *bona fide* coordinates on the sphere. Fortunately, the polar singularities do not interfere with the overall smoothness of the sphere. Nevertheless, one must always be careful at or near these two distinguished points.

The curves $\{\,\varphi = c\,\}$ where the azimuthal angle takes a prescribed constant value are the circular parallels of constant latitude — except for the north and south poles which are merely points. The equator is at $\varphi = \frac{1}{2}\,\pi$, while the tropics of Cancer and Capricorn are $23\frac{1}{2}^\circ \approx 0.41$ radians above and below the equator. The curves $\{\,\theta = c\,\}$ where the meridial angle is constant are the semi-circular meridians of constant longitude stretching from north to south pole. Note that $\theta = 0$ and $\theta = 2\,\pi$ describe the same meridian. In terrestrial navigation, latitude is the angle, in degrees, measured from the equator, while longitude is the angle measured from the Greenwich meridian.

Example B.9. A *torus* is a surface of the form of an inner tube. One convenient parametrization of a particular toroidal surface is

$$\mathbf{x}(\psi,\theta) = (\,(2+\cos\psi)\cos\theta, (2+\cos\psi)\sin\theta, \sin\psi\,)^T \qquad \text{for} \qquad 0 \le \psi, \theta \le 2\,\pi. \tag{B.28}$$

Note that the parametrization is $2\,\pi$ periodic in both ψ and θ. If we introduce cylindrical coordinates

$$x = r\,\cos\theta, \qquad\qquad y = r\,\sin\theta, \qquad\qquad z,$$

then the torus is parametrized by

$$r = 2 + \cos\psi, \qquad\qquad z = \sin\psi.$$

Therefore, the relevant values of (r,z) all lie on the circle

$$(r-2)^2 + z^2 = 1 \tag{B.29}$$

of radius 1 centered at $(2,0)$. As the polar angle θ increases from 0 to $2\,\pi$, the circle rotates around the z axis, and thereby sweeps out the torus.

Remark: The sphere and the torus are examples of *closed surfaces*. The requirements for a surface to be closed are that it be simple and bounded, and, moreover, have no boundary. In general, a subset $S \subset \mathbb{R}^3$ is *bounded* provided it does not stretch off infinitely far away. More precisely, boundedness is equivalent to the existence of a fixed number $R > 0$ which bounds the norm $\| \mathbf{x} \| < R$ of all points $\mathbf{x} \in S$.

Tangents to Surfaces

Consider a surface S parameterized by $\mathbf{x}(p, q)$ where $(p, q) \in \Omega$. Each parametrized curve $(p(t), q(t))$ in the parameter domain Ω will be mapped to a parametrized curve $C \subset S$ contained in the surface. The curve C is parametrized by the composite map

$$\mathbf{x}(t) = \mathbf{x}(p(t), q(t)) = \big(x(p(t), q(t)), y(p(t), q(t)), z(p(t), q(t)) \big)^T .$$

The tangent vector

$$\frac{d\mathbf{x}}{dt} = \frac{\partial \mathbf{x}}{\partial p}\,\frac{dp}{dt} + \frac{\partial \mathbf{x}}{\partial q}\,\frac{dq}{dt} \tag{B.30}$$

to such a curve will be tangent to the surface. The set of all possible tangent vectors to curves passing through a given point in the surface traces out the *tangent plane* to the surface at that point, as in Figure tp■. Note that the tangent vector (B.30) is a linear combination of the two *basis tangent vectors*

$$\mathbf{x}_p = \frac{\partial \mathbf{x}}{\partial p} = \left(\frac{\partial x}{\partial p}, \frac{\partial y}{\partial p}, \frac{\partial z}{\partial p} \right)^T , \qquad \mathbf{x}_q = \frac{\partial \mathbf{x}}{\partial q} = \left(\frac{\partial x}{\partial q}, \frac{\partial y}{\partial q}, \frac{\partial z}{\partial q} \right)^T , \tag{B.31}$$

which therefore span the tangent plane to the surface at the point $\mathbf{x}(p, q) \in S$. The first basis vector is tangent to the curves where $q =$ constant, while the second is tangent to the curves where $p =$ constant.

Example B.10. Consider the torus T parametrized as in (B.28). The basis tangent vectors are

$$\frac{\partial \mathbf{x}}{\partial \psi} = \begin{pmatrix} -(2 + \cos\theta) \sin\psi \\ (2 + \cos\theta) \cos\psi \\ 0 \end{pmatrix} , \qquad \frac{\partial \mathbf{x}}{\partial \theta} = \begin{pmatrix} -\sin\theta\, \cos\psi \\ -\sin\theta\, \sin\psi \\ \cos\theta \end{pmatrix} . \tag{B.32}$$

They serve to span the tangent plane to the torus at the point $\mathbf{x}(\theta, \psi)$. For example, at the point $\mathbf{x}(0, 0) = (3, 0, 0)^T$ corresponding to the particular parameter values $\theta = \psi = 0$, the basis tangent vectors are

$$\mathbf{x}_\psi(0, 0) = (0, 3, 0)^T = 3\,\mathbf{e}_2 , \qquad \mathbf{x}_\theta(0, 0) = (0, 0, 1)^T = \mathbf{e}_3 ,$$

and so the tangent plane at this particular point is the (y, z)–plane spanned by the standard basis vectors $\mathbf{e}_2, \mathbf{e}_3$.

The tangent to any curve contained within the torus at the given point will be a linear combination of these two vectors. For instance, the toroidal knot (B.16) corresponds to the straight line

$$\psi(t) = 2t, \qquad 0 \le t \le 2\pi, \qquad \theta(t) = 3t,$$

in the parameter space. Its tangent vector

$$\frac{d\mathbf{x}}{dt} = \begin{pmatrix} -(4 + 2\cos 3t)\sin 2t - 3\,\sin 3t\,\cos 2t \\ (4 + 2\cos 3t)\cos 2t - 3\,\sin 3t\,\sin 2t \\ 3\,\cos 3t \end{pmatrix}$$

lies in the tangent plane to the torus at each point. In particular, at $t = 0$, the knot passes through the point $\mathbf{x}(0,0) = (\,3, 0, 0\,)^T$, and has tangent vector

$$\frac{d\mathbf{x}}{dt} = \begin{pmatrix} 0 \\ 6 \\ 3 \end{pmatrix} = 2\,\mathbf{x}_\psi(0,0) + 3\,\mathbf{x}_\theta(0,0) \qquad \text{since} \qquad \frac{d\psi}{dt} = 2,\ \frac{d\theta}{dt} = 3.$$

A point $\mathbf{x}(p,q) \in S$ on the surface is said to be *nonsingular* provided the basis tangent vectors $\mathbf{x}_p(p,q), \mathbf{x}_q(p,q)$ are linearly independent. Thus the point is nonsingular if and only if the tangent vectors span a full two-dimensional subspace of $\mathbb{R}^3$ — the tangent plane to the surface at the point. Nonsingularity ensures the smoothness of the surface at each point, which is a consequence of the general Implicit Function Theorem, [**126**]. Singular points, where the tangent vectors are linearly dependent, can take the form of corners, cusps and folds in the surface. From now on, we shall always assume that our surface is *nonsingular* meaning every point is a nonsingular point.

Linear independence of the tangent vectors is equivalent to the requirement that their cross product is a nonzero vector:

$$\mathbf{N} = \frac{\partial \mathbf{x}}{\partial p} \wedge \frac{\partial \mathbf{x}}{\partial q} = \left(\frac{\partial(y,z)}{\partial(p,q)}, \frac{\partial(z,x)}{\partial(p,q)}, \frac{\partial(x,y)}{\partial(p,q)} \right)^T \neq \mathbf{0}. \tag{B.33}$$

In this formula, we have adopted the standard notation

$$\frac{\partial(x,y)}{\partial(p,q)} = \det \begin{pmatrix} x_p & x_q \\ y_p & y_q \end{pmatrix} = \frac{\partial x}{\partial p}\,\frac{\partial y}{\partial q} - \frac{\partial x}{\partial q}\,\frac{\partial y}{\partial p} \tag{B.34}$$

for the *Jacobian determinant* of the functions x, y with respect to the variables p, q, which we already encountered in the change of variables formula (A.52) for double integrals. The cross-product vector $\mathbf{N}$ in (B.33) is orthogonal to both tangent vectors, and hence orthogonal to the entire tangent plane. Therefore, $\mathbf{N}$ defines a *normal* vector to the surface at the given (nonsingular) point.

Example B.11. Consider a surface S parametrized as the graph of a function $z = u(x,y)$, and so, as in Example B.7

$$\mathbf{x}(x,y) = (\,x, y, u(x,y)\,)^T, \qquad (x,y) \in \Omega.$$

The tangent vectors

$$\frac{\partial \mathbf{x}}{\partial x} = \left(1, 0, \frac{\partial u}{\partial x}\right)^T, \qquad \frac{\partial \mathbf{x}}{\partial y} = \left(0, 1, \frac{\partial u}{\partial y}\right)^T,$$

span the tangent plane sitting at the point $(x, y, u(x, y)$ on S. The normal vector is

$$\mathbf{N} = \frac{\partial \mathbf{x}}{\partial x} \wedge \frac{\partial \mathbf{x}}{\partial y} = \left(-\frac{\partial u}{\partial x}, -\frac{\partial u}{\partial y}, 1 \right)^T,$$

and points upwards, as in Figure graphN■. Note that every point on the graph is nonsingular.

The *unit normal* to the surface at the point is a unit vector orthogonal to the tangent plane, and hence given by

$$\mathbf{n} = \frac{\mathbf{N}}{\| \mathbf{N} \|} = \frac{\mathbf{x}_p \wedge \mathbf{x}_q}{\| \mathbf{x}_p \wedge \mathbf{x}_q \|}. \tag{B.35}$$

In general, the direction of the normal vector $\mathbf{N}$ depends upon the order of the two parameters p, q. Computing the cross product in the reverse order, $\mathbf{x}_q \wedge \mathbf{x}_p = -\mathbf{N}$, reverses the sign of the normal vector, and hence switches its direction. Thus, there are two possible unit normals to the surface at each point, namely $\mathbf{n}$ and $-\mathbf{n}$. For a closed surface, one normal points outwards and one points inwards.

When possible, a consistent (meaning continuously varying) choice of a unit normal serves to define an *orientation* of the surface. All closed surfaces, and most other surfaces can be oriented. The usual convention for closed surfaces is to choose the orientation defined by the outward normal. The simplest example of a non-orientable surface is the Möbius strip obtained by gluing together the ends of a twisted strip of paper; see Exercise ■.

Example B.12. For the sphere of radius r parametrized by the spherical angles as in (B.27), the tangent vectors are

$$\frac{\partial \mathbf{x}}{\partial \varphi} = \begin{pmatrix} r \cos\varphi \cos\theta \\ r \sin\varphi \cos\theta \\ -r \sin\varphi \end{pmatrix}, \qquad \frac{\partial \mathbf{x}}{\partial \theta} = \begin{pmatrix} -r \sin\varphi \sin\theta \\ r \sin\varphi \cos\theta \\ 0 \end{pmatrix}.$$

These vectors are tangent to, respectively, the meridians of constant longitude, and the parallels of constant latitude. The normal vector is

$$\mathbf{N} = \frac{\partial \mathbf{x}}{\partial \varphi} \wedge \frac{\partial \mathbf{x}}{\partial \theta} = \begin{pmatrix} r^2 \sin^2\varphi \cos\theta \\ r^2 \sin^2\varphi \sin\theta \\ r^2 \cos\varphi \sin\varphi \end{pmatrix} = r \sin\varphi \, \mathbf{x}. \tag{B.36}$$

Thus $\mathbf{N}$ is a non-zero multiple of the radial vector $\mathbf{x}$, except at the north or south poles when $\varphi = 0$ or π. This reconfirms our earlier observation that the poles are problematic points for the spherical angle parametrization. The unit normal

$$\mathbf{n} = \frac{\mathbf{N}}{\| \mathbf{N} \|} = \frac{\mathbf{x}}{r}$$

determined by the spherical coordinates φ, θ is the outward pointing normal. Reversing the order of the angles, θ, φ, would lead to the outwards normal $-\mathbf{n} = -\mathbf{x}/r$.

Remark: As we already saw in the example of the hemisphere, a given surface can be parametrized in many different ways. In general, to change parameters

$$p = g(\widetilde{p}, \widetilde{q}), \qquad\qquad q = h(\widetilde{p}, \widetilde{q}),$$

requires a smooth, invertible map between the two parameter domains $\widetilde{\Omega} \to \Omega$. Many interesting surfaces, particularly closed surfaces, cannot be parametrized in a single consistent manner that satisfies the smoothness constraint (B.33) on the entire surface. In such cases, one must assemble the surface out of pieces, each parametrized in the proper manner. The key problem in cartography is to find convenient parametrizations of the globe that do not significantly distort the geographical features of the planet.

A surface is *piecewise smooth* if it can be constructed by gluing together a finite number of smooth parts, joined along piecewise smooth curves. For example, a cube is a piecewise smooth surface, consisting of squares joined along straight line segments. We shall rely on the reader's intuition to formalize these ideas, leaving a rigorous development to a more comprehensive treatment of surface geometry, e.g., [**49**].

B.5. Surface Integrals.

As with spatial line integrals, there are two important types of surface integral. The first is the integration of a scalar field with respect to surface area. A typical application is to compute the area of a curved surface or the mass and center of mass of a curved shell of possibly variable density. The second type is the surface integral that computes the flux associated with a vector field through an oriented surface. Applications appear in fluid mechanics, electromagnetism, thermodynamics, gravitation, and many other fields.

Surface Area

According to (B.10), the length of the cross product of two vectors measures the area of the parallelogram they span. This observation underlies the proof that the length of the normal vector to a surface (B.35), namely

$$\| \mathbf{N} \| = \| \mathbf{x}_p \wedge \mathbf{x}_q \|,$$

is a measure of the infinitesimal element of surface area, denoted

$$dS = \| \mathbf{N} \| \, dp \, dq = \| \mathbf{x}_p \wedge \mathbf{x}_q \| \, dp \, dq. \tag{B.37}$$

The total area of the surface is found by summing up these infinitesimal contributions, and is therefore given by the double integral

$$\begin{aligned} \text{area } S = \iint_S dS &= \iint_\Omega \| \mathbf{x}_p \wedge \mathbf{x}_q \| \, dp \, dq \\ &= \iint_\Omega \sqrt{\left(\frac{\partial(y,z)}{\partial(p,q)}\right)^2 + \left(\frac{\partial(z,x)}{\partial(p,q)}\right)^2 + \left(\frac{\partial(x,y)}{\partial(p,q)}\right)^2} \; dp \, dq. \end{aligned} \tag{B.38}$$

The surface's area does not depend upon the parametrization used to compute the integral. In particular, if the surface is parametrized by x, y as the graph $z = u(x, y)$ of a function over a domain $(x, y) \in \Omega$, then the surface area integral reduces to the familiar form

$$\text{area } S = \iint_S \, dS = \iint_\Omega \sqrt{1 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)} \, dx\, dy. \tag{B.39}$$

A detailed justification of these formulae can be found in [**9**, **58**].

Example B.13. The well-known formula for the surface area of a sphere is a simple consequence of the integral formula (B.38). Using the parametrization by spherical angles (B.27) and the formula (B.36) for the normal, we find

$$\text{area } S_r = \iint_{S_r} \, dS = \int_0^{2\,\pi} \int_0^{\pi} r^2 \, \sin\varphi \, d\varphi \, d\theta = 4\,\pi\, r^2. \tag{B.40}$$

Fortunately, the problematic poles do not cause any difficulty in the computation, since they contribute nothing to the surface area integral.

Alternatively, we can compute the area of one hemisphere S_r^+ by realizing it as a graph

$$z = \sqrt{r^2 - x^2 - y^2} \qquad \text{for} \qquad x^2 + y^2 \le 1,$$

over the disk of radius r, and so, by (B.39),

$$\begin{aligned} \text{area } S_r^+ &= \iint_\Omega \sqrt{1 + \frac{x^2}{r^2 - x^2 - y^2} + \frac{y^2}{r^2 - x^2 - y^2}} \, dx\, dy \\ &= \iint_\Omega \frac{r}{\sqrt{r^2 - x^2 - y^2}} \, dx\, dy = \int_0^r \int_0^{2\,\pi} \frac{r\,\rho}{\sqrt{r^2 - \rho^2}} \, d\theta\, d\rho = 2\,\pi\, r^2, \end{aligned}$$

where we used polar coordinates $x = \rho\cos\theta, y = \rho\sin\theta$ to evaluate the final integral. The area of the entire sphere is twice the area of the hemisphere.

Example B.14. Similarly, to compute the surface area of the torus T parametrized in (B.28), we use the tangent vectors in (B.32) to compute the normal to the torus:

$$\mathbf{N} = \mathbf{x}_\psi \wedge \mathbf{x}_\theta = \begin{pmatrix} (2 + \cos\psi)\cos\psi\cos\theta \\ (2 + \cos\psi)\cos\psi\sin\theta \\ (2 + \cos\psi)\sin\psi \end{pmatrix}, \qquad \text{with} \qquad \| \, \mathbf{x}_\psi \wedge \mathbf{x}_\theta \, \| = 2 + \cos\psi.$$

Therefore,

$$\text{area } T = \int_0^{2\,\pi} \int_0^{2\,\pi} (2 + \cos\psi) \, d\psi \, d\theta = 8\,\pi^2.$$

If $S \subset \mathbb{R}^3$ is a surface with finite area, the *mean* or *average* of a scalar function $f(x, y, z)$ over S is given by

$$\mathrm{M}_S[\, f \,] = \frac{1}{\text{area } S} \iint_S f \, dS. \tag{B.41}$$

For example, the mean of a function over a sphere $S_r = \{ \| \mathbf{x} \| = r \}$ of radius r is explicitly given by

$$\mathrm{M}_{S_r}[\, f \,] = \frac{1}{4\,\pi\, r^2} \iint_{\| \mathbf{x} \| = r} f(\mathbf{x})\, dS = \frac{1}{4\,\pi} \int_0^{2\,\pi} \int_0^{\pi} F(r, \varphi, \theta) \sin \varphi \, d\varphi \, d\theta, \tag{B.42}$$

where $F(r, \varphi, \theta)$ is the spherical coordinate expression for the scalar function f. As usual, the mean lies between the maximum and minimum values of the function on the surface:

$$\min_S f \le \mathrm{M}_S[\, f \,] \le \max_S f.$$

In particular, the *center of mass* $\mathbf{C}$ of a surface (assuming it has constant density) is equal to the mean of the coordinate functions $\mathbf{x} = (\, x, y, z \,)^T$, so

$$\mathbf{C} = (\, \mathrm{M}_S[\, x \,], \mathrm{M}_S[\, y \,], \mathrm{M}_S[\, z \,] \,)^T = \frac{1}{\text{area } S} \left(\iint_S x \, dS, \iint_S y \, dS, \iint_S z \, dS \right)^T. \tag{B.43}$$

Thus, the center of mass of a hemisphere is ∎

More generally, the integral of a scalar field $u(x, y, z)$ over the surface is given by

$$\iint_S u \, dS = \iint_\Omega u(x(p,q), y(p,q), z(p,q)) \, \| \mathbf{x}_p \wedge \mathbf{x}_q \| \, dp \, dq. \tag{B.44}$$

If S represents a thin curved shell, and $u = \rho(\mathbf{x})$ the density of the material at position $\mathbf{x} \in S$, then the surface integral (B.44) represents the total mass of the shell. For example, the integral of $u(x, y, z)$ over a hemisphere S_r^+ of radius r can be evaluated by either of the formulae

$$\begin{aligned} \iint_{S_r^+} u \, dS &= \int_0^{2\,\pi} \int_0^{\pi/2} u(r \cos\theta \sin\varphi, r \sin\theta \sin\varphi, r \cos\varphi) \; r^2 \sin\varphi \; d\varphi \, d\theta \\ &= \iint_{x^2 + y^2 \le r^2} \frac{r}{\sqrt{r^2 - x^2 - y^2}} \; u(x, y, \sqrt{r^2 - x^2 - y^2}\,) \; dx \, dy, \end{aligned} \tag{B.45}$$

depending upon whether one prefers spherical or graphical coordinates.

Flux Integrals

Now assume that S is an oriented surface with chosen unit normal $\mathbf{n}$. If $\mathbf{v} = (\, u, v, w \,)^T$ is a vector field, then the *surface integral*

$$\iint_S \mathbf{v} \cdot \mathbf{n} \; dS = \iint_\Omega \big(\mathbf{v} \cdot \mathbf{x}_p \wedge \mathbf{x}_q \big) \; dp \, dq = \iint_\Omega \det \begin{pmatrix} u & x_p & x_q \\ v & y_p & y_q \\ w & z_p & z_q \end{pmatrix} dp \, dq \tag{B.46}$$

of the normal component of $\mathbf{v}$ over the entire surface measures its *flux* through the surface. An alternative common notation for the flux integral is

$$\iint_S \mathbf{v} \cdot \mathbf{n} \; dS = \iint_S u \, dy \, dz + v \, dz \, dx + w \, dx \, dy \tag{B.47}$$

$$= \iint_\Omega \left(u(x,y,z)\, \frac{\partial(y,z)}{\partial(p,q)} + v(x,y,z)\, \frac{\partial(z,x)}{\partial(p,q)} + w(x,y,z)\, \frac{\partial(x,y)}{\partial(p,q)} \right) dx\, dy,$$

Note how the Jacobian determinant notation (B.34) seamlessly interacts with the integration. In particular, if the surface is the graph of a function $z = h(x,y)$, then the surface integral reduces to the particularly simple form

$$\iint_S \mathbf{v} \cdot \mathbf{n}\, dS = \iint_\Omega \left(u(x,y,z)\, \frac{\partial z}{\partial x} + v(x,y,z)\, \frac{\partial z}{\partial y} + w(x,y,z) \right) dp\, dq \qquad (B.48)$$

The flux surface integral relies upon the consistent choice of an orientation or unit normal on the surface. Thus, flux only makes sense through an oriented surface — it doesn't make sense to speak of "flux through a Möbius band". If we switch normals, using, say, the inward instead of the outward normal, then the surface integral changes sign — just like a line integral if we reverse the orientation of a curve. Similarly, if we decompose a surface into the union of two or more parts, with only their boundaries in common, then the surface integral similarly decomposes into a sum of surface integrals. Thus,

$$\begin{aligned} \iint_{-S} \mathbf{v} \cdot \mathbf{n}\, dS &= -\iint_S \mathbf{v} \cdot \mathbf{n}\, dS, \\ \iint_S \mathbf{v} \cdot \mathbf{n}\, dS &= \iint_{S_1} \mathbf{v} \cdot \mathbf{n}\, dS \;+\; \iint_{S_2} \mathbf{v} \cdot \mathbf{n}\, dS, \qquad S = S_1 \cup S_2. \end{aligned} \qquad (B.49)$$

In the first formula, $-S$ denotes the surface S with the reverse orientation. In the second formula, S_1 and S_2 are only allowed to intersect along their boundaries; moreover, they must be oriented in the same manner as S, i.e., have the same unit normal direction.

