Lecture Notes in Engineering Mathematics

George Nakos

Johns Hopkins University

Engineering Programs for Professionals

Contents

Chapter 1

Linear Algebra

1.1 Matrices and Vectors

Matrices

A **matrix** is a rectangular arrangement of numbers called **entries**. A matrix has **rows** that are numbered top to bottom and **columns** that are numbered left to right. The (i, j) entry is the entry at the ith row and jth column.

A matrix has **size** $m \times n$ (pronounced 'm by n'), if it has m rows and n columns. If $m = n$, then the matrix is called **square**. In this case, n is the **size** of the square matrix.

For our purposes, the entries of a matrix are usually real numbers. Sometimes we use complex numbers and, occasionally, mathematical functions.

Example 1.1.1. The following are matrices of respective sizes 4×2, 2×3, 3×3, 5×1, 1×2, and 2×2.

$$\begin{bmatrix} 1 & -2 \\ -3 & 5 \\ 0 & 6 \\ 2 & -8 \end{bmatrix}, \quad \begin{bmatrix} 7 & 21 & 1/2 \\ -9 & \sqrt{5} & 4+i \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{23} & a_{33} \end{bmatrix},$$

$$\begin{bmatrix} 7.1 \\ 3.2 \\ -1.5 \\ 4.9 \\ 6.9 \end{bmatrix}, \quad \begin{bmatrix} a & b \end{bmatrix}, \quad \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

The $(3,2)$ entry of the first matrix is 6. The third and sixth matrices are square of respective sizes 3 and 2.

A general matrix A of size $m \times n$ with (i,j) entry a_{ij} is denoted by

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

This is abbreviated by

$$A = [a_{ij}]$$

where i and j are indices such that $1 \leq i \leq m$ and $1 \leq j \leq n$.

Notation The set of all $m \times n$ matrices with real entries is denoted by $M_{m\times n}$.

Vectors

If $n = 1$, then A is called a **column matrix**, or a m-**vector**, or a **vector**. If $m = 1$, then A is called a **row matrix**, or a n-**row vector**, or a **row vector**. The entries of vectors are usually called **components**.

Example 1.1.2. The following are vectors. The first is a 2-vector, the second is a 4-vector, and the third is a n-vector.

$$\begin{bmatrix} 7 \\ -3 \end{bmatrix}, \qquad \begin{bmatrix} 4 \\ -3 \\ 2 \\ -1 \end{bmatrix}, \qquad \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

Here are some row vectors.

$$[-3], \quad \begin{bmatrix} 1.2 & \sqrt{3} \end{bmatrix}, \qquad \begin{bmatrix} a & b & c \end{bmatrix}, \qquad \begin{bmatrix} -2 & 3 & 0 & 4 & 1 \end{bmatrix}$$

Notation The set of all n-vectors with real entries is denoted by $\mathbf{R}^n$.

Zero Matrices; Equal Matrices

A **zero** matrix, denoted by $\mathbf{0}$, is a matrix with zero entries. Here are some examples.

$$\mathbf{0} = [0], \quad \mathbf{0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{0} = \begin{bmatrix} 0 & 0 \end{bmatrix}, \quad \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We say that two matrices A and B are **equal** and we write $A = B$, if A and B have the same size and their corresponding entries are equal. So, if

$$A = \begin{bmatrix} 1 & 2 \\ a & b \end{bmatrix}, \quad B = \begin{bmatrix} c & d \\ 3 & 4 \end{bmatrix}$$

then $A = B$, only if $a = 3$, $b = 4$, $c = 1$, and $d = 2$.

Matrix Addition

We can **add** two matrices of the same size by adding the corresponding entries. The resulting matrix is the **sum** of the two matrices.

Example 1.1.3. We have

$$\begin{bmatrix} 1 & -3 & 0 \\ 2 & -4 & 7 \end{bmatrix} + \begin{bmatrix} 0 & 4 & 5 \\ -1 & 4 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 5 \\ 1 & 0 & 5 \end{bmatrix}$$

In general, if $A = [a_{ij}]$ and $B = [b_{ij}]$, for $1 \leq i \leq m$ and $1 \leq j \leq n$, then

$$A + B = [a_{ij} + b_{ij}]$$

Scalar Multiplication

We also multiply any real number c, times a matrix A, by multiplying all entries of A by c.

Example 1.1.4. We have

$$2 \begin{bmatrix} 1 & 0 \\ -3 & 4 \\ 5 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ -6 & 8 \\ 10 & -2 \end{bmatrix}$$

In general, if $A = [a_{ij}]$, then

$$cA = [ca_{ij}]$$

This operation is called **scalar multiplication**. The multiplier c is often called a **scalar**, because it *scales* A.

Matrices: Opposite, Difference

The matrix $(-1)A$ is called the **opposite** of A and it is denoted by $-A$.

Example 1.1.5. We have

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

The matrix $A+(-1)B$ is denoted by $A-B$ and it is called the **difference** between A and B. This is the **subtraction** operation.

$$A - B = A + (-1)B$$

Example 1.1.6. We have

$$\begin{bmatrix} 1 & -2 \\ 7 & 4 \\ 5 & -5 \\ 8 & 0 \end{bmatrix} - \begin{bmatrix} 1 & -1 \\ 6 & 3 \\ 7 & 0 \\ -3 & 7 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \\ -2 & -5 \\ 11 & -7 \end{bmatrix}$$

Properties of Operations

Theorem 1.1.1. *For any matrices A, B, and C of the same size and for any scalars a, b, and c, we have the following.*

1. $(A + B) + C = A + (B + C)$ ***(Associative Law)***
2. $A + B = B + A$ ***(Commutative Law)***
3. $A + \mathbf{0} = \mathbf{0} + A = A$
4. $A + (-A) = (-A) + A = \mathbf{0}$
5. $c(A + B) = cA + cB$ ***(Distributive Law)***

6. $(a+b)C = aC + bC$ *(**Distributive Law**)*

7. $(ab)C = a(bC) = b(aC)$

8. $1A = A$

9. $0A = \mathbf{0}$

1.2 Matrix Transpose; Special Matrices

Matrix Transpose

Let A be any $m \times n$ matrix. The **transpose** of A, denoted by A^T, is the $n \times m$ matrix obtained from A by switching all columns of A to rows and maintaining the same order.

Example 1.2.1. We have

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}, \quad \begin{bmatrix} a & b & c & d \end{bmatrix}^T = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \quad \begin{bmatrix} \frac{1}{3} \\ -8 \end{bmatrix}^T = \begin{bmatrix} \frac{1}{3} & -8 \end{bmatrix}$$

Properties of Transposition

Theorem 1.2.1. *Let A and B be $m \times n$ matrices and let c be any scalar. Then*

1. $(A+B)^T = A^T + B^T$

2. $(cA)^T = cA^T$

3. $\left(A^T\right)^T = A$

Symmetric and skew-Symmetric Matrices

A matrix A such that $A^T = A$ is called **symmetric**.

Example 1.2.2. The matrices are symmetric.

$$\begin{bmatrix} 5 & -7 \\ -7 & 6 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 & 3 \\ -1 & 4 & 9 \\ 3 & 9 & 6 \end{bmatrix}, \quad \begin{bmatrix} a & b & c & d \\ b & e & f & g \\ c & f & h & i \\ d & g & i & j \end{bmatrix}$$

Note the mirror symmetry of a symmetric matrix with respect to the upper-left to lower-right diagonal line, called the **the main diagonal**.

A matrix A such that $A^T = -A$ is called **skew-symmetric**.

Example 1.2.3. The matrices are skew-symmetric.

$$\begin{bmatrix} 0 & 7 \\ -7 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & -1 & 3 \\ 1 & 0 & -9 \\ -3 & 9 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & -b & -c & -d \\ b & 0 & -f & -g \\ c & f & 0 & -i \\ d & g & i & 0 \end{bmatrix}$$

Notice that a skew-symmetric matrix has zeros on the main diagonal. Also, it has an opposite mirror symmetry with respect to main diagonal.

Note Every square matrix A can be written as the sum of a symmetric matrix B, plus a skew-symmetric C, as follows

$$A = B + C, \quad \text{where} \quad B = \frac{1}{2}(A + A^T) \text{ and } C = \frac{1}{2}(A - A^T).$$

The Standard Basis Matrices $E_{i,j}$ and Vectors $\mathbf{e}_i$

We denote by $E_{i,j}$ in $M_{m\times n}$ the $m \times n$ matrix whose (i, j) entry is 1 and the remaining entries are zero. The matrices $E_{i,j}$ are called the **standard basis matrices** of $M_{m\times n}$. In $M_{2\times 2}$, we have the standard basis matrices

$$E_{1,1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \; E_{1,2} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \; E_{2,1} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \; E_{2,2} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

We also denote by $\mathbf{e}_i$ in $\mathbf{R}^n$ the n-vector whose ith entry is 1 and the remaining entries are zero. The vectors $\mathbf{e}_i$ are called the **standard basis vectors** of $\mathbf{R}^n$. In $\mathbf{R}^4$, we have the standard basis vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \; \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \; \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \; \mathbf{e}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Special Square Matrices

Let A be a square matrix of size n. Recall that the entries a_{ii}, $1 \leq i \leq n$ form the **main diagonal**. We have the following definitions.

- A is **upper triangular**, if all entries below the main diagonal are zero, i.e., if $a_{ij} = 0$ for $j < i$.
- A is **lower triangular**, if the entries above the main diagonal are all zero, so $a_{ij} = 0$ for $i < j$.
- If the main diagonal is also zero, we talk about **strictly upper triangular** and **strictly lower triangular** matrices.
- A is **diagonal**, if all its nondiagonal entries are zero.
- A is **scalar**, if all it is diagonal and all diagonal entries are equal.

Example 1.2.4. For the matrices below we have

$$A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}, \; B = \begin{bmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{bmatrix}, \; C = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix},$$

$$D = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, \; E = \begin{bmatrix} 7 & 0 \\ 0 & 7 \end{bmatrix}$$

- A, D, E are upper triangular.
- B, C, D, E are lower triangular.
- C is strictly lower triangular.
- Matrices D and E are diagonal.
- Matrix E is a scalar matrix.

The Identity Matrix

A scalar matrix of size n with common diagonal entry 1 is called an **identity matrix** and it is denoted by I_n, or by I.

$$I = I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \ldots, \quad I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

1.3 Matrix Multiplication

Let A be a $m \times k$ matrix and B be a $k \times n$ matrix. The **product** AB is the $m \times n$ matrix $C = [c_{ij}] = AB$, with entries c_{ij} are given by

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ \boxed{\begin{matrix} a_{i1} & a_{i2} & \cdots & a_{ik} \end{matrix}} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{bmatrix} \qquad B = \begin{bmatrix} b_{11} & \cdots & \boxed{\begin{matrix} b_{1j} \\ b_{2j} \\ \vdots \\ b_{kj} \end{matrix}} & \cdots & b_{1n} \\ b_{21} & \cdots & & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ b_{k1} & \cdots & & \cdots & b_{kn} \end{bmatrix}$$

$$c_{ij} = a_{i1}\, b_{1j} \; + \; a_{i2}\, b_{2j} \; + \cdots + a_{ik}\, b_{kj} = \sum_{r=1}^{k} a_{ir} b_{rj}$$

Example 1.3.1. We have

$$\begin{bmatrix} 2 & 0 & 1 \\ 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 2 & 4 \\ -2 & 4 & 5 \\ 0 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 6 & 7 & 6 \\ 4 & 14 & 9 \end{bmatrix}$$

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

$$\begin{bmatrix} 1 \\ -2 \\ 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 & -1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & -1 & -2 & 1 \\ -8 & 2 & 4 & -2 \\ 12 & -3 & -6 & 3 \\ 20 & -5 & -10 & 5 \end{bmatrix}$$

Properties of Matrix Multiplication

Theorem 1.3.1. *For any matrices A, B, and C of compatible sizes and for any scalar a, we have*

1. $(AB)C = A(BC)$ *(**Associative law**)*
2. $A(B+C) = AB + AC$ *(**Left Distributive law**)*
3. $(B+C)A = BA + CA$ *(**Right Distributive law**)*
4. $a(BC) = (aB)C = B(aC)$
5. $I_m A = AI_n = A$ *(**Multiplicative identity**)*
6. $\mathbf{0}A = \mathbf{0}$ *and* $A\mathbf{0} = \mathbf{0}$
7. $(AB)^T = B^T A^T$

Caution with Matrix Multiplication

AB may not equal BA. In fact, if AB is defined, then BA may not be defined. If BA is defined, then it may not have the same size as AB. If it does have the same size, it may still not equal AB.

1. We say that matrix multiplication is **noncommutative**.
2. If two matrices A and B satisfy $AB = BA$, then we say that they **commute**.

Example 1.3.2. $A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix}$ commute.

Powers of Square Matrix

Let A be a square matrix. The product AA is also denoted by A^2. Likewise, $AAA = A^3$ and $AA\cdots A = A^n$ for n factors of A. In addition, we write $A^1 = A$ and if A is nonzero, we write $A^0 = I$.

$$A^n = \underbrace{AA\cdots A}_{n \text{ factors}}, \qquad A^1 = A, \qquad A^0 = I$$

Example 1.3.3. We have

$$A^1 = \begin{bmatrix} 1 & -1 \\ -2 & 3 \end{bmatrix}, \quad A^2 = \begin{bmatrix} 3 & -4 \\ -8 & 11 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 11 & -15 \\ -30 & 41 \end{bmatrix}, \cdots$$

$$B^1 = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}, \quad B^2 = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}, \quad B^3 = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}, \quad \cdots$$

$$C^1 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad C^2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad C^3 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \cdots$$

Motivation for Matrix Multiplication

Let $\mathbf{x} = (x_1, x_2)$, $\mathbf{y} = (y_1, y_2)$, and $\mathbf{z} = (z_1, z_2)$ be coordinate frames. Suppose we go from frame $\mathbf{y}$ to frame $\mathbf{z}$ by using the linear transformation

$$\begin{aligned} z_1 &= a_{11}y_1 + a_{12}y_2 \\ z_2 &= a_{21}y_1 + a_{22}y_2 \end{aligned}$$

and from frame $\mathbf{x}$ to frame $\mathbf{y}$ by the linear transformation

$$\begin{aligned} y_1 &= b_{11}x_1 + b_{12}x_2 \\ y_2 &= b_{21}x_1 + b_{22}x_2 \end{aligned}$$

If we want to go from frame $\mathbf{x}$ to frame $\mathbf{z}$, we substitute

$$\begin{aligned} z_1 &= a_{11}(b_{11}x_1 + b_{12}x_2) + a_{12}(b_{21}x_1 + b_{22}x_2) \\ z_2 &= a_{21}(b_{11}x_1 + b_{12}x_2) + a_{22}(b_{21}x_1 + b_{22}x_2) \end{aligned}$$

and rearrange to get

$$\begin{aligned} z_1 &= (a_{11}b_{11} + a_{12}b_{21})x_1 + (a_{11}b_{12} + a_{12}b_{22})x_2 \\ z_2 &= (a_{21}b_{11} + a_{22}b_{21})x_1 + (a_{21}b_{12} + a_{22}b_{22})x_2 \end{aligned}$$

Now if A and B are coefficient matrices of the first two transformations and C is the coefficient matrix of the last one, then we see that $C = AB$.

$$\begin{aligned} C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} &= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = AB \end{aligned}$$

1.4 Linear Systems and Gauss Elimination

Linear Equations

Definition A **linear equation** in n **unknowns** $x_1, \ldots, x_n$, is an equation that can be written in the form

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b \tag{1.1}$$

where the a_i and b are given numbers. The a_i are called the **coefficients** of the equation and b is called the **constant term**. The unknowns are also called **variables**, or **indeterminants**. If $b = 0$, then the equation is called **homogeneous**.

The first variable with nonzero coefficient of a linear equation is called the **leading variable**. The remaining variables are called **free variables**. An equation which is not linear is called **nonlinear**.

Example 1.4.1. (a) The equation

$$x_1 + x_2 + 4x_3 - 6x_4 - 1 = x_1 - x_2 + 2$$

is linear, because it can be written in the form (1.1) as

$$0x_1 + 2x_2 + 4x_3 - 6x_4 = 3$$

The leading variable is x_2. The free variables are x_1, x_3, and x_4.

(b) The next three equations are also linear.

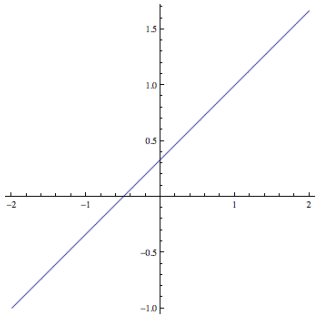
$$x_1 + 2x_2 - \sqrt{5}x_3 - x_4 = 0, \qquad x - 4y + 9z = \tan 4, \qquad F = \frac{9}{5}C + 32$$

(c) The following equations are nonlinear, due to $x_1^2, 1/x_2$, and $\sin x_1$.

$$x_1^2 - x_2 = 7, \qquad \frac{x_1}{x_2} - 3x_3 = 2, \qquad \sin x_1 + x_2 = 0$$

A **(particular) solution** of equation (1.1) is an n-tuple of numbers $(r_1, r_2, \ldots, r_n)$ such that the equation is satisfied when we substitute $x_1 = r_1, \ldots, x_n = r_n$. The set of all possible solutions is the **solution set**.

For example, $(0, \frac{1}{3})$ is a particular solution of $2x_1 - 3x_2 = -1$. The solution set of this equation can written as $\{(\frac{3}{2}t - \frac{1}{2}, t),\ t \text{ any real}\}$. Geometrically, this represents the straight line in the plane shown here.



Solving Linear Equations

Note In general, the solution set of the linear equation $a_1x_1 + a_2x_2 = b$ represents a **straight line** in the plane and the solution set of the linear equation $a_1x_1 + a_2x_2 + a_3x_3 = b$ represents a **plane** in space.

To find all solutions of a linear equation, we just solve for the leading variable in terms of the free variables and let let the free variables take on any values. For clarity, we usually rename the free variables. The new names are called **parameters**.

Example 1.4.2. The solutions of $x_2 - x_3 + x_4 = 2$ in the variables $x_1, \ldots, x_4$ are

$$x_1 = r,\ x_2 = 2 + s - t,\ x_3 = s,\ x_4 = t$$

where the parameters r, s, t are any real numbers.

Linear Systems; Gauss Elimination

A **linear system** of m equations in n **unknowns** $x_1, \ldots, x_n$, is a set of m linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{1.2}$$

The unknowns are also called **variables**, or **indeterminants**. The numbers a_{ij} are the **coefficients** and the numbers b_i are the **constant terms**. If all constant terms are zero, then the system is called **homogeneous**. The

homogeneous system that has the same coefficients as system (1.2) is said to be **associated** with (1.2). If $m = n$, then the system is called **square**.

Examples of Linear Systems

Example 1.4.3. The system

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & & & = & -3 \\ 2x_1 & + & 3x_2 & - & 2x_3 & = & -10 \\ -x_1 & & & + & 6x_3 & = & 9 \end{array} \tag{1.3}$$

is linear square with coefficients $1, 2, 0, 2, 3, -2, -1, 0, 6$, constant terms $-3, -10, 9$, and associated homogeneous system

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & & & = & 0 \\ 2x_1 & + & 3x_2 & - & 2x_3 & = & 0 \\ -x_1 & & & + & 6x_3 & = & 0 \end{array}$$

Example 1.4.4. The following three systems are linear. The first is from an ancient Chinese text.[1]

$$\begin{array}{l} 3x + 2y + z = 39 \\ 2x + 3y + z = 34 \\ x + 2y + 3z = 26 \end{array} \qquad \begin{array}{l} x_1 + x_2 = 5 \\ x_1 - 2x_2 = 6 \\ -3x_1 + x_2 = 1 \end{array} \qquad \begin{array}{l} y_1 + y_2 + y_3 = -2 \\ y_1 - 2y_2 + 7y_3 = 6 \end{array}$$

The Solution Set of a Linear System

For a linear system we have the following definitions.

- A **(particular) solution** of system (1.2) (p. 24) is an n-tuple $(r_1, r_2, \ldots, r_n)$ of numbers that is a common solution to each linear equation of (1.2).
- The set of all possible solutions is the **solution set**.
- Any generic element of the solution set is called the **general solution**.
- If a system has solutions, it is called **consistent**, otherwise it is called **inconsistent**.

[1] A third century BC book titled *Nine Chapters of Mathematical Art*. See Carl Boyer's *A History of Mathematics* (New York: Wiley).

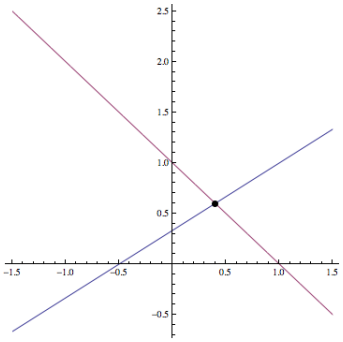
- Two linear systems with the same solution sets are called **equivalent**.
- A solution that consists only of zeros is called a **trivial solution**.

Note The trivial solution is always a solution of a homogeneous system. A homogeneous system may also have nontrivial solutions.

Example 1.4.5. It is easy to see that $\left(\frac{2}{5}, \frac{3}{5}\right)$ is the only solution of the system

$$\begin{aligned} 2x_1 - 3x_2 &= -1 \\ x_1 + x_2 &= 1 \end{aligned}$$

Geometrically, the solution is the intersection of the two lines defined by the equations of the system.



Back-Substitution

The easiest systems to solve are those in triangular form. A system is in **echelon form**, or in **triangular form**, if the leading variable in each equation occurs to the right of the leading variable of the equation above it.

To solve such systems we first solve for the leading variable of the last equation, then substitute the value found into the equation above it, and repeat. This method is called **back-substitution**.

Example 1.4.6. Solve the linear system by back-substitution.

$$\begin{array}{rcrcrcr} x_1 & + & 5x_2 & + & x_3 & = & -4 \\ & - & 2x_2 & + & 4x_3 & = & 14 \\ & & & & 3x_3 & = & 9 \end{array}$$

Solution: Going from the bottom up, the last equation yields $x_3 = 3$, the second $x_2 = -1$ and the first $x_1 = -2$. Hence, the only solution is $x_1 = -2, x_2 = -1, x_3 = 3$.

In this example all variables were leading variables. This need not be always the case. The next example has free variables. In such a case the leading variables are computed in terms of the free variables. The free variables are renamed as the **parameters** and they can take on any values.

Example 1.4.7. Solve the system.

$$\begin{array}{rcrcrcrcrcl} x_1 & - & x_2 & + & x_3 & - & x_4 & + & 2x_5 & - & x_6 & = & 1 \\ & & & - & x_3 & & & + & x_5 & & & = & 1 \\ & & & & & & & - & x_5 & + & x_6 & = & 3 \end{array} \tag{1.4}$$

Solution: We solve for the leading variables x_5, x_3, x_1 in terms of the free variables x_6, x_4, x_2 which can take on any parameter values, say $x_6 = r, x_4 = s, x_2 = t$. By back-substitution we get the general solution

$$\begin{array}{l} x_1 = -2r + s + t + 11 \\ x_2 = t \\ x_3 = r - 4 \\ x_4 = s \\ x_5 = r - 3 \\ x_6 = r \end{array} \qquad \text{for all } r, s, t \in \mathbf{R} \tag{1.5}$$

Augmented and Coefficient Matrix

The matrix that consists of the coefficients and constant terms, is called the **augmented matrix** of the system. The augmented matrix of system (1.3) is

$$\begin{bmatrix} 1 & 2 & 0 & -3 \\ 2 & 3 & -2 & -10 \\ -1 & 0 & 6 & 9 \end{bmatrix}$$

The matrix with entries the coefficients is the **coefficient matrix** of the system. The vector of all constant terms is the **vector of constants**. The coefficient matrix and the vector of constants of system (1.3) are

$$\begin{bmatrix} 1 & 2 & 0 \\ 2 & 3 & -2 \\ -1 & 0 & 6 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} -3 \\ -10 \\ 9 \end{bmatrix}$$

Matrix Form of Linear System

Though a linear system is a set of linear equations, it can also be viewed as a single vector equation, by using the matrix-vector product. System (1.2) (p. 24) can take the equivalent expression as equality of between two vectors as follows:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

This can be further abbreviated as

$$A\mathbf{x} = \mathbf{b} \tag{1.6}$$

where A is the coefficient matrix, $\mathbf{x}$ is the vector of the unknowns, and $\mathbf{b}$ is the vector of constants.

Example 1.4.8. Write the linear system in matrix-vector product form.

$$\begin{aligned} 7x_1 + 4x_2 + 5x_3 &= 1 \\ 2x_1 - 3x_2 + 9x_3 &= -8 \end{aligned}$$

Solution: We have

$$\begin{bmatrix} 7 & 4 & 5 \\ 2 & -3 & 9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ -8 \end{bmatrix}$$

Elementary Row Operations

One basic way to solve a linear system is by **elimination** or **Gauss elimination**. This method eliminates unknowns from equations so that an equivalent echelon form system is obtained. The latter is solved by back-substitution. To get the original system to echelon form we perform a sequence of *elementary equation operations*: (a) add to an equation a multiple of another, (b) multiply an equation by a nonzero scalar, and (c) switch two equations. These operations do not change the solutions of the original system.

For economy, these operations are performed on the augmented matrix.

The **elementary row operations** of any matrix are:

Elimination: *add a constant multiple of one row to another:* $\boxed{R_i + cR_j \to R_i}$

Scaling: *multiply a row by a nonzero constant:* $\boxed{cR_i \to R_i}$

Interchange: *interchange two rows:* $\boxed{R_i \leftrightarrow R_j}$

Example of Gauss Elimination

Example 1.4.9. Solve the system by Gauss elimination.

$$\begin{aligned} x_1 + 2x_2 &= -3 \\ 2x_1 + 3x_2 - 2x_3 &= -10 \\ -x_1 + 6x_3 &= 9 \end{aligned}$$

Solution: We have

$$\begin{bmatrix} 1 & 2 & 0 & -3 \\ 2 & 3 & -2 & -10 \\ -1 & 0 & 6 & 9 \end{bmatrix} \boxed{\begin{matrix} R_2 - 2R_1 \to R_2 \\ R_3 + R_1 \to R_3 \end{matrix}} \begin{bmatrix} 1 & 2 & 0 & -3 \\ 0 & -1 & -2 & -4 \\ 0 & 2 & 6 & 6 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 0 & -3 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & 2 & -2 \end{bmatrix} \boxed{R_3 + 2R_2 \to R_3} \begin{bmatrix} 1 & 2 & 0 & -3 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & 2 & -2 \end{bmatrix}$$

The system is in triangular form. Start at the bottom and work upwards to eliminate unknowns *above* the leading variables (first variables with nonzero coefficients) of each equation (back-substitution).

$$\begin{bmatrix} 1 & 2 & 0 & -3 \\ 0 & -1 & -2 & -4 \\ 0 & 0 & 2 & -2 \end{bmatrix} \boxed{R_2 + R_3 \to R_2} \begin{bmatrix} 1 & 2 & 0 & -3 \\ 0 & -1 & 0 & -6 \\ 0 & 0 & 2 & -2 \end{bmatrix} \boxed{R_1 + 2R_2 \to R_1}$$

$$\begin{bmatrix} 1 & 0 & 0 & -15 \\ 0 & -1 & 0 & -6 \\ 0 & 0 & 2 & -2 \end{bmatrix} \boxed{\begin{matrix} (-1)\, R_2 \to R_2 \\ (1/2) R_3 \to R_3 \end{matrix}} \begin{bmatrix} 1 & 0 & 0 & -15 \\ 0 & 1 & 0 & 6 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

$$x_1 = -15, \quad x_2 = 6, \quad x_3 = -1$$

Infinitely Many Solutions

Example 1.4.10 (Infinitely Many Solutions)**.** Find the intersection of the three planes.

$$\begin{aligned} x + 2y - z &= 4 \\ 2x + 5y + 2z &= 9 \\ x + 4y + 7z &= 6 \end{aligned}$$

Solution: By elimination the augmented matrix of the system reduces to

$$\begin{bmatrix} 1 & 0 & -9 & 2 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

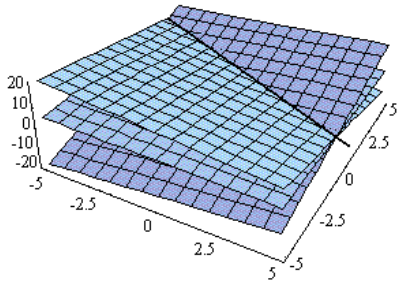
We get $x - 9z = 2$, $y + 4z = 1$. Rewriting this as $x = 9z + 2$, $y = -4z + 1$, and using parameters we get the following infinitely many solutions which are conveniently written in the form

$$\begin{aligned} x &= 9r + 2 \\ y &= -4r + 1 \quad , \quad r \in \mathbf{R} \\ z &= r \end{aligned}$$

It is illuminating to write the solution of the last system in vector form:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 9r + 2 \\ -4r + 1 \\ r \end{bmatrix} = r \begin{bmatrix} 9 \\ -4 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

As r varies, the vector $r = [9, -4, 1]^T$ covers the line in 3-space passing through the origin in the direction of $[9, -4, 1]^T$. This line is translated by the vector $[2, 1, 0]^T$. Hence, the solution set is the line through $(2, 1, 0)$ in the direction of $[9, -4, 1]^T$. So, the three planes intersect along this line.



No Solutions

Example 1.4.11 (No Solutions)**.** Find the intersection of the three planes in the (p, q, k)-coordinate system.

$$\begin{aligned} q - 2k &= -5 \\ 2p - q + k &= -2 \\ 4p - q &= -4 \end{aligned}$$

Solution: The augmented matrix of the system reduces to

$$\left[\begin{array}{cccc} 2 & -1 & 1 & -2 \\ 0 & 1 & -2 & -5 \\ 0 & 0 & 0 & 5 \end{array}\right]$$

The last row corresponds to the false expression $0 = 5$. Hence, the system is inconsistent. Therefore, the planes do not have a common intersection.

Echelon Forms

A **zero row** of a matrix is a row that consists entirely of zeros. The first nonzero entry of a nonzero row is called a **leading entry**. If a leading entry happens to be 1, we call it a **leading 1**. Similarly, we can talk about zero columns.

Definitions Consider the following conditions on a matrix A.

1. All zero rows are at the bottom of the matrix.
2. The leading entry of each nonzero row after the first occurs to the right of the leading entry of the previous row.
3. The leading entry in any nonzero row is 1.
4. All entries in the column above and below a leading 1 are zero.

If A satisfies the first two conditions, we call it **row echelon form**. If it satisfies all four conditions, we call it **reduced row echelon form**. We often omit the word "row" and just say *echelon form*, or *reduced echelon form*.

Example 1.4.12. Consider the following matrices.

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 & -6 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$D = \begin{bmatrix} 1 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & 4 \end{bmatrix}, \quad E = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} 1 & 7 & 0 & 9 & 0 \\ 0 & 0 & 1 & -8 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}$$

Matrices A, B, D, F, G, H are in echelon form. Out of these, A, B, D, F are in reduced echelon form. Matrices G and H are not in reduced echelon form. For G condition 4 fails. For H condition 3 fails. Matrices C and E are not in echelon form. For C condition 2 fails. For E condition 1 fails.

If a matrix B can be obtained from a matrix A by using elementary row operations, we say that A and B are **row equivalent**. It turns out that any matrix A is row equivalent to a unique matrix in reduced row echelon form, which we denote by $\text{rref}(A)$. The following algorithm finds $\text{rref}(A)$.

Algorithm 1.4.1 (Gauss Elimination)**.** *To reduce any matrix to reduced row echelon form apply the following steps.*

1. *Find the leftmost nonzero column.*
2. *If the first row has a zero in the column of step 1, interchange it with one that has a nonzero entry in the same column.*
3. *Obtain zeros below the leading entry by adding suitable multiples of the top row to the rows below that.*
4. *Cover the top row and repeat the same process starting with step 1 applied to the leftover submatrix. Repeat this process with the rest of the rows, until the matrix is in echelon form.*
5. *Starting with the last nonzero row work upward: For each row obtain a leading 1 and introduce zeros above it, by adding suitable multiples to the corresponding rows.*

Example 1.4.13. Apply Gauss elimination to find a reduced echelon form of the matrix.

$$\begin{bmatrix} 0 & 3 & -6 & -4 & -3 \\ -1 & 3 & -10 & -4 & -4 \\ 4 & -9 & 34 & 0 & 1 \\ 2 & -6 & 20 & 8 & 8 \end{bmatrix}$$

Solution:

$$\boxed{R_1 \leftrightarrow R_2} \quad \begin{bmatrix} \boxed{-1} & 3 & -10 & -4 & -4 \\ 0 & 3 & -6 & -4 & -3 \\ 4 & -9 & 34 & 0 & 1 \\ 2 & -6 & 20 & 8 & 8 \end{bmatrix}$$

The pivot now is -1, at pivot position $(1, 1)$.

$$\boxed{\begin{array}{l} R_3 + 4R_1 \to R_3 \\ R_4 + 2R_1 \to R_4 \end{array}} \quad \begin{bmatrix} -1 & 3 & -10 & -4 & -4 \\ 0 & 3 & -6 & -4 & -3 \\ 0 & 3 & -6 & -16 & -15 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 3 & -10 & -4 & -4 \\ \boxed{\begin{array}{cccc} 0 & 3 & -6 & -4 & -3 \\ 0 & 3 & -6 & -16 & -15 \\ 0 & 0 & 0 & 0 & 0 \end{array}} \end{bmatrix} \xrightarrow{\text{Step 1}} \begin{bmatrix} -1 & 3 & -10 & -4 & -4 \\ 0 & \boxed{\begin{array}{cccc} 3 & -6 & -4 & -3 \\ 3 & -6 & -16 & -15 \\ 0 & 0 & 0 & 0 \end{array}} \end{bmatrix}$$

The next pivot is 3, at position $(2, 2)$.

$$\begin{bmatrix} -1 & 3 & -10 & -4 & -4 \\ 0 & \boxed{3} & -6 & -4 & -3 \\ 0 & 3 & -6 & -16 & -15 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \boxed{R_3 - R_2 \to R_3} \quad \begin{bmatrix} -1 & 3 & -10 & -4 & -4 \\ 0 & 3 & -6 & -4 & -3 \\ 0 & 0 & 0 & \boxed{-12} & -12 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

STEP 5: Starting with the last nonzero row work upward: For each row obtain a leading 1 and introduce zeros above it, by adding suitable multiples to the corresponding rows.

$$\boxed{(-1/12)R_3 \to R_3} \quad \begin{bmatrix} -1 & 3 & -10 & -4 & -4 \\ 0 & 3 & -6 & -4 & -3 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \boxed{\begin{array}{l} R_2 + 4R_3 \to R_2 \\ R_1 + 4R_3 \to R_1 \end{array}}$$

We continue in the same fashion to get

$$\begin{bmatrix} -1 & 3 & -10 & 0 & 0 \\ 0 & 3 & -6 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \boxed{(1/3)R_2 \to R_2} \begin{bmatrix} -1 & 3 & -10 & 0 & 0 \\ 0 & 1 & -2 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \boxed{R_1 - 3R_2 \to R_1}$$

$$\begin{bmatrix} -1 & 0 & -4 & 0 & -1 \\ 0 & 1 & -2 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \boxed{(-1)R_1 \to R_1} \begin{bmatrix} 1 & 0 & 4 & 0 & 1 \\ 0 & 1 & -2 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The last matrix is the unique reduced row echelon form of the given matrix.

1.5 Gauss Elimination for Linear Systems

To solve any linear system we apply Gauss elimination to the augmented matrix of the system as follows:

Algorithm 1.5.1 (Solution of Linear System)**.** *To solve any linear system:*

1. *Apply Gauss elimination to the augmented matrix of the system (forward pass). If during any stage of this process it is found that the last column is a pivot column (i.e., it has a pivot), stop. In this case the system is inconsistent. Otherwise, continue with step 2.*

2. *Complete Gauss elimination to reduced row echelon form. Write the system whose augmented matrix is the reduced echelon form matrix, ignoring any zero equations.*

3. *Separate the variables of the reduced system into leading and free (if any). Write the free variables as parameters. Solve the leading variables in terms of the parameters and/or numbers.*

Example 1.5.1 (General Solution of Linear System)**.** Find the general solution of the system

$$\begin{aligned} 3x_2 - 6x_3 - 4x_4 - 3x_5 &= -5 \\ -x_1 + 3x_2 - 10x_3 - 4x_4 - 4x_5 &= -2 \\ 2x_1 - 6x_2 + 20x_3 + 2x_4 + 8x_5 &= -8 \end{aligned}$$

Solution: Gauss elimination on the augmented matrix of the system yields

$$\left[\begin{array}{cccccc} 1 & 0 & 4 & 0 & 1 & -3 \\ 0 & 1 & -2 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 2 \end{array}\right]$$

Therefore, the original system reduces to the equivalent system

$$\begin{array}{rrrrrcr} x_1 & & +4\,x_3 & & +x_5 & = & -3 \\ & x_2 & -2x_3 & & -x_5 & = & 1 \\ & & & x_4 & & = & 2 \end{array}$$

Next, we use parameters for the free variables and we solve for the leading variables to get the two-parameter infinite set:

$$\begin{array}{ll} x_1 = -4s - r - 3 & \\ x_2 = 2s + r + 1 & \\ x_3 = s & \text{for any } r, s \in \mathbf{R} \\ x_4 = 2 & \\ x_5 = r & \end{array}$$

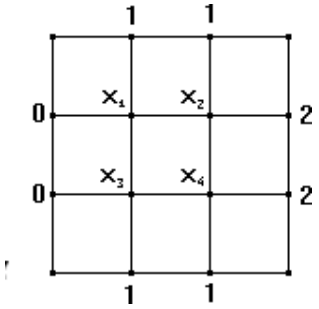
It is often useful to write the solution in vector form as follows:

$$\left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\right] = \left[\begin{array}{c} -4s - r - 3 \\ 2s + r + 1 \\ s \\ 2 \\ r \end{array}\right] = s\left[\begin{array}{c} -4 \\ 2 \\ 1 \\ 0 \\ 0 \end{array}\right] + r\left[\begin{array}{c} -1 \\ 1 \\ 0 \\ 0 \\ 1 \end{array}\right] + \left[\begin{array}{c} -3 \\ 1 \\ 0 \\ 2 \\ 0 \end{array}\right]$$

1.6 Application to Heat Conduction

Mean Value Property for Heat Conduction The temperature at any interior point of a square mesh covering a plate is the average of the temperatures of its neighbouring points.

Suppose, for simplicity, we have a rectangular metal plate with only four interior points with unknown temperatures x_1, x_2, x_3, x_4 and 12 boundary points (not named) with temperatures indicated in the figure below.



We want to compute the interior temperatures x_1, x_2, x_3, x_4.

Example 1.6.1 (Heat Conduction)**.** Given the temperatures on the boundary as indicated use the heat conduction principle to compute the interior temperatures x_1, x_2, x_3, x_4.

Solution: According to the Mean Value Property we have

$$\begin{aligned} x_1 &= \tfrac{1}{4}(x_2 + x_3 + 1) \\ x_2 &= \tfrac{1}{4}(x_1 + x_4 + 3) \\ x_3 &= \tfrac{1}{4}(x_1 + x_4 + 1) \\ x_4 &= \tfrac{1}{4}(x_2 + x_3 + 3) \end{aligned}$$

The system in standard form is

$$\begin{aligned} 4x_1 - x_2 - x_3 &= 1 \\ -x_1 + 4x_2 - x_4 &= 3 \\ -x_1 + 4x_3 - x_4 &= 1 \\ -x_2 - x_3 + 4x_4 &= 3 \end{aligned}$$

The augmented matrix

$$\begin{bmatrix} 4 & -1 & -1 & 0 & 1 \\ -1 & 4 & 0 & -1 & 3 \\ -1 & 0 & 4 & -1 & 1 \\ 0 & -1 & -1 & 4 & 3 \end{bmatrix}$$

reduces to

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 3/4 \\ 0 & 1 & 0 & 0 & 5/4 \\ 0 & 0 & 1 & 0 & 3/4 \\ 0 & 0 & 0 & 1 & 5/4 \end{bmatrix}$$

Hence, the interior temperatures are $x_1 = 3/4$, $x_2 = 5/4$, $x_3 = 3/4$, and $x_4 = 5/4$.

1.7 Linear Combinations of Vectors

Definition Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ be given n-vectors and let $c_1, c_2, \ldots, c_k$ be any scalars. The n-vector $\mathbf{v}$

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

is called a **linear combination** of $\mathbf{v}_1, \ldots, \mathbf{v}_k$. The scalars $c_1, \ldots, c_k$ are called the **coefficients** of the linear combination. If not all c_i are zero, we have a **nontrivial linear combination**. If all c_i are zero, we have the **trivial linear combination**. The trivial linear combination represents the zero vector.

The concept of linear combination is simple: we scale a few vectors and then we add them.

Notice that in Example 1.5.1 we wrote the solutions of a linear system as linear combinations of certain vectors.

Example 1.7.1. Check that the following are linear combinations of the vectors $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$.

$$-\mathbf{v}_1 + 3\mathbf{v}_2 + 4\mathbf{v}_3, \quad \mathbf{v}_1 + 1.5\mathbf{v}_2 - 9\mathbf{v}_3, \quad \mathbf{v}_1 - \mathbf{v}_3$$

Solution: We have

$$\begin{aligned} -\mathbf{v}_1 + 3\mathbf{v}_2 + 4\mathbf{v}_3 &= (-1)\,\mathbf{v}_1 + 3\mathbf{v}_2 + 4\mathbf{v}_3 \\ \mathbf{v}_1 + 1.5\mathbf{v}_2 - 9\mathbf{v}_3 &= 1\mathbf{v}_1 + (1.5)\,\mathbf{v}_2 + (-9)\,\mathbf{v}_3 \\ \mathbf{v}_1 - \mathbf{v}_3 &= 1\mathbf{v}_1 + 0\mathbf{v}_2 + (-1)\,\mathbf{v}_3 \end{aligned}$$

Note that the difference $\mathbf{v}_1 - \mathbf{v}_2$ between two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is the linear combination $1\mathbf{v}_1 + (-1)\,\mathbf{v}_2$ with coefficients 1 and -1.

Note Linear combinations of vectors are intimately connected with matrix-vector multiplication. In fact if $[\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_k]$ is the matrix with columns $\mathbf{v}_i$, then

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = [\mathbf{v}_1|\mathbf{v}_2|\ldots|\mathbf{v}_k] \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix}$$

Example 1.7.2. We have

$$a\begin{bmatrix}1\\2\end{bmatrix}+b\begin{bmatrix}3\\4\end{bmatrix}=\begin{bmatrix}1&3\\2&4\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix}$$

Example 1.7.3 (Linear Combination)**.** Is $\begin{bmatrix}0\\10\\-16\end{bmatrix}$ a linear combination of the vectors $\begin{bmatrix}1\\2\\3\end{bmatrix},\begin{bmatrix}2\\0\\1\end{bmatrix},\begin{bmatrix}-1\\4\\-5\end{bmatrix}$?

Solution: This is true, if there exist scalars c_1, c_2, c_3 such that

$$\begin{bmatrix}0\\10\\-16\end{bmatrix}=c_1\begin{bmatrix}1\\2\\3\end{bmatrix}+c_2\begin{bmatrix}2\\0\\1\end{bmatrix}+c_3\begin{bmatrix}-1\\4\\-5\end{bmatrix}$$

or

$$\begin{bmatrix}1&2&-1\\2&0&4\\3&1&-5\end{bmatrix}\begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix}=\begin{bmatrix}0\\10\\-16\end{bmatrix}$$

Solving the linear system yields $c_1=-1, c_2=2, c_3=3$. So, yes, the vector is a linear combination of the three given vectors.

1.8 The Span of Vectors

The set of all linear combinations of given n-vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ is called the **span** of these vectors and it is denoted by $\text{Span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$. The span consists of all vectors of the form

$$c_1\mathbf{v}_1+c_2\mathbf{v}_2+\cdots+c_k\mathbf{v}_k.$$

where the coeffients c_i may take on any real values.

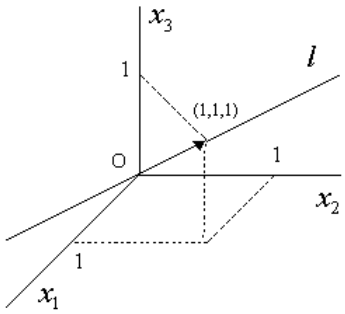
Notes

- The span always contains the zero vector (choose all $c_i=0$).
- The span is an **infinite set**, unless all the $\mathbf{v}_i$ are $\mathbf{0}$.

- The span of one n-vector $\mathbf{v}$ consists of all scalar products of $\mathbf{v}$.

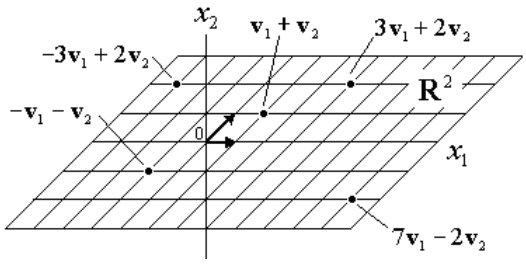
$$\text{Span}\,\{\mathbf{v}\} = \{c\mathbf{v}, \quad c \in \mathbf{R}\}$$

For example, the span of $\mathbf{v} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$ in $\mathbf{R}^3$ is the space line through the origin and $\mathbf{v}$. (See figure below.)



It is geometrically clear that the span of two 2-vectors, or two 3-vectors, that are not multiples of each each other is the unique plane through the origin containing these vectors.

For example, Span $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\} = \mathbf{R}^2$. (See figure below.)



1.9 Linear Dependence and Independence

Linear Dependence

Definition The sequence of m-vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is **linearly dependent** (or the vectors are linearly dependent), if there are scalars $c_1, \ldots, c_k$ *not all zero* such that

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \tag{1.7}$$

This is the same as saying that there is a nontrivial linear combination of the $\mathbf{v}_i$s representing the zero vector. Equation (1.7) with not all c_i zero is called a **linear dependence relation** of the $\mathbf{v}_i$s.

Special Cases

- **One Vector:** Vector $\mathbf{v}$ is linearly dependent if and only if $\mathbf{v} = \mathbf{0}$.
- **Two Vectors:** Vectors $\mathbf{v_1}, \mathbf{v_2}$ are linearly dependent if and only if one vector is a scalar multiple of the other. This because $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \mathbf{0}$ with, say, $c_1 \neq 0$ is equivalent to $\mathbf{v}_1 = (-c_2/c_1)\mathbf{v}_2$.

Example 1.9.1. The vectors

$$\begin{bmatrix} 1 \\ -1 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 \\ 14 \\ -6 \\ 4 \end{bmatrix}$$

are linearly dependent, because if we let $c_1 = 2$, $c_2 = -6$, and $c_3 = 1$, then

$$2\begin{bmatrix} 1 \\ -1 \\ 3 \\ 4 \end{bmatrix} + (-6)\begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \end{bmatrix} + 1\begin{bmatrix} 4 \\ 14 \\ -6 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Note In order to determine linear dependence we do not have to guess, as we just did, but check whether the homogeneous linear system with coefficient columns the given vectors has non-trivial solutions. This is because

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \Leftrightarrow [\mathbf{v}_1|\ldots|\mathbf{v}_k]\mathbf{c} = \mathbf{0}$$

where $[\mathbf{v}_1|\ldots|\mathbf{v}_k]$ is the matrix with columns $\mathbf{v}_i$ and $\mathbf{c}$ is the vector with components c_i.

Example 1.9.2. Let $S = \left\{ \begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 7 \end{bmatrix}, \begin{bmatrix} 3 \\ 14 \\ 9 \end{bmatrix} \right\}$.

(a) Show that S is linearly dependent.

(b) Find a linear dependence relation.

Solution: (a) We seek c_1, c_2, c_3 not all zero such that

$$c_1 \begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 2 \\ 7 \end{bmatrix} + c_3 \begin{bmatrix} 3 \\ 14 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Equivalently, we seek nontrivial solutions to the *homogeneous* linear system

$$\begin{bmatrix} 0 & 1 & 3 \\ -2 & 2 & 14 \\ 3 & 7 & 9 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We solve this system to get $c_1 = 4r$, $c_2 = -3r$, $c_3 = r$. There are nontrivial solutions, hence the set is linearly dependent.
(b) To get a particular linear dependence relation we assign a nonzero value to the parameter r. For example, if $r = 1$, then we have

$$4 \begin{bmatrix} 0 \\ -2 \\ 3 \end{bmatrix} + (-3) \begin{bmatrix} 1 \\ 2 \\ 7 \end{bmatrix} + 1 \begin{bmatrix} 3 \\ 14 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

This is one of infinitely many linear dependence relations.

From the definition of linear dependence we see that the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are linearly dependent if and only if at least one of them is a linear combination of the remaining. This because if, say, some $c_i \neq 0$, then we can use the linear dependence relation to solve for $\mathbf{v}_i$ in terms of the other vectors. For example, if $c_1 \neq 0$, we have

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \Leftrightarrow \mathbf{v}_1 = \left(-\frac{c_2}{c_1}\right)\mathbf{v}_2 + \cdots + \left(-\frac{c_k}{c_1}\right)\mathbf{v}_k$$

Remark Linearly dependent sets have vectors that are **redundant**. If we drop a vector that is a linear combination in the remaining, then the *span does not change*. The dropped vector can be recovered from the appropriate linear combination of the remaining.

So for example, in $\mathbf{R}^3$ a set of three linearly dependent vectors must be on the same plane, because if we drop one vector, then the span of the other two defines either a plane or a line.

Linear Independence

Linear independence of vectors is the opposite of linear dependence. This time the vectors are not redundant: if we drop any one of them the span changes. More precisely, we have:

Definition The set of m-vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is called **linearly independent**, if it is not linearly dependent. This is the same as saying that *there is no linear dependence relation* among $\mathbf{v}_1, \ldots, \mathbf{v}_k$. So, *all nontrivial linear combinations of the* $\mathbf{v}_i$*s yield nonzero vectors.* Equivalently, we have

$$\text{if} \quad c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0}, \quad \text{then} \quad c_1 = 0, \ldots, c_k = 0$$

This is equivalent to saying that the homogeneous system $[\mathbf{v}_1|\cdots|\mathbf{v}_k]\,\mathbf{c} = \mathbf{0}$ has only the trivial solution (where, $\mathbf{c} = [c_1 \ldots c_k]^T$).

Special Cases

- **One Vector:** Vector $\mathbf{v}$ is linearly independent if and only if $\mathbf{v} \neq \mathbf{0}$.
- **Two Vectors:** Vectors $\mathbf{v_1}, \mathbf{v_2}$ are linearly independent if and only if none is a scalar multiple of the other.

Example 1.9.3. Show that $\left\{ \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \begin{bmatrix} 5 \\ 3 \end{bmatrix} \right\}$ is linearly independent in $\mathbf{R}^2$.

Solution: Let c_1 and c_2 be scalars such that $c_1\mathbf{e}_1 + c_2\mathbf{e}_2 = \mathbf{0}$. In other words,

$$c_1 \begin{bmatrix} 1 \\ -2 \end{bmatrix} + c_2 \begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We solve the system with augmented matrix $\begin{bmatrix} 1 & 5 & 0 \\ -2 & 3 & 0 \end{bmatrix}$ to get $c_1 = 0$ and $c_2 = 0$. Therefore, the set is linearly independent.

Note To check vectors $\mathbf{v}_1, \cdots, \mathbf{v}_k$ for linear independence we do not need to actually solve the system $[\mathbf{v}_1|\cdots|\mathbf{v}_k]\,\mathbf{c} = \mathbf{0}$. It suffices to row reduce the matrix $[\mathbf{v}_1|\cdots|\mathbf{v}_k]$ to **any** echelon form: if each column has a pivot, then the set must be linearly independent.

Example 1.9.4. Show that S is linearly independent.

$$S = \left\{ \begin{bmatrix} 2 \\ 3 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 8 \\ -6 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} -4 \\ 3 \\ 1 \\ -6 \end{bmatrix} \right\}$$

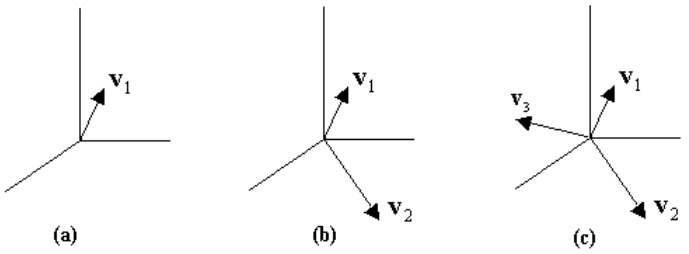
Solution: We only need to count the number of pivots of the coefficient matrix.

$$\begin{bmatrix} 2 & 8 & -4 \\ 3 & -6 & 3 \\ 2 & 5 & 1 \\ 4 & 0 & -6 \end{bmatrix} \sim \begin{bmatrix} 2 & 8 & -4 \\ 0 & -3 & 5 \\ 0 & 0 & -21 \\ 0 & 0 & 0 \end{bmatrix}$$

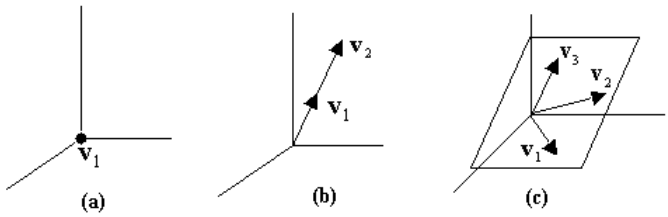
This number is 3, the same as the number of columns, so the set is linearly independent.

Geometry of Linear Dependence/Independence

If $\mathbf{v}_1, \cdots, \mathbf{v}_k$ are linearly independent, then none of $\mathbf{v}_i$ is a linear combination of the remaining. So, for example, in three-dimensional space two such vectors cannot be on the same line and three of them cannot be on the same plane. The picture below shows cases of one, two, and three linearly independent vectors.



In contrast, we have the following for linearly dependent vectors:



1.10 Rank of Matrix

Row Space; Column Space

Let A be any size matrix. The span of all its columns is called the **column space** of A. The span of all its rows is called the **row space** of A. Let us

denote the column space by $\text{Col}(A)$ and the row space by $\text{Row}(A)$.

When we apply elementary row operations to a matrix we do not change its row space, because the rows of the new matrix are linear combinations of the original matrix and vice versa. So, *row reduction does not change the row space*. In particular, we have that $\text{Row}(A) = \text{Row}(B)$, where B is any row echelon form of A (in particular B can be $\text{rref}(A)$). Let S be the set that consists of all nonzero rows of B. Then S is linearly independent, because the pivots in each row occur at different positions. S is rather special because it is both linearly independent and its span is the entire row space of B, thus the entire row space of A. If we add to S any element of the row space of A, then the new set will no longer be linearly independent. If we take any vector out of S, then the new set has a span that is smaller than the row space of A. We summarize in the following theorem.

Theorem 1.10.1. *Let A be any matrix and let B be any echelon form of A. If S is the set of nonzero rows of B, then*

1. *S is linearly independent.*
2. *The span of S equals Row(A).*
3. *S is a maximal linearly independent set for Row(A). I.e., if we add any element of Row(A) to S, then the new set is no longer linearly independent.*
4. *S is a minimal spanning set for Row(A). I.e., if we take out elements from S, the new set no longer spans Row(A).*

The number of elements r of S in the theorem is called the **rank** of A. This number is the same as the number of pivots of A and it is independent of the particular echelon form B.

From our previous discussion we conclude the following:

The **rank** of any matrix A equals

1. the number of the pivots of A.
2. the number of nonzero rows of any echelon form of A.
3. the number of pivot rows of any echelon form of A.

4. the number of pivot columns of any echelon form of A.

5. the maximum number of linearly independent rows of A.

6. the minimum number of spanning rows of Row(A).

In addition, it can be proved that

1. A and A^T have the same rank.

2. The rank of A is the maximum number of linearly independent columns of A.

Example 1.10.1. Find the rank of $A = \begin{bmatrix} 1 & 2 & 2 & -1 \\ 1 & 3 & 1 & -2 \\ 1 & 1 & 3 & 0 \\ 0 & 1 & -1 & -1 \\ 1 & 2 & 2 & -1 \end{bmatrix}$.

Solution: A row reduces to the echelon form B:

$$B = \begin{bmatrix} 1 & 2 & 2 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

B has two nonzero rows, so the rank of A is 2.

1.11 Vector Spaces

Definition Let V be a set equipped with two operations named **addition** and **scalar multiplication**. Addition is a map that associates any two elements $\mathbf{u}$ and $\mathbf{v}$ of V with a third one, called the **sum** of $\mathbf{u}$ and $\mathbf{v}$ and denoted by $\mathbf{u} + \mathbf{v}$.

$$V \times V \to V, \qquad (\mathbf{u}, \mathbf{v}) \to \mathbf{u} + \mathbf{v}$$

Scalar multiplication is a map that associates any real scalar c and any element $\mathbf{u}$ of V with another element of V, called the **scalar multiple** of $\mathbf{u}$ by c and denoted by $c\mathbf{u}$.

$$\mathbf{R} \times V \to V, \qquad (c, \mathbf{u}) \to c\mathbf{u}$$

Such a set V is called a (real) **vector space**, if the two operations satisfy the following properties, known as **axioms** for a vector space.

Addition

(A1) $\mathbf{u}+\mathbf{v}$ belongs to V for all $\mathbf{u},\mathbf{v}\in V$.

(A2) $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$ for all $\mathbf{u},\mathbf{v}\in V$. (**Commutative Law**)

(A3) $(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w})$ for all $\mathbf{u},\mathbf{v},\mathbf{w}\in V$. (**Associative Law**)

(A4) There exists a unique element $\mathbf{0}\in V$, called the **zero** of V, such that for all $\mathbf{u}$ in V

$$\mathbf{u}+\mathbf{0}=\mathbf{0}+\mathbf{u}=\mathbf{u}$$

(A5) For each $\mathbf{u}\in V$ there exists a unique element $-\mathbf{u}\in V$, called the **negative** or **opposite** of $\mathbf{u}$, such that

$$\mathbf{u}+(-\mathbf{u})=(-\mathbf{u})+\mathbf{u}=\mathbf{0}$$

Scalar Multiplication

(M1) $c\,\mathbf{u}$ belongs to V for all $\mathbf{u}\in V$ and all $c\in\mathbf{R}$.