Example B.15. Let S denote the triangular surface given by that portion of the plane $x + y + z = 1$ that lies inside the positive orthant $\{\, x \geq 0,\ y \geq 0,\ z \geq 0 \,\}$, as in Figure tri3. The flux of the vector field $\mathbf{v} = (\, y, x\, z, 0\,)^T$ through S equals the surface integral

$$\iint_S y\, dy\, dz + x\, z\, dz\, dx,$$

where we orient S by choosing the upwards pointing normal. To compute, we note that S can be identified as the graph of the function $z = 1 - x - y$ lying over the triangle $T = \{\, 0 \leq x \leq 1, 0 \leq y \leq 1 - x \,\}$. Therefore, by (B.47),

$$\begin{aligned} \iint_S y\, dy\, dz + x\, z\, dz\, dx &= \iint_T \left[y\, \frac{\partial(y, 1-x-y)}{\partial(x,y)} + x\,(1-x-y)\, \frac{\partial(1-x-y, x)}{\partial(x,y)} \right] dx\, dy \\ &= \int_0^1 \int_0^{1-x} (1-x)\,(y+x)\, dy\, dx = \int_0^1 \left(\tfrac{1}{2} + \tfrac{1}{2}\, x - \tfrac{1}{2}\, x^2 + \tfrac{1}{2}\, x^3 \right) dx = \tfrac{17}{24}. \end{aligned}$$

If $\mathbf{v}$ represents the velocity vector field for a steady state fluid flow, then its flux integral (B.46) tells us the total volume of fluid passing through S per unit time. Indeed, at each point on S, the volume fluid that flows across a small part the surface in unit time

will fill a thin cylinder whose base is the surface area element dS and whose height $\mathbf{v} \cdot \mathbf{n}$ is the normal component of the fluid velocity $\mathbf{v}$, as pictured in Figure fluxs. Summing (integrating) all these flux cylinder volumes over the surface results in the flux integral. The choice of orientation or unit normal specifies the convention for measuring the direction of positive flux through the surface. If S is a closed surface, and we choose $\mathbf{n}$ to be the unit outward normal, then the flux integral (B.46) represents the net amount of fluid flowing *out* of the solid region bounded by S per unit time.

Example B.16. The vector field $\mathbf{v} = (\,0, 0, 1\,)^T$ represents a fluid moving with constant velocity in the vertical direction. Let us compute the fluid flux through a hemisphere

$$S_r^+ = \left\{ \, z = \sqrt{r^2 - x^2 - y^2} \;\middle|\; x^2 + y^2 \leq 1 \, \right\},$$

sitting over the disk D_r of radius r in the x, y plane. The flux integral over S_r^+ is computed using (B.48), so

$$\iint_{S_r^+} \mathbf{v} \cdot \mathbf{n} \; dS = \iint_{S_r^+} dx \wedge dy = \iint_{D_r} dx \, dy = \pi \, r^2.$$

The resulting double integral is just the area of the disk. Indeed, in this case, the value of the flux integral is the *same* for all surfaces $z = h(x, y)$ sitting over the disk D_r.

This example provides a particular case of a surface-independent flux integral, which are defined in analogy with the path-independent line integrals that we encountered earlier. In general, a flux integral is called *surface-independent* if

$$\iint_{S_1} \mathbf{v} \cdot \mathbf{n} \; dS = \iint_{S_2} \mathbf{v} \cdot \mathbf{n} \; dS \tag{B.50}$$

whenever the surfaces S_1 and S_2 have a common boundary $\partial S_1 = \partial S_2$. In other words, the value of the integral depends only upon the boundary of the surface. The veracity of (B.50) requires that the surfaces be oriented in the "same manner". For instance, if they do not cross, then the combined surface $S = S_1 \cup S_2$ is closed, and one uses the outward pointing normal on one surface and the inward pointing normal on the other. In more complex situations, one checks that the two surfaces induce the same orientation on their common boundary. (We defer a discussion of the boundary orientation until later.) Finally, applying (B.49) to the closed surface $S = S_1 \cup S_2$ and using the prescribed orientations, we deduce an alternative characterization of surface-independent vector fields.

Proposition B.17. *A vector field leads to a surface-independent flux integral if and only if*

$$\iint_S \mathbf{v} \cdot \mathbf{n} \; dS = 0 \tag{B.51}$$

for every closed surface S contained in the domain of definition of $\mathbf{v}$.

A fluid is *incompressible* when its volume is unaltered by the flow. Therefore, in the absence of sources or sinks, there cannot be any net inflow or outflow across a simple closed surface bounding a region occupied by the fluid. Thus, the flux integral over a closed surface

must vanish: $\iint_S \mathbf{v} \cdot \mathbf{n}\, dS = 0$. Proposition B.17 implies that the fluid velocity vector field defines a surface-independent flux integral. Thus, the flux of an incompressible fluid flow through any surface depends only on the (oriented) boundary curve of the surface!

B.6. Volume Integrals.

Volume or triple integrals take place over domains $\Omega \subset \mathbb{R}^3$ representing solid three-dimensional bodies. A simple example of such a domain is a *ball*

$$\mathrm{B}_r(\mathbf{a}) = \{\, \mathbf{x} \mid \ \| \mathbf{x} - \mathbf{a} \| < r \,\} \tag{B.52}$$

of radius $r > 0$ centered at a point $\mathbf{a} \in \mathbb{R}^3$. Other examples of domains include solid cubes, solid cylinders, solid tetrahedra, solid tori (doughnuts and bagels), solid cones, etc.

In general, a subset $\Omega \subset \mathbb{R}^3$ is *open* if, for every point $\mathbf{x} \in \Omega$, a small open ball $\mathrm{B}_\varepsilon(\mathbf{a}) \subset \Omega$ centered at $\mathbf{a}$ of radius $\varepsilon = \varepsilon(\mathbf{a}) > 0$, which may depend upon $\mathbf{a}$, is also contained in Ω. In particular, the ball (B.52) is open. The *boundary* $\partial\Omega$ of an open subset Ω consists of all limit points which are not in the subset. Thus, the boundary of the open ball $\mathrm{B}_r(\mathbf{a})$ is the sphere $\mathrm{S}_r(\mathbf{a}) = \{\, \| \mathbf{x} - \mathbf{a} \| = r \,\}$ of radius r centered at the point a. An open subset is called a *domain* if its boundary $\partial\Omega$ consists of one or more simple, piecewise smooth surfaces. We are allowing corners and edges in the bounding surfaces, so that an open cube will be a perfectly valid domain.

A subset $\Omega \subset \mathbb{R}^3$ is *bounded* provided it fits inside a sphere of some (possibly large) radius. For example, the solid ball $\mathrm{B}_r = \{\, \| \mathbf{x} \| < R \,\}$ is bounded, while its exterior $\mathrm{E}_r = \{\, \| \mathbf{x} \| > R \,\}$ is an unbounded domain. The sphere $S_R = \{\, \| \mathbf{x} \| = R \,\}$ is the common boundary of the two domains: $S_R = \partial B_r = \partial E_R$. Indeed, any simple closed surface separates $\mathbb{R}^3$ into two domains that have a common boundary — its *interior*, which is bounded, and its unbounded *exterior*.

The boundary of a bounded domain consists of one or more closed surfaces. For instance, the solid annular domain

$$A_{r,R} = \big\{\, 0 < r < \| \mathbf{x} \| < R \,\big\} \tag{B.53}$$

consisting of all points lying between two concentric spheres of respective radii r and R has boundary given by the two spheres: $\partial A_{r,R} = S_r \cup S_R$. On the other hand, setting $r = 0$ in (B.53) leads to a *punctured ball* of radius R whose center point has been removed. A punctured ball is *not* a domain, since the center point is part of the boundary, but is not a *bona fide* surface.

If the domain $\Omega \subset \mathbb{R}^3$ represents a solid body, and the scalar field $\rho(x, y, z)$ represents its density at a point $(x, y, z) \in \Omega$, then the triple integral

$$\iiint_\Omega \rho(x, y, z)\, dx\, dy\, dz \tag{B.54}$$

equals the total mass of the body. In particular, the volume of Ω is equal to

$$\operatorname{vol} \Omega = \iiint_\Omega dx\, dy\, dz. \tag{B.55}$$

Triple integrals can be directly evaluated when the domain has the particular form

$$\Omega = \big\{\, \xi(x,y) < z < \eta(x,y), \quad \varphi(x) < y < \varphi(x), \quad a < x < b \,\big\} \tag{B.56}$$

where the z coordinate lies between two graphical surfaces sitting over a common domain in the (x,y)–plane that is itself of the form of (A.47) used to evaluate double integrals; see Figure triple■. In such cases we can evaluate the triple integral by iterated integration first with respect to z, then with respect to y and, finally, with respect to x:

$$\iiint_\Omega u(x,y,z)\,dx\,dy\,dz = \int_a^b \left(\int_{\varphi(x)}^{\psi(x)} \left(\int_{\xi(x,y)}^{\eta(x,y)} u(x,y,z)\,dz \right) dy \right) dx. \tag{B.57}$$

A similar result holds for other orderings of the coordinates.

Fubini's Theorem, [**126**, **125**], assures us that the result of iterated integration does not depend upon the order in which the variables are integrated. Of course, the domain must be of the requisite type in order to write the volume integral as repeated single integrals. More general triple integrals can be evaluated by chopping the domain up into disjoint pieces that have the proper form.

Example B.18. The volume of a solid ball B_R of radius R can be computed as follows. We express the domain of integration $x^2 + y^2 + z^2 < R^2$ in the form

$$-R < x < R, \;\; -\sqrt{R^2 - x^2} < y < \sqrt{R^2 - x^2}\,, \;\; -\sqrt{R^2 - x^2 - y^2} < z < \sqrt{R^2 - x^2 - y^2}\,.$$

Therefore, in accordance with (B.57),

$$\begin{aligned}
\iiint_{B_R} dx\,dy\,dz &= \int_{-R}^{R} \left(\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \left(\int_{-\sqrt{R^2-x^2-y^2}}^{\sqrt{R^2-x^2-y^2}} dz \right) dy \right) dx \\
&= \int_{-R}^{R} \left(\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} 2\sqrt{R^2 - x^2 - y^2}\,dy \right) dx \\
&= \int_{-R}^{R} \left(y\sqrt{R^2 - x^2 - y^2} + (R^2 - x^2)\sin^{-1}\frac{y}{\sqrt{R^2 - x^2}} \right) \Bigg|_{y=-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} dx \\
&= \int_{-R}^{R} \pi(R^2 - x^2)\,dx = \pi\left(R^2 x - \frac{x^2}{3} \right) \Bigg|_{x=-R}^{R} = \frac{4}{3}\,\pi\,R^3,
\end{aligned}$$

recovering the standard formula, as it should.

Change of Variables

Sometimes, an inspired change of variables can be used to simplify a volume integral. If

$$x = f(p,q,r), \qquad y = g(p,q,r), \qquad z = h(p,q,r), \tag{B.58}$$

is an invertible change of variables — meaning that each point (x,y,z) corresponds to a unique point (p,q,r) — then

$$\iiint_\Omega u(x,y,z)\,dx\,dy\,dz = \iiint_D U(p,q,r) \left| \frac{\partial(x,y,z)}{\partial(p,q,r)} \right| dp\,dq\,dr. \tag{B.59}$$

Here

$$U(p,q,r) = u(x(p,q,r), y(p,q,r), z(p,q,r))$$

is the expression for the integrand in the new coordinates, while D is the domain consisting of all points (p,q,r) that map to points $(x,y,z) \in \Omega$ in the original domain. Invertibility requires that each point in D corresponds to a unique point in Ω. The change in volume is governed by the absolute value of the three-dimensional *Jacobian determinant*

$$\frac{\partial(x,y,z)}{\partial(p,q,r)} = \det \begin{pmatrix} x_p & x_q & x_r \\ y_p & y_q & y_r \\ z_p & z_q & z_r \end{pmatrix} = \mathbf{x}_p \cdot \mathbf{x}_q \wedge \mathbf{x}_r \tag{B.60}$$

for the change of variables. The identification of the vector triple product (B.60) with an (infinitesimal) volume element lies behind the justification of the change of variables formula; see [**9**, **58**] for a detailed proof.

By far, the two most important cases are cylindrical and spherical coordinates. *Cylindrical coordinates* correspond to replacing the x and y coordinates by their polar counterparts, while retaining the vertical z coordinate unchanged. Thus, the change of coordinates has the form

$$x = r\,\cos\theta, \qquad y = r\,\sin\theta, \qquad z = z. \tag{B.61}$$

The Jacobian determinant for cylindrical coordinates is

$$\frac{\partial(x,y,z)}{\partial(r,\theta,z)} = \det \begin{pmatrix} x_r & x_\theta & x_z \\ y_r & y_\theta & y_z \\ z_r & z_\theta & z_z \end{pmatrix} = \det \begin{pmatrix} \cos\theta & -r\,\sin\theta & 0 \\ \sin\theta & r\,\cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} = r. \tag{B.62}$$

Therefore, the general change of variables formula (B.59) tells us the formula for a triple integral in cylindrical coordinates:

$$\iiint f(x,y,z)\,dx\,dy\,dz = \iiint f(r\,\cos\theta, r\,\sin\theta, z)\,r\,dr\,d\theta\,d\varphi. \tag{B.63}$$

Example B.19. For example, consider an ice cream cone

$$C_h = \big\{\, x^2 + y^2 < z^2, \;\; 0 < z < h \,\big\} = \big\{\, r < z, \;\; 0 < z < h \,\big\}$$

of height h plotted in Figure cone. To compute its volume, we express the domain in terms of the cylindrical coordinates, leading to

$$\iiint_{C_h} dx\,dy\,dz = \int_0^h \int_0^{2\,\pi} \int_0^z r\,dr\,d\theta\,dz = \int_0^h \pi\,z^2\,dz = \tfrac{1}{3}\,\pi\,h^3.$$

Spherical coordinates are denoted by r, φ, θ, where

$$x = r\,\sin\varphi\,\cos\theta, \qquad y = r\,\sin\varphi\,\sin\theta, \qquad z = r\,\cos\varphi. \tag{B.64}$$

Here $r = \| \mathbf{x} \| = \sqrt{x^2 + y^2 + z^2}$ represents the radius, $0 \le \varphi \le \pi$ is the azimuthal angle or latitude, while $0 \le \theta < 2\,\pi$ is the meridial angle or longitude. The reader may recall that

we already encountered these coordinates in our parametrization (B.27) of the sphere. It is important to distinguish between the spherical r, θ and the cylindrical r, θ — even though the same symbols are used, they represent *different* quantities.

A short computation proves that the spherical coordinate Jacobian determinant is

$$\begin{aligned} \frac{\partial(x,y,z)}{\partial(r,\varphi,\theta)} &= \det \begin{pmatrix} x_r & x_\varphi & x_\theta \\ y_r & y_\varphi & y_\theta \\ z_r & z_\varphi & z_\theta \end{pmatrix} \\ &= \det \begin{pmatrix} \sin\varphi\, \cos\theta & r\, \cos\varphi\, \cos\theta & -r\, \sin\varphi\, \sin\theta \\ \sin\varphi\, \sin\theta & r\, \cos\varphi\, \sin\theta & r\, \sin\varphi\, \cos\theta \\ \cos\varphi & -r\, \sin\varphi & 0 \end{pmatrix} = r^2 \sin\varphi. \end{aligned} \tag{B.65}$$

Therefore, a triple integral is evaluated in spherical coordinates according to the formula

$$\iiint f(x,y,z)\, dx\, dy\, dz = \iiint F(r,\varphi,\theta)\, r^2 \sin\varphi\; dr\, d\varphi\, d\theta, \tag{B.66}$$

where we rewrite the integrand

$$F(r,\varphi,\theta) = f(r\, \sin\varphi\, \cos\theta, r\, \sin\varphi\, \sin\theta, r\, \cos\varphi) \tag{B.67}$$

as a function of the spherical coordinates.

Example B.20. The integration required in Example B.18 to compute the volume of a ball B_R of radius R can be considerably simplified by switching over to spherical coordinates. The ball is given by $B_R = \big\{\, 0 \le r < R, 0 \le \varphi \le \pi, 0 \le \theta < 2\,\pi \,\big\}$. Thus, using (B.66), we compute

$$\iiint_{B_R} dx\, dy\, dz = \int_0^R \int_0^\pi \int_0^{2\,\pi} r^2\, \sin\varphi\; d\theta\, d\varphi\, dr = \int_0^R 4\,\pi\, r^2\, dr = \tfrac{4}{3}\,\pi\, R^3. \tag{B.68}$$

The reader may note that the next-to-last integrand represents the surface area of the sphere of radius R. Thus, we are, in effect, computing the volume by summing up (i.e., integrating) the surface areas of concentric thin spherical shells.

Remark: Sometimes, we will be sloppy and use the same letter for a function in an alternative coordinate system. Thus, we may use $f(r,\varphi,\theta)$ to represent the spherical coordinate form (B.67) of a function $f(x,y,z)$. Technically, this is not correct! However, the clarity and intuition sometimes outweighs the pedantic use of a new letter each time we change coordinates. Moreover, in geometry and modern physical theories, [**dg**], the symbol "f" represents an intrinsic scalar field, and $f(x,y,z)$ and $f(r,\varphi,\theta)$ merely its incarnations in two different coordinate charts on $\mathbb{R}^3$. Hopefully, this will be clear from the context.

B.7. Gradient, Divergence, and Curl.

There are three important vector differential operators that play a ubiquitous role in three-dimensional vector calculus, known as the gradient, divergence and curl. We have already encountered their two-dimensional counterparts in Chapter A.

The Gradient

We begin with the three-dimensional version of the *gradient* operator

$$\nabla u = \begin{pmatrix} u_x \\ u_y \\ u_z \end{pmatrix}. \tag{B.69}$$

The gradient defines a linear operator that maps a scalar function $u(x, y, z)$ to the vector field whose components are its partial derivatives with respect to the Cartesian coordinates.

If $\mathbf{x}(t) = (\, x(t), y(t), z(t)\,)^T$ is any parametrized curve, then the rate of change in the function u as we move along the curve is given by the inner product

$$\frac{d}{dt}\, u(x(t), y(t), z(t)) = \frac{\partial u}{\partial x}\,\frac{dx}{dt} + \frac{\partial u}{\partial y}\,\frac{dy}{dt} + \frac{\partial u}{\partial z}\,\frac{dz}{dt} = \nabla u \cdot \dot{\mathbf{x}} \tag{B.70}$$

between the gradient and the tangent vector to the curve. Therefore, as we reasoned earlier in the planar case, the gradient ∇u points in the direction of steepest increase in the function u, while its negative $-\nabla u$ points in the direction of steepest decrease. For example, if $u(x, y, z)$ represents the temperature at a point (x, y, z) in space, then ∇u points in the direction in which temperature is getting the hottest, while $-\nabla u$ points in the direction it gets the coldest. Therefore, if one wants to cool down as rapidly as possible, one should move in the direction of $-\nabla u$ at each instant, which is the direction of the flow of heat energy. Thus, the path $\mathbf{x}(t)$ to be followed for the fastest cool down will be a solution to the gradient flow equations

$$\dot{\mathbf{x}} = -\nabla u, \tag{B.71}$$

or, explicitly,

$$\frac{dx}{dt} = -\frac{\partial u}{\partial x}\,(x, y, z), \qquad \frac{dy}{dt} = -\frac{\partial u}{\partial y}\,(x, y, z), \qquad \frac{dz}{dt} = -\frac{\partial u}{\partial z}\,(x, y, z).$$

A solution $\mathbf{x}(t)$ to such a system of ordinary differential equations will experience continuously decreasing temperature. In Chapter 19, we will learn how to use such gradient flows to locate and numerically approximate the minima of functions.

The set of all points where a scalar field $u(x, y, z)$ has a given value,

$$u(x, y, z) = c \tag{B.72}$$

for some fixed constant c, is known as a *level set* of u. If u measures temperature, then its level sets are the isothermal surfaces of equal temperature. If u is sufficiently smooth, most of its level sets are smooth surfaces. In fact, if $\nabla u \neq 0$ at a point, then one can prove that all nearby level sets are smooth surfaces near the point in question. This important fact is a consequence of the general Implicit Function Theorem, [**126**]. Thus, if $\nabla u \neq 0$ at all points on a level set, then the level set is a smooth surface, and, if bounded, a simple closed surface. (On the other hand, finding an explicit parametrization of a level set may be quite difficult!)

Theorem B.21. *If nonzero, the gradient vector $\nabla u \neq \mathbf{0}$ defines the normal direction to the level set $\{\, u = c \,\}$ at each point.*

Proof: Indeed, suppose $\mathbf{x}(t)$ is any curve contained in the level set, so that

$$u(x(t), y(t), z(t)) = c \qquad \text{for all} \qquad t.$$

Since c is constant, the derivative with respect to t is zero, and hence, by (B.70),

$$\frac{d}{dt}\, u(x(t), y(t), z(t)) = \nabla u \cdot \dot{\mathbf{x}} = 0,$$

which implies that the gradient vector ∇u is orthogonal to the tangent vector $\dot{\mathbf{x}}$ to the curve. Since this holds for all such curves contained within the level set, the gradient must be orthogonal to the entire tangent plane at the point, and hence, if nonzero, defines a normal direction to the level surface. *Q.E.D.*

Physically, Theorem B.21 tells us that the direction of steepest increase in temperature is perpendicular to the isothermal surfaces at each point. Consequently, the solutions to the gradient flow equations (B.71) form an orthogonal system of curves to the level set surfaces of u, and one should follow these curves to minimize the temperature as rapidly as possible. Similarly, in a steady state fluid flow, the fluid potential is represented by a scalar field $\varphi(x, y, z)$. Its gradient $\mathbf{v} = \nabla \varphi$ determines the fluid velocity at each point. The streamlines followed by the fluid particles are the solutions to the gradient flow equations $\dot{\mathbf{x}} = \mathbf{v} = \nabla \varphi$, while the level sets of φ are the equipotential surfaces. Thus, fluid particles flow in a direction orthogonal to the equipotential surfaces.

Example B.22. The level sets of the radial function $u = x^2 + y^2 + z^2$ are the concentric spheres centered at the origin. Its gradient $\nabla u = (\, 2\,x, 2\,y, 2\,z \,)^T = 2\mathbf{x}$ points in the radial direction, orthogonal to each spherical level set. Note that $\nabla u = \mathbf{0}$ only at the origin, which is a level set, but not a smooth surface.

The radial vector also specifies the direction of fastest increase (decrease) in the function u. Indeed, the solution to the associated gradient flow system (B.71), namely

$$\dot{\mathbf{x}} = -\,2\mathbf{x} \qquad \text{is} \qquad \mathbf{x}(t) = \mathbf{x}_0\, e^{-2\,t},$$

where $\mathbf{x}_0 = \mathbf{x}(0)$ is the initial position. Therefore, to decrease the function u as rapidly as possible, one should follow a radial ray into the origin.

Example B.23. An implicit equation for the torus (B.28) is obtained by replacing $r = \sqrt{x^2 + y^2}$ in (B.29). In this manner, we are led to consider the level sets of the function

$$u(x, y, z) = x^2 + y^2 + z^2 - 4\sqrt{x^2 + y^2} \;=\; c, \tag{B.73}$$

with the particular value $c = -3$ corresponding to (B.28). The gradient of the function is

$$\nabla u(x, y, z) = \left(2\,x - \frac{4\,x}{\sqrt{x^2 + y^2}},\; 2\,y - \frac{4\,y}{\sqrt{x^2 + y^2}},\; 2\,z \right)^T, \tag{B.74}$$

which is well-define except on the z axis, where $x = y = 0$. Note that $\nabla F \neq \mathbf{0}$ unless $z = 0$ and $x^2 + y^2 = 4$. Therefore, the level sets of u are smooth, toroidal surfaces except for z axis and the circle of radius 2 in the (x, y) plane.

Divergence and Curl

The second important vector differential operator is the *divergence*,

$$\operatorname{div} \mathbf{v} = \nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z} \ . \tag{B.75}$$

The divergence maps a vector field $\mathbf{v} = (\, v_1, v_2, v_3 \,)^T$ to a scalar field $\mathbf{f} = \nabla \cdot \mathbf{v}$. For example, the radial vector field $\mathbf{v} = (\, x, y, z \,)^T$ has constant divergence $\nabla \cdot \mathbf{v} = 3$.

In fluid mechanics, the divergence measures the local, instantaneous change in the volume of a fluid packet as it moves. Thus, a steady state fluid flow is *incompressible*, with unchanging volume, if and only if its velocity vector field is divergence-free: $\nabla \cdot \mathbf{v} \equiv 0$. The connection between incompressibility and the earlier zero-flux condition will be addressed in the Divergence Theorem B.36 below.