(M2) $c(\mathbf{u}+\mathbf{v})=c\mathbf{u}+c\mathbf{v}$ for all $\mathbf{u},\mathbf{v}\in V$ and all $c\in\mathbf{R}$.(**Distributive Law**)

(M3) $(c+d)\mathbf{u}=c\mathbf{u}+d\mathbf{u}$ for all $\mathbf{u}\in V$ and all $c,d\in\mathbf{R}$.(**Distributive Law**)

(M4) $c(d\mathbf{u})=(cd)\mathbf{u}$ for all $\mathbf{u}\in V$ and all $c,d\in\mathbf{R}$.

(M5) $1\mathbf{u}=\mathbf{u}$ for all $\mathbf{u}\in V$.

The elements of a vector space are called **vectors**. Axioms (A1) and (M1) are also expressed by saying that V **is closed under addition** and **is closed under scalar multiplication**. Note that *a vector space is a nonempty set,* because it has a zero by (A4).

Examples of Vector Spaces

Example 1.11.1. The following are examples of vector spaces.

1. The set $\mathbf{R}^n$ of all n-vectors with real components.

   Operations: The usual vector addition and scalar multiplication. *Zero:* The zero n-vector $\mathbf{0}$. *Axioms:* For the axioms see the Properties of Matrix Operations Theorem.

2. The set M_{mn} of all $m \times n$ matrices with real entries.

   Operations: The usual matrix addition and scalar multiplication. *Zero:* The $m \times n$ zero matrix $\mathbf{0}$. *Axioms:* For the axioms see the Properties of Matrix Operations Theorem.

3. The set P of all polynomials with real coefficients.

   Operations:

   (a) *Addition:* The sum of two polynomials is formed by adding the coefficients of the same powers of x of the polynomials. Explicitly, if

   $$p_1 = a_0 + a_1 x + \cdots + a_n x^n, \quad p_2 = b_0 + b_1 x + \cdots + b_m x^m, \quad n \geq m$$

   we write p_2 as $p_2 = b_0 + b_1 x + \cdots + b_n x^n$, by adding zeros if necessary, and form the sum

   $$p_1 + p_2 = (a_0 + b_0) + (a_1 + b_1)\, x + \cdots + (a_n + b_n)\, x^n$$

   (b) *Scalar multiplication:* This is multiplication of a polynomial through by a constant.

   $$cp_1 = (ca_0) + (ca_1)\, x + \cdots + (ca_n)\, x^n$$

   (c) *Zero:* The zero polynomial, $\mathbf{0}$, is the polynomial with zeros as coefficients.

   (d) *Axioms:* The verification of the axioms is left as exercise.

4. The set $F(\mathbf{R})$ of all real valued functions defined on $\mathbf{R}$.

   Operations: Let f and g be two real valued functions with domain $\mathbf{R}$ and let c be any scalar.

(a) *Addition:* We define the sum $f + g$ of f and g as the function whose values are given by

$$(f + g)(x) = f(x) + g(x) \qquad \text{for all } x \in \mathbf{R}$$

(b) *Scalar multiplication:* The scalar product cf is defined by

$$(c\,f)(x) = c\,f(x) \qquad \text{for all } x \in \mathbf{R}$$

(c) *Zero:* The zero function $\mathbf{0}$ is the function whose values are all zero.

$$\mathbf{0}(x) = 0 \qquad \text{for all } x \in \mathbf{R}$$

(d) *Axioms:* The verification of the axioms is left as exercise.

Example 1.11.2. Is $\mathbf{R}^2$ with the usual addition and the following scalar multiplication, denoted by $\odot$, a vector space?

$$c \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} ca_1 \\ a_2 \end{bmatrix}$$

Solution: It is *not* a vector space, because

$$(c + d) \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} (c + d)\,a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} ca_1 + da_1 \\ a_2 \end{bmatrix}$$

and

$$c \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + d \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} ca_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} da_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} ca_1 + da_1 \\ 2a_2 \end{bmatrix}$$

So, $(c + d) \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \neq c \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + d \odot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ and axiom (M3) fails.

Subspaces

Definition A subset W of a vector space V is called a **subspace** of V, if W itself is a vector space under the same addition and scalar multiplication as V. In particular, a subspace always contains the zero element.

Theorem 1.11.1 (Criterion for Subspace)**.** *Let W be a nonempty subset W of a vector space V. Then W is a subspace of V if and only if it is closed under addition (axiom (A1)) and scalar multiplication (axiom (M1)), that is, if and only if*

1. *If $\mathbf{u}$ and $\mathbf{v}$ are in W, then $\mathbf{u}+\mathbf{v}$ is in W.*
2. *If c is any scalar and $\mathbf{u}$ is in W, then $c\,\mathbf{u}$ is in W.*

Example 1.11.3 (Subspace?)**.** Check to see if the set S is a vector subspace of $\mathbf{R}^2$.

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x_1 + 3x_2 = 0 \right\} \subset \mathbf{R}^2$$

Solution: Since S is already in $\mathbf{R}^2$ which is a vector space, we only need to check if S is closed under addition and scalar multiplication. For addition, let $\mathbf{u} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ be in S. Hence, by the definition fo S we have $x_1 + 3x_2 = 0$ and $y_1 + 3y_2 = 0$. We must verify that $\mathbf{u}+\mathbf{v}$ is also in S. We have

$$\mathbf{u}+\mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \end{bmatrix}$$

Now, $(x_1 + y_1) + 3\,(x_2 + y_2) = (x_1 + 3x_2) + (y_1 + 3y_2) = 0 + 0 = 0$. Hence, $\mathbf{u}+\mathbf{v}$ is in S.

For scalar multiplication, for any scalar c and any $\mathbf{u} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ in S (so $x_1 + 3x_2 = 0$) we have

$$c\mathbf{u} = c \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} cx_1 \\ cx_2 \end{bmatrix}$$

and $cx_1 + 3\,(cx_2) = c\,(x_1 + 3x_2) = c0 = 0$. Hence, $c\mathbf{v}$ is in S.

Therefore, S is closed both under addition and scalar multiplication. Hence, it is a subspace of $\mathbf{R}^2$.

Example 1.11.4 (Subspace?)**.** Check to see if the set S is a vector subspace of $\mathbf{R}^3$.

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad x_1 x_2 x_3 = 0 \right\} \subset \mathbf{R}^3$$

Solution: This set is not closed under vector addition. Indeed, the computation

$$\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

shows that the vectors $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ which are both in S have a sum which is not in S. Therefore, S is not a subspace of $\mathbf{R}^3$.

Example 1.11.5. We have the following examples of subspaces.

1. The set $W = \{c\mathbf{v},\ c \in \mathbf{R}\}$ of all scalar multiples of the fixed vector $\mathbf{v}$ of a vector space V is a subspace of V.
2. $\{\mathbf{0}\}$ and V are subspaces of V. These are the **trivial subspaces** of V. $\{\mathbf{0}\}$ is called the **zero subspace**.
3. The set P_n that consists of all polynomials of degree $\leq n$ and the zero polynomial is a subspace of P.
4. The set $C(\mathbf{R})$ of all continuous real valued functions defined on $\mathbf{R}$ is a subspace of $F(\mathbf{R})$.

Linear Combinations and Span

If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are vectors from a vector space V and $c_1, \ldots, c_n$ are scalars, then the expression

$$c_1\,\mathbf{v}_1 + \cdots + c_n\ \mathbf{v}_n$$

is well defined and is called a **linear combination** of $\mathbf{v}_1, \ldots, \mathbf{v}_n$. If not all c_i are zero, we have a **nontrivial linear combination**. If all c_i are zero, we have the **trivial linear combination**. The trivial linear combination represents the zero vector.

The set of all linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is called the **span** of these vectors and it is denoted by

$$\text{Span}\,\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$$

If $V = \text{Span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, we say that $\mathbf{v}_1, \ldots, \mathbf{v}_k$ **span** V and that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is a **spanning set** of V.

Example 1.11.6. Let V be a vector space and let $\mathbf{v}_1, \mathbf{v}_2$ be in V. The following vectors are in $\text{Span}\{\mathbf{v}_1, \mathbf{v}_2\}$.

$$\mathbf{0}, \quad \mathbf{v}_1, \quad \mathbf{v}_2\,, \quad \mathbf{v}_1 + \mathbf{v}_2\,, \quad -2\mathbf{v}_1\,, \quad 3\mathbf{v}_1 - 2\mathbf{v}_2$$

Example 1.11.7. Let V be a vector space and $\mathbf{v}$ be in V. $\text{Span}\{\mathbf{v}\}$ is the set of all scalar multiples of $\mathbf{v}$.

$$\text{Span}\{\mathbf{v}\} = \{c\mathbf{v}\ ,\ c \in \mathbf{R}\}$$

Example 1.11.8. Let $p = -1+x-2x^2$ in P_3. Show that $p \in \text{Span}\{p_1, p_2, p_3\}$, where

$$p_1 = x - x^2 + x^3, \quad p_2 = 1 + x + 2x^3, \quad p_3 = 1 + x$$

Solution: Let c_1, c_2, c_3 be scalars such that

$$-1 + x - 2x^2 = c_1\left(x - x^2 + x^3\right) + c_2\left(1 + x + 2x^3\right) + c_3\left(1 + x\right)$$

Then

$$-1 + x - 2x^2 = (c_2 + c_3) + (c_1 + c_2 + c_3)\,x - c_1x^2 + (c_1 + 2c_2)\,x^3$$

Equating coefficients of the same powers of x yields the linear system

$$c_2 + c_3 = -1,\ c_1 + c_2 + c_3 = 1,\ -c_1 = -2,\ c_1 + 2c_2 = 0$$

with solution $c_1 = 2, c_2 = -1, c_3 = 0$. Therefore, p is in the span of p_1, p_2, p_3.

Note We can define the span of an *infinite* subset S of vectors from a vector space V. We simply take as the span to be the set of all **finite linear combinations** of elements of S.

The most important property of the span is that it is a subspace.

Theorem 1.11.2 (The Span is a Subspace)**.** *Let* $\mathbf{v}_1, \ldots, \mathbf{v}_k, \ldots$ *be vectors in a vector space* V*. Then the span,* $\text{Span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k \ldots\}$ *is a subspace of* V*.*

Proof. The sum of two linear combinations of elements of S is also a linear combination. And the any scalar product of linear combinations is again a linear combination. So the span is closed under the two vector space operations. Hence, it is a subspace of V.

This theorem often makes it easy to prove that a subset is a subspace. We try to rewrite the subset as the span of vectors.

Example 1.11.9. Show that $S = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad x_1 + x_2 = 0 \right\}$ is subspace of $\mathbf{R}^3$.

Solution:

$$S = \left\{ \begin{bmatrix} x_1 \\ -x_1 \\ x_3 \end{bmatrix}, \quad x_1, x_3 \text{ any} \right\} = \left\{ x_1 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad x_1, x_3 \text{ any} \right\}$$

$$= \text{Span} \left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

So S is a subspace because it is the span of the two fixed vectors, namely, $\begin{bmatrix} 1 & -1 & 0 \end{bmatrix}^T$ and $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$.

Linear Independence

Definition Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be vectors of a vector space V. Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is **linearly dependent**, if there are scalars $c_1, \ldots, c_n$ *not all zero* such that

$$c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \mathbf{0} \tag{1.8}$$

So, *there are nontrivial linear combinations that represent the zero vector.* Equation (1.8) with not all c_i zero is called a **linear dependence relation** of the $\mathbf{v}_i$s.

Example 1.11.10. Show that the set $\{2 - x + x^2, 2x + x^2, 4 - 4x + x^2\}$ is linearly dependent in P_3.

Solution: This true, because

$$2(2 - x + x^2) + (-1)(2x + x^2) + (-1)\left(4 - 4x + x^2\right) = \mathbf{0}$$

Definition The set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ from a vector space V is called **linearly independent**, if it is not linearly dependent. This is the same as saying that *there is no linear dependence relation* among $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Equivalently,

$$c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k = \mathbf{0} \Rightarrow c_1 = 0, \ldots, c_k = 0$$

So, *every nontrivial linear combination is nonzero.*

Example 1.11.11. Show that the standard basis matrices set $\{E_{11}, E_{12}, E_{21}, E_{22}\}$ is linearly independent in M_{22}.

Solution: Let

$$c_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c_3 \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + c_4 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Hence, $c_1 = c_2 = c_3 = c_4 = 0$. So, the set is linearly independent.

Example 1.11.12. Are $1 + x, -1 + x, 4 - x^2, 2 + x^3$ linearly independent in P_3?

Solution: If a linear combination in these polynomials is the zero polynomial, then

$$c_1 (1 + x) + c_2 (-1 + x) + c_3 (4 - x^2) + c_4 (2 + x^3) = \mathbf{0} \quad \Rightarrow$$
$$(c_1 - c_2 + 4c_3 + 2c_4) + (c_1 + c_2) x + (-c_3) x^2 + c_4 x^3 = \mathbf{0}$$

Equating coefficients yields,

$$c_1 - c_2 + 4c_3 + 2c_4 = 0, \quad c_1 + c_2 = 0, \quad -c_3 = 0, \quad c_4 = 0$$

We solve this linear system to get $c_1 = c_2 = c_3 = c_4 = 0$. So, the vectors are linearly independent in P_3.

Basis of a Vector Space

Definition A subset $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of a nonzero vector space V is a basis of V, if

1. it is linearly independent, and
2. it spans V.

The empty set is, by definition, the only basis of the zero vector space $\{\mathbf{0}\}$.

Example 1.11.13. Some examples of bases are

1. The set of the standard basis vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ in $\mathbf{R}^n$ is a basis of $\mathbf{R}^n$. This is called the **standard basis** of $\mathbf{R}^n$.

2. $\{1, x, x^2, \ldots, x^n\}$ is a basis of P_n, called the **standard basis** of P_n.

3. The set of the standard basis matrices $\{E_{11}, E_{12}, E_{13}, \ldots, E_{mn}\}$ of M_{mn} is a basis of M_{mn}. This is called the **standard basis** of M_{mn}.

Example 1.11.14 (Basis)**.** Show that $\mathcal{B} = \{1 + x, -1 + x, x^2\}$ is a basis of P_2.

Solution:

(1) To show that $\mathcal{B}$ spans P_2, we want every polynomial $p = a + bx + cx^2$ to be a linear combination in $\mathcal{B}$. So, we look for scalars c_1, c_2, c_3 such that

$$\begin{gathered} c_1(1 + x) + c_2(-1 + x) + c_3x^2 = a + bx + cx^2 \Rightarrow \\ (c_1 - c_2) + (c_1 + c_2)x + c_3x^2 = a + bx + cx^2 \end{gathered}$$

which leads to the system $c_1 - c_2 = a$, $c_1 + c_2 = b$, $c_3 = c$. We have

$$\begin{bmatrix} 1 & -1 & 0 & a \\ 1 & 1 & 0 & b \\ 0 & 0 & 1 & c \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & \frac{1}{2}a + \frac{1}{2}b \\ 0 & 1 & 0 & -\frac{1}{2}a + \frac{1}{2}b \\ 0 & 0 & 1 & c \end{bmatrix}$$

so the system is consistent for all choices of a, b, c. Thus, $\mathcal{B}$ spans P_2.

(2) To show that $\mathcal{B}$ is linearly independent, let c_1, c_2, c_3 be such that

$$\begin{gathered} c_1(1 + x) + c_2(-1 + x) + c_3x^2 = \mathbf{0} \Rightarrow \\ (c_1 - c_2) + (c_1 + c_2)x + c_3x^2 = \mathbf{0} \end{gathered}$$

Hence, we have the system $c_1 - c_2 = 0$, $c_1 + c_2 = 0$, $c_3 = 0$. Now

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

So the system has only the trivial solution $c_1 = c_2 = c_3 = 0$. Thus, $\mathcal{B}$ is linearly independent.

One of the main characterizations of a basis is described in the following theorem.

Theorem 1.11.3. *A subset $\mathcal{B} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of a vector space V is a basis of V if and only if for each vector $\mathbf{v}$ in V there are **unique** scalars $c_1, \ldots, c_n$ such that*

$$\mathbf{v} = c_1\mathbf{v}_1 + \cdots + c_n\mathbf{v}_n$$

Note The unique coeffiecients c_i in the representation of vector $\mathbf{v}$ as linear combination in the basis $\mathcal{B}$ form a vector called the **coordinate vector** $[\mathbf{v}]_{\mathcal{B}}$ of $\mathbf{v}$ with respect to the basis $\mathcal{B}$.

Example 1.11.15. Find the coordinate vector $[\mathbf{v}]_{\mathcal{B}}$ of $\mathbf{v} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$ with respect to the basis $\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 5 \end{bmatrix} \right\}$ of $\mathbf{R}^2$.

Solution: We have

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix} = (9/7)\begin{bmatrix} 1 \\ 2 \end{bmatrix} + (-5/7)\begin{bmatrix} -1 \\ 5 \end{bmatrix}$$

Therefore, $[\mathbf{v}]_{\mathcal{B}} = \begin{bmatrix} 9/7 \\ -5/7 \end{bmatrix}$.

Dimension

Definitions If a vector space V has a basis with n elements, then V is called **finite dimensional** and we say that n is the **dimension** of V. We write

$$\dim(V) = n$$

Note that the dimension is a well defined number and does not depend on the choice of basis.

The dimension of the zero space $\{\mathbf{0}\}$ is *defined* to be zero. A vector space that has no finite spanning set it is called **infinite dimensional**.

Example 1.11.16. By counting the number of elements of the standard bases we see that

1. $\dim(\mathbf{R}^n) = n$

2. $\dim(P_n) = n + 1$

3. $\dim(M_{mn}) = m \cdot n$

Note *The rank of a matrix is the dimension of its column space.*

Finding a Basis the Dimension of Subspace

Example 1.11.17. Show that S is a subspace of $\mathbf{R}^3$ and find a basis and its dimension.

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \ -5x_1 + x_2 = 0 \right\}$$

Solution: We have

$$S = \left\{ \begin{bmatrix} x_1 \\ 5x_1 \\ x_3 \end{bmatrix}, \ x_1, x_3 \text{ any} \right\} = \left\{ x_1 \begin{bmatrix} 1 \\ 5 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \ x_1, x_3 \text{ any} \right\}$$

$$= \text{Span} \left\{ \begin{bmatrix} 1 \\ 5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

Since S is the span of the vectors $\begin{bmatrix} 1 & 5 & 0 \end{bmatrix}^T$ and $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$, it is a subspace of $\mathbf{R}^3$. In addition, these vectors are linearly independent, hence they are a **basis** of S. The dimension of S is two because a basis of it has two elements.

Example 1.11.18. It is easy to see that

$$S = \left\{ \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^T, \ x_1 + x_2 + x_3 = 0 \right\}$$

is a subspace of $\mathbf{R}^4$. Furthermore, S has a basis that consists of the vectors $\begin{bmatrix} 1 & 0 & -1 & 0 \end{bmatrix}^T$, $\begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}^T$, and $\begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$. So, S is 3-dimensional.

1.12 Determinants

Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. The determinant, $\det(A)$, of A is the number

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

Let A be

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

The determinant of A in terms of 2×2 determinants is the number

$$\det(A) = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

In the same manner we can define determinants of 4×4 matrices.

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} & a_{24} \\ a_{32} & a_{33} & a_{34} \\ a_{42} & a_{43} & a_{44} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} & a_{24} \\ a_{31} & a_{33} & a_{34} \\ a_{41} & a_{43} & a_{44} \end{vmatrix} +$$
$$+ a_{13} \begin{vmatrix} a_{21} & a_{22} & a_{24} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44} \end{vmatrix} - a_{14} \begin{vmatrix} a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{vmatrix}$$

Example 1.12.1. Find $\det(C)$, if

$$C = \begin{bmatrix} 1 & 2 & 0 & 1 \\ -1 & 1 & 2 & 0 \\ -2 & 1 & 0 & -2 \\ 1 & 0 & 2 & -1 \end{bmatrix}$$

Solution: $\det(C)$ equals

$$1 \begin{vmatrix} 1 & 2 & 0 \\ 1 & 0 & -2 \\ 0 & 2 & -1 \end{vmatrix} - 2 \begin{vmatrix} -1 & 2 & 0 \\ -2 & 0 & -2 \\ 1 & 2 & -1 \end{vmatrix} + 0 \begin{vmatrix} -1 & 1 & 0 \\ -2 & 1 & -2 \\ 1 & 0 & -1 \end{vmatrix} - 1 \begin{vmatrix} -1 & 1 & 2 \\ -2 & 1 & 0 \\ 1 & 0 & 2 \end{vmatrix}$$
$$= 1 \cdot 6 - 2 \cdot (-12) + 0 \cdot (-3) - 1 \cdot 0 = 30$$

Cofactor Expansion

We have introduced what is known as the *cofactor expansion of a determinant about its first row*. Each entry of the first row is multiplied by the corresponding minor and each such product is multiplied by ± 1 depending on the position of the entry. The signed products were added together. Actually, instead of the first row can use *any* row or column. Here is how: Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

First we assign the sign $(-1)^{i+j}$ to the entry a_{ij} of A. This is a checkerboard pattern of $\pm$'s.

$$\begin{bmatrix} + & - & + & \cdots \\ - & + & - & \cdots \\ + & - & + & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Then we pick a row or column and multiply each entry a_{ij} of it by the corresponding *signed* minor $(-1)^{i+j}M_{ij}$. Lastly, we add all these products.

The signed minor $(-1)^{i+j}M_{ij}$ is called the (i,j) **cofactor**, of A and is denoted by C_{ij}.

$$C_{ij} = (-1)^{i+j}M_{ij}$$

1. **Cofactor Expansion about the ith row** The determinant of A can be expanded about the ith row in terms of the cofactors as follows.

$$\det A = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in}$$

2. **Cofactor Expansion about the jth column** The determinant of A can be expanded about the jth column in terms of the cofactors as follows.

$$\det A = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj}$$

This method of computing determinants by using cofactors is called the **cofactor**, or **Laplace expansion** and it is attributed to Vandermonde and to Laplace.

Properties of Determinants

1. A and its transpose have the same determinant, $\det(A) = \det(A^T)$. For example,
$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}$$

2. Let B be obtained from A by multiplying one of its rows (or columns) by a nonzero constant. Then $\det(B) = k \det(A)$. For example,
$$\begin{vmatrix} a_1 & a_2 & a_3 \\ kb_1 & kb_2 & kb_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = k \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}, \quad \begin{vmatrix} a_1 & a_2 & ka_3 \\ b_1 & b_2 & kb_3 \\ c_1 & c_2 & kc_3 \end{vmatrix} = k \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}$$

3. Let B be obtained from A by interchanging any two rows (or columns). Then $\det(B) = -\det(A)$. For example,
$$\begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = - \begin{vmatrix} b_1 & b_2 & b_3 \\ a_1 & a_2 & a_3 \\ c_1 & c_2 & c_3 \end{vmatrix}, \quad \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = - \begin{vmatrix} a_3 & a_2 & a_1 \\ b_3 & b_2 & b_1 \\ c_3 & c_2 & c_1 \end{vmatrix}$$

4. Let B be obtained from A by adding a multiple of one row (or column) to another. Then $\det(B) = \det(A)$. For example,
$$\begin{vmatrix} a_1 & a_2 & a_3 \\ ka_1 + b_1 & ka_2 + b_2 & ka_3 + b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}$$

Note that

1. Elimination $R_i + cR_j \to R_i$, does not change the determinant.

2. Scaling, $cR_i \to R_i$, scales the determinant by c.

3. Interchange, $R_i \leftrightarrow R_j$, changes the sign of the determinant.

The properties of determinants can be used to compute a determinant as follows. We convert it to triangular form by Gauss elimination and then multiply the diagonal entries of the triangular form.

Determinants by Row Reduction

$$\begin{vmatrix} 1 & 2 & 3 & -1 & 8 \\ 0 & 0 & 4 & 2 & -1 \\ 0 & -5 & 5 & 3 & 7 \\ 0 & 0 & 0 & 1 & 6 \\ 1 & 2 & 3 & -2 & -9 \end{vmatrix} = \begin{vmatrix} 1 & 2 & 3 & -1 & 8 \\ 0 & 0 & 4 & 2 & -1 \\ 0 & -5 & 5 & 3 & 7 \\ 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & -1 & -17 \end{vmatrix} \qquad \boxed{-R_1 + R_5 \to R_5}$$

$$= - \begin{vmatrix} 1 & 2 & 3 & -1 & 8 \\ 0 & -5 & 5 & 3 & 7 \\ 0 & 0 & 4 & 2 & -1 \\ 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & -1 & -17 \end{vmatrix} \qquad \boxed{R_2 \leftrightarrow R_3}$$

$$= - \begin{vmatrix} 1 & 2 & 3 & -1 & 8 \\ 0 & -5 & 5 & 3 & 7 \\ 0 & 0 & 4 & 2 & -1 \\ 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 0 & -11 \end{vmatrix} \qquad \boxed{R_4 + R_5 \to R_5}$$

$$= 1 \cdot (-5) \cdot 4 \cdot 1 \cdot (-11)$$
$$= -220$$

1.13 Cramer's Rule

Let $A\mathbf{x} = \mathbf{b}$ be a square system with

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$$

Let A_i denote the matrix obtained from A by replacing the ith column with $\mathbf{b}$.

$$A_i = \begin{bmatrix} a_{11} & \cdots & a_{1,i-1} & b_1 & a_{1,i+1} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{n,i-1} & b_n & a_{n,i+1} & \cdots & a_{nn} \end{bmatrix}$$

Cramer's Rule gives *an explicit formula* for the solution of a consistent square system.

Cramer's Rule. If $\det(A) \neq 0$, then the system $A\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x} = (x_1, \ldots, x_n)$ given by

$$x_1 = \frac{\det(A_1)}{\det(A)}, \quad x_2 = \frac{\det(A_2)}{\det(A)}, \ . \quad .. \ , \ x_n = \frac{\det(A_n)}{\det(A)}$$

Example 1.13.1. Use Cramer's Rule to solve the system.

$$\begin{aligned} x_1 + x_2 - x_3 &= 2 \\ x_1 - x_2 + x_3 &= 3 \\ -x_1 + x_2 + x_3 &= 4 \end{aligned}$$

Solution: We have

$$\det(A) = \begin{vmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{vmatrix} = -4, \quad \det(A_1) = \begin{vmatrix} 2 & 1 & -1 \\ 3 & -1 & 1 \\ 4 & 1 & 1 \end{vmatrix} = -10$$

$$\det(A_2) = \begin{vmatrix} 1 & 2 & -1 \\ 1 & 3 & 1 \\ -1 & 4 & 1 \end{vmatrix} = -12, \quad \det(A_3) = \begin{vmatrix} 1 & 1 & 2 \\ 1 & -1 & 3 \\ -1 & 1 & 4 \end{vmatrix} = -14$$

Hence,

$$x_1 = \frac{\det(A_1)}{\det(A)} = \frac{5}{2}, \quad x_2 = \frac{\det(A_2)}{\det(A)} = 3, \quad x_3 = \frac{\det(A_3)}{\det(A)} = \frac{7}{2}$$

1.14 Matrix Inverse

Definition An $n \times n$ matrix A is **invertible**, if there exists a matrix B such that

$$AB = I \qquad \text{and} \qquad BA = I$$

In such case B is called an **inverse** of A. If no such B exists for A, then we say that A is **noninvertible**. Another name for invertible is **nonsingular** and another name for noninvertible is **singular**.

Note that the definition forces B to be square of size n (why?).

Theorem 1.14.1. *An invertible matrix has only one inverse.*

Proof: Suppose that the invertible matrix A has two inverses B and C. Then

$$B = BI_n = B(AC) = (BA)C = I_nC = C$$

Therefore, $B = C$.

The unique inverse of an invertible matrix A is denoted by A^{-1}. So

$$AA^{-1} = I \qquad \text{and} \qquad A^{-1}A = I$$

Next we see how to compute the inverse of an invertible matrix A. The idea is simple: If A^{-1} has unknown columns $\mathbf{x}_i$, then $AA^{-1} = I$ takes the form

$$[A\mathbf{x}_1 \ \cdots \ A\mathbf{x}_n] = [\mathbf{e}_1 \ \cdots \ \mathbf{e}_n]$$

This matrix equation splits into n linear systems

$$A\mathbf{x}_1 = \mathbf{e}_1, \ \ldots \ , \ A\mathbf{x}_n = \mathbf{e}_n$$

which we solve to find each column $\mathbf{x}_i$ of A^{-1}. These systems have the same coefficient matrix A. Solving each system separately would amount into $n-1$ unnecessary row reductions of A. It is smarter to solve the systems simultaneously, by simply row reducing the matrix

$$[A : I]$$

If we get a matrix of the form $[I : B]$, then the ith column of B would be $\mathbf{x}_i$. Thus, $B = A^{-1}$. So, in order to compute A^{-1}, we just row reduce $[A : I]$.

Example 1.14.1. Compute, if possible, A^{-1}, for $A = \begin{bmatrix} 1 & 0 & -1 \\ 3 & 4 & -2 \\ 3 & 5 & -2 \end{bmatrix}$.

Solution: We row reduce $[A : I]$.

$$\begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 3 & 4 & -2 & 0 & 1 & 0 \\ 3 & 5 & -2 & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 4 & 1 & -3 & 1 & 0 \\ 0 & 5 & 1 & -3 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 4 & 1 & -3 & 1 & 0 \\ 0 & 0 & -\frac{1}{4} & \frac{3}{4} & -\frac{5}{4} & 1 \end{bmatrix} \sim$$

$$\begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 4 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1 & -3 & 5 & -4 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & -2 & 5 & -4 \\ 0 & 4 & 0 & 0 & -4 & 4 \\ 0 & 0 & 1 & -3 & 5 & -4 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & -2 & 5 & -4 \\ 0 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -3 & 5 & -4 \end{bmatrix}$$

Therefore,

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

It is easy to check that $AA^{-1} = I_3$ and that $A^{-1}A = I_3$.

Example 1.14.2. Compute, if possible, A^{-1}, for $A = \begin{bmatrix} 1 & 0 & -1 \\ 3 & 4 & -2 \\ -3 & -4 & 2 \end{bmatrix}$.

Solution: We row reduce $[A : I]$.

$$\begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 3 & 4 & -2 & 0 & 1 & 0 \\ -3 & -4 & 2 & 0 & 0 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 4 & 1 & -3 & 1 & 0 \\ 0 & -4 & -1 & 3 & 0 & 1 \end{bmatrix}$$

$$\sim \begin{bmatrix} 1 & 0 & -1 & 1 & 0 & 0 \\ 0 & 4 & 1 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

There is no inverse, because we cannot proceed to get the identity matrix on the left hand side, because there is a row of zeros in the left half.

Theorem 1.14.2 (Properties of Inversion)**.** *Let A and B be invertible $n \times n$ matrices and let c be a nonzero scalar. Then*

1. *AB is invertible and*
$$(AB)^{-1} = B^{-1}A^{-1}$$
2. *A^{-1} is invertible and*
$$(A^{-1})^{-1} = A$$
3. *cA is invertible and*
$$(cA)^{-1} = \frac{1}{c}A^{-1}$$
4. *A^T is invertible and*
$$(A^T)^{-1} = \left(A^{-1}\right)^T$$

Cancellation Laws

Recall that $AB = AC$ does not imply that $B = C$. However, if A is invertible, then the implication is true.

Theorem 1.14.3. *Let A, B, and C be $n \times n$ matrices and A is invertible. Then the cancellation laws hold:*

$$AB = AC \Rightarrow B = C, \qquad BA = CA \Rightarrow B = C$$

Proof: Let $AB = AC$. Since A^{-1} exists, we can multiply on the left by A^{-1} to get

$$A^{-1}(AB) = A^{-1}(AC) \Rightarrow (A^{-1}A)B = (A^{-1}A)C \Rightarrow IB = IC \Rightarrow B = C$$

The second implication is proved similarly.

Determinants and Inversion

Cauchy's Theorem. *The determinant of a product of two $n \times n$ matrices is the product of the determinants of the factors.*

$$\det(AB) = \det(A)\det(B)$$

Cauchy's Theorem 1.14 has the following implication.

Theorem 1.14.4. *A square matrix is invertible if and only if its determinant is nonzero.*

Furthermore, if A is invertible, then

$$\det(A^{-1}) = \frac{1}{\det(A)}$$

Invertibility and Linear Systems

Theorem 1.14.5. *Let A be an invertible matrix, so $\det(A) \neq 0$. Then*

1. *$A\mathbf{x} = \mathbf{b}$ has a unique solution given by*

$$\mathbf{x} = A^{-1}\mathbf{b}$$

2. $A\mathbf{x} = \mathbf{0}$ *has only the trivial solution.*

Theorem 1.14.6. *Let A be a $n \times n$ matrix. Then the following are equivalent.*

1. $\det(A) = 0$
2. $A\mathbf{x} = \mathbf{0}$ *has nontrivial solutions*

The Adjoint of a Square Matrix

Definition Let A be an $n \times n$ matrix. The matrix whose (i, j) entry is the cofactor C_{ij} of A is the **matrix of cofactors** of A. Its transpose is the **adjoint of** A and it is denoted by $\text{Adj}(A)$.

$$\text{Adj}(A) = \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix}$$

Example 1.14.3. Find the adjoint of A, where

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

Solution: The cofactors of A are

$$\begin{aligned} C_{11} &= 9, \quad C_{12} = -2, \quad C_{13} = 15 \\ C_{21} &= -6, \quad C_{22} = 7, \quad C_{23} = 10 \\ C_{31} &= -10, \quad C_{32} = 6, \quad C_{33} = -11 \end{aligned}$$

Hence,

$$\text{Adj}(A) = [C_{ij}]^T = \begin{bmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \end{bmatrix} = \begin{bmatrix} 9 & -6 & -10 \\ -2 & 7 & 6 \\ 15 & -10 & -11 \end{bmatrix}$$

Adjoint and Inverse

Theorem 1.14.7. *1. Let A be an $n \times n$ matrix. Then*

$$A\,\mathrm{Adj}(A) = \det(A) I_n = \mathrm{Adj}(A)\,A$$

2. Let A be an invertible matrix. Then

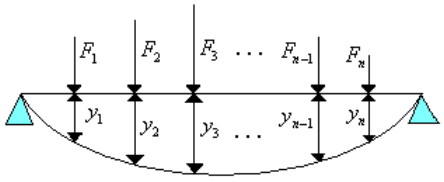
$$A^{-1} = \frac{1}{\det(A)}\mathrm{Adj}(A)$$

Example 1.14.4. For the above A, we have $\det(A) = 17$. Hence,

$$A^{-1} = \frac{1}{\det(A)}\mathrm{Adj}(A) = \frac{1}{17}\begin{bmatrix} 9 & -6 & -10 \\ -2 & 7 & 6 \\ 15 & -10 & -11 \end{bmatrix} = \begin{bmatrix} \frac{9}{17} & -\frac{6}{17} & -\frac{10}{17} \\ -\frac{2}{17} & \frac{7}{17} & \frac{6}{17} \\ \frac{15}{17} & -\frac{10}{17} & -\frac{11}{17} \end{bmatrix}$$

1.15 Application to Stiffness of Beam

We consider an elastic beam supported on the edges. We choose points $P_1, \ldots, P_n$ on which parallel forces $F_1, \ldots, F_n$ are applied causing displacements $y_1, \ldots, y_n$.



We assume the **principle of linear superposition**, or **Hooke's law**:

1. If two systems of forces are applied, then the corresponding displacements are added.

2. If the magnitudes of all forces are multiplied by a scalar c, then the displacements are multiplied by c.

Let a_{ik} be the displacement of P_i under the action of the unit force at P_k. Then under the action of all the forces, the displacements are given by the formulae

$$\sum_{k=1}^{n} a_{ik}F_k = y_i, \quad i = 1, \ldots, n \tag{1.9}$$

If A is the matrix $A = [a_{ik}]$, $\mathbf{y}$ is the vector of displacements $\mathbf{y} = (y_1, \ldots, y_n)$, and $\mathbf{F}$ is the vector of forces, $\mathbf{F} = (F_1, \ldots, F_n)$, then equations (1.9) become

$$A\mathbf{F} = \mathbf{y}$$

Matrix A is called the **flexibility matrix**. Given the flexibility matrix and the displacements, we can calculate the forces F_i by inverting A.

$$\mathbf{F} = A^{-1}\mathbf{y}$$

The inverse A^{-1} is called the **stiffness matrix**.

1.16 Matrix Transformations

A **matrix transformation** $T : \mathbf{R}^n \to \mathbf{R}^m$, is a map from $\mathbf{R}^n$ to $\mathbf{R}^m$ for which the images $T(\mathbf{x})$ of n-vectors $\mathbf{x}$ are of the form $A\mathbf{x}$ for a fixed $m \times n$ matrix A. In other words, there is an $m \times n$ matrix A such that

$$T(\mathbf{x}) = A\mathbf{x}$$

for all $\mathbf{x}$ in $\mathbf{R}^n$. The matrix A is called the **standard matrix** of T.

Example 1.16.1. Find the standard matrix A of the transformation $T : \mathbf{R}^2 \to \mathbf{R}^3$ given by the formula: $T\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 - 4x_2 \\ 3x_1 + 5x_2 \\ 2x_1 \end{bmatrix}$.

Answer: $A = \begin{bmatrix} 1 & -4 \\ 3 & 5 \\ 2 & 0 \end{bmatrix}$.

Example 1.16.2. Find a formula for the image of any vector $\mathbf{x}$ for the matrix transformation $T : \mathbf{R}^3 \to \mathbf{R}^2$ with standard matrix $A = \begin{bmatrix} 3 & -7 & 8 \\ 2 & 1 & -4 \end{bmatrix}$.

Also, find the image of the vector $\begin{bmatrix} -2 \\ 3 \\ 5 \end{bmatrix}$.

Solution:

$$T\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 & -7 & 8 \\ 2 & 1 & -4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3x_1 - 7x_2 + 8x_3 \\ 2x_1 + x_2 - 4x_3 \end{bmatrix}$$

and

$$T\begin{bmatrix} -2 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 3 & -7 & 8 \\ 2 & 1 & -4 \end{bmatrix}\begin{bmatrix} -2 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 13 \\ -21 \end{bmatrix}$$

The set of all images of a matrix transformation is called the **image** or **image set** of it.

Example 1.16.3 (Image of Matrix Transformation)**.** Consider the matrix transformation $T\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 + 2x_2 + 2x_3 \\ x_2 + x_3 \end{bmatrix}$. Which of the vectors $\mathbf{v}_1 = \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ is in its image?

Solution: We need to find vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ such that $T(\mathbf{u}_1) = \mathbf{v}_1$ and $T(\mathbf{u}_2) = \mathbf{v}_2$. So, we need to solve the linear systems

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 0 & 1 & 1 \end{bmatrix}\mathbf{u}_1 = \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 0 & 1 & 1 \end{bmatrix}\mathbf{u}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The following reduction

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 2 & -1 \\ 0 & 1 & 1 & -1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

shows that the first system is solvable. Hence, $\mathbf{v}_1$ is in the image of T.

On the other hand, the reduction

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

shows that and the second system is not solvable. Therefore, $\mathbf{v}_2$ is not in the image of T.

Example 1.16.4. In the last example we saw that the vector $\mathbf{v}_1 = \begin{bmatrix} 0 & -1 & -1 \end{bmatrix}^T$ is in the image of the matrix transformation $T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 + 2x_2 + 2x_3 \\ x_2 + x_3 \end{bmatrix}$. Find all vectors that map to $\mathbf{v}_1$. Then specify one such vector $\mathbf{u}$.

Solution: The augmented matrix of the linear system $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 0 & 1 & 1 \end{bmatrix} \mathbf{u}_1 = \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}$ reduced to $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$. So the general solution is $x_1 = 1, x_2 = -1 - r, x_3 = r$, r any. So all the vectors of the form $\begin{bmatrix} 1 & -1 - r & r \end{bmatrix}^T$ map to $\mathbf{v}_1$. One one these vectors is $\mathbf{u} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}^T$ (where we chose $r = 0$).

Linearity of Matrix Transformations

The most important property of any matrix transformation $T : \mathbf{R}^n \to \mathbf{R}^m$ is that it is **linear**. That is it satisfies the following **linearity** conditions:

For all vectors $\mathbf{u}$ and $\mathbf{v}$ of V and any scalar c, we have

1. $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$
2. $T(c\mathbf{u}) = cT(\mathbf{u})$

The addition in $\mathbf{u}+\mathbf{v}$ is addition in $\mathbf{R}^n$, whereas the addition in $T(\mathbf{u})+T(\mathbf{v})$ is addition in $\mathbf{R}^m$. Likewise, scalar multiplications $c\mathbf{u}$ and $cT(\mathbf{u})$ occur in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively.

Linearity follows from the properties of matrix multiplication. If A is the standard matrix of T, then

$$T(\mathbf{u} + \mathbf{v}) = A(\mathbf{u} + \mathbf{v}) = A(\mathbf{u}) + A(\mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$$

and

$$T(c\mathbf{u}) = A(c\mathbf{u}) = c(A\mathbf{u}) = cT(\mathbf{u})$$

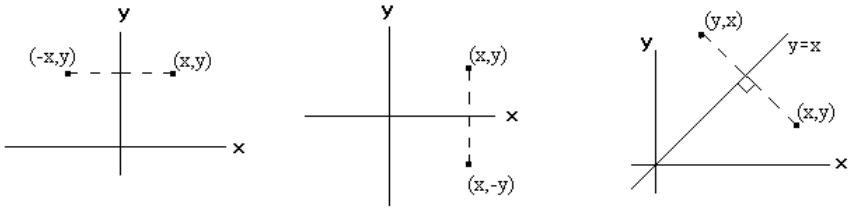
1.17 Transformations of Computer Graphics

Some matrix transformations of the plane $T : \mathbf{R}^2 \to \mathbf{R}^2$, or of the space $T : \mathbf{R}^3 \to \mathbf{R}^3$ have interesting geometric interpretations. They are reflections, compressions, rotations, shears, projections, etc. Such transformations are useful all around and they are indispensable in computer graphics.

Reflections

In two dimensions it is easy to see that **reflections** about the y-axis, the x-axis, the line $y = x$, and the origin are matrix transformations with respective standard matrices

$$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

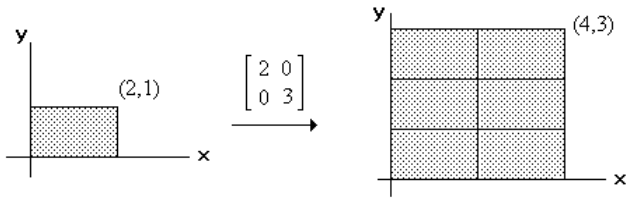


Compressions-Expansions

Compressions and expansions are scalings along the coordinate axes. If c and d are positive scalars then (cx, y), (x, cy), and (cx, dy) represent scalings along the x-axis, along the y-axis, and along both axes. These scalings define matrix transformations with corresponding standard matrices

$$\begin{bmatrix} c & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix}, \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix}$$

If the scalars are less than 1, we have **compressions**. If they are greater than 1 we have **expansions**.



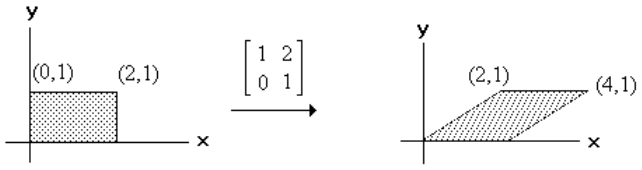
Shears

A **shear** along the x-axis is a transformation of the form

$$T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & c \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x+cy \\ y \end{bmatrix}$$

Each point is moved along the x-direction by an amount proportional to the distance from the x-axis. We also have shears along the y-axis

$$T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ cx+y \end{bmatrix}$$



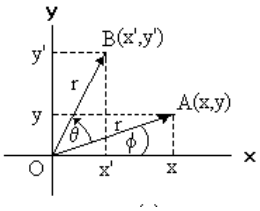
Rotations

Another common plane transformation is **rotation** about any point in the plane. We are interested in rotations about the origin.

Show that the transformation $T : \mathbf{R}^2 \to \mathbf{R}^2$ defined by

$$T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \end{bmatrix}$$

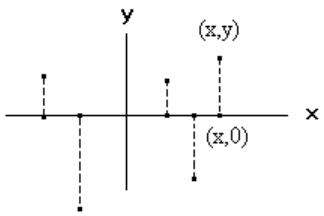
rotates each vector counterclockwise θ radians about the origin.

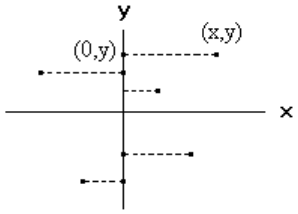


Projections

Projections of the plane onto a line are also transformations of the plane. The most important of these are the orthogonal projections onto lines through the origin, especially onto the axes. For example, the **orthogonal projections** onto the x-axis and the y-axis are matrix transformations given by

$$T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}, \quad T\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$





Application to Computer Graphics

We can use matrix-vector products to compute images of matrix transformations of the plane. Here we find a set of images by using matrix multiplication.

Example 1.17.1. Find the images of the vertices

$$(1.0, 0), (0.7, 0.7), (0, 1.0), (-0.7, 0.7),$$

$$(-1.0, 0), (-0.7, -0.7), (0, -1.0), (0.7, -0.7)$$

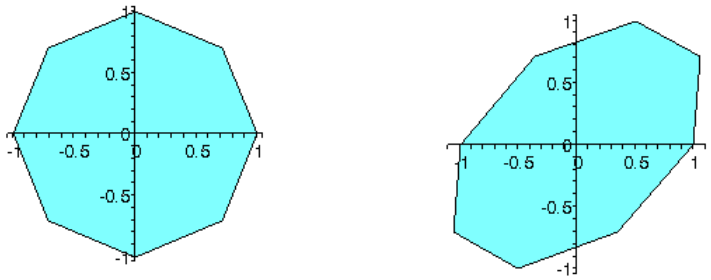
of an octagon under the shear transformation $T(\mathbf{x}) = A\mathbf{x}$, where

$$A = \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix}$$

Solution: We form a 2×8 matrix P with columns the vertices of the octagon and compute the product AP.

$$AP = \begin{bmatrix} 1 & 0.5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.0 & 0.7 & 0 & -0.7 & -1.0 & -0.7 & 0 & 0.7 \\ 0 & 0.7 & 1.0 & 0.7 & 0 & -0.7 & -1.0 & -0.7 \end{bmatrix}$$
$$= \begin{bmatrix} 1.0 & 1.05 & 0.5 & -0.35 & -1.0 & -1.05 & -0.5 & 0.35 \\ 0 & 0.7 & 1.0 & 0.7 & 0 & -0.7 & -1.0 & -0.7 \end{bmatrix}$$

The columns of AP are the transformed vertices of the octagon.



1.18 The Dot Product

The **dot product** $\mathbf{u} \cdot \mathbf{v}$ of two n-vectors $\mathbf{u} = (u_1, ..., u_n)$ and $\mathbf{v} = (v_1, ..., v_n)$ is the matrix-vector product

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v}$$

The matrix in this case is the row vector obtained by transposing $\mathbf{u}$. In terms of components the dot product is the *number*

$$\mathbf{u} \cdot \mathbf{v} = [u_1 \cdots u_n] \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = u_1 v_1 + \cdots + u_n v_n \tag{1.10}$$

If the dot product of two vectors is zero, we call these vectors **orthogonal**.

Note that in equation (1.10) for convenience we identified a 1×1 matrix $[a]$ with its single entry a.

Example 1.18.1. Let $\mathbf{u} = (-3, 2, 1)$, $\mathbf{v} = (4, -1, 5)$, and $\mathbf{w} = (-2, 1, -8)$.

1. Find $\mathbf{u} \cdot \mathbf{v}$.
2. Are $\mathbf{u}$ and $\mathbf{w}$ are orthogonal?

Solution:

(1) We have

$$\mathbf{u} \cdot \mathbf{v} = \begin{bmatrix} -3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ -1 \\ 5 \end{bmatrix} = (-3)\,4 + 2\,(-1) + (1)\,(5) = -9$$

(2) Vectors $\mathbf{u}$ and $\mathbf{w}$ are orthogonal, because

$$\mathbf{u} \cdot \mathbf{w} = (-3, 2, 1) \cdot (-2, 1, -8) = 0$$

Definition The **norm**, or **length**, or **magnitude** of an n-vector $\mathbf{u} = (u_1, \ldots, u_n)$ is the positive number (or zero)

$$\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \left(u_1^2 + \cdots + u_n^2\right)^{\frac{1}{2}}$$

The (**Euclidean**) **distance** between two n-vectors $\mathbf{u}$ and $\mathbf{v}$ is

$$\|\mathbf{u} - \mathbf{v}\|$$

A n-vector is a **unit** vector, if its norm is 1.

Example 1.18.2. Let $\mathbf{v} = (1, 2, -3, 1)$ and $\mathbf{u} = \left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\right)$.

1. Find the length of $\mathbf{v}$.
2. Find the distance between $\mathbf{v}$ and $\mathbf{u}$.
3. Is $\mathbf{u}$ a unit vector?

Solution: We have

(1) $\|\mathbf{v}\| = \left(1^2 + 2^2 + (-3)^2 + 1^2\right)^{\frac{1}{2}} = \sqrt{15}$

(2) $\|\mathbf{v}-\mathbf{u}\| = \left\|\left(\frac{1}{2}, \frac{5}{2}, -\frac{7}{2}, \frac{3}{2}\right)\right\| = \sqrt{21}$

(3) $\|\mathbf{u}\| = \left\|\left(\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\right)\right\| = 1$. So, $\mathbf{u}$ is a unit vector.

The dot product for plane and space vectors is related to the length and angle between the vectors by the following formula

$$\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\ \|\mathbf{v}\|\cos\theta \tag{1.11}$$

This can be seen by using the *law of cosines* on the triangle OPQ with $OP = \mathbf{u}$ and $OQ = \mathbf{v}$.

$$\begin{aligned}\|\mathbf{u}\|\ \|\mathbf{v}\|\ \cos\theta &= \frac{1}{2}\left(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|PQ\|^2\right)\\ &= \frac{1}{2}\left(\sum_{i=1}^{3} u_i^2 + \sum_{i=1}^{3} v_i^2 - \sum_{i=1}^{3}(v_i-u_i)^2\right)\\ &= \sum_{i=1}^{3} u_i v_i = \mathbf{u}\cdot\mathbf{v}\end{aligned}$$

Main Properties of Dot Product

Theorem 1.18.1. *Let* $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ *be any* n*-vectors and* c *be any scalar. Then*

1. $\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}$ ***(Symmetry)***
2. $\mathbf{u}\cdot(\mathbf{v}+\mathbf{w}) = \mathbf{u}\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{w}$ ***(Additivity)***
3. $c\,(\mathbf{u}\cdot\mathbf{v}) = (c\mathbf{u})\cdot\mathbf{v} = \mathbf{u}\cdot(c\mathbf{v})$ ***(Homogeneity)***
4. $\mathbf{u}\cdot\mathbf{u} \geq 0$. *Also,* $\mathbf{u}\cdot\mathbf{u} = 0$ *if and only if* $\mathbf{u} = \mathbf{0}$. ***(Positive Definiteness)***
5. ***(Pythagorean Theorem)*** $\mathbf{u}$ *and* $\mathbf{v}$ *are orthogonal if and only if*
$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$
6. ***(Cauchy-Bunyakovsky-Schwarz Inequality)***
$$|\mathbf{u}\cdot\mathbf{v}| \leq \|\mathbf{u}\|\ \|\mathbf{v}\| \tag{1.12}$$

Inner Product

Definition An **inner product** on a (real) vector space V is a function that to each pair of vectors $\mathbf{u}$ and $\mathbf{v}$ of V associates a real number, denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$.

$$\langle \_, \_ \rangle : V \times V \to \mathbf{R}, \qquad (\mathbf{u}, \mathbf{v}) \to \langle \mathbf{u}, \mathbf{v} \rangle$$

This function satisfies the following properties, or **axioms**.

For any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ of V and any scalar c, we have

1. $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$ (**Symmetry**)
2. $\langle \mathbf{u} + \mathbf{w}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle$ (**Additivity**)
3. $\langle c\mathbf{u}, \mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$ (**Homogeneity**)
4. $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$. Furthermore, $\langle \mathbf{u}, \mathbf{u} \rangle = 0$ if and only if $\mathbf{u} = \mathbf{0}$. (**Positivity**)

A real vector space with an inner product is called an **inner product space**.

Theorem 1.18.2. *Let* $\mathbf{u}$, $\mathbf{v}$, *and* $\mathbf{w}$ *be any vectors in an inner product space and let* c *be any scalar. Then*

1. $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$
2. $\langle \mathbf{u}, c\mathbf{v} \rangle = c\langle \mathbf{u}, \mathbf{v} \rangle$
3. $\langle \mathbf{u} - \mathbf{w}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{w}, \mathbf{v} \rangle$
4. $\langle \mathbf{u}, \mathbf{v} - \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{w} \rangle$
5. $\langle \mathbf{0}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{0} \rangle = 0$

Example 1.18.3. The following are examples of inner products.

1. The dot product of $\mathbf{R}^n$ is an inner product.
2. (Weighted Dot Product) Let $w_1, \ldots, w_n$ be any *positive* numbers and let $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ be any n-vectors. The following defines an inner product in $\mathbf{R}^n$.

$$\langle \mathbf{u}, \mathbf{v} \rangle = w_1 u_1 v_1 + \cdots + w_n u_n v_n \tag{1.13}$$

3. Let A and B be 2×2 matrices with real entries.

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}$$

The following function defines an inner product in M_{22}.

$$\langle A, B \rangle = a_1b_1 + a_2b_2 + a_3b_3 + a_4b_4$$

4. Let $f(x)$ and $g(x)$ be in $C[a, b]$, the vector space of the continuous real-valued functions defined on $[a, b]$. Then the following defines an inner product on $C[a, b]$.

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,dx$$

Length and Orthogonality

Let V be an inner product space. Two vectors $\mathbf{u}$ and $\mathbf{v}$ are called **orthogonal** if their inner product is zero.

$$\mathbf{u} \text{ and } \mathbf{v} \text{ are orthogonal if } \langle \mathbf{u}, \mathbf{v} \rangle = 0$$

The **norm** (or **length**, or **magnitude**) of $\mathbf{v}$ is the nonnegative number $\|\mathbf{v}\|$, defined by

$$\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} \tag{1.14}$$

We also define the **distance**, $d(\mathbf{u}, \mathbf{v})$, between two vectors $\mathbf{u}$ and $\mathbf{v}$ by

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| \tag{1.15}$$

Note that

$$d(\mathbf{0}, \mathbf{v}) = d(\mathbf{v}, \mathbf{0}) = \|\mathbf{v}\|$$

A vector with norm 1 is called a **unit** vector. The set $\mathbf{S}$ of all unit vectors of V is called the **unit circle** or the **unit sphere**.

$$\mathbf{S} = \{\mathbf{v}\,,\ \mathbf{v} \in V \text{ and } \|\mathbf{v}\| = 1\} \tag{1.16}$$

The norm in an inner product space V satisfies the following basic properties.

Theorem 1.18.3. *For all vectors u and v of V and all scalars c, we have*

1. $\|c\mathbf{u}\| = |c| \, \|\mathbf{u}\|$
2. $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ *(the **Triangle Inequality**)*
3. $\|\mathbf{u}\| \geq 0$ *and* $\|\mathbf{u}\| = 0$ *if and only if* $\mathbf{u} = \mathbf{0}$

Example 1.18.4. Consider the vector space $C[-\pi, \pi]$ with the integral inner product $\langle f, g \rangle = \int_{-\pi}^{\pi} f(x) \, g(x) \, dx$.

1. Are the functions $\sin(x)$ and $\sin(2x)$ orthogonal in $C[-\pi, \pi]$?
2. What is the norm of $\sin(2x)$ with respect to this inner product?

Solution:

1. We have

$$\begin{aligned} \langle \sin(x), \sin(2x) \rangle &= \int_{-\pi}^{\pi} \sin(x) \sin(2x) \, dx = \frac{1}{2} \int_{-\pi}^{\pi} (\cos x - \cos 3x) \, dx \\ &= \frac{1}{2} \left(\sin x - \frac{1}{3} \sin 3x \right) \Big|_{-\pi}^{\pi} = 0 \end{aligned}$$

   so the functions are orthogonal.

2. The norm is

$$\|\sin(2x)\| = \left(\int_{-\pi}^{\pi} \sin^2(2x) \, dx \right)^{1/2} = \left(\frac{1}{2} \int_{-\pi}^{\pi} (1 - \cos(4x)) \, dx \right)^{1/2} = \sqrt{\pi}$$

Note the trigonometric identities: (used in the above calculations)

$$\sin(a) \sin(b) = \frac{1}{2} (\cos(a - b) - \cos(a + b))$$

$$\sin^2(a) = \frac{1}{2} (1 - \cos 2a)$$

1.19 Linear Transformations

Definition A **linear transformation** or **linear map** from a vector space V to a vector space W is a transformation $T : V \to W$ such that for all vectors $\mathbf{u}$ and $\mathbf{v}$ of V and any scalar c, we have

1. $T(\mathbf{u}+\mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$
2. $T(c\mathbf{u}) = cT(\mathbf{u})$

The addition in $\mathbf{u}+\mathbf{v}$ is addition in V, whereas the addition in $T(\mathbf{u})+T(\mathbf{v})$ is addition in W. Likewise, scalar multiplications $c\mathbf{u}$ and $cT(\mathbf{u})$ occur in V and W, respectively. In the special case where $V = W$, the linear transformation $T : V \to V$ is called a **linear operator** of V.

Examples of Linear Transformations

- Matrix transformations. If A is the matrix of the transformation, then

$$T(\mathbf{x}_1 + \mathbf{x}_2) = A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2 = T(\mathbf{x}_1) + T(\mathbf{x}_2)$$

and

$$T(c_1\mathbf{x}_1) = A(c_1\mathbf{x}_1) = c_1A\mathbf{x}_1 = c_1T(\mathbf{x}_1)$$

- The special matrix transformations with matrices

$$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

are linear and represent reflection about the y-axis and the x-axis, reflection about the origin and rotation by θ radians about the origin.

- $T : M_{22} \to P_3$, $T\begin{bmatrix} a & b \\ c & d \end{bmatrix} = d + cx + (b-a)x^3$ is linear.

$$T\left(\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} + \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}\right) = T\begin{bmatrix} a_1+a_2 & b_1+b_2 \\ c_1+c_2 & d_1+d_2 \end{bmatrix}$$
$$= (d_1+d_2) + (c_1+c_2)x + \{(b_1+b_2) - (a_1+a_2)\}x^3$$
$$= \{d_1 + c_1x + (b_1-a_1)x^3\} + \{d_2 + c_2x + (b_2-a_2)x^3\}$$
$$= T\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} + T\begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix}$$

and

$$T\left(c\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}\right) = T\begin{bmatrix} ca_1 & cb_1 \\ cc_1 & cd_1 \end{bmatrix}$$
$$= cd_1 + cc_1x + (cb_1 - ca_1)\,x^3$$
$$= c\left[d_1 + c_1x + (b_1 - a_1)\,x^3\right]$$
$$= cT\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}$$

Example 1.19.1. The transformation $T : M_{22} \to P_3,\ T\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a^2 + bx^3$ is **nonlinear**.

Solution: The verification is left as exercise.

1.20 Eigenvalues

Definition Let A be an $n \times n$ matrix. A nonzero vector $\mathbf{v}$ is called an **eigenvector** of A, if for some scalar λ

$$A\mathbf{v} = \lambda\mathbf{v} \tag{1.17}$$

The scalar λ (which may zero) is called an **eigenvalue** of A *corresponding to* (or *associated with*) the eigenvector $\mathbf{v}$.

Geometrically, if $\mathbf{v}$ is an eigenvector of A, then $\mathbf{v}$ and $A\mathbf{v}$ are on the same line through the origin.

Example 1.20.1. Let

$$A = \begin{bmatrix} 2 & 2 \\ 2 & -1 \end{bmatrix}, \quad \mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

1. Show that $\mathbf{v}_1$ and $\mathbf{v}_2$ are eigenvectors of A.

2. What are the eigenvalues corresponding to $\mathbf{v}_1$ and $\mathbf{v}_2$?

Solution: We have

$$A\mathbf{v}_1 = \begin{bmatrix} 2 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 3\mathbf{v}_1$$

$$A\mathbf{v}_2 = \begin{bmatrix} 2 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} -2 \\ 4 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ -2 \end{bmatrix} = -2\mathbf{v}_2$$

Therefore, $\mathbf{v}_1$ is an eigenvector with corresponding eigenvalue $\lambda = 3$ and $\mathbf{v}_2$ is an eigenvector with corresponding eigenvalue $\lambda = -2$.

Example 1.20.2. Find all the eigenvalues and eigenvectors of A geometrically, if

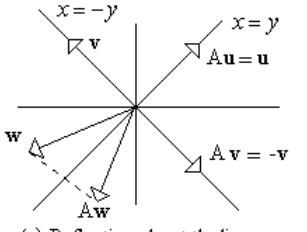
1. $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

2. A is the standard matrix of the rotation by $30°$ in $\mathbf{R}^3$ about the z-axis in the positive direction.

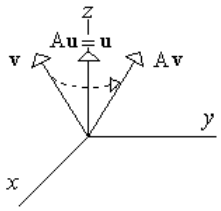
Solution:

1. $A\mathbf{x}$ is the reflection of $\mathbf{x}$ about the line $y = x$. The only vectors that remain on the same line after rotation are the vectors along the lines $y = x$ and $y = -x$. These without the origin are the only eigenvectors. For $\mathbf{v}$ along $y = x$ we have $A\mathbf{v} = 1\mathbf{v}$, so $\mathbf{v}$ is an eigenvector with corresponding eigenvalue 1. For $\mathbf{v}$ along $y = -x$, $A\mathbf{v} = -1\mathbf{v}$, so $\mathbf{v}$ is an eigenvector with corresponding eigenvalue -1.

2. The only vectors that remain on the same line after rotation are all vectors along the z-axis. These without the origin are the only eigenvectors. The corresponding eigenvalue is 1.



(a) Reflection about the line $x = y$



(b) Rotation about the z-axis.

Computation of Eigenvalues

Theorem 1.20.1. *Let A be a square matrix.*

1. *A vector $\mathbf{v}$ is an eigenvector of A corresponding to eigenvalue λ if and only if $\mathbf{v}$ is a nontrivial solution of the system*

$$(A - \lambda I)\mathbf{v} = \mathbf{0} \tag{1.18}$$

2. *A scalar λ is an eigenvalue of A if and only if*

$$\det(A - \lambda I) = 0 \tag{1.19}$$

Equation (1.19) is called the **characteristic equation** of A. The determinant $\det(A - \lambda I)$ is a polynomial of degree n in λ and is called the **characteristic polynomial** of A. The matrix $A - \lambda I$ is called the **characteristic matrix** of A. If an eigenvalue λ is a root of the characteristic equation of multiplicity k, we say that λ has **algebraic multiplicity** k.

Proof of Theorem 1.20.1

1. We have

$$\begin{aligned} A\mathbf{v} = \lambda\mathbf{v} &\Rightarrow A\mathbf{v} = \lambda I\mathbf{v} \\ &\Rightarrow A\mathbf{v} - \lambda I\mathbf{v} = \mathbf{0} \\ &\Rightarrow (A - \lambda I)\mathbf{v} = \mathbf{0} \end{aligned}$$

   Hence, $\mathbf{v}$ is an eigenvector if and only if it is a nontrivial solution of the homogeneous system $(A - \lambda I)\mathbf{v} = \mathbf{0}$.

2. The homogeneous linear system (1.18) has a nontrivial solution if and only if the determinant of the coefficient matrix is zero. Thus, λ is an eigenvalue of A if and only if $\det(A - \lambda I) = 0$.

Eigenspace

Theorem 1.20.2. *Let A be a $n \times n$ and let λ be an eigenvalue of A. Let E_λ be the set that consists of all eigenvectors of A corresponding to λ and the zero n-vector. Then E_λ is a subspace of $\mathbf{R}^n$.*

The subspace E_λ of $\mathbf{R}^n$ mentioned above consisting of the zero vector and the eigenvectors of A with eigenvalue λ is called an **eigenspace** of A. It is the eigenspace with eigenvalue λ. The dimension of E_λ is called the **geometric multiplicity** of λ.

In the next three examples we compute the eigenvalues, the eigenvectors and find bases for each eigenspace of the given matrix A.

Example 1.20.3. $A = \begin{bmatrix} 1 & -1 & -1 \\ -2 & 0 & 4 \\ -2 & 6 & -2 \end{bmatrix}$.

Solution: The characteristic equation is

$$\begin{vmatrix} 1-\lambda & -1 & -1 \\ -2 & 0-\lambda & 4 \\ -2 & 6 & -2-\lambda \end{vmatrix} = -\lambda^3 - \lambda^2 + 30\lambda = -\lambda\,(\lambda - 5)\,(\lambda + 6) = 0$$

Hence, the eigenvalues are

$$\lambda_1 = 0, \quad \lambda_2 = 5, \quad \lambda_3 = -6$$

Next, we find the eigenevectors. For $\lambda_1 = 0$ we have

$$[A - 0I : \mathbf{0}] = \begin{bmatrix} 1 & -1 & -1 & 0 \\ -2 & 0 & 4 & 0 \\ -2 & 6 & -2 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -2 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The general solution is $(2r, r, r)$ for $r \in \mathbf{R}$. Hence,

$$E_0 = \left\{ \begin{bmatrix} 2r \\ r \\ r \end{bmatrix},\ r \in \mathbf{R} \right\} = \text{Span} \left\{ \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \right\}$$

and eigenvector $\mathbf{v}_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}$ defines the basis $\{\mathbf{v}_1\}$ of E_0.

For $\lambda_2 = 5$ we have

$$[A - 5I : \mathbf{0}] = \begin{bmatrix} -4 & -1 & -1 & 0 \\ -2 & -5 & 4 & 0 \\ -2 & 6 & -7 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 1/2 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The general solution is $(-r/2, r, r)$ for $r \in \mathbf{R}$. Hence,

$$E_5 = \left\{ \begin{bmatrix} -r/2 \\ r \\ r \end{bmatrix},\ r \in \mathbf{R} \right\} = \text{Span} \left\{ \begin{bmatrix} -1/2 \\ 1 \\ 1 \end{bmatrix} \right\}$$

Any nonzero vector of E_5 is a basis of E_5. We may choose a fraction free one. So, $\mathbf{v}_2 = \begin{bmatrix} -1 \\ 2 \\ 2 \end{bmatrix}$ defines the basis $\{\mathbf{v}_2\}$ of E_5.

For $\lambda = -6$ we have

$$[A - (-6)\, I : \mathbf{0}] = \begin{bmatrix} 7 & -1 & -1 & 0 \\ -2 & 6 & 4 & 0 \\ -2 & 6 & 4 & 0 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & -1/20 & 0 \\ 0 & 1 & 13/20 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The general solution is $(r/20, -13r/20, r)$ for $r \in \mathbf{R}$. Hence,

$$E_{-6} = \left\{ \begin{bmatrix} r/20 \\ -13r/20 \\ r \end{bmatrix},\ r \in \mathbf{R} \right\} = \text{Span} \left\{ \begin{bmatrix} 1/20 \\ -13/20 \\ 1 \end{bmatrix} \right\}$$

Any nonzero vector of E_{-6} is a basis of E_{-6}. For example, $\mathbf{v}_3 = \begin{bmatrix} 1 \\ -13 \\ 20 \end{bmatrix}$ defines the basis $\{\mathbf{v}_3\}$ of E_{-6}.

Example 1.20.4. $A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Solution: The characteristic equation is

$$\det(A - \lambda I) = \begin{vmatrix} -\lambda & 0 & 1 \\ 0 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{vmatrix} = -\lambda\,(1-\lambda)^2 = 0$$

Hence, the eigenvalues are

$$\lambda_1 = 0 \ , \ \lambda_2 = \lambda_3 = 1$$

Next, we find the eigenevectors. For $\lambda_1 = 0$ we have

$$[A - 0I : \mathbf{0}] = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \sim \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The general solution is $(r, 0, 0)$ for $r \in \mathbf{R}$. Hence,

$$E_0 = \left\{ \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix}, \ r \in \mathbf{R} \right\} = \text{Span} \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

and eigenvector $\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ defines the basis $\{\mathbf{v}_1\}$ of E_0.

For $\lambda_2 = \lambda_3 = 1$ with algebraic multiplicity 2, we have

$$[A - 1I : \mathbf{0}] = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The general solution is (r, s, r) for $r \in \mathbf{R}$. But $(r, s, r) = r(1, 0, 1) + s(0, 1, 0)$, so

$$E_1 = \left\{ \begin{bmatrix} r \\ s \\ r \end{bmatrix}, \ r \in \mathbf{R} \right\} = \text{Span} \left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

The spanning eigenvectors $\mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, $\mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ are linearly independent. So, $\{\mathbf{v}_2, \mathbf{v}_3\}$ is a basis for E_1 and the geometric multiplicity of $\lambda = 1$ is 2.

NOTE If $A = [a_{ij}]$ is a triangular matrix, then so is $A - \lambda I$. Hence, in this case

$$\det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda)$$

We conclude that *the eigenvalues of a triangular matrix are the diagonal entries.*

Example 1.20.5. $A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -4 & 2 \\ 0 & 0 & -2 \end{bmatrix}$.

A is triangular, so the eigenvalues are the diagonal entries $1, -2, -4$. By row reducing $[A - 1I : \mathbf{0}]$, $[A - (-2)I : \mathbf{0}]$, and $[A - (-4)I : \mathbf{0}]$ we get

$$E_1 = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}, \quad E_{-2} = \text{Span}\left\{\begin{bmatrix} 1/3 \\ 1 \\ 1 \end{bmatrix}\right\}, \quad E_{-4} = \text{Span}\left\{\begin{bmatrix} 1/5 \\ 1 \\ 0 \end{bmatrix}\right\}$$

The spanning eigenvectors define bases for the corresponding eigenspaces. We may also use different bases that are free of fractions. For example,

$$E_1 = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}, \quad E_{-2} = \text{Span}\left\{\begin{bmatrix} 1 \\ 3 \\ 3 \end{bmatrix}\right\}, \quad E_{-4} = \text{Span}\left\{\begin{bmatrix} 1 \\ 5 \\ 0 \end{bmatrix}\right\}$$

1.21 Diagonalization

Matrix arithmetic with diagonal matrices is easier than with any other matrices. This is most notable in matrix multiplication. For example, a diagonal matrix D does not mix the components of $\mathbf{x}$ in the product $D\mathbf{x}$.

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2a \\ 3b \end{bmatrix}$$

Also, it does not mix rows of A in a product DA (or columns in AD).

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} = \begin{bmatrix} 2a & 2b & 2c \\ 3d & 3e & 3f \end{bmatrix}$$

Moreover, it is very easy to compute the powers D^k.

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}^k = \begin{bmatrix} 2^k & 0 \\ 0 & 3^k \end{bmatrix}$$

We study matrices that can be transformed to diagonal matrices and take advantage of the easy arithmetic. We use eigenvalues to develop criteria that identify these matrices and we explore their basic properties.

Definition Let A and B be two $n \times n$ matrices. We say that B **is similar to** A if there exists an invertible matrix P such that

$$B = P^{-1}AP$$

Definition If a $n \times n$ matrix A is similar to a *diagonal* matrix D, then it is called **diagonalizable**. We also say that A **can be diagonalized**. This means that there exists an invertible $n \times n$ matrix P such that $P^{-1}AP$ is a diagonal matrix D.

$$P^{-1}AP = D$$

The process of finding matrices P and D is called **diagonalization**. We say that P and D **diagonalize** A.

The answer of how to diagonalize a matrix is provided in the next theorem.

Theorem 1.21.1. *Let A be an $n \times n$ matrix.*

1. *A is diagonalizable if and only if it has n linearly independent eigenvectors.*

2. *If A is diagonalizable with $P^{-1}AP = D$, then the columns of P are eigenvectors of A and the diagonal entries of D are the corresponding eigenvalues.*

3. *If $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ are linearly independent eigenvectors of A with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$, then A can be diagonalized by*

$$P = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_n] \quad \text{and} \quad D = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$$

Theorem 1.21.2. *Let A be a $n \times n$ matrix. The following are equivalent.*

1. *A is diagonalizable.*

2. *$\mathbf{R}^n$ has a basis of eigenvectors of A.*

Example 1.21.1. $A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Solution: We found before that $\lambda_1 = 0$, $\lambda_2 = \lambda_3 = 1$ and

$$E_0 = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}, \quad E_1 = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\right\}$$

A has 3 linearly independent eigenvectors so it is diagonalizable. We may take

$$P = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We may check this by

$$P^{-1}AP = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = D$$

Example 1.21.2. $A = \begin{bmatrix} 1 & -1 & 0 \\ 0 & -4 & 2 \\ 0 & 0 & -2 \end{bmatrix}$.

Solution: We have found that $\lambda_1 = 1$, $\lambda_2 = -2$, $\lambda_3 = -4$ and

$$E_1 = \text{Span}\left\{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right\}, \quad E_{-2} = \text{Span}\left\{\begin{bmatrix} \frac{1}{3} \\ 1 \\ 1 \end{bmatrix}\right\}, \quad E_{-4} = \text{Span}\left\{\begin{bmatrix} \frac{1}{5} \\ 1 \\ 0 \end{bmatrix}\right\}$$

A has 3 linearly independent eigenvectors so it is diagonalizable. We may take

$$P = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{5} \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -4 \end{bmatrix}$$

Theorem 1.21.3. *Let $\lambda_1, \ldots, \lambda_l$ be **any** distinct eigenvalues of an $n \times n$ matrix A.*

1. *Then any corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_l$ are linearly independent.*
2. *If $\mathcal{B}_1, \ldots, \mathcal{B}_l$ are bases for the corresponding eigenspaces, then*

$$\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_l$$

is linearly independent.

3. *Let l be the number of* all *distinct eigenvalues of A. Then A is diagonalizable, if and only if $\mathcal{B}$ in part 2 has exactly n elements.*

Example 1.21.3. Is $A = \begin{bmatrix} 1 & 0 & 3 \\ 1 & -1 & 2 \\ -1 & 1 & -2 \end{bmatrix}$ diagonalizable?

Solution: We have found that $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = -2$ and

$$E_0 = \text{Span}\left\{ \begin{bmatrix} -3 \\ -1 \\ 1 \end{bmatrix} \right\}, \quad E_{-2} = \text{Span}\left\{ \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} \right\}$$

This time A has at most 2 (< 3) linearly independent eigenvectors, so it is **not** diagonalizable, by part 2 of the theorem.

Powers of Diagonalizable Matrices

Let A be diagonalizable $n \times n$ matrix, diagonalized by P and D, so $A = PDP^{-1}$. We have $A^2 = (PDP^{-1})(PDP^{-1}) = PD^2P^{-1}$. We iterate to get

$$A^k = PD^kP^{-1}$$

Example 1.21.4. Find a formula for A^k, $k = 0, 1, 2, \ldots$, where

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 3 & 0 & 3 \end{bmatrix}$$

Solution: A has eigenvalues $0, 2, 4$ and the corresponding basic eigenvectors $(-1, 0, 1)$, $(0, 1, 0)$, $(1, 0, 3)$ are linearly independent. Hence,

$$\begin{aligned} A^k &= \begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{bmatrix}^k \begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2^k & 0 \\ 0 & 0 & 4^k \end{bmatrix} \begin{bmatrix} -\frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 1 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} \end{bmatrix} \\ &= \begin{bmatrix} 4^{k-1} & 0 & 4^{k-1} \\ 0 & 2^k & 0 \\ 3 \cdot 4^{k-1} & 0 & 3 \cdot 4^{k-1} \end{bmatrix} \end{aligned}$$

An Important Change of Variables

Let us now discuss an idea that is in the core of most applications of diagonalization. Let A be diagonalizable, diagonalized by P and D. Often a matrix-vector equation $f(A, \mathbf{x}) = \mathbf{0}$ can be substantially simplified, if we replace $\mathbf{x}$ by the new vector $\mathbf{y}$ such that

$$\mathbf{x} = P\mathbf{y} \quad \text{or} \quad \mathbf{y} = P^{-1}\mathbf{x} \tag{1.20}$$

and replace A with PDP^{-1} to get an equation of the form $g(D, \mathbf{y}) = \mathbf{0}$ that involves the diagonal matrix D and the new vector $\mathbf{y}$.

To illustrate suppose we have a linear system $A\mathbf{x} = \mathbf{b}$. Then we can convert this system into a diagonal system as follows. We consider the new variable vector $\mathbf{y}$ defined by $\mathbf{y} = P\mathbf{x}$. We have

$$\begin{aligned} A\mathbf{x} = \mathbf{b} &\Leftrightarrow PA\mathbf{x} = P\mathbf{b} \\ &\Leftrightarrow PAP^{-1}\mathbf{y} = P\mathbf{b} \\ &\Leftrightarrow D\mathbf{y} = P\mathbf{b} \end{aligned}$$

The last equation defines a **diagonal system**.

1.22 Orthogonal Matrices

Definition A *square* matrix A is called **orthogonal** if it is has *orthonormal* columns. This means that every pair of columns is orthogonal and each column is a unit vector.

Note that a *nonsquare* matrix with orthonormal columns is *not* called orthogonal. (Perhaps a better name for orthogonal matrix would be "orthonormal".)

Orthogonal matrices are **invertible**, because they are square with linearly independent columns. We have the following important theorem.

Theorem 1.22.1. *Let A be a square matrix. The following are equivalent.*

1. *A is orthogonal.*
2. *$A^T A = I$*
3. *$A^{-1} = A^T$*

Examples of Orthogonal Matrices

Example 1.22.1. Show that the rotation matrix A is orthogonal

$$A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

and find its inverse.

Solution: A is orthogonal because

$$AA^T = \begin{bmatrix} \cos^2\theta + \sin^2\theta & 0 \\ 0 & \cos^2\theta + \sin^2\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Note that the inverse of an orthogonal matrix is its transpose. Hence,

$$A^{-1} = A^T = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

Example 1.22.2. It is easy to see that B is also orthogonal.

$$B = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \end{bmatrix}$$

Theorem 1.22.2. *For A a $n \times n$ matrix, the following statements are equivalent.*

1. *A is orthogonal.*
2. *$A\mathbf{u} \cdot A\mathbf{v} = \mathbf{u} \cdot \mathbf{v}$ for any n-vectors $\mathbf{u}$ and $\mathbf{v}$ (Preservation of dot products).*
3. *$\|A\mathbf{v}\| = \|\mathbf{v}\|$ for any n-vector $\mathbf{v}$ (Preservation of lengths).*

REMARK The matrix transformation $T(\mathbf{x}) = A\mathbf{x}$ defined by an orthogonal matrix A is also called orthogonal. By the last theorem we see that *orthogonal matrix transformations preserve dot products. Hence, they preserve lengths and angles.*

Theorem 1.22.3. *If λ is an eigenvalue of an orthogonal matrix A, then $|\lambda| = 1$.*

Proof: If $\mathbf{v}$ an eigenvector of A, then by part 3 of the last theorem

$$\|\mathbf{v}\| = \|A\mathbf{v}\| = \|\lambda\mathbf{v}\| = |\lambda| \, \|\mathbf{v}\|$$

Hence, $|\lambda| = 1$, since $\|\mathbf{v}\| \neq 0$.

This theorem also holds for complex eigenvalues of A.

Eigenvalues of Symmetric Matrices

Theorem 1.22.4. *We have the following.*

1. *A real symmetric matrix has only real eigenvalues.*
2. *A real skew-symmetric matrix eigenvalues that are either pure imaginary or zero.*

Example 1.22.3. The symmetric matrix A

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

has real eigenvalues: $1, 4, -1$.
The skew-symmetric matrix B

$$B = \begin{bmatrix} 0 & 2 & 1 \\ -2 & 0 & 1 \\ -1 & -1 & 0 \end{bmatrix}$$

has pure imaginary or zero eigenvalues: $0, i\sqrt{6}, -i\sqrt{6}$.

1.23 Hermitian and Unitary Matrices

Definitions Let A be a square complex matrix. Then

1. A is called **Hermitian**, if $\overline{A}^T = A$.

2. A is called **skew-Hermitian**, if $\overline{A}^T = -A$.

3. A is called **unitary**, if $\overline{A}^T = A^{-1}$.

Example 1.23.1. Show that matrix A is Hermitian, matrix B is skew-Hermitian, and matrix C is unitary.

$$A = \begin{bmatrix} 4 & 2+i \\ 2-i & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 2-i \\ -2-i & -4i \end{bmatrix}, \quad C = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2}i \\ -\frac{\sqrt{3}}{2}i & \frac{1}{2} \end{bmatrix}$$

Solution: That A is Hermitian because

$$\overline{A}^T = \overline{\begin{bmatrix} 4 & 2+i \\ 2-i & 0 \end{bmatrix}}^T = \begin{bmatrix} 4 & 2-i \\ 2+i & 0 \end{bmatrix}^T = \begin{bmatrix} 4 & 2+i \\ 2-i & 0 \end{bmatrix} = A$$

B is skew-Hermitian, because

$$\overline{B}^T = \overline{\begin{bmatrix} 0 & 2-i \\ -2-i & -4i \end{bmatrix}}^T = \begin{bmatrix} 0 & 2+i \\ -2+i & 4i \end{bmatrix}^T = \begin{bmatrix} 0 & -2+i \\ 2+i & 4i \end{bmatrix} = -B$$

To show that C is unitary, it suffices to check that $\overline{C}^T C = I$. We have

$$\overline{C}^T C = \overline{\begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2}i \\ -\frac{\sqrt{3}}{2}i & \frac{1}{2} \end{bmatrix}}^T \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2}i \\ -\frac{\sqrt{3}}{2}i & \frac{1}{2} \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2}i \\ \frac{\sqrt{3}}{2}i & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2}i \\ -\frac{\sqrt{3}}{2}i & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$$

REMARKS

1. For a *real* skew-Hermitian matrix A, we have $\overline{A} = -A$. Such a matrix is called **skew-symmetric**.

2. For a *real* unitary matrix A, we have $\overline{A} = A^{-1}$. Hence, *a real unitary matrix is orthogonal.*

3. The main diagonal of a Hermitian matrix consists of real numbers.

4. The main diagonal of a skew-Hermitian matrix consists of 0s, or pure imaginary numbers.

5. Equivalent statements for A being a unitary matrix are: $\overline{A}^T A = I$ and also by taking the transpose

$$A^T \overline{A} = I$$

Theorem 1.23.1. *Let A be a complex square matrix. Then*

1. *If A is Hermitian, then its eigenvalues are real. (Thus, this holds for symmetric matrices.)*

2. *If A is skew-Hermitian, then its eigenvalues are pure imaginary, or 0. (Thus, this holds for skew-symmetric matrices.)*

3. *If A is unitary, then its eigenvalues have absolute value 1. (Thus, this holds for real orthogonal matrices.)*

Note Let A be a $n \times n$ unitary matrix. Then for any complex n-vectors $\mathbf{u}$ and $\mathbf{v}$, we have with respect to the complex dot product:

1. $A\mathbf{u} \cdot A\mathbf{v} = \mathbf{u} \cdot \mathbf{v}$ (**Preservation of the complex dot product**)

2. $\|A\mathbf{v}\| = \|\mathbf{v}\|$ (**Preservation of complex norm**)

Note The complex dot product is given by $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \overline{\mathbf{v}} = u_1\overline{v_1} + \cdots + u_n\overline{v_n}$.

Chapter 2

Laplace Transforms

2.1 Laplace Transform; Inverse Transform

Let $f(t)$ be a function with domain all $t \geq 0$. The **Laplace transform** of $f(t)$ is the function $F(s)$ in s given by the integral:

$$F(s) = L(f) = \int_0^\infty e^{-st}\ f(t)\ dt$$

We assume that the above indefinite integral exists. It certainly exists for all $f(t)$ that are useful in applications.

Next, we compute the Laplace transforms of some basic functions.

Example 2.1.1. Show that

$$L(1) = \frac{1}{s}, \quad s > 0 \tag{2.1}$$

Solution: We have

$$\begin{aligned} L(1) &= \int_0^\infty e^{-st}\ 1\ dt \\ &= \lim_{k\to\infty} \frac{-1}{s} e^{-st}\Big|_{t=0}^{t=k} \\ &= \frac{-1}{s} \lim_{k\to\infty} \left(e^{-sk} - e^0\right) \\ &= \frac{1}{s}, \quad s > 0 \end{aligned}$$

We impose $s > 0$, because only then e^{-sk} goes to zero as k goes to infinity.

Example 2.1.2. Show that

$$L(t) = \frac{1}{s^2}, \quad s > 0 \tag{2.2}$$

Solution: By integration by parts and Example 2.1.1, we get

$$\begin{aligned} L(t) &= \int_0^\infty e^{-st}\ t\ dt \\ &= \frac{-e^{-st}}{s} t\Big|_0^\infty - \frac{-1}{s}\int_0^\infty e^{-st}\ dt \\ &= 0 + \frac{1}{s} L(t) \\ &= \frac{1}{s^2}, \quad s > 0 \end{aligned}$$

By iteration of the method in Example 2.1.2 and mathematical induction we can easily compute $L(t^n)$ to get

$$L(t^n) = \frac{n!}{s^{n+1}}, \quad s > 0 \tag{2.3}$$

Example 2.1.3. Show that for a constant a

$$L\left(e^{at}\right) = \frac{1}{s-a}, \quad s > a \tag{2.4}$$

Solution: We have

$$\begin{aligned} L\left(e^{at}\right) &= \int_0^\infty e^{-st}\ e^{at}\ dt \\ &= \lim_{k\to\infty} \frac{-1}{s-a} e^{-t(s-a)}\Big|_{t=0}^{t=k} \\ &= \frac{-1}{s-a} \lim_{k\to\infty} \left(e^{-(s-a)k} - e^0\right) \\ &= \frac{1}{s-a}, \quad s > a \end{aligned}$$

Again, the assumption $s > a$ is necessary so that $e^{-(s-a)k}$ vanishes as $k \to \infty$.

Example 2.1.4. Show that for a constant a

$$L(\cos\ at) = \frac{s}{s^2+a^2}, \quad L(\sin\ at) = \frac{a}{s^2+a^2}, \quad s > 0 \tag{2.5}$$

Solution: Let $x = L(\cos\ at)$ and $y = L(\sin\ at)$. By integrating by parts we get

$$\begin{aligned} x &= \int_0^\infty e^{-st}\ \cos\ at\ dt \\ &= -\frac{e^{-st}}{s}\cos\ at\Big|_0^\infty - \frac{a}{s}\int_0^\infty e^{-st}\ \sin\ at\ dt \\ &= \frac{1}{s} - \frac{a}{s}y \end{aligned}$$

and

$$\begin{aligned} y &= \int_0^\infty e^{-st}\ \sin\ at\ dt \\ &= -\frac{e^{-st}}{s}\sin\ at\Big|_0^\infty + \frac{a}{s}\int_0^\infty e^{-st}\ \cos\ at\ dt \\ &= \frac{a}{s}x \end{aligned}$$

provided that $s > 0$. Substituting $y = ax/s$ into $x = 1/s - ay/s$, yields

$$x = \frac{1}{s} - \left(\frac{a}{s}\right)^2 x$$

or

$$x = \frac{s}{s^2 + a^2}$$

Therefore,

$$y = \frac{a}{s}x = \frac{a}{s}\frac{s}{s^2 + a^2} = \frac{a}{s^2 + a^2}$$

So, the formulas are proved.

Example 2.1.5. Find $L(f(t))$, where

$$f(t) = \begin{cases} e^t, & \text{if} \quad 0 < t < 1 \\ 0, & \text{if} \quad t > 1 \end{cases}$$

Solution: We have

$$\begin{aligned} L(f(t)) &= \int_0^\infty e^{-st}\ f(t)\ dt \\ &= \int_0^1 e^{-st}\ e^t\ dt + \int_1^\infty e^{-st}\ 0\ dt \\ &= \frac{1}{1-s} e^{-t(s-1)}\Big|_{t=0}^{t=1} + 0 \\ &= \frac{1-e^{1-s}}{s-1} \end{aligned}$$

Linearity of Laplace Transform

The most important property of Laplace transforms is *linearity*. This is expressed in the following theorem.

Theorem 2.1.1. *The Laplace transform is* linear. *I.e., if $f_1(t)$ and $f_2(t)$ are functions whose Laplace transforms exist and c_1 and c_2 are any constants, then*

$$L\{c_1 f_1(t) + c_2 f_2(t)\} = c_1 L\{f_1(t)\} + c_2 L\{f_2(t)\}$$

Proof. We have

$$\begin{aligned} L\{c_1 f_1(t) + c_2 f_2(t)\} &= \int_0^\infty e^{-st}\ [c_1 f_1(t) + c_2 f_2(t)]\ dt \\ &= c_1 \int_0^\infty e^{-st}\ f_1(t)\ dt + c_2 \int_0^\infty e^{-st}\ f_2(t)\ dt \\ &= c_1 L\{f_1(t)\} + c_2 L\{f_2(t)\} \end{aligned}$$

Linearity may be used to compute the Laplace transforms of more complicated functions.

Example 2.1.6. Find

$$L\left(7 + 2e^{-3t} + 5t^2 - \cos\ 6t\right)$$

Solution: By linearity this expression equals

$$7L(1) + 2L\left(e^{-3t}\right) + 5L\left(t^2\right) - L(\cos\ 6t) = \frac{7}{s} + \frac{2}{s+3} + \frac{10}{s^3} - \frac{s}{s^2+36}$$

2.2 Inverse Laplace Transform

If $F(s) = L(f)$, then we say that the **inverse Laplace transform** of $F(s)$ is $f(t)$ and we write $f(t) = L^{-1}(F)$. Hence,

$$f(t) = L^{-1}(F) \Longleftrightarrow F(s) = L(f)$$

Note that $L(L^{-1}(F)) = F$ and $L^{-1}(L(f)) = f$.

Just like Laplace transforms, inverse Laplace transforms are also *linear*.

$$L^{-1}\{c_1F_1(s) + c_2F_2(s)\} = c_1L^{-1}\{F_1(s)\} + c_2L^{-1}\{F_2(s)\}$$

Example 2.2.1. Find $L^{-1}\left(\frac{1}{s^4} + \frac{1}{s^2+4} + \frac{1}{5s-1}\right)$.

Solution: By linearity we have

$$\begin{aligned} L^{-1}\left(\frac{1}{s^4} + \frac{1}{s^2+4} + \frac{1}{5s-1}\right) &= \frac{1}{3!}L^{-1}\left(\frac{3!}{s^4}\right) + \frac{1}{2}L^{-1}\left(\frac{2}{s^2+2^2}\right) \\ &\quad + \frac{1}{5}L^{-1}\left(\frac{1}{s-\frac{1}{5}}\right) \\ &= \frac{1}{6}t^3 + \frac{1}{2}\sin(2t) + \frac{1}{5}e^{\frac{t}{5}} \end{aligned}$$

Example 2.2.2. Find $L^{-1}\left(\frac{2}{(s+1)(s-1)}\right)$.

Solution: By partial fractions $\frac{2}{(s+1)(s-1)} = \frac{1}{s-1} - \frac{1}{s+1}$. By linearity,

$$L^{-1}\left(\frac{2}{(s+1)(s-1)}\right) = L^{-1}\left(\frac{1}{s-1}\right) - L^{-1}\left(\frac{1}{s+1}\right) = e^t - e^{-t}$$

Example 2.2.3. Find $L^{-1}\left(\frac{3s+1}{s^2+4}\right)$.

Solution: By linearity we have

$$L^{-1}\left(\frac{3s+1}{s^2+4}\right) = 3L^{-1}\left(\frac{s}{s^2+4}\right) + \frac{1}{2}L^{-1}\left(\frac{2}{s^2+4}\right) = 3\cos 2t + \frac{1}{2}\sin 2t$$

Example 2.2.4. Find $L^{-1}\left(\frac{1}{s(s^2+1)}\right)$.

Solution: By partial fractions $\frac{1}{s(s^2+1)} = \frac{1}{s} - \frac{s}{s^2+1}$. Hence, by linearity,

$$L^{-1}\left(\frac{1}{s(s^2+1)}\right) = L^{-1}\left(\frac{1}{s}\right) - L^{-1}\left(\frac{s}{s^2+1}\right) = 1 - \cos t$$

2.3 Exponential Shifting

Exponential functions of the form e^{at} play a special role in the theory of Laplace transforms. As an example, knowing the Laplace of $f(t)$ immediately yields the Laplace of $e^{at}f(t)$, according to the following useful theorem.

Theorem 2.3.1 (Exponential Shifting or the First Shifting Theorem)**.** *Let* $F(s) = L(f(t))$. *Then*

$$L\left\{e^{at}\ f(t)\right\} = F(s-a)$$

and

$$L^{-1}\left\{F(s-a)\right\} = e^{at}\ f(t)$$

Proof. We have

$$\begin{aligned} L\left\{e^{at}\ f(t)\right\} &= \int_0^\infty e^{-st}\ e^{at}\ f(t)\ dt = \int_0^\infty e^{-(s-a)t}\ f(t)\ dt \\ &= L\left\{f(t)\right\}(s-a) = F(s-a) \end{aligned}$$

Example 2.3.1. Find $L(t^3e^{5t})$.

Solution: By exponential shifting, we get

$$L\left(t^3e^{5t}\right) = L\left\{t^3\right\}(s-5) = \left.\frac{3!}{s^4}\right|_{s-5} = \frac{6}{(s-5)^4}$$

Example 2.3.2. Find $L^{-1}\left(\frac{1}{(s+7)^3}\right)$.

Solution: We have

$$L^{-1}\left(\frac{1}{(s+7)^3}\right) = \frac{1}{2}L^{-1}\left(\frac{2!}{(s+7)^3}\right) = \frac{1}{2}L^{-1}\left(\left.\frac{2!}{s^3}\right|_{s+7}\right) = \frac{1}{2}t^2e^{-7t}$$

Example 2.3.3 (Completion of Square)**.** Find $L^{-1}\left(\frac{s+3}{s^2+2s+2}\right)$.

Solution: The quadratic in the denominator has complex roots. So it cannot be factored over the real numbers. However, we may complete the square and

use exponential shifting to get

$$\begin{aligned} L^{-1}\left(\frac{s+3}{s^2+2s+2}\right) &= L^{-1}\left(\frac{s+3}{(s+1)^2+1}\right) = \\ &= L^{-1}\left(\frac{s+1}{(s+1)^2+1}\right) + L^{-1}\left(\frac{2}{(s+1)^2+1}\right) \\ &= L^{-1}\left(\frac{s}{s^2+1}\Big|_{s+1}\right) + 2L^{-1}\left(\frac{1}{s^2+1}\Big|_{s+1}\right) \\ &= e^{-t}\ \cos\ t + 2e^{-t}\ \sin\ t \end{aligned}$$

2.4 A Table of Laplace Transforms

We collect here the Laplace transforms of some basic functions for future reference.

$f(t)$	$L(f)$
1	$\frac{1}{s}$ $(s>0)$
t^n	$\frac{n!}{s^{n+1}}$ $(s>0)$
e^{at}	$\frac{1}{s-a}$ $(s>a)$
$\cos\ at$	$\frac{s}{s^2+a^2}$ $(s>a)$
$\sin\ at$	$\frac{a}{s^2+a^2}$ $(s>0)$
$\cosh\ at$	$\frac{s}{s^2-a^2}$ $(s>a)$
$\sinh\ at$	$\frac{a}{s^2-a^2}$ $(s>a)$
$e^{at}\cos\ bt$	$\frac{s-a}{(s-a)^2+b^2}$ $(s>a)$
$e^{at}\sin\ bt$	$\frac{b}{(s-a)^2+b^2}$ $(s>a)$

Fact A function has Laplace transform if it does not grow faster that an exponential of the form Me^{kt}. More precisely we have the following.

Theorem 2.4.1 (Existence of Laplace Transform)**.** *Let $f(t)$ be defined for all $t \geq 0$ and be piecewise continuous on all finite sunintervals of $[0, \infty]$. If for all $t \geq 0$ there are constants M and k such that*

$$|f(t)| \leq Me^{kt}, \quad t \geq 0 \tag{2.6}$$

then $L(f)$ exists and it is defined for all $s > k$.

2.5 Transforms of Derivatives and Integrals; ODES

The following theorem that computes the transform of the derivative is key to solving initial value problems by using Laplace transforms.

Theorem 2.5.1 (Laplace of Derivatives)**.** *Let $f, f', \ldots, f^{(n-1)}$ be continuous and satisfy (2.6) and let $f^{(n)}$ be piecewise continuous on all finite subintervals of $[0, \infty]$. Then*

$$\begin{aligned} L(f') &= sL(f) - f(0) \\ L(f'') &= s^2 L(f) - sf(0) - f'(0) \\ &\vdots \\ L\left(f^{(n)}\right) &= s^n L(f) - s^{n-1} f(0) - s^{n-2} f'(0) - \cdots - f^{(n-1)}(0) \end{aligned}$$

Proof. We prove the formula for the first derivative. The remaining follow by induction. By integration by parts, we get

$$L(f') = \int_0^\infty e^{-st}\ f'(t)\ dt = \left[e^{-st} f(t)\right]\Big|_0^\infty + s\int_0^\infty e^{-st}\ f(t)\ dt$$

Now f satisfies (2.6), thus, $\lim_{t\to\infty} e^{-st} f(t) = 0$, provided $s > k$. Hence, the only contribution of the first term is $-e^0 f(0) = -f(0)$. The second term is $sL(f)$.

Application to ODEs

Example 2.5.1. Solve the initial value problem.

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

Solution: Let $Y = L(y)$. We apply Laplace transform to both sides of the differential equation to get

$$\begin{aligned}
L(y'') + L(y) = \tfrac{2}{s^2} &\Rightarrow \\
s^2 Y - sy(0) - y'(0) + Y = \tfrac{2}{s^2} &\Rightarrow \\
s^2 Y - s + 1 + Y = \tfrac{2}{s^2} &\Rightarrow \\
(s^2+1) Y = s + \tfrac{2}{s^2} - 1 &\Rightarrow \\
Y = \tfrac{s}{s^2+1} + \tfrac{2}{s^2(s^2+1)} - \tfrac{1}{s^2+1} &\Rightarrow \\
Y = \tfrac{s}{s^2+1} + \left[\tfrac{2}{s^2} - \tfrac{2}{s^2+1}\right] - \tfrac{1}{s^2+1} &\Rightarrow \\
Y = \tfrac{s}{s^2+1} - \tfrac{3}{s^2+1} + \tfrac{2}{s^2}
\end{aligned}$$

Then we take the inverse Laplace of both sides of the last equation to get

$$y(t) = \cos\ t - 3\sin\ t + 2t$$

Transform of the Integral

Theorem 2.5.2 (The Laplace of an Integral)**.** *Let $f(t)$ be defined for all $t \geq 0$ be piecewise continuous on $[0, \infty]$ and satisfy (2.6). If $F(s) = L(f)$, then for $s > \max\{k, 0\}$ and $t > 0$ we have*

$$L\left(\int_0^t f(\tau)\ d\tau\right) = \frac{1}{s} F(s)$$

and in inverse form

$$L^{-1}\left\{\frac{1}{s} F(s)\right\} = \int_0^t f(\tau)\ d\tau$$

Example 2.5.2. Find $L^{-1}\left(\frac{1}{s(s^2+1)}\right)$.

Solution: By Theorem 2.5.2 we have

$$L^{-1}\left(\frac{1}{s}\frac{1}{(s^2+1)}\right) = \int_0^t \sin(\tau)\ d\tau = -\cos(\tau)\Big|_0^t = 1 - \cos\ t$$

This answer agrees with the one in Example 2.2.4 where we used partial fractions to compute $L^{-1}\left(\frac{1}{s(s^2+1)}\right)$.

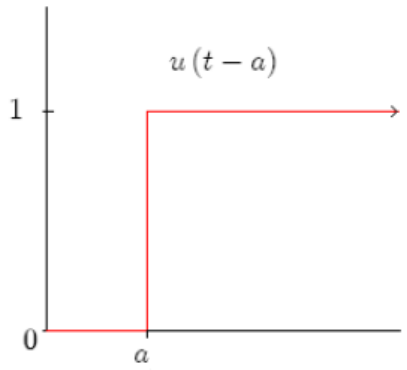
2.6 Unit Step Functions; ODEs

The material in this section is very useful. It addresses the cases where the right-hand side of the differential equation is a piecewise continuous function, something rather common in physics and engineering.

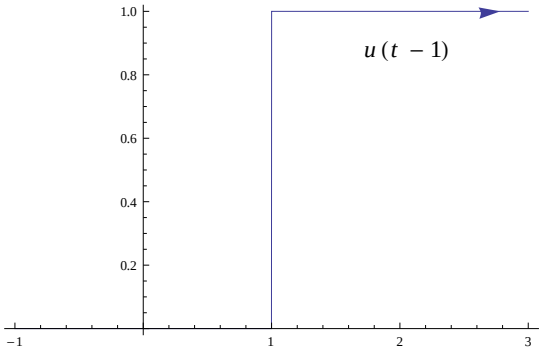
The **unit step function** (or **Heaviside function**), $u(t-a)$, about $t=a$, $(a>0)$ is the function in t that has a unit jump at $t=a$. I.e.,

$$u(t-a)=\begin{cases}1, & t>a\\ 0, & t<a\end{cases}$$

For the special case with $a=0$ we simply write $u(t)$.



Example 2.6.1. The function $u(t-1)$ is depicted in the figure below.



Theorem 2.6.1 (Laplace of Unit Step Function)**.** *We have*

$$L\{u(t-a)\}=\frac{e^{-as}}{s}$$

and

$$L^{-1}\left\{\frac{e^{-as}}{s}\right\} = u\left(t-a\right)$$

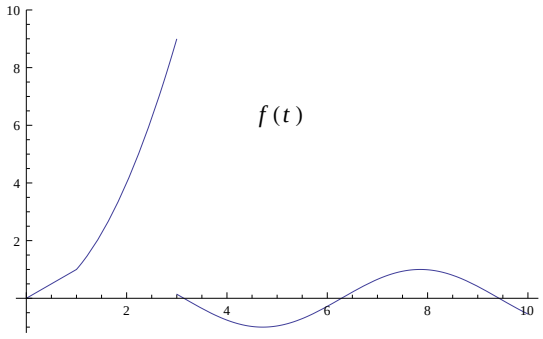
Proof.

$$\begin{aligned} L\left\{u\left(t-a\right)\right\} &= \int_0^\infty e^{-st}\ u\left(t-a\right)\ dt \\ &= \int_0^a e^{-st}\ 0\ dt + \int_a^\infty e^{-st}\ 1\ dt \\ &= \left.\frac{e^{-st}}{-s}\right|_a^\infty \\ &= \frac{e^{-as}}{s} \end{aligned}$$

Piecewise continuous functions can be written in terms of unit step functions. This can be very useful in practice.

Example 2.6.2. Write the function $f\left(t\right)$ in terms of unit step functions for $t \geq 0$.

$$f\left(t\right) = \begin{cases} t & \text{if} \quad 0 < t < 1 \\ t^2 & \text{if} \quad 1 < t < 3 \\ \sin\ t & \text{if} \quad t > 3 \end{cases}$$



Solution: An easy way to proceed is to see the points $1, 3$ as "switches". At the beginning we "turn on" the function t by starting with t. Then at "time" $t = 1$ we "turn off" t by subtracting $tu\left(t-1\right)$ and "turn on" the new function t^2 by adding $t^2 u\left(t-1\right)$. We continue like this to get

$$f\left(t\right) = t - tu\left(t-1\right) + t^2 u\left(t-1\right) - t^2 u\left(t-3\right) + \sin\ t\ u\left(t-3\right)$$

Theorem 2.6.2 (Second Shifting Theorem)**.** *If* $F(s) = L(f)$, *then*

$$L\{f(t-a)\ u(t-a)\} = e^{-as}F(s)$$

and

$$L^{-1}\left\{e^{-as}F(s)\right\} = f(t-a)\ u(t-a)$$

Proof. Using the change of variables $v = t - a$ we get,

$$\begin{aligned} L\{f(t-a)\ u(t-a)\} &= \int_0^\infty e^{-st}\ u(t-a)\ f(t-a)\ dt \\ &= \int_a^\infty e^{-st}\ f(t-a)\ dt \\ &= \int_0^\infty e^{-s(v+a)}\ f(v)\ dv \\ &= e^{-as}\int_0^\infty e^{-sv}\ f(v)\ dv \\ &= e^{-as}F(s) \end{aligned}$$

Example 2.6.3. Find $L(g(t))$, where

$$g(t) = \begin{cases} e^t, & \text{if} \quad 0 < t < 1 \\ 0, & \text{if} \quad t > 1 \end{cases}$$

Solution: We have $g(t) = e^t - e^t u(t-1)$. Hence,

$$\begin{aligned} L(g(t)) &= L\left(e^t\right) - L\left(e^t u(t-1)\right) \\ &= \frac{1}{s-1} - L\left(e^t u(t-1)\right) \end{aligned}$$

To compute the second term we use the second shifting theorem with $a = 1$ and $f(t-1) = e^t$. Therefore, $f(t) = e^{t+1} = e^t e$. Thus,

$$\begin{aligned} L\left\{e^t u(t-1)\right\} &= e^{-s}L\left(ee^t\right) \\ &= e^{-s}eL\left(e^t\right) \\ &= \frac{e^{1-s}}{s-1} \end{aligned}$$

Therefore,

$$\begin{aligned} L\left(g(t)\right) &= \frac{1}{s-1} - \frac{e^{1-s}}{s-1} \\ &= \frac{1-e^{1-s}}{s-1} \end{aligned}$$

Note that this is the same answer as in Example 2.1.5

Example 2.6.4. Find $L\left\{(2t-1)\,u\left(t-1\right)\right\}$.

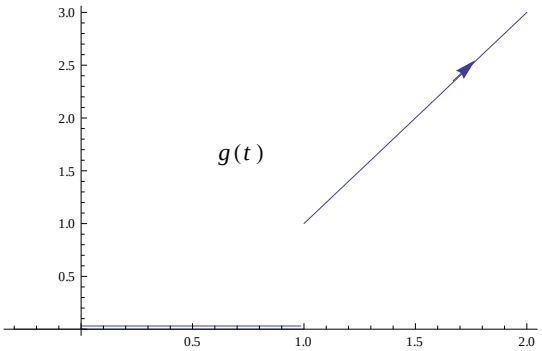
Solution: We have $a=1$ and $f\left(t-1\right)=2t-1$. Hence, $f\left(t\right)=2\left(t+1\right)-1=2t+1$. Thus,

$$L\left\{(2t-1)\,u\left(t-1\right)\right\}=e^{-s}L\left(2t+1\right)=e^{-s}\left(\frac{2}{s^2}+\frac{1}{s}\right)$$

Example 2.6.5. Solve the initial value problem

$$y'+y=g\left(t\right),\ \ y\left(0\right)=0$$

where $g\left(t\right)$ is the function with value 0 for t between 0 and 1. And for $t>1$ the value $g\left(t\right)$ is the same as that of the straight line through the points $(1,1)$ and $(2,3)$.



Solution: The function g is $g\left(t\right)=\begin{cases} 0 & \text{if} \quad 0<t<1 \\ 2t-1 & \text{if} \quad t>1 \end{cases}$. Hence, $g\left(t\right)=(2t-1)\,u\left(t-1\right)$. By Example 2, we have $L\left(g\right)=e^{-s}\left(\frac{2}{s^2}+\frac{1}{s}\right)$. Thus, applying Laplace transforms to the differential equation yields (with $Y=L\left(y\right)$)

$$sY-0+Y=e^{-s}\left(\frac{2}{s^2}+\frac{1}{s}\right)\Rightarrow Y=e^{-s}\frac{s+2}{s^2\left(s+1\right)}$$

Hence, by partial fractions

$$Y = e^{-s}\left(\frac{1}{s+1} - \frac{1}{s} + \frac{2}{s^2}\right)$$

Taking inverse Laplace we get

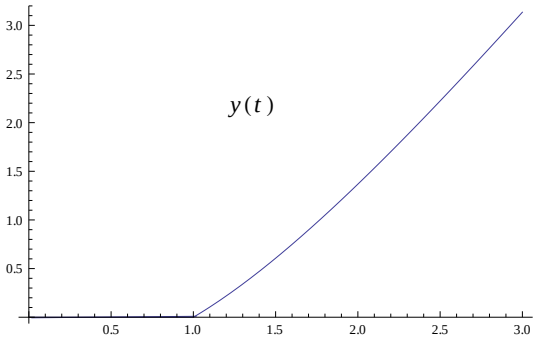
$$y = e^{-(t-1)}u\left(t-1\right) - u\left(t-1\right) + 2\left(t-1\right)u\left(t-1\right)$$

or,

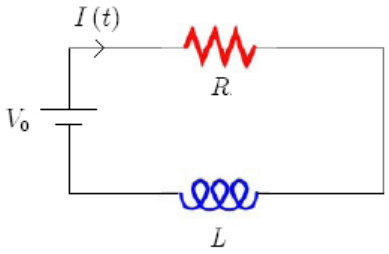
$$y\left(t\right) = \left(e^{1-t} + 2t - 3\right)u\left(t-1\right)$$

or

$$y\left(t\right) = \begin{cases} 0, & \text{if} \quad t < 1 \\ e^{1-t} + 2t - 3, & \text{if} \quad t > 1 \end{cases}$$



RL-Circuit Response to a Rectangular Wave



In the RL series circuit with resistor of resistance R Ohm, inductor with inductance L Henry, we apply constant voltage V_0 only between times $t = a$ and $t = b$. If the initial current is zero, $I\left(0\right) = 0$ we want to find $I\left(t\right)$.

By Kirchhoff's vlotage law we have

$$L\frac{dI}{dt} + RI = V(t)$$

In this case

$$L\frac{dI}{dt} + RI = V_0\{u(t-a) - u(t-b)\}$$

Applying Laplace transforms we get (using J for the Laplace of I)

$$L(sJ - 0) + RJ = V_0\left(\frac{e^{-as}}{s} - \frac{e^{-bs}}{s}\right)$$

Hence,

$$J = V_0\left(\frac{e^{-as}}{s(sL+R)} - \frac{e^{-bs}}{s(sL+R)}\right)$$

By partial fractions we get

$$J = \frac{V_0}{R}\left(\frac{1}{s} - \frac{1}{s+R/L}\right)e^{-as} - \frac{V_0}{R}\left(\frac{1}{s} - \frac{1}{s+R/L}\right)e^{-bs}$$

Taking inverse Laplace we get

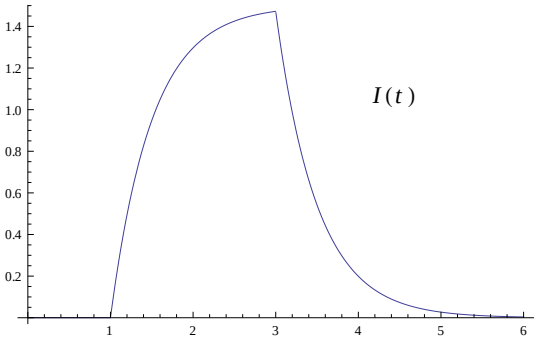
$$I(t) = \frac{V_0}{R}\left(1 - e^{-(R/L)(t-a)}\right)u(t-a) - \frac{V_0}{R}\left(1 - e^{-(R/L)(t-b)}\right)u(t-b)$$

or

$$I(t) = \begin{cases} 0 & \text{if} \quad t < a \\ \frac{V_0}{R}\left(1 - e^{-(R/L)(t-a)}\right) & \text{if} \quad a < t < b \\ \frac{V_0}{R}\left(e^{-(R/L)(t-b)} - e^{-(R/L)(t-a)}\right) & \text{if} \quad t > b \end{cases}$$

For example, for $a = 1$, $b = 3$, $V_0 = 3$, $R = 2$, and $L = 1$ we have the following solution.

$$I(t) = \begin{cases} 0 & \text{if} \quad t < 1 \\ \frac{3}{2}\left(1 - e^{-2(t-1)}\right) & \text{if} \quad 1 < t < 3 \\ \frac{3}{2}\left(e^{-2(t-3)} - e^{-2(t-1)}\right) & \text{if} \quad t > 3 \end{cases}$$

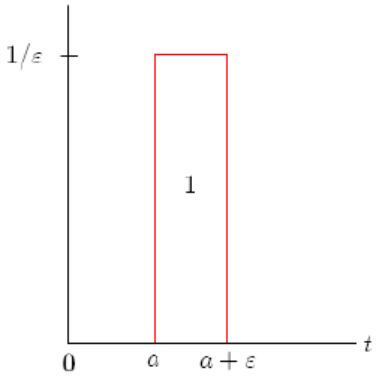


2.7 Dirac's Delta Function

We need a function to model sudden and large changes in quantities that we interested in. For example, if a weight attached to a spring is hit by a hammer, or a baseball is hit by a bat. Such cases may be modelled by using Dirac's delta function. We start with the function

$$d_\varepsilon(t-a) = \begin{cases} 1/\varepsilon & \text{if } \ a < t < a+\varepsilon \\ 0 & \text{otherwise} \end{cases}$$

This function's graph is a rectangular wave of width ε and height $1/\varepsilon$ starting at $t = a$. So, its area is 1.



If we let ε take on small values we get a sequence of rectangular wave functions d_ε whose width goes to zero and the height goes to infinity but the area remains always 1. This limit is denoted by $\delta(t-a)$ and it is called

Dirac's delta function or **unit impulse function**.

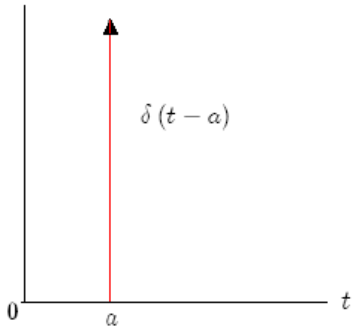
$$\delta(t-a) = \lim_{n\to\infty} d_\varepsilon(t-a)$$

So δ has one value which is ∞ at $t=a$ and it is zero if $t \neq a$, yet, the area under the graph of it is 1. Strictly speaking, Dirac's delta is not a function in the usual sense, but a so-called **generalized function**, or a **distribution**. One way to view $\delta(t-a)$ is as a "function" that takes the value ∞, if $t=a$, also it takes the value 0, if $t \neq a$,

$$\delta(t-a) = \begin{cases} \infty & \text{if} \quad t=a \\ 0 & \text{if} \quad t \neq a \end{cases}$$

and at the same time

$$\int_0^\infty \delta(t-a)\ dt = 1$$



For any continuous function $f(t)$ defined at least for $t \geq 0$, we have the following property.

$$\int_0^\infty f(t)\ \delta(t-a)\ dt = f(a)$$

for $a \geq 0$.

We compute the Laplace transform of $\delta(t-a)$ by first computing the

$L\{d_\varepsilon(t-a)\}$

$$\begin{aligned} L\{d_\varepsilon(t-a)\} &= L\left\{\frac{1}{\varepsilon}u(t-a)-\frac{1}{\varepsilon}u(t-(a+n))\right\} \\ &= \frac{1}{\varepsilon}\left\{\frac{e^{-as}}{s}-\frac{e^{-(a+n)s}}{s}\right\} \\ &= e^{-as}\left\{\frac{1-e^{-\varepsilon s}}{\varepsilon s}\right\} \end{aligned}$$

and then taking the limit as $\varepsilon \to 0$.

$$\begin{aligned} L\{\delta(t-a)\} &= \lim_{\varepsilon\to 0} L\{d_\varepsilon(t-a)\} \\ &= \lim_{\varepsilon\to 0} e^{-as}\left\{\frac{1-e^{-\varepsilon s}}{\varepsilon s}\right\} \\ &= e^{-as}\lim_{\varepsilon\to 0}\left\{\frac{1-e^{-\varepsilon s}}{\varepsilon s}\right\} \\ &= e^{-as}\lim_{\varepsilon\to 0}\left\{\frac{\frac{d}{d\varepsilon}(1-e^{-\varepsilon s})}{\frac{d}{d\varepsilon}(\varepsilon s)}\right\} \\ &= e^{-as}\lim_{\varepsilon\to 0}\left\{\frac{se^{-s\varepsilon}}{s}\right\} \\ &= e^{-as}\lim_{\varepsilon\to 0}\left\{e^{-s\varepsilon}\right\} \\ &= e^{-as} \end{aligned}$$

where in the third equality we used l'Hopital's rule. So, we have the following theorem

Theorem 2.7.1 (Laplace of the Delta Function)**.** *For a constant a we have*

$$L\{\delta(t-a)\} = e^{-as}$$

and

$$L^{-1}\left(e^{-as}\right) = \delta(t-a)$$

Example 2.7.1 (Strike of Mass attached to Spring)**.** Solve the initial value problem

$$y'' + y = 5\delta(t-1),\ y(0) = 2,\ y'(0) = 0$$

This models the displacement $y(t)$ of a mass of 1 mass unit attached to spring where at time $t = 1$ a hammer blow applies "five times the unit infinite force" during an infinitesimal time interval.