The composition of divergence and gradient

$$\nabla \cdot \nabla u = \Delta u = u_{xx} + u_{yy} + u_{zz}$$

produces the Laplacian operator, just as in two dimensions. Indeed, as we shall see, except for the missing minus sign and the all-important boundary conditions, this is effectively the same as the self-adjoint form of the three-dimensional Laplacian:

$$\nabla^* \circ \nabla u = - \nabla \cdot \nabla u = - \Delta u.$$

See (Lapsa3■) below for details.

The third important vector differential operator is the *curl*, which, in three dimensions, maps vector fields to vector fields. It is most easily memorized in the form of a (formal) 3×3 determinant

$$\operatorname{curl} \mathbf{v} = \nabla \wedge \mathbf{v} = \begin{pmatrix} \dfrac{\partial v_3}{\partial y} - \dfrac{\partial v_2}{\partial z} \\ \dfrac{\partial v_1}{\partial z} - \dfrac{\partial v_3}{\partial x} \\ \dfrac{\partial v_2}{\partial x} - \dfrac{\partial v_1}{\partial y} \end{pmatrix} = \det \begin{pmatrix} \partial_x & v_1 & \mathbf{e}_1 \\ \partial_y & v_2 & \mathbf{e}_2 \\ \partial_z & v_3 & \mathbf{e}_3 \end{pmatrix}, \tag{B.76}$$

in analogy with the determinantal form (B.6) of the cross product. For instance, the radial vector field $\mathbf{v} = (\, x, y, z \,)^T$ has zero curl:

$$\nabla \wedge \mathbf{v} = \det \begin{pmatrix} \partial_x & x & \mathbf{e}_1 \\ \partial_y & y & \mathbf{e}_2 \\ \partial_z & z & \mathbf{e}_3 \end{pmatrix} = \mathbf{0}.$$

This is indicative of the lack or any rotational effect of the induced flow.

If $\mathbf{v}$ represents the velocity vector field of a steady state fluid flow, its curl $\nabla \wedge \mathbf{v}$ measures the instantaneous rotation of the fluid flow at a point, and is known as the *vorticity* of the flow. When non-zero, the direction of the vorticity vector represents the axis of rotation, while its magnitude $\| \nabla \wedge \mathbf{v} \|$ measures the instantaneous angular velocity

of the swirling flow. Physically, if we place a microscopic turbine in the fluid so that its shaft points in the direction specified by a unit vector $\mathbf{n}$, then its rate of spin will be proportional to component of the vorticity vector $\nabla \wedge \mathbf{v}$ in the direction of its shaft. This is equal to the dot product

$$\mathbf{n} \cdot (\nabla \wedge \mathbf{v}) = \| \nabla \wedge \mathbf{v} \| \cos \varphi,$$

where φ is the angle between $\mathbf{n}$ and the curl vector. Therefore, the maximal rate of spin will occur when $\varphi = 0$, and so the shaft of the turbine lines up with the direction of the vorticity vector $\nabla \wedge \mathbf{v}$. In this orientation, the angular velocity of the turbine will be proportional to its magnitude $\| \nabla \wedge \mathbf{v} \|$. On the other hand, if the axis of the turbine is orthogonal to the direction of the vorticity, then it will not rotate. If $\nabla \wedge \mathbf{v} \equiv \mathbf{0}$, then there is no net motion of a turbine, not matter which orientation it is placed in the fluid flow. Thus, a flow with zero curl is *irrotational*. The precise connection between this definition and the earlier zero circulation condition will be explained shortly.

Example B.24. Consider a helical fluid flow with velocity vector

$$\mathbf{v} = (\, -y, x, 1 \,)^T .$$

Integrating the ordinary differential equations $\dot{\mathbf{x}} = \mathbf{v}$, namely

$$\dot{x} = -y, \qquad \dot{y} = x, \qquad \dot{z} = 1,$$

with initial conditions $x(0) = x_0$, $y(0) = y_0$, $z(0) = z_0$ gives the flow

$$x(t) = x_0 \cos t - y_0 \sin t, \qquad y(t) = x_0 \sin t + y_0 \cos t, \qquad z(t) = z_0 + t. \quad (B.77)$$

Therefore, the fluid particles move along helices spiraling up the z axis, as illustrated in Figure hel■.

The divergence of the vector field $\mathbf{v}$ is

$$\nabla \cdot \mathbf{v} = \frac{\partial}{\partial x}(-y) + \frac{\partial}{\partial y} x + \frac{\partial}{\partial z} 1 = 0,$$

and hence the flow is incompressible. Indeed, any fluid packet will spiral up the z axis unchanged in shape, and so its volume does not change.

The vorticity or curl of the velocity is

$$\nabla \wedge \mathbf{v} = \begin{pmatrix} \dfrac{\partial}{\partial y} 1 - \dfrac{\partial}{\partial z} x \\ \dfrac{\partial}{\partial z}(-y) - \dfrac{\partial}{\partial x} 1 \\ \dfrac{\partial}{\partial x} x - \dfrac{\partial}{\partial y}(-y) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix},$$

which points along the z-axis. This reflects the fact that the flow is spiraling up the z-axis. If a turbine is placed in the fluid at an angle φ with the z-axis, then its rate of rotation will be proportional to $2 \cos \varphi$.

Example B.25. Any planar vector field $\mathbf{v} = (\, v_1(x,y), v_2(x,y)\,)^T$ can be identified with a three-dimensional vector field

$$\mathbf{v} = (\, v_1(x,y), v_2(x,y), 0\,)^T$$

that has no vertical component. If $\mathbf{v}$ represents a fluid velocity, then the fluid particles remain on horizontal planes $\{\, z = c\,\}$, and the individual planar flows are identical. Its three-dimensional curl

$$\nabla \wedge \mathbf{v} = \left(0, 0, \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right)^T$$

is a purely vertical vector field, whose third component agrees with the scalar two-dimensional curl (A.18) of $\mathbf{v}$. This provides the direct identification between the two- and three-dimensional versions of the curl operation. Indeed, our analysis of flows around airfoils in Chapter 16 directly relied upon this identification between two- and three-dimensional flows.

Interconnections and Connectedness

The three basic vector differential operators — gradient, curl and divergence — are intimately inter-related. The proof of the key identities relies on the equality of mixed partial derivatives, which in turn requires that the functions involved are sufficiently smooth. We leave the explicit verification of the key result to the reader.

Proposition B.26. *If u is a smooth scalar field, then $\nabla \wedge \nabla u \equiv \mathbf{0}$. If $\mathbf{v}$ is a smooth vector field, then $\nabla \cdot (\nabla \wedge \mathbf{v}) \equiv 0$.*

Therefore, the curl of any gradient vector field is automatically zero. As a consequence, all gradient vector fields represent irrotational flows. Also, the divergence of any vector field that is a curl is also automatically zero. Thus, all curl vector fields represent incompressible flows. On the other hand, the divergence of a gradient vector field is the Laplacian of the underlying potential, as we previously noted, and hence is zero if and only if the potential is a harmonic function.

The converse statements are almost true. As in the two-dimensional case, the precise statement of this result depends upon the topology of the underlying domain. In two dimensions, we only had to worry about whether or not the domain contained any holes, i.e., whether or not the domain was simply connected. Similar concerns arise in three dimensions. Moreover, there are two possible classes of "holes" in a solid domain, and so there are two different types of connectivity. For lack of a better terminology, we introduce the following definition.

Definition B.27. A domain $\Omega \subset \mathbb{R}^3$ is said to be

(a) 0–*connected* or *pathwise connected* if there is a curve $C \subset \Omega$ connecting any two points $\mathbf{x}_0, \mathbf{x}_1 \in \Omega$, so that[†] $\partial C = \{\, \mathbf{x}_0, \mathbf{x}_1 \,\}$.

[†] We use the notation ∂C to denote the endpoints of a curve C.

(*b*) 1–*connected* if every unknotted simple closed curve $C \subset \Omega$ is the boundary, $C = \partial S$ of an oriented surface $S \subset \Omega$.

(*c*) 2–*connected* if every simple closed surface $S \subset \Omega$ is the boundary, $S = \partial D$ of a subdomain $D \subset \Omega$.

Remark: The unknotted condition is to avoid considering "wild" curves that fail to bound any oriented surface $S \subset \mathbb{R}^3$ whatsoever.

For example, $\mathbb{R}^3$ is both 0, 1 and 2–connected, as are all solid balls, cubes, tetrahedra, solid cylinders, and so on. A disjoint union of balls is not 0–connected, although it does remain both 1 and 2–connected. The domain $\Omega = \left\{ 0 \leq r < \sqrt{x^2 + y^2} < R \right\}$ lying between two cylinders is not 1–connected since it has a "one-dimensional" hole drilled through it. Indeed, if $C \subset \Omega$ is any closed curve that encircles the inner cylinder, then every bounding surface S with $\partial S = C$ must pass across the inner cylinder and hence will not lie entirely within the domain. On the other hand, this cylindrical domain Ω is both 0 and 2–connected — even an annular surface that encircles the inner cylinder will bound a solid annular domain contained inside Ω. Similarly, the domain $\Omega = \{ 0 \leq r < \| \mathbf{x} \| < R \}$ between two concentric spheres is 0 and 1–connected, but not 2–connected owing to the spherical cavity inside. Any closed curve $C \subset \Omega$ will bound a surface $S \subset \Omega$; for instance, a circle going around the equator of the inner sphere will still bound a hemispherical surface that does not pass through the spherical cavity. On the other hand, a sphere that lies between the inner and outer spheres will not bound a solid domain contained within the domain. A full discussion of the topology underlying the various types of connectivity, the nature of holes and cavities, and their connection with the existence of scalar and vector potentials, must be deferred to a more advanced course in differential topology, [**21**, **68**].

We can now state the basic theorem relating the connectivity of domains to the kernels of the fundamental vector differential operators; see [**21**] for details.

Theorem B.28. *Let $\Omega \subset \mathbb{R}^3$ be a domain.*

(*a*) *If Ω is 0–connected, then a scalar field $u(x, y, z)$ defined on all of Ω has vanishing gradient, $\nabla u \equiv \mathbf{0}$, if and only if $u(x, y, z) =$* constant.

(*b*) *If Ω is 1–connected, then a vector field $\mathbf{v}(x, y, z)$ defined on all of Ω has vanishing curl, $\nabla \wedge \mathbf{v} \equiv \mathbf{0}$, if and only if there is a scalar field φ, known as a* scalar potential *for $\mathbf{v}$, such that $\mathbf{v} = \nabla \varphi$.*

(*c*) *If Ω is 2–connected, then a vector field $\mathbf{v}(x, y, z)$ defined on all of Ω has vanishing divergence, $\nabla \cdot \mathbf{v} \equiv 0$, if and only if there is a vector field $\mathbf{w}$, known as a* vector potential *for $\mathbf{v}$, such that $\mathbf{v} = \nabla \wedge \mathbf{w}$.*

If $\mathbf{v}$ represents the velocity vector field of a steady-state fluid flow, then the curl-free condition $\nabla \wedge \mathbf{v} \equiv \mathbf{0}$ corresponds to an irrotational flow. Thus, on a 2–connected domain, every irrotational flow field $\mathbf{v}$ has a scalar potential φ with $\nabla \varphi = \mathbf{v}$. The divergence-free condition $\nabla \cdot \mathbf{v} \equiv \mathbf{0}$ corresponds to an incompressible flow. If the domain is 1–connected, every incompressible flow field $\mathbf{v}$ has a vector potential $\mathbf{w}$ that satisfies $\nabla \wedge \mathbf{w} = \mathbf{v}$. The vector potential can be viewed as the three-dimensional analog of the stream function for planar flows. If the fluid is both irrotational and incompressible, then its scalar potential

satisfies

$$0 = \nabla \cdot \mathbf{v} = \nabla \cdot \nabla \varphi = \Delta \varphi,$$

which is Laplace's equation! Thus, just as in the two-dimensional case, the scalar potential to an irrotational, incompressible fluid flow is a harmonic function. This fact is used in modeling many problems arising in physical fluids, including water waves, [**Lighthill**]. Unfortunately, in three dimensions there is no counterpart of complex function theory to represent the solutions of the Laplace equation, or to connect the vector and scalar potentials.

Example B.29. The vector field

$$\mathbf{v} = (\, -y, x, 1\,)^T$$

that generates the helical flow (B.77) satisfies $\nabla \cdot \mathbf{v} = 0$, and so is divergence-free, reconfirming our observation that the flow is incompressible. Since $\mathbf{v}$ is defined on all of $\mathbb{R}^3$, Theorem B.28 assures us that there is a vector potential $\mathbf{w}$ that satisfies $\nabla \wedge \mathbf{w} = \mathbf{v}$. One candidate for the vector potential is

$$\mathbf{w} = \big(\, y, 0, \tfrac{1}{2}\, x^2 + \tfrac{1}{2}\, y^2 \,\big)^T .$$

The helical flow is not irrotational, and so it does not admit a scalar potential.

Remark: The construction of a vector potential is not entirely straightforward, but we will not dwell on this problem. Unlike a scalar potential which, when it exists, is uniquely defined up to a constant, there is, in fact, quite a bit of ambiguity in a vector potential. Adding in *any* gradient,

$$\widetilde{\mathbf{w}} = \mathbf{w} + \nabla \varphi$$

will give an equally valid vector potential. Indeed, using Proposition B.26, we have

$$\nabla \wedge \widetilde{\mathbf{w}} = \nabla \wedge \mathbf{w} + \nabla \wedge \nabla \varphi = \nabla \wedge \mathbf{w}.$$

Thus, any vector field of the form

$$\mathbf{w} = \left(y + \frac{\partial \varphi}{\partial x}\,,\ \frac{\partial \varphi}{\partial y}\,,\ \frac{x^2}{2} + \frac{y^2}{2} + \frac{\partial \varphi}{\partial z} \right)^T ,$$

where $\varphi(x, y, z)$ is an *arbitrary* function, is also a valid vector potential for the helical vector field $\mathbf{v} = (\, -y, x, 1\,)^T$.

B.8. The Fundamental Integration Theorems.

In three-dimensional vector calculus there are 3 fundamental differential operators — gradient, curl and divergence. There are also 3 types of integration — line, surface and volume integrals. And, not coincidentally, there are 3 basic theorems that generalize the Fundamental Theorem of Calculus to line, surface and volume integrals in three-dimensional space. In all three results, the integral of some differentiated quantity over a curve, surface, or domain is related to an integral of the quantity over its boundary. The

first theorem relates the line integral of a gradient over a curve to the values of the function at the boundary or endpoints of the curve. Stokes' Theorem relates the surface integral of the curl of a vector field to the line integral of the vector field around the boundary curve of the surface. Finally, the Divergence Theorem, also known as Gauss' Theorem, relates the volume integral of the divergence of a vector field to the surface integral of that vector field over the boundary of the domain.

The Fundamental Theorem for Line Integrals

We begin with the Fundamental Theorem for line integrals. This is identical to the planar version, as stated earlier in Theorems A.20 and A.21. We do not need to reproduce its proof again here.

Theorem B.30. *Let $C \subset \mathbb{R}^3$ be a curve that starts at the endpoint* $\mathbf{a}$ *and goes to the endpoint* $\mathbf{b}$*. Then the line integral of a gradient of a function along C is given by*

$$\int_C \nabla u \cdot \, d\mathbf{x} = u(\mathbf{b}) - u(\mathbf{a}). \tag{B.78}$$

Since its value only depends upon the endpoints, the line integral of a gradient is independent of path. In particular, if C is a closed curve, then $\mathbf{a} = \mathbf{b}$, and so the endpoint contributions cancel out:

$$\oint_C \nabla u \cdot \, d\mathbf{x} = 0.$$

Conversely, if $\mathbf{v}$ is any vector field with the property that its integral around any closed curve vanishes,

$$\oint_C \mathbf{v} \cdot \, d\mathbf{x} = 0, \tag{B.79}$$

then $\mathbf{v} = \nabla\varphi$ admits a potential. Indeed, as long as the domain is 0–connected, one can construct a potential $\varphi(\mathbf{x})$ by integrating over any convenient curve C connecting a fixed point $\mathbf{a} \in \Omega$ to the point $\mathbf{x}$

$$\varphi(\mathbf{x}) = \int_{\mathbf{a}}^{\mathbf{x}} \mathbf{v} \cdot \, d\mathbf{x}.$$

The proof that this is a well-defined potential is similar to the planar version discussed in Chapter A.

Example B.31. Line integrals over cylindrical and spherical domains. ∎

If $\mathbf{v}$ represents the velocity vector field of a three-dimensional steady state fluid flow, then its line integral around a closed curve C, namely

$$\oint_C \mathbf{v} \cdot d\mathbf{x} = \oint_C \mathbf{v} \cdot \mathbf{t} \, ds$$

is the integral of the tangential component of the velocity vector field. This represents the circulation of the fluid around the curve C. In particular, if the circulation line integral is 0 for every closed curve, then the fluid flow will be *irrotational* because $\nabla \wedge \mathbf{v} = \nabla \wedge \nabla\varphi \equiv \mathbf{0}$.

Stokes' Theorem

The second of the three fundamental integration theorems is known as *Stokes' Theorem*. This important result relates the circulation line integral of a vector field around a closed curve with the integral of its curl over any bounding surface. Stokes' Theorem first appeared in an 1850 letter from Lord Kelvin (William Thompson) written to George Stokes, who made it into on an undergraduate exam question for the Smith Prize at Cambridge University in England.

Theorem B.32. *Let $S \subset \mathbb{R}^3$ be an oriented, bounded surface whose boundary ∂S consists of one or more piecewise smooth simple closed curves. Let $\mathbf{v}$ be a smooth vector field defined on S. Then*

$$\oint_{\partial S} \mathbf{v} \cdot d\mathbf{x} = \iint_S (\nabla \wedge \mathbf{v}) \cdot \mathbf{n} \, dS. \tag{B.80}$$

To make sense of Stokes' formula (B.80), we need to assign a consistent orientation to the surface — meaning a choice of unit normal $\mathbf{n}$ — and to its boundary curve — meaning a direction to go around it. The proper choice is described by the following *left hand rule*: If we walk along the boundary ∂S with the normal vector $\mathbf{n}$ on S pointing upwards, then the surface should be on our left hand side; see Figure Stokes. For example, if $S \subset \{ z = 0 \}$ is a planar domain and we choose the upwards normal $\mathbf{n} = (0, 0, 1)^T$, then C should be oriented in the usual, counterclockwise direction. Indeed, in this case, Stokes' Theorem B.32 reduces to Green's Theorem A.25!

Stokes' formula (B.80) can be rewritten using the alternative notations (B.21), (B.47), for surface and line integrals in the form

$$\begin{aligned} &\oint_{\partial S} u\, dx + v\, dy + w\, dz = \\ &\qquad \iint_S \left(\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} \right) dy\, dz + \left(\frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} \right) dz\, dx + \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) dx\, dy. \end{aligned} \tag{B.81}$$

Recall that a closed surface is one without boundary: $\partial S = \varnothing$. In this case, the left hand side of Stokes' formula (B.80) is zero, and we find that integrals of curls vanish on closed surfaces.

Proposition B.33. *If the vector field $\mathbf{v} = \nabla \wedge \mathbf{w}$ is a curl, then $\displaystyle\iint_S \mathbf{v} \cdot \mathbf{n} \, dS = 0$ for every closed surface S.*

Thus, every curl vector field defines a surface-independent integral.

Example B.34. Let $S = \{ x + y + z = 1, \ x > 0, \ y > 0, \ z > 0 \}$ denote the triangular surface considered in Example B.15. Its boundary $\partial S = L_x \cup L_y \cup L_z$ is a triangle composed of three line segments

$$\begin{aligned} L_x &= \{ x = 0, \ y + z = 1, \ y \geq 0, \ z \geq 0 \}, \\ L_y &= \{ y = 0, \ x + z = 1, \ x \geq 0, \ z \geq 0 \}, \\ L_z &= \{ z = 0, \ x + y = 1, \ x \geq 0, \ y \geq 0 \}. \end{aligned}$$

To compute the line integral

$$\oint_{\partial S} \mathbf{v} \cdot d\mathbf{x} = \oint_{\partial S} y^2\, dx + x\, z^2\, dy$$

of the vector field $\mathbf{v} = \big(\, y^2, x\, z^2, 0 \,\big)^T$, we could proceed directly, but this would require evaluating three separate integrals over the three sides of the triangle. As an alternative, we can use Stokes formula (B.80), and compute the integral of its curl $\nabla \wedge \mathbf{v} = (\, 2\, y, 2\, x\, z, 0\,)^T$ over the triangle, which is

$$\oint_{\partial S} \mathbf{v} \cdot d\mathbf{x} = \iint_S (\nabla \wedge \mathbf{v}) \cdot \mathbf{n}\; dS = \iint_S 2\, y\, dy\, dz + 2\, x\, z\, dz\, dx = \frac{17}{12}\,,$$

where this particular surface integral was already computed in Example B.15.

We remark that Stokes' Theorem B.32 is consistent with Theorem B.28. Suppose that $\mathbf{v}$ is a curl-free vector field, so $\nabla \wedge \mathbf{v} = \mathbf{0}$, which is defined on a 1–connected domain $\Omega \subset \mathbb{R}^3$. Since every simple (unknotted) closed curve $C \subset \Omega$ bounds a surface, $C = \partial S$, with $S \subset \Omega$ also contained inside the domain, then, Stokes' formula (B.80) implies

$$\oint_C \mathbf{v} \cdot d\mathbf{x} = \iint_S (\nabla \wedge \mathbf{v}) \cdot \mathbf{n}\; dS = 0.$$

Since this happens for *every*[†] $C \subset \Omega$, then the path-independence condition (B.79) is satisfied, and hence $\mathbf{v} = \nabla \varphi$ admits a potential.

Example B.35. The Newtonian gravitational force field

$$\mathbf{v}(\mathbf{x}) = \frac{\mathbf{x}}{\| x \|^3} = \frac{(\, x, y, z\,)^T}{(x^2 + y^2 + z^2)^{3/2}}$$

is well defined on $\Omega = \mathbb{R}^3 \setminus \{\mathbf{0}\}$, and is divergence-free: $\operatorname{div} \mathbf{v} \equiv 0$. Nevertheless, this vector field does not admit a vector potential. Indeed, on the sphere $S_a = \{\, \| \mathbf{x} \| = a \,\}$ of radius a, the unit normal vector at a point $\mathbf{x} \in S_a$ is $\mathbf{n} = \mathbf{x}/\| \mathbf{x} \|$. Therefore,

$$\iint_{S_a} \mathbf{v} \cdot \mathbf{n}\; dS = \iint_{S_a} \frac{\mathbf{x}}{\| \mathbf{x} \|^3} \cdot \frac{\mathbf{x}}{\| \mathbf{x} \|}\; dS = \iint_{S_a} \frac{1}{\| \mathbf{x} \|^2}\; dS = \frac{1}{a^2} \iint_{S_a}\; dS = 4\, \pi,$$

since S_a has surface area $4\, \pi\, a^2$. Note that this result is independent of the radius of the sphere. If $\mathbf{v} = \nabla \wedge \mathbf{w}$, this would contradict Proposition B.33.

The problem is, of course, that the domain Ω is not 2–connected, and so Theorem B.28 does not apply. However, it would apply to the vector field $\mathbf{v}$ on any 2–connected subdomain, for example the domain $\widetilde{\Omega} = \mathbb{R}^3 \setminus \{\, x = y = 0, z \leq 0 \,\}$ obtained by omitting the negative z-axis. Exercise ■ asks you to construct a vector potential in this case.

We further note that $\mathbf{v}$ is curl free: $\nabla \wedge \mathbf{v} \equiv \mathbf{0}$. Since the domain of definition Ω is 1–connected, Theorem B.28 tells us that $\mathbf{v}$ admits a scalar potential — the Newtonian gravitational potential. Indeed, $\nabla\big(\, \| \mathbf{x} \|^{-1}\,\big) = \mathbf{v}$, as the reader can check.