Solution: Applying Laplace transforms we get

$$s^2Y - 2s - 0 + Y = 5e^{-s}$$

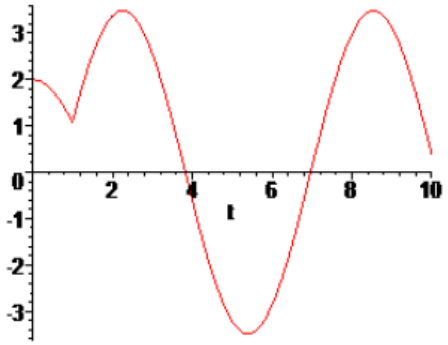
or

$$Y = \frac{2s}{s^2+1} + \frac{5e^{-s}}{s^2+1}$$

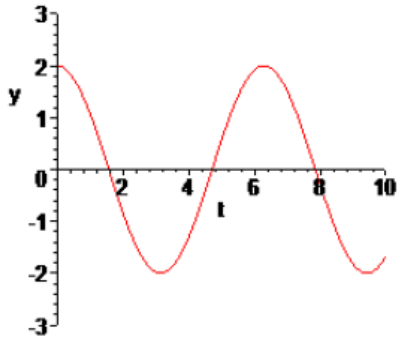
Thus,

$$y(t) = 2\cos t + 5\sin(t-1)u(t-1)$$

So, we start out with the function $2\cos t$ and at time $t = 1$ when the strike occurs, the function $5\sin(t-1)$ is added to it.



$y(t) = 2\cos t + 5\sin(t-1)\,u(t-1)$



$y(t) = 2\cos t$

2.8 Laplace and Systems of ODES; Applications

Review: Ordinary Differential Equation

Differential equations express relations between unknown functions and some of their derivatives. For example,

$$\frac{dx}{dt} = 2x \tag{2.7}$$

is a differential equation with unknown function $x(t)$.

Solving a differential equations amounts to finding all possible unknown functions $x(t)$ that satisfy the equation. In the case of equation (2.7) we seek a function with derivative 2 times the function itself. It is easy to guess such function, namely e^{2t}. In fact, for any constant c the function $x(t) = ce^{2t}$ is a solution, because

$$\frac{dx}{dt} = \frac{d\left(ce^{2t}\right)}{dt} = c\left(2e^{2t}\right) = 2\left(ce^{2t}\right) = 2x$$

So, a differential equation may have a **family of solutions** that depend on constants, such as c above, called **parameters**.

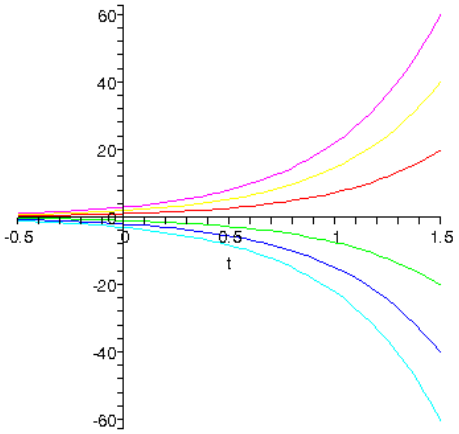


Figure 2.1: A family of solutions for $x'(t) = 2x$.

A **system of ordinary differential equations** is a set of differential equations in several unknown functions. A solution to a system is a set of functions that is a common solution to all its differential equations.

Example 2.8.1. One may easily verify that for any constants c_1 and c_2 the functions

$$y_1(t) = c_1\cos(t) + c_2\sin(t) \quad \text{and} \quad y_2(t) = c_2\cos(t) - c_1\sin(t)$$

are solutions of the system

$$\begin{aligned} \frac{dy_1}{dt} &= y_2 \\ \frac{dy_2}{dt} &= -y_1 \end{aligned}$$

Ordinary Differential Equations and Laplace Transforms

Laplace transforms can be used to solve linear systems of differential equations with given initial conditions.

Example 2.8.2. Solve the homogeneous system of differential equations for the unknown functions $y_1(t)$, $y_2(t)$, and $y_3(t)$.

$$\begin{aligned} \frac{dy_1}{dt} &= -y_1 + 8y_3 \\ \frac{dy_2}{dt} &= -y_2 + y_3 \\ \frac{dy_3}{dt} &= y_1 + y_3 \end{aligned}$$

subject to initial conditions

$$y_1(0) = -4, \quad y_2(0) = 4, \quad y_3(0) = -2$$

Solution: Applying Laplace transforms and using $Y_1 = L(y_1)$, $Y_2 = L(y_2)$, and $Y_3 = L(y_3)$, we get

$$\begin{aligned} sY_1 + 4 &= -Y_1 + 8Y_3 \\ sY_2 - 4 &= -Y_2 + Y_3 \\ sY_3 + 2 &= Y_1 + Y_3 \end{aligned}$$

or

$$\begin{aligned} (s+1)Y_1 - 8Y_3 &= -4 \\ (s+1)Y_2 - Y_3 &= 4 \\ -Y_1 + (s-1)Y_3 &= -2 \end{aligned}$$

Solving the linear system by Cramer's rule yields

$$Y_1 = -\frac{4}{s-3}, \quad Y_2 = \frac{4s-14}{s^2-2s-3}, \quad Y_3 = -\frac{2}{s-3}$$

But

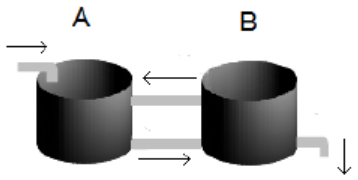
$$Y_2 = \frac{4s-14}{s^2-2s-3} = \frac{4s-14}{(s+1)(s-3)} = \frac{9}{2(s+1)} - \frac{1}{2(s-3)}$$

So taking inverse Laplace transforms yields

$$y_1(t) = -4e^{3t}, \quad y_2(t) = \frac{9}{2}e^{-t} - \frac{1}{2}e^{3t}, \quad y_3(t) = -2e^{3t}$$

Applications to Mixing

Example 2.8.3. Tanks A and B each of volume 100 gallons are full of brine. Brine flows into tank A at the rate of 6 gallons per minute at salt concentration of 2 pounds per gallon. Brine also flows from tank A to tank B at the rate of 8 gallons per minute and from tank B back to tank A at the rate of 2 gallons per minute. Finally, brine flows out of tank B at the rate of 6 gallons per minute. Find $x_1(t)$ and $x_2(t)$, the amounts of salt in A and B at time t, given that the initial amount of salt in each tank was zero.



Solution: For each tank the net rate of salt equals the rate of salt into the tank minus the rate out. Also, each tank remains at the same volume of 100 gallons at all times. So we have,

$$\begin{aligned}
\frac{dx_1}{dt} &= R_{\text{in}} - R_{\text{out}} \\
&= \left(6\frac{\text{gal}}{\text{min}}\right)\left(2\frac{\text{lb}}{\text{gal}}\right) + \left(2\frac{\text{gal}}{\text{min}}\right)\left(\frac{x_2}{100}\frac{\text{lb}}{\text{gal}}\right) - \left(8\frac{\text{gal}}{\text{min}}\right)\left(\frac{x_1}{100}\frac{\text{lb}}{\text{gal}}\right) \\
\frac{dx_2}{dt} &= R_{\text{in}} - R_{\text{out}} \\
&= \left(8\frac{\text{gal}}{\text{min}}\right)\left(\frac{x_1}{100}\frac{\text{lb}}{\text{gal}}\right) - \left(2\frac{\text{gal}}{\text{min}}\right)\left(\frac{x_2}{100}\frac{\text{lb}}{\text{gal}}\right) - \left(6\frac{\text{gal}}{\text{min}}\right)\left(\frac{x_2}{100}\frac{\text{lb}}{\text{gal}}\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{dx_1}{dt} &= 12 - \frac{2}{25}x_1 + \frac{1}{50}x_2 \\
\frac{dx_2}{dt} &= \frac{2}{25}x_1 - \frac{2}{25}x_2
\end{aligned}$$

subject to

$$x_1(t) = 0, \qquad x_2(t) = 0$$

Using Laplace transforms with $X_1 = L(x_1)$ and $X_2 = L(x_2)$ we get

$$\begin{aligned} sX_1 &= \frac{12}{s} - \frac{2}{25}X_1 + \frac{1}{50}X_2 \\ sX_2 &= \frac{2}{25}X_1 - \frac{2}{25}X_2 \end{aligned}$$

We solve to get

$$X_1 = \frac{7500s + 600}{3s + 100s^2 + 625s^3} = \frac{200}{s} - \frac{150}{s + 1/25} - \frac{50}{s + 3/25}$$

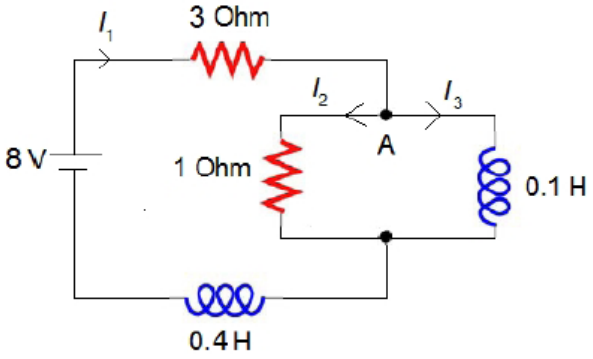
$$X_2 = \frac{600}{3s + 100s^2 + 625s^3} = \frac{200}{s} - \frac{300}{s + 1/25} + \frac{100}{s + 3/25}$$

Taking inverse transfroms yields

$$\begin{aligned} x_1(t) &= 200 - 150e^{-t/25} - 50e^{-3t/25} \\ x_2(t) &= 200 - 300e^{-t/25} + 100e^{-3t/25} \end{aligned}$$

Applications to Electrical Circuits

Example 2.8.4. Find the currents $I_1(t)$, $I_2(t)$, and $I_3(t)$ in the following circuit, given that all initial currents are zero.



Solution: For the loop with currents $I_1(t)$ and $I_2(t)$, we have $0.4I_1' + 3I_1 + I_2 = 8$. For the loop with currents $I_2(t)$ and $I_3(t)$, we have $0.1I_3' - I_2 = 0$. At node A, we have $I_1 = I_2 + I_3$. Hence,

$$\begin{aligned} 0.4I_1' + 3I_1 + I_2 &= 8 \\ 0.1I_3' - I_2 &= 0 \\ I_1 - I_2 - I_3 &= 0 \end{aligned}$$

We apply Laplace transforms and use the notation $J_1 = L(I_1)$, $J_2 = L(I_2)$, and $J_3 = L(I_3)$. Given that $I_1(0) = 0$, $I_2(0) = 0$, and $I_3(0) = 0$, we have

$$\begin{aligned} J_1 - J_2 - J_3 &= 0 \\ (0.4s + 3) J_1 + J_2 &= 8/s \\ -J_2 + 0.1sJ_3 &= 0 \end{aligned}$$

We solve the linear system to get

$$\begin{aligned} J_1 &= \frac{20s + 200}{75s + 20s^2 + s^3} = \frac{8}{3s} - \frac{2}{s+5} - \frac{2}{3(s+15)} \\ J_2 &= \frac{20}{75 + 20s + s^2} = \frac{2}{s+5} - \frac{2}{s+15} \\ J_3 &= \frac{200}{75s + 20s^2 + s^3} = \frac{8}{3s} - \frac{4}{s+5} + \frac{4}{3(s+15)} \end{aligned}$$

Finally, we apply inverse transfroms to get

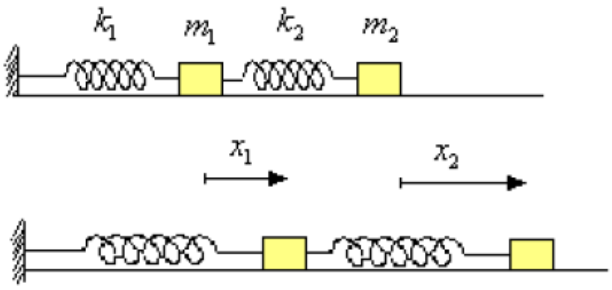
$$\begin{aligned} I_1(t) &= \frac{8}{3} - 2e^{-5t} - \frac{2}{3}e^{-15t} \\ I_2(t) &= 2e^{-5t} - 2e^{-15t} \\ I_3(t) &= \frac{8}{3} - 4e^{-5t} + \frac{4}{3}e^{-15t} \end{aligned}$$

Applications to Springs

Example 2.8.5. Consider the following two-spring two-mass system. Let $x_1(t)$ and $x_2(t)$ be the displacements of the masses m_1 and m_2 at time t measured respectively from the positions of m_1 and m_2 where the springs are neither stretched nor compressed, i.e., when the system is at equilibrium.

1. Set up a system of differential equations involving x_1 and x_2 as unknown functions.

2. Solve for $x_1(t)$ and $x_2(t)$ using the special values $m_1 = 3$ and $m_2 = 2$, $k_1 = k_2 = 1$. Use initial displacements $x_1(0) = x_2(0) = 1$ and initial velocities $x_1'(0) = x_2'(0) = 0$.



Solution: 1. We use Newton's second law and Hooke's spring law for each of the masses to get

$$\begin{aligned} m_1 x_1'' &= -(k_1+k_2)\,x_1 + k_2 x_2 \\ m_2 x_2'' &= k_2 x_1 - k_2 x_2 \end{aligned}$$

2. For the given values we have

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

subject to

$$x_1(0) = 1, \quad x_2(0) = 1, \quad x_1'(0) = 0, \quad x_2'(0) = 0$$

We apply Laplace transforms to get

$$\begin{aligned} 3\left(s^2 X_1 - s\right) &= -2X_1 + X_2 \\ 2\left(s^2 X_2 - s\right) &= X_1 - X_2 \end{aligned}$$

We solve for X_1 and X_2 to get

$$\begin{aligned} X_1 &= \frac{5s+6s^3}{1+7s^2+6s^4} = \frac{5s+6s^3}{(6s^2+1)(s^2+1)} = \frac{1}{5}\frac{s}{s^2+1} + \frac{24}{5}\frac{s}{6s^2+1} \\ X_2 &= \frac{7s+6s^3}{1+7s^2+6s^4} = \frac{7s+6s^3}{(6s^2+1)(s^2+1)} = \frac{36}{5}\frac{s}{6s^2+1} - \frac{1}{5}\frac{s}{s^2+1} \end{aligned}$$

We apply the inverse transforms to get

$$\begin{aligned} x_1(t) &= \frac{1}{5}\cos t + \frac{4}{5}\cos\frac{t}{\sqrt{6}} \\ x_2(t) &= -\frac{1}{5}\cos t + \frac{6}{5}\cos\frac{t}{\sqrt{6}} \end{aligned}$$

Chapter 3

Dynamical Systems

3.1 Review: Linear Homogeneous Equations with Constant Coefficients

Let $y = y(x)$ be an unknown function of the independent variable x.

A nth order differential equation with constant coefficients is of the form

$$a_n \frac{d^n y}{dx^n} + a_{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1 \frac{dy}{dx} + a_0 y = f(x) \tag{3.1}$$

where all a_i are constants and $f(x)$ is a given function.

If $f(x)$ is nonzero (3.1) is called **nonhomogeneous**.

If the function $f(x)$ is the zero function, then we have a **homogeneous** differential equation:

$$a_n \frac{d^n y}{dx^n} + a_{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1 \frac{dy}{dx} + a_0 y = 0 \tag{3.2}$$

Equation (3.2) is called the **associated homogeneous** equation of (3.1).

In order to solve (3.2), we seek solutions of the form $y = e^{rx}$, where r is a constant. Given that

$$\frac{d^k}{dx^k}\left(e^{rx}\right) = r^k e^{rx}$$

substitution into (3.2) yields

$$\begin{aligned}
&a_n\left(r^n e^{rx}\right) + a_{n-1}\left(r^{n-1} e^{rx}\right) + \cdots + a_1\left(r e^{rx}\right) + a_0\left(e^{rx}\right) = 0 \\
&\Rightarrow e^{rx}\left(a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0\right) = 0 \\
&\Rightarrow a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0 = 0
\end{aligned}$$

So to find solutions of the form $y = e^{rx}$ it suffices to solve the *auxiliary* polynomial equation

$$a_n r^n + a_{n-1} r^{n-1} + \cdots + a_1 r + a_0 = 0 (3.3) \tag{3.3}$$

The roots of (3.3) can be real, or complex, or repeated.

For the special case of a second order equation the general solution is discussed the following theorem.

Theorem 3.1.1. *Let $y = y(x)$ be an unknown function*

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} + cy = 0 \tag{3.4}$$

with auxiliary

$$ar^2 + br + c = 0 \tag{3.5}$$

1. *If (3.5) has two distict real roots r_1 and r_2, then the general real solution of (3.4) is given by*

$$y(x) = c_1 e^{r_1 x} + c_2 e^{r_2 x}$$

for any constants c_1 and c_2.

2. *If (3.5) has a double real root r, then the general real solution of (3.4) is given by*

$$y(x) = c_1 e^{rx} + c_2 x e^{rx}$$

for any constants c_1 and c_2.

3. *If (3.5) has a complex conjugate pair of roots $a \pm ib$, then the general real solution of (3.4) is given by*

$$y(x) = c_1 e^{ax} \cos(bx) + c_2 e^{ax} \sin(bx)$$

for any constants c_1 and c_2.

Example 3.1.1. Solve $2y'' - 7y' + 3y = 0$.

Solution: We have $2r^2 - 7r + 3 = 0$, so $r = \frac{1}{2}, 3$. Hence,

$$y(x) = c_1 e^{x/2} + c_2 e^{3x}$$

Example 3.1.2. Solve $y'' - 8y' + 16y = 0$.

Solution: We have $r^2 - 8r + 16$, so $r = 4, 4$. Hence,

$$y(x) = c_1 e^{4x} + c_2 x e^{4x}$$

Example 3.1.3. $y'' - 8y' + 20y = 0$.

Solution: We have $r^2 - 8r + 20 = 0$, so $r = 4 - 2i, 4 + 2i$. Therefore,

$$y(x) = c_1 e^{4x} \cos(2x) + c_1 e^{4x} \sin(2x)$$

An initial value problem (IVP) is a set of differential equations and initial conditions (IC's).

Example 3.1.4. Solve the IVP.

$$y'' + 16y = 0, \quad y(\pi) = 3, \quad y'(\pi) = -4$$

Solution: We have $r^2 + 16 = 0$. So, $r = \pm 4i$. Hence, $y(x) = c_1 \cos(4x) + c_2 \sin(4x)$. Differentiate to get $y'(x) = -4c_1 \sin 4x + 4c_2 \cos 4x$. Now $y(\pi) = 3$ yields $c_1 = 3$ and $y'(\pi) = -4$ yields $c_2 = -\frac{4}{4} = -1$. So the solution is

$$y(x) = 3\cos(4x) - \sin(4x)$$

3.2 Systems of Ordinary Differential Equations

We recall that a **system of ordinary differential equations** is a set of differential equations in some unknown functions $x_1(t), x_2(t), \ldots, x_n(t)$ each being a function in the same independent variable t. Such a system is of n**th order** if it consists of n unknown functions and n equations. So we would have a system of the following type.

$$\begin{aligned} x_1' &= f_1(t, x_1, x_2, \ldots, x_n) \\ x_2' &= f_2(t, x_1, x_2, \ldots, x_n) \\ &\vdots \qquad\qquad \vdots \\ x_n' &= f_n(t, x_1, x_2, \ldots, x_n) \end{aligned} \tag{3.6}$$

Here the derivatives are with respect to t. So $x_i' = \frac{dx_i}{dt}$. The functions f_i are given or determined by the specific problem.

It is useful to think of the independent variable t as **time**.

A solution to a system is a set of functions that is a common solution to all its differential equations.

In the special case where the functions f_i do not depend on t, the system takes the form

$$\begin{aligned} x_1' &= f_1(x_1, x_2, \ldots, x_n) \\ x_2' &= f_2(x_1, x_2, \ldots, x_n) \\ &\vdots \qquad\quad \vdots \\ x_n' &= f_n(x_1, x_2, \ldots, x_n) \end{aligned} \tag{3.7}$$

and it is called **autonomous**.

Systems of the form (3.6) or (3.7) can rarely be solved in terms of known functions. However, one must try to develop solution methods such as approximations of solutions or qualitative methods of solutions, because such systems are ubiquitous in every day life and science.

The area of dynamical systems studies the qualitative behavior of solutions of differential systems. For example, we may be interested in finding the orbits of three objects under Newtonian gravity. This problem is very difficult to solve. However, we may are able to say what the long term behavior of the orbits are depending on various initial conditions. This very important point of view of solutions of systems was developed by Henry Poincaré. In this point of view geometry plays an important role.

Linear Systems

The only systems for which there are satisfactory methods of solution are the linear systems.

A system is **linear**, if the functions f_i are linear in $x_1, \ldots, x_n$. This means that the system is of form

$$\begin{aligned} x_1' &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + g_1(t) \\ x_2' &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n + g_2(t) \\ &\vdots \qquad\qquad\qquad\quad \vdots \\ x_n' &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + g_n(t) \end{aligned} \tag{3.8}$$

In the special case where all $g_i(t)$ are the zero function then the system is called **homogeneous**.

$$\begin{aligned} x_1' &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n \\ x_2' &= a_{21}(t)x_1 + a_{22}(t)x_2 + \cdots + a_{2n}(t)x_n \\ &\vdots \qquad\qquad\quad \vdots \\ x_n' &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n \end{aligned} \tag{3.9}$$

Converting a Higher Order Equation to a System

An nth order linear differential equation can be converted to nth order system, by introducing as many unknown new functions as different derivatives present.

For example, the equation

$$y''' - \sin(t)y'' + 2y' - ty = e^t$$

can be converted to a third order system by introducing new unknown functions $x_i(t)$ by $x_1 = y$, $x_2 = y'$, $x_3 = y''$. Then we have the following equivalent system

$$\begin{aligned} x_1' &= x_2 \\ x_2' &= x_3 \\ x_3' &= tx_1 - 2x_2 + \sin(t)x_3 + e^t \end{aligned} \tag{3.10}$$

Often there is merit in converting a differential equation into a system. For example, there are several methods of approximating numerically solutions of systems.

Homogeneous Linear Systems with Constant Coefficients

Let $x_1(t), \ldots, x_n(t)$ be unknown functions in t. A **homogeneous nth order linear system of n differential equations with constant coefficients** is a set of linear differential equations of the form.

$$\begin{aligned} x_1' &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ x_2' &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ &\vdots \qquad\qquad\qquad \vdots \\ x_n' &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{aligned} \tag{3.11}$$

where a_{ij} are given constants called the **coefficients** of the system. Such a system is **autonomous**, since A does not depend on t.

Our goal is solve such a system for the unknown functions $x_i(t)$ over some interval I. It is convenient to use matrices to study such systems.

If

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & & a_{nn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad D\mathbf{x} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}$$

then System (3.11) is written in matrix notation as

$$D\mathbf{x} = A\mathbf{x} \tag{3.12}$$

Example 3.2.1. Consider the homogeneous system.

$$\begin{aligned} \frac{dx_1}{dt} &= -2x_1 + 5x_2 \\ \frac{dx_2}{dt} &= 5x_1 - 2x_2 \end{aligned}$$

(a) Write the system in matrix form.

(b) Show that $\mathbf{h}_1(t) = \begin{bmatrix} -e^{-7t} \\ e^{-7t} \end{bmatrix}$ and $\mathbf{h}_2(t) = \begin{bmatrix} e^{3t} \\ e^{3t} \end{bmatrix}$ are solutions.

Solution: (a) We have

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} -2 & 5 \\ 5 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

(b) $\mathbf{h}_1(t)$ is a solution, because

$$\begin{bmatrix} (-e^{-7t})' \\ (e^{-7t})' \end{bmatrix} = \begin{bmatrix} 7e^{-7t} \\ -7e^{-7t} \end{bmatrix} = \begin{bmatrix} -2 & 5 \\ 5 & -2 \end{bmatrix} \begin{bmatrix} -e^{-7t} \\ e^{-7t} \end{bmatrix}$$

Similarly, it can be shown that $\mathbf{h}_2(t)$ is also a solution.

Theorem 3.2.1. *Any linear combination of solutions of system (3.12) is also a solution.*

Proof: Let $\mathbf{h}_1(t), \ldots, \mathbf{h}_k(t)$ be solutions of system (3.12). Hence, $D\mathbf{h}_i(t) = A\mathbf{h}_i(t)$. Let $\mathbf{h}(t) = c_1\mathbf{h}_1(t) + \cdots + c_k\mathbf{h}_k(t)$, be any linear combination. We have

$$\begin{aligned} D\mathbf{h}(t) &= D\left(c_1\mathbf{h}_1(t) + \cdots + c_k\mathbf{h}_k(t)\right) \\ &= c_1 D\mathbf{h}_1(t) + \cdots + c_k D\mathbf{h}_k(t) \\ &= c_1 A\mathbf{h}_1(t) + \cdots + c_k A\mathbf{h}_k(t) \\ &= A\left(c_1\mathbf{h}_1(t) + \cdots + c_k\mathbf{h}_k(t)\right) \\ &= A\mathbf{h}(t) \end{aligned}$$

Therefore, $\mathbf{h}(t)$ is a solution. $\square$

To solve system (3.12) we mimic the case of equation (2.7) and seek exponential solutions. Let us find solutions of the form $\mathbf{h}(t) = e^{\lambda t}\mathbf{v}$ for some scalar λ and some fixed vector $\mathbf{v}$. Then

$$\begin{aligned} & D\left(e^{\lambda t}\mathbf{v}\right) = A\left(e^{\lambda t}\mathbf{v}\right) \\ \Leftrightarrow\ & \lambda e^{\lambda t}\mathbf{v} = e^{\lambda t}A\mathbf{v} \\ \Leftrightarrow\ & A\mathbf{v} = \lambda\mathbf{v} \end{aligned}$$

So, $\mathbf{v}$ is an eigenvector of A and λ is the corresponding eigenvalue.

We conclude that we can find solutions of the form $\mathbf{h}(t) = e^{\lambda t}\mathbf{v}$ where $\mathbf{v}$ is an eigenvector of A and λ is the corresponding eigenvalue.

Now that we know how to find solutions, the question arises: how do we find all possible solutions or the general solution? By a general solution we mean a formula that encompasses all possible solutions as special cases.

It is a basic fact in the theory of differential equations that if $\mathbf{h}_1(t), \ldots, \mathbf{h}_n(t)$ are n linearly independent solutions of $\mathbf{x}' = A\mathbf{x}$ (A is $n \times n$) on an interval I, then

$$\mathbf{h}(t) = c_1\mathbf{h}_1(t) + \cdots + c_n\mathbf{h}_n(t)$$

is the general solution. A set of vector functions $\mathbf{h}_1(t), \ldots, \mathbf{h}_n(t)$ is **linearly independent** on an interval I if

$$c_1\mathbf{h}_1(t) + \cdots + c_n\mathbf{h}_n(t) = \mathbf{0} \quad \text{all } t \in I \quad \Longrightarrow \quad c_1 = \cdots = c_n = 0$$

Given this fact we can find the general solution in the case when A is diagonalizable. This is done in the following theorem. First recall that if A is diagonalizable, then A has n linearly independent eigenvectors.

Theorem 3.2.2. *Let A be a $n \times n$ diagonalizable matrix and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be linearly independent eigenvectors of A. If $\lambda_1, \ldots, \lambda_n$ are the corresponding eigenvalues, then the general solution on any interval I of the homogeneous linear system*

$$D\mathbf{x} = A\mathbf{x}$$

is given by

$$\mathbf{h}(t) = c_1e^{\lambda_1 t}\mathbf{v}_1 + \cdots + c_ne^{\lambda_n t}\mathbf{v}_n$$

Proof: It is sufficient to show that $e^{\lambda_1 t}\mathbf{v}_1, \ldots, e^{\lambda_n t}\mathbf{v}_n$ are linearly independent solutions on the interval I. Let c_i be constants such that

$$c_1e^{\lambda_1 t}\mathbf{v}_1 + \cdots + c_ne^{\lambda_n t}\mathbf{v}_n = \mathbf{0}$$

Then

$$c_1 e^{\lambda_1 t} = 0, \ldots, c_n e^{\lambda_n t} = 0$$

because the eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are linearly independent. But $e^{\lambda_i t} \neq 0$, therefore

$$c_1 = \cdots = c_n = 0$$

So, $e^{\lambda_1 t}\mathbf{v}_1, \ldots, e^{\lambda_n t}\mathbf{v}_n$ are linearly independent. □

Example 3.2.2. First find the general solution of the system of Example 3.2.1. Then find the particular solution with initial values $x_1(0) = 1$ and $x_2(0) = -2$.

Solution: The coefficient matrix $\begin{bmatrix} -2 & 5 \\ 5 & -2 \end{bmatrix}$ is diagonalizable with eigenvalues and corresponding eigenvectors:

$$\lambda_1 = -7,\ \mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \lambda_2 = 3,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Hence, the general solutions is

$$\mathbf{h}(t) = c_1 e^{-7t} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + c_2 e^{3t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Therefore,

$$x_1(t) = -c_1 e^{-7t} + c_2 e^{3t}$$
$$x_2(t) = c_1 e^{-7t} + c_2 e^{3t}$$

To find the solution with $x_1(0) = -1$ and $x_2(0) = 2$, we need to find c_1 and c_2 such that the initial condition is satisfied.

$$\mathbf{h}(0) = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Solving the linear system for c_1, c_2 yields $c_1 = -\frac{1}{2}$, $c_2 = -\frac{3}{2}$. So, we get the following functions (Fig. 2).

$$x_1(t) = \frac{3}{2}e^{-7t} - \frac{1}{2}e^{3t}$$
$$x_2(t) = -\frac{3}{2}e^{-7t} - \frac{1}{2}e^{3t}$$

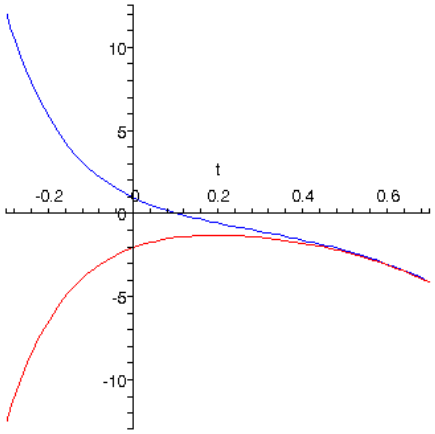


Figure 3.1: The solutions with $x_1(0) = -1$ and $x_2(0) = 2$ in Example 3.2.2.

It is often illuminating to consider a pair of solutions $(x_1(t), x_2(t))$ of a 2×2 system as the points of a parametric curve. We vary t and plot the pairs $(x_1(t), x_2(t))$. Such a curve is called a **solution curve** or an **integral curve** or a **trajectory** of the system.

The pair of solutions found in Example 3.2.2 is depicted as a trajectory in Fig. 3.2.

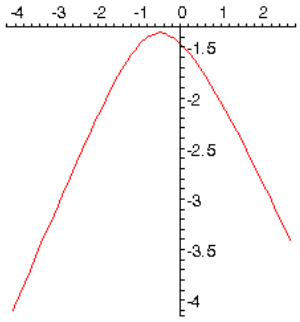


Figure 3.2: The unique pair of solutions in Example 3.2.2 as a trajectory of the system.

Example 3.2.3. Find the general solution of the system.

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} -2 & 0 & 5 \\ 1 & 1 & 0 \\ 5 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Solution: The coefficient matrix has the following eigenvalues and corresponding eigenvectors.

$$\lambda_1 = 1,\ \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \lambda_2 = 3,\ \mathbf{v}_2 = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}, \quad \lambda_3 = -7,\ \mathbf{v}_3 = \begin{bmatrix} -8 \\ 1 \\ 8 \end{bmatrix}$$

All eigenvectors are linearly independent, so by Theorem 3.2.2, the general solution is

$$\mathbf{h}(t) = c_1 e^t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c_2 e^{3t} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} + c_3 e^{-7t} \begin{bmatrix} -8 \\ 1 \\ 8 \end{bmatrix}$$

In individual functions we have

$$\begin{aligned} x_1(t) &= 2c_2 e^{3t} - 8c_3 e^{-7t} \\ x_2(t) &= c_1 e^t + c_2 e^{3t} + c_3 e^{-7t} \\ x_3(t) &= 2c_2 e^{3t} + 8c_3 e^{-7t} \end{aligned}$$

Complex Eigenvalues

Theorem 3.2.2 is valid whether or not the eigenvalues are real. In the case of complex eigenvalues we get complex exponential solutions $e^{\lambda_i t}\mathbf{v}_i$. The complex exponential e^z is defined by using **Euler's Formula**. If $z = a + ib$, then

$$e^z = e^{a+ib} = e^a(\cos b + i \sin b)$$

Example 3.2.4. Find the real and imaginary parts of the vector function $\mathbf{h}(t)$.

$$\mathbf{h}(t) = e^{it} \begin{bmatrix} 1 \\ 1 - i \end{bmatrix}$$

Solution: By Euler's formula $e^{it} = \cos t + i \sin t$. Hence,

$$\mathbf{h}(t) = (\cos t + i \sin t) \begin{bmatrix} 1 \\ 1 - i \end{bmatrix} = \begin{bmatrix} \cos t \\ \cos t + \sin t \end{bmatrix} + i \begin{bmatrix} \sin t \\ \sin t - \cos t \end{bmatrix}$$

Therefore,

$$\mathrm{Re}(\mathbf{h}(t)) = \begin{bmatrix} \cos t \\ \cos t + \sin t \end{bmatrix}, \quad \mathrm{Im}(\mathbf{h}(t)) = \begin{bmatrix} \sin t \\ \sin t - \cos t \end{bmatrix}$$

Example 3.2.5. Find the complex general solution of the homogeneous system.

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Solution: The coefficient matrix has eigenvalues and eigenvectors

$$\lambda_1 = i,\ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 1-i \end{bmatrix}, \quad \lambda_2 = -i,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1+i \end{bmatrix}$$

So, the general solution for complex scalars c_1 and c_2 is

$$\mathbf{h}(t) = c_1 e^{it} \begin{bmatrix} 1 \\ 1-i \end{bmatrix} + c_2 e^{-it} \begin{bmatrix} 1 \\ 1+i \end{bmatrix}$$

We are often interested in the real solutions of a system with real coefficient matrix that has complex eigenvalues. To get real solutions we simply take the real and imaginary parts of the complex solutions. This is expressed in the following theorem whose proof is omitted.

Theorem 3.2.3. *Let A be a $n \times n$ diagonalizable matrix. Then the homogeneous linear system*

$$D\mathbf{x} = A\mathbf{x}$$

has n linearly independent real solutions yielding the general solution on I. These can be obtained as follows. For each real eigenvalue λ with eigenvector $\mathbf{v}$ we use $e^{\lambda t}\mathbf{v}$. For each complex conjugate nonreal pair $\lambda, \overline{\lambda}$ we use the real and imaginary parts of the complex solution $e^{\lambda t}\mathbf{v}$, or of the complex solution $e^{\overline{\lambda} t}\overline{\mathbf{v}}$, but not both.

Example 3.2.6. Find the real general solution of the system of Example 3.2.5.

Solution: In Example 3.2.5, we found a complex conjugate pair eigenvalues $\lambda = \pm i$. We choose one, say $\lambda = i$, and from the corresponding complex solution

$$e^{it} \begin{bmatrix} 1 \\ 1-i \end{bmatrix}$$

found in Example 3.2.5, we compute the real and imaginary parts. This computation was carried out in Example 3.2.4. So the real general solution is

$$\mathbf{h}(t) = c_1 \begin{bmatrix} \cos t \\ \cos t + \sin t \end{bmatrix} + c_2 \begin{bmatrix} \sin t \\ \sin t - \cos t \end{bmatrix}$$

for real scalars c_1 and c_2. Some trajectories of these solutions are shown in Fig. 3.3.

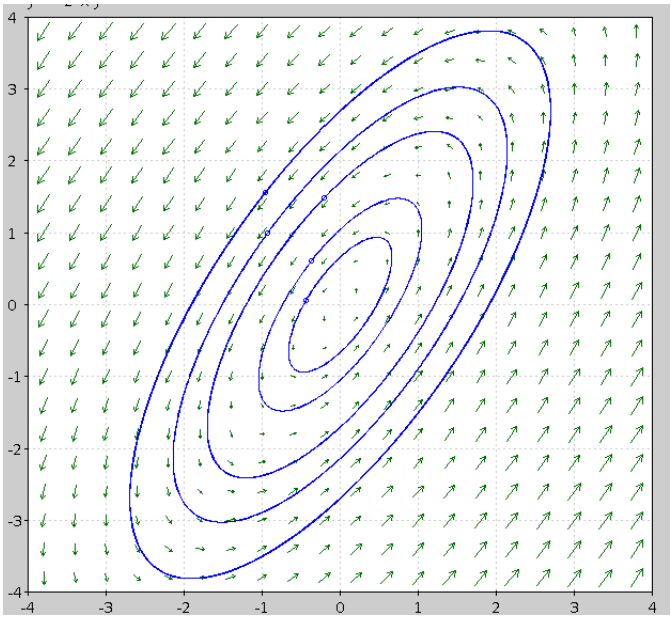


Figure 3.3: Some trajectories of the system in Example 3.2.6.

Example 3.2.7. Find the real general solution of the system.

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Solution: The coefficient matrix has complex eigenvalues and eigenvectors.

$$\lambda_1 = -1 + 2i,\ \mathbf{v}_1 = \begin{bmatrix} 1 \\ -i \end{bmatrix}, \quad \lambda_2 = -1 - 2i,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ i \end{bmatrix}$$

We only need the $(\lambda_1, \mathbf{v}_1)$ and compute the real and imaginary parts of $e^{\lambda_1 t}\mathbf{v}_1$.

We have

$$\begin{aligned} e^{(-1+2i)t}\begin{bmatrix} 1 \\ -i \end{bmatrix} &= e^{-t}e^{i2t}\begin{bmatrix} 1 \\ -i \end{bmatrix} \\ &= e^{-t}(\cos(2t)+i\sin(2t))\begin{bmatrix} 1 \\ -i \end{bmatrix} \\ &= e^{-t}\begin{bmatrix} \cos(2t)+i\sin(2t) \\ \sin(2t)-i\cos(2t) \end{bmatrix} \\ &= \begin{bmatrix} e^{-t}\cos(2t) \\ e^{-t}\sin(2t) \end{bmatrix} + i\begin{bmatrix} e^{-t}\sin(2t) \\ -e^{-t}\cos(2t) \end{bmatrix} \end{aligned}$$

Using the real and imaginary parts we get the real general solution

$$\mathbf{h}(t) = c_1\begin{bmatrix} e^{-t}\cos(2t) \\ e^{-t}\sin(2t) \end{bmatrix} + c_2\begin{bmatrix} e^{-t}\sin(2t) \\ -e^{-t}\cos(2t) \end{bmatrix}$$

where c_1 and c_2 are any real constants. Some trajectories of these solutions are shown in Fig. 3.4.

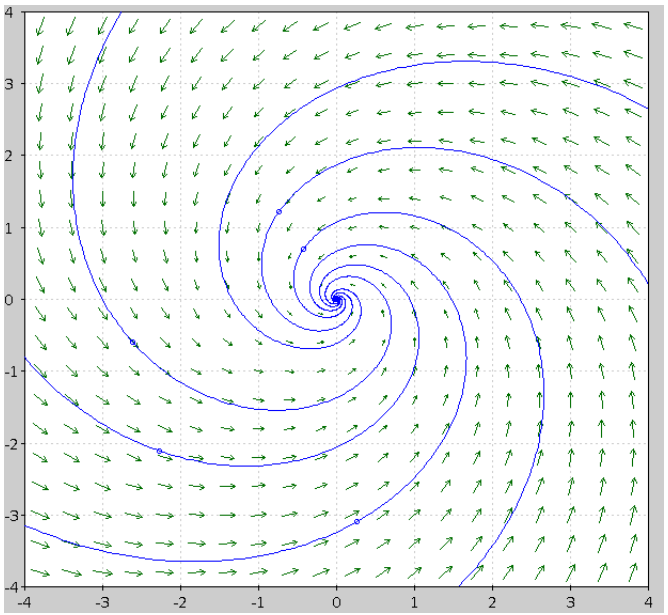


Figure 3.4: Some trajectories of the system in Example 3.2.7.

Example 3.2.8. Find the real general solution of the system.

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \\ 2 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Solution: The coefficient matrix has eigenvalues and eigenvectors.

$$\lambda_1 = 2,\ \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \lambda_2 = i,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1-i \end{bmatrix}, \quad \lambda_3 = -i,\ \mathbf{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 1+i \end{bmatrix}$$

From the pair $i, -i$ we keep $i, \mathbf{v}_2$ and compute the real and imaginary parts of $e^{it}\mathbf{v}_2$. We get the following general solution.

$$\mathbf{h}(t) = c_1 \begin{bmatrix} 0 \\ e^{2t} \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} \cos t \\ 0 \\ \cos t + \sin t \end{bmatrix} + c_3 \begin{bmatrix} \sin t \\ 0 \\ \sin t - \cos t \end{bmatrix}$$

Fig. 3.5 below shows the trajectory space curve $(x_1(t), x_2(t), x_3(t))$ with $c_1 = c_2 = c_3 = 1$.

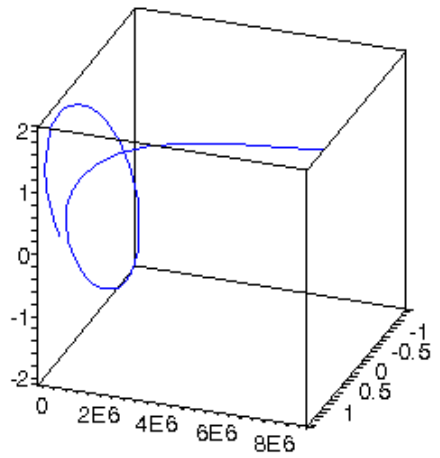


Figure 3.5: One trajectory of the system in Example 3.2.8.

NOTE

An **attractor** is a point to which trajectories are attracted. A **repeller** is a point from which all trajectories move away. A **saddle point** is a point that attracts some trajectories and repels others.

3.3 Phase Portraits of Linear Systems

In this section we concentrate on the second order linear homogeneous system with constant coefficients

$$\begin{aligned} \frac{dx}{dt} &= a_1 x + b_1 y \\ \frac{dy}{dt} &= a_2 x + b_2 y \end{aligned} \tag{3.13}$$

which for convenience we also write in matrix form as

$$D\mathbf{x} = A\mathbf{x}$$

where

$$D\mathbf{x} = \begin{bmatrix} \frac{dx}{dt} \\ \frac{dy}{dt} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}, \qquad A = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$$

Our discussion focuses on second order systems only for clarity and to be able to get graphical solutions easily. The ideas, however, apply to systems of higher order.

In Section 3.2 we introduced the concept of a **trajectory** or an **integral curve** of a system. This is a directed curve with components the unknown functions. In our case, this is the vector valued function

$$\mathbf{x}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$$

which is considered as a parametrized curve.

Our first goal is to try to sketch **graphs** of trajectories of the system. Integral curves provide much less information than actually finding exact solutions of the system but they can be very useful. An exact solution would not only give us all the points of a trajectory but also the time at which we have arrived at a given point.

Our second goal is to try to determine the **direction** in which the trajectory is traversed.

A set of directed trajectories of a system is called the **phase portrait** of the system.

Phase portraits of systems yield information about the qualitative behavior of solutions of systems such as the long term behavior of the solutions.

In this section we concentrate on the system (3.13). First we have the following fact.

Fact Distinct trajectories of the constant coefficient system $D\mathbf{x} = A\mathbf{x}$ do not intersect. Furthermore, if $\mathbf{x}(t)$ is a solution, then $\mathbf{x}(t + t_0)$ is also a solution and these two solutions determine the same trajectory: if $\mathbf{x}(t)$ reaches a point at time $t = t_1$, then $\mathbf{x}(t + t_0)$ reaches the same point at time $t = t_1 - t_0$.

Real Eigenvalues

Example 3.3.1. Sketch the phase portrait of $D\mathbf{x} = A\mathbf{x}$ with

$$A = \begin{bmatrix} 1 & 3 \\ 1 & -1 \end{bmatrix}$$

Solution: The eigenvalues and corresponding eigenvectors of A are

$$\lambda_1 = -2,\ \mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \lambda_2 = 2,\ \mathbf{v}_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

and determined the general solutions to be

$$\mathbf{h}\,(t) = c_1 e^{-2t} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + c_2 e^{2t} \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

The solutions that correspond to the eigenvalues and eigenvectors are

$$\mathbf{x}_1(t) = e^{\lambda_1 t}\mathbf{v}_1 \quad \text{and} \quad \mathbf{x}_2(t) = e^{\lambda_2 t}\mathbf{v}_2$$

As t takes on any real values, these vector-valued functions will cover all positive multiples of $\mathbf{v}_1$ and $\mathbf{v}_2$. So these trajectories are **half lines** determined by the eigenvectors. The opposites of these solutions are also solutions, so the half lines opposite to the eigenvectors are also trajectories. Now *the origin is also a trajectory by itself*. So each of these two eigenlines defines trajectories. However, each line consists of **three** distinct trajectories: the **two half lines** and the **origin**.

We may determine how these trajectories are traced. For $\mathbf{x}_1(t) = e^{\lambda_1 t}\mathbf{v}_1$, since $\lambda_1 = -2 < 0$, as t goes to ∞, $\mathbf{x}_1(t)$ approaches $(0, 0)$ along the eigenline of $\mathbf{v}_1$. As t goes to $-\infty$, $\mathbf{x}_1(t)$ comes goes out to infinite distance from the origin along this eigenline. So as time t increases the trajectory approaches the origin along this eigenline.

The opposite is true with the solution $\mathbf{x}_2(t) = e^{\lambda_2 t}\mathbf{v}_2$. Since $\lambda_2 = 2 > 0$, as t goes to ∞, $\mathbf{x}_2(t)$ moves away from the origin along the eigenline of $\mathbf{v}_2$.

These observations are depicted in Fig. 3.6.

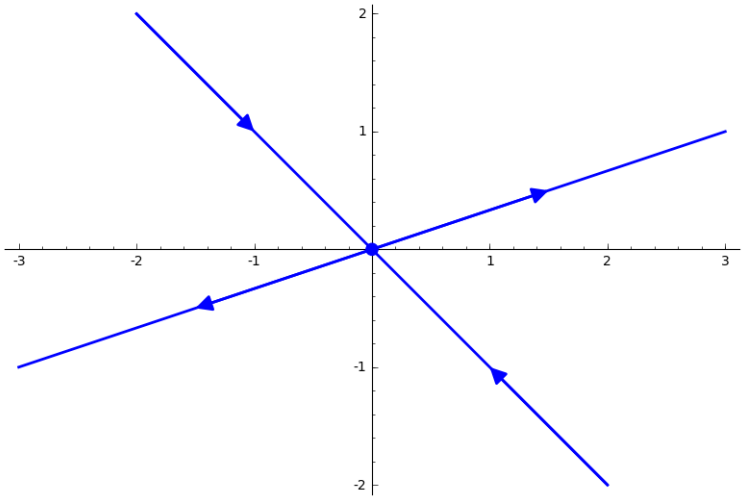


Figure 3.6: The five trajectories defined by the eigenvectors in Example 3.3.1.

The remaining of the solutions are of the form

$$\begin{aligned}\mathbf{x}(t) &= c_1 e^{\lambda_1 t}\mathbf{v}_1 + c_2 e^{\lambda_2 t}\mathbf{v}_2 \\ &= c_1 e^{-2t}\mathbf{v}_1 + c_2 e^{2t}\mathbf{v}_2\end{aligned}$$

for nonzero constants c_1 and c_2. As t increases the first term approaches zero whereas the second term grows. So the trajectory defined by $\mathbf{x}(t)$ is asymptotic to the eigenline of $\mathbf{v}_2$. To get a complete picture we should trace the curve $\mathbf{x}(t)$ backward. If we let $t \to -\infty$, then the first term grows and the second approaches zero. So the trajectory defined by $\mathbf{x}(t)$ is asymptotic to the eigenline of $\mathbf{v}_1$. Based on these observations we may get a phase portrait of our system as seen in Fig. 3.7.

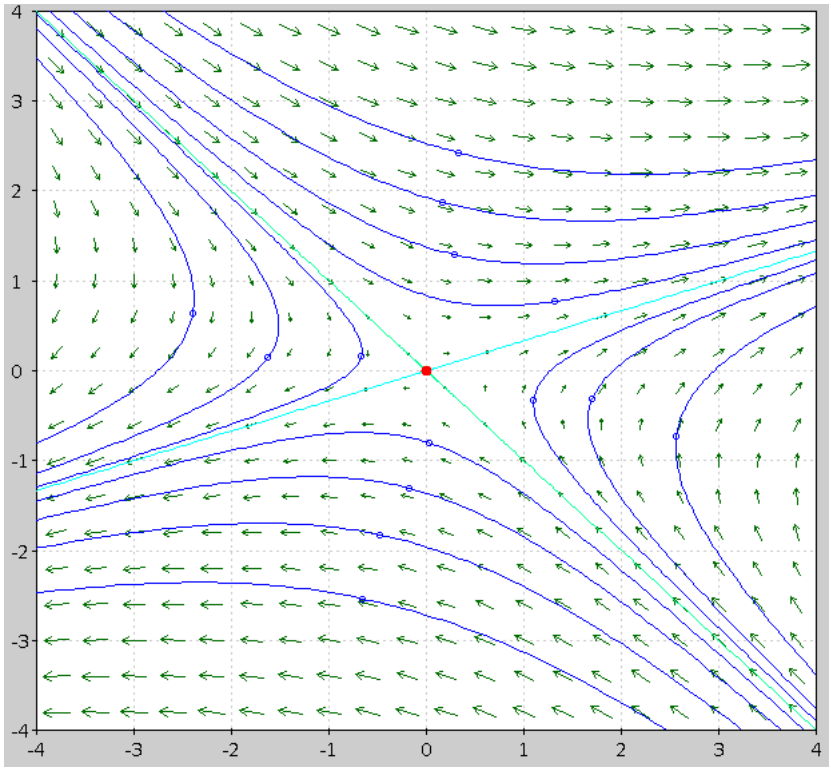


Figure 3.7: A phase portrait of $D\mathbf{x} = A\mathbf{x}$, $\lambda_1 < 0 < \lambda_2$.

□

Fig. 3.7 is a typical phase portrait of a second order system with A having opposite sign eigenvalues. We have the following theorems.

Theorem 3.3.1. *If $(\lambda, \mathbf{v})$ is an eigenpair of the matrix A with $\lambda \neq 0$, then the phase portrait of $D\mathbf{x} = A\mathbf{x}$ includes the line through the origin defined by $\mathbf{v}$. This line determines three trajectories: The two half lines and the origin.*

Theorem 3.3.2. *If λ_1 and λ_2 are opposite sign eigenvalues of the 2×2 matrix A, $\lambda_1 < 0 < \lambda_2$, then the phase portrait of $D\mathbf{x} = A\mathbf{x}$ includes two lines through the origin defined by the corresponding eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Every other trajectory is asymptotic to the eigenline of $\mathbf{v}_1$ as $t \to -\infty$ and asymptotic to the eigenline of $\mathbf{v}_2$ as $t \to \infty$.*

Example 3.3.2. Sketch the phase portrait of $D\mathbf{x} = A\mathbf{x}$ with

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

Solution: The eigenvalues and corresponding eigenvectors of A are

$$\lambda_1 = 1,\ \mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \lambda_2 = 5,\ \mathbf{v}_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

The eigenvectors are the same in Example 3.3.1. This time though both eigenvalues are positive. So the motion along the eigenlines is now away from the origin for all four half line trajectories. The one-point trajectory $\{(0,0)\}$ remains stationary. These observations are depicted in Fig. 3.8.

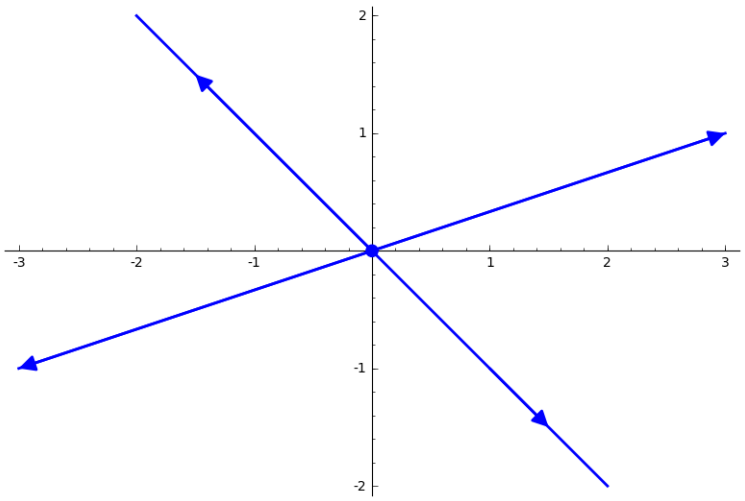


Figure 3.8: The five trajectories defined by the eigenvectors in Example 3.3.2.

The remaining of the solutions are of the form $\lambda_2 = 5$)

$$\begin{aligned} \mathbf{x}(t) &= c_1 e^{\lambda_1 t}\mathbf{v}_1 + c_2 e^{\lambda_2 t}\mathbf{v}_2 \\ &= e^{\lambda_2 t}(c_1 e^{(\lambda_1 - \lambda_2)t}\mathbf{v}_1 + c_2\mathbf{v_2}) \\ &= e^{5t}(c_1 e^{-4t}\mathbf{v}_1 + c_2\mathbf{v_2}) \end{aligned}$$

for nonzero constants c_1 and c_2. As t increases the first term inside the parentheses goes to zero whereas the second term remains constant, namely, $c_2\mathbf{v}_2$. So the slope of $\mathbf{x}(t)$ approaches the slope of $\mathbf{v}_2$. Therefore, as $t \to \infty$ the trajectories tend to become almost parallel to the eigenvector with the highest eigenvalue. The opposite happens as $t \to -\infty$: the trajectories tend to become parallel to the eigenvector with the smallest eigenvalue. The phase portrait of this system is seen in Fig. 3.9.

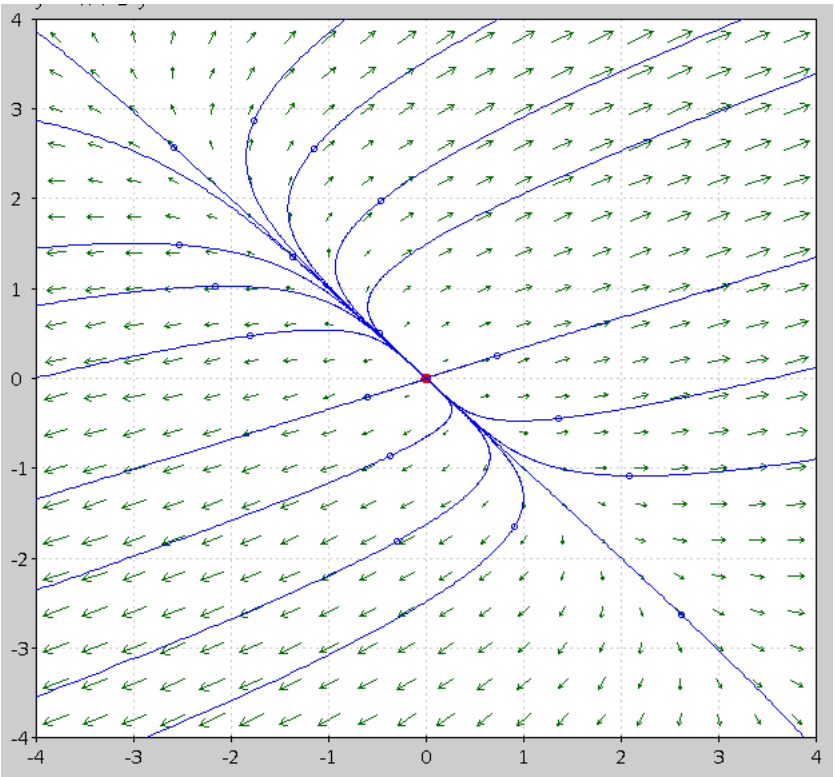


Figure 3.9: A phase portrait of $D\mathbf{x} = A\mathbf{x}$, $0 < \lambda_1 < \lambda_2$.

□

Fig. 3.8 is a typical phase portrait of a second order system with A having the same sign nonzero and not equal eigenvalues. If both eigenvalues are negative all trajectories have direction towards the origin. We have the following theorem.

Theorem 3.3.3. *If λ_1 and λ_2 are real nonzero eigenvalues such that $\lambda_1 < \lambda_2$, of the 2×2 matrix A, then the phase portrait of $D\mathbf{x} = A\mathbf{x}$ includes two lines through the origin defined by the corresponding eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$. Every other trajectory is asymptotic to the eigenline of $\mathbf{v}_1$ as $t \to -\infty$ and asymptotic to the eigenline of $\mathbf{v}_2$ as $t \to \infty$.*

Example 3.3.3. Sketch the phase portrait of $D\mathbf{x} = A\mathbf{x}$ with

$$A = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}$$

Solution: A has only one (repeated) eigenvalue and two corresponding linearly independent eigenvectors.

$$\lambda = -2, \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

So we have a repeated nonzero eigenvalue with two independent eigenvectors $\mathbf{v_1}$ and $\mathbf{v}_2$ span the plane, every nonzero vector is an eigenvector. So the phase portrait includes all the lines through the origin. The trajectories are all half lines that start at the origin plus the origin. The direction of each trajectory is towards the origin since the eigenvalue is negative. See Fig. 3.10.

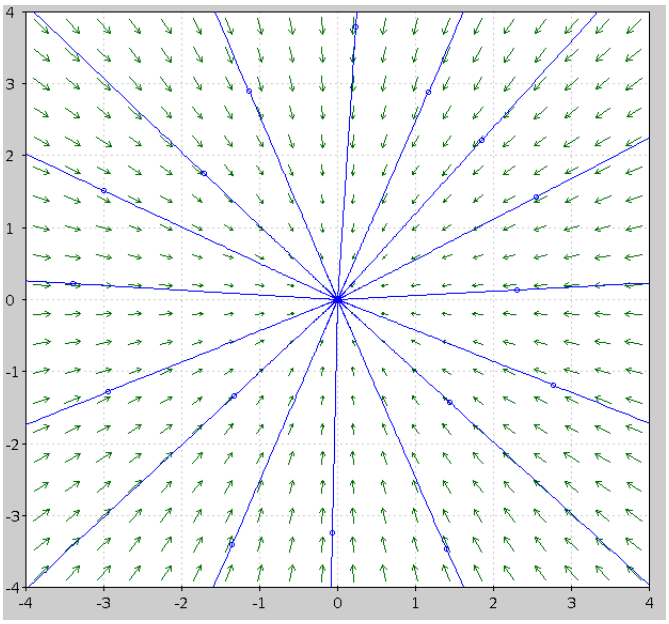


Figure 3.10: A phase portrait of $D\mathbf{x} = A\mathbf{x}$, with one repeated eigevalue and two linearly independent eigenvectors.