[†] It suffices to know this for unknotted curves to conclude it for arbitrary closed curves.

The Divergence Theorem

The last of the three fundamental integral theorems is the *Divergence Theorem*, also known as *Gauss' Theorem*. This result relates a surface flux integral over a closed surface to a volume integral over the domain it bounds.

Theorem B.36. *Let $\Omega \subset \mathbb{R}^3$ be a bounded domain whose boundary $\partial\Omega$ consists of one or more piecewise smooth simple closed surfaces. Let* $\mathbf{n}$ *denote the unit* outward *normal to the boundary of Ω. Let* $\mathbf{v}$ *be a smooth vector field defined on Ω and continuous up to its boundary. Then*

$$\iint_{\partial\Omega} \mathbf{v} \cdot \mathbf{n} \; dS = \iiint_{\Omega} \nabla \cdot \mathbf{v} \; dx\, dy\, dz. \tag{B.82}$$

In terms of the alternative notation (B.47) for surface integrals, the divergence formula (B.82) can be rewritten in the form

$$\iint_S u \; dy\, dz + v\, dz\, dx + w\, dx\, dy = \iiint_{\Omega} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) dx\, dy\, dz. \tag{B.83}$$

Example B.37. Let us compute the surface integral

$$\iint_S x\, y\, dz\, dx + z\, dx\, dy$$

of the vector field $\mathbf{v} = (\, 0, x\, y, z\,)^T$ over the sphere $S = \{\, \| \mathbf{x} \| = 1 \,\}$ of radius 1. A direct evaluation in either graphical or spherical coordinates is not so pleasant. But the divergence formula (B.83) immediately gives

$$\begin{aligned} \iint_S x\, y\, dz\, dx + z\, dx\, dy &= \iiint_{\Omega} \left(\frac{\partial (x\, y)}{\partial y} + \frac{\partial z}{\partial z} \right) dx\, dy\, dz \\ &= \iiint_{\Omega} (x+1)\, dx\, dy\, dz = \iiint_{\Omega} x\, dx\, dy\, dz + \iiint_{\Omega} dx\, dy\, dz = \tfrac{4}{3}\, \pi, \end{aligned}$$

where $\Omega = \{\, \| \mathbf{x} \| < 1 \,\}$ is the unit ball with boundary $\partial\Omega = S$. The final two integrals are, respectively, the x coordinate of the center of mass of the sphere multiplied by its volume, which is clearly 0, plus the volume of the spherical ball.

Example B.38. Suppose $\mathbf{v}(t, \mathbf{x})$ is the velocity vector field of a time-dependent fluid flow. Let $\rho(t, \mathbf{x})$ represent the density of the fluid at time t and position $\mathbf{x}$. Then the surface flux integral $\displaystyle\iint_S (\rho\, \mathbf{v}) \cdot \mathbf{n}\, dS$ represents the mass flux of fluid through the surface $S \subset \mathbb{R}^3$. In particular, if $S = \partial\Omega$ represents a closed surface bounding a domain Ω, then, by the Divergence Theorem B.36,

$$\iint_{\partial\Omega} (\rho\, \mathbf{v}) \cdot \mathbf{n}\, dS = \iiint_{\Omega} \nabla \cdot (\rho\, \mathbf{v})\, dx\, dy\, dz$$

represents the net mass flux out of the domain Ω at time t. On the other hand, this must equal the rate of change of mass in the domain, namely

$$-\frac{\partial}{\partial t}\iiint_{\Omega} \rho\, dx\, dy\, dz = -\iiint_{\Omega} \frac{\partial \rho}{\partial t}\, dx\, dy\, dz,$$

the minus sign coming from the fact that we are measuring net mass loss due to outflow. Equating these two, we discover that

$$\iiint_{\Omega} \left(\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho\, \mathbf{v}) \right) dx\, dy\, dz = 0$$

for *every* domain occupied by the fluid. Since the domain is arbitrary, this can only happen if the integrand vanishes, and hence

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho\, \mathbf{v}) = 0. \tag{B.84}$$

The latter is the basic *continuity equation* of fluid mechanics, which takes the form of a conservation law.

For a steady state fluid flow, the left hand side of the divergence formula (B.82) measures the fluid flux through the boundary of the domain $\partial\Omega$, while the left hand side integrates the divergence over the domain Ω. As a consequence, the divergence must represent the net local change in fluid volume at a point under the flow. In particular, if $\nabla \mathbf{v} = 0$, then there is no net flux, and the fluid flow is incompressible.

The Divergence Theorem B.36 is also consistent with Theorem B.28. Let $\mathbf{v}$ is a divergence-free vector field, $\nabla \cdot \mathbf{v} = 0$, defined on a 2–connected domain $\Omega \subset \mathbb{R}^3$. Every simple closed surface $S \subset \Omega$ bounds a subdomain, so $S = \partial D$, with $D \subset \Omega$ also contained inside the domain of definition of $\mathbf{v}$. Then, by the divergence formula (B.82),

$$\iint_{S} \mathbf{v} \cdot \mathbf{n}\; dS = \iiint_{\Omega} \nabla \cdot \mathbf{v}\; dx\, dy\, dz = 0.$$

Therefore, by Theorem B.28, $\mathbf{v} = \nabla \wedge \mathbf{w}$ admits a vector potential.

Remark: The proof of all three of the fundamental integral theorems, can, in fact, be reduced to the Fundamental Theorem of (one-variable) Calculus. They are, in fact, all special cases of the general Stokes' Theorem, which forms the foundation of the profound theory of integration on manifolds, [**2**, **21**, **58**]. Stokes' Theorem has deep and beautiful connections with topology — and is of fundamental importance in modern mathematics and physics. However, the full ramifications lie beyond the scope of this introductory text.

Appendix C

Infinite Series

When applied to partial differential equation in higher dimensions, the separation of variables method often results in ordinary differential equations of a non-elementary type. Their solutions are expressed in terms of certain remarkable and important non-elementary functions — the Bessel functions. These so-called *special functions* do not appear in elementary calculus, but do play a starring role in more advanced applications in physics, engineering and mathematics. Most interesting special functions arise as solutions to certain second order, self-adjoint boundary value problems of Sturm–Liouville type. As such, they obey basic orthogonality relations, and thus can be used in place of the trigonometric sines and cosines that form the foundations of elementary Fourier analysis. Thus, the series solutions of higher dimensional partial differential equations lead naturally to the study of Fourier–Bessel series. In this appendix, we collect together the required results about the most important classes of special functions, including a short presentation of the series approach for solving non-elementary ordinary differential equations.

C.1. Power Series.

By definition, a power series

$$f(x) \quad \sim \quad c_0 + c_1\, x + \, \cdots \, + c_n\, x^n + \, \cdots \, = \, \sum_{k=0}^{\infty} c_k\, x^k$$

can be viewed as an infinite linear combination of the basic monomials $1, x, x^2, x^3, \ldots$. According to Taylor's formula, (C.8), the coefficients are given in terms of the derivatives of the function at the origin,

$$c_k = \frac{f^{(k)}(0)}{k!}\,,$$

not by an inner product. The partial sums

$$s_n(x) = c_0 + c_1\, x + \, \cdots \, + c_n\, x^n \; = \; \sum_{k=0}^{n} c_k\, x^k$$

of a power series are ordinary polynomials, and the same convergence questions arise.

A power series either converges everywhere, or on an interval centered at 0, or nowhere except at 0. (See Section 16.2 for details.) A Fourier series can converge on quite bizarre sets. In fact, a detailed analysis of the convergence of Fourier series led Georg Cantor to establish the foundations of modern set theory, and, thus, had a seminal impact on the

very foundations of mathematics and logic. Secondly, when a power series converges, it converges to an analytic function, which is infinitely differentiable, and whose derivatives are also represented by power series, obtained by termwise differentiation.

Taylor's Theorem

Taylor's theorem with remainder.

Theorem C.1. *The Taylor expansion of $u(x+h)$, where h is viewed as small, is*

$$u(x+h) = u(x) + u'(x)\,h + u''(x)\,\frac{h^2}{2} + \cdots + u^{(n)}(x)\,\frac{h^n}{n!} + R_n(x,h), \tag{C.1}$$

where the remainder term has various forms, of which the most relevant is Cauchy's form

$$R_n(x,h) = u^{(n+1)}(\xi)\,\frac{h^{n+1}}{(n+1)!}, \tag{C.2}$$

where ξ is a point lying between x and $x+h$. Note that the error is of order h^{n+1}. If we are not interested in the precise form of the error term, we write

$$R_n(x,h) = \mathrm{O}(h^{n+1}),$$

to indicate its order, which is $n+1$, in the small parameter h.

For example, at

A *Taylor series* is

Example C.2. Expansions of powers $(1+x)^r$.

The coefficients are called binomial coefficients, and denoted

$$\binom{r}{i} = \frac{r\,(r-1)\,(r-2)\,\cdots\,(r-i+1)}{1\cdot 2\cdot 3\cdot\,\cdots\,\cdot i}. \tag{C.3}$$

If $r = n$ is a non-negative integer, then the Taylor expansion terminates, and reduces to the well-known Binomial Theorem

$$(x+y)^n = x^n + n\,x^{n-1}\,y + \frac{n\,(n-1)}{2}\,x^{n-2}\,y^2 + \cdots + n\,x\,y^{n-1} + y^n = \sum_{i=0}^{n} \binom{n}{i}\,x^{n-i}\,y^i. \tag{C.4}$$

We can also write

$$\binom{n}{i} = \frac{n\,(n-1)\,(n-2)\,\cdots\,(n-i+1)}{1\cdot 2\cdot 3\cdot\,\cdots\,\cdot i} = \frac{n!}{i!\,(n-i)!}$$

Corollary: Mean Value Theorem

Theorem C.3. *Suppose $f(u)$ is continuously differentiable. Then*

$$f(x) - f(a) = f'(\xi)\,(x-a) \qquad \text{for some } \xi \text{ between } x \text{ and } a. \tag{C.5}$$

The first order Taylor expansion

$$u(x+h) = u(x) + u'(x)\,h + \mathrm{O}(h) = u(x) + u'(x)\,h + R_1(x,h), \tag{C.6}$$

See (Taylor2iter■)

$$g(u) = g(u^\star) + g'(u^\star)\,(u-u^\star) + \tfrac{1}{2}\,g''(w)\,(u-u^\star)^2,$$

immediately produces the simplest finite difference approximation to the first derivative

$$u'(x) \;\approx\; \frac{u(x+h) - u(x)}{h}\,.$$

Thus, computing the usual difference quotient at two nearby points, x and $x+h$, produces a simple approximation to the derivative $u'(x)$. Geometrically, we are approximating the slope of the tangent line to the graph of u at x by the sloe of a secant line through two nearby points on the graph. The error in the approximation, meaning the difference between the two sides of (14.129), is

$$\frac{1}{h}\,R_1(x,h) = \frac{1}{2}\,u''(\xi)h = \mathrm{O}(h).$$

We say the finite difference approximation (14.132) is *first order* because the error is proportional to h, and write

$$u'(x) = \frac{u(x+h) - u(x)}{h} + \mathrm{O}(h).$$

The second order Taylor expansion

$$u(x+h) \;\approx\; u(x) + u'(x)\,h + \tfrac{1}{2}\,u''(x)\,h^2. \tag{C.7}$$

The remainder is

Recall that a function $f(x)$ is called *analytic* at a point a if it is smooth, and, moreover, its Taylor series

$$f(a) + f'(a)\,(x-a) + \tfrac{1}{2}\,f''(a)\,(x-a)^2 + \;\cdots\; = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}\,(x-a)^n \tag{C.8}$$

converges to $f(x)$ for all x sufficiently close to a. (It does not have to converge on the entire interval I.) Not every smooth function is analytic, and so $\mathcal{A}(I) \subsetneq \mathrm{C}^\infty(I)$. An explicit example is the function

$$f(x) = \begin{cases} e^{-1/x}, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{C.9}$$

It can be shown that every derivative of this function at 0 exists and equals zero: $f^{(n)}(0) = 0$, $n = 0, 1, 2, \ldots$, and so the function is smooth. However, its Taylor series at $a = 0$ is

$0 + 0\,x + 0\,x^2 + \cdots \equiv 0$, which converges to the zero function, not to $f(x)$. Therefore $f(x)$ is *not* analytic at $a = 0$.

In the vector-valued case first order Taylor expansion of a vector-valued function at a point $\mathbf{u}^\star$ takes the form

$$\mathbf{g}(\mathbf{u}) = \mathbf{g}(\mathbf{u}^\star) + \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) + R(\mathbf{u} - \mathbf{u}^\star). \tag{C.10}$$

C.2. Laurent Series.

As we know, any function $f(z)$ which is analytic at a point $z = a$ has a power series expansion (16.26). If $f(z)$ has a pole at the point a, then power series expansion of the usual form at the point a. Nevertheless, there is a more general expansion, named after the French mathematician Laurent, that converges to $f(z)$ at all nearby points z (except, of course, at the singularity itself $z = a$).

If $f(z)$ has a pole of order n at $z = a$, then, by definition, the function

$$g(z) = (z-a)^n f(z) = b_0 + b_1(z-a) + b_2(z-a)^2 + \cdots$$

is analytic at $z = a$, and hence can be expanded in an ordinary power series. Dividing both sides of this equation by $(z-a)^n$ and relabeling the coefficients $c_k = b_{k+n}$ for clarity, we find

$$\begin{aligned} f(z) &= \frac{p(z)}{(z-a)^n} + h(z) \qquad\qquad (C.11)\\ &= \frac{c_{-n}}{(z-a)^n} + \frac{c_{1-n}}{(z-a)^{n-1}} + \cdots + \frac{c_{-1}}{z-a} + c_0 + c_1(z-a) + c_2(z-a)^2 + \cdots . \end{aligned}$$

Here

$$\begin{aligned} p(z) &= c_{-n} + c_{1-n}(z-a) + \cdots + c_{-1}(z-a)^{n-1},\\ h(z) &= c_0 + c_1(z-a) + c_2(z-a)^2 + \cdots , \end{aligned}$$

are, respectively, a polynomial of degree $\le n-1$ and an analytic function near $z = a$.

The series (C.11) is the *Laurent series* for $f(z)$ at the pole $z = a$. The highest negative power $-n$ indicates the order n of the pole. In particular, $f(z)$ is analytic at $z = a$ if and only if $n = 0$ and the Laurent series reduces to a standard Taylor series. Laurent series only apply to poles. See [**4**] for generalizations that apply at branch points (Puiseux series) and essential singularities (Weierstrass ∎).

Example C.4. The function $f(z) = \dfrac{1}{1+z^2}$ has a simple pole at $z = \mathrm{i}$. To expand it in a Laurent series, we multiply by $z - \mathrm{i}$ and expand the resulting analytic function in a geometric series:

$$\frac{z-\mathrm{i}}{1+z^2} = \frac{1}{z+\mathrm{i}} = \frac{-\frac{\mathrm{i}}{2}}{1 - \frac{\mathrm{i}}{2}(z-\mathrm{i})} = -\sum_{n=1}^{\infty} \frac{\mathrm{i}^{\,n+1}}{2^{n+1}}(z-\mathrm{i})^n = -\frac{\mathrm{i}}{2} + \frac{1}{4}(z-\mathrm{i}) + \frac{\mathrm{i}}{8}(z-\mathrm{i})^2 + \cdots .$$

Therefore, the required Laurent series is

$$\frac{1}{1+z^2} = -\sum_{n=1}^{\infty} \frac{\mathrm{i}^{\,n+1}}{2^{n+1}}(z-\mathrm{i}\,)^{n-1} = -\frac{\mathrm{i}}{2(z-\mathrm{i}\,)} + \frac{1}{4} + \frac{\mathrm{i}}{8}(z-\mathrm{i}\,) + \cdots . \qquad (C.12)$$

This converges for $0 < |\, z - \mathrm{i}\,| < 2$.

Remark: Laurent series have a radius of convergence ρ which, just like power series, measures the distance from the pole to the nearest other singularity fo the function. The series converges for all z satisfying $0 < |\, z - a\,| < \rho$, and, possibly, some of the points on the boundary $|\, z - a\,| = \rho$ of the convergence disk.

Remark: One can manipulate Laurent series in a very similar fashion as power series. Thus, addition, multiplication, division, and so on are defined in the evident fashion.

According to the basic residue identity (16.139), the only term in the series that has a nontrivial residue is the one involving $1/(z-a)$. Therefore, we can immediately read off the residue at a pole of a complex function from its Laurent series expansion.

Lemma C.5. *The residue of $f(z)$ at $z = a$ is the coefficient of $(z-a)^{-1}$ in its Laurent series expansion* (C.11)*:*

$$\operatorname*{Res}_{z=a} f(z) = c_{-1}.$$

For example, by inspection of the Laurent series (C.12), we immediately find that

$$\operatorname*{Res}_{z=\mathrm{i}} \frac{1}{1+z^2} = -\frac{\mathrm{i}}{2},$$

which is the coefficient of the initial term.

C.3. Special Functions.

Very few differential equations can be solved explictly in closed form. Even for linear ordinary differential equations, once one tries to move beyond the simplest constant coefficient equations, there are not very many examples with explicit solutions. One important example that has been solved are the Euler equations (3.76) But many other fairly simple second order equations, including the Bessel and Legendre equations that arose as a result of our separation of variables solution to partial differential equations, do not have elementary functions as solutions. These and other equations that appear in a number of key applications lead to new types of "special functions" that occur over and over again in applications.

Just as the student learned to become familiar with exponential and trigonometric functions, thus, at a more advanced level, applications in physics, engineering and mathematics require gaining some familiarity with the properties of these functions. The purpose of this section is to introduce the student to some basic properties of the most important special functions, including the gamma function, the Airy functions, the Legendre functions and, finally the Bessel functions. Lack of space will prevent us from introducing

additional important special functions, such as hypergeometric functions, confluent hypergeometric functions, parabolic cylinder functions, the zeta function, elliptic functions, and many others. The interested reader can consult more advanced texts, such as [**116**, **149**], and the handbook [**3**], as well as the soon to appear update [**48**] and its web site, for the latest information on this fascinating and very active field of mathematics and applications. We should remark that there is no precise definition of the term "special function" — it merely designates a function that plays a distinguished role in a range of applications and whose properties and evaluation are therefore of particular interest.

Most special functions arise most naturally as solutions to second order linear ordinary differential equations with variable coefficients. One method of gaining anlytical insight into their properties is to formulate them as power series. Therefore, we will learn how to construct power series solutions to differential equations when closed form solutions are not available. As we shall see, although computationally messy at times, the power series method is straightforward to implement in practice. When we are at a regular point for the differential equation, the solutions can be obtained as ordinary power series. At so-called regular singular points, a more general type of series known as a Frobenius expansion is required. More general singular points require more advanced techniques, and will not be discussed here.

The Gamma Function

The first special function that we shall treat does not, in fact arise as the solution to a differential equation. Rather, it forms a generalization of the factorial function from integers to arbitrary real and complex numbers. As such, it will often appear in power series solutions to differential equations when parameters take on non-integral values.

First recall that the factorial of a non-negative integer n is defined inductively by the iterative formula

$$n\,! = n \cdot (n-1)\,!, \qquad \text{starting with} \qquad 0\,! = 1. \tag{C.13}$$

Thus, if n is a non-negative integer, the iteration based on the second formula terminates, and yields the familiar expression

$$n\,! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1. \tag{C.14}$$

If n is not a non-negative integer, then the iteration will not terminate, and we cannot use it to compute the factorial. Our goal is to introduce a function $f(x)$, defined for all values of x, that will play the role of such a factorial. The function should satisfy the functional equation

$$f(x) = x\, f(x-1) \tag{C.15}$$

where defined. If, in addition, $f(0) = 1$, then we know $f(n) = n\,!$ whenever n is a non-negative integer, and hence such a function will extend the definition of factorials to more general real and complex numbers.

A moment's thought should convince the student that there are many possible ways to construct such a function. The most important method relies on an integral formula, and leads to the definition of the gamma function, originally discovered by Euler.

Definition C.6. The *Gamma function* is defined by

$$\Gamma(z) = \int_0^\infty e^{-t}\, t^{z-1}\, dt. \tag{C.16}$$

For real z, the gamma function integral will converge provided $z > 0$; otherwise the singularity of t^{z-1} is too severe to permit convergence of the improper integral at $t = 0$. The key property that turns the gamma function into a substitute for the factorial function relies on an elementary integration by parts:

$$\Gamma(z+1) = \int_0^\infty e^{-t}\, t^z\, dt = -\, e^{-t}\, t^z \,\big|_{t=0}^{\infty} + z \int_0^\infty e^{-t}\, t^{z-1}\, dt.$$

The boundary terms vanish whenever $z > 0$, while the final integral is merely $\Gamma(z)$. Therefore, the gamma function satisfies the recurrence relation

$$\Gamma(z+1) = z\,\Gamma(z) \qquad \text{provided} \qquad z > 0. \tag{C.17}$$

If we set $f(x) = \Gamma(x+1)$, then (C.17) is the same as (C.15)! Moreover, by direct integration

$$\Gamma(1) = \int_0^\infty e^{-t}\, dt = 1.$$

Combining this with the recurrence relation (C.17), we deduce that

$$\Gamma(n+1) = n! \tag{C.18}$$

whenever $n \geq 0$ is a non-negative integer. Therefore, we can identify $x!$ with the value $\Gamma(x+1)$ whenever $x > -1$ is *any* real number.

Remark: The student may legitimately ask why not replace t^{z-1} by t^z in the definition of $\Gamma(z)$, which would avoid the $n-1$ in (C.18). There is no simple answer; we are merely following a well-established precedent set originally by Euler.

One important value of the gamma function is at $z = \frac{1}{2}$. Using the substitution $t = x^2$, with $dt = 2\,x\,dx = 2\,t^{1/2}$, we find

$$\Gamma\left(\tfrac{1}{2}\right) = \int_0^\infty e^{-t}\, t^{-1/2}\, dt = \int_0^\infty 2\, e^{-x^2}\, dx = \sqrt{\pi}, \tag{C.19}$$

where the final integral was evaluated earlier, (intex2■). Thus, using the identification with the factorial function, we identify this value with $\left(-\frac{1}{2}\right)! = \sqrt{\pi}$. The recurrence relation (C.17) will then fix the value of the gamma function at all half-integers $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$. For example,

$$\Gamma\left(\tfrac{3}{2}\right) = \tfrac{1}{2}\,\Gamma\left(\tfrac{1}{2}\right) = \tfrac{1}{2}\,\sqrt{\pi}, \tag{C.20}$$

and hence $\frac{1}{2}! = \frac{1}{2}\sqrt{\pi}$. Further properties of the gamma function are outlined in the exercises. A graph of the gamma function appear in Figure C.1. Note the appearance of singularities at negative integer values of $x = -1, -2, \ldots$.

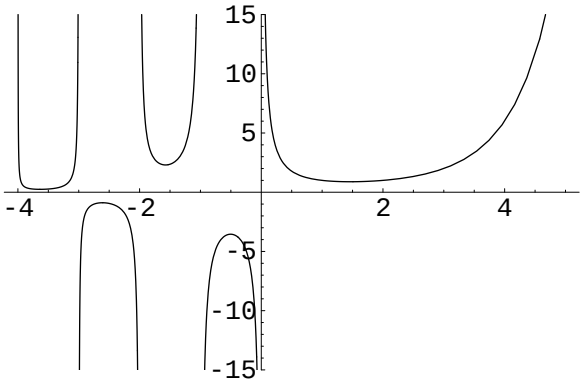


Figure C.1. The Gamma Function.