□

Fact Fig. 3.10 is typical to the case of one nonzero repeated eigenvalue with two linearly independent eigenvectors. If the eigenvalue is negative, the trajectories move towards the origin. If the eigenvalue is positive, then the trajectories move radially away from the origin.

Example 3.3.4. Sketch the phase portrait of $D\mathbf{x} = A\mathbf{x}$ with

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

Solution: There is only one repeated eigenvalue and one linearly independent eigenvector, namely

$$\lambda = -1, \ \mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

In this case an analysis similar of Example 3.3.2 shows that the phase portrait includes the only eigenline which consists of three trajectories: the two half lines and the origin. The remaining trajectories have slopes that become parallel to the eigenline both when $t \to -\infty$ and $t \to \infty$. Since the eigenvalue is negative, all trajectories approach the origin. See Fig. 3.11. To determine whether the trajectories approach the eigenline in a left-handed or a right-handed direction we find the motion at a point outside the eigenline. For example, when the trajectory reaches, say the point

$$\mathbf{x}(t) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

then

$$D\mathbf{x} = A\mathbf{x} = \begin{bmatrix} 0 & -1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

Since the top coordinate $-1 < 0$, the rate of change of the x-coordinate is negative

$$\frac{dx}{dt} = -1 < 0$$

so, the trajectory at this point is towards the left.

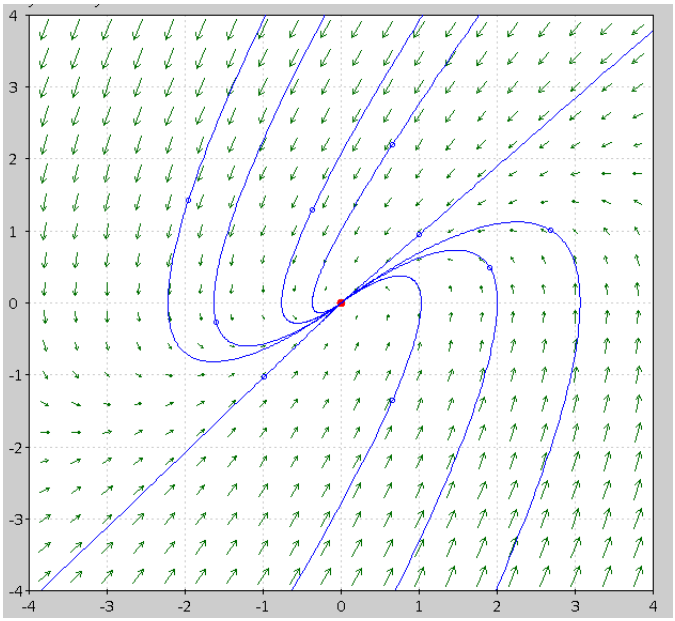


Figure 3.11: A phase portrait of $D\mathbf{x} = A\mathbf{x}$, with one nonzero repeated eigevalue and one linearly independent eigenvector.

□

Example 3.3.5. Sketch the phase portrait of $D\mathbf{x} = A\mathbf{x}$ with

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Solution: We leave it to the reader to see that the phase portrait is the one of Fig. 3.12.

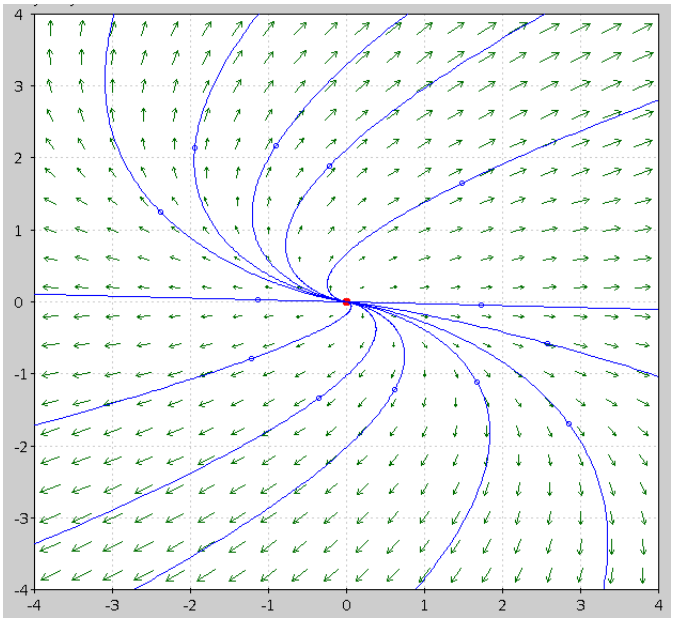


Figure 3.12: A phase portrait of $D\mathbf{x} = A\mathbf{x}$, with one nonzero repeated eigevalue and one linearly independent eigenvector.

□

Theorem 3.3.4. *If $\lambda \neq 0$ is the only nonzero eigenvalue of the 2×2 matrix A, then the phase portrait of $D\mathbf{x} = A\mathbf{x}$ has trajectories that are half lines radially moving towards the origin (if $\lambda < 0$) or away from the origin (if $\lambda > 0$), if there are two linearly independent eigenvectors of λ. If λ has only one linearly independent eigenvector, then the phase portrait includes the two rays of the eigenline and trajectories that have slopes that become parallel to the eigenline both when $t \to -\infty$ and $t \to \infty$. Also the origin defines a separate trajectory.*

For the extreme cases when one eigenvalue is zero we have the following theorem.

Theorem 3.3.5. *Let the 2×2 matrix A have 0 as an eigenvalue. Let $\mathbf{v}$ a corresponding eigenvector and let $L_{\mathbf{v}}$ be the eigenline of $\mathbf{v}$. Then for the phase portrait of system $D\mathbf{x} = A\mathbf{x}$, we have*

1. *Every single point of $L_{\mathbf{v}}$ is a trajectory.*

2. *If* 0 *is the only eigenvalue, then all other trajectories are lines parallel to* $L_{\mathbf{v}}$*. The trajectories in opposite sides of* $L_{\mathbf{v}}$ *move in opposite directions.*

3. *If* A *has another eigenvalue* $\lambda \neq 0$*, with eigenvector* $\mathbf{w}$*, then every trajectory off of* $L_{\mathbf{v}}$ *is a half line parallel to* $\mathbf{w}$ *with endpoint on* $L_{\mathbf{v}}$*. The motion of these trajectories is toward* $L_{\mathbf{v}}$*, if* $\lambda < 0$*, and away from it, if* $\lambda > 0$*.*

Example 3.3.6. Sketch the phase portrait of $D\mathbf{x} = A\mathbf{x}$ with

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Solution: The eigenvalues and corresponding eigenvectors of A are

$$\lambda_1 = 0,\ \mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \lambda_2 = 2,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Applying Theorem 3.3.5 we find the phase portrait of Fig. 3.13.

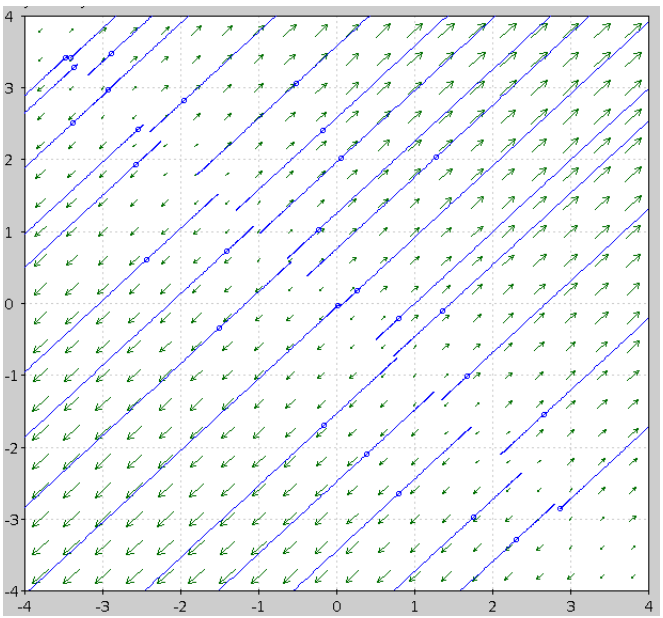


Figure 3.13: A phase portrait of $D\mathbf{x} = A\mathbf{x}$, with one zero eigevalue and one nonzero eigenvalue.

□

Complex Eigenvalues

Let us consider now the case that the 2×2 matrix A has a complex conjugate pair of eigenvalues. We have seen that two linearly independent real solutions are obtained by using only one eigenpair $(\lambda, \mathbf{v})$ and finding the real and imaginary parts of the complex function $e^{\lambda t}\mathbf{v}$ by using Euler's formula. If $\lambda = \alpha + i\beta$, we have

$$e^{(\alpha+i\beta)t}\mathbf{v} = e^{\alpha t}(\cos(\beta t) + i\sin(\beta t))\mathbf{v}$$

The last two factors turn the complex vector $\mathbf{v}$ around, whereas the real exponential $e^{\alpha t}$ scales it. If $\alpha = 0$, then there is no scaling, so the trajectory is a closed orbit around the origin. If $\alpha \neq 0$, then the trajectories spiral away or towards the origin.

Theorem 3.3.6. *Let the 2×2 matrix A have complex eigenvalues $\alpha \pm i\beta$.*

1. *If $\alpha \neq 0$, then the trajectories of $D\mathbf{x} = A\mathbf{x}$ spiral around the origin. If $\alpha < 0$, then the trajectories spiral towards the origin. If $\alpha > 0$, then the trajectories spiral away from the origin.*

2. *If $\alpha = 0$, then the trajectories of $D\mathbf{x} = A\mathbf{x}$ are closed loops (ellipses or circles) around the origin.*

Example 3.3.7. In Example 3.2.7 we found the real general solution of the system

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} -1 & -2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

to be

$$\mathbf{x}(t) = c_1 \begin{bmatrix} e^{-t}\cos(2t) \\ e^{-t}\sin(2t) \end{bmatrix} + c_2 \begin{bmatrix} e^{-t}\sin(2t) \\ -e^{-t}\cos(2t) \end{bmatrix}$$

where c_1 and c_2 are any real constants. We also found some trajectories of these solutions. Here we repeat the phase portrait found in Example 3.2.7 as an illustration of Theorem 3.3.6.

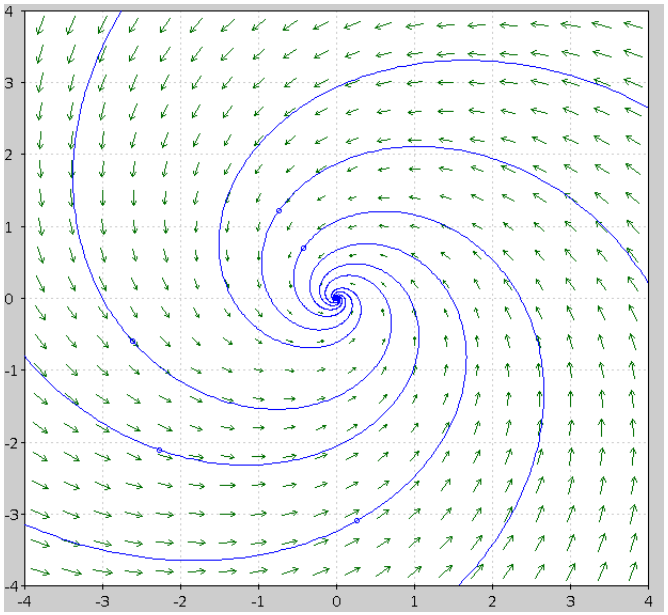


Figure 3.14: Some trajectories of the system in Example 3.3.7.

Example 3.3.8. In Example 3.2.6 we found the real general solution of the system

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

to be

$$\mathbf{h}(t) = c_1 \begin{bmatrix} \cos t \\ \cos t + \sin t \end{bmatrix} + c_2 \begin{bmatrix} \sin t \\ \sin t - \cos t \end{bmatrix}$$

for real scalars c_1 and c_2. Some trajectories of these solutions were shown in Fig. 3.3.We repeat this phase portrait here for convenience.

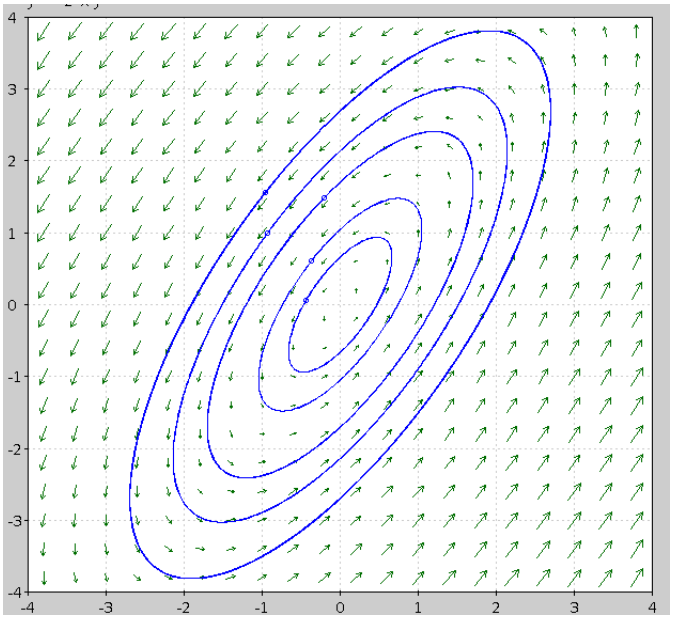


Figure 3.15: Some trajectories of the system in Example 3.3.8.

NOTE

Depending on the matrix A, the spirals and the closed trajectories may be traced either clockwise or counterclockwise.

3.4 Linearization and Stability

In Section 3.3 we studied phase portraits of linear systems. However, most systems we encounter in practice are nonlinear. A general second order autonomous system is of the form

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned} \tag{3.14}$$

We assume that the functions f and g have nice enough properties so that every initial value problem with differential equations (3.14) has a unique solution for all t in some interval I. At the end of this section we shall state a theorem that guarantees existence and uniqueness of solutions to such initial value problems.

In trying to understand the solutions of (3.14) the most important step is to find the **constant** solutions. A constant solution of the system (3.14), or more generally of the system (3.7), is called an **equilibrium**.

For a second order system, an equilibrium is a pair of numbers (a, b) such that for $x = a$ and $y = b$ the system (3.14) is satisfied. Often an equilibrium $\mathbf{c}$ is viewed as a point in the plane or as a vector from the origin to the point (a, b) and we write $\mathbf{c} = \begin{bmatrix} a \\ b \end{bmatrix}$.

Since an equilibrium (x, y) is a constant solution, $\frac{dx}{dt} = 0$ and $\frac{dy}{dt} = 0$. Hence, $f(x, y) = 0$ and $g(x, y) = 0$. So, to find the equilibria we solve for x and y the system

$$\begin{aligned} f(x, y) &= 0 \\ g(x, y) &= 0 \end{aligned} \tag{3.15}$$

Example 3.4.1. Find the equilibria of the system

$$\begin{aligned} \frac{dx}{dt} &= x - x^2 - 3xy \\ \frac{dy}{dt} &= y - y^2 - 2xy \end{aligned} \tag{3.16}$$

Solution: We solve the polynomial system

$$\begin{aligned} x - x^2 - 3xy &= x(1 - x - 3y) = 0 \\ y - y^2 - 2xy &= y(1 - y - 2x) = 0 \end{aligned}$$

Hence,

$$x = 0 \quad \text{or} \quad 1 - x - 3y = 0$$

and

$$y = 0 \quad \text{or} \quad 1 - y - 2x = 0$$

If $x = 0$, then either $y = 0$ or $y = 1$. If $1 - x - 3y = 0$, then $y = 0$ yields $x = 1$. If $1 - x - 3y = 0$, then $1 - y - 2x = 0$ yields $x = 2/5$ and $y = 1/5$. Hence, there are four equilibria, namely,

$$(0, 0), \quad (1, 0), \quad (0, 1), \quad \left(\frac{2}{5}, \frac{1}{5}\right)$$

□

Let $\mathbf{x} = \mathbf{c}$ be an equilibrium of (3.14), say

$$\mathbf{c} = \begin{bmatrix} a \\ b \end{bmatrix}$$

We have the following definitions.

1. We say that $\mathbf{c}$ is an **attractor**, if there is circle centered at the point (a, b) such that any solution $\mathbf{x}(t)$ that starts inside the circle it approaches $\mathbf{c}$ as $t \to \infty$.

2. We say that $\mathbf{c}$ is **stable**, if every solution that starts near $\mathbf{c}$ remains close to $\mathbf{c}$. An attractor is an example of a stable equilibrium.

3. If $\mathbf{c}$ has the property that there is a circle centered at $\mathbf{c}$ such that some solutions that start inside the circle leave and stay out of the circle, then $\mathbf{c}$ is called **unstable**.

4. If all solutions that are starting inside the circle (except $\mathbf{x} = \mathbf{c}$) eventually leave the circle and stay out, then $\mathbf{c}$ is called **repeller**. A repeller is an example of an unstable equilibrium.

The above concepts generalize to higher order systems. Instead of a circle we have a sphere or a higher dimensional version of a sphere.

The examples and theorems of Section 3.3 lead us to the following conclusions about the linear system $D\mathbf{x} = A\mathbf{x}$.

Theorem 3.4.1. *For the linear system $D\mathbf{x} = A\mathbf{x}$ we have*

1. *The zero vector $\mathbf{x} = \mathbf{0}$ is an equilibrium.*

2. *If all the eigenvalues of A have negative real part, then $\mathbf{x} = \mathbf{0}$ is an attractor.*

3. *If all the eigenvalues of A have positive real part, then $\mathbf{x} = \mathbf{0}$ is a repeller.*

4. *If A has some eigenvalues with positive real parts and some with negative real parts, or if there is a zero eigenvalue, or if there is a pure imaginary eigenvalue, then $\mathbf{x} = \mathbf{0}$ is neither an attractor nor a repeller.*

To analyze the phase portrait of a nonlinear system we begin with the equilibria. We try to find locally how the phase portrait looks around an equilibrium. At an equilibrium we **linearize** the system. i.e., we approximate it by a linear one and use its phase portrait.

Let us discuss some justification why such approach should work.

We use Taylor's expansion about the point (a, b) for the functions $f(x, y)$ and $g(x, y)$. We have

$$\begin{aligned} f(x,y) &= f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) + \gamma_1(x,y) \\ g(x,y) &= g(a,b) + g_x(a,b)(x-a) + g_y(a,b)(y-b) + \gamma_2(x,y) \end{aligned}$$

where γ_i are functions that very small compared to the distance from (x, y) to (a, b). More precisely

$$\lim_{(x,y)\to(a,b)} \frac{\gamma_i(x,y)}{((x-a)^2+(y-b)^2)^{1/2}} = 0, \qquad i = 1, 2$$

Here, we used the usual convention that f_x denotes the partial derivative $\frac{\partial f}{\partial x}$.

For (x, y) near (a, b) we approximate $f(x, y)$ and $g(x, y)$ by dropping the small functions γ_i. If in addition $\mathbf{c}$ is an equilibrium, thus $f(a, b) = 0$ and $g(a, b) = 0$, we get the **linear** system

$$\begin{aligned} \frac{dx}{dt} &= f_x(a,b)(x-a) + f_y(a,b)(y-b) \\ \frac{dy}{dt} &= g_x(a,b)(x-a) + g_y(a,b)(y-b) \end{aligned} \tag{3.17}$$

which we consider as an approximation of the original system near the equilibrium $\mathbf{c}$.

If we let $\mathbf{y} = \mathbf{x} - \mathbf{c}$ and let

$$A_{\mathbf{c}} = \begin{bmatrix} f_x(a,b) & f_y(a,b) \\ g_x(a,b) & g_y(a,b) \end{bmatrix}$$

then since $D\mathbf{x} = D\mathbf{y}$, the linear system that approximates (3.14) near $\mathbf{c}$ can take the form

$$D\mathbf{y} = A_{\mathbf{c}}\mathbf{y} \tag{3.18}$$

The matrix $A_{\mathbf{c}}$ is called the **linearization matrix** of (3.14) near $\mathbf{c}$.

The following important theorem tells us to what extend the phase portrait of the linearized system approximates that of the original system.

Theorem 3.4.2 (Hartman-Grobman Theorem)**.** *If the linearization matrix $A_{\mathbf{c}}$ has no zero or pure imaginary eigenvalues, then the phase portrait of (3.14) near the equilibrium* **c** *can be obtained from the phase portrait of the linearized system (3.18) by a* ***homeomorphism****, i.e., a continuous one-to-one and onto map that has a continuous inverse.*

Intuitively, two phase portraits being homeomorphic means that locally we can map one to the other such that trajectories map to trajectories and one is slightly distorted version of the other. Bending and warping is allowed. For example, straight lines may get curved. However, no ripping is allowed. For example, closed curves must remain closed.

Example 3.4.2. Discuss the equilibria and stability of the system

$$\begin{aligned} \frac{dx}{dt} &= x - x^2 - 3xy \\ \frac{dy}{dt} &= y - y^2 - 2xy \end{aligned}$$

Solution: In Example 3.4.1 we found the following equilibria.

$$(0,0), \quad (1,0), \quad (0,1), \quad \left(\frac{2}{5}, \frac{1}{5}\right)$$

To find the linearizations at the equilibria we first compute the matrix of partial derivatives, also known as the **jacobian** or the **jacobian matrix** for $f(x,y) = x - x^2 - 3xy$ and $g(x,y) = y - y^2 - 2xy$.

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} -2x - 3y + 1 & -3x \\ -2y & -2x - 2y + 1 \end{bmatrix}$$

At the equilibria, the linearization matrices and their eigenvalues, basic eigenvectors, and phase portraits obtained by the methods of Section 3.3 are as follows.

$$A_{(0,0)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \lambda = 1, \ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

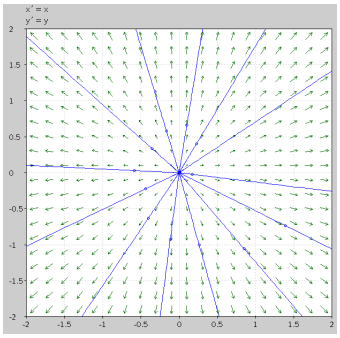


Figure 3.16: The phase portrait of $D\mathbf{y} = A_{(0,0)}\mathbf{y}$.

$$A_{(1,0)} = \begin{bmatrix} -1 & -3 \\ 0 & -1 \end{bmatrix}, \quad \lambda = -1, \ \mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

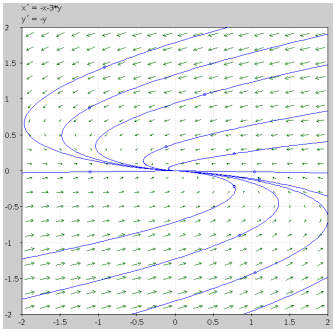


Figure 3.17: The phase portrait of $D\mathbf{y} = A_{(1,0)}\mathbf{y}$.

$$A_{(0,1)} = \begin{bmatrix} -2 & 0 \\ -2 & -1 \end{bmatrix}, \quad \lambda_1 = -2, \ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \lambda_2 = -1, \ \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

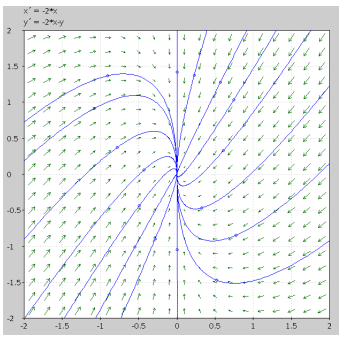


Figure 3.18: The phase portrait of $D\mathbf{y} = A_{(0,1)}\mathbf{y}$.

$$A_{(2/5,1/5)} = \begin{bmatrix} -\frac{2}{5} & -\frac{6}{5} \\ -\frac{3}{5} & -\frac{1}{5} \end{bmatrix}, \quad \lambda_1 = \frac{2}{5},\ \mathbf{v}_1 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \lambda_2 = -1,\ \mathbf{v}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

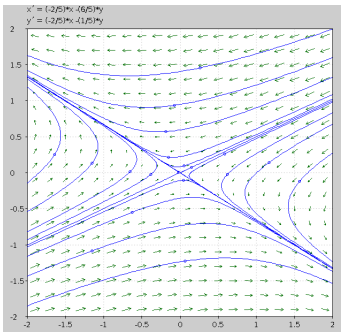


Figure 3.19: The phase portrait of $D\mathbf{y} = A_{(2/5,1/5)}\mathbf{y}$.

Since none of the eigenvalues was zero or pure imaginary the Hartman-Grobman Theorem tells us that the phase portrait of the original system near each equilibrium is similar to the phase portrait of the linearized system.

Since both eigenvalues of $A_{(0,0)}$ are positive, $(0, 0)$ is a repeller. All eigenvalues of $A_{(1,0)}$ and $A_{(0,1)}$ are negative, so the points $(1, 0)$ and $(0, 1)$ are both attractors. The eigenvalues of $A_{(2/5,1/5)}$ have opposite signs so the equilibrium is unstable but not a repeller.

In Figure 3.20 we have sketched the phase portrait of the given system by using computer software. Basically, the program draws a vector field of the solutions at several points and then follows the field to draw a few trajectories. We should notice how similar the phase portraits at the equilibria are to those of the linearized systems. One notable difference is that some straight lines are now curved.

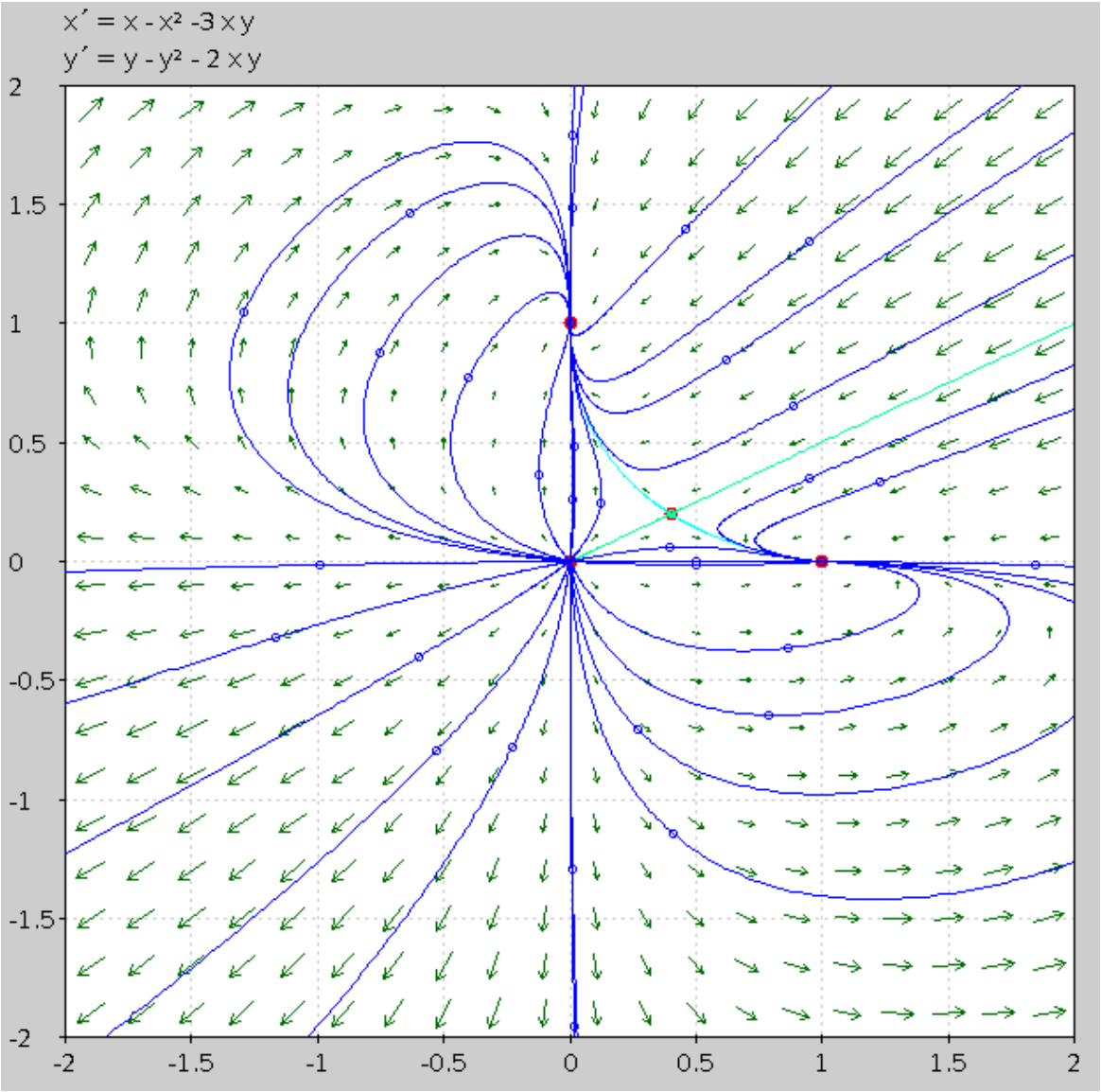


Figure 3.20: The phase portrait of the system in Example 3.4.2.

The figure shows that each trajectory that starts in a certain quadrant it remains in the quadrant at all times. In this example the coordinate axis define trajectories and since trajectories do not intersect, we expect each trajectory to be confined entirely in one quadrant.

We also observe that most solutions in the first quadrant tend to one of the two attractors $(1, 0)$ and $(0, 1)$. Actually, the only solutions that do

not tend to one of the attractors in the first quadrant are those that come from the other two equilibria $(0,0)$ and $(2/5,1/5)$ and two trajectories one going from $(0,0)$ to $(2/5,1/5)$ and one coming from "infinity" to $(2/5,1/5)$. These trajectories are exceptional in that they separate first quadrant into two regions in one of which the trajectories approach one attactor and in the other where the trajectories approach the other attractor. Because of this property each of these curves is called a **separatrix**. If we move backwards in time we see that most of the trajectories either come from infinity or from the repeller $(0,0)$. These two types of curves are separated by the two trajectories that go from $(2/5,1/5)$ to the attractors. These two exceptional trajectories are also called separatrices.

In quadrants 2, 3, and 4 the behavior of the trajectories is simpler. In the second quadrant trajectories that start about the x-axis approach the attractor $(0,1)$ as $t \to \infty$ and approach the repeller $(0,0)$ as $t \to -\infty$. As similar behavior is seen in the fourth quadrant with the attractor $(1,0)$ been the target of trajectories. In the third qudrant the solutions go out to infinity. □

Example 3.4.3. Discuss the equilibria and stability of the system

$$\frac{dx}{dt} = 4x - 4x^2 - xy$$
$$\frac{dy}{dt} = y - y^2 - 4xy$$

Solution: The equilibria are

$$(0,0), \quad (1,0), \quad (0,1)$$

The jacobian matrix is

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} -8x - y + 4 & -x \\ -4y & -4x - 2y + 1 \end{bmatrix}$$

At the equilibria, the linearization matrices and their eigenvalues, basic eigenvectors, and phase portraits are as follows.

$$A_{(0,0)} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \quad \lambda_1 = 1, \ \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \lambda_2 = 4, \ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

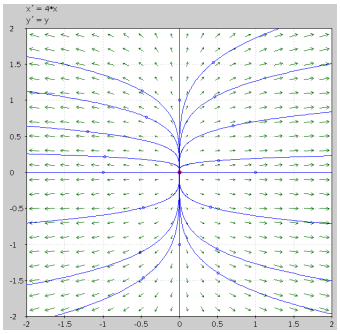


Figure 3.21: The phase portrait of $D\mathbf{y} = A_{(0,0)}\mathbf{y}$.

$$A_{(1,0)} = \begin{bmatrix} -4 & -1 \\ 0 & -3 \end{bmatrix}, \quad \lambda_1 = -4,\ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda_2 = -3,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

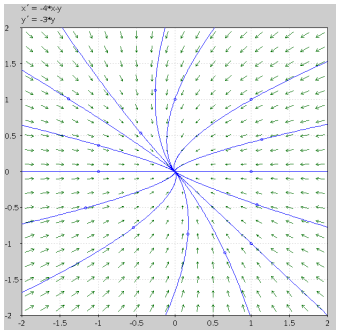


Figure 3.22: The phase portrait of $D\mathbf{y} = A_{(1,0)}\mathbf{y}$.

$$A_{(0,1)} = \begin{bmatrix} 3 & 0 \\ -4 & -1 \end{bmatrix}, \quad \lambda_1 = -1,\ \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \lambda_2 = 3,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

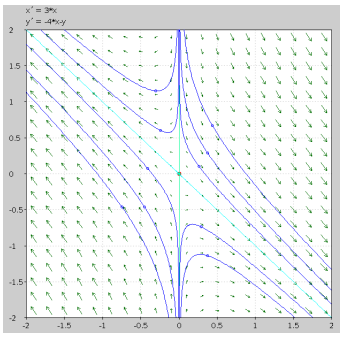


Figure 3.23: The phase portrait of $D\mathbf{y} = A_{(0,1)}\mathbf{y}$.

Since none of the eigenvalues was zero or pure imaginary the Hartman-Grobman Theorem implies that the phase portrait of the original system near each equilibrium is similar to the phase portrait of the linearized system.

By using graphical methods we get the phase portrait of the given system as seen in Figure 3.24.

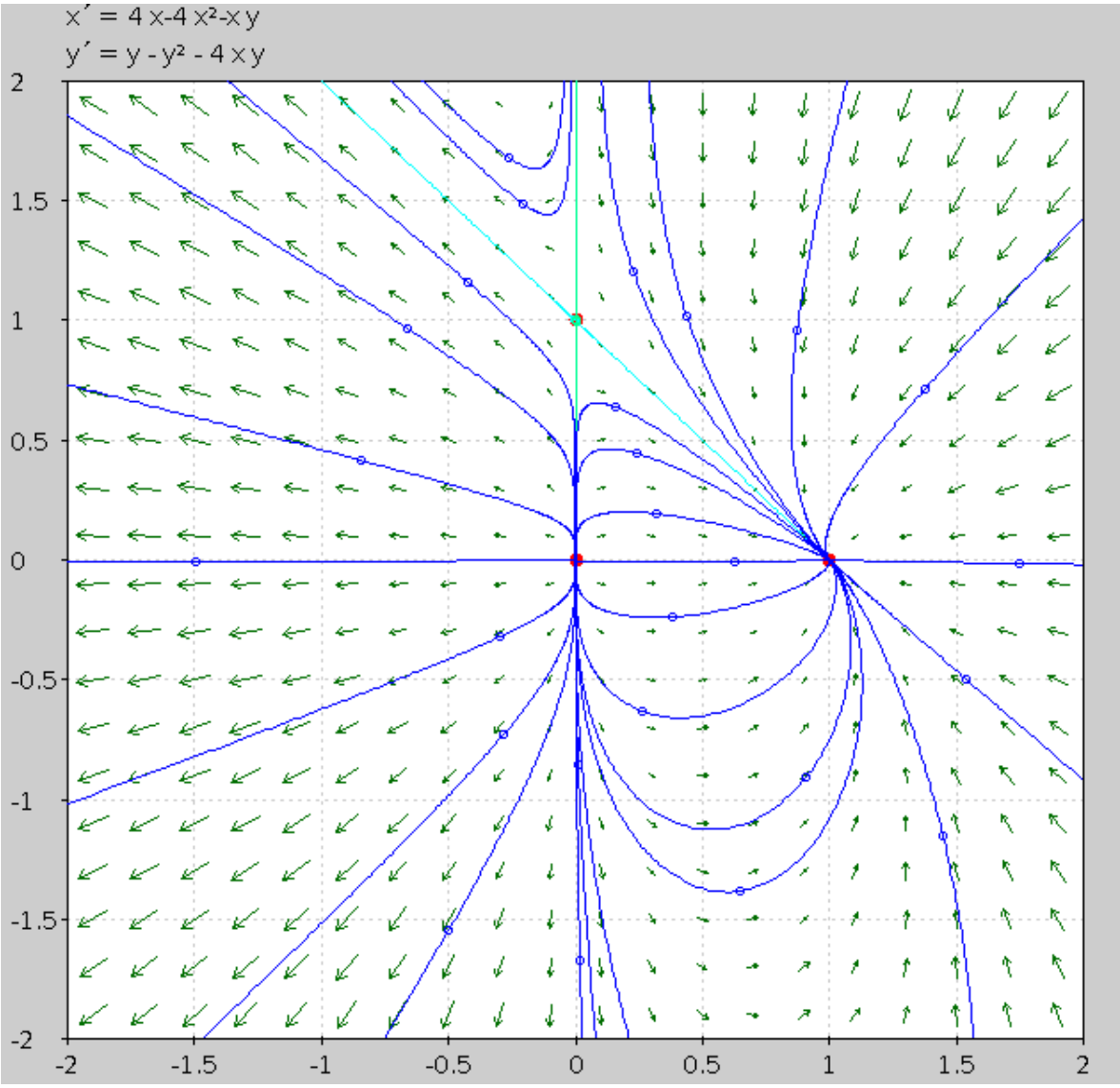


Figure 3.24: The phase portrait of the system in Example 3.4.3.

Note that near the equilibria the phase portrait is similar to the corresponding phase portrait of the linearized system. Some of the straight lines are now curved.

Since the eigenvalues of $A_{(0,0)}$ are both positive, $(0,0)$ is a repeller. The eigenvalues of $A_{(1,0)}$ are both negative, so $(1,0)$ is an attractor. The eigenvalues of $A_{(0,1)}$ have opposite signs, so $(0,1)$ is unstable but not a repeller.

Note that the axes define trajectories and just as before each trajectory is confined in one quadrant. In the first quadrant almost all solutions go to the attractor $(1,0)$ as $t \to \infty$. As $t \to -\infty$, the solutions come from either infinity or from the repeller $(0,0)$. In the second quadrant almost all solutions become unbounded as $t \to \infty$ and either come from infinity or from the repeller $(0,0)$ as $t \to -\infty$.

The solutions in the third quadrant become unbounded as $t \to \infty$ and go to $(0,0)$ as $t \to -\infty$.

In the fourth quadrant solutions that start to the y-axis approach the attractor $(0,0)$ as $t \to \infty$ and become unbounded as $t \to -\infty$.

Example 3.4.4. Discuss the equilibria and stability of the system

$$\begin{aligned} \frac{dx}{dt} &= x - x^2 - xy \\ \frac{dy}{dt} &= -y + 4xy \end{aligned}$$

Solution: The equilibria are

$$(0,0), \quad (1,0), \quad \left(\frac{1}{4}, \frac{3}{4}\right)$$

The jacobian matrix is

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} -2x - y + 1 & -x \\ 4y & 4x - 1 \end{bmatrix}$$

At the equilibria, the linearization matrices and their eigenvalues, basic eigenvectors, and phase portraits are as follows.

$$A_{(0,0)} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \lambda_1 = -1, \ \mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \lambda_2 = 1, \ \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

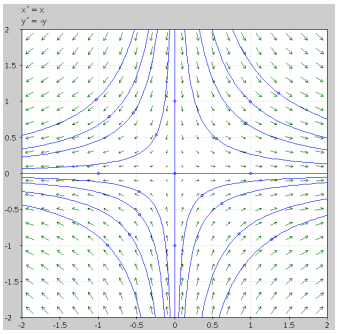


Figure 3.25: The phase portrait of $D\mathbf{y} = A_{(0,0)}\mathbf{y}$.

$$A_{(1,0)} = \begin{bmatrix} -1 & -1 \\ 0 & 3 \end{bmatrix}, \quad \lambda_1 = -1,\ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda_2 = 3,\ \mathbf{v}_2 = \begin{bmatrix} 1 \\ -4 \end{bmatrix}$$

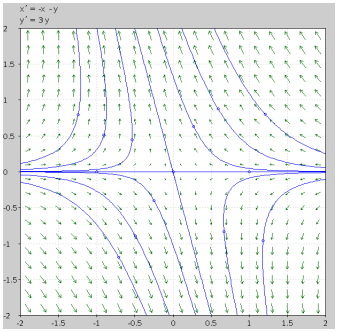


Figure 3.26: The phase portrait of $D\mathbf{y} = A_{(1,0)}\mathbf{y}$.

$$A_{(1/4,3/4)} = \begin{bmatrix} -\frac{1}{4} & -\frac{1}{4} \\ 3 & 0 \end{bmatrix}, \quad \lambda = -\frac{1}{8} \pm \frac{\sqrt{47}}{8}i$$

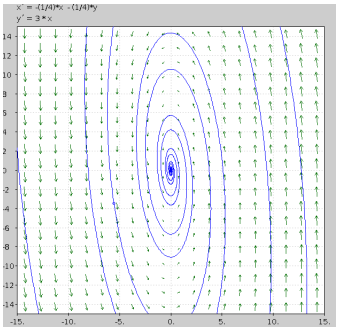


Figure 3.27: The phase portrait of $D\mathbf{y} = A_{(1/4,3/4)}\mathbf{y}$.

None of the eigenvalues was zero or pure imaginary so we may again apply the Hartman-Grobman Theorem.

Again, by using graphical methods we get the phase portrait of the given system as seen in Figure 3.28.

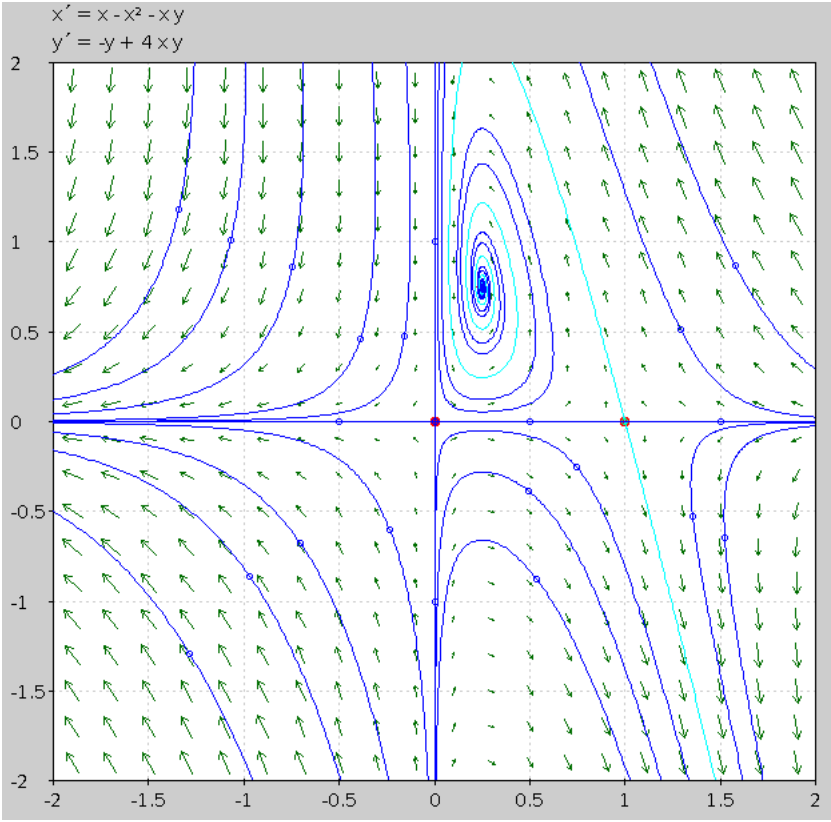


Figure 3.28: The phase portrait of the system in Example 3.4.4.

Each of $A_{(0,0)}$ and $A_{(1,0)}$ have opposite sign eigenvalues, so $(0,0)$ and $(1,0)$ are unstable but not repellers. Since $A_{(1/4,3/4)}$ has complex eigenvalues with negative real part, $(1/4,3/4)$ is an attractor.

Most of the trajectories in the first quadrant spiral around the attractor $(1/4,3/4)$ as $t \to \infty$. Exceptions are: the positive x-axis where the trajectories on this axis approach $(1,0)$ and the positive y-axis where the trajectories on this axis approach $(0,0)$. As $t \to -\infty$ most of the trajectories in the first quadrant become unbounded. The exception are the trajectory going from $(0,0)$ to $(1,0)$ and the trajectory going from $(1,0)$ and spirals towards the attractor $(1/4,3/4)$.

Most of the solutions of the second, third, and fourth quadrants become unbounded as $t \to \pm\infty$. Exception are the trajectories along the axes and the trajectory going from $(1, 0)$ to infinity.

Example 3.4.5. Use linearization and discuss the stability of the system

$$\begin{aligned}\frac{dx}{dt} &= xy + 2y + z^2\\ \frac{dy}{dt} &= x - y\\ \frac{dz}{dt} &= (x-y)^2 - 4z\end{aligned}$$

Solution: The equilibria are the points that satisfy the system

$$\begin{aligned}xy + 2y + z^2 &= 0\\ x - y &= 0\\ (x-y)^2 - 4z &= 0\end{aligned}$$

The second equation implies $x = y$. Hence, the third equation implies that $z = 0$. Then the first equation implies that either $x = y = 0$ or $x = y = -2$. Hence, the equilibria are

$$(0, 0, 0) \quad \text{and} \quad (-2, -2, 0)$$

The jacobian matrix is

$$\begin{bmatrix} f_x & f_y & f_z \\ g_x & g_y & g_z \\ h_x & h_y & h_z \end{bmatrix} = \begin{bmatrix} y & x+2 & 2z \\ 1 & -1 & 0 \\ 2x-2y & -2x+2y & -4 \end{bmatrix}$$

Hence, the linearization matrices and their eigenvalues are:

$$A_{(0,0,0)} = \begin{bmatrix} 0 & 2 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -4 \end{bmatrix}, \quad \lambda_1 = -4,\ \lambda_2 = -2,\ \lambda_3 = 1.$$

and

$$A_{(-2,-2,0)} = \begin{bmatrix} -2 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -4 \end{bmatrix}, \quad \lambda_1 = -4,\ \lambda_2 = -2,\ \lambda_3 = -1.$$

One eigenvalue of $A_{(0,0,0)}$ is positive and the other two are negative. Hence, $(0,0,0)$ is an unstable equilibrium that is not a repeller.

All eigenvalues of $A_{(-2,-2,0)}$ are negative. Therefore, $(-2,-2,0)$ is an attractor.

Existence and Uniqueness of Solutions

In this paragraph we state a fundamental theorem that guarantees the existence and uniqueness of solutions of the autonomous system

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2, \ldots, x_n) \\ \frac{dx_2}{dt} &= f_2(x_1, x_2, \ldots, x_n) \\ &\vdots \\ \frac{dx_n}{dt} &= f_n(x_1, x_2, \ldots, x_n) \end{aligned} \tag{3.19}$$

which we abrreviate in vector notation by

$$D\mathbf{x} = \mathbf{f}(\mathbf{x}) \tag{3.20}$$

with

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}$$

Theorem 3.4.3 (Existence and Uniqueness of Solutions)**.** *Consider the initial value problem*

$$D\mathbf{x} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{3.21}$$

Suppose that $\mathbf{f}$ *is continuous, i.e., all* f_i *are continuous and that all its partial derivatives* $\partial f_i/\partial x_j$ *exist and are continuous for all* $\mathbf{x}$ *in an open connected set* $D \subset \mathbf{R}^n$*. Then for* $\mathbf{x}_0$ *in* D*, the initial value problem (3.21) has a solution* $\mathbf{x}(t)$ *for* t *in some interval* $(-\tau, \tau)$ *and this solution is unique.*

3.5 Constants of Motion; Pendulum Lotka-Volterra Equations

The Undamped Pendulum

We start this section with the familiar equation of a simple pendulum.

A mass m is attached to a thin solid rod of length L and of negligible weight. The other end of the rod is attached to a point that allows it to turn freely. See Fig. (3.29).

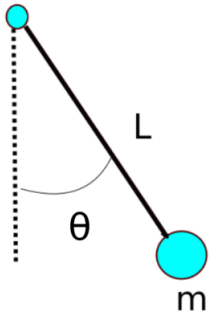


Figure 3.29: Undamped pendulum.

If we assume that the only acting force on the mass is gravity and that we measure the angular displacement of the mass by the angle θ from the vertical line through the support of the rod to the rod, then θ satisfies the **undamped pendulum equation**

$$\frac{d^2\theta}{dt^2} + \frac{g}{L}\sin(\theta) = 0 \tag{3.22}$$

This is a nonlinear equation due to the sine term. One could use the approximation that for small θ we have $\sin(\theta) \approx \theta$, but then we miss the behavior of the large oscillations.

We convert this equation to a system by using the substitution

$$x = \theta \qquad \text{and} \qquad y = \frac{d\theta}{dt}$$

to get

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= -\frac{g}{L}\sin(x) \end{aligned} \tag{3.23}$$

The equilibria are obtained from solving

$$\begin{aligned} y &= 0 \\ -\frac{g}{L}\sin(x) &= 0 \end{aligned}$$

to get the infinitely many

$$\mathbf{c}_n = \begin{bmatrix} n\pi \\ 0 \end{bmatrix}$$

one for each integer n.

Next we compute the jacobian matrix

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L}\cos(x) & 0 \end{bmatrix}$$

and substitute the equilibria to get the linearization matrices.

$$A_n = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L}\cos(n\pi) & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ (-1)^{n+1}\frac{g}{L} & 0 \end{bmatrix}$$

since $\cos(n\pi) = (-1)^n$ for integer n.

If n is odd we get

$$A_n = \begin{bmatrix} 0 & 1 \\ \frac{g}{L} & 0 \end{bmatrix}, \qquad n \text{ odd}$$

The eigenvalues are $\pm\sqrt{g/L}$. Therefore, the Hartman-Grobman Theorem applies and the trajectories near these equilibria are like

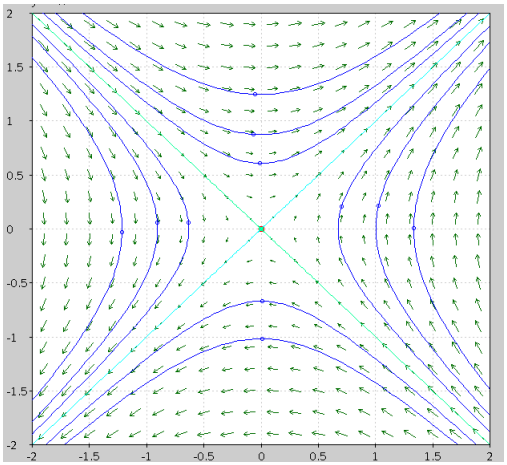


Figure 3.30: Linearization at $\mathbf{c}_n$, n odd and $g = L$.

If n is even we get

$$A_n = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L} & 0 \end{bmatrix}, \qquad n \text{ even}$$

The eigenvalues are $\pm i\sqrt{g/L}$. These are pure imaginary so we cannot use the Hartman-Grobman Theorem.

To study the phase portrait of (3.23) we shall use some physics. We analyze the **energy** of the pendulum. In our case the force on the mass depends only on its position and not on its velocity or on time, so the **potential energy** is defined up to a constant as the negative antiderivative of the force

$$E_p(x) = \int \frac{g}{L} \sin(x)\, dx = -\frac{g}{L} \cos(x)$$

The **kinetic energy** is defined by

$$E_k(y) = \frac{1}{2} y^2$$

where $y = \frac{dx}{dt}$ is the (angular) velocity. Hence, the total energy $E(x, y)$ is

$$E(x, y) = E_p(x) + E_k(y) = -\frac{g}{L} \cos(x) + \frac{1}{2} y^2$$

For each solution $(x(t), y(t))$ of the system (3.23) the energy along the solution is

$$E(t) = E(x(t), y(t))$$

Differentiation with respect to t yields, by Chain Rule

$$\begin{aligned}\frac{dE}{dt} &= \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt} \\ &= \frac{g}{L}\sin(x)\frac{dx}{dt} + y\frac{dy}{dt}\end{aligned}$$

Along a solution we have $dx/dt = y$ and $dy/dt = -(g/L)\sin(x)$. Therefore,

$$\frac{dE}{dt} = \frac{g}{L}\sin(x)y + y\left(-\frac{g}{L}\sin(x)\right) = 0$$

So the energy is *constant along solutions*. Therefore, the solutions must lie along the *level curves* $E(x, y) = c$, where c is a constant. So we may sketch level curves using

$$-\frac{g}{L}\cos(x) + \frac{1}{2}y^2 = c$$

for several c to get an idea how the trajectories look like. (Fig. 3.31.)

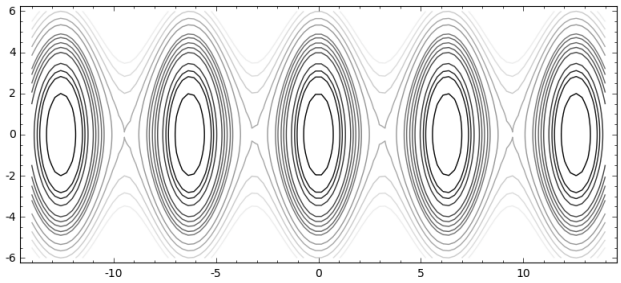


Figure 3.31: Level curves for the pendulum energy function.

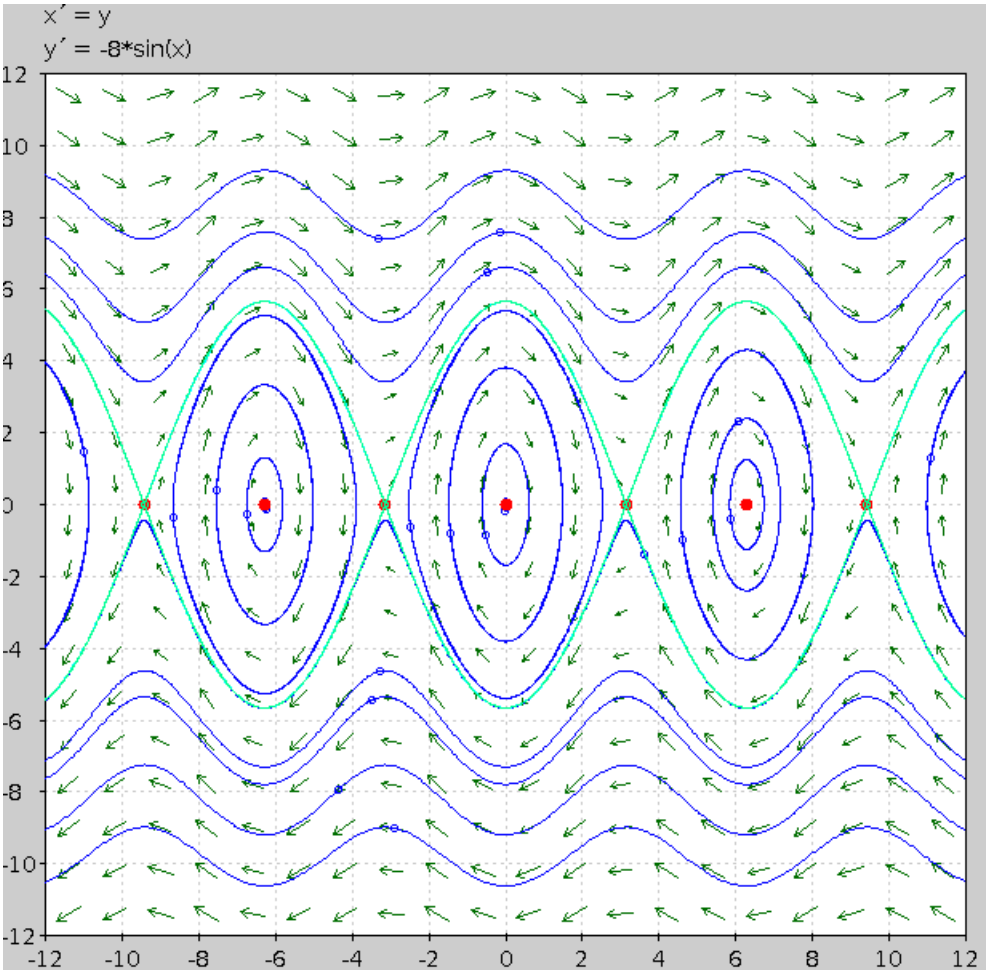


Figure 3.32: A pendulum phase portrait.

There are three kinds of level curves of the energy function that depend on c. If $c > g/L$, there are two branches symmetrically above the below the x-axis. These define distinct trajectories of the phase portrait. If $c = g/L$, the level curves are pairs of functions passing through the equilibria for n odd. For $-g/L < c < g/L$ the level curves are closed orbits around the equilibria for n even. If $c = -g/L$, we get the points $(n\pi, 0)$ for n even. There are no level curves for $c < -g/L$, since $-1 \leq \cos(x) \leq 1$ and $y^2 \geq 0$.

The phase portrait of the system is seen in Fig. 3.32. We observe that the equilibria $(n\pi, 0)$ for n odd have local phase portraits as expected by the Hartman-Grobman Theorem. Surprisingly, the equilibria $(n\pi, 0)$ for n even

also have local phase portraits (closed trajectories around them) just as they would, if the Hartman-Grobman Theorem were applicable. This is not always the case.

The equilibria for n even are stable represent the cases where the pendulum rests in the straight down position. For n odd the equilibria are unstable and represent the cases where the pendulum rests in the straight up position.

We also observe that at high enough energy $c > g/L$, $\theta(= x)$ is continuously increasing, if the angular velocity $y > 0$ and continuously decreasing, if $y < 0$. The solutions in this case become unbounded as $t \to \pm\infty$. The pendulum performs continuously complete rotations either in the counterclockwise direction ($y > 0$) or in the clockwise direction ($y < 0$).

At lower energy levels when $-g/L < c < g/L$, the solutions are periodic. The angle θ oscillates about an equilibrium for n even. The pendulum now swings continuously and symmetrically about the straight down position.

At energy $c = g/L$, if n is odd we have unstable equilibria. Some of the trajectories move towards the equilibrium and some away. Physically, the pendulum swings around the top vertical position. If the angle θ_0 and velocity y_0 are such tha the point (θ_0, y_0) lies in a trajectory moving towards the equilibrium, then the pendulum moves towards the straight up position approaching it as $t \to \infty$ without ever reaching it.

In the example of the pendulum we examined a method of finding the trajectories of a nonlinear system by using the energy function $E(x, y)$ that has the property that it is constant along the solutions of the system. In general, a function $E(x, y)$ that is constant along every solution of the system is called a **constant of motion**.

We have the following theorem about constants of motion.

Theorem 3.5.1. *Let $E(x, y)$ have continuous partial derivatives $\partial E/\partial x$ and $\partial E/\partial y$. Then $E(x, y)$ is a constant of motion of the system*

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned} \tag{3.24}$$

if and only if

$$f(x, y)\frac{\partial E}{\partial x} + g(x, y)\frac{\partial E}{\partial y} = 0$$

for all (x, y) in the domain of E.