One of the most useful formulas involving the gamma function is *Stirling's Formula*,

$$\Gamma(n+1) = n! \quad \sim \quad \sqrt{2\,\pi\,n}\,\frac{n^n}{e^n}\,, \qquad\qquad n \longrightarrow \infty, \tag{C.21}$$

which gives the asympotitic values of the factorial function for large n. A proof is outlined in the exercises.

Series Solutions of Ordinary Differential Equations

When confronted with a novel differential equation, there are a few standard options for making progress in solving and understanding the solutions. One of these is the "look-up" method, that relies on published collections of differential equations and their solutions. One of the most useful references that collects together many solved differential equations is the classic German compendium written by Kamke, [**89**]. Two more recent English-language handbooks are [**156**, **158**].

Of course, numerical integration — see Chapter 20 for a presentation of basic methods — is always an option for approximating the solution. Numerical methods do, however, have their limitations, and are best accompanied by some understanding of the underlying theory, coupled with qualitative or quantitative expectations of how the solutions should behave. Furthermore, numerical methods provide less than adequate insight into the nature of the special functions that appear as solutions of the particular differential equations arising in separation of variables. A numerical approximation cannot, in itself, be used to establish rigorous mathematical properties of the solutions of the differential equation.

A more classical means of constructing and approximating the solutions of differential equations is based on their power series or Taylor series expansions. The Taylor expansion of a solution at a point x_0 is found by substituting a general power series into the differential equation and equating coefficients of the various powers of $(x - x_0)$. Using the initial conditions at x_0, the resulting system of equations serves to uniquely determine the coefficients and hence the derivatives of the solution at the initial point. The Taylor expansion of a special function can be used to deduce many of the key properties of the solution, as well as provide reasonable numerical approximations to its values within the radius of convergence of the series. (Howver, serious numerical computations are often performed through other methods, such as asymptotic expansions, [**116**].)

In this section, we provide a brief introduction to the basic series solution techniques for ordinary differential equations, concentrating on second order linear differential equations, since these form by far the most important class of examples arising in applications. When x_0 is a regular point, the method will construct a standard Taylor expansion for the solution, while so-called regular singular points require a slightly more general series expansion. Generalizations to higher order equations, nonlinear equations, and even (nonlinear) systems are left to other more detailed texts, including [**85**].

Regular Points

We shall concentrate on solving a homogeneous linear differential equation of the form

$$p(x)\,\frac{d^2u}{dx^2} + q(x)\,\frac{du}{dx} + r(x)\,u = 0. \tag{C.22}$$

Throughout this section, the coefficients $p(x), q(x), r(x)$ are assumed to be analytic functions where defined. This means that, at a point x_0, they admit convergent power series expansions

$$\begin{aligned} p(x) &= p_0 + p_1\,(x - x_0) + p_2\,(x - x_0)^2 + \cdots, \\ q(x) &= q_0 + q_1\,(x - x_0) + q_2\,(x - x_0)^2 + \cdots, \\ r(x) &= r_0 + r_1\,(x - x_0) + r_2\,(x - x_0)^2 + \cdots. \end{aligned} \tag{C.23}$$

We expect that the solutions to the differential equation will also be analytic functions. This expectation is valid provided that the equation is *regular* at the point x_0, in the sense that it is of genuinely second order, meaning that the coefficient $p(x)$ of the second derivative does not vanish at x_0.

Definition C.7. A point $x = x_0$ is a *regular point* of a second order linear ordinary differential equation (C.22) provided the leading coefficient does not vanish there:

$$p_0 = p(x_0) \neq 0.$$

A point where $p(x_0) = 0$ is known as a *singular point*.

Remark: The definition of a singular point assumes that the other coefficients do not both vanish there, i.e., either $q(x_0) \neq 0$ or $r(x_0) \neq 0$. If all three functions happen to vanish at x_0, we would factor out a common factor $(x - x_0)^k$, and hence, without loss of generality, can assume at least one of the coefficients is nonzero at x_0.

The basic existence theorem for differential equations at regular points follows. See [**85**; ■] for a proof.

Theorem C.8. *If x_0 is a regular point for the second order homogeneous linear ordinary differential equation* (C.22)*, then there exists a unique solution $u(x)$ to the initial value problem*

$$u(x_0) = a, \qquad u'(x_0) = b, \tag{C.24}$$

which is an analytic function for x sufficiently close to x_0.

Remark: It can be proved that the radius of convergence of any analytic solution $u(x)$ is equal to the distance from the regular point to the nearest singular point in the complex plane.

Therefore, every solution to an analytic differential equation at a regular point is an analytic function, and so can be expanded in an ordinary power series

$$u(x) = u_0 + u_1 (x - x_0) + u_2 (x - x_0)^2 + \cdots = \sum_{n=0}^{\infty} u_n (x - x_0)^n \qquad (C.25)$$

at the point. We remark that the power series is the same as the Taylor series for $u(x)$, and hence the coefficients

$$u_n = \frac{u^{(n)}(x_0)}{n!}$$

are multiples of the derivatives of the function at the point x_0. (Some authors prefer to keep the $n!$'s in the power series; this is purely a matter of taste.) In particular, the first two coefficients

$$u_0 = u(x_0) = a, \qquad u_1 = u'(x_0) = b. \qquad (C.26)$$

are prescribed by the initial conditions. Once the initial conditions have been specified, the remaining coefficients must be uniquely prescribed since there is only one solution to the initial value problem.

The basic method for constructing the power series solution to the initial value problem is quite straightforward. One substitutes the known power series (C.23) for the coefficient functions and the unknown power series (C.25) for the solution into the differential equation (C.22). Multiplying out the formulae will result in a (complicated) power series that must be equated to zero. At this point, one analyzes the individual coefficients. We rely on the basic observation that

Two power series are equal if and only if their individual coefficients are equal,

generalizing the standard test for equality of polynomials. In particular, a power series represents the zero function† if and only if all its coefficients are 0.

Thus, the power series solution method continues by equating, in order, the coefficients of the resulting power series to zero, starting with the lowest order (constant) and working upwards. The lowest order terms are multiples of $(x - x_0)^0 = 1$, i.e., the constant terms in the differential equation, lead to a linear recurrence relation

$$u_2 = R_2(u_0, u_1) = R_2(a, b)$$

that prescribes the coefficient u_2 in terms of the initial data. The coefficients of $(x - x_0)$ lead to a linear recurrence relation

$$u_3 = R_3(u_0, u_1, u_2) = R_3(a, b, R_2(a, b))$$

† Here it is essential that we work with analytic functions, since this result is *not* true for C^∞ functions! For example, the function e^{-1/x^2} has identically zero power series at $x_0 = 0$; see (e2x■).

that prescribes the coefficient u_3 in terms of the initial data and the prviously computed coefficient u_2. And so on. At the n^{th} stage of the procedure, the coefficients of $(x - x_0)^n$ lead to the n^{th} linear *recurrence relation*

$$u_{n+2} = R_n(u_0, u_1, \dots, u_{n+1}), \qquad n = 0, 1, 2, \dots , \tag{C.27}$$

that will prescribe the $(n + 2)^{\text{nd}}$ order coefficient in terms of the previous ones. Once the coefficients u_0 and u_1 have been specified by the initial conditions, the remaining coefficients $u_2, u_3, u_4, \dots$ are successively fixed by the recurrence relations (C.27). In this fashion, we easily deduce the existence of a formal power series solution to the differential equation at a regular point. The one remaining issue is whether the resulting power series actually converges. This can be proved with a detailed analysis, [**85**], and will serve to complete the proof of the general existence Theorem C.8.

At any regular point, the second order differential equation (C.22) admits two linearly independent analytic solutions, which we denote by $u(x)$ and $\widetilde{u}(x)$. The general solution can be written as a linear combination of the two basis solutions:

$$u(x) = a\, u(x) + b\, \widetilde{u}(x). \tag{C.28}$$

A standard choice for the two basis solutions is to take the first to satisfy the initial conditions

$$u(x_0) = 1, \qquad u'(x_0) = 0, \tag{C.29}$$

and the second to satisfy

$$\widetilde{u}(x_0) = 0, \qquad \widetilde{u}'(x_0) = 1, \tag{C.30}$$

although other choices may be used depending upon particular circumstances. With this choice, the linear combination automatically satisfies the initial conditions (C.24).

Rather than continue in generality, the best way to learn the method is to investigate simple examples.

The Airy Equation

A particularly easy case to analyze is the *Airy equation*

$$u'' = x\, u. \tag{C.31}$$

This second order ordinary differential equation arises in optics, dispersive waves, caustics (focusing of light waves as with a magnifying glass) and diffraction. It was first derived by the English mathematician Airy in 1839, [**5**]. In Exercise ■, we saw how it arises in a separation of variables solution to the Tricomi equation arising in supersonic fluid motion.

The solutions to the Airy equation are known as *Airy functions*. While Airy functions cannot be written in terms of the standard elementary functions, it is relatively straightforward to determine their power series expansion. Since the leading coefficient $p(x) \equiv 1$ is constant, and every point x_0 is a regular point of the Airy equation. For simplicity, we only treat the case $x_0 = 0$, and therefore consider a Maclaurin series

$$u(x) = u_0 + u_1\, x + u_2\, x^2 + u_3\, x^3 + \cdots \;= \sum_{n=0}^{\infty} u_n\, x^n$$

for the solution Term-by-term differentiation yields the series expansions[†]

$$\begin{aligned} u'(x) &= u_1 + 2\,u_2\,x + 3\,u_3\,x^2 + 4\,u_4\,x^3 + \cdots &&= \sum_{n=0}^{\infty} (n+1)\,u_{n+1}\,x^n, \\ u''(x) &= u_2 + 6\,u_3\,x + 12\,u_4\,x^2 + 20\,u_5\,x^3 + \cdots &&= \sum_{n=0}^{\infty} (n+1)(n+2)\,u_{n+2}\,x^n, \end{aligned} \tag{C.32}$$

for its derivatives. On the other hand,

$$x\,u(x) = u_0\,x + u_1\,x^2 + u_2\,x^3 + \cdots = \sum_{n=1}^{\infty} u_{n-1}\,x^n. \tag{C.33}$$

Substituting (C.33) and the formula for $u''(x)$ back into the Airy equation (C.31), and then equating the various powers of x leads to the following recurrence relations relating the coefficients of our power series. The left column indicates the power of x, while the right column tells us the recurrence relation:

$$\begin{array}{cc} 1 & u_2 = 0, \\ x & 6\,u_3 = u_0, \\ x^2 & 12\,u_4 = u_1, \\ x^3 & 20\,u_5 = u_2, \\ x^4 & 30\,u_6 = u_3, \\ \vdots & \vdots \\ x^n & (n+1)(n+2)\,u_{n+2} = u_{n-1}. \end{array}$$

We solve the recurrence relations in order. The first equation determines u_2. The second prescribes $u_3 = \frac{1}{6}\,u_0$ in terms of u_0. Next we find $u_4 = \frac{1}{12}\,u_1$ in terms of u_1. Next, $u_5 = \frac{1}{20}u_2 = 0$. Then $u_6 = \frac{1}{30}\,u_3 = \frac{1}{180}\,u_0$ is first given in terms of u_3, but we already know the latter in terms of u_0. And so on. At the n^{th} stage of the recursion, stage we determine u_{n+2} using our previsously generated formula for u_{n-1}.

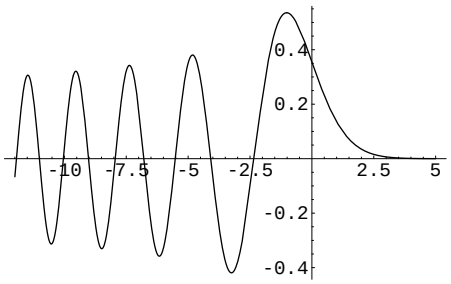
The only coefficients that are not determined by this procedure are the first two, u_0 and u_1. These correspond to the value of the solution and its derivative at $x_0 = 0$, as in (C.24). Let us construct the two basis solutions. The first uses the initial conditions

$$u_0 = u(0) = 1, \qquad u_1 = u'(0) = 0.$$

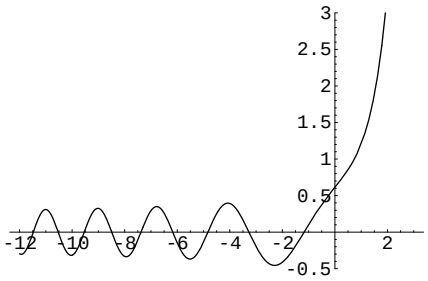
The recurrence relations then show that the only nonvanishing coefficients c_n are when $n = 3k$ is a multiple of 3; all others are zero. Moreover,

$$c_{3k} = \frac{c_{3k-3}}{3k(3k-1)}$$

[†] If we choose to work with the series in summation form, we need to re-index appropriately in order to display the term of degree n.



Ai(x)



Bi(x)

Figure C.2. The Airy Functions.

A straightforward induction proves that

$$c_{3k} = \frac{1}{3k(3k-1)(3k-3)(3k-4)\cdots 6\cdot 5\cdot 3\cdot 2}\,.$$

The resulting solution is known as the *Airy function of the first kind*, and denoted by

$$\mathrm{Ai}(x) = 1 + \tfrac{1}{6}\,x^3 + \tfrac{1}{180}\,x^6 + \cdots = \sum_{k=1}^{\infty} \frac{x^{3k}}{3k(3k-1)(3k-3)(3k-4)\cdots 6\cdot 5\cdot 3\cdot 2}\,. \quad (C.34)$$

Note that the denominator is similar to a factorial, except every third term is omitted. The "Ai" is read as a single symbol — not the product of A and i!

Similarly, starting with the initial conditions

$$u_0 = u(0) = 0, \qquad u_1 = u'(0) = 1,$$

we find that the only nonvanishing coefficients c_n are when $n = 3k+1$ leaves a remainder of 1 when divided by 3. The recurrence relation

$$c_{3k+1} = \frac{c_{3k-2}}{(3k+1)(3k)} \qquad \text{yields} \qquad c_{3k+1} = \frac{1}{(3k+1)(3k)(3k-2)(3k-3)\cdots 7\cdot 6\cdot 4\cdot 3}\,.$$

The resulting solution

$$\mathrm{Bi}(x) = x + \tfrac{1}{12}\,x^4 + \tfrac{1}{504}\,x^7 + \cdots = \sum_{k=1}^{\infty} \frac{x^{3k+1}}{(3k+1)(3k)(3k-2)(3k-3)\cdots 7\cdot 6\cdot 4\cdot 3} \quad (C.35)$$

is known as the *Airy function of the second kind*. Again, the denominator skips every third term in the product. Every solution to the Airy equation can be written as a linear combination

$$u(x) = a\ \mathrm{Ai}(x) + b\ \mathrm{Bi}(x), \qquad \text{where} \qquad a = u(0), \qquad b = u'(0)$$

correspond to the initial conditions of $u(x)$ at $x = 0$. The power series (C.34), (C.35), converge quite rapidly for all values of x, and so the first few terms provide a reasonable approximation to the two Airy functions for moderate values of x.

A graph of the two Airy functions appears in Figure C.2. Both functions oscillate for negative values of x, with a slowly decreasing amplitude. An intuitive explanation is

that when $x < 0$ the Airy equation (C.31) corresponds to a constant coefficient differential equation of the form $u'' = -k^2\, u$, which has oscillatory trigonometric solutions. On the other hand, when $x > 0$, the Airy equation is more like a constant coefficient differential equation of the form $u'' = +k^2\, u$, whose basis solutions $e^{k\,x}$ and $e^{-k\,x}$ are, respectively, exponentially growing and exponentially decaying. Indeed, as $x \to \infty$, the first Airy function $\mathrm{Ai}(x)$ decays very rapidly, whereas the second $\mathrm{Bi}(x)$ grows even more dramatically. Actually, the growth/decay rates are faster than exponential. It can be shown that

$$\mathrm{Ai}(x) \;\sim\; \begin{cases} \dfrac{e^{-\,2\,x^{3/2}/3}}{2\sqrt{\pi}\,x^{1/4}}\,, & x \to +\infty, \\[2ex] \dfrac{\sin\left(\dfrac{2}{3}\,(-x)^{3/2} + \dfrac{\pi}{4}\right)}{\sqrt{\pi}\,(-x)^{1/4}}\,, & x \to -\infty, \end{cases}$$

$$\mathrm{Bi}(x) \;\sim\; \begin{cases} \dfrac{e^{2\,x^{3/2}/3}}{2\sqrt{\pi}\,x^{1/4}}\,, & x \to +\infty, \\[2ex] \dfrac{\cos\left(\dfrac{2}{3}\,(-x)^{3/2} + \dfrac{\pi}{4}\right)}{\sqrt{\pi}\,(-x)^{1/4}}\,, & x \to -\infty. \end{cases}$$

Detailed investigations into the properties and numerical computation of the Airy functions can be found in [**3**, **48**, **116**].

The Legendre Equation

A particularly important example is the *Legendre equation*

$$(1-t^2)^2\,\frac{d^2P}{dt^2} - 2\,t\,(1-t^2)\,\frac{dP}{dt} + \big[\,\lambda\,(1-t^2) - m^2\,\big]\,P = 0. \qquad (C.36)$$

The integer m govenrs the *order* of the Legendre equation, while λ plays the role of an eigenvalue. As we learned in the preceding sections, this differential equation arises in the solutions to a wide variety of partial differential equations in spherical coordinates. The boundary conditions that serve to specify the eigenvalues are that the solution remain bounded at the two singular points $t = \pm 1$, leading to the boundary conditions

$$|\,P(-1)\,| < \infty, \qquad\qquad |\,P(+1)\,| < \infty. \qquad (C.37)$$

The point $t = 0$ is a regular point of the Legendre equation. Indeed, the only singular points are the boundary points $t = \pm 1$. Therefore, we can determine the solutions to the Legendre equation by the method of power series based at $t_0 = 0$. However, the general recurrence relations are rather complicated to solve in closed form, and we use some tricks to get a handle on the solutions.

Consider first the case $m = 0$. The Legendre equation of order 0 is

$$(1-t^2)\,\frac{d^2P}{dt^2} - 2\,t\,\frac{dP}{dt} + \lambda\,P = 0. \qquad (C.38)$$

As we noted above, the eigenfunctions are the *Legendre polynomials*

$$P_n(t) = \frac{1}{2^n\, n!}\, \frac{d^n}{dt^n}\, (t^2-1)^n.$$

They clearly satisfy the boundary conditions (C.37). To verify that they are indeed solutions to the differential equation (C.38), we let

$$q_n(t) = (t^2-1)^n.$$

By the chain rule, the derivative of $q_n(t)$ is

$$q_n' = 2n\,t\,(t^2-1)^{n-1} \qquad \text{and hence} \qquad (t^2-1)q_n' = 2n\,t\,(t^2-1)^n = 2n\,t\,q_n.$$

Differentiating the latter formula,

$$(t^2-1)q_n'' + 2\,t\,q_n' = 2n\,t : q_n' + 2n\,q_n, \qquad \text{or} \qquad (t^2-1)q_n'' = 2(n-1)t\,q_n' + 2n\,q_n.$$

A simple induction proves that the k^{th} order derivative $q_n^{(k)}(t) = d^k q_n/dt^k$ satisfies

$$\begin{aligned}(t^2-1)\,q_n^{(k+2)} &= 2(n-k-1)\,t\,q_n^{(k+1)} + 2[n + (n-1) + \cdots + (n-k)]\,q_n^{(k)} \\ &= 2(n-k-1)\,t\,q_n^{(k+1)} + (k+1)(2n-k)\,q_n^{(k)}.\end{aligned} \tag{C.39}$$

In particular, when $k = n$, this reduces to

$$(t^2-1)q_n^{(n+2)} = -2\,t\,q_n^{(n+1)} + n(n+1)\,q_n^{(n)} = 0,$$

and so $v_n = q_n^{(n)}$ satisfies

$$(1-t^2)v_n'' - 2\,t\,v_n' + n(n+1)\,v_n = 0,$$

which is precisely the order 0 Legendre equation (C.38) with eigenvalue parameter $\lambda = n(n+1)$. The Legendre polynomial P_n is a constant multiple of v_n, and hence it too satisfies the order 0 Legendre equation and hence forms an eigenfunction for the Legendre boundary value problem (C.36), (C.37). While it is not immediately apparent that the Legendre polynomials form a complete system of eigenfunctions, this is the case. This is the result of a general theory of eigenfunctions of Sturm–Liouville boundary value problems, [**34**], or, more particlarly, the theory of orthogonal polynomials, [**OP**]. Indeed, the orthogonality of the Legendre polynomials that was noted in Chapter 5 is, in fact, a consequence of the fact that they are eigenfunctions for this self-adjoint boundary value problem.

More generally, if we substitute $k = m + n$ in (C.39), we have

$$(1-t^2)\,w_n'' - 2(m+1)\,t\,w_n' + (m+n+1)(n-m)\,w_n = 0, \qquad \text{where} \qquad w_n = q_n^{(m+n)}. \tag{C.40}$$

This is *not* the order m Legendre equation, but can be converted into it by setting

$$w_n = (1-t^2)^{-m/2}\, z_n.$$

Differentiating, we find

$$\begin{aligned} w_n' &= (1-t^2)^{-m/2}\, z_n' - m\,t\,(1-t^2)^{-m/2-1}\, z_n, \\ w_n'' &= (1-t^2)^{-m/2}\, z_n'' - 2m\,t\,(1-t^2)^{-m/2-1}\, z_n' + (m + m(m+1)t^2)\,(1-t^2)^{-m/2-2}\, z_n. \end{aligned}$$

Therefore, after a little algebra, equation (C.40) takes the alternative form

$$(1-t^2)^{-m/2+1}\, z_n'' - 2\,t\,(1-t^2)^{-m/2}\, z_n' + (n(n+1)(1-t^2) - m^2)\,(1-t^2)^{-m/2-1}\, z_n = 0,$$

which, when multiplied by $(1-t^2)^{m/2+1}$, is precisely the order m Legendre equation (C.36) with eigenvalue parameter $\lambda = n(n+1)$. We conclude that

$$z_n(t) = (1-t^2)^{m/2} w_n(t) = (1-t^2)^{m/2}\, \frac{d^{n+m}}{dt^{n+m}}\, (t^2-1)^n$$

is a solution to the order m Legendre equation. Moreover, $z_n(\pm 1) = 0$, and hence $z_n(t)$ is an eigenfunction for the order m Legendre boundary value problem. Indeed, $z_n(t)$ is a constant multiple of the associated Legendre function $P_n^m(t)$, as defined in (18.27). With some more work, it can be proved that the associated Legendre functions form a complete system of eigenfunctions for the the order m Legendre boundary value problem.

Regular Singular Points

In certain situations, one is primarily interested in the behavior of solutions to a differential equation near a singular point. In most cases, a power series expansion (C.25) will not, in general, produce a solution at a singular point. As before, we write the differential equation as

$$p(x)\,\frac{d^2u}{dx^2} + q(x)\,\frac{du}{dx} + r(x)\,u = 0, \tag{C.41}$$

and assume that the functions p, q, r are analytic at x_0, where now we assume that $p(x_0) = 0$, but at least one of $q(x_0)$, $r(x_0)$ is non-zero. If the singular point is not too "wild", one can construct solutions using a relatively simple modification of the basic power series.

In order to formulate the key definition, we rewrite the differential equation in solved form

$$\frac{d^2u}{dx^2} = g(x)\,\frac{du}{dx} + h(x)\,u$$

where

$$g(x) = -\,\frac{q(x)}{p(x)}\,, \qquad\qquad h(x) = -\,\frac{r(x)}{p(x)}\,.$$

If $p(x_0) = 0$, then, typically, the functions $g(x), h(x)$ will have singularities at $x = x_0$, and we need to ensure that these singularities are not too bad.

Definition C.9. A singular point is called a *regular singular point* if

$$g(x) = \frac{k(x)}{(x-x_0)}\,, \qquad\qquad h(x) = \frac{\ell(x)}{(x-x_0)^2}\,, \tag{C.42}$$

where $k(x)$ and $\ell(x)$ are analytic at $x = x_0$.