Furthermore, if $E(x, y)$ is a constant of motion of the system, then the trajectories of the system lie in the level curves $E(x, y) = c$ of E.

In the example of the pendulum, the equilibria occured at the critical points of the energy $E(x, y)$, i.e, at the points where $\partial E/\partial x = 0$ and $\partial E/\partial y = 0$. In cases when this happens for a constant of motion $E(x, y)$ we may obtain information about the equilibria by examining the kinds of critical points of E.

Let (x_0, y_0) be an equlibrium of system 3.24 that is also a critical point of a constant of motion $E(x, y)$.

Let us assume that (x_0, y_0) is a **strict local minimum**. This means that there is a circle centered at (x_0, y_0) such that $E(x, y) < E(x_0, y_0)$ for all (x, y) inside the circle.

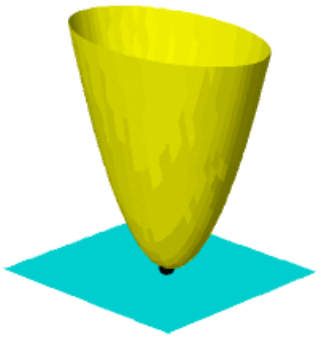


Figure 3.33: A local minimum of a constant of motion.

In Figure 3.33 we show a graph of a function $E(x, y)$ as a surface $z = E(x, y)$ with a strict local minimum. Let $c_0 = E(x_0, y_0)$ be the minimum value. Then if we choose a c slightly larger than c_0, the plane $z = c$ will intersect the surface at a small closed curve just above the equilibrium. (See Figure 3.34.) The level curves $E(x, y) = c$, hence, the trajectories of the system will be closed orbits around the equilibrium and the equilibrium is stable. (See Figure 3.35.)

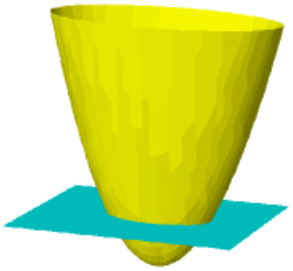


Figure 3.34: The plane $z = c$ intersects the surface at a small closed curve above the equlibrium.

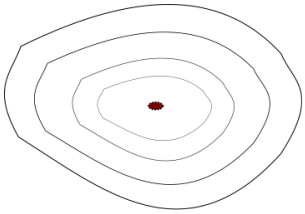


Figure 3.35: Level curves around an equlibrium for some constant of motion.

A similar phase portrait occurs, if the constant of motion has a **strict local maximum** at the equilibrium of system (3.24).

In Calculus we classified the critical points of a function $E(x, y)$ by using the **second partial derivatives test** which makes use of the **discriminant** $\Delta(x, y)$ defined by

$$\Delta(x, y) = \frac{\partial^2 E}{\partial x^2}\,\frac{\partial^2 E}{\partial y^2} - \left(\frac{\partial^2 E}{\partial x\,\partial y}\right)^2$$

More precisely we have the theorem.

Theorem 3.5.2 (Second Partial Derivatives Test)**.** *Let $E(x, y)$ have continuous all first and second partial derivatives. Let (x_0, y_0) be a critical point of $E(x, y)$. So, $\partial E/\partial x\,|_{(x_0,y_0)} = 0 = \partial E/\partial y\,|_{(x_0,y_0)}$.*

1. *If $\Delta(x_0, y_0) > 0$, then E has a strict local extremum at (x_0, y_0). This is a strict local maximum, if $\partial^2 E/\partial x^2\,|_{(x_0,y_0)} < 0$ and a strict local minimum, if $\partial^2 E/\partial x^2\,|_{(x_0,y_0)} > 0$.*

2. *If $\Delta(x_0, y_0) < 0$, then E has neither local maximum nor local minimum at (x_0, y_0). We say, that (x_0, y_0) is a* ***saddle point****.*

3. *If $\Delta(x_0, y_0) = 0$, the test is inconclusive. In this case (x_0, y_0) is called a* ***degenerate critical point****.*

We recall from Calculus that in the case of the saddle point (second case of Theorem 3.5.2) the graph of E resembles a "saddle" near (x_0, y_0). It has a local minimum on a plane parallel to one of the axes and a local maximum on a plane parallel to the other axis. If we slice the surface with planes just below the point $(x_0, y_0, E(x_0, y_0))$ we get a curve with two branches that looks like a hyperbola. If we move the plane slightly above the point, then we get a different curve with two branches that looks like a hyperbola that is at a right angle with the first one.

Figure 3.36: Saddle point.

If we draw level curves around a saddle point we get two curves that cross at the saddle point and divide the plane into four "quadrants". In each of the quadrants we get curves that remain in their quadrant, as seen in Figure 3.37.

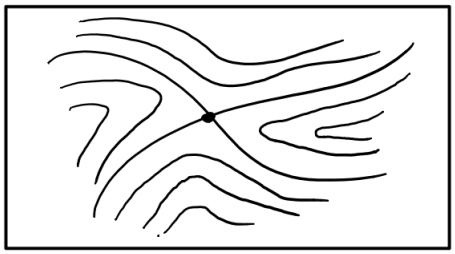


Figure 3.37: Level curves near a saddle point.

In the case when the saddle point is also an equilibrium of system 3.24, then the local phase portrait of the system resembles the one seen in Figure 3.38.

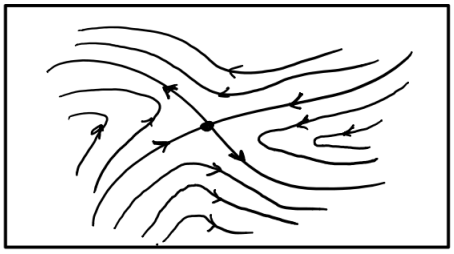


Figure 3.38: Trajectories near a saddle point equilibrium.

Such an equilibrium we also call a **saddle point equilibrium**.

Let us summarize our observations about critical points and equilibria.

Theorem 3.5.3. *Suppose that $E(x, y)$ has continuous first and second partial derivatives and is a constant of motion of the system*

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned} \tag{3.25}$$

Let (3.25) have an equilibrium at a point (x_0, y_0) with the property that there is a neighborhood around this equilibrium that contains no other equilibria. Then we have

1. *If (x_0, y_0) is a critical point of $E(x, y)$ with $\Delta(x_0, y_0) > 0$, then the equilibrium at (x_0, y_0) is stable and has closed trajectories around it. The solutions oscillate periodically around the equilibrium.*

2. *If (x_0, y_0) is a critical point of $E(x, y)$ with $\Delta(x_0, y_0) < 0$, then the equilibrium at (x_0, y_0) is a saddle point and it is unstable. If $c_0 = E(x_0, y_0)$, then any trajectories at the level sets $E(x, y) = c_0$ are separatrices that approach the equilibrium either as $t \to \infty$ or as $t \to -\infty$. All other trajectories near (x_0, y_0) leave any neighborhood of (x_0, y_0) as $t \to \infty$ and also as $t \to -\infty$.*

The Lotka-Volterra Equations

The **Lotka-Volterra equations** or **predator-prey equations** are two nonlinear equations modeling the populations of certain types of competing species labeled one as a "predator" and the other as "prey". The solutions of these equations explain how the species interact and how their populations change over time. These equations form the nonlinear system

$$\begin{aligned} \frac{dx}{dt} &= ax - bxy \\ \frac{dy}{dt} &= -cy + dxy \end{aligned} \tag{3.26}$$

The constants a, b, c, d are assumed to be positive.

Here $x(t)$ is the number of prey and $y(t)$ is the number of predators. We view these functions as continuous in the time variable t and not as integers. Since x and y model populations we are interested only in the cases where $x \geq 0$ and $y \geq 0$.

The equilibria of system 3.26 are

$$(0,0) \quad \text{and} \quad \left(\frac{c}{d}, \frac{a}{b}\right)$$

Next we compute the jacobian matrix

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} a - by & -bx \\ dy & -c + dx \end{bmatrix}$$

and substitute the equilibria to get the linearization matrices.

For $(0,0)$, we have

$$A_{(0,0)} = \begin{bmatrix} a & 0 \\ 0 & -c \end{bmatrix}, \quad \lambda_1 = a, \ \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \lambda_2 = -c, \ \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

The eigenvalues have opposite signs, hence the Hartman-Grobman Theorem applies. The origin is a saddle point. The axes are eigenlines and they define trajectories. For $x > 0$ the trajectories along the x-axis move away from the origin as $t \to \infty$. This means that if there are no predators ($y = 0$), the population of the prey will keep increasing. For $y > 0$ the trajectories along the y-axis move towards the origin as $t \to \infty$. This means that if there is no prey ($x = 0$), the predators will eventually become extinct. So in the extreme

cases of the absence of one of the species the model predicts the expected behavior.

For $(c/d, a/b)$, we have

$$A_{(c/d,a/b)} = \begin{bmatrix} 0 & -\frac{bc}{d} \\ \frac{ad}{b} & 0 \end{bmatrix}, \qquad \lambda = \pm i\sqrt{ac}$$

The eigenvalues are pure imaginary, so the Hartman-Grobman Theorem does not apply at this equilibrium.

We seek a constant of motion. We note that the trajectories have slope

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{-cy + dxy}{ax - bxy}$$

So we have a separable differential equation

$$\frac{dy}{dx} = \frac{(-c + dx)y}{(a - by)x}$$

which can be solved by separation of variables. We get the implicit solution.

$$x^c e^{-dx} y^a e^{-by} = C$$

Hence, a constant of motion is given by the function

$$E(x, y) = x^c e^{-dx} y^a e^{-by}$$

Next we compute the partial derivatives in order to find the discriminant at the equilibrium. We get

$$\frac{\partial E}{\partial x} = \left(\frac{c}{x} - d\right) E(x, y)$$
$$\frac{\partial E}{\partial y} = \left(\frac{a}{y} - b\right) E(x, y)$$

The only critical point of $E(x, y)$ that is not on the axes is the already found equilibrium $(c/d, a/b)$.

Next we have the second partial derivatives.

$$\frac{\partial^2 E}{\partial x^2} = \left(\left(\frac{c}{x} - d \right)^2 - \frac{c}{x^2} \right) E(x,y)$$
$$\frac{\partial^2 E}{\partial y^2} = \left(\left(\frac{a}{y} - b \right)^2 - \frac{a}{y^2} \right) E(x,y)$$
$$\frac{\partial^2 E}{\partial y \partial x} = \left(\frac{c}{x} - d \right) \left(\frac{a}{y} - b \right) E(x,y)$$

Finally, we evaluate the discriminant $\Delta((c/d, a/b))$ to get

$$\Delta((c/d, a/b)) = \frac{b^2 d^2}{a^2 c^2} \left(E(c/d, a/b) \right)^2 > 0$$

Therefore, by Theorem 3.5.3, $(c/d, a/b)$ is a stable equilibrium that has closed trajectories, i.e., periodic orbits, around it.

In fact, what really happens is that *all* trajectories in the first quadrant that are off the axes are closed trajectories around this equilibrium. (See Fig. 3.39.)

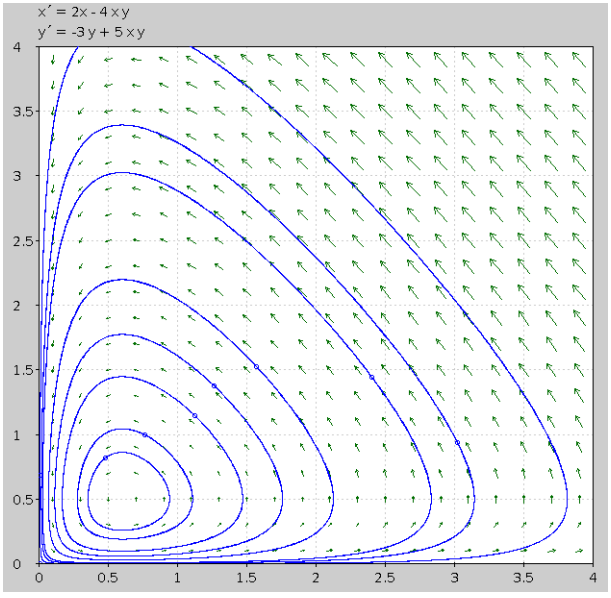


Figure 3.39: Trajectories of the Lotka-Volterra equations.

We see that the populations of predator and prey are periodic around the equilibrium. At the equilibrium of course they remain fixed. What is interesting is that as long as neither population is zero even if the starting point of a population is small it will eventually recover and the numbers will grow before they become small again.

3.6 Limit Cycles

The Van der Pol equation is the following second order differential equation

$$\frac{d^2x}{dt^2} + \epsilon\left(x^2 - 1\right)\frac{dx}{dt} + x = 0 \tag{3.27}$$

which was introduced by Dutch engineer Van der Pol to model the current in a triode vacuum tube.

We are interested in the parameter value $\epsilon = 1$. We convert the equation to a system to get

$$\begin{aligned} \frac{dx}{dt} &= y \\ \frac{dy}{dt} &= \left(1 - x^2\right)y - x \end{aligned} \tag{3.28}$$

This is a second order nonlinear system.

If we set $y = 0$ and $(1 - x^2)\, y - x = 0$, we see that $(0, 0)$ is the only equilibrium.

Next, we compute the jacobian matrix

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 - 2xy & 1 - x^2 \end{bmatrix}$$

and substitute the equilibrium to get the linearization matrix and eigenvalues.

$$A_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \qquad \lambda = \frac{1}{2} \pm i\frac{\sqrt{3}}{2}$$

The eigenvalues are complex with positive real part. Hence, the Hartman-Grobman Theorem applies and the equilibrium $(0, 0)$ is a repeller. The trajectories locally spiral away from the origin. The phase portrait of the linearized system is seen in Figure 3.40.

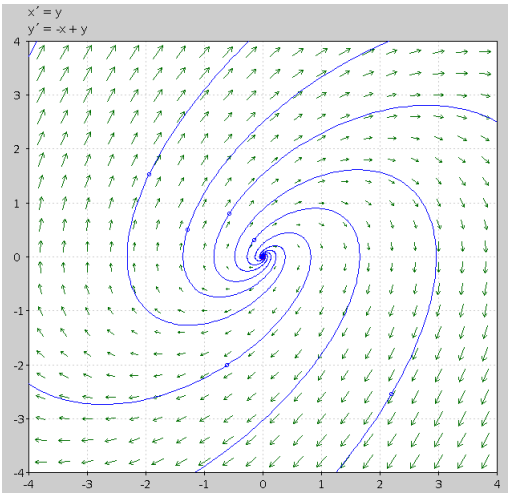


Figure 3.40: The Van der Pol equation with $\epsilon = 1$, linearization at repeller equilibrium $(0, 0)$.

Based on the local behavior of the trajectories near the equilibrium, we may be tempted to deduce that the trajectories as they spiral out away from the origin, they become unbounded. However, experiments show that the current of an unforced triode always eventually settles to periodic values.

If this is the case and the theoretical model is accurate, then clearly the *local behavior does not determine the global behavior of the trajectories.*

First we observe that if $\epsilon = 0$ in the original equation instead of $\epsilon = 1$, the term $(1 - x^2)y$ is absent and the system is linear with trajectories that are circles around the equilibrium $(0, 0)$. What changes this linear behavior is the nonlinear term $(1 - x^2)y$. We think of the term as one that changes the vertical component of the rotational motion in a way that is proportional to y and $1 - x^2$. For values $|x| < 1$ the factor $1 - x^2$ is positive, but for $|x| > 1$ it is negative and can have arbitrarily large magnitude.

Let us examine the direction of the motion around a large circle, say $x^2 + y^2 = 400$. In the small band $-1 \leq x \leq 1$, the motion is outward since the y-component is increasing, having positive derivative. The motion is upward, if $y > 0$ and downward, if $y < 0$.

Outside the band $-1 \leq x \leq 1$ the motion is inward since the y-component is decreasing, having negative derivative.

So, except for a few small values of x the motion is strongly inward. So as trajectory goes around it moves inward.

This analysis is not rigorous but it seems that trajectories far out wind

around inward, whereas we know that near the origin trajectories spiral out away from the origin. We may then conclude that there must be a closed trajectory that trajectories that start outside wind around towards it and trajectories that start inside wind out toward it. Such trajectory is called a **limit cycle**.

Figure 3.41 shows a phase portrait of the system showing the limiting cycle.

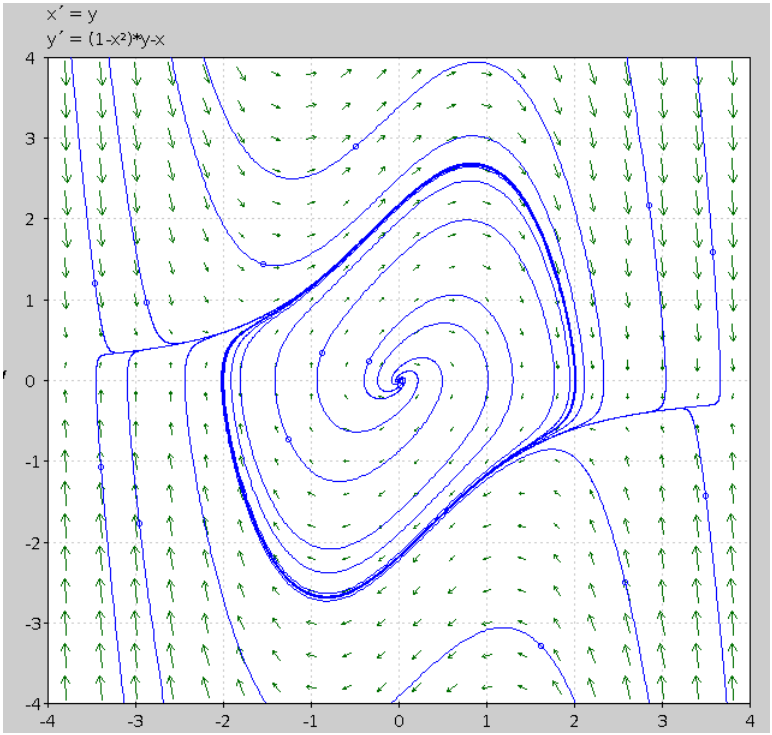


Figure 3.41: The Van der Pol equation with $\epsilon = 1$.

Our analysis of the Van der Pol equation was rather intuitive. In the remaining of this section we discuss some important theorems that help us determine where or not limiting cycles exist.

Theorem 3.6.1. *If the system*

$$\begin{aligned} \frac{dx}{dt} &= f(x,y) \\ \frac{dy}{dt} &= g(x,y) \end{aligned} \tag{3.29}$$

has a closed trajectory, then this trajectory has an equilibrium in its interior.

If there is only one equilibrium in the interior of the trajectory, then the linearization matrix at this unique equilibrium cannot have real eigenvalues of opposite sign.

Example 3.6.1. Show that the system

$$\begin{aligned}\frac{dx}{dt} &= x - x^2 - 3xy \\ \frac{dy}{dt} &= y - y^2 - 2xy\end{aligned}$$

has no closed trajectories.

Solution: In Example 3.4.2 we saw that equilibria are

$$(0,0), \quad (1,0), \quad (0,1), \quad \left(\frac{2}{5}, \frac{1}{5}\right)$$

and that the coordinate axis define trajectories. Since trajectories do not intersect any trajectory that has points in one quadrant, it is entirely contained in the quadrant. If there is a closed trajectory C, then it must have an equilibrium in its interior, by Theorem 3.6.1. Since the only equilibium that does not lie on an axis is $\left(\frac{2}{5}, \frac{1}{5}\right)$, then this equilibrium must be enclosed by C. However, in Example 3.4.2 we showed that this equilibrium has real eigenvalues of opposite signs. Hence, C cannot enclose this equilibrium, again by Theorem 3.6.1. We conclude that the phase portrait has no closed trajectories.

□

For our next theorem we need to know that a region D in the plane is called a **simply connected domain**, if the interior of every simple closed curve in D is entirely in D. In other words, D has "no holes".

Theorem 3.6.2. *Let $f(x,y)$ and $g(x,y)$ be functions with continuous first partial derivatives. Suppose that the function*

$$\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y}$$

is always positive or always negative at all points of a simply connected domain D of the plane. Then the system

$$\begin{aligned} \frac{dx}{dt} &= f(x,y) \\ \frac{dy}{dt} &= g(x,y) \end{aligned} \tag{3.30}$$

has no closed trajectory in D.

Example 3.6.2. Show that the system

$$\begin{aligned} \frac{dx}{dt} &= x - 2x^2 - xy \\ \frac{dy}{dt} &= -y + 4xy \end{aligned}$$

has no closed trajectories.

Solution: The equilibria are

$$(0,0), \quad \left(\frac{1}{2}, 0\right), \quad \left(\frac{1}{4}, \frac{1}{2}\right)$$

and the axes determine trajectories. Therefore, each trajectory is confined in one and only quadrant. Just as in Example 3.6.1 a closed trajectory C must be in the interior of one of the quadrants. Furthermore, C must enclose an equilibrium by Theorem 3.6.1. So C must enclose the equilibrium $\left(\frac{1}{4}, \frac{1}{2}\right)$ which is the only available that is not on the axes. Now in the interior of the first quadrant we have

$$\frac{\partial}{\partial x}(x - 2x^2 - xy) + \frac{\partial}{\partial y}(-y + 4xy) = -y < 0$$

Hence, by Theorem 3.6.2 this system cannot not have a closed trajectory.

The theorems so far gave us some conditions of non existence of closed trajectories. The following important theorem give us conditions under which a closed trajectory exists.

Theorem 3.6.3 (Poincaré-Bendixson Theorem)**.** *Let $f(x,y)$ and $g(x,y)$ be functions with continuous first partial derivatives. Let B be a bounded region*

of the plane that contains its boundary and that does not contain any equilibria of the system

$$\begin{aligned} \frac{dx}{dt} &= f(x,y) \\ \frac{dy}{dt} &= g(x,y) \end{aligned} \tag{3.31}$$

If (3.31) has a solution that is in B for all $t \geq t_0$, then B contains a closed trajectory of (3.31).

Example 3.6.3. Show that the system

$$\begin{aligned} \frac{dx}{dt} &= -y + x - x^3 - 2xy^2 \\ \frac{dy}{dt} &= x + y - x^2y - 2y^3 \end{aligned}$$

has a closed trajectory.

Solution: It can be shown that the only equilibrium of the system is at $(0,0)$.

The distance r from the origin as a point moves along a trajectory satisfies the equation

$$r^2 = x^2 + y^2$$

If we differentiate implicitly with respect to t, we get

$$2r\frac{dr}{dt} = 2x\frac{dx}{dt} + 2y\frac{dy}{dt}$$

Hence,

$$\begin{aligned} \frac{dr}{dt} &= \frac{1}{r}\left(x(-y + x - x^3 - 2xy^2) + y(x + y - x^2y - 2y^3)\right) \\ &= -\frac{1}{r}\left(x^2 + 2\,y^2 - 1\right)\left(x^2 + y^2\right) \\ &= -r(r^2 + y^2 - 1) \end{aligned}$$

Now on the circle centered at $(0,0)$ of radius 1 we have

$$\frac{dr}{dt} = -y^2 < 0$$

since $r = 1$. Hence, a solution that reaches this circle must move inward and cannot cross the circle again, because the distance from the origin keeps decreasing.

On the circle centered at $(0,0)$ of radius $\frac{1}{2}$ we have

$$\frac{dr}{dt} = -\frac{1}{2}\left(y^2 - \frac{3}{4}\right) > 0$$

since $r = \frac{1}{2}$, hence $y^2 < \frac{1}{4}$. Hence, a solution that reaches this circle must move outward and cannot cross the circle again, because the distance from the origin keeps inreasing.

We conclude that a trajectory that starts inside the annulus between the two circles will remain in this region. Hence, by the Poincaré-Bendixson Theorem there is a closed trajectory that lies in this region.

Figure 3.42 is a phase portrait of the system. We see that that there is a closed trajectory with the property that solutions that start outside remain outside and solutions that start inside remain inside. The graph also shows that eventually all other solutions approach the closed trajectory by winding around it. This trajectory "attracts" all other solutions.

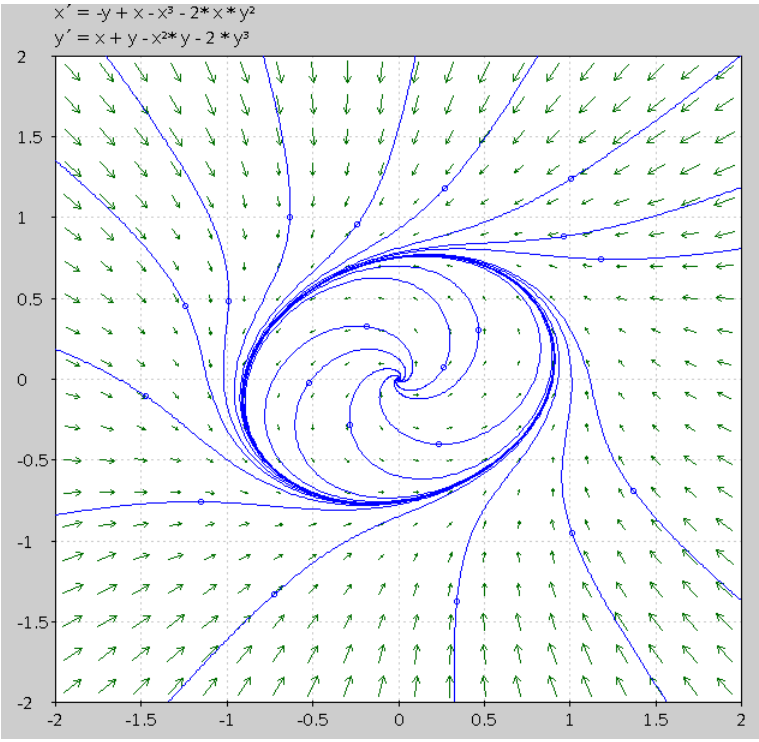


Figure 3.42: The phase portrait of the system of Example 3.6.3 has a closed trajectory.

Chapter 4

Partial Differential Equations

4.1 Some Trigonometric Identities

The following trigonometric identities will be useful in several of the remaining sections in connection to the integration of some trigonometric functions.

1. $\sin(a+b) = \sin a \cos b + \cos a \sin b$
2. $\cos(a+b) = \cos a \cos b - \sin a \sin b$
3. $\sin(a-b) = \sin a \cos b - \cos a \sin b$
4. $\cos(a-b) = \cos a \cos b + \sin a \sin b$
5. $\sin(2a) = 2 \sin a \cos a$
6. $\cos(2a) = 2\cos^2 a - 1$
7. $\cos^2(a) = \dfrac{1+\cos 2a}{2}$
8. $\sin^2(a) = \dfrac{1-\cos 2a}{2}$
9. $\sin a \cos b = \dfrac{1}{2}\sin(a+b) + \dfrac{1}{2}\sin(a-b)$
10. $\sin a \sin b = \dfrac{1}{2}\cos(a-b) - \dfrac{1}{2}\cos(a+b)$

11. $\cos a \cos b = \frac{1}{2}\cos(a-b) + \frac{1}{2}\cos(a+b)$

12. $\cos(k\pi) = (-1)^k$, k integer.

13. $\sin(k\pi) = 0$, k integer.

14. $\cos\left((2k-1)\frac{\pi}{2}\right) = 0$, k integer.

4.2 Orthogonal Sets of Functions

Orthonormal Sets of Functions

We consider continuous real-valued functions defined on an interval $[a,b]$. So they are in the set $C[a,b]$.

Definitions

1. We say that the distinct functions $g_m(x)$ and $g_n(x)$ are **orthogonal** on $[a,b]$, if their integral inner product is zero. I.e., if

$$\langle g_m, g_n\rangle = \int_a^b g_m(x)\, g_n(x)\, dx = 0, \qquad \text{for } m \neq n$$

2. We say that the sequence of distinct functions $g_1(x), g_2(x), \ldots, g_n(x), \ldots$ is an **orthogonal set** on $[a,b]$, if all functions are pairwise orthogonal. I.e., if

$$\langle g_m, g_n\rangle = 0, \qquad \text{for all } \ m \neq n$$

Recall that the **norm** or **length** of each g_m on $[a,b]$ under this inner product is

$$\|g_m\| = \sqrt{\langle g_m, g_m\rangle} = \sqrt{\int_a^b g_m(x)\, g_m(x)\, dx} = \sqrt{\int_a^b g_m^2(x)\, dx}$$

Orthonormal Sets of Functions

Definition We say that the sequence of distinct functions $g_1(x), g_2(x), \ldots, g_n(x), \ldots$ is an **orthonormal set** on $[a, b]$, if

1. The set is orthogonal: $\langle g_m, g_n \rangle = 0$ for $m \neq n$, and
2. All functions are unit: $\|g_m\| = 1$.

Because $\|g_m\| = 1$ is equivalent to $\|g_m\|^2 = 1$, the above definition is equivalent to saying

$$\langle g_m, g_n \rangle = \int_a^b g_m(x)\, g_n(x)\, dx = \begin{cases} 0, & \text{for all } m \neq n \\ 1, & \text{for all } m = n \end{cases}$$

Note: From an orthogonal set we may obtain an orthonormal set by dividing each function by its own norm. So we replace g_m with $\frac{1}{\|g_m\|} g_m$.

This is because if $\|g_m\| \neq 1, 0$, then

$$\left\| \frac{1}{\|g_m\|} g_m \right\| = \left| \frac{1}{\|g_m\|} \right| \|g_m\| = \frac{1}{\|g_m\|} \|g_m\| = 1$$

Assumptions

Standing assumptions: From now on we assume that

a. all functions we discuss are bounded on $[a, b]$,

b. their integrals over $[a, b]$ are finite, and

c. their norms are nonzero.

Examples of Orthogonal Sets

Example 4.2.1. Consider the set of functions

$$\{1, \cos(\pi x), \sin(\pi x)\}$$

Is this set orthogonal on

1. $[-1, 1]$?

2. $[0, 1]$?

Solution: We check all pairs for orthogonality.

1. On $[-1, 1]$

$$\langle 1, \cos(\pi x)\rangle = \int_{-1}^{1} (1)\cos(\pi x)\, dx = \left.\frac{\sin(\pi x)}{\pi}\right|_{-1}^{1} = 0$$

$$\langle 1, \sin(\pi x)\rangle = \int_{-1}^{1} (1)\sin(\pi x)\, dx = \left.-\frac{\cos(\pi x)}{\pi}\right|_{-1}^{1} = 0$$

$$\begin{aligned}\langle \cos(\pi x), \sin(\pi x)\rangle &= \int_{-1}^{1} \cos(\pi x)\sin(\pi x)\, dx \\ &= \frac{1}{2}\int_{-1}^{1} \sin(2\pi x)\, dx \\ &= \left.-\frac{\cos(2\pi x)}{4\pi}\right|_{-1}^{1} = 0\end{aligned}$$

All pairs are orthogonal. So, the set is orthogonal on $[-1, 1]$.

2. On $[0, 1]$

$$\langle 1, \cos(\pi x)\rangle = \int_{0}^{1} (1)\cos(\pi x)\, dx = \left.\frac{\sin(\pi x)}{\pi}\right|_{0}^{1} = 0$$

$$\langle 1, \sin(\pi x)\rangle = \int_{0}^{1} (1)\sin(\pi x)\, dx = \left.-\frac{\cos(\pi x)}{\pi}\right|_{0}^{1} = \frac{2}{\pi} \neq 0$$

We need not go further. The set is not orthogonal on $[0, 1]$.

Example 4.2.2. Consider the set of functions

$$\sin(x),\ \sin(2x),\ \ldots,\ \sin(nx), \ldots$$

1. Show that this set is orthogonal on $[-\pi, \pi]$.

2. Find each norm on $[-\pi, \pi]$ and find the corresponding orthonormal set.

3. Show that this set orthogonal on $[0, \pi]$.

4. Find each norm on $[0, \pi]$ and find the corresponding orthonormal set.

5. Is the set orthogonal on $[0, 1]$? Why or why not?

Solution: Let $g_n(x) = \sin(nx)$, $n = 1, 2, \ldots$

1. For the interval $[-\pi, \pi]$, if $m \neq n$, then

$$\begin{aligned}
\langle g_m, g_n \rangle &= \int_{-\pi}^{\pi} \sin(mx) \sin(nx)\, dx \\
&= \frac{1}{2} \int_{-\pi}^{\pi} [\cos((m-n)x) - \cos((m+n)x)]\, dx \\
&= \left. \frac{\sin((m-n)x)}{2(m-n)} \right|_{-\pi}^{\pi} - \left. \frac{\sin((m+n)x)}{2(m+n)} \right|_{-\pi}^{\pi} \\
&= 0 - 0 \\
&= 0
\end{aligned}$$

2. For the interval $[-\pi, \pi]$, each norm is computed from

$$\begin{aligned}
\|g_m\|^2 &= \int_{-\pi}^{\pi} \sin^2(mx)\, dx \\
&= \frac{1}{2} \int_{-\pi}^{\pi} (1 - \cos(2mx))\, dx \\
&= \frac{1}{2} \left(x - \frac{\sin(2mx)}{2m} \right) \Bigg|_{-\pi}^{\pi} \\
&= \frac{1}{2}(2\pi) \\
&= \pi
\end{aligned}$$

So

$$\|\sin(mx)\| = \sqrt{\pi}, \quad \text{for} \quad m = 1, 2, \ldots$$

So, the corresponding orthonormal set is

$$\frac{\sin(x)}{\sqrt{\pi}}, \ \frac{\sin(2x)}{\sqrt{\pi}}, \ldots, \ \frac{\sin(mx)}{\sqrt{\pi}}, \ldots$$

3. For the interval $[0, \pi]$, if $m \neq n$, then $\langle g_m, g_n \rangle = \int_0^\pi \sin(mx) \sin(nx)\, dx$ is computed as in part 1, except that the lower limit of the integral is 0 and $\sin(0) = 0$ which proves the claim.

4. For the interval $[0, \pi]$, each norm is computed from $\|g_m\|^2 = \int_0^\pi \sin^2(mx)\, dx$ just as in part 2 but now the lower limit of the integral is 0 and we get $\pi/2$.

$$\|g_m\|^2 = \int_0^\pi \sin^2(mx)\, dx = \frac{\pi}{2}$$

So

$$\|\sin(mx)\| = \sqrt{\frac{\pi}{2}}, \quad \text{for} \quad m = 1, 2, \ldots$$

So, the corresponding orthonormal set is

$$\frac{\sin(x)}{\sqrt{\pi/2}}, \ \frac{\sin(2x)}{\sqrt{\pi/2}}, \ldots, \ \frac{\sin(mx)}{\sqrt{\pi/2}}, \ldots$$

5. For the interval $[0, 1]$, if $m \neq n$, then just as before we get

$$\langle g_m, g_n \rangle = \int_0^1 \sin(mx) \sin(nx)\, dx = \frac{\sin(m-n)}{2(m-n)} - \frac{\sin(m+n)}{2(m+n)}$$

The last expression is not zero. For example, if $m = 2$ and $n = 1$ we get

$$\frac{1}{2}\sin(1) - \frac{1}{6}\sin(3) \simeq 0.397\,22 \neq 0$$

So, the set is not orthogonal on $[0, 1]$.

Parts 3, and 4 of Example 4.2.2 can be generalized to the following useful example. The integrations are similar to those of Example 4.2.2 (also obtained easily by the change of variables $u = \pi x/L$).

Example 4.2.3. For $L > 0$, consider the set of functions

$$\sin\left(\frac{\pi x}{L}\right), \ \sin\left(\frac{2\pi x}{L}\right), \ \ldots, \ \sin\left(\frac{n\pi x}{L}\right), \ldots$$

1. This set is orthogonal on $[0, L]$.

2. Each norm on $[0, L]$ is equal to $\sqrt{L/2}$.

$$\left\|\sin\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{\frac{L}{2}}, \qquad n = 1, 2, 3, \ldots$$

Example 4.2.4. For $L > 0$, consider the set of functions

$$1,\ \cos\left(\frac{\pi x}{L}\right),\ \cos\left(\frac{2\pi x}{L}\right),\ \ldots,\ \cos\left(\frac{n\pi x}{L}\right), \ldots$$

1. Show that this set is orthogonal on $[0, L]$.
2. Show that the norms on $[0, L]$ are

$$\|1\| = \sqrt{L}$$

and

$$\left\|\cos\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{\frac{L}{2}}, \qquad n = 1, 2, 3, \ldots$$

Solution:

1. For orthogonality, we have

$$\begin{aligned}
\left\langle 1, \cos\left(\frac{n\pi x}{L}\right)\right\rangle &= \int_0^L (1)\cos\left(\frac{n\pi x}{L}\right) dx \\
&= \left.\frac{\sin\left(n\pi x/L\right)}{n\pi/L}\right|_0^L \\
&= \frac{\sin\left(n\pi\right)}{n\pi/L} - \frac{\sin\left(0\right)}{n\pi/L} \\
&= 0 - 0 \\
&= 0
\end{aligned}$$

and for $m \neq n$

$$\begin{aligned}
\left\langle \cos\left(\frac{m\pi x}{L}\right), \cos\left(\frac{n\pi x}{L}\right)\right\rangle &= \int_0^L \cos\left(\frac{m\pi x}{L}\right)\cos\left(\frac{n\pi x}{L}\right) dx \\
&= \frac{1}{2}\int_0^L \left(\cos\left(\left(m-n\right)\pi x/L\right) + \cos\left(\left(m+n\right)\pi x/L\right)\right) dx \\
&= \frac{1}{2}\left.\frac{\sin\left(\left(m-n\right)\pi x/L\right)}{\left(\left(m-n\right)\pi/L\right)} + \frac{\sin\left(\left(m+n\right)\pi x/L\right)}{\left(\left(m+n\right)\pi/L\right)}\right|_0^L \\
&= 0 - 0 \\
&= 0
\end{aligned}$$

2. For the norms we have

$$\|1\|^2 = \int_0^L 1dx = x|_0^L = L$$

and

$$\begin{aligned}\left\|\cos\left(\frac{n\pi x}{L}\right)\right\|^2 &= \int_0^L \cos^2\left(\frac{n\pi x}{L}\right) dx \\ &= \frac{1}{2}\int_0^L \left(1+\cos\left(2n\pi/L\right)\right) dx \\ &= \frac{1}{2}\left(x + \frac{\sin\left(2n\pi/L\right)}{2n\pi/L}\right)\Bigg|_0^L \\ &= \frac{L}{2}\end{aligned}$$

and we take square roots.

Example 4.2.5. Consider the set of functions

$$1,\ \cos(x),\ \sin(x),\ \cos(2x),\ \sin(2x),\ldots,\ \cos(nx),\ \sin(nx),\ldots$$

1. Show that this set forms an orthogonal set on $[-\pi,\pi]$ and also on $[0,2\pi]$.
2. Find each norm and the corresponding orthonormal set.

Solution:

1. We have

   a.
   $$\langle 1, \cos(nx)\rangle = \int_{-\pi}^{\pi} (1)\cos(nx)\,dx = \frac{\sin(nx)}{n}\Bigg|_{-\pi}^{\pi} = 0$$
   This computation also works for $[0,2\pi]$.

   b.
   $$\langle 1, \sin(nx)\rangle = \int_{-\pi}^{\pi} (1)\sin(nx)\,dx = -\frac{\cos(nx)}{n}\Bigg|_{-\pi}^{\pi} = 0$$
   This computation also works for the interval $[0,2\pi]$.

c. If $m \neq n$, then $\langle \sin(mx), \sin(nx) \rangle = 0$. This was proved before. Similarly, this also works for works for the interval $[0, 2\pi]$.

d. If $m \neq n$, then

$$\begin{aligned}
\langle \cos(mx), \sin(nx) \rangle &= \int_{-\pi}^{\pi} \cos(mx) \sin(nx)\, dx \\
&= \frac{1}{2} \int_{-\pi}^{\pi} (\sin((m+n)x) - \sin((m-n)x))\, dx \\
&= \frac{-\cos((m+n)x)}{2(m+n)}\Big|_{-\pi}^{\pi} + \frac{\cos((m-n)x)}{2(m-n)}\Big|_{-\pi}^{\pi} \\
&= 0 + 0 = 0
\end{aligned}$$

This computation also works for the interval $[0, 2\pi]$.

e.

$$\begin{aligned}
\langle \cos(mx), \sin(mx) \rangle &= \int_{-\pi}^{\pi} \cos(mx) \sin(mx)\, dx \\
&= \frac{1}{2} \int_{-\pi}^{\pi} \sin(2mx)\, dx \\
&= -\frac{1}{2} \frac{\cos(2mx)}{2m}\Big|_{-\pi}^{\pi} = 0
\end{aligned}$$

This computation also works for the interval $[0, 2\pi]$.

f. If $m \neq n$, then

$$\begin{aligned}
\langle \cos(mx), \cos(nx) \rangle &= \int_{-\pi}^{\pi} \cos(mx) \cos(nx)\, dx \\
&= \frac{1}{2} \int_{-\pi}^{\pi} (\cos((m-n)x) + \cos((m+n)x))\, dx \\
&= \frac{\sin((m-n)x)}{2(m-n)}\Big|_{-\pi}^{\pi} + \frac{\sin((m+n)x)}{2(m+n)}\Big|_{-\pi}^{\pi} \\
&= 0 + 0 \\
&= 0
\end{aligned}$$

This computation also works for the interval $[0, 2\pi]$.

2. The norms are computed from

   a.
   $$\|1\|^2_{[-\pi,\pi]} = \int_{-\pi}^{\pi} 1dx = x|_{-\pi}^{\pi} = 2\pi$$
   and
   $$\|1\|^2_{[0,2\pi]} = \int_{0}^{2\pi} 1dx = x|_{0}^{2\pi} = 2\pi$$

   b.
   $$\begin{aligned}\|\cos(nx)\|^2_{[-\pi,\pi]} &= \int_{-\pi}^{\pi} \cos^2(nx)\,dx \\ &= \frac{1}{2}\int_{-\pi}^{\pi} (1+\cos(2nx))\,dx \\ &= \frac{1}{2}\left(x + \frac{\sin(2nx)}{2n}\right)\Bigg|_{-\pi}^{\pi} \\ &= \pi\end{aligned}$$

   We also get the same answer for $\|\cos(nx)\|^2_{[0,2\pi]} = \pi$.

   c. $\|\sin(mx)\|^2 = \pi$ was proved in Example 4.2.2 part 2. We also get the same answer for $\|\sin(nx)\|^2_{[0,2\pi]} = \pi$. So, the norms are for both $[-\pi,\pi]$ and $[0,2\pi]$

   $$\|1\| = \sqrt{2\pi}, \quad \|\cos(nx)\| = \sqrt{\pi}, \quad \|\sin(nx)\| = \sqrt{\pi}, \quad \text{for } n = 1,2,\ldots$$

   So, the orthonormal set is for both $[-\pi,\pi]$ and $[0,2\pi]$

   $$\frac{1}{\sqrt{2\pi}}, \frac{\cos(x)}{\sqrt{\pi}}, \frac{\sin(x)}{\sqrt{\pi}}, \frac{\cos(2x)}{\sqrt{\pi}}, \frac{\sin(2x)}{\sqrt{\pi}}, \ldots, \frac{\cos(nx)}{\sqrt{\pi}}, \frac{\sin(nx)}{\sqrt{\pi}}, \ldots$$

Example 4.2.5 can be easily generalized to the following example.

Example 4.2.6. The set of functions

$$1, \cos\left(\frac{\pi x}{L}\right), \sin\left(\frac{\pi x}{L}\right), \cos\left(\frac{2\pi x}{L}\right), \sin\left(\frac{2\pi x}{L}\right), \ldots, \cos\left(\frac{n\pi x}{L}\right), \sin\left(\frac{n\pi x}{L}\right), \ldots$$

is orthogonal on $[-L,L]$ (for $L>0$) and also on $[0,2L]$. The norms for both intervals are the same and they are

$$\|1\| = \sqrt{2L}, \quad \left\|\cos\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{L}, \quad \left\|\sin\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{L}, \quad \text{for } n = 1,2,\ldots$$

To see why Example 4.2.6 is valid the reader may either perform similar integrations to those of Example 4.2.5 or perform the simple integration change of variables $u = \pi x/L$ and reuse the results of Example 4.2.5.

4.3 Generalized Fourier Series

Orthogonal sets are very important because if $f(x)$ is a given function defined on $[a,b]$ and $g_1(x), g_2(x), \ldots, g_n(x), \ldots$ orthogonal on $[a,b]$, then in general $f(x)$ can be represented as a convergent series of the $g_n(x)$.

$$f(x) = \sum_{n=1}^{\infty} a_n g_n(x) = a_1 g_1(x) + \cdots + a_n g_n(x) + \cdots \tag{4.1}$$

where the a_n are constants that depend on the function $f(x)$.

Equation (4.1) is called the **generalized Fourier series** of $f(x)$ with respect to the orthogonal set $g_n(x)$ $n = 1, 2, \ldots$. The coefficients a_n are the **generalized Fourier coefficients** of $f(x)$.

Under some general conditions the series in (4.1) converges to the function $f(x)$.

Under the convergence conditions the generalized Fourier coefficients can be computed as follows.

$$\langle f, g_n \rangle = \left\langle \sum_{m=1}^{\infty} a_m g_m, g_n \right\rangle = \sum_{m=1}^{\infty} a_m \langle g_m, g_n \rangle = a_n \langle g_n, g_n \rangle$$

because $\langle g_m, g_n \rangle = 0$ for $m \neq n$, by orthogonality. Therefore,

$$a_n = \frac{\langle f, g_n \rangle}{\langle g_n, g_n \rangle} = \frac{\langle f, g_n \rangle}{\|g_n\|^2} \tag{4.2}$$

Or, by the definition of the integral inner product, we have

$$a_n = \frac{1}{\|g_n\|^2} \int_a^b f(x) g_n(x)\, dx = \frac{\int_a^b f(x) g_n(x)\, dx}{\int_a^b g_n^2(x)\, dx} \tag{4.3}$$

Example: The (Classical) Fourier Series

In Example 4.2.6 we saw that the set of functions

$$1, \cos\left(\frac{\pi x}{L}\right), \sin\left(\frac{\pi x}{L}\right), \cos\left(\frac{2\pi x}{L}\right), \sin\left(\frac{2\pi x}{L}\right), \ldots, \cos\left(\frac{n\pi x}{L}\right), \sin\left(\frac{n\pi x}{L}\right), \ldots$$

is orthogonal on $[-L, L]$. If a function $f(x)$ is defined on $[-L, L]$, then the special generalized Fourier series

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right) \right) \tag{4.4}$$

is called the (classical) **Fourier series** of $f(x)$. The coefficients are computed by using (4.3) to get

$$\begin{aligned} a_0 &= \frac{1}{L} \int_{-L}^{L} f(x)\, dx \\ a_n &= \frac{1}{L} \int_{-L}^{L} f(x) \cos\left(\frac{n\pi x}{L}\right) dx, \quad n = 1, 2, \ldots \\ b_n &= \frac{1}{L} \int_{-L}^{L} f(x) \sin\left(\frac{n\pi x}{L}\right) dx, \quad n = 1, 2, \ldots \end{aligned} \tag{4.5}$$

because we have already seen in Example 4.2.6 that the norms were

$$\|1\| = \sqrt{2L}, \quad \left\|\cos\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{L}, \quad \left\|\sin\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{L}, \quad \text{for } n = 1, 2, \ldots$$

Notice that it is convenient to use $\frac{a_0}{2}$ instead of a_0 in (4.4) so that the coefficient, $1/L$, of the first integral is the same as the remaining ones. The extra factor of 2 comes from the fact that the square norm of the function 1 is twice the size of the other square norms.

Relations (4.5) are the (classical) **Fourier coefficients** of $f(x)$.

If we use the interval $[0, 2L]$ instead of $[-L, L]$ for the same set of functions we get another version of the classical Fourier Series which for $[0, 2L]$ is given by

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{n\pi x}{L}\right) + b_n \sin\left(\frac{n\pi x}{L}\right) \right) \tag{4.6}$$

and

$$\begin{aligned} a_0 &= \frac{1}{L} \int_{0}^{2L} f(x)\, dx \\ a_n &= \frac{1}{L} \int_{0}^{2L} f(x) \cos\left(\frac{n\pi x}{L}\right) dx, \quad n = 1, 2, \ldots \\ b_n &= \frac{1}{L} \int_{0}^{2L} f(x) \sin\left(\frac{n\pi x}{L}\right) dx, \quad n = 1, 2, \ldots \end{aligned} \tag{4.7}$$

Example: The Fourier Sine Series

In Example 4.2.3 we saw that the set of functions

$$\sin\left(\frac{\pi x}{L}\right),\ \sin\left(\frac{2\pi x}{L}\right),\ldots,\ \sin\left(\frac{n\pi x}{L}\right),\ldots$$

is orthogonal on $[0,L]$. If a function $f(x)$ is defined on $[0,L]$, then the special generalized Fourier series

$$f(x)=\sum_{n=1}^{\infty}b_n\sin\left(\frac{n\pi x}{L}\right) \tag{4.8}$$

is called the **Fourier Sine series** of $f(x)$. The coefficients are computed by using (4.3) to get

$$b_n=\frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx,\quad n=1,2,\ldots \tag{4.9}$$

because we have already seen in Example 4.2.3 that the norms were

$$\left\|\sin\left(\frac{n\pi x}{L}\right)\right\|=\sqrt{\frac{L}{2}},\quad \text{for}\ \ n=1,2,\ldots$$

Relations (4.9) are the **Fourier Sine series coefficients** of $f(x)$.

Example: The Fourier Cosine Series

In Example 4.2.4 we saw that the set of functions

$$1,\ \cos\left(\frac{\pi x}{L}\right),\ \ \cos\left(\frac{2\pi x}{L}\right),\ldots,\cos\left(\frac{n\pi x}{L}\right),\ldots$$

is orthogonal on $[0,L]$. If a function $f(x)$ is defined on $[0,L]$, then the special generalized Fourier series

$$f(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left(a_n\cos\left(\frac{n\pi x}{L}\right)+b_n\sin\left(\frac{n\pi x}{L}\right)\right) \tag{4.10}$$

is called the **Fourier Cosine series** of $f(x)$. The coefficients are computed by using (4.3) to get

$$a_0 = \frac{2}{L}\int_0^L f(x)\,dx \tag{4.11}$$
$$a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)dx, \quad n = 1, 2, \ldots$$

because we have already seen in Example 4.2.4 that the norms were

$$\|1\| = \sqrt{L}, \qquad \left\|\cos\left(\frac{n\pi x}{L}\right)\right\| = \sqrt{\frac{L}{2}} \quad , \quad \text{for} \quad n = 1, 2, \ldots$$

Notice that just as with the classical Fourier series it is convenient to use $\frac{a_0}{2}$ instead of a_0 in (4.11) so that the coefficient, $2/L$, of the first integral is the same as the remaining ones.

Relations (4.11) are the **Fourier Cosine series coefficients** of $f(x)$.

Orthogonality with Respect to a Weight Function

Let $p(x)$ be a positive function defined on $[a, b]$. I.e., $p(x) > 0$ for all x in $[a, b]$. The assignment

$$(f, g) \to \langle f, g\rangle = \int_a^b p(x)\, f(x)\, g(x)\, dx$$

defines as inner product on the vector space of all real-valued continuous functions $C[a, b]$ defined on $[a, b]$. The norm defined by this inner product is

$$\|f\| = \sqrt{\int_a^b p(x)\, f^2(x)\, dx}$$

Definition Let $p(x)$ be a positive function defined on $[a, b]$. The sequence of functions $g_1(x), g_2(x), \ldots$ is an **orthogonal set** on $[a, b]$ **with respect to the weight function** $p(x)$, if the inner product with weight $p(x)$ is zero.

$$\langle g_m, g_n\rangle = \int_a^b p(x)\, g_m(x)\, g_n(x)\, dx = 0, \qquad \text{for } m \neq n$$

If each function in an orthogonal set has norm 1 with respect to the weight function $p(x)$, then the set is **orthonormal with respect to the weight function** $p(x)$.

Notes

1. Orthogonality is the same as orthogonality with respect to the weight function $p(x) = 1$, for all x in $[a, b]$.

2. If the $g_1(x), g_2(x), \ldots$ is orthogonal with respect to weight $p(x)$ and we set $h_n(x) = \sqrt{p(x)} g_n(x)$, then by the weighted orthogonality we get

$$\begin{aligned}\int_a^b h_m(x) h_n(x)\, dx &= \int_a^b \sqrt{p(x)} g_m(x) \sqrt{p(x)} g_n(x)\, dx \\ &= \int_a^b p(x) g_m(x) g_n(x)\, dx \\ &= 0\end{aligned}$$

So the functions $h_1(x), h_2(x), \ldots$ are orthogonal in the usual sense.

4.4 Sturm-Liouville Theory

A Sturm-Liouville problem (S-L) defined on $[a, b]$ is a boundary value problem for a second order homogeneous differential equation in unknown function $y = y(x)$ that can be written in the form

$$[r(x) y']' + [q(x) + \lambda p(x)] y = 0$$

with two boundary conditions of the form

$$\begin{aligned}k_1 y(a) + k_2 y'(a) &= 0 \\ l_1 y(b) + l_2 y'(b) &= 0\end{aligned}$$

where the constants k_1, k_2 are not both zero and the constants l_1, l_2 are also not both zero. The number λ is called the **parameter** of the S-L problem.

Note that a S-L problem has always the trivial solution $y(x) = 0$ for all x in $[a, b]$. If λ is a scalar such that the S-L problem has a nontrivial solution $y(x)$, λ is called an eigenvalue of the problem and the nontrivial $y(x)$ is called an **eigenfunction** corresponding to λ.

Example 4.4.1. Find the eigenvalues and eigenfunctions of the S-L problem.

$$y'' + \lambda y = 0, \qquad y(0) = 0, \qquad y(\pi) = 0$$

Solution: We have the following cases:

Case 1 Let $\lambda < 0$, say $\lambda = -v^2$ for $v > 0$. Then we have

$$y'' - v^2 y = 0 \Rightarrow r^2 - v^2 = 0 \Rightarrow r = \pm v$$

This is a case of two real roots. So

$$y(x) = c_1 e^{vx} + c_2 e^{-vx}$$

Using the boundary conditions: The first yields $y(0) = 0 = c_1 + c_2$, so $c_2 = -c_1$. The second condition yields $y(\pi) = c_1(e^{v\pi} - e^{-v\pi}) = 0$. So, $c_1 = 0$. Hence, $c_2 = 0$. We only get the trivial solution.

Case 2 Let $\lambda = 0$. Then $y''(x) = 0$. Hence, $y(x) = c_1 x + c_2$, by integration. Using the boundary conditions we get $y(0) = 0 = c_1$. Hence, $y(x) = c_2$. Now $y(\pi) = c_2 = 0$. Thus, we again get the trivial solution.

Case 3 Let $\lambda > 0$, say $\lambda = v^2$ for $v > 0$. Then we have

$$y'' + v^2 y = 0 \Rightarrow r^2 + v^2 = 0 \Rightarrow r = \pm vi$$

This is a case of two complex conjugate roots. So

$$y(x) = c_1 \cos(vx) + c_2 \sin(vx)$$

Using the boundary conditions we get $y(0) = 0 = c_1$. So $y(x) = c_2 \sin(vx)$. Now $y(\pi) = c_2 \sin(v\pi) = 0$. If $c_2 = 0$, we get the trivial solution. If $c_2 \neq 0$, then $\sin(v\pi) = 0$. Hence, $v\pi = n\pi$, where n is any integer. Therefore, $v = n$ is an integer. So there are infinitely many eigenvalues

$$\lambda_n = n^2$$

with corresponding eigenfunctions

$$y_n(x) = \sin(nx), \quad n = 1, 2, 3, \ldots$$

Exercise Find the eigenvalues and eigenfunctions of the S-L problem.

$$y'' + \lambda y = 0, \quad y(\pi) = y(-\pi) \quad y'(\pi) = y'(-\pi)$$

Orthogonality of Eigenfunctions

Theorem 4.4.1 (Orthogonality of Eigenfunctions)**.** *If $p(x), q(x), r(x)$, and $r'(x)$ are real-valued continuous functions defined on $[a,b]$ for an S-L problem.*

$$\begin{aligned}[r(x)y']' + [q(x) + \lambda p(x)]y &= 0\\ k_1 y(a) + k_2 y'(a) &= 0\\ l_1 y(b) + l_2 y'(b) &= 0\end{aligned}$$

Let $y_m(x)$ and $y_n(x)$ be two eigenfunctions corresponding to two different eigenvalues λ_m and λ_n. Then $y_m(x)$ and $y_n(x)$ are orthogonal with respect to weight function $p(x)$. Furthermore:

1. *If $r(a) = 0$, then the first boundary condition can be dropped.*
2. *If $r(b) = 0$, then the second boundary condition can be dropped.*
3. *If $r(a) = r(b)$, then the two boundary conditions can be replaced by*

$$y(a) = y(b), \quad y'(a) = y'(b)$$

Theorem 4.4.2 (Real Eigenvalues)**.** *If $p(x), q(x), r(x)$, and $r'(x)$ are real-valued continuous functions defined on $[a,b]$ for*

$$\begin{aligned}[r(x)y']' + [q(x) + \lambda p(x)]y &= 0\\ k_1 y(a) + k_2 y'(a) &= 0\\ l_1 y(b) + l_2 y'(b) &= 0\end{aligned}$$

and $p(x)$ is either positive in the entire interval $[a,b]$, or negative in the entire interval $[a,b]$, then all the eigenvalues are real numbers.

Example: Periodic Boundary Conditions

Example 4.4.2 (Periodic Ends)**.** Find the eigenvalues and eigenfunctions of the S-L problem.

$$y'' + \lambda y = 0, \quad y(0) = y(2\pi), \quad y'(0) = y'(2\pi)$$

Solution: We have

Case 1 Let $\lambda < 0$, say $\lambda = -v^2$. Then the auxiliary is $r^2 - v^2 = 0$. We get $r = \pm v$. So, $y(x) = c_1 e^{vx} + c_2 e^{-vx}$. Using the boundary conditions we have $y' = vc_1 e^{vx} - vc_2 e^{-vx}$. Hence,

$$\begin{aligned} y(0) = c_1 + c_2 = y(2\pi) = c_1 e^{2\pi v} + c_2 e^{-2\pi v} \\ y'(0) = vc_1 - vc_2 = y'(2\pi) = vc_1 e^{2\pi v} - vc_2 e^{-2\pi v} \end{aligned}$$

Thus, we get the homogoenuous linear system in c_1 and c_2

$$\begin{aligned} c_1\left(1 - e^{2\pi v}\right) + c_2\left(1 - e^{-2\pi v}\right) = 0 \\ c_1\left(1 - e^{2\pi v}\right) + c_2\left(e^{-2\pi v} - 1\right) = 0 \end{aligned}$$

The coefficient matrix has determinant

$$\begin{vmatrix} 1 - e^{2\pi v} & 1 - e^{-2\pi v} \\ 1 - e^{2\pi v} & e^{-2\pi v} - 1 \end{vmatrix} = 2e^{-2\pi v} + 2e^{2\pi v} - 4 = 2\left(e^{\pi v} - e^{-\pi v}\right)^2 \neq 0$$

So the system has only the trivial solution.

Case 2 Let $\lambda = 0$. Then $y''(x) = 0$. Hence, $y(x) = c_1 x + c_2$, by integration. Using the boundary conditions we get

$$\begin{aligned} y(0) = c_2 = y(2\pi) = 2\pi c_1 + c_2 \\ y'(0) = c_1 = y'(2\pi) = c_1 \end{aligned}$$

Hence, $c_1 = 0$. However, there is no restriction on c_2. So $y(x)$ can be any constant. Say, $y(x) = a_0$.

Case 3 Let $\lambda < 0$, say $\lambda = -v^2$. Then the auxiliary is $r^2 + v^2 = 0$. We get $r = \pm iv$. So, $y(x) = c_1 \cos(vx) + c_2 \sin(vx)$. Using the boundary conditions we have $y' = -vc_1 \sin vx + vc_2 \cos vx$. Hence,

$$\begin{aligned} y(0) = c_1 = y(2\pi) = c_1 \cos(2\pi v) + c_2 \sin(2\pi v) \\ y'(0) = vc_2 = y'(2\pi) = -vc_1 \sin(2\pi v) + vc_2 \cos(2\pi v) \end{aligned}$$

Thus, we get the homogoenuous linear system in c_1 and c_2

$$\begin{aligned} c_1(1 - \cos(2\pi v)) + c_2(-\sin(2\pi v)) = 0 \\ c_1(\sin(2\pi v)) + c_2(1 - \cos(2\pi v)) = 0 \end{aligned}$$

For nontrivial solutions the coefficient determinant must be zero.

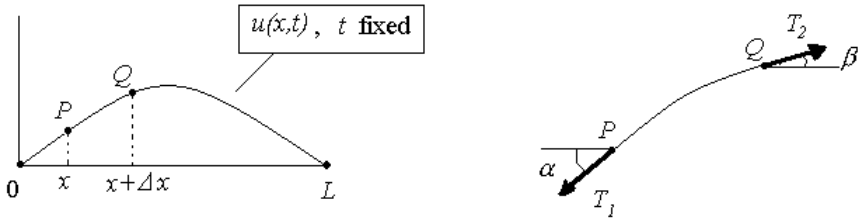
$$\begin{vmatrix} 1-\cos(2\pi v) & -\sin(2\pi v) \\ \sin(2\pi v) & 1-\cos(2\pi v) \end{vmatrix} = 2-2\cos(2\pi v) = 4\sin^2(\pi v) = 0$$

Therefore, $\pi v = n\pi$, where n is an integer. So the system has eigenfunctions $c_1\cos(nx) + c_2\sin(nx)$, n is any integer.

So all eigenfunctions are nontrivial linear combinations of the eigenfunctions

$$1, \cos x, \cos(2x), \cos(3x), \ldots, \sin x, \sin(2x), \sin(3x), \ldots$$

4.5 Modeling the Vibrating String



We consider a string of length L attached to fixed points with x-coordinates 0 and L. Let $u(x,t)$ be the **deflection** or **displacement** (signed vertical distance from the x-axis) of the string at location x at time t.

Goal: Calculate $u(x,t)$, given (a) the ends of the strings are fixed and (b) initial displacement $u(x,0)$ and initial velocity $u_t(x,0)$.

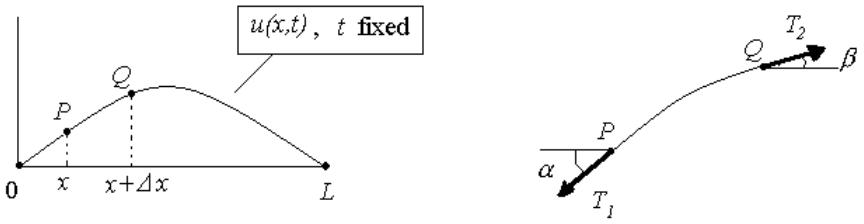
We need assumptions to simplify the partial differential equation for $u(x,t)$. This is because PDEs are very hard or impossible to solve exactly.

Assumptions:

1. The mass of the string per unit length is constant (homogeneous string).
2. The string is completely elastic. It does not resist to bending.
3. The tension caused by stretching is much greater that gravity. So, gravity is not a factor here.
4. The string performs a small transverse motion in the vertical plane, so that both the deflection $u(x,t)$ and its slope $u_x(x,t)$ are small.

Modeling the Vibrating String

Forces:



Consider forces acting on small portions of the string.

Since there is no resistance to bending, the tension is tangential to the curve of the string at each point.

Let T_1 and T_2 be the tensions at P and Q.

Horizontal direction: There is no motion in the horizontal direction, so the horizontal component must be constant, say T. So

$$T_1 \cos\alpha = T_2 \cos\beta = T \tag{4.12}$$

Vertical direction: In the vertical direction we have two forces, the vertical components $-T_1 \sin\alpha$ and $T_2 \sin\beta$.

Let ρ be the *linear mass density* of the string, i.e., mass per unit length. By Newton's second law the resultant force is mass $\rho\Delta x$ times acceleration $\frac{\partial^2 u}{\partial t^2}$ evaluated at some point between x and $x + \Delta x$.

$$T_2 \sin\beta - T_1 \sin\alpha = \rho\Delta x \frac{\partial^2 u}{\partial t^2}$$

Using (4.12) we get

$$\frac{T_2 \sin\beta}{T_2 \cos\beta} = \frac{T_1 \sin\alpha}{T_1 \cos\alpha} = \tan\beta - \tan\alpha = \rho\frac{\Delta x}{T}\frac{\partial^2 u}{\partial t^2}$$

But $\tan\alpha = \left(\frac{\partial u}{\partial x}\right)_x$, $\tan\beta = \left(\frac{\partial u}{\partial x}\right)_{x+\Delta x}$ are the slopes at x and $x + \Delta x$. So we have

$$\frac{1}{\Delta x}\left(\left(\frac{\partial u}{\partial x}\right)_{x+\Delta x} - \left(\frac{\partial u}{\partial x}\right)_x\right) = \frac{\rho}{T}\frac{\partial^2 u}{\partial t^2}$$

Now we take the limit as $\Delta x \to 0$ to get

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{4.13}$$

where $c^2 = \frac{T}{\rho}$.
Equation (4.13) is the **one-dimensional wave equation**.

4.6 The One-Dimensional Wave Equation

Solving the One-Dimensional Wave Equation

We solve the one-dimensional **wave equation**

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{4.14}$$

subject to the fixed-ends **boundary conditions**

$$u(0,t) = 0, \qquad u(L,t) = 0, \qquad t \geq 0 \tag{4.15}$$

and initial conditions specifying an initial deflection $f(x)$ and initial velocity $g(x)$, for x such that $0 \leq x \leq L$.

$$u(x,0) = f(x), \qquad \left.\frac{\partial u}{\partial t}\right|_{t=0} = g(x), \qquad 0 \leq x \leq L \tag{IC}$$

Method of solution

Stage 1: Separation of Variables

First we seek nontrivial solutions of the system (W-1), (4.15). Notice that the trivial solution is already a solution.To solve (W-1), (4.15) we use the method of separation of variables. I.e., we seek solutions of the form.

$$u(x,t) = X(x)\,T(t)$$

where $X = X(x)$ is a function of x only and $T = T(t)$ is a function of t only. Substitution into (W-1) yields

$$XT'' = c^2 X''T$$

where by X' we mean $\frac{dX}{dx}$ and by T' we mean $\frac{dT}{dt}$. Now we separate the variables by dividing both sides by c^2XT to get

$$\frac{T''}{c^2T} = \frac{X''}{X}$$

Now x and t are completely independent variables, one being location and one being time. So the only way the functions $\frac{T''}{c^2T}$ of t and $\frac{X''}{X}$ of x is if they are both the same constant, say $-\lambda$. So,

$$\frac{T''}{c^2T} = \frac{X''}{X} = -\lambda$$

Therefore, we get a system of two ordinary differential equations homogeneous with constant coefficients: one in X only and only in T only.

$$T'' + c^2\lambda T = 0, \qquad X'' + \lambda X = 0$$

These can be readily solved, provided we know the constant λ.

We use $X'' + \lambda X = 0$ and the boundary conditions (4.15) to find λ and $X(x)$. The boundary conditions (4.15) are written in terms of X and T. For all $t \geq 0$

$$u(0,t) = X(0)\,T(t) = 0, \qquad u(L,t) = X(L)\,T(t) = 0$$

T cannot be identically zero ($T(t) = 0$, for all t), or else $u(x,t)$ would be zero for all x and t, hence we would get the trivial solution. So we must have $X(0) = 0$ and $X(L) = 0$. We get the Sturm-Liouville problem

$$X'' + \lambda X = 0, \qquad X(0) = 0, \qquad X(L) = 0$$

which we have essentially solved before. We have the following cases:

Case 1 Let $\lambda < 0$, say $\lambda = -v^2$ for $v > 0$. Then we have

$$X'' - v^2X = 0 \Rightarrow r^2 - v^2 = 0 \Rightarrow r = \pm v$$

This is a case of two real roots. So

$$X(x) = c_1e^{vx} + c_2e^{-vx}$$

Using the boundary conditions we get $X(0) = 0 = c_1 + c_2$ and $X(L) = c_1e^{vL} + c_2e^{-vL} = 0$. So $c_2 = -c_1$. Hence, $X(L) = c_1\left(e^{vL} - e^{-vL}\right) = 0$. Thus, $c_1 = 0$. So, $c_2 = 0$ and we get the trivial solution.

Case 2 Let $\lambda = 0$. Then $X''(x) = 0$. Hence, $X(x) = c_1 x + c_2$, by integration. Using the boundary conditions we get $X(0) = 0 = c_2$. Hence, $X(x) = c_1 x$. Now $X(L) = c_1 L = 0$. So, $c_1 = 0$. Thus, we again get the trivial solution.

Case 3 Let $\lambda > 0$, say $\lambda = v^2$ for $v > 0$. Then we have

$$X'' + v^2 X = 0 \Rightarrow r^2 + v^2 = 0 \Rightarrow r = \pm vi$$

This is a case of two complex conjugate roots. So

$$X(x) = c_1 \cos(vx) + c_2 \sin(vx)$$

Using the boundary conditions we get $y(0) = 0 = c_1$. So $X(x) = c_2 \sin(vx)$. Now $X(L) = c_2 \sin(v\pi) = 0$. If $c_2 = 0$, we get the trivial solution. If $c_2 \neq 0$, then $\sin(vL) = 0$. Hence, $vL = n\pi$, where n is any integer. Therefore, $v = n\pi/L$. So there are infinitely many eigenvalues

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, 3, \ldots$$

with corresponding eigenfunctions

$$X_n(x) = \sin\left(\frac{n\pi x}{L}\right), \quad n = 1, 2, 3, \ldots$$

Note that since $\sin(-x) = -\sin(x)$ and $\sin(0) = 0$, so we need not keep any negative integer values for n. These signs can be absorbed by the constant coefficients.

Now that we know X and λ we turn to T. The equation $T'' + c^2 \lambda T = 0$ takes the form

$$T_n'' + \left(\frac{cn\pi}{L}\right)^2 T_n = 0$$

which can be solved right away, becuase $\left(\frac{cn\pi}{L}\right)^2 > 0$. The auxiliary is $r^2 + \left(\frac{cn\pi}{L}\right)^2 = 0$. So, $r = \pm\left(\frac{cn\pi}{L}\right) i$. We have for any constants a_n and b_n

$$T_n = a_n \cos\left(\frac{cn\pi t}{L}\right) + b_n \sin\left(\frac{cn\pi t}{L}\right), \quad n = 1, 2, 3, \ldots$$

Hence, $u = X_n T_n$ becomes

$$u_n(x, t) = \left(a_n \cos\left(\frac{cn\pi t}{L}\right) + b_n \sin\left(\frac{cn\pi t}{L}\right)\right) \sin\left(\frac{n\pi x}{L}\right), \quad n = 1, 2, 3, \ldots$$

Note that since the system (W-1), (4.15) is homogeneous, any finite sum of solutions u_n is also a solution.

$$u(x,t) = \sum_{n=1}^{k} u_n(x,t)$$

Under certain conidtions an infinite sum of solutions is also a solution

$$u(x,t) = \sum_{n=1}^{\infty} u_n(x,t)$$

So we may have a general solution of the form

$$u(x,t) = \sum_{n=1}^{\infty} \left(a_n \cos\left(\frac{cn\pi t}{L}\right) + b_n \sin\left(\frac{cn\pi t}{L}\right) \right) \sin\left(\frac{n\pi x}{L}\right) \tag{4.16}$$

Stage 2: Fourier Analysis

Solution (4.16) is the kind of solution this method produces, provided we know the coefficients a_n and b_n. These can be computed by using the boundary conditions $u(x,0) = f(x)$ and $\frac{\partial u}{\partial t}\Big|_{t=0} = g(x)$.

Using the first condition and 4.16) with $t = 0$ yields

$$u(x,0) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right) = f(x)$$

Now recall that the functions $s_n(x) = \sin\left(\frac{n\pi}{L}x\right)$ were eigenfunctions to the S-L problem on $[0,L]$. Therefore, by Theorem 1 on S-L problems, these eigenfunctions must be *orthogonal*. So we can use the generic formula $a_n = \langle f, s_n\rangle / \langle s_n, s_n\rangle$ to find a_n. We have

$$\begin{aligned} a_n &= \frac{\langle f, s_n\rangle}{\langle s_n, s_n\rangle} = \frac{\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx}{\int_0^L \sin^2\left(\frac{n\pi x}{L}\right)dx} = \frac{\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx}{\frac{1}{2}\int_0^L \left(1-\cos\left(\frac{2n\pi x}{L}\right)\right)dx} \\ &= \frac{\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx}{L/2} = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx \end{aligned}$$

Using the second condition and 4.16) with $t = 0$ yields

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = \sum_{n=1}^{\infty} \left(b_n \frac{cn\pi}{L}\right) \sin\left(\frac{n\pi x}{L}\right) = g(x)$$

The functions $s_n(x) = \sin\left(\frac{n\pi}{L}x\right)$ are orthogonal, So we can use the generic formula $b_n \frac{cn\pi}{L} = \langle g, s_n\rangle / \langle s_n, s_n\rangle$ to find b_n. We have

$$b_n = \frac{L}{cn\pi}\frac{\langle f, s_n\rangle}{\langle s_n, s_n\rangle} = \frac{L}{cn\pi}\frac{\int_0^L g(x)\sin\left(\frac{n\pi x}{L}\right)dx}{\int_0^L \sin^2\left(\frac{n\pi x}{L}\right)dx} = \frac{L}{cn\pi}\frac{\int_0^L g(x)\sin\left(\frac{n\pi x}{L}\right)dx}{L/2}$$
$$= \frac{2}{cn\pi}\int_0^L g(x)\sin\left(\frac{n\pi x}{L}\right)dx$$

So the method of separation of variables yields the following solution to the one-dimensional wave equation.

$$u(x,t) = \sum_{n=1}^{\infty}\left(a_n\cos\left(\frac{cn\pi t}{L}\right) + b_n\sin\left(\frac{cn\pi t}{L}\right)\right)\sin\left(\frac{n\pi x}{L}\right) \qquad (4.17)$$
$$a_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx, \qquad n = 1, 2, \ldots$$
$$b_n = \frac{2}{cn\pi}\int_0^L g(x)\sin\left(\frac{n\pi x}{L}\right)dx, \qquad n = 1, 2, \ldots$$

Normal Modes

The terms in the series solution for $u(x,t)$

$$u_n(x,t) = \left[a_n\cos\left(\frac{cn\pi t}{L}\right) + b_n\sin\left(\frac{cn\pi t}{L}\right)\right]\sin\left(\frac{n\pi x}{L}\right), \qquad n = 1, 2, \ldots$$

are called the **eigenfunctions** or **characteristic functions** of the problem. The numbers $\lambda_n = \frac{cn\pi}{L}$ are the **eigenvalues** of the problem. The set of all eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n, \ldots\}$ is the **spectrum**.

Each $u_n(x,t)$ represents a harmonic motion of period $\frac{\lambda_n}{2\pi} = \frac{cn}{2L}$. The eigenfunction $u_n(x,t)$ is also called the nth **normal mode**. If $n = 1$, we have the **fundamental mode**. If $n > 1$, we have the **overtones**.

The nth normal mode has $n-1$ **nodes**, i.e., points on the string that never move. The nodes can be computed as follows:

We want $u_n(x,t) = 0$ for all $t \geq 0$. Hence, $\sin\left(\frac{n\pi x}{L}\right) = 0$, $n = 1, 2, \ldots$. Therefore,

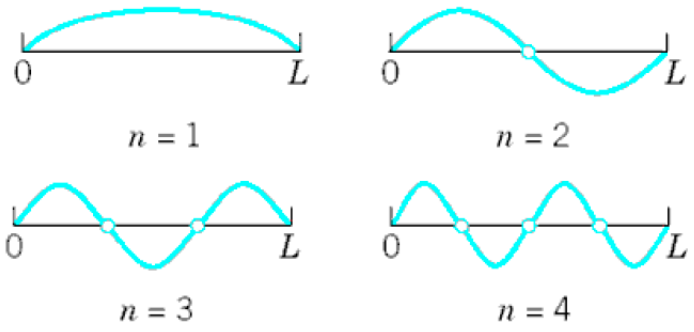
$$\frac{n\pi x}{L} = k\pi, \qquad k \text{ integer}$$

So

$$x = \frac{k}{n}L, \qquad k \text{ integer for fixed } n = 1, 2, \ldots$$

Now $0 \leq x \leq L$. Hence, $k = 0, 1, \ldots, n-1$ and the nodes plus the endpoints are

$$x = 0, \quad \frac{L}{n}, \quad \frac{2L}{n}, \quad \cdots, \quad \frac{(n-1)\,L}{n}, \quad L$$



Normal modes of the vibrating string

4.7 One Dimensional Wave Equation: Examples

Recall that

$\dfrac{\partial^2 u}{\partial t^2} = c^2 \dfrac{\partial^2 u}{\partial x^2}$	The Wave Equation
$u(0,t) = 0, \quad t \geq 0$	Boundary Condition (fixed left end)
$u(L,t) = 0, \quad t \geq 0$	Boundary Condition (fixed right end)
$u(x,0) = f(x), \qquad 0 \leq x \leq L$	Initial Condition (initial displacement)
$\left.\dfrac{\partial u}{\partial t}\right\|_{t=0} = g(x), \qquad 0 \leq x \leq L$	Initial Condition (initial velocity)

By using separation of variables we arrived at the following series solution.

$$u(x,t) = \sum_{n=1}^{\infty} \left[a_n \cos\left(\frac{cn\pi t}{L}\right) + b_n \sin\left(\frac{cn\pi t}{L}\right) \right] \sin\left(\frac{n\pi x}{L}\right)$$

$$a_n = \frac{2}{L}\int_0^L f(x) \sin\left(\frac{n\pi x}{L}\right) dx, \qquad n = 1, 2, \ldots$$

$$b_n = \frac{2}{cn\pi}\int_0^L g(x) \sin\left(\frac{n\pi x}{L}\right) dx, \qquad n = 1, 2, \ldots$$

Example 4.7.1. Consider the fixed ends vibrating string problem

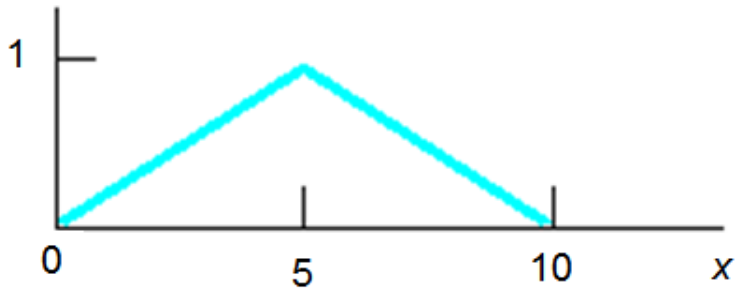
$$\frac{\partial^2 u}{\partial t^2} = 4\frac{\partial^2 u}{\partial x^2}$$
$$u(0,t) = u(10,t) = 0$$

with initial velocity zero and initial displacement the tent function

$$u(x,0) = f(x) = \begin{cases} x/5 & \text{if} \quad 0 < x < 5 \\ (10-x)/5 & \text{if} \quad 5 \leq x < 10 \end{cases}$$

1. Solve for the displacement $u(x,t)$.

2. Use the first three nonzero terms of the answer to approximate the displacement u at location 2 units and time 1 unit.



A tent function

Solution:

1. We have $L = 10$, $c = 2$, and $b_n = 0$ since the initial velocity is zero. So we are left with

$$u(x,t) = \sum_{n=1}^{\infty} a_n \cos\left(\frac{2n\pi t}{10}\right) \sin\left(\frac{n\pi x}{10}\right) = \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi t}{5}\right) \sin\left(\frac{n\pi x}{10}\right)$$
$$a_n = \frac{2}{10}\int_0^{10} f(x) \sin\left(\frac{n\pi x}{10}\right) dx, \qquad n = 1, 2, \ldots$$

We have

$$a_n = \frac{1}{5}\int_0^{10} f(x) \sin\left(\frac{n\pi x}{10}\right) dx = \frac{1}{5}\int_0^{5} \frac{x}{5} \sin\left(\frac{n\pi x}{10}\right) dx + \frac{1}{5}\int_5^{10} \frac{10-x}{5} \sin\left(\frac{n\pi x}{10}\right) dx$$
$$= \frac{1}{25}\int_0^{5} x \sin\left(\frac{n\pi x}{10}\right) dx + \frac{1}{25}\int_5^{10} (10-x) \sin\left(\frac{n\pi x}{10}\right) dx$$

We integrate by parts with $u = x$ for the first integral and $u = 10 - x$ for the second to get

$$\frac{1}{25}\left(-\frac{10}{n\pi} x \cos\left(\frac{n\pi x}{10}\right)\Big|_0^5 - \left(\frac{-10}{n\pi}\right)\int_0^5 \cos\left(\frac{n\pi x}{10}\right) dx\right) +$$
$$+ \frac{1}{25}\left(-\frac{10}{n\pi}(10-x) \cos\left(\frac{n\pi x}{10}\right)\Big|_5^{10} - \left(\frac{-10}{n\pi}\right)\int_5^{10} (-1) \cos\left(\frac{n\pi x}{10}\right) dx\right)$$
$$= \frac{1}{25}\left(\frac{-50}{n\pi}\cos\left(\frac{n\pi}{2}\right) + \frac{100}{n^2\pi^2}\sin\left(\frac{n\pi x}{10}\right)\Big|_0^5\right) +$$
$$+ \frac{1}{25}\left(\frac{50}{n\pi}\cos\left(\frac{n\pi}{2}\right) - \frac{100}{n^2\pi^2}\sin\left(\frac{n\pi x}{10}\right)\Big|_5^{10}\right)$$
$$= \frac{8}{n^2\pi^2}\sin\left(\frac{n\pi}{2}\right)$$

Therefore,

$$u(x,t) = \frac{8}{\pi^2}\sum_{n=1}^{\infty} \frac{1}{n^2} \sin\left(\frac{n\pi}{2}\right) \cos\left(\frac{n\pi t}{5}\right) \sin\left(\frac{n\pi x}{10}\right)$$

2. Expanding the sum yields

$$\begin{aligned} u(x,t) &= \frac{8}{\pi^2}\Big(\frac{1}{1^2}\sin\left(\frac{\pi}{2}\right)\cos\left(\frac{\pi t}{5}\right)\sin\left(\frac{\pi x}{10}\right) + \frac{1}{2^2}\sin(\pi)\cos\left(\frac{2\pi t}{5}\right)\sin\left(\frac{\pi x}{5}\right) \\ &\quad + \frac{1}{3^2}\sin\left(\frac{3\pi}{2}\right)\cos\left(\frac{3\pi t}{5}\right)\sin\left(\frac{3\pi x}{10}\right) + \frac{1}{4^2}\sin(2\pi)\cos\left(\frac{4\pi t}{5}\right)\sin\left(\frac{2\pi x}{5}\right) \\ &\quad + \frac{1}{5^2}\sin\left(\frac{5\pi}{2}\right)\cos(\pi t)\sin\left(\frac{\pi x}{2}\right) + \cdots\Big) \\ &\simeq \frac{8}{\pi^2}\left(\cos\left(\frac{\pi t}{5}\right)\sin\left(\frac{\pi x}{10}\right) - \frac{1}{9}\cos\left(\frac{3\pi t}{5}\right)\sin\left(\frac{3\pi x}{10}\right) + \frac{1}{25}\cos(\pi t)\sin\left(\frac{\pi x}{2}\right)\right) \end{aligned}$$

So,

$$u(x,t) \simeq \frac{8}{\pi^2}\left(\cos\left(\frac{\pi t}{5}\right)\sin\left(\frac{\pi x}{10}\right) - \frac{1}{9}\cos\left(\frac{3\pi t}{5}\right)\sin\left(\frac{3\pi x}{10}\right) + \frac{1}{25}\cos(\pi t)\sin\left(\frac{\pi x}{2}\right)\right)$$

Therefore,

$$\begin{aligned} u(2,1) &\simeq \frac{8}{\pi^2}\left(\cos\left(\frac{\pi}{5}\right)\sin\left(\frac{2\pi}{10}\right) - \frac{1}{9}\cos\left(\frac{3\pi}{5}\right)\sin\left(\frac{6\pi}{10}\right) + \frac{1}{25}\cos(\pi)\sin(\pi)\right) \\ &\simeq 0.411\,92 \quad \text{height units} \end{aligned}$$

More Problems on the Wave Equation

In problems 1 and 2 solve the initial boundary value problem by finding $u(x,t)$.

$$\frac{\partial^2 u}{\partial t^2} = 9\frac{\partial^2 u}{\partial x^2}, \quad u(0,t) = 0, \quad u(4,t) = 0 \ (t > 0), \quad u(x,0) = f(x), \quad \left.\frac{\partial u}{\partial t}\right|_{t=0} = 0$$

Problem 1 $f(x)$ is given by

$$f(x) = \begin{cases} 2 - x & \text{if} \quad 0 \le x \le 2 \\ 0 & \text{if} \quad 2 \le x \le 4 \end{cases}$$

Problem 2 $f(x)$ is given by

$$f(x) = -2\sin(2\pi x) + 7\sin(5\pi x)$$

Solution: We have $c=3$, $L=4$ and $b_n=0$ because the initial velocity is zero. So

$$u(x,t)=\sum_{n=1}^{\infty}a_n\cos\left(\frac{3n\pi t}{4}\right)\sin\left(\frac{n\pi x}{4}\right),\qquad a_n=\frac{2}{4}\int_0^4 f(x)\sin\left(\frac{n\pi x}{4}\right)dx$$

Solution of Problem 1:

$$a_n=\frac{1}{2}\int_0^4 f(x)\sin\left(\frac{n\pi x}{4}\right)dx=\frac{1}{2}\int_0^2(2-x)\sin\left(\frac{n\pi x}{4}\right)dx+\frac{1}{2}\int_2^4(0)\sin\left(\frac{n\pi x}{4}\right)dx$$
$$=\frac{1}{2}\int_0^2(2-x)\sin\left(\frac{n\pi x}{4}\right)dx$$

integration by parts with $u=2-x$ and $dv=\sin(n\pi x/4)\,dx$ yields ($du=-dx$, $v=-4\cos(n\pi x/4)/(n\pi)$)

$$a_n=\frac{-2}{n\pi}(2-x)\cos\left(\frac{n\pi x}{4}\right)\Big|_0^2-\left(\frac{-2}{n\pi}\right)\int_0^2\cos\left(\frac{n\pi x}{4}\right)(-dx)$$
$$=\frac{4}{\pi n}-\frac{8}{n^2\pi^2}\sin\left(\frac{n\pi x}{4}\right)\Big|_0^2=\frac{4}{\pi n}-\frac{8}{n^2\pi^2}\sin\left(\frac{n\pi}{2}\right)$$

Therefore,

$$a_n=\frac{4}{\pi n}-\frac{8}{n^2\pi^2}\sin\left(\frac{n\pi}{2}\right)$$

Hence,

$$u(x,t)=\frac{4}{\pi}\sum_{n=1}^{\infty}\left(\frac{1}{n}-\frac{2}{n^2\pi}\sin\left(\frac{n\pi}{2}\right)\right)\cos\left(\frac{3n\pi t}{4}\right)\sin\left(\frac{n\pi x}{4}\right)$$

Solution of Problem 2: Since the orthogonal functions $\sin\left(\frac{n\pi x}{4}\right)$ include the functions $\sin(2\pi x)$ and $\sin(5\pi x)$ of $f(x)$ we may use superposition (comparison) and avoid integration altogether. We have $a_8=-2$, $a_{20}=7$, $a_{rest}=0$. Hence,

$$u(x,t)=(-2)\cos\left(\frac{3(8)\pi t}{4}\right)\sin\left(\frac{8\pi x}{4}\right)+7\cos\left(\frac{3(20)\pi t}{4}\right)\sin\left(\frac{20\pi x}{4}\right)$$

Hence,

$$u(x,t)=-2\cos(6\pi t)\sin(2\pi x)+7\cos(15\pi t)\sin(5\pi x)$$

4.8 The Principle of Superposition

Let $g_1(x), g_2(x), \ldots, g_n(x), \ldots$ be an orthogonal set of functions on $[a,b]$. If we have two representations of a function $f(x)$ as generalized Fourier Series

$$f(x) = \sum_{n=1}^{\infty} a_n g_n(x), \quad f(x) = \sum_{n=1}^{\infty} b_n g_n(x)$$

then it must be that

$$a_n = b_n, \quad n = 1, 2, \ldots$$

So we have uniqueness of Fourier coefficients a_n of $f(x)$. (The proof: $a_n = \frac{\langle f, g_n \rangle}{\langle g_n, g_n \rangle} = b_n$)

This observation has the following application in computing the Fourier coefficients: If the function is a linear combination of some of the $g_n(x)$, then computing the Fourier coefficients a_n can be done by **comparison**, also known **superposition**. Thus, integration can be completely avoided.

Example 4.8.1 (Superposition)**.** Solve the $u(x,t)$ the deflection of a string with $L = 2$, $c^2 = 4$, initial velocity zero, and initial deflection $u(x,0) = 5\sin(2\pi x) - \frac{1}{2}\sin(3\pi x)$.

Solution: Because $c = 2$, $L = 2$, and the initial velocity being zero yields $b_n = 0$, we have

$$u(x,t) = \sum_{n=1}^{\infty} a_n \cos(n\pi t) \sin\left(\frac{n\pi x}{2}\right)$$

It remains to calculate the a_n.

$$u(x,0) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{2}\right) = 5\sin(2\pi x) - \frac{1}{2}\sin(3\pi x)$$

Since $\sin(2\pi x)$ and $\sin(3\pi x)$ appear on the left-hand side as part of the orthogonal set with $n = 4$ and $n = 6$. So, we may use superposition and compare coefficients:

$$a_4 = 5, \quad a_6 = -\frac{1}{2}, \quad a_{\text{rest}} = 0$$

Therefore,

$$u(x,t) = 5\cos(4\pi t)\sin(2\pi x) - \frac{1}{2}\cos(6\pi t)\sin(3\pi x)$$

Of course we could have used integration

$$\begin{aligned} a_n &= \frac{2}{2}\int_0^2 \left(5\sin(2\pi x) - \frac{1}{2}\sin(3\pi x)\right)\sin\left(\frac{n\pi x}{2}\right)dx \\ &= 5\int_0^2 \sin(2\pi x)\sin\left(\frac{n\pi x}{2}\right)dx - \frac{1}{2}\int_0^2 \sin(3\pi x)\sin\left(\frac{n\pi x}{2}\right)dx \end{aligned}$$

but we would have to exercise **caution**. If $n \neq 4$, then the first integral is zero, as seen by using the appropriate trigonometric identities. However, if $n = 4$, then the first integral becomes

$$\int_0^2 \sin^2(2\pi x)\,dx = \frac{1}{2}\int_0^2 (1 - \cos(2\pi x))\,dx = \frac{1}{2}\left(x - \frac{\sin(2\pi x)}{2\pi}\right)\Bigg|_0^2 = 1$$

So, the answer for the first term is 5. Likewise, for the second integral, if $n \neq 6$ we get zero, but if $n = 6$, we get 1, so the second term is $-\frac{1}{2}$ as asserted.

4.9 One Dimensional Heat Equation

The **one-dimensional heat equation** is given by

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}$$

and it models the temperature $u(x,t)$ at location x at time t of a thin metal rod.

We also have the **two-dimensional heat equation** given by

$$\frac{\partial u}{\partial t} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

modeling the temperature $u(x,y,t)$ at location (x,y) at time t of a thin metal plate, and the **three-dimensional heat equation** given by

$$\frac{\partial u}{\partial t} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right)$$

modeling the temperature $u(x,y,z,t)$ at location (x,y,z) at time t of a solid.

In general, the heat equation for a function $u(x_1, x_2, \ldots, x_n, t)$ is given by

$$\frac{\partial u}{\partial t} = c^2\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}\right)$$

also written as

$$\frac{\partial u}{\partial t} = c^2 \nabla^2 u$$

where $\nabla^2 u$ is the **Laplacian** of u defined by $\nabla^2 u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}$

Here we concentrate on the one-dimensional heat equation.

One Dimensional Heat Equation: Zero Ends

We consider a thin metal rod of length L placed along the x-axis so that its left end is at $x = 0$. We assume that temperature $u(x,t)$ is always zero at the two ends of the rod. Therefore, $u(0,t) = 0$ and $u(L,t) = 0$ for all $t \geq 0$. These are the boundary conditions of the problem. In addition, we assume that the initial temperature distribution is given, say $u(x,0) = f(x)$ for $0 \leq x \leq L$. This is the initial condition. So, we have the following IBVP.

$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}$	The 1-Dimensional Heat Equation
$u(0,t) = 0, \quad t \geq 0$	Boundary Condition (zero left end)
$u(L,t) = 0, \quad t \geq 0$	Boundary Condition (zero right end)
$u(x,0) = f(x), \quad 0 \leq x \leq L$	Initial Condition (initial temperature)

Note that we only have one initial condition, because we have only a first derivative in t.

To solve this IBVP we use just as before, separation of variables.

Solution: We start with

$$\begin{aligned} \frac{\partial u}{\partial t} &= c^2 \frac{\partial^2 u}{\partial x^2} \\ u(0,t) &= u(L,t) = 0 \\ u(x,0) &= f(x) \end{aligned}$$

and seek solutions of the form $u = X(x)T(t) = XT$. Then the PDE becomes

$$XT' = c^2 X''T \Rightarrow \frac{T'}{c^2 T} = \frac{X''}{X} = -k$$

Thus

$$X'' + kX = 0, \quad T' + c^2 kT = 0$$

The boundary conditions become $u(0,t) = X(0)T(t) = 0$ and $u(L,t) = X(L)T(t) = 0$. Hence, $X(0) = 0$ and $X(L) = 0$, since $T(t)$ cannot be zero for all t, or else we would get the trivial solution. We start with X.

$$X'' + kX = 0, \quad X(0) = 0, \quad X(L) = 0$$

Case 1: $k < 0$, say $k = -v^2$. Then $X(x) = c_1 e^{vx} + c_2 e^{-vx}$. But $X(0) = c_1 + c_2 = 0$, so $c_2 = -c_1$. And $X(L) = c_1 e^{vL} + c_2 e^{-vL} = 0$. Hence, $X(L) = c_1 \left(e^{vL} - e^{-vL}\right) = 0$. Since the second factor is not 0, we have $c_1 = 0$. But then $c_2 = 0$ and we get the trivial solution.
Case 2: $k = 0$. Then $X(x) = c_1 x + c_2$. But $X(0) = c_1(0) + c_2 = 0$, so $c_2 = 0$. And $X(L) = c_1 L = 0$. Hence, c_1 and we get the trivial solution.
Case 3: $k > 0$, say $k = v^2$. Then $X(x) = c_1 \cos(vx) + c_2 \sin(vx)$. But $X(0) = c_1(1) + c_2(0) = 0$, so $c_1 = 0$. Thus, $X(x) = c_2 \sin(vx)$. Also $X(L) = c_2 \sin(vL) = 0$. If $c_2 = 0$, we get the trivial solution. If $c_2 \neq 0$, then $\sin(vL) = 0$. Thus, $vL = n\pi$, $n = 1, 2, 3, \ldots$. Therefore,

$$v = \frac{n\pi}{L}, \quad k = \left[\frac{n\pi}{L}\right]^2, \quad n = 1, 2, 3, \ldots$$

and

$$X = X_n = c_2 \sin\left(\frac{n\pi x}{L}\right)$$

Now

$$T' + c^2 kT = 0 \Rightarrow T' + \left[\frac{cn\pi}{L}\right]^2 T = 0 \Rightarrow T = ce^{-\left[\frac{cn\pi}{L}\right]^2 t}$$

Thus,

$$u = u_n = X_n T_n = b_n \sin\left(\frac{n\pi x}{L}\right) e^{-\left[\frac{cn\pi}{L}\right]^2 t}$$

We keep all solutions by

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) e^{-\left[\frac{cn\pi}{L}\right]^2 t}$$

By the initial condition $u(x,0) = f(x)$. Hence

$$u(x,0) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) = f(x)$$

Therefore, b_n is the Fourier sine series coefficient of $f(x)$. Thus,

$$b_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx$$

Hence,

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{L}\right) e^{-\left[\frac{cn\pi}{L}\right]^2 t} \tag{4.18}$$

$$b_n = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{n\pi x}{L}\right)dx$$

Example 4.9.1. Solve the heat flow problem, by finding $u(x,t)$.

$$\frac{\partial u}{\partial t} = 3\frac{\partial^2 u}{\partial x^2}$$
$$u(0,t) = u(1,t) = 0$$
$$u(x,0) = f(x) = x(1-x)$$

Then approximate the temperature $u(x,t)$ at $x = 0.25$, at time $t = 1$ second, by using two nonzero terms.

Solution: We need the coefficients. Using equation (4.18) we have (check!)

$$b_n = \frac{2}{1}\int_0^1 x(1-x)\sin(n\pi x)\,dx = 4\frac{1-(-1)^n}{n^3\pi^3}$$

Therefore,

$$u(x,t) = \frac{4}{\pi^3}\sum_{n=1}^{\infty}\frac{1-(-1)^n}{n^3}\sin(n\pi x)\,e^{-3n^2\pi^2 t}$$

(b) $u(0.25,1)$ is approximately

$$\begin{aligned} u(0.25,1) &\simeq \frac{4}{\pi^3}\frac{1-(-1)^1}{1^3}\sin\left(\frac{1\pi}{2}\right)e^{-3\pi^2} + \frac{4}{\pi^3}\frac{1-(-1)^2}{2^3}\sin\left(\frac{2\pi}{2}\right)e^{-3\left(2^2\right)\pi^2} + \\ &+ \frac{4}{\pi^3}\frac{1-(-1)^3}{3^3}\sin\left(\frac{3\pi}{2}\right)e^{-3\left(3^2\right)\pi^2} \\ &= \frac{4}{\pi^3}\left(2e^{-3\pi^2} - \frac{2}{27}e^{-27\pi^2}\right) = 3.5702\times 10^{-14} \text{ degrees} \end{aligned}$$

We had to use three terms because the middle term was zero.

One Dimensional Heat Equation: Insulated Ends

Next we consider the case where the ends of the rod are insulated. So there is no heat flow through the ends. Mathematically, this is expressed by the conditions $\frac{\partial u}{\partial x}(0,t)=0$ and $\frac{\partial u}{\partial x}(L,t)=0$. We now have the following IBVP.

$$\begin{aligned}
\frac{\partial u}{\partial t} &= c^2\frac{\partial^2 u}{\partial x^2} && \text{The 1-Dimensional Heat Equation}\\
\frac{\partial u}{\partial x}(0,t) &= 0, \quad t\geq 0 && \text{Boundary Condition (insulated left end)}\\
\frac{\partial u}{\partial x}(L,t) &= 0, \quad t\geq 0 && \text{Boundary Condition (insulated right end)}\\
u(x,0) &= f(x), \qquad 0\leq x\leq L && \text{Initial Condition (initial temperature)}
\end{aligned}$$

To solve this IBVP we use just as before, separation of variables.

We start with

$$\begin{aligned}
\frac{\partial u}{\partial t} &= c^2\frac{\partial^2 u}{\partial x^2}\\
\frac{\partial u}{\partial x}(0,t) &= \frac{\partial u}{\partial x}(L,t)=0\\
u(x,0) &= f(x)
\end{aligned}$$

and seek solutions of the form $u=X(x)T(t)=XT$. Then the PDE becomes

$$XT'=c^2X''T \Rightarrow \frac{T'}{c^2T}=\frac{X''}{X}=-k$$

Thus

$$X''+kX=0, \qquad T'+c^2kT=0$$

The boundary conditions become $\frac{\partial u}{\partial x}(0,t)=X'(0)T(t)=0$ and $\frac{\partial u}{\partial x}(L,t)=X'(L)T(t)=0$. Hence, $X'(0)=0$ and $X'(L)=0$, since $T(t)$ cannot be zero for all t, or else we would get the trivial solution. We start with X.

$$X''+kX=0, \quad X'(0)=0, \quad X'(L)=0$$

Case 1: $k<0$, say $k=-v^2$. Then $X(x)=c_1e^{vx}+c_2e^{-vx}$. So, $X'(x)=c_1ve^{vx}-c_2ve^{-vx}$. But $X'(0)=c_1-c_2=0$, so $c_2=c_1$. And $X'(L)=c_1ve^{vL}-c_1ve^{-vL}=0$. Hence, $X'(L)=c_1v\left(e^{vL}-e^{-vL}\right)=0$. Since the last

two factors are not 0, we have $c_1 = 0$. But then $c_2 = 0$ and we get the trivial solution.

Case 2: $k = 0$. Then $X(x) = c_1 x + c_2$. So, $X'(x) = c_1$. But, $X'(0) = c_1 = 0$. So $X = c_2$. And $X'(L) = 0$ poses no restriction on c_2, so we keep $X = c_2$. In this case the equation for T becomes $T' = 0$. Thus, $T = T_0$ a constant. So, the solution u is the constant $u = c_2 T_0$. Let's call this constant $\frac{a_0}{2}$.

$$u = \frac{a_0}{2}$$

Case 3: $k > 0$, say $k = v^2$. Then $X(x) = c_1 \cos(vx) + c_2 \sin(vx)$. So, $X'(x) = -c_1 v \sin(vx) + c_2 v \cos(vx)$. But $X'(0) = -c_1 v(0) + c_2 v(1) = 0$, so $c_2 = 0$. Thus, $X(x) = c_1 \cos(vx)$. Also $X'(L) = -c_1 v \sin(vL) = 0$. If $c_1 = 0$, we get the trivial solution. If $c_1 \neq 0$, then $\sin(vL) = 0$. Thus, $vL = n\pi$, $n = 1, 2, 3, \ldots$. Therefore,

$$v = \frac{n\pi}{L}, \quad k = \left[\frac{n\pi}{L}\right]^2, \quad n = 1, 2, 3, \ldots$$

and

$$X = X_n = c_1 \cos\left(\frac{n\pi x}{L}\right)$$

Now in this case

$$T' + c^2 k T = 0 \Rightarrow T' + \left[\frac{cn\pi}{L}\right]^2 T = 0 \Rightarrow T = ce^{-\left[\frac{cn\pi}{L}\right]^2 t}$$

Thus,

$$u = u_n = X_n T_n = a_n \cos\left(\frac{n\pi x}{L}\right) e^{-\left[\frac{cn\pi}{L}\right]^2 t}$$

We keep all solutions from cases 2 and 3 by

$$u(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) e^{-\left[\frac{cn\pi}{L}\right]^2 t}$$

By the initial condition $u(x,0) = f(x)$. Hence

$$u(x,0) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) = f(x)$$

Therefore, b_n is the Fourier cosine series coefficient of $f(x)$. Thus,

$$a_0 = \frac{2}{L}\int_0^L f(x)\,dx$$
$$a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)dx$$

Hence, the complete solution to the insulated ends heat flow problem is

$$u(x,t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi x}{L}\right) e^{-\left[\frac{cn\pi}{L}\right]^2 t} \tag{4.19}$$
$$a_0 = \frac{2}{L}\int_0^L f(x)\,dx$$
$$a_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{n\pi x}{L}\right)dx, n = 1, 2, \ldots$$

Example 4.9.2. Solve the heat flow problem, by finding $u(x,t)$.

$$\frac{\partial u}{\partial t} = 4\frac{\partial^2 u}{\partial x^2}$$
$$\frac{\partial u}{\partial x}(0,t) = \frac{\partial u}{\partial x}(2,t) = 0$$
$$u(x,0) = f(x) = \begin{cases} 3 & \text{if} \quad 0 \le x \le 1 \\ 0 & \text{if} \quad 1 \le x \le 2 \end{cases}$$

Solution: The coefficients are computed by

$$a_0 = \frac{2}{2}\int_0^2 f(x)\,dx = \int_0^1 3dx + \int_1^2 0dx = 3$$

and

$$a_n = \frac{2}{2}\int_0^2 f(x)\cos\left(\frac{n\pi x}{2}\right)dx = \int_0^1 3\cos\left(\frac{n\pi x}{2}\right)dx + \int_1^2 0\cos\left(\frac{n\pi x}{2}\right)dx$$
$$= \frac{6}{n\pi}\sin\left(\frac{n\pi}{2}\right)$$

Thus,

$$u(x,t) = \frac{3}{2} + \frac{6}{\pi}\sum_{n=1}^{\infty}\frac{1}{n}\sin\left(\frac{n\pi}{2}\right)\cos\left(\frac{n\pi x}{2}\right)e^{-n^2\pi^2 t}$$

Superposition Example

Example 4.9.3 (Superposition)**.** Find the temperature $u(x,t)$ of a zero ends heated rod of length $L = 2$ and $c^2 = 4$, and initial temperature $u(x,0) = 5\sin(2\pi x) - \frac{1}{2}\sin(3\pi x)$.

Solution: Because $c = 2$ and $L = 2$, we have

$$u(x,t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{2}\right) e^{-n^2\pi^2 t}$$

It remains to calculate the b_n. We have

$$u(x,0) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{2}\right) = 5\sin(2\pi x) - \frac{1}{2}\sin(3\pi x)$$

Just as in Example 4.8.1, since $\sin(2\pi x)$ and $\sin(3\pi x)$ appear on the left-hand side as part of the orthogonal set with $n = 4$ and $n = 6$, we may use **superposition** and compare coefficients:

$$b_4 = 5, \qquad b_6 = -\frac{1}{2}, \qquad b_{\text{rest}} = 0$$

Therefore,

$$u(x,t) = 5\sin(2\pi x)\, e^{-16\pi^2 t} - \frac{1}{2}\sin(3\pi x)\, e^{-36\pi^2 t}$$

Of course, we could have used integration

$$\begin{aligned} b_n &= \frac{2}{2}\int_0^2 \left(5\sin(2\pi x) - \frac{1}{2}\sin(3\pi x)\right)\sin\left(\frac{n\pi x}{2}\right) dx \\ &= 5\int_0^2 \sin(2\pi x)\sin\left(\frac{n\pi x}{2}\right) dx - \frac{1}{2}\int_0^2 \sin(3\pi x)\sin\left(\frac{n\pi x}{2}\right) dx \end{aligned}$$

but we would have to exercise **caution**. If $n \neq 4$, then the first integral is zero, as seen by using the appropriate trigonometric identities. However, if $n = 4$, then the first integral becomes

$$\int_0^2 \sin^2(2\pi x)\, dx = \frac{1}{2}\int_0^2 (1 - \cos(2\pi x))\, dx = \frac{1}{2}\left(x - \frac{\sin(2\pi x)}{2\pi}\right)\Bigg|_0^2 = 1$$

So, the answer for the first term is 5. Likewise, for the second integral, if $n \neq 6$ we get zero, but if $n = 6$, we get 1, so the second term is $-\frac{1}{2}$ as asserted.

4.10 Steady State Two Dimensional Heat Equation

The **two-dimensional heat equation** is given by

$$\frac{\partial u}{\partial t} = c^2 \nabla^2 u = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

The heat is called **steady-state**, if the temperature is independent of time and it depends only on the location (x, y). Hence, in this case $\frac{\partial u}{\partial t} = 0$. So the heat equation becomes

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

This PDE is very important and it is called **Laplace's equation**.

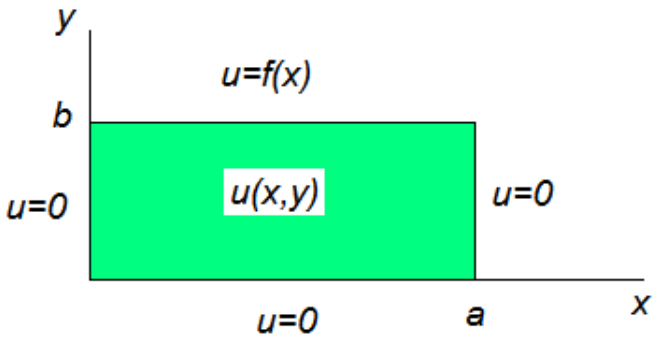
We consider conditions on the boundary C of some region R, where u must be computed.

- If u is prescribed on C, then we have a **Dirichlet problem**.
- If the normal derivative $u_n = \frac{\partial u}{\partial n}$ is prescribed on C, then we have a **Neumann problem**.
- If u is prescribed on part of C, and u_n is prescribed on the remaining part of C, then we have a **Mixed problem**.

In this course we look only at Dirichlet problems.

Steady-State Dirichlet Problem on a Rectangle

We want to find $u(x, y)$, the steady-state temperature over a rectangle R with boundary temperatures as shown, where $f(x)$ is a given function.



We need to solve the boundary value problem consisting of one PDE, Laplace's equation, and four boundary conditions.

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}&=0\\ u(0,y)&=0\\ u(a,y)&=0\\ u(x,0)&=0\\ u(x,b)&=f(x)\end{aligned}$$

where, $0\leq x\leq a$ and $0\leq y\leq b$.

To find $u(x,y)$, we use separation of variables. Let $u(x,y)=X(x)Y(y)=XY$. Substitution into the PDE yields $X''Y+XY''=0$. Hence,

$$\frac{X''}{X}=-\frac{Y''}{Y}=-k$$

Therefore,

$$X''+kX=0,\qquad Y''-kY=0$$

The first two boundary conditions imply

$$u(0,y)=X(0)Y(y)=0,\quad u(a,y)=X(a)Y(y)=0$$

for all $0\leq y\leq b$. Hence, $X(0)=0$ and $X(a)=0$. Therefore,

$$\begin{aligned}X''+kX&=0\\ X(0)&=0\\ X(a)&=0\end{aligned}$$

This S-L problem has been solved before: We get

$$k = k_n = \left(\frac{n\pi}{a}\right)^2, n = 1, 2, \ldots$$

and

$$X_n(x) = \sin\left(\frac{n\pi x}{a}\right)$$

up to constant coefficients. Now,

$$Y'' - \left(\frac{n\pi}{a}\right)^2 Y = 0$$

Its auxiliary qudratic $r^2 - \left(\frac{n\pi}{a}\right)^2 = 0$ yields $r = \pm\left(\frac{n\pi}{a}\right)$. Therefore,

$$Y_n(y) = a_n e^{\frac{n\pi}{a}y} + b_n e^{-\frac{n\pi}{a}y}$$

But the third condition yields, $u(x,0) = X(x)Y(0) = 0, \quad 0 \leq x \leq a$. Hence, $Y(0) = 0$. So, $Y(0) = a_n e^0 + b_n e^0$. Thus, $a_n = -b_n$. Therefore,

$$Y_n(x) = a_n\left(e^{\frac{n\pi}{a}y} - e^{-\frac{n\pi}{a}y}\right)$$

This can be simplified by using the hyperbolic sign identity $\sinh(a) = \frac{1}{2}(e^a - e^a)$ to get

$$Y_n(x) = c_n \sinh\left(\frac{n\pi y}{a}\right)$$

where $c_n = 2a_n$. Hence,

$$u_n(x,y) = X_n Y_n = c_n \sin\left(\frac{n\pi x}{a}\right)\sinh\left(\frac{n\pi y}{a}\right)$$

Now, we keep all solutions by using the infinite series

$$u(x,y) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi x}{a}\right)\sinh\left(\frac{n\pi y}{a}\right) \tag{4.20}$$

Lastly, we use the fourth boundary condition to compute the c_n.

$$u(x,b) = \sum_{n=1}^{\infty}\left[c_n \sinh\left(\frac{n\pi b}{a}\right)\right]\sin\left(\frac{n\pi x}{a}\right) = f(x)$$

This is the Fourier sine series of $f(x)$ with Fourier coefficeints $c_n \sinh\left(\frac{n\pi b}{a}\right)$. Hence,

$$c_n \sinh\left(\frac{n\pi b}{a}\right) = \frac{2}{a}\int_0^a f(x)\sin\left(\frac{n\pi x}{a}\right)dx$$

Therefore,

$$c_n = \frac{2}{a\sinh\left(\frac{n\pi b}{a}\right)}\int_0^a f(x)\sin\left(\frac{n\pi x}{a}\right)dx \tag{4.21}$$

Equations (4.20) and (4.21) completely solve the 2-dimensional steady-state heat flow in question.

4.11 Two Dimensional Wave Equation (Rectangular Membrane)

The general wave equation for a function $u(x_1, x_2, \ldots, x_n, t)$ is

$$\frac{\partial^2 u}{\partial t^2} = c^2\nabla^2 u = c^2\left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}\right)$$

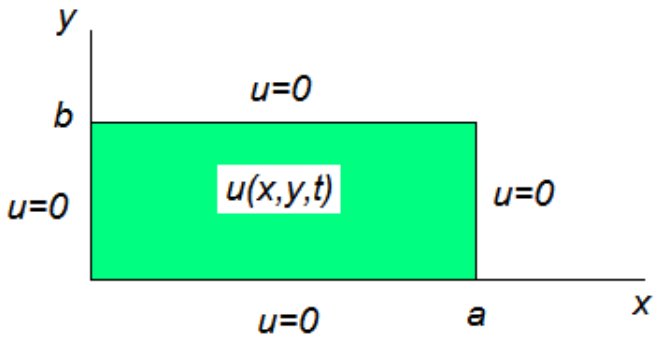
In this section we study the **two-dimensional wave equation**

$$\frac{\partial^2 u}{\partial t^2} = c^2\nabla^2 u = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

The function $u(x, y, t)$ models the displacement or deflection of an elastic membrane occupying a 2-dimensional region R. In our case the region R is a rectangle.

Rectangular Membrane with Fixed Ends

Let us assume that the membrane is attached to the boundary of R. So, $u = 0$ on the boundary.



Also, it is assumed that the initial deflection is a known function $f(x,y)$ and the initial veliocity is a known function $g(x,y)$. The goal is to find $u(x,y,t)$ for all $t \geq 0$, $0 \leq x \leq a, 0 \leq y \leq b$. We have the IBVP consisting of the PDE four boundary conditions and two initial conditions.

$$\begin{aligned}
\frac{\partial^2 u}{\partial t^2} &= c^2\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)\\
u(0,y,t) &= 0, \quad 0 \leq y \leq b, \quad t \geq 0\\
u(a,y,t) &= 0, \quad 0 \leq y \leq b, \quad t \geq 0\\
u(x,0,t) &= 0, \quad 0 \leq x \leq a, \quad t \geq 0\\
u(x,b,t) &= 0, \quad 0 \leq x \leq a, \quad t \geq 0\\
u(x,y,0) &= f(x,y), \quad 0 \leq x \leq a, \quad 0 \leq y \leq b\\
\frac{\partial u}{\partial t}(x,y,0) &= g(x,y), \quad 0 \leq x \leq a, \quad 0 \leq y \leq b
\end{aligned}$$

We solve this problem as usual by separation of varialbes. Let $u(x,y,t) = X(x)Y(y)T(t) = XYT$. Substitution into the PDE yields $XYT'' = c^2(X''YT + XY''T)$. So,

$$\frac{T''}{c^2T} = \frac{X''}{X} + \frac{Y''}{Y}$$

We see that all the fractions must be constants. Let α and β be such that

$$\frac{X''}{X} = -\alpha, \quad \frac{Y''}{Y} = -\beta$$

Then

$$\frac{T''}{T} = -c^2(\alpha+\beta)$$

So, we get

$$X'' + \alpha X = 0, \quad Y'' + \beta Y = 0, \quad T'' + c^2 (\alpha + \beta) T = 0$$

The boundary conditions imply

$$X(0) = 0, \quad X(a) = 0, \quad Y(0) = 0, \quad Y(b) = 0$$

So we have two familiar S-L problems.

$$\begin{aligned} X'' + \alpha X &= 0 \\ X(0) &= 0 \\ X(a) &= 0 \end{aligned}$$

and

$$\begin{aligned} Y'' + \beta Y &= 0 \\ Y(0) &= 0 \\ Y(b) &= 0 \end{aligned}$$

Hence,

$$\begin{aligned} X_m &= \sin\left(\frac{m\pi x}{a}\right), \qquad \alpha = \left(\frac{m\pi}{a}\right)^2, \quad m = 1, 2, \ldots \\ Y_n &= \sin\left(\frac{n\pi y}{a}\right), \qquad \beta = \left(\frac{n\pi}{b}\right)^2, \quad n = 1, 2, \ldots \end{aligned}$$

So, the equation in T is now

$$T'' + c^2 \left(\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 \right) T = 0$$

Hence,

$$T_{mn}(t) = a_{mn} \cos(\lambda_{mn} t) + b_{mn} \sin(\lambda_{mn} t)$$

where

$$\lambda_{mn} = c\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2}, \qquad m = 1, 2, \ldots \quad n = 1, 2, \ldots$$

Therefore,

$$u_{mn} = (a_{mn} \cos(\lambda_{mn} t) + b_{mn} \sin(\lambda_{mn} t)) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right)$$

are the eigenfunctions or **normal modes**. We put them all together in a double series

$$u(x,y,t)=\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\left(a_{mn}\cos(\lambda_{mn}t)+b_{mn}\sin(\lambda_{mn}t)\right)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) \tag{4.22}$$

$$\lambda_{mn}=c\sqrt{\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2},\qquad m=1,2,\ldots\quad n=1,2,\ldots$$

What emains to compute is the coefficients a_{mn} and b_{mn} by using the intial conditions (the Fourier analysis part). From the first initial condition we have

$$u(x,y,0)=\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}a_{mn}\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)=f(x,y)$$

This is a double Fourier series. Let for each fixed m

$$k_m(y)=\sum_{n=1}^{\infty}a_{mn}\sin\left(\frac{n\pi y}{b}\right)$$

Then for fixed m, a_{mn} is the Fourier sine series coefficient of $k_m(y)$. So

$$a_{mn}=\frac{2}{b}\int_0^b k_m(y)\sin\left(\frac{n\pi y}{b}\right)\,dy \tag{4.23}$$

On the other hand,

$$f(x,y)=\sum_{m=1}^{\infty}k_m(y)\sin\left(\frac{m\pi x}{a}\right) \tag{4.24}$$

So, $k_m(y)$ is the mth Fourier sine series coefficient of $f(x,y)$ for each fixed y. Hence,

$$k_m(y)=\frac{2}{a}\int_0^a f(x,y)\sin\left(\frac{m\pi x}{a}\right)\,dx \tag{4.25}$$

By equations (4.23), (4.24), and (4.25) we get

$$a_{mn}=\frac{4}{ab}\int_0^b\int_0^a f(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)\,dx\,dy \tag{4.26}$$

Next we by using the second initial condition we get

$$\frac{\partial u}{\partial t}(x,y,0)=\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\lambda_{mn}b_{mn}\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)=g(x,y)$$

This time we find $\lambda_{mn}b_{mn}$ exactly as we found a_{mn} and then we divide by λ_{mn} to find b_{mn}. We get

$$b_{mn}=\frac{4}{\lambda_{mn}ab}\int_0^b\int_0^a g(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)\,dx\,dy \tag{4.27}$$

Equations (4.22), (4.26), and (4.27) completely solve the deflection problem on a rectangle by Fourier series methods.