Thus, in the language of complex analysis, the point x_0 is a regular singular point provided $g(x)$ has a pole of order at most 1, while $h(x)$ has a pole of order at most 2 at

$x = x_0$. In terms of the original coefficients, the regularity conditions (C.42) require that we can write the differential equation in the form

$$(x - x_0)^2\, s(x)\, \frac{d^2u}{dx^2} + (x - x_0)\, t(x)\, \frac{du}{dx} + r(x)\, u = 0, \tag{C.43}$$

where $s(x), t(x)$ and $r(x)$ are analytic at $x = x_0$ and, moreover, $s(x_0) \neq 0$.

Fortunately, almost all ordinary differential equations arising in applications have only regular singular points. The irregular singular points are much harder to deal with, and must be relegated to an advanced treatment, e.g., [**85**, **79**, **XX**].

The simplest example of an equation with a regular singular point is the Euler equation

$$a\, x^2\, u'' + b\, x\, u' + c\, u = 0, \tag{C.44}$$

where $a \neq 0, b, c$ are constants. The point $x = 0$ is a regular singular point; indeed, the solved form of the Euler equation is

$$u'' = -\,\frac{b}{a}\, x\, u' - \frac{c}{a}\, u,$$

and hence satisfies (C.42). All other points $x_0 \neq 0$ are regular points for the Euler equation.

As discussed in Example 7.34, Euler equations are solved by substituting the power ansatz $u(x) = x^r$ into the equation. As a result, the exponent r is determined by the associated characteristic equation (7.43), namely

$$a\, r\, (r - 1) + b\, r + c = 0.$$

If this quadratic equation has two distinct roots $r_1 \neq r_2$, we obtain two linearly independent (possibly complex) solutions $u(x) = x^{r_1}$ and $\widetilde{u}(x) = x^{r_2}$. The general solution $u(x) = c_1\, x^{r_1} + c_2\, x^{r_2}$ is a linear combination of these two basis solutions. Note that unless r_1 and r_2 are non-negative integers, the solutions have a singularity — either a pole or branch point — at the singular point $x = 0$. A repeated root, $r_1 = r_2$, requires an additional logarithmic term, $\widetilde{u}(x) = x^{r_1} \log x$, in the second solution, and the general solution has the form $u(x) = c_1\, x^{r_1} + c_2\, x^{r_1} \log x$.

The series solution method at more general regular singular points is modeled on the simple example of the Euler equation. One now seeks a solution that has a series expansion of the form

$$u(x) = (x - x_0)^r \sum_{n=0}^{\infty} u_n (x - x_0)^n = u_0 (x - x_0)^r + u_1 (x - x_0)^{r+1} + u_2 (x - x_0)^{r+2} + \cdots. \tag{C.45}$$

The full theory was established by the German mathematician Georg Frobenius in the late 1800's, and the series are sometimes known as *Frobenius expansions*. The exponent r is known as the *index* of the expansion.

Remark: If the inex $r = -n$ is a negative integer, then (C.45) has the form of a Laurent series expansion, as in (C.11). But r can be non-integral, or even complex, and the resulting expansion is known in complex analysis as a *Puiseux expansion*, [**complex**].

We can assume, without any loss of generality, that the leading coefficient $u_0 \neq 0$. Indeed, if $u_k \neq 0$ is the first non-zero coefficient, then the series begins with $u_k(x-x_0)^{r+k}$, and we replace r by $r+k$ to write it in the preceding form. Moreover, since any scalar multiple of a solution is a solution, we can divide by u_0 and assume that $u_0 = 1$ or any other convenient non-zero value, as desired.

Warning: Unlike ordinary power series expansions, the coefficients u_0 and u_1 are *not* prescribed by the initial conditions at the point x_0. Indeed, as we learned in our study of the Bessel and Legendre equations, one cannot typically impose specific initial values for the solutions at a singular point. Often, mere boundedness will suffice to distinguish a solution. Here, the solution is usually completely detremined by the index r and the leading coefficient u_0.

The Frobenius solution method proceeds by substituting the series (C.45) into the differential equation (C.43). Since

$$\begin{aligned} u(x) &= (x-x_0)^r + u_1(x-x_0)^{r+1} + \cdots, \\ (x-x_0)\,u'(x) &= r\,(x-x_0)^r + (r+1)u_1(x-x_0)^{r+1} + \cdots, \\ (x-x_0)\,u''(x) &= r\,(r-1)\,(x-x_0)^r + (r+1)\,r\,u_1(x-x_0)^{r+1} + \cdots, \end{aligned}$$

the lowest order terms are multiples of $(x-x_0)^r$. Equating this particular coefficient to zero leads to a quadratic equation of the form

$$s_0\,r\,(r-1) + t_0\,r + r_0 = 0, \tag{C.46}$$

where

$$s_0 = s(x_0) = \tfrac{1}{2}\,p''(x_0), \qquad t_0 = t(x_0) = q'(x_0), \qquad r_0 = r(x_0),$$

are the leading coefficients in the power series expansions of the coefficients of the differential equation. The quadratic equation (C.46) is known as the *indicial equation*, since it determines the possible indices r in the Frobenius expansion of a solution.

Therefore, just as in the Euler equation, it turns out that (typically) there are two allowable indices, say r_1 and r_2, which are the roots of the quadratic indicial equation. If the indices are distinct, then one expects to find two different Frobenius expansions. Usually, this assumption is valid, but there is an important exception, which occurs when the roots differ by an integer. The general result is summarized in the following list.

(*i*) If $r_2 - r_1$ is not an integer, then there are two linearly independent solutions $u(x)$ and $\widetilde{u}(x)$, each having a convergent Frobenius expansions of the form (C.45).

(*ii*) If $r_1 = r_2$, then there is only one solution with a convergent Frobenius expansion.

(*iii*) Finally, if $r_2 = r_1 + k$, where $k > 0$ is a positive integer, then there is a solution with a convergent Frobenius expansion corresponding to the smaller index r_1. The solution associated with the larger index r_2 may or may not have a convergent Frobenius expansion.

Thus, in every case the differential equation has at least one solution with a Frobenius expansion. When the leading coefficient $u_0 = 1$ is fixed, then the remaining coefficients $u_1, u_2, \ldots$ are uniquely prescribed by the recurrence relations stemming from substitution

of the expansion into the differential equation. If the second solution does not have a Frobenius expansion, then it has an additional logarithmic term, as with the Euler equation, of a well-prescribed form. Details appear in the exercises. Rather than try to develop the theory in any more detail here, we suffice with consideration of some particular examples.

Example C.10. Consider the second order ordinary differential equation

$$u'' + \left(\frac{1}{x} + \frac{x}{2} \right) u' + u = 0 \tag{C.47}$$

that we needed to solve for finding the fundamental solution to the heat equation; see (rheatode■). We look for series solutions based at $x = 0$. Since the coefficient of u' has a simple pole, the point $x = 0$ is a regular singular point, and thus we can work with a Frobenius expansion as in (C.49). Substituting into the differential equation, we find that the coefficients of x^r lead to the indicial equation

$$r^2 = 0.$$

There is only one root, $r = 0$, and hence even though we are at a singular point, we are dealing with an ordinary power series. The next term tells us that $u_1 = 0$. Since $r = 0$, the general recurrence relation is

$$(n+2)^2 u_{n+2} + \tfrac{1}{2}(n+2) u_n = 0,$$

and hence

$$u_{n+2} = - \frac{u_n}{2(n+2)} \, .$$

Therefore, the odd coefficients $u_{2k+1} = 0$ are all zero, while the even ones are

$$u_{2k} = - \frac{u_{2k-2}}{4k} = \frac{u_{2k-4}}{4k(4k-4)} = - \frac{u_{2k-6}}{4k(4k-4)(4k-8)} = \cdots = \frac{(-1)^k}{4^k \, k!} \qquad \text{since} \qquad u_0 = 1.$$

The resulting power series takes a familiar form:

$$u(x) = \sum_{k=1}^{\infty} u_{2k} x^{2k} = \sum_{k=1}^{\infty} \frac{1}{k!} \left(- \frac{x^2}{4} \right)^k = e^{-x^2/4},$$

reconfirming (rheatodesol■).

The second solution will require a logarithmic term. However, it can be found directly by a general reduction method. Once we know one solution to a second order ordinary differential equation, the second solution can be found by substituting the ansatz

$$\widetilde{u}(x) = u(x) \, v(x) = e^{-x^2/4} \, v(x)$$

into the equation. Thus,

$$\begin{aligned} \widetilde{u}'' + \left(\frac{1}{x} + \frac{x}{2} \right) \widetilde{u}' + \widetilde{u} &= \left[u'' + \left(\frac{1}{x} + \frac{x}{2} \right) u' + u \right] v + u \, v'' + 2 \, u' \, v' + \left(\frac{1}{x} + \frac{x}{2} \right) u \, v' \\ &= e^{-x^2/4} \left(v'' + \frac{1}{x} \, v' \right). \end{aligned}$$

Therefore, v' satisfies a linear first order ordinary differential equation:

$$v'' + \frac{v'}{x} = 0, \qquad \text{and hence} \qquad v' = c\,\frac{1}{x}, \qquad v = c\log x + d.$$

The general solution to the original differential equation is

$$\widetilde{u}(x) = u(x)\,v(x) = e^{-x^2/4}\,(c\log x + d).$$

Bessel's Equation

Perhaps the most important "non-elementary" ordinary differential equation is

$$x^2\,u'' + x\,u' + (x^2 - m^2)\,u = 0, \tag{C.48}$$

known as Bessel's equation of order m. We assume here that the order $m \geq 0$ is a non-negative real number; see Exercise ■ for the Bessel equation of imaginary order. As we have seen, the Bessel equation arises from separation of variables in a remarkable number of partial differential equations, including the Laplace, heat and wave equations on a disk, a cylinder, and a spherical ball. Interestingly, the solutions to the Bessel equation were first discovered by the German mathematician Bessel in a completely different context: the study of celestial mechanics, i.e., the Newtonian theory of planets orbiting around a central sun; see (19.16).

The Bessel equation cannot (except in a few particular instances) be solved in terms of elementary functions, and so the use of power series is natural. The leading coefficient $p(x) = x^2$ is nonzero *except* when $x = 0$, and so all points except the origin are regular points. Therefore, at all nonzero points $x_0 \neq 0$, the standard power series construction can be used to produce the appropriate power series solutions of the Bessel equation. However, the recurrence relations for the coefficients are not particularly easy to solve in clsoed form. Moreover, applications tend to demand understanding the behavior of the solutions to the Bessel equation at the singular point $x_0 = 0$. Writing the Bessel equation in solved form

$$u'' = -\,\frac{1}{x}\,u' + \left(\frac{m^2}{x^2} - 1\right) u,$$

we immediately see that $x = 0$ satisfies the conditions to qualify as a regular singular point. Consequently, we are led to seek a solution in the form of a Frobenius expansion. We first compute the expressions for the first two derivatives

$$\begin{aligned} u(x) &= x^r + u_1 x^{r+1} + u_2 x^{r+2} + \cdots \\ u'(x) &= r x^{r-1} + (r+1)u_1 x^r + (r+2)u_2 x^{r+1} + \cdots \\ u''(x) &= r(r-1)x^{r-2} + (r+1)r u_1 x^{r-1} + (r+2)(r+1)u_2 x^r + \cdots, \end{aligned} \tag{C.49}$$

of our purported solution. Substituting these expressions into (C.48), we find

$$\begin{aligned} &\big[\, r(r-1)\,x^r + (r+1)r\,u_1\,x^{r+1} + (r+2)(r+1)\,u_2\,x^{r+2} + \cdots \,\big] + \\ &+ \big[\, r\,x^r + (r+1)\,u_1\,x^{r+1} + (r+2)\,u_2\,x^{r+2} + \cdots \,\big] + \\ &+ \big[\, x^{r+2} + u_1\,x^{r+3} + u_2\,x^{r+4} + \cdots \,\big] - \big[\, m^2\,x^r + m^2\,u_1\,x^{r+1} + m^2\,u_2\,x^{r+2} + \cdots \,\big] = 0, \end{aligned}$$

We equate the coefficients of the various powers of x to zero. The coefficient of the lowest order power, x^r, is the indicial equation

$$r(r-1) + r - m^2 = r^2 - m^2 = 0.$$

There are two solutions to the indicial equation, $r = \pm m$, unless $m = 0$ in which case there is only one possible index $r = 0$.

The higher powers of x lead to recurrence relations for the successive coefficients u_n. If we replace m^2 by r^2, we find the following constraints:

$$\begin{aligned}
x^{r+1}: &\qquad \big[\,(r+1)^2 - r^2\,\big]u_1 = (2r+1)u_1 = 0, && u_1 = 0,\\
x^{r+2}: &\qquad \big[\,(r+2)^2 - r^2\,\big]u_2 + 1 = (4r+4)u_2 + 1 = 0, && u_2 = -\frac{1}{4r+4},\\
x^{r+3}: &\qquad \big[\,(r+3)^2 - r^2\,\big]u_3 + u_1 = (6r+9)u_3 + u_1 = 0, && u_3 = -\frac{u_1}{6r+9} = 0,
\end{aligned}$$

and, in general,

$$x^{r+n}: \qquad \big[\,(r+n)^2 - r^2\,\big]u_n + u_{n-2} = n(2r+n)u_n + u_{n-2} = 0.$$

Thus, the basic recurrence relation is

$$u_n = -\frac{1}{n(2r+n)}\,u_{n-2}, \qquad\qquad n = 2, 3, 4, \dots . \tag{C.50}$$

Starting with $u_0 = 1$, $u_1 = 0$, it is easy to deduce that all $u_n = 0$ for all odd $n = 2k+1$, while for even $n = 2k$,

$$\begin{aligned}
u_{2k} &= -\frac{u_{2k-2}}{4k(k+r)} = \frac{u_{2k-4}}{16\,k(k-1)(r+k)(r+k-1)} = \cdots\\
&= \frac{(-1)^k}{2^{2k}\,k(k-1)\cdots 3\cdot 2\,(r+k)(r+k-1)\cdots(r+2)(r+1)}\,.
\end{aligned}$$

Therefore, the series solution is

$$u(x) = \sum_{k=0}^{\infty} u_{2k}\,x^{m+2k} = \sum_{k=0}^{\infty} \frac{(-1)^k\,x^{m+2k}}{2^{2k}\,k(k-1)\cdots 3\cdot 2\,(r+k)(r+k-1)\cdots(r+2)(r+1)}\,. \tag{C.51}$$

So far, we not paid attention to the precise values of the indices $r = \pm m$, or whether our solution to the recurrence relations is valid. In order to continue the recurrence, we need to ensure that the recurrence relation (C.50) is legitimate, meaning that the denominator is never 0. Since $n > 0$, this will *not* be the case if and only if $2r + n = 0$, which requires that $r = -\frac{1}{2}n$ be either a negative integer $-1, -2, -3, \dots$, or half-integer, $-\frac{1}{2}, -\frac{3}{2}, -\frac{5}{2}, \dots$. These cases occur when the order $m = -r = \frac{1}{2}n$ is either an integer or a half-integer. Indeed, these cases are precisely the cases when the two indices, namely $r_1 = -m$ and $r_2 = m$, differ by an integer, $r_2 - r_1 = n$, and so we are in the tricky case (*iii*) of the Frobenius method.

There is, in fact, a key distinction between the integral and the half integral cases. Recall that the odd coefficients $u_{2\,k+1} = 0$ in the Frobenius series automatically vanish,

and so we only have to worry about the recurrence relation (C.50) for *even* values of n. Thus, for even $n = 2k$, the factor $2r + n = 2(r + k) = 0$ vanishes only when $r = -k$ is a negative integer; the half integral values do not, in fact cause problems. Therefore, if the order $m \geq 0$ is *not* a non-negative integer, then the Bessel equation of order m admits two linearly independent Frobenius solutions, given by the expansions (C.51) with exponents $r = +m$ and $r = -m$. If m is an integer, however, there is only one Frobeius solution, namely the expansion (C.51) with $r = +m$ given by the positive exponent. The second independent solution has an additional logarithmic term in its formula; details appear in Exercise ∎.

By convention, the standard *Bessel function* of order m is obtained by multiplying this solution by

$$\frac{1}{2^m\, m!} \qquad \text{or, rather,} \qquad \frac{1}{2^m\, \Gamma(m+1)}\,, \tag{C.52}$$

where the first factorial form can be used if m is a non-negative integer, while the more general gamma function expression must be employed for non-integral values of m. The result

$$J_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k\, x^{m+2k}}{2^{2k+m}\, k!\, (m+k)!}\,, \tag{C.53}$$

where, for non-integral m we interpret the factorial $(m + k)!$ as the gamma function $\Gamma(m + k + 1)$. The series is well-defined for all[†] m except when $m = -1, -2, -3, \ldots$ is a negative integer. We conclude that

Theorem C.11. *If $m \geq 0$ is* not *an integer, then the two linearly independent solutions to the Bessel equation of order m are the Bessel functions $J_m(x)$ and $J_{-m}(x)$. If $m = 0, 1, 2, 3, \ldots$ is an integer, the the Bessel function $J_m(x)$ is a solution to the Bessel equation. The second solution, traditionally denoted $Y_m(x)$, can be found as a limiting case*

$$Y_m(x) = \lim_{\nu \to m} Y_\nu(x) = \lim_{\nu \to m} \frac{\cos \nu\pi\, J_\nu(x) - J_{-\nu}(x)}{\sin \nu\pi} \tag{C.54}$$

of a certain linear combination of Bessel functions of non-integral order ν.

The justification of the last statement of the theorem can be found in Exercise ∎. We note that for $\nu \neq m$, the linear combination of Bessel functions in the limiting expression is a solution to the Bessel equation of order ν which is indepedent from $J_\nu(x)$. It can be proved that this continues to hold in the limit. The series formula for $Y_m(x)$ is quite complicated, [**116**, **145**], and its derivation is left to a more advanced course.

Example C.12. Consider the particular case when $m = \frac{1}{2}$. There are two indices, $r = \pm\frac{1}{2}$, for the Bessel equation of order $m = \frac{1}{2}$, leading to two solutions $J_{1/2}(x)$ and

[†] Actually, if m is a negative integer, the first $2m + 1$ terms in the series vanish because $\Gamma(-n) = \infty$ at negative integer values. The series $J_{-m}(x) = J_m(x)$ then actually coincides with its positive sibling.

$J_{-1/2}(x)$ obtained by the Frobenius method. For the first, with $r = \frac{1}{2}$, the recurrence relation (C.50) takes the form

$$u_n = -\frac{1}{(n+1)n}\, u_{n-2}\,.$$

Starting with $u_0 = 1$ and $u_1 = 0$, the general formula is easily found to be

$$u_n = \begin{cases} \dfrac{(-1)^k}{(n+1)!}\,, & n = 2\,k \text{ even}, \\ 0 & n = 2\,k+1 \text{ odd}. \end{cases}$$

Therefore, the resulting solution is

$$u(x) = \sqrt{x}\, \sum_{k=0}^{\infty} \frac{(-1)^k}{(2\,k+1)!}\, x^{2k} = \frac{1}{\sqrt{x}}\, \sum_{k=0}^{\infty} \frac{(-1)^k}{(2\,k+1)!}\, x^{2k+1} = \frac{\sin x}{\sqrt{x}}\,.$$

According to (C.52), the Bessel function of order $\frac{1}{2}$ is obtained by dividing this function by

$$\sqrt{2}\,\Gamma\left(\tfrac{3}{2}\right) = \sqrt{\frac{\pi}{2}}\,,$$

where we used (C.20) to evaluate the gamma function at $\frac{3}{2}$. Therefore,

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\, \sin x\,. \tag{C.55}$$

Similarly, for the other index $r = -\frac{1}{2}$, the recurrence relation

$$u_n = -\frac{1}{n(n-1)}\, u_{n-2}$$

leads to the formula

$$u_n = \begin{cases} \dfrac{(-1)^k}{n!}\,, & n = 2\,k \text{ even}, \\ 0 & n = 2\,k+1 \text{ odd}, \end{cases}$$

for the coefficients, corresponding to the solution

$$u(x) = x^{-1/2}\, \sum_{k=0}^{\infty} \frac{(-1)^k}{(2\,k)!}\, x^{2k} = \frac{\cos x}{\sqrt{x}}\,.$$

Therefore, using (C.52), (C.19), the Bessel function of order $-\frac{1}{2}$ is

$$J_{-1/2}(x) = \frac{\sqrt{2}}{\Gamma\left(\frac{1}{2}\right)}\, \frac{\cos x}{\sqrt{x}} = \sqrt{\frac{2}{\pi x}}\, \cos x\,. \tag{C.56}$$

Remark: If we now substitute (C.55) into the defining formula (18.85) for the spherical Bessel functions, we prove our earlier elementary formula (18.86) for the spherical Bessel function of order 0.

Finally, we demonstrate how Bessel functions of different orders are related by an important recurrence relation.

Proposition C.13. *The Bessel functions are interconnected by the following recurrence formulae:*

$$\frac{dJ_m}{dx} + \frac{m}{x}\,J_m = J_{m-1}, \qquad \frac{dJ_m}{dx} - \frac{m}{x}\,J_m = -\,J_{m+1}. \tag{C.57}$$

Proof: Let us differentiate the power series

$$x^m\,J_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k\,x^{2m+2k}}{2^{2k+m}\,k!\,(m+k)!}.$$

We find

$$\begin{aligned}\frac{d}{dx}\,[\,x^m\,J_m(x)\,] &= \sum_{k=0}^{\infty} \frac{(-1)^k\,2\,(m+k)x^{2m+2k-1}}{2^{2k+m}\,k!\,(m+k)!}\\ &= x^m \sum_{k=0}^{\infty} \frac{(-1)^k\,x^{m-1+2k}}{2^{2k+m-1}\,k!\,(m-1+k)!} = x^m\,J_{m-1}(x).\end{aligned}$$

Expansion of the left hand side of this formula leads to

$$x^m\,\frac{dJ_m}{dx} + m\,x^{m-1}\,J_m(x) = \frac{d}{dx}\,[\,x^m\,J_m(x)\,] = x^m\,J_{m-1}(x),$$

which proves the first recurrence formula (C.57). The second formula is proved by a similar manipulation involving differentiation of $x^{-m}\,J_m(x)$. *Q.E.D.*

Example C.14. For instance, we can use (C.57) to find the corresponding recurrence formulae for the spherical Bessel functions

$$S_n(x) = \sqrt{\frac{\pi}{2\,x}}\;J_{n+1/2}(x).$$

Differentiating and using the second recurrence relation, we find

$$\begin{aligned}\frac{dS_n}{dx} &= \sqrt{\frac{\pi}{2\,x}}\;\frac{dJ_{n+1/2}}{dx} - \frac{1}{2}\sqrt{\frac{\pi}{2}}\;\frac{1}{x^{3/2}}\;J_{n+1/2}(x)\\ &= -\sqrt{\frac{\pi}{2\,x}}\;\left(\,J_{n+3/2}(x) + \frac{n+\frac{1}{2}}{x}\;J_{n+1/2}(x)\,\right) - \frac{1}{2}\sqrt{\frac{\pi}{2}}\;\frac{1}{x^{3/2}}\;J_{n+1/2}(x)\\ &= -\sqrt{\frac{\pi}{2\,x}}\;J_{n+3/2}(x) + \frac{n}{x}\,\sqrt{\frac{\pi}{2\,x}}\;J_{n+1/2}(x) = -\,S_{n+1}(x) + \frac{n}{x}\,S_n(x).\end{aligned}$$

This completes the proof of the spherical Bessel recurrence formula (18.87).

With this, we conclude our brief introduction to the method of Frobenius and the theory of Bessel functions. The reader interested in further delving into either the general method, or the host of additional properties of Bessel functions is encouraged to consult the texts [**149**, **116**, **85**, **145**].

References

[**1**] Ablowitz, M.J., and Clarkson, P.A., *Solitons, Nonlinear Evolution Equations and the Inverse Scattering Transform*, L.M.S. Lecture Notes in Math., vol. 149, Cambridge University Press, Cambridge, 1991.

[**2**] Abraham, R., Marsden, J.E., and Ratiu, T., *Manifolds, Tensor Analysis, and Applications*, Springer–Verlag, New York, 1988.

[**3**] Abramowitz, M., and Stegun, I., *Handbook of Mathematical Functions*, National Bureau of Standards Appl. Math. Series, #55, U.S. Govt. Printing Office, Washington, D.C., 1970.

[**4**] Ahlfors, L., *Complex Analysis*, McGraw–Hill, New York, 1966.

[**5**] Airy, G.B., On the intensity of light in the neighborhood of a caustic, *Trans. Cambridge Phil. Soc.* **6** (1838), 379–402.

[**6**] Aki, K., and Richards, P.G., *Quantitative Seismology*, W.H. Freeman, San Francisco, 1980.

[**7**] Alligood, K.T., Sauer, T.D., and Yorke, J.A., *Chaos. An Introduction to Dynamical Systems*, Springer-Verlag, New York, 1997.

[**8**] Antman, S.S., *Nonlinear Problems of Elasticity*, Appl. Math. Sci., vol. 107, Springer–Verlag, New York, 1995.

[**9**] Apostol, T.M., *Calculus*, Blaisdell Publishing Co., Waltham, Mass., 1967–69.

[**10**] Apostol, T.M., *Linear Algebra*, John Wiley & Sons, Inc., New York, 1997.

[**11**] Tannenbaum, P., and Arnold, R., *Excursions in Modern Mathematics*, 5th ed., Prentice–Hall, Inc., Englewood Cliffs, N.J., 2004.

[**12**] Baker, G.A., Jr., and Graves–Morris, P., *Padé Approximants*, Encyclopedia of Mathematics and Its Applications, v. 59, Cambridge Univ. Press, Cambridge, 1996.

[**13**] Ball, J.M., and James, R.D., Fine phase mixtures as minimizers of energy, *Arch. Rat. Mech. Anal.* **100** (1987), 13–52.

[**14**] Ball, J.M., and Mizel, V.J., One-dimensional variational problem whose minimizers do not satisfy the Euler-Lagrange equation, *Arch. Rat. Mech. Anal.* **90** (1985), 325–388.

[**15**] Batchelor, G.K., *An Introduction to Fluid Dynamics*, Cambridge Univ. Press, Cambridge, 1967.

[**16**] Bateman, H., Some recent researches on the motion of fluids, *Monthly Weather Rev.* **43** (1915), 63–170.

[**17**] Berest, Y., and Winternitz, P., Huygens' principle and separation of variables, *Rev. Math. Phys.* **12** (2000), 159–180.

[**18**] Birkhoff, G., *Hydrodynamics — A Study in Logic, Fact and Similitude*, 1st ed., Princeton Univ. Press, Princeton, 1950.

[**19**] Birkhoff, G., and Rota, G.–C., *Ordinary Differential Equations*, Blaisdell Publ. Co., Waltham, Mass., 1962.

[**20**] Blanchard, P., Devaney, R.L., and Hall, G.R., *Differential Equations*, Brooks–Cole Publ. Co., Pacific Grove, Calif., 1998.

[**21**] Bott, R., and Tu, L.W., *Differential Forms in Algebraic Topology*, Springer–Verlag, New York, 1982.

[**22**] Boussinesq, J., Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond, *J. Math. Pures Appl.* **17** (2) (1872), 55–108.

[**23**] Boussinesq, J., Essai sur la théorie des eaux courants, *Mém. Acad. Sci. Inst. Nat. France* **23** (1) (1877), 1–680.

[**24**] Boyce, W.E., and DiPrima, R.C., *Elementary Differential Equations and Boundary Value Problems*, 7th ed., John Wiley & Sons, Inc., New York, 2001.

[**25**] Braun, M., *Differential Equations and their Applications: an Introduction to Applied Mathematics*, Springer–Verlag, New York, 1993.

[**26**] Brigham, E.O., *The Fast Fourier Transform*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1974.

[**27**] Briggs, W.L., *Henson, V.E.*, The DFT. An Owner's Manual for the Discrete Fourier Transform; SIAM, Philadelphia, PA, 1995.

[**28**] Brown, J.W., and Churchill, R.V., *Fourier Series and Boundary Value Problems*, McGraw-Hill, New York, 1993.

[**29**] Buhmann, M.D., Radial basis functions, *Acta Numer.* **9** (2000), 1–38.

[**30**] Burden, R.L., and Faires, J.D., *Numerical Analysis*, Seventh Edition, Brooks/Cole, Pacific Grove, CA, 2001.

[**31**] Bürgisser, P., Clausen, M., and Shokrollahi, M.A., *Algebraic Complexity Theory*, Springer–Verlag, New York, 1997.

[**32**] Cantwell, B.J., *Introduction to Symmetry Analysis*, Cambridge University Press, Cambridge, 2003.

[**33**] Chartrand, G., and Lesniak, L., *Graphs & Digraphs*, 3rd ed., Chapman & Hall, London, 1996.

[**34**] Coddington, E.A., and Levinson, N., *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[**35**] Cole, J.D., On a quasilinear parabolic equation occurring in aerodynamics, *Q. Appl. Math.* **9** (1951), 225–236.

[**36**] Cooley, J.W., and Tukey, J.W., An algorithm for the machine computation of complex Fourier series, *Math. Comp.* **19** (1965), 297–301.

[**37**] Copson, E.T., *Partial Differential Equations*, Cambridge University Press, Cambridge, 1975.

[**38**] Courant, R., *Differential and Integral Calculus*, 2nd ed., Interscience Publ., New York, 1937.

[**39**] Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, vol. I, Interscience Publ., New York, 1953.

[**40**] Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, vol. II, Interscience Publ., New York, 1953.

[**41**] Crowe, M.J., *A History of Vector Analysis*, Dover Publ., New York, 1985.

[**42**] Daubechies, I., Orthonormal bases of compactly supported wavelets, *Commun. Pure Appl. Math.* **41** (1988), 909–996.

[**43**] Daubechies, I., *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[**44**] Devaney, R.L., *An Introduction to Chaotic Dynamical Systems*, Addison–Wesley, Redwood City, Calif., 1989.

[**45**] Dewdney, A.K., *The Planiverse. Computer Contact with a Two-dimensional World*, Copernicus, New York, 2001.

[**46**] Diacu, F., *An Introduction to Differential Equations*, W.H. Freeman and Co., New York, 2000.

[**47**] Dirac, P.A.M., *The Principles of Quantum Mechanics*, Third Edition, Clarendon Press, Oxford, 1947.

[**48**] DLMF Project ■ .

[**49**] do Carmo, M.P., *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Englewood Cliffs, N.J., 1976.

[**50**] Drazin, P.G., and Johnson, R.S., *Solitons: An Introduction*, Cambridge University Press, Cambridge, 1989.

[**51**] Dym, H., and McKean, H.P., *Fourier Series and Integrals*, Academic Press, New York, 1972.

[**52**] Farin, G.E., *Curves and Surfaces for CAGD: A Practical Guide*, Academic Press, London, 2002.

[**53**] Feigenbaum, M.J., Qualitative universality for a class of nonlinear transformations, *J. Stat. Phys.* **19** (1978), 25–52.

[**54**] Fermi, E., Pasta, J., and Ulam, S., Studies of nonlinear problems. I., preprint, Los Alamos Report LA 1940, 1955; in: *Nonlinear Wave Motion*, A.C. Newell, ed., Lectures in Applied Math., vol. 15, American Math. Soc., Providence, R.I., 1974, pp. 143–156.

[**55**] Field, J.V., *The Invention of Infinity: Mathematics and Art in the Renaissance*, Oxford University Press, Oxford, 1997.

[**56**] Fletcher, N.H., and Rossing, T.D., *The Physics of Musical Instruments*, Second Edition, Springer–Verlag, New York, 1998.

[**57**] Fine, B., and Rosenberger, G., *The Fundamental Theorem of Algebra*, Undergraduate Texts in Mathematics, Springer–Verlag, New York, 1997.

[**58**] Fleming, W.H., *Functions of Several Variables*, 2d ed., Springer–Verlag, New York, 1977.

[**59**] Forsyth, A.R., *The Theory of Differential Equations*, Cambridge University Press, Cambridge, 1890, 1900, 1902, 1906.

[**60**] Fourier, J., *The Analytical Theory of Heat*, Dover Publ., New York, 1955.

[**61**] Francis, J.G.F., The QR transformation I, II, *Comput. J.* **4** (1961–2), 265–271, 332–345.

[**62**] Gaal, L., *Classical Galois theory*, 4th ed., Chelsea Publ. Co., New York, 1988.

[**63**] Garabedian, P., *Partial Differential Equations*, 2nd ed., Chelsea Publ. Co., New York, 1986.

[**64**] Gel'fand, I.M., and Fomin, S.V., *Calculus of Variations*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1963.

[**65**] Goldstein, H., *Classical Mechanics*, Second Edition, Addison–Wesley, Reading, Mass., 1980.

[**66**] Golub, G.H, and Van Loan, C.F., *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, 1989.

[**67**] Goode, S.W., *Differential Equations and Linear Algebra*, Second Ed., Prentice Hall, Upper Saddle River, NJ, 2000.

[**68**] Guillemin, V., and Pollack, A., *Differential Topology*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1974.

[**69**] Gurtin, M.E., *An Introduction to Continuum Mechanics*, Academic Press, New York, 1981.

[**70**] Haar, A., Zur Theorie der orthogonalen Funktionensysteme, *Math. Ann.* **69** (1910), 331–371.

[**71**] Haberman, R., *Elementary Applied Partial Differential Equations*, Third Edition, Prentice Hall, Upper Saddle River, NJ, 1998.

[**72**] Hale, J.K., *Ordinary Differential Equations*, Second Edition, R. E. Krieger Pub. Co., Huntington, N.Y., 1980.

[**73**] Hall, R.W., and Josić, K., Planetary motion and the duality of force laws, *SIAM Review* **42** (2000), 115–124.

[**74**] Hall, R.W., and Josić, K., The mathematics of musical instruments, *Amer. Math. Monthly* **108** (2001), 347–357.

[**75**] Hamming, R.W., *Numerical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1962.

[**76**] Henrici, P., *Applied and Computational Complex Analysis*, vol. 1, J. Wiley & Sons, New York, 1974.

[**77**] Herstein, I.N., *Abstract Algebra*, John Wiley & Sons, Inc., New York, 1999.

[**78**] Hestenes, M.R., and Stiefel, E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards* **49** (1952), 409–436.

[**79**] Hille, E., *Ordinary Differential Equations in the Complex Domain*, John Wiley & Sons, New York, 1976.

[**80**] Hirsch, M.W., and Smale, S., *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[**81**] Hobson, E.W., *The Theory of Spherical and Ellipsoidal Harmonics*, Chelsea Publ. Co., New York, 1965.

[**82**] Hoggatt, V.E., Jr., and Lind, D.A., The dying rabbit problem, *Fib. Quart.* **7** (1969), 482–487.

[**83**] Hopf, E., The partial differential equation $u_t + u\,u_x = \mu\,u$, *Commun. Pure Appl. Math.* **3** (1950), 201–230.

[**84**] Hydon, P.E., *Symmetry Methods for Differential Equations*, Cambridge Texts in Appl. Math., Cambridge University Press, Cambridge, 2000.

[**85**] Ince, E.L., *Ordinary Differential Equations*, Dover Publ., New York, 1956.

[**86**] Isaacson, E., and Keller, H.B.., *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[**87**] Jolliffe, I.T., *Principal Component Analysis*, 2nd ed., Springer–Verlag, New York, 2002.

[**88**] Kailath, T., Sayed, A.H., and Hassibi, B., *Linear Estimation*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2000.

[**89**] Kamke, E., *Differentialgleichungen Lösungsmethoden und Lösungen*, vol. 1, Chelsea, New York, 1971.

[**90**] Kammler, D.W., *A First Course in Fourier Analysis*, Prentice Hall, Upper Saddle River, NJ, 2000.

[**91**] Kaplansky, I., *An Introduction to Differential Algebra*, 2nd ed., Hermann, Paris, 1976.

[**92**] Kauffman, L.H., *Knots and Physics*, 2nd ed., World Scientific, Singapore, 1993.

[**93**] Keener, J.P., *Principles of Applied Mathematics. Transformation and Approximation*, Addison–Wesley Publ. Co., New York, 1988.

[**94**] Kemeny, J., and Snell, J., *Finite Markov Chains*, Springer–Verlag, New York, 1976.

[**95**] Kevorkian, J., *Partial Differential Equations*, Second Edition, Texts in Applied Mathematics, vol. 35, Springer–Verlag, New York, 2000.

[**96**] Korteweg, D.J., and de Vries, G., On the change of form of long waves advancing in a rectangular channel, and on a new type of long stationary waves, *Phil. Mag.* (5) **39** (1895), 422–443.

[**97**] Krall, A.M., *Applied Analysis*, D. Reidel Publishing Co., Boston, 1986.

[**98**] Kreysig, E., *Advanced Engineering Mathematics*, Eighth Edition, J. Wiley & Sons, New York, 1998.

[**99**] Kublanovskaya, V.N., On some algorithms for the solution of the complete eigenvalue problem, *USSR Comput. Math. Math. Phys.* **3** (1961), 637–657.

[**100**] Landau, L.D., and Lifshitz, E.M., *Quantum Mechanics* (*Non-relativistic Theory*), Course of Theoretical Physics, vol. 3, Pergamon Press, New York, 1977.

[**101**] Lanford, O., A computer-assisted proof of the Feigenbaum conjecture, *Bull. Amer. Math. Soc.* **6** (1982), 427–434.

[**102**] Mandelbrot, B.B., *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983.

[**103**] Marsden, J.E., and Tromba, A.J., *Vector Calculus*, 4th ed., W.H. Freeman, New York, 1996.

[**104**] Messiah, A., *Quantum Mechanics*, John Wiley & Sons, New York, 1976.

[**105**] Miller, W., Jr., *Symmetry and Separation of Variables*, Encyclopedia of Mathemtics and Its Applications, vol. 4, Addison–Wesley Publ. Co., Reading, Mass., 1977.

[**106**] Misner, C.W., Thorne, K.S., and Wheeler, J.A., *Gravitation*, W.H. Freeman, San Francisco, 1973.

[**107**] Moon, F.C., *Chaotic Vibrations*, John Wiley & Sons, New York, 1987.

[**108**] Moon, P., and Spencer, D.E., *Field Theory Handbook*, Springer-Verlag, New York, 1971.

[**109**] Morgan, F., *Geometric Measure Theory: a Beginner's Guide*, Academic Press, New York, 2000.

[**110**] Morse, P.M., and Feshbach, H., *Methods of Theoretical Physics*, McGraw–Hill, New York, 1953.

[**111**] Murray, R.N., Li, Z.X., and Sastry, S.S., *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1994.

[**112**] Nitsche, J.C.C., *Lectures on Minimal Surfaces*, Cambridge Univ. Press, Cambridge, 1988.

[**113**] Noether, E., Invariante Variationsprobleme, *Nachr. Konig. Gesell. Wissen. Gottingen, Math.–Phys. Kl.* (1918), 235–257. (See *Transport Theory and Stat. Phys.* **1** (1971), 186–207 for an English translation.)

[**114**] Oberhettinger, F., *Tables of Fourier Transforms and Fourier Transforms of Distributions*, Springer-Verlag, New York, 1990.

[**115**] Oberhettinger, F., and Badii, L., *Tables of Laplace Transforms*, Springer-Verlag, New York, 1973.

[**116**] Olver, F.W.J., *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[**117**] Olver, P.J., *Applications of Lie Groups to Differential Equations*, 2nd ed., Graduate Texts in Mathematics, vol. 107, Springer–Verlag, New York, 1993.

[**118**] O'Neil, P.V., *Advanced Engineering Mathematics*, Fourth Edition, Wadsworth Publ. Co., Belmont, Ca., 1995.

[**119**] Ortega, J.M., *Numerical Analysis; A Second Course*, Academic Press, New York, 1972.

[**120**] Peitgen, H.-O., and Richter, P.H., *The Beauty of Fractals: Images of Complex Dynamical Systems*,Springer–Verlag, New York, 1986.

[**121**] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in C*: *The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, 1995.

[**122**] Reed, M., and Simon, B., *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

[**123**] Renardy, M., and Rogers, R.C., *An Introduction to Partial Differential Equations*, Academic Press, New York, 1993.

[**124**] Richards, I., and Youn, H., *Theory of Distributions: a Non-Technical Introduction*, Cambridge University Press, Cambridge, 1990.

[**125**] Royden, H.L., *Real Analysis*, Macmillan Co., New York, 1988.

[**126**] Rudin, W., *Principles of Mathematical Analysis*, McGraw–Hill, New York, 1976.

[**127**] Saff, E.B., and Snider, A.D., *Fundamentals of Complex Analysis*, Third Ed., Prentice–Hall, Inc., Upper Saddle River, N.J., 2003.

[**128**] Sapiro, G., *Geometric Partial Differential Equations and Image Analysis*, Cambridge Univ. Press, Cambridge, 2001.

[**129**] Schumaker, L.L., *Spline Functions: Basic Theory*, John Wiley & Sons, New York, 1981.

[**130**] Schwartz, L., *Théorie des distributions*, Hermann, Paris, 1957.

[**131**] Scott Russell, J., On waves, *in*: *Report of the 14*$^{\text{th}}$ *Meeting*, British Assoc. Adv. Sci., 1845, pp. 311–390.

[**132**] Seshadri, R., and Na, T.Y., *Group Invariance in Engineering Boundary Value Problems*, Springer–Verlag, New York, 1985.

[**133**] Sethares, W.A., *Tuning, Timbre, Spectrum, Scale*, Springer–Verlag, New York, 1999.

[**134**] Strang, G., *Introduction to Applied Mathematics*, Wellesley Cambridge Press, Wellesley, Mass., 1986.

[**135**] Strang, G., *Linear Algebra and its Applications*, Third Ed., Harcourt, Brace, Jovanovich, San Diego, 1988.

[**136**] Strang, G., *Calculus*, Wellesley Cambridge Press, Wellesley, Mass., 1991.

[**137**] Strang, G., Wavelet transforms versus Fourier transforms, *Bull. Amer. Math. Soc.* **28** (1993), 288–305.

[**138**] Strang, G., and Borre, K., *Linear Algebra, Geodesy, and GPS*, Wellesley Cambridge Press, Wellesley, Mass., 1997.

[**139**] Strang, G., and Fix, G.J., *An Analysis of the Finite Element Method*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1973.

[**140**] Strang, G., and Nguyen, T., *Wavelets and Filter Banks*, Wellesley Cambridge Press, Wellesley, Mass., 1996.

[**141**] Titchmarsh, E. C., *Theory of Functions*, Oxford University Press, London, 1968.

[**142**] Tychonov, A.N., and Samarski, A.A., *Partial Differential Equations of Mathematical Physics*, Holden–Day, San Francisco, 1964.

[**143**] Ugural, A.C., and Fenster, S.K., *Advanced Strength and Applied Elasticity*, 3rd ed., Prentice–Hall, Inc., Englewood Cliffs, N.J., 1995.

[**144**] Varga, R.S., *Matrix Iterative Analysis*, 2nd ed., Springer–Verlag, New York, 2000.

[**145**] Watson, G.N., *A Treatise on the Theory of Bessel Functions*, Cambridge Univ. Press, Cambridge, 1952.

[**146**] Weinberger, H.F., *A First Course in Partial Differential Equations*, Ginn and Co., Waltham, Mass., 1965.

[**147**] Whitham, G.B., *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.

[**148**] Whittaker, E.T., *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Cambridge Univ. Press, Cambridge, 1937.

[**149**] Whittaker, E.T., and Watson, G.N., *A Course of Modern Analysis*, Cambridge Univ. Press, Cambridge, 1990.

[**150**] Widder, D.V., *The Heat Equation*, Academic Press, New York, 1975.

[**151**] Wolfram, S., *The Mathematica Book*, Third Edition, Cambridge University Press, Cambridge, 1996.

[**152**] Yaglom, I.M., *Felix Klein and Sophus Lie*, Birkhäuser, Boston, 1988.

[**153**] Yale, P.B., *Geometry and Symmetry*, Holden–Day, San Francisco, 1968.

[**154**] Young, D.M., *Iterative Solution of Laarge Linear Systems*, Academic Press, New York, 1971.

[**155**] Zabusky, N.J., and Kruskal, M.D., Interaction of "solitons" in a collisionless plasma and the recurrence of initial states, *Phys. Rev. Lett.* **15** (1965), 240–243.

[**156**] Zaitsev, V.F., and Polyanin, A.D., *Handbook of Exact Solutions for Ordinary Differential Equations*, CRC Press, Boca Raton, Fl., 1995.

[**157**] Zienkiewicz, O.C., and Taylor, R.L., *The Finite Element Method*, 4th ed., McGraw–Hill, New York, 1989.

[**158**] Zwillinger, D., *Handbook of Differential Equations*, Academic Press, Boston, 1992.

[**159**] Zygmund, A., *Trigonometric Series*, Cambridge University Press, Cambridge, 1968.

Appendix D

Canonical Forms and Matrix Exponentials

In this appendix, we survey the Jordan canonical form, that replaces diagonalization for incomplete matrices, and the matrix exponential, which provides and alternative, useful approach to the solution of systems of linear ordinary differential equations.

D.1. The Jordan Canonical Form.

As we have seen, particularly in Chapter 7, an intinsic mathematical quantity, e.g., a linear function, a quadratic form, etc., can have quite different explicit forms when written in different coordinate systems, or, equivalently, a choice of basis on the underlying vector space. Mathematicians use the term *canonical form* to refer to a particularly simple coordinate representative of the given object. For example, in Theorems ■ and ■, we saw that a quadratic form on $\mathbb{R}^n$ could always be diagonalized through a choice of coordinates, and the diagonal version consitutes a canonical form for a quadratic form. Another example is the very simple canonical form (a■) for a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^m$, in which we are allowed to independently choose bases on both the domain space $\mathbb{R}^n$ and the target space $\mathbb{R}^m$.

In applications to ordinary differential equations and discrete dynamical systems, a key issue is the choice of a canonical form for a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^n$ in terms of a given basis on $\mathbb{R}^n$. As we know, once a basis has been chosen, the linear transformation assumes the form of an $n \times n$ matrix. for example suppose that $A = \big(a_{ij} \big)$ is the $n \times n$ matrix representing L in terms of the standard basis $\mathbf{e}_1, \dots, \mathbf{e}_n$ of $\mathbb{R}^n$, meaning that $L[\mathbf{e}_i] = \displaystyle\sum_{j=1}^{n} a_{ij}\, \mathbf{e}_j$. If we choose a different basis, $\mathbf{v}_1, \dots, \mathbf{v}_n$, then the matrix form of the linear transformation is

$$B = S^{-1} A S, \qquad \text{where} \qquad S = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{pmatrix}^T \tag{D.1}$$

is the matrix whose columns are the chosen basis vectors. Thus, the problem is to find a basis, or, equivalently an nonsingular matrix S, such that the resulting matrix (D.1) is in a simple canonical form.

In most cases, one can diagonalize the matrix A. More specifically, if A is complete and so has a basis of eigenvectors, then choosing S to be the eigenvector basis matrix will result in a diagonal matrix B; see Theorem 8.20. We regard a diagonal matrix as "simple", and so, when applicable, as provinding a canonical form for the linear transformation or original matrix. Diagonalizability may require us to move to the complex domain, and regard our matrix as a linear transformation on $\mathbb{C}^n$. This avoids the question of finding

a real canonical form for a real linear transformation on $\mathbb{R}^n$, but this will be discussed below.

Unfortunately, not all matrices are diagonalizable, and so we must find an alternative canonical form for the incomplete cases. The Jordan canonical form, named after the nineteenth century French mathematician Camille Jordan, provides one such simplification, but it is, in fact, but one of a number of interesting and useful canonical forms for matrices and linear transformations. A more complete exposition can be found in [**Gantmacher**]. The starting point is the simplest non-diagonalizable matrices.

Definition D.1. A $n \times n$ matrix of the form[†]

$$J_{\lambda,n} = \begin{pmatrix} \lambda & 1 & & & & \\ & \lambda & 1 & & & \\ & & \lambda & 1 & & \\ & & & \ddots & \ddots & \\ & & & & \lambda & 1 \\ & & & & & \lambda \end{pmatrix}, \tag{D.2}$$

in which λ is a real or complex number is known as a *Jordan block*.

In particular, a 1×1 Jordan block is merely a scalar $J_{\lambda,1} = \lambda$. Theorem 8.12 tells us that every matrix has at least one eigenvector. Therefore, according to Exercise ■, the Jordan block matrices have the least possible number of eigenvectors.

Lemma D.2. *The Jordan block matrix* (D.2) *has a single eigenvalue, λ, and a single independent eigenvector, $\mathbf{e}_n$.*

Definition D.3. A *Jordan matrix* is a square matrix of block diagonal form

$$J = \operatorname{diag}(J_{\lambda_1,n_1}, J_{\lambda_2,n_2}, \dots, J_{\lambda_k,n_k}) = \begin{pmatrix} J_{\lambda_1,n_1} & \mathrm{O} & \mathrm{O} & \dots & \mathrm{O} & \mathrm{O} \\ \mathrm{O} & J_{\lambda_2,n_2} & \mathrm{O} & \dots & \mathrm{O} & \mathrm{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{O} & \mathrm{O} & \mathrm{O} & \dots & J_{\lambda_{k-1},n_{k-1}} & \mathrm{O} \\ \mathrm{O} & \mathrm{O} & \mathrm{O} & \dots & \mathrm{O} & J_{\lambda_k,n_k} \end{pmatrix}, \tag{D.3}$$

in which one or more Jordan blocks, not necessarily of the same size, lie along the diagonal, while all off-diagonal blocks are zero matrices of the appropriate sizes.

[†] All non-dsiplaed entries are zero.

For example, the 6×6 matrices

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

are all Jordan matrices; the first is a diagonal matrix, consisting of 6 distinct 1×1 Jordan blocks; the second has a 4×4 jordan block followed by a 2×2 block that happen to have the same diagonal entries; the last has three 2×2 Jordan blocks. As a simple corollary of Lemma D.2 coupled with the block structure, we obtain a complete classification of the eigenvectors and eigenvalues of a Jordan matrix.

Lemma D.4. *A Jordan matrix of the form* (D.3) *has eigenvalues* $k_1, \dots, k_\lambda$ *and* k *linearly independent eigenvectors, namely the standard basis vectors* $e_{j_1}, \dots, e_{j_k}$ *whose indices* j_κ *correspond to the index of the lower right entry of the* κ^{th} *Jordan block* $J_{\lambda_\kappa, n_\kappa}$.

Thus, in the preceding examples of Jordan matrices, the first has eigenvalues $1, 2, 3$ and linearly independent eigenvalues $\mathbf{e}_1, \dots, \mathbf{e}_6$; the second has only one eigenvalue, -1, but two linearly independent eigenvectors, $\mathbf{e}_4$ and $\mathbf{e}_6$. The last has eigenvalues $0, 1, 2$ and eigenvectors $\mathbf{e}_2, \mathbf{e}_4, \mathbf{e}_6$. In each case, the index of the eigenvector corresponds to the last entry of its associated Jordan block.

Definition D.5. A nonzero vector $\mathbf{w} \neq \mathbf{0}$ that satisfies

$$(A - \lambda \, \mathrm{I})^k \, \mathbf{w} = \mathbf{0} \tag{D.4}$$

for some $k > 0$ and $\lambda \in \mathbb{C}$ is called a *generalized eigenvector* of the matrix A.

Note that every ordinary eigenvector is automatically a generalized eigenvector, since we can just take $k = 1$, but the converse is not necessarily valid. We shall call the minimal value of k for which (D.4) holds the *index* of the generalized eigenvector. Thus, an ordinary eigenvector is a generalized eigenvector of index 0. Since $A - \lambda \, \mathrm{I}$ is nonsingular whenever λ is not an eigenvalue of A, its k^{th} power $(A - \lambda \, \mathrm{I})^k$ is also nonsingular. Therefore, generalized eigenvectors can only exist when λ is an ordinary eigenvalue of A — there are no "generalized eigenvalues".

Example D.6. Consider the 3×3 Jordan block $A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$. The only eigenvalue is $\lambda = 2$, and $A - 2 \, \mathrm{I} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$ is a nilpotent matrix. First, $\ker(A - 2 \, \mathrm{I})$ is spanned by $\mathbf{e}_3$, which, up to constant multiple, is the only eigenvector or generalized eigenvector of index 1. Secondly, the kernel of $(A - 2 \, \mathrm{I})^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ is spanned by $\mathbf{e}_2$

and $\mathbf{e}_3$, and so any vector of the form $\mathbf{w} = b\mathbf{e}_2 + c\mathbf{e}_3$ satisfies the generalized eigenvector condition $(A - 2\,\mathrm{I})^2\,\mathbf{w} = \mathbf{0}$; those with $b \neq 0$ have index 2, while those iwth $b = 0, c \neq 0$ have index 1. As with eigenvectors, we shall refer to $\mathbf{e}_2$ as "the" generalized eigenvector of index 2, which in reality means it and the ordinary eigenvector $\mathbf{e}_3$ span the space of generalized eigenvectors of index ≤ 2. Finally, $(A - 2\,\mathrm{I})^3 = \mathrm{O}$ is the zero matrix, and so every (nonzero) vector $\mathbf{w} = a\mathbf{e}_1 + b\mathbf{e}_2 + c\mathbf{e}_3$ is a generalized eigenvector; those with $a \neq 0$ have index 2. As before, we call $\mathbf{e}_1$ the index 2 generalized eigenvector.

Generalizing this example, it is easy to see that the $n \times n$ Jordan block (D.2) has a basis of generalized eigenvectors, namely $\mathbf{e}_1, \dots, \mathbf{e}_n$ where $\mathbf{e}_k$ is of index $n - k + 1$. More generally, any Jordan matrix also has the standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ forming a basis of generalized eigenvectors. The fundamental Jordan canonical form theorem states that, over the complex numbers, this holds in general, and thereby forms the generlization of the diagonalization Theorem 8.20.

Theorem D.7. *Let A be an $n \times n$ real or complex matrix. Then there is a basis $\mathbf{w}_1, \dots, \mathbf{w}_n$ of $\mathbb{C}^n$ consisting of generalized eigenvectors of A. Moreover, one can choose the basis so that the corresponding matrix $S = (\,\mathbf{w}_1 \;\; \mathbf{w}_2 \;\; \dots \;\; \mathbf{w}_n\,)^T$ places the matrix in* Jordan canonical form

$$S^{-1} A S = J = \operatorname{diag}(J_{\lambda_1,n_1}, J_{\lambda_2,n_2}, \dots, J_{\lambda_k,n_k}). \tag{D.5}$$

The diagonal entries of the resulting Jordan matrix J are the eigenvalues of A. The Jordan canonical form of A is uniquely determined, up to a permutation of the diagonal Jordan blocks.

The sizes of the Jordan blocks are prescribed by the *generalized eigenspaces*

$$W_{\lambda,k} = \ker(A - \lambda\,\mathrm{I})^k.$$

Note that for a fixed eigenvalue λ, the subspaces satisfy $\{\mathbf{0}\} \equiv W_{\lambda,0} \subsetneq W_{\lambda,1} \subseteq W_{\lambda,2} \subseteq \cdots$. Note that the elements of $W_{\lambda,k} \setminus W_{\lambda,k-1}$ are the generalized eigenvectors of index k. Let $m_{\lambda,k} = \dim W_{\lambda,k}$. It can be proved that

$$0 = m_{\lambda,0} < m_{\lambda,1} < \cdots < m_{\lambda,j-1} < m_{\lambda,j} = m_{\lambda,j+1} = m_{\lambda,j+2} = \cdots,$$

so that j is the maximal index for the generalized eigenvectors associated with λ. The Jordan blocks with λ on the diagonal in the Jordan canonical form of A are as follows, and there are exactly

$$b_{\lambda,k} = 2m_{\lambda,k} - m_{\lambda,k+1} - m_{\lambda,k-1} \tag{D.6}$$

Jordan blocks of size k that have λ along the diagonal. For example, a 11×11 Jordan matrix with just one eigenvalue λ and blocks of respective sizes $1, 3, 3, 4$ would have

$$\begin{aligned}
&W_{\lambda,1} = \{\mathbf{0}\}, && m_{\lambda,1} = 0,\\
&W_{\lambda,1} = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_5, \mathbf{e}_8\} && m_{\lambda,1}, = 4,\\
&W_{\lambda,2} = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_5, \mathbf{e}_6, \mathbf{e}_8, \mathbf{e}_9\}, && m_{\lambda,1} = 7,\\
&W_{\lambda,3} = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6, \mathbf{e}_7, \mathbf{e}_8, \mathbf{e}_9, \mathbf{e}_{10}\}, && m_{\lambda,3} = 7,\\
&W_{\lambda,k} = \operatorname{span}\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5, \mathbf{e}_6, \mathbf{e}_7, \mathbf{e}_8, \mathbf{e}_9, \mathbf{e}_{10}, \mathbf{e}_{11}\}, && m_{\lambda,k} = 11, \qquad k \geq 4,
\end{aligned}$$

and so we reconfirm the number of Jordan blocks of each size:

$$b_{\lambda,1} = 2 \cdot 4 - 0 - 7 = 1, \qquad b_{\lambda,3} = 2 \cdot 10 - 7 - 11 = 2,$$
$$b_{\lambda,2} = 2 \cdot 7 - 4 - 10 = 0, \qquad b_{\lambda,4} = 2 \cdot 11 - 10 - 11 = 1.$$

We refer the reader to [**X**] for a proof of Theorem D.7.

Example D.8. Consider the matrix ∎

D.2. Inhomogeneous Linear Systems.

We now direct our attention to general inhomogeneous linear systems of ordinary differential equations. For simplicity, we consider only first order† systems of the form

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u} + \mathbf{f}(t), \tag{D.7}$$

where A is a constant $n \times n$ matrix and $\mathbf{f}(t)$ is a vector of functions that represents external forcing to the system. According to our general Theorem 7.37, the solution to the inhomogeneous system will have the form

$$\mathbf{u}(t) = \mathbf{u}^\star(t) + \mathbf{z}(t)$$

where $\mathbf{u}^\star(t)$ is a particular solution and $\mathbf{z}(t)$ is a general solution to the homogeneous system (8.2). Physically, one interprets the solution as a combination of a particular response to the external forcing coupled with the system's own internal motion.

Since we already know how to find the solution $\mathbf{z}(t)$ to the homogeneous system, the only task is to determine one particular solution to the inhomogeneous system. The method used to construct the solution is known as *variation of parameters*, and will work even when the matrix A depends on t. The student may have encoutered the scalar version of this method in a first course on ordinary differential equations, and the same basic idea applies to systems. As we know, the general solution to the associated homogeneous system

$$\frac{d\mathbf{u}}{dt} = A\,\mathbf{u} \tag{D.8}$$

has the form

$$\mathbf{u}(t) = c_1\mathbf{u}_1(t) + \,\cdots\, + c_n\mathbf{u}_n(t), \tag{D.9}$$

where $\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)$ are n linearly independent solutions and $c_1, \dots, c_n$ are arbitrary constants. In the method of variation of parameters, one replaces the constants by scalar functions of t, and tries a the variable coefficient linear combination

$$\mathbf{u}^\star(t) = h_1(t)\mathbf{u}_1(t) + \,\cdots\, + h_n(t)\mathbf{u}_n(t) \tag{D.10}$$

for the particular solution. We then plug this ansatz into the system (D.7) and solve the resulting equations for the coefficient functions $h_i(t)$.

† Higher order systems can, as remarked earlier, always be converted into first order systems involving additional variables.

Before doing this, it will help to rewrite the variation of parameters formula (D.10) in a more convenient matrix form, based on our standard matrix formula (2.14) for linear combinations. To this end, we introduce the $n \times n$ matrix

$$U(t) = (\mathbf{u}_1(t), \dots, \mathbf{u}_n(t)) \tag{D.11}$$

whose columns are the independent solutions of the homogeneous system, along with the column vector

$$\mathbf{h}(t) = (\, h_1(t), \dots, h_n(t) \,)^T$$

representing the unknown coefficients. The variation of parameters formula (D.10) can then be written in matrix form

$$\mathbf{u}^\star(t) = U(t)\, \mathbf{h}(t). \tag{D.12}$$

The matrix (D.11) is known as a *fundamental matrix solution* to the homogeneous system. Indeed, it satisfies a matrix form of the system (D.8)

$$\frac{dU}{dt} = A\, U(t) \tag{D.13}$$

where we differentiate U entry by entry. This follows directly from the rules of matrix multiplication — specifically, formula (1.11). Moreover, since the solutions forming its columns are linearly independent, $U(t)$ is a nonsingular matrix for all t. Consequently, we can write the general solution (D.9) to the homogeneous system in the matrix form

$$\mathbf{u}(t) = U(t)\, \mathbf{c}, \qquad \text{where} \qquad \mathbf{c} = (\, c_1, c_2, \dots, c_n \,)^T \tag{D.14}$$

is a vector containing the arbitrary constants. To solve for initial conditions

$$\mathbf{u}(t_0) = \mathbf{b}, \tag{D.15}$$

we merely invert (or, better, use Gaussian elimination) to solve

$$U(t_0)\, \mathbf{c} = \mathbf{b}, \qquad \text{so that} \qquad \mathbf{c} = U(t_0)^{-1} \mathbf{b}.$$

Therefore, the solution to the homogeneous initial value problem (D.8), (D.15) is

$$\mathbf{u}(t) = U(t) U(t_0)^{-1} \mathbf{b}. \tag{D.16}$$

Example D.9. For the system

$$\frac{d\mathbf{u}}{dt} = A\, \mathbf{u} \qquad \text{where} \qquad A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix},$$

a fundamental matrix solution is obtained by assembling the two independent exponential solutions into a 2×2 matrix:

$$U(t) = \begin{pmatrix} e^{4t} & -e^{2t} \\ e^{4t} & e^{2t} \end{pmatrix}. \tag{D.17}$$

Note that

$$\frac{dU}{dt} = \begin{pmatrix} 4e^{4t} & -2e^{2t} \\ 4e^{4t} & 2e^{2t} \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} e^{4t} & -e^{2t} \\ e^{4t} & e^{2t} \end{pmatrix} = A\, U,$$

verifying (D.13) in this case. To solve the initial value problem (19.1), which had initial conditions $\mathbf{u}(0) = \mathbf{b} = (\,1, -2\,)^T$ we can use the formula (D.16)

$$\begin{aligned}\mathbf{u}(t) &= U(t)U(0)^{-1}\mathbf{b} = \begin{pmatrix} e^{4t} & -e^{2t} \\ e^{4t} & e^{2t} \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ -2 \end{pmatrix} \\ &= \begin{pmatrix} e^{4t} & -e^{2t} \\ e^{4t} & e^{2t} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2}\,e^{4t} + \frac{3}{2}\,e^{2t} \\ -\frac{1}{2}\,e^{4t} - \frac{3}{2}\,e^{2t} \end{pmatrix},\end{aligned}$$

which reproduces (19.5).

Now, let us return to the business at hand — the computation of a particular solution to the inhomogeneous system using the variation of parameters ansatz — in matrix form (D.12). We need to differentiate $\mathbf{u}^\star$, and this relies on the matrix version of the Leibniz rule for differentiating products.

Lemma D.10. *If $A(t)$ and $B(t)$ are differentiable matrix-valued functions, and their product $C(t) = A(t)\,B(t)$ is defined, then $C(t)$ is differentiable, and*

$$\frac{dC}{dt} = \frac{d}{dt}\,(AB) = \frac{dA}{dt}\;B + A\;\frac{dB}{dt}\,. \tag{D.18}$$

The only difference with the scalar Leibniz rule is the noncommutativity of the matrix version, which means that the factors in the two terms on the right hand side of (D.18) must stay in the correct order. The proof of Lemma D.10 is a simple consequence of the basic laws of matrix multiplication, and is left to the reader to fill in the details.

Thus, when we differentiate (D.12) we obtain

$$\frac{d\mathbf{u}}{dt} = \frac{dU}{dt}\;\mathbf{h} + U\,\frac{d\mathbf{h}}{dt}\,.$$

We now use the fact that U solves the matrix version of the system, and so

$$\frac{d\mathbf{u}}{dt} = A\,U\,\mathbf{h} + U\,\frac{d\mathbf{h}}{dt} = A\,\mathbf{u} + U\,\frac{d\mathbf{h}}{dt}\,.$$

comparing with (D.7), we see that our solution ansatz (D.12) will be a solution to the inhomogeneous system if and only if

$$U\,\frac{d\mathbf{h}}{dt} = \mathbf{f}(t) \qquad \text{or} \qquad \frac{d\mathbf{h}}{dt} = U(t)^{-1}\mathbf{f}(t). \tag{D.19}$$

Integrating equation (D.19), we have

$$\mathbf{h}(t) = \int_{t_0}^{t} U(s)^{-1}\,\mathbf{f}(s)\,ds, \tag{D.20}$$

where the integration is done entry by entry. Therefore, a particular solution to (D.7) is obtained by substituting (D.20) into (D.12), yielding

$$\mathbf{u}^\star(t) = U(t)\;\int_{t_0}^{t} U(s)^{-1}\,\mathbf{f}(s)\,ds = \int_{t_0}^{t} U(t)\,U(s)^{-1}\,\mathbf{f}(s)\,ds. \tag{D.21}$$

There are two alternative versions of the integrated matrix product:

$$\mathbf{u}^\star(t) = \int_{t_0}^{t} U(t-(s-t_0))\,U(t_0)^{-1}\,\mathbf{f}(s)\,ds = \int_0^{t-t_0} U(t-s)\,U(t_0)^{-1}\,\mathbf{f}(s)\,ds, \qquad (D.22)$$

which is a consequence of the formula

$$U(t)\,U(s)^{-1} = U(t-a)\,U(s-a)^{-1}, \qquad (D.23)$$

which is valid for any $t, s, a \in \mathbb{R}$. The proof relies on properties of the matrix exponential to be discussed below; see Exercise ■.

We have therefore established the following general formula for the solution to an first order, inhomogeneous linear system with constant coefficient matrix.

Theorem D.11. *Let $U(t)$ be a fundamental matrix solution to the homogeneous system* (D.8)*. Then the general solution to the inhomogeneous linear system* (D.7) *is given by*

$$\mathbf{u}(t) = U(t)\,\mathbf{c} + \mathbf{u}^\star(t) = U(t)\,\mathbf{c} + \int_0^{t-t_0} U(t-s)\,U(t_0)^{-1}\,\mathbf{f}(s)\,ds, \qquad (D.24)$$

where $\mathbf{c} \in \mathbb{R}^n$ is any constant vector, uniquely detemined by the initial conditions, and t_0 is any convenient limit for the integration. The unique solution to the initial value problem $\mathbf{u}(t_0) = \mathbf{b}$ is

$$\mathbf{u}(t) = U(t)U(t_0)^{-1}\,\mathbf{b} + \int_0^{t-t_0} U(t-s)\,U(t_0)^{-1}\,\mathbf{f}(s)\,ds. \qquad (D.25)$$

Example D.12. Consider the initial value problem

$$\begin{aligned} \dot u_1 &= 2u_1 - u_2, & u_1(0) &= 1, \\ \dot u_2 &= 4u_1 - 3u_2 + e^t & u_2(0) &= 0. \end{aligned} \qquad (D.26)$$

The eigenvalues and eigenvectors of the coefficient matrix $A = \begin{pmatrix} 2 & -1 \\ 4 & -3 \end{pmatrix}$ are

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad\qquad \lambda_2 = -2, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}.$$

We use these to form the fundamental matrix solution

$$U(t) = \begin{pmatrix} e^t & e^{-2t} \\ e^t & 4e^{-2t} \end{pmatrix}.$$

We can compute the solution directly from (D.25). First, note that

$$U(t)U(0)^{-1} = \begin{pmatrix} \frac{4}{3}e^t - \frac{1}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^t - \frac{4}{3}e^{-2t} & -\frac{1}{3}e^t + \frac{4}{3}e^{-2t} \end{pmatrix}. \qquad (D.27)$$

The integral in (D.25) is

$$\int_0^t U(t-s)\,U(0)^{-1}\,\mathbf{f}(s)\,ds = \int_0^t \begin{pmatrix} \frac{4}{3}e^{t-s} - \frac{1}{3}e^{-2(t-s)} & -\frac{1}{3}e^{t-s} + \frac{1}{3}e^{-2(t-s)} \\ \frac{4}{3}e^{t-s} - \frac{4}{3}e^{-2(t-s)} & -\frac{1}{3}e^{t-s} + \frac{4}{3}e^{-2(t-s)} \end{pmatrix} \begin{pmatrix} 0 \\ e^s \end{pmatrix} ds$$

$$= \begin{pmatrix} \displaystyle\int_0^t -\tfrac{1}{3}e^{t} + \tfrac{1}{3}e^{-2t+3s}\,ds \\ \displaystyle\int_0^t -\tfrac{1}{3}e^{t} + \tfrac{4}{3}e^{-2t+3s}\,ds \end{pmatrix} = \begin{pmatrix} -\frac{1}{3}te^t + \frac{1}{9}(e^t - 1) \\ -\frac{1}{3}te^t + \frac{4}{9}(e^t - 1) \end{pmatrix}.$$

This is the particular solution for the homogeneous initial conditions $\mathbf{u}(0) = \mathbf{0}$. To obtain the solution that satisfies the given initial conditions, we compute the first term in (D.25)

$$U(t)U(0)^{-1}\,\mathbf{b} = \begin{pmatrix} \frac{4}{3}e^{t} - \frac{1}{3}e^{-2t} & -\frac{1}{3}e^{t} + \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^{t} - \frac{4}{3}e^{-2t} & -\frac{1}{3}e^{t} + \frac{4}{3}e^{-2t} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{4}{3}e^{t} - \frac{1}{3}e^{-2t} \\ \frac{4}{3}e^{t} - \frac{4}{3}e^{-2t} \end{pmatrix},$$

which is the solution to the homogeneous system for the given nonzero initial conditions. We sum these two to finally obtain the solution to our initial value problem:

$$\mathbf{u}(t) = \begin{pmatrix} \frac{4}{3}e^{t} - \frac{1}{3}e^{-2t} - \frac{1}{3}te^t + \frac{1}{9}(e^t - 1) \\ \frac{4}{3}e^{t} - \frac{4}{3}e^{-2t} - \frac{1}{3}te^t + \frac{4}{9}(e^t - 1) \end{pmatrix}.$$

D.2.1. (*a*) $\dot{u} + a\,u = \sin t$, $u(0) = 1$. (*b*) Decompose the solution into a *transient* that eventually disappears, and a response to the forcing.

* D.2.2. In chemical processes, the *reaction-rate equations* model the rate at which the reagants are produced. For the following system of reaction-rate equations, solve the system for the rate of production of the reagants x_1, x_2, and x_3: $\dfrac{dx_1}{dt} = 7x_1 - x_2 - 65.5$, $\dfrac{dx_2}{dt} = 3x_2 - 2x_3 - 5.5$, $\dfrac{dx_3}{dt} = 0.2x_1 + 2x_2 - x_3 - 1.5$. What is the equilibrium?