We have the following complete solution of the two-dimensional wave equation in terms of a double Fourier series in x and y.

$$u(x,y,t)=\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\left(a_{mn}\cos(\lambda_{mn}t)+b_{mn}\sin(\lambda_{mn}t)\right)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) \tag{4.28}$$

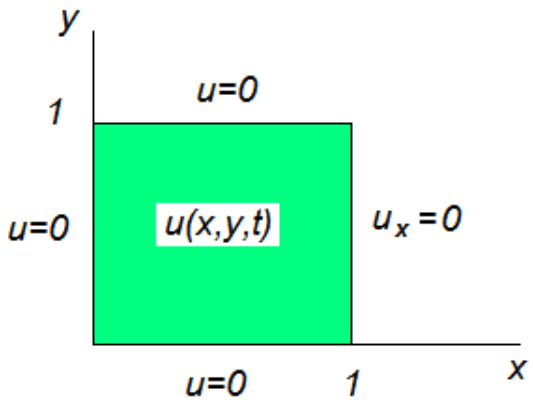
$$\lambda_{mn}=c\sqrt{\left(\frac{m\pi}{a}\right)^2+\left(\frac{n\pi}{b}\right)^2},\qquad m=1,2,\ldots\quad n=1,2,\ldots$$

$$a_{mn}=\frac{4}{ab}\int_0^b\int_0^a f(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)\,dx\,dy \tag{4.29}$$

$$b_{mn}=\frac{4}{\lambda_{mn}ab}\int_0^b\int_0^a g(x,y)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)\,dx\,dy \tag{4.30}$$

Square Membrane with One Loose End

We consider now the case $a=b=1$, $c^2=1$, and the right side of the square is loose. Thus, instead of $u(1,y,t)=0$, we have $\frac{\partial u}{\partial t}(1,y,t)=0$. Furthermore, we assume that there is intitial deflection $f(x,y)$, but that there is no initial velocity.



Again we want to find $u(x,y,t)$ for all $t \geq 0$, $0 \leq x \leq 1, 0 \leq y \leq 1$. We have the IBVP consisting of the PDE four boundary conditions and two initial conditions.

$$\begin{aligned}
\frac{\partial^2 u}{\partial t^2} &= \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \\
u(0,y,t) &= 0, \quad 0 \leq y \leq 1, \quad t \geq 0 \\
\frac{\partial u}{\partial t}(1,y,t) &= 0, \quad 0 \leq y \leq 1, \quad t \geq 0 \\
u(x,0,t) &= 0, \quad 0 \leq x \leq 1, \quad t \geq 0 \\
u(x,1,t) &= 0, \quad 0 \leq x \leq 1, \quad t \geq 0 \\
u(x,y,0) &= f(x,y), \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1 \\
\frac{\partial u}{\partial t}(x,y,0) &= 0, \quad 0 \leq x \leq 1, \quad 0 \leq y \leq 1
\end{aligned}$$

We solve this problem as usual by separation of varialbes. Let $u(x,y,t) = X(x)Y(y)T(t) = XYT$. Substitution into the PDE yields $XYT'' = X''YT + XY''T$. So,

$$\frac{T''}{T} = \frac{X''}{X} + \frac{Y''}{Y}$$

We see that all the fractions must be constants. Let α and β be such that

$$\frac{X''}{X} = -\alpha, \quad \frac{Y''}{Y} = -\beta$$

Then

$$\frac{T''}{T} = -(\alpha + \beta)$$

So, we get

$$X'' + \alpha X = 0, \quad Y'' + \beta Y = 0, \quad T'' + (\alpha + \beta) T = 0$$

The boundary conditions imply

$$X(0) = 0, \quad X(a) = 0, \quad Y(0) = 0, \quad Y(b) = 0$$

So we have two familiar S-L problems.

$$\begin{aligned} X'' + \alpha X &= 0 \\ X(0) &= 0 \\ X'(1) &= 0 \end{aligned}$$

and

$$\begin{aligned} Y'' + \beta Y &= 0 \\ Y(0) &= 0 \\ Y(1) &= 0 \end{aligned}$$

The system in Y has been solved several times before, we get (with $L = 1$)

$$Y_n = \sin(n\pi y), \qquad \beta = (n\pi)^2, \quad n = 1, 2, \ldots$$

The system in X has a derivative in the second boundary condition.

To solve we have

$$X'' + \alpha X = 0, \quad X(0) = 0, \quad X'(1) = 0$$

Case 1: $\alpha < 0$, say $\alpha = -v^2$. Then $X(x) = c_1 e^{vx} + c_2 e^{-vx}$. So, $X'(x) = c_1 v e^{vx} - c_2 v e^{-vx}$. But $X(0) = c_1 + c_2 = 0$, so $c_2 = -c_1$. And $X'(1) = c_1 v e^{v} - c_2 v e^{-v} = 0$. Hence, $X'(1) = c_1 v (e^{v} + e^{-v}) = 0$. Thus,$c_1 = 0$ and so $c_2 = 0$. So we get the trivial solution.

Case 2: $\alpha = 0$. Then $X(x) = c_1 x + c_2$. So, $X'(x) = c_1$. But, $X(0) = c_2 = 0$. So $X = c_1 x$. And $X'(1) = c_1 = 0$.. So, again, we get the trivial solution.

Case 3: $\alpha > 0$, say $\alpha = v^2$. Then $X(x) = c_1 \cos(vx) + c_2 \sin(vx)$. So, $X'(x) = -c_1 v \sin(vx) + c_2 v \cos(vx)$. But $X(0) = c_1 = 0$. Thus, $X(x) =$

$c_2 \sin(vx)$. Also $X'(1) = c_2 v \cos(v) = 0$. If $c_2 = 0$, we get the trivial solution. If $c_2 \neq 0$, then $\cos(v) = 0$. Thus, $v = (2m-1)\frac{\pi}{2}$, $m = 1, 2, 3, \ldots$. Therefore,

$$v = \frac{(2m-1)\pi}{2}, \quad \alpha = \left[\frac{(2m-1)\pi}{2}\right]^2, \quad m = 1, 2, 3, \ldots$$

and up to coefficient

$$X = X_m = \sin\left(\frac{(2m-1)\pi x}{2}\right)$$

So, the equation in T is now

$$T'' + \left[\left(\frac{(2m-1)\pi}{2}\right)^2 + (n\pi)^2\right] T = 0$$

We solve for T to get,

$$T_{mn}(t) = a_{mn}\cos(\lambda_{mn} t) + b_{mn}\sin(\lambda_{mn} t)$$

where

$$\lambda_{mn} = \pi\sqrt{\frac{(2m-1)^2}{4} + n^2}, \qquad m = 1, 2, \ldots \quad n = 1, 2, \ldots$$

Therefore,

$$u_{mn} = (a_{mn}\cos(\lambda_{mn} t) + b_{mn}\sin(\lambda_{mn} t))\sin\left(\frac{(2m-1)\pi x}{2}\right)\sin(n\pi y)$$

are the eigenfunctions or **normal modes**. We put them all together in a double series

$$u(x, y, t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\left[a_{mn}\cos(\lambda_{mn} t) + b_{mn}\sin(\lambda_{mn} t)\right]\sin\left(\frac{(2m-1)\pi x}{2}\right)\sin(n\pi y)$$

$$\lambda_{mn} = \pi\sqrt{\frac{(2m-1)^2}{4} + n^2}, \qquad m = 1, 2, \ldots \quad n = 1, 2, \ldots$$

Because the initial velocity is zero we have $b_{mn} = 0$. So,

$$u(x,y,t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{mn}\cos(\lambda_{mn}t)\sin\left(\frac{(2m-1)\pi x}{2}\right)\sin(n\pi y) \tag{4.31}$$

$$\lambda_{mn} = \pi\sqrt{\frac{(2m-1)^2}{4}+n^2}, \qquad m = 1,2,\ldots \qquad n = 1,2,\ldots$$

We use the first initial condition to compute a_{mn}.

$$u(x,y,0) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{mn}\sin\left(\frac{(2m-1)\pi x}{2}\right)\sin(n\pi y) = f(x,y)$$

Using double Fourier series just as before we get

$$a_{mn} = 4\int_0^1\int_0^1 f(x,y)\sin\left(\frac{(2m-1)\pi x}{2}\right)\sin(n\pi y)\ dx\ dy \tag{4.32}$$

Equations (4.31) and (4.32) completely solve this deflection problem on a square by Fourier series methods.

4.12 The Cauchy-Euler Equation

Let a, b, c be constants. The second order differential equation

$$ax^2y'' + bxy' + cy = 0 \tag{4.33}$$

is called a Cauchy-Euler equation in the unknown function $y = y(x)$ of x. Although this ODE has variable coefficients, it is easy to solve.

We seek solutions of the form $y = x^v$. Then $y' = vx^{v-1}$, $y'' = v(v-1)x^{v-2}$. Substitution into the ODE yields

$$ax^2v(v-1)x^{v-2} + bxvx^{v-1} + cx^v = 0$$

or

$$(av(v-1) + bv + c)x^v = 0$$

Therefore,

$$av(v-1) + bv + c = 0$$

or

$$av^2 + (b-a)\,v + c = 0 \tag{4.34}$$

This is the corresponding auxiliary equation. It is a quadratic in v. We solve it to get in general two solutions , say, v_1 and v_2. If these solutions are real and distinct we see that x^{v_1} and x^{v_2} are not scalar multiples of each other, thus, they are linearly independent. So

$$y(x) = c_1 x^{v_1} + c_2 x^{v_2} \tag{4.35}$$

is the general solution of the Cauchy-Euler equation.

Example 4.12.1. Find the general solution of

$$2x^2 y'' + 3xy' - y = 0$$

Solution: This is a Cauchy-Euler equation. The auxiliary equation is computed by using (4.34). We have

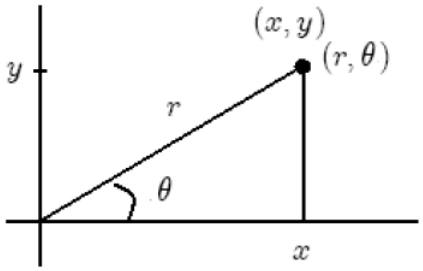
$$2x^2 + x - 1 = (x+1)(2x-1) = 0$$

Thus, $v_1 = -1$ and $v_2 = 1/2$. Hence, the general solution is

$$y(x) = \frac{c_1}{x} + c_2\sqrt{x}$$

4.13 Laplacian in Polar Coordinates

Some boundary value problems that involve circular types of regions can be simplified, if polar coordinates are used.



Recall that

$$x = r\cos(\theta), \quad y = \sin(\theta) \tag{4.36}$$
$$r^2 = x^2 + y^2, \quad \tan(\theta) = \frac{y}{x}$$

A function $u(x, y)$ in Cartesian coordinates may be also considered as a function $u(r, \theta)$ in polar coordinates.

In this section we write the Laplacian of a function in polar coordinates and solve Laplace's equation

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{4.37}$$

by using polar coordinates.

Laplacian in Polar Coordinates

Let $u(x, y)$ be a function in Cartesian coordinates and $u(r, \theta)$ be the same function in terms of polar coordinates. We have the following theorem.

Theorem 4.13.1. *The Laplacian in polar coordinates is*

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} \tag{4.38}$$

Proof: (Optional) Chain Rule in vector-matrix form yields

$$\begin{bmatrix} u_r \\ u_\theta \end{bmatrix} = \begin{bmatrix} x_r & y_r \\ x_\theta & y_\theta \end{bmatrix} \begin{bmatrix} u_x \\ u_y \end{bmatrix} = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -r\sin(\theta) & r\cos(\theta) \end{bmatrix} \begin{bmatrix} u_x \\ u_y \end{bmatrix}$$

Hence, by matrix inversion

$$\begin{bmatrix} u_x \\ u_y \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -r^{-1}\sin(\theta) \\ \sin(\theta) & r^{-1}\cos(\theta) \end{bmatrix} \begin{bmatrix} u_r \\ u_\theta \end{bmatrix}$$

The first row as an operator is

$$\frac{\partial}{\partial x} = \cos(\theta)\frac{\partial}{\partial r} - \frac{1}{r}\sin(\theta)\frac{\partial}{\partial \theta}$$

Applying this operator twice to u yields

$$\begin{aligned}
u_{xx} &= \cos(\theta)\frac{\partial}{\partial r}\left\{\cos(\theta)u_r - \frac{1}{r}\sin(\theta)u_\theta\right\} - \frac{1}{r}\sin(\theta)\frac{\partial}{\partial\theta}\left\{\cos(\theta)u_r - \frac{1}{r}\sin(\theta)u_\theta\right\} \\
&= \cos(\theta)\left(\cos(\theta)u_{rr} - \frac{1}{r}\sin(\theta)u_{r\theta} - \frac{1}{r^2}\sin(\theta)u_\theta\right) \\
&- \frac{1}{r}\sin(\theta)\left(-\sin(\theta)u_r + \cos(\theta)u_{r\theta} - \frac{1}{r}\cos(\theta)u_\theta - \frac{1}{r}\sin(\theta)u_{\theta\theta}\right) \\
&= \cos^2(\theta)u_{rr} + \frac{1}{r}\sin^2(\theta)u_r - \frac{2}{r}\sin(\theta)\cos(\theta)u_{r\theta} \\
&+ \frac{2}{r^2}\sin(\theta)\cos(\theta)u_\theta + \frac{1}{r^2}\sin^2(\theta)u_{\theta\theta}
\end{aligned}$$

The second row as an operator is

$$\frac{\partial}{\partial y} = \sin(\theta)\frac{\partial}{\partial r} + \frac{1}{r}\cos(\theta)\frac{\partial}{\partial\theta}$$

Applying this operator twice to u yields

$$\begin{aligned}
u_{xx} &= \sin(\theta)\frac{\partial}{\partial r}\left\{\sin(\theta)u_r + \frac{1}{r}\cos(\theta)u_\theta\right\} + \frac{1}{r}\cos(\theta)\frac{\partial}{\partial\theta}\left\{\sin(\theta)u_r + \frac{1}{r}\cos(\theta)u_\theta\right\} \\
&= \sin(\theta)\left(\sin(\theta)u_{rr} + \frac{1}{r}\cos(\theta)u_{r\theta} - \frac{1}{r^2}\cos(\theta)u_\theta\right) \\
&+ \frac{1}{r}\cos(\theta)\left(\cos(\theta)u_r + \sin(\theta)u_{r\theta} - \frac{1}{r}\sin(\theta)u_\theta + \frac{1}{r}\cos(\theta)u_{\theta\theta}\right) \\
&= \sin^2(\theta)u_{rr} + \frac{1}{r}\cos^2(\theta)u_r + \frac{2}{r}\sin(\theta)\cos(\theta)u_{r\theta} \\
&- \frac{2}{r^2}\sin(\theta)\cos(\theta)u_\theta + \frac{1}{r^2}\cos^2(\theta)u_{\theta\theta}
\end{aligned}$$

We add the two to get

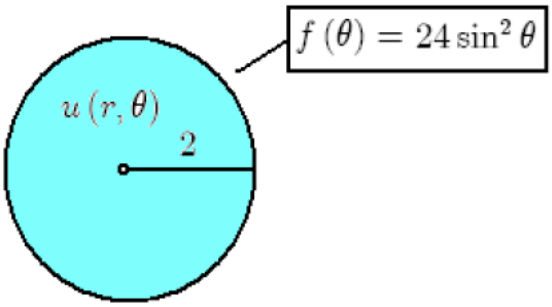
$$\begin{aligned}
u_{xx} + u_{yy} &= \left(\cos^2(\theta) + \sin^2(\theta)\right)u_{rr} + \frac{1}{r}\left(\cos^2(\theta) + \sin^2(\theta)\right)u_r + \frac{1}{r^2}\left(\cos^2(\theta) + \sin^2(\theta)\right)u_{\theta\theta} \\
&= u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}
\end{aligned}$$

Steady-State Temperature in a Disk: Example

We now solve Laplace's equation in the case where $u = u(x, y) = u(r, \theta)$ is a function defined on a disk. The values of u on the boundary are given.

For example, such u may represent the steady-state temperature of a disk with given boundary temperature. To be more concrete, let us work with the following example.

Example 4.13.1. Find the steady-state temperatute $u(r,\theta)$ on a disk of radius 2 with boundary temperature $f(\theta) = 24\sin^2(\theta)$.



Solution: From the heat equation we have

$$\nabla^2 u = c^2 \frac{\partial u}{\partial t} = 0$$

because u is independent of time. Thus, we need to solve the boundary value problem

$$\begin{aligned} \nabla^2 u &= \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} = 0 \\ u(2,\theta) &= 24\sin^2(\theta) \end{aligned}$$

We use separation of variables. We seek solutions of the form

$$u(r,\theta) = R(r)\,\Theta(\theta) = R\Theta$$

Because of the nature of polar coordinates, there are restrictions on the functions $R(r)$ and $\Theta(\theta)$. For example, since the solutions depend on points on the disk and not on their representations in polar coordinates, we want $\Theta(\theta)$ and its derivatives to be periodic with period 2π. In particular, we impose the restrictions

$$\Theta(0) = \Theta(2\pi), \quad \Theta'(0) = \Theta'(2\pi) \tag{4.39}$$

Also, we want the solutions to remain finite near the pole, so that we require that

$$R(r) \text{ is bounded as } r \to 0$$

Substitution $u = R\Theta$ into the PDE yields

$$R''\Theta + \frac{1}{r}R'\Theta + \frac{1}{r^2}R\Theta'' = 0$$

We separate R from Θ to get

$$r^2\frac{R''}{R} + r\frac{R'}{R} = -\frac{\Theta''}{\Theta} = k$$

where k is the common constant of the function in r on the left and the function of θ on the right. Therefore,

$$r^2R'' + rR' - kR = 0 \tag{4.40}$$

and

$$\Theta'' + k\Theta = 0 \tag{4.41}$$

Now Θ must be periodic, therefore, $k \geq 0$, or else we would have exponentials. Let $k = v^2$ for some $v \geq 0$. The equation (4.40) becomes

$$r^2R'' + rR' - v^2R = 0$$

This is a Cauchy-Euler equation. We solve it to get

$$R(r) = c_1r^v + c_2r^{-v}$$

Since, $R(r)$ must remain bounded as $r \to 0$, we must have $c_2 = 0$. Hence,

$$R(r) = c_1r^v, \quad v \geq 0$$

The equation in Θ becomes

$$\Theta'' + v^2\Theta = 0, \quad v \geq 0$$

If $v = 0$, we have $\Theta'' = 0$ and $R = c_1$. Hence, $\Theta(\theta) = a\theta + b$. But then $a = 0$, since Θ is periodic. Thus, Θ is a constant. Therefore, u is a constant, say

$$u = u_0 = \frac{a_0}{2}$$

Now let $v > 0$. Then

$$\Theta(\theta) = k_1 \cos(v\theta) + k_2 \sin(v\theta)$$

Hence,

$$\Theta'(\theta) = -vk_1 \sin(v\theta) + vk_2 \cos(v\theta)$$

Using (4.39) we get

$$\begin{aligned} k_1 &= k_1 \cos(2\pi v) + k_2 \sin(2\pi v) \\ vk_2 &= -vk_1 \sin(2\pi v) + vk_2 \cos(2\pi v) \end{aligned}$$

This yields a 2×2 homogeneous linear system in k_1 and k_2.

$$\begin{aligned} (\cos(2\pi v) - 1) k_1 + \sin(2\pi v) k_2 &= 0 \\ -\sin(2\pi v) k_1 + (\cos(2\pi v) - 1) k_2 &= 0 \end{aligned}$$

This system has nontrivial solutions if and only if the coefficient determinant is zero.

$$\begin{aligned} 0 &= \begin{vmatrix} \cos(2\pi v) - 1 & \sin(2\pi v) \\ -\sin(2\pi v) & \cos(2\pi v) - 1 \end{vmatrix} \\ &= \cos^2(2\pi v) - 2\cos(2\pi v) + \sin^2(2\pi v) + 1 \\ &= 2(1 - \cos(2\pi v)) \end{aligned}$$

Hence, $\cos(2\pi v) = 1$. Therefore, $v = n$, where n is any integer. Thus, for any integer n, $\Theta = \Theta_n$ is a linear combination of $\cos(n\theta)$ and $\sin(n\theta)$. I.e.,

$$\Theta(\theta) = \Theta_n(\theta) = a_n \cos(n\theta) + b_n \sin(n\theta)$$

We may restrict n to positive integers, because $\cos(-n\theta) = \cos(n\theta)$ and $\sin(-n\theta) = \sin(n\theta)$ and the extra sign can be passed to the constant b_n.

Thus, up to constant $R(r) = R_n(r) = r^n$. Therefore, we get

$$u = u_n = (a_n \cos(n\theta) + b_n \sin(n\theta)) r^n$$

Putting all the instances of u together in an infinite series we get

$$u(r, \theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(n\theta) + b_n \sin(n\theta)) r^n \tag{4.42}$$

Next, we take into account the boundary condition.

$$\begin{aligned} u(2,\theta) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(n\theta) + b_n \sin(n\theta)\right) 2^n = 24\sin^2(\theta) \\ &= 12 - 12\cos(2\theta) \end{aligned}$$

In the last equality we used the half-angle formula. By superposition, we get

$$\frac{a_0}{2} = 12, \quad 2^2 a_2 = -12, \quad 2^{\text{rest}} a_{\text{rest}} = 0, \quad 2^n b_n = 0$$

Hence,

$$a_0 = 24, \quad a_2 = -3, \quad a_{\text{rest}} = 0, \quad b_n = 0$$

Therefore,

$$u(r,\theta) = 12 - 3\cos(2\theta) r^2$$

Steady-State Temperature in a Disk: General Case

Right up to equation (4.42) we did not use the fact that the disk was of radius 2 and also we did not use the specific $f(\theta)$ on the boundary. So, equation (4.42) is valid in general. Now assume that our disk has any radius R and that $f(\theta)$ on the boundary is general. So we have

$$u(R,\theta) = f(\theta)$$

Hence,

$$u(R,\theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(n\theta) + b_n \sin(n\theta)\right) R^n = f(\theta)$$

or by disrtibuting R^n we have

$$u(R,\theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n R^n\right) \cos(n\theta) + \left(b_n R^n\right) \sin(n\theta) = f(\theta) \tag{4.43}$$

Now since the functions

$$1, \ \cos(\theta), \ \sin(\theta), \ \cos(2\theta), \ \sin(2\theta), \ \ldots, \cos(n\theta), \ \sin(n\theta), \ \ldots$$

are orthogonal on $[0, 2\pi]$, equation (4.43) represents a classical Fourier series with Fourier coefficients $a_n R^n$ and $b_n R^n$. The Fourier series coefficient were computed by equations (4.7). In our case $L = \pi$. So, we have

$$\begin{aligned} a_0 &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\, d\theta \\ a_n R^n &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\cos(n\theta)\, d\theta, \quad n = 1, 2, \ldots \\ b_n R^n &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\sin(n\theta)\, d\theta, \quad n = 1, 2, \ldots \end{aligned}$$

Therefore, the complete solution to the general case for a disk of radius R and boundary temperature $f(\theta)$ we have

$$\begin{aligned} u(r,\theta) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos(n\theta) + b_n \sin(n\theta)\right) r^n \\ a_0 &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\, d\theta \\ a_n &= \frac{1}{R^n \pi}\int_0^{2\pi} f(\theta)\cos(n\theta)\, d\theta, \quad n = 1, 2, \ldots \\ b_n &= \frac{1}{R^n \pi}\int_0^{2\pi} f(\theta)\sin(n\theta)\, d\theta, \quad n = 1, 2, \ldots \end{aligned}$$

The above set of equations is often written as follows with different notation for the coefficients.

$$\begin{aligned} u(r,\theta) &= \frac{A_0}{2} + \sum_{n=1}^{\infty} \left(A_n \cos(n\theta) + B_n \sin(n\theta)\right) \left(\frac{r}{R}\right)^n \qquad (4.44) \\ A_0 &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\, d\theta \\ A_n &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\cos(n\theta)\, d\theta, \quad n = 1, 2, \ldots \\ B_n &= \frac{1}{\pi}\int_0^{2\pi} f(\theta)\sin(n\theta)\, d\theta, \quad n = 1, 2, \ldots \end{aligned}$$

Here, $A_0 = a_0$, $A_n = a_n R^n$, and $B_n = b_n R^n$.

We urge the reader to practice using formulas (4.44) by reproving the formula $u(r,\theta) = 12 - 3\cos(2\theta)\, r^2$ of Example (4.13.1).

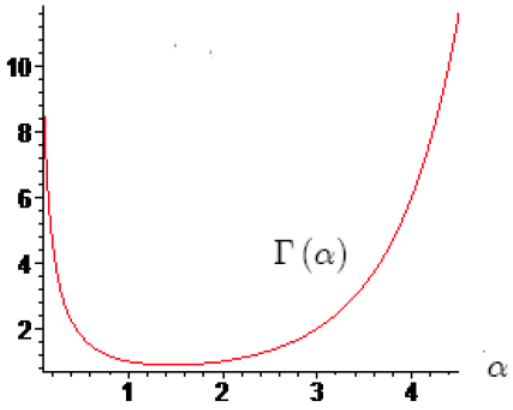
4.14 The Gamma Function

The gamma function, introduced by Euler, is one of the useful functions of mathematics. It is an extension of the factorial function $n!$ over the real numbers and also over the complex numbers. It appears in probability, in statistics, in combinatorics, and in several applications, such as the vibrations of a circular membrane.

For a positive number α we define the gamma function $\Gamma(\alpha)$ by the improper integral

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$$

This integral converges absolutely for $\alpha > 0$, so $\Gamma(\alpha)$ is well-defined.



First we compute $\Gamma(1)$. We have

$$\Gamma(1) = 1$$

This is seen from the computation

$$\Gamma(1) = \int_0^\infty e^{-t} dt = \lim_{r\to\infty} \left(-e^{-t}\right)\Big|_0^r = \lim_{r\to\infty} \left(1 - e^{-r}\right) = 1$$

The most important property of the gamma function is expressed in the following equation.

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{4.45}$$

Equation (4.45) is proved by intergation by parts. Using $u = t^{\alpha}$ and $dv = e^{-t}dt$, we get

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^{\infty} t^{\alpha} e^{-t} dt \\ &= \left(-e^{-t}t^{\alpha}\right)\Big|_0^{\infty} - \int_0^{\infty} \alpha t^{\alpha-1}\left(-e^{-t}\right) dt \\ &= 0 + \alpha \int_0^{\infty} t^{\alpha-1} e^{-t} dt \\ &= \alpha\Gamma(\alpha)\end{aligned}$$

Property (4.45) is used to prove many other properties of the gamma function. For example, we prove that for any positive integer n

$$\Gamma(n+1) = n! \tag{4.46}$$

This shows that the gamma function generalizes the factorial function. Equation (4.46) can be proved by induction and using (4.45) as follows

$$\begin{aligned}\Gamma(1) &= 1 \\ \Gamma(2) &= \Gamma(1+1) = 1\Gamma(1) = 1! \\ \Gamma(3) &= \Gamma(2+1) = 2\Gamma(2) = 2! \\ \Gamma(4) &= \Gamma(3+1) = 3\Gamma(3) = 3! \\ &\vdots \\ \Gamma(n+1) &= n\Gamma(n) = n\left[(n-1)!\right] = n!\end{aligned}$$

We may also compute other values of $\Gamma(\alpha)$, such as $\Gamma(1/2)$. To see this, we have

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{\infty} \frac{e^{-t}}{\sqrt{t}} dt = 2\int_0^{\infty} e^{-u^2} du = 2\frac{\sqrt{\pi}}{2} = \sqrt{\pi}$$

where in the first intergral we changed the variable to $u = \sqrt{t}$ and then used the known and useful fact that $\int_0^{\infty} e^{-x^2} dx = \sqrt{\pi/2}$. [1] The formula for $\Gamma(1/2)$ can used to compute $\Gamma(3/2)$.

$$\Gamma\left(\frac{3}{2}\right) = \Gamma\left(1+\frac{1}{2}\right) = \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = \frac{\sqrt{\pi}}{2}$$

[1] For a proof, we compute $\int_0^{\infty} e^{-t^2} dt \int_0^{\infty} e^{-u^2} du = \int_0^{\infty}\int_0^{\infty} e^{-\left(t^2+u^2\right)} dt\ du = \int_0^{\pi/2}\int_0^{\infty} e^{-r^2} r dr d\theta = \pi/4$, by switching to polar coordinates.

In general, we have

$$\Gamma\left(n+\frac{1}{2}\right)=\frac{1\cdot 3\cdot 5\cdot 7\ldots(2n-1)}{2^n}\sqrt{\pi},\qquad n=1,2,3,\ldots$$

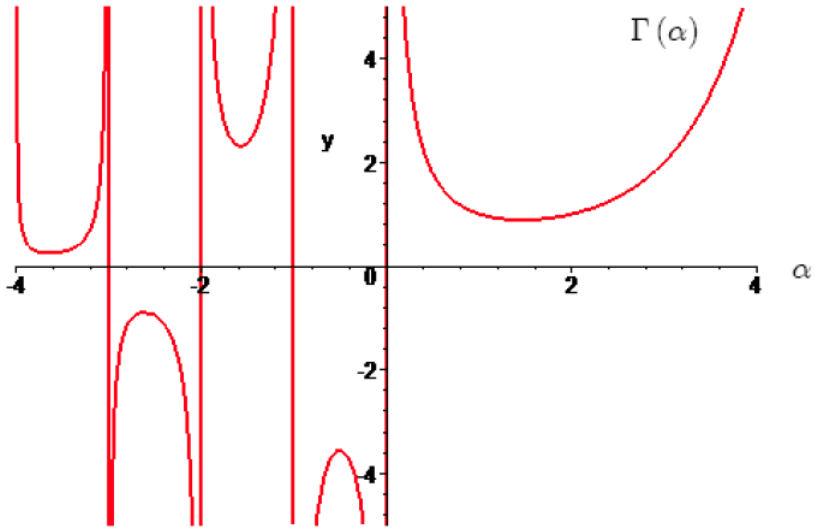
Note The gamma function $\Gamma(\alpha)$ can be extended to negative values of α, provided that α is not a negative integer. This is done by using the following formula for $-1<\alpha<0$.

$$\Gamma(\alpha)=\frac{\Gamma(\alpha+1)}{\alpha}$$

This identity is iterated again to define $\Gamma(\alpha)$ on the negative x-axis except for the negative integers $\alpha=-1,-2,\ldots$. For any α that is not zero or a negative integer we have the formula

$$\Gamma(\alpha)=\frac{\Gamma(\alpha+n)}{\alpha(\alpha+1)\ldots(\alpha+n-1)},\quad \alpha+n>0$$

The following graph shows the extended gamma function over negative values of α.



4.15 Bessel's Equation

Bessel's differential equation is the following second order homogeneous differential equation with variable coefficients

$$x^2y'' + xy' + \left(x^2 - v^2\right) y = 0, \quad v \geq 0 \tag{4.47}$$

where v is a given nonnegative number.

Bessel's equation is important. It appears in electromagnetic theory, in heat conduction, in the vibrations of a circular membrane, in diffusion problems, and elsewhere.

The solutions to Bessel's equations involve new kinds of functions that cannot be written in terms of elementary functions such as polynomials, rational functions, exponentials, logarithms, trigonometric functions, etc.

Bessel Functions of the First Kind: $J_v(x)$

We find power series representations of these functions by using the method of Frobenius.

Let a solution to (4.47) be in the form

$$y(x) = \sum_{m=0}^{\infty} a_m x^{m+r}, \quad a_0 \neq 0$$

where r and a_m are to be determined. Substitution into (4.47) yields

$$\begin{aligned}
&\sum_{m=0}^{\infty} (m+r)(m+r-1) a_m x^{m+r} + \sum_{m=0}^{\infty} (m+r) a_m x^{m+r} \\
&+ \sum_{m=0}^{\infty} a_m x^{m+r+2} - v^2 \sum_{m=0}^{\infty} a_m x^{m+r} \\
&= 0
\end{aligned}$$

Next, we compare the coefficients of both sides of the above equality. The coefficients of x^r yield

$$r(r-1)a_0 + ra_0 - v^2 a_0 = 0 \tag{4.48}$$

The coefficients of x^{r+1} yield

$$r(r+1)a_1 + (r+1)a_1 - v^2 a_1 = 0 \tag{4.49}$$

Likewise, the coefficients of x^{r+j} for $j = 2, 3, \ldots$, yield

$$(j+r)(j+r-1)a_j + (j+r)a_j + a_{j-2} - v^2 a_j = 0 \tag{4.50}$$

Equation (4.48) implies $(r^2 - v^2)a_0 = 0$. Hence, $r = \pm v$, because $a_0 \neq 0$. Let $r = v$. Then (4.49) yields $(2v+1)a_1 = 0$. Thus, $a_1 = 0$, because $2v + 1 > 0$. Now (4.50) yields

$$j(j+2v)a_j + a_{j-2} = 0, \qquad j = 2, 3, \ldots$$

Therefore,

$$a_1 = a_3 = a_5 = \cdots = a_{\text{odd}} = \cdots = 0$$

and for $j = 2k$ even we get the following recursive formula

$$a_{2k} = -\frac{1}{2^2 k(v+k)} a_{2k-2}, \qquad k = 1, 2, \ldots$$

This formula can be applied to any starting value of a_0. The following choice for a_0 is useful in practice.

$$a_0 = \frac{1}{2^v \Gamma(v+1)}$$

Then for $k = 1$, we have

$$a_2 = -\frac{a_0}{2^2(v+1)} = -\frac{1}{2^2(v+1)2^v\Gamma(v+1)} = -\frac{1}{2^{2+v}\Gamma(v+2)}$$

and for $k = 2$, we have

$$a_4 = -\frac{a_2}{2^2 2(v+2)} = \frac{-1}{2^2 2(v+2)} \frac{-1}{2^{2+v}\Gamma(v+2)} = \frac{1}{2^{4+v} 2! \Gamma(v+3)}$$

We continue in the same fashion and by induction we get

$$a_{2k} = \frac{(-1)^k}{2^{2k+v} k! \Gamma(v+k+1)}$$

Therefore, we obtain a particular solution to Bessel's equation in form of a series.

$$J_v(x) = x^v \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k+v} k! \Gamma(v+k+1)} \tag{4.51}$$

The functions $J_v(x)$ are called **Bessel functions of the first kind**. They are well-defined for all real values of x because the defining series converges, as seen by the Ratio Test.

The special case $v = n$ $(n = 0, 1, 2, \dots)$ is of particular interest. In this case we have

$$J_n(x) = x^n \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k+n} k!\,(n+k)!} \tag{4.52}$$

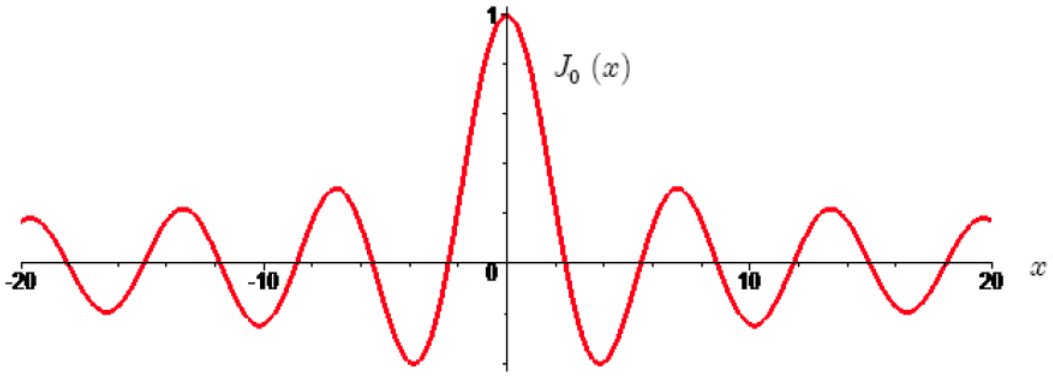
The first thing to notice from the above formula is that

$$J_{2r}(x) \text{ is an even function for } r = 0, 1, 2, \dots$$

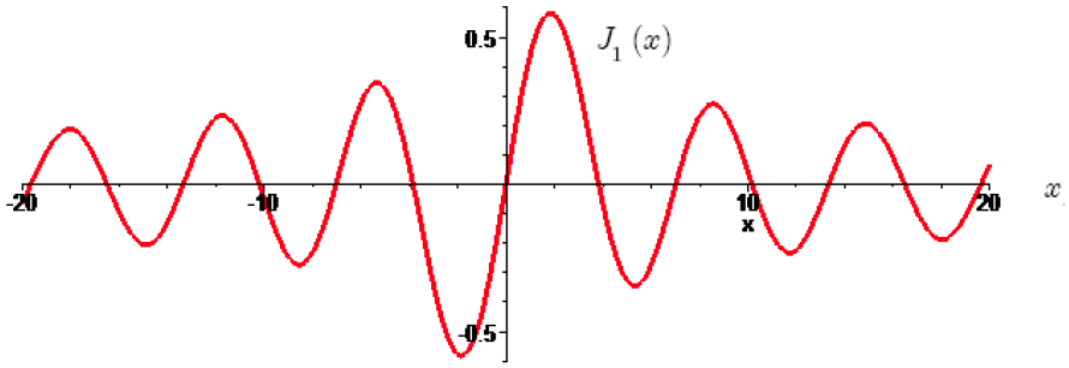
and

$$J_{2r-1}(x) \text{ is an odd function for } r = 1, 2, \dots$$

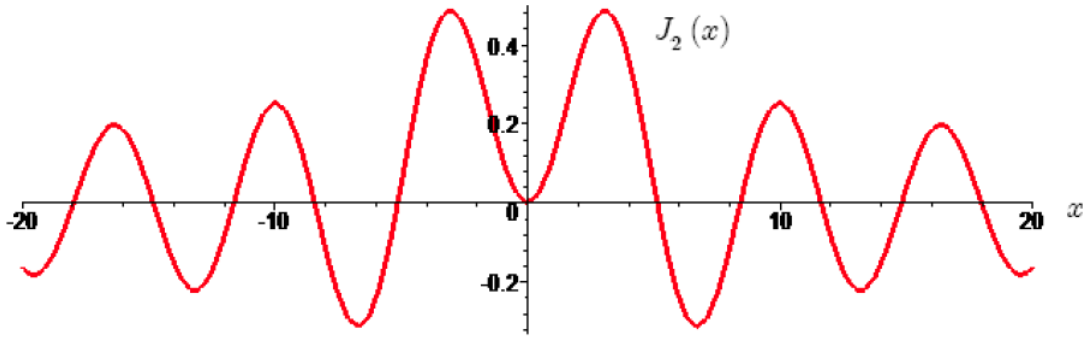
$J_0(x)$ is plotted in the following graph.



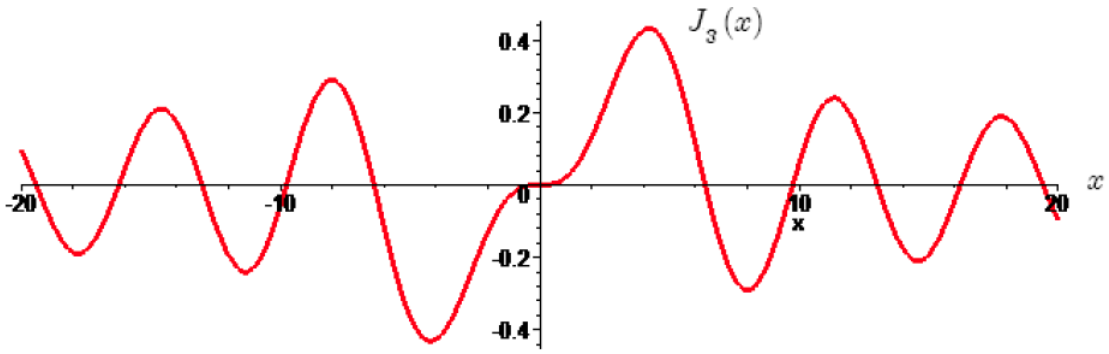
A plot of $J_1(x)$ follows.



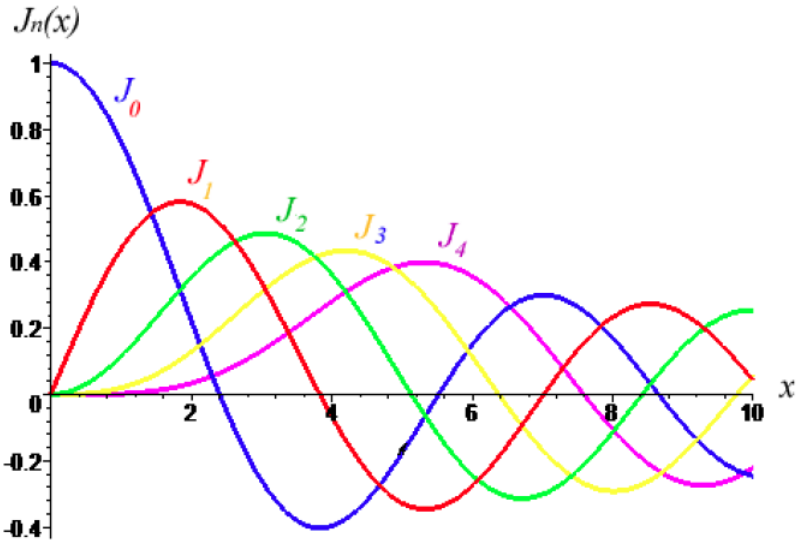
Next, we have a plot of $J_2(x)$



and a plot of $J_3(x)$.



Next, we plot $J_n(x)$ for $n = 0, 1, 2, 3, 4$ for $x > 0$.



Remark It should be noted that

- $J_0(x)$ is the only of $J_n(x)$ with $J_0(0) = 1$. For the remaining $J_n(x)$ we have $J_n(0) = 0$ $(n > 0)$.

- Each $J_n(x)$ has infinitely many zeros. These zeros are not evenly spaced. However, they are asymptotically evenly spaced.
- The zeros of two consecutive $J_n(x)$'s are interlaced.
- The graphs oscillate like sines and cosines, but they decay proportinally to $\sqrt{2/(\pi x)}$ as $x \to \infty$ or as $x \to -\infty$.

One interesting simplification of the $J_v(x)$ occurs when $v = 1/2$. By using the definition and (4.51) we get

$$\begin{aligned} J_{1/2}(x) &= x^{1/2} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k+1/2} k! \Gamma\left(\frac{1}{2} + k + 1\right)} \\ &= \sqrt{\frac{2}{x}} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^{2k+1} k!} \frac{2^{k+1}}{1 \cdot 3 \cdot 5 \ldots (2k+1)\sqrt{\pi}} x^{2k+1} \\ &= \sqrt{\frac{2}{\pi x}} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^k k!} \frac{1}{1 \cdot 3 \cdot 5 \ldots (2k+1)} x^{2k+1} \\ &= \sqrt{\frac{2}{\pi x}} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1} \\ &= \sqrt{\frac{2}{\pi x}} \sin(x) \end{aligned}$$

We have proved that

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}} \sin(x) \tag{4.53}$$

Now we get back to the point where we consider the case $r = -v$ in equations (4.49) and (4.50). Now equation (4.49) yields $(1 - 2v) a_1 = 0$. If $v \neq 1/2$, then $a_1 = 0$ and equation (4.50) yields

$$j(j - 2v) a_j + a_{j-2} = 0, \qquad j = 2, 3, \ldots$$

Therefore,

$$a_1 = a_3 = a_5 = \cdots = a_{\text{odd}} = \cdots = 0$$

and for $j = 2k$ even we get the following recursive formula

$$a_{2k} = -\frac{1}{2^2 k (k - v)} a_{2k-2}, \qquad k = 1, 2, \ldots$$

Now let us assume that v is not an integer. This time we choose for a_0 the following value.

$$a_0 = \frac{1}{2^{-v}\Gamma(1-v)}$$

Just as before, the recursive relation yields

$$a_{2k} = \frac{(-1)^k}{2^{2k-v}k!\Gamma(k-v+1)}$$

and we obtain a second particular solution to Bessel's equation, that we call $J_{-v}(x)$, given by

$$J_{-v}(x) = x^{-v}\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k}}{2^{2k-v}k!\Gamma(k-v+1)}$$

This solution is well defined provided that $v \neq 1/2$ and that v is not an integer. However, we may extend the scope of the above notation and define $J_{-n}(x)$ for $n = 1, 2, \ldots$ by the formula

$$J_{-n}(x) = x^{-n}\sum_{k=0}^{\infty}\frac{(-1)^k x^{2k}}{2^{2k-n}k!\Gamma(k-n+1)}$$

by agreeing that the terms for which $k \leq n-1$ are zero. This would make sense because for these values of k the gamma function would become infinity. The relation between $J_n(x)$ and $J_{-n}(x)$ is given in the following theorem.

Theorem 4.15.1. *We have*

$$J_{-n}(x) = (-1)^n J_n(x), \qquad n = 1, 2, \ldots$$

Proof. We have

$$\begin{aligned} J_{-n}(x) &= x^{-n} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k-n} k! \Gamma(k-n+1)} \\ &= x^{-n} \sum_{k=n}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k-n} k! \Gamma(k-n+1)} \\ &= x^{-n} \sum_{k'=0}^{\infty} \frac{(-1)^{k'+n} x^{2(k'+n)}}{2^{2(k'+n)-n} (k'+n)! \Gamma(k'+1)} \\ &= (-1)^n x^n \sum_{k'=0}^{\infty} \frac{(-1)^{k'} x^{2k'}}{2^{2k'+n} (k'+n)! k'!} \\ &= (-1)^n J_n(x) \end{aligned}$$

where, $k' = k - n$. □

Finally, if $r = -v$ and $v = 1/2$, then equation (4.49) becomes

$$j(j-1) a_j + a_{j-2} = 0$$

Again, we get a recursive relation.

$$a_j = -\frac{1}{j(j-1)} a_{j-2}$$

Therefore, we get

$$a_{2k} = \frac{(-1)^k}{(2k)!} a_0, \qquad k = 0, 1, 2, \ldots$$

and

$$a_{2k-1} = \frac{(-1)^k}{(2k-1)!} a_1 \qquad k = 1, 2, \ldots$$

Hence, the general solution to Bessel's equation in this case is

$$\begin{aligned} y(x) &= x^{-1/2} \sum_{m=0}^{\infty} a_m x^m \\ &= x^{-1/2} \sum_{k=0}^{\infty} a_{2k} x^{2k} + x^{-1/2} \sum_{k=1}^{\infty} a_{2k-1} x^{2k-1} \\ &= a_0 x^{-1/2} \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} x^{2k} + a_1 x^{-1/2} \sum_{k=1}^{\infty} \frac{(-1)^k}{(2k-1)!} x^{2k-1} \\ &= x^{-1/2} \left(a_0 \cos(x) + a_1 \sin(x) \right) \end{aligned}$$

The choice $a_0 = 0$ and a_1 given by

$$a_1 = \frac{1}{2^{1/2} \Gamma\left(\frac{1}{2} + 1\right)} = \frac{1}{2^{1/2} \frac{1}{2} \Gamma\left(\frac{1}{2}\right)} = \sqrt{\frac{2}{\pi}}$$

yields $\sqrt{\frac{2}{\pi x}} \sin(x) = J_{1/2}(x)$. Likewise, the choice $a_0 = \sqrt{2/\pi}$ and $a_1 = 0$ yields $\sqrt{\frac{2}{\pi x}} \cos(x)$. We define this function to be $J_{-1/2}(x)$. Hence, we have

$$J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}} \cos(x)$$

For v not an integer $J_v(x)$ and $J_{-v}(x)$ are not scalar multiples of each other. Hence, they form a linearly independent set of solutions of Bessel's differential equation. Therefore, in this case the general solution to Bessel's differential equation is

$$y(x) = c_1 J_v(x) + c_2 J_{-v}(x), \qquad v > 0, \qquad v \neq 1, 2, \ldots$$

Four Basic Properties of $J_v(x)$

We conclude this section with four useful properties of $J_v(x)$. We have the following theorem.

Theorem 4.15.2. *The following identities hold for $v > 0$.*

1. $(x^v J_v(x))' = x^v J_{v-1}(x)$

2. $\left(x^{-v}J_v(x)\right)' = -x^{-v}J_{v+1}(x)$

3. $J_{v-1}(x) + J_{v+1}(x) = \frac{2v}{x}J_v(x)$

4. $J_{v-1}(x) - J_{v+1}(x) = 2J_v'(x)$

Proof. We prove only the first identity. The proof of the remaining identities is left as exercise. We have

$$\begin{aligned}
\left(x^v J_v(x)\right)' &= \left(\sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+2v}}{2^{2k+v}k!\Gamma(v+k+1)}\right)' \\
&= \sum_{k=0}^{\infty} \frac{(2k+2v)(-1)^k x^{2k+2v-1}}{2^{2k+v}k!(v+k)\Gamma(v+k)} \\
&= x^v \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+(v-1)}}{2^{2k+(v-1)}k!\Gamma(v+k)} \\
&= x^v J_{v-1}(x)
\end{aligned}$$

The remaining identities can be proved similarly. □

Notation We denote by α_{kn} $(k = 1, 2, \dots)$ the infinitely many positive zeros of $J_n(x)$.

$$\alpha_{1n} < \alpha_{2n} < \cdots < \alpha_{kn} < \cdots$$

4.16 Bessel's Functions Y_v

In the last section we found solutions to Bessel's equation

$$x^2y'' + xy' + \left(x^2 - v^2\right)y = 0, \quad v \geq 0$$

If $v = n$ an integer, then Theorem 4.15.1 of the last section shows that $J_n(x)$ and $J_{-n}(x)$ are linearly dependent. So, we still need to find a second independent solution. We address the question of the second solution here.

For noninteger α we define

$$Y_\alpha(x) = \frac{J_\alpha(x)\cos(\alpha\pi) - J_{-\alpha}(x)}{\sin(\alpha\pi)}$$

For an integer n we define $Y_n(x)$ as the limit of $Y_\alpha(x)$ as α approaches n through noninteger values.

$$Y_n(x) = \lim_{\alpha \to \infty} Y_\alpha(x), \qquad \alpha \ \text{ non integer}$$

The functions $Y_\alpha(x)$ (including integer values of α) are called **Bessel's functions of the second kind**. They are solutions of Bessel's equation that are linearly independ of $J_\alpha(x)$. Therefore, the general solution to Bessel's equation can be written as

$$y(x) = c_1 J_\alpha(x) + c_2 Y_\alpha(x) \tag{4.54}$$

Both functions $J_\alpha(x)$ and $Y_\alpha(x)$ are well-understood and their values can be approximated for any given α.

The functions $Y_n(x)$ can be defined independently of the limit above by seeking a series representation of a solution of the form

$$Y_n(x) = A_0 J_n(x) \ln(x) + \sum_{k=1}^{\infty} A_k x^k$$

Substitution into Bessel's equation eventually yields

$$Y_n(x) = \frac{2}{\pi} J_n(x) \left(\ln\left(\frac{x}{2}\right) + \gamma \right) + \frac{x^n}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^{k-1} (g_k + g_{k+n})}{2^{2k+n} k! (k+n)!} x^{2k} \tag{4.55}$$
$$- \frac{x^{-n}}{\pi} \sum_{k=0}^{n-1} \frac{(n-k-1)!}{2^{2k-n} k!} x^{2k}$$

where

$$n = 0, 1, 2, \ldots, \qquad g_0 = 0, \quad g_r = 1 + \frac{1}{2} + \cdots + \frac{1}{r}$$

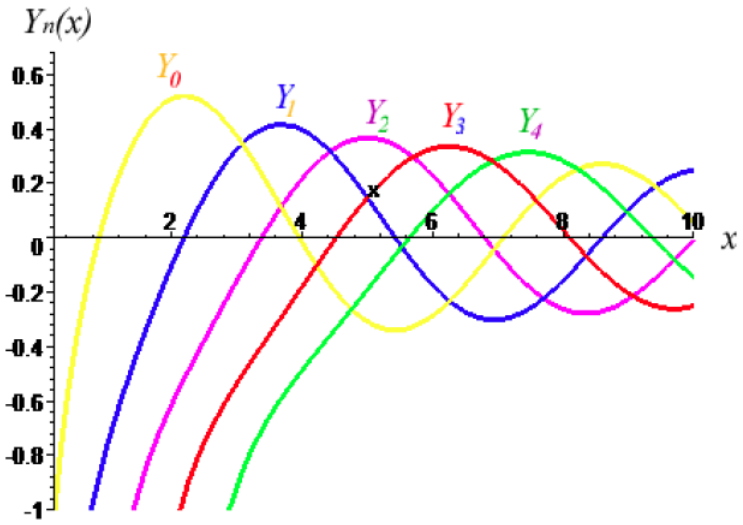
and $\gamma \approx 0.5772156649\ldots$ is Euler's constant defined as

$$\gamma = \lim_{r \to \infty} \left(1 + \frac{1}{2} + \cdots + \frac{1}{r} - \ln(r) \right)$$

Equation (4.55) shows that the values of $Y_n(x)$ near $x = 0$ go to $-\infty$ because of $\ln(x/2)$. This is in sharp contrast with the behavior of $J_n(x)$ near the origin.

$$\lim_{x \to 0} Y_n(x) = -\infty$$

The following is a graph of $Y_n(x)$, $n = 0, 1, 2, 3$ in a positive range of x.



Just as in the case of $J_n(x)$, it should be noted that $Y_n(x)$ has infinitely many zeros and that the values approach zero as x approaches infinity.

Note When we solve Bessel's equation we consider the Bessel functions $J_\alpha(x)$ and $Y_\alpha(x)$ as known functions (just as as we do with sines and cosines) and we rarely use the power series representations of $J_\alpha(x)$ and $Y_\alpha(x)$. Typically we accept equation (4.54) as a complete general solution to Bessel's equation.

Example 4.16.1. Find the general solution of each of the differential equations.

1. $x^2y'' + xy' + x^2y = 0$
2. $x^2y'' + xy' + (x^2 - 16)\,y = 0$
3. $4x^2y'' + 4xy' + (100x^2 - 9)\,y = 0$

Solution

1. This equation is Bessel with $v = 0$. So its general solution is given by

$$y(x) = c_1 J_0(x) + c_2 Y_0(x)$$

2. This equation is Bessel with $v = 4$. Its general solution is given by

$$y(x) = c_1 J_4(x) + c_2 Y_4(x)$$

3. We divide the equation by 4 to get

$$x^2 y'' + xy' + \left(25x^2 - \frac{9}{4}\right) y = 0 \tag{4.56}$$

This equation is not Bessel, but it can be transformed to a Bessel one by a simple change of variables. Let $z = 5x$. Then by Chain Rule we have

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = 5\frac{dy}{dz}$$

Hence,

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(5\frac{dy}{dz}\right) = 5\frac{d}{dz}\left(\frac{dy}{dx}\right)\frac{dz}{dx} = 25\frac{d^2y}{dz^2}$$

Substitution into (4.56) yields

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} + \left(z^2 - \frac{9}{4}\right) y = 0$$

This is Bessel in $y(z)$ with $v = 3/2$. Hence,

$$y(z) = c_1 J_{3/2}(z) + c_2 Y_{3/2}(z)$$

Therefore,

$$y(x) = c_1 J_{3/2}(5x) + c_2 Y_{3/2}(5x)$$

4.17 Orthogonality of Bessel Functions

Bessel functions of the first kind satisfy an interesting orthogonality condition that is useful in finding exact solutions for the displacement function of a circular membrane. The function $J_n = J_n(t)$ satisfies Bessel's equation. Hence, we have the identity

$$t^2\ddot{J}_n + t\dot{J}_n + \left(t^2 - n^2\right) J_n = 0 \tag{4.57}$$

where $\dot{J}_n = \frac{dJ_n}{dt}$. Next, we rewrite this equation as a Sturm-Liouville equation by using a change of variables (as in part 3 of Example 4.16.1 of the last section). Let $t = \lambda x$. Then by Chain rule (using $J_n' = \frac{dJ_n}{dx}$) we have

$$\dot{J}_n = J_n' \frac{dt}{dx} = \lambda J_n'$$

Hence,

$$\ddot{J}_n = \frac{d}{dt}\left(\lambda J_n'\right) = \lambda \frac{dJ_n'}{dx}\frac{dx}{dt} = \lambda^2 J_n''$$

Substitution into (4.57) yields

$$x^2 J_n''\left(\lambda x\right) + x J_n'\left(\lambda x\right) + \left(\lambda^2 x^2 - n^2\right) J_n\left(\lambda x\right) = 0$$

which we may rewrite as

$$\left[x J_n'\left(\lambda x\right)\right]' + \left(-\frac{n^2}{x} + \lambda^2 x\right) J_n\left(\lambda x\right) = 0 \tag{4.58}$$

This is a Sturm-Liouville equation with parameter λ^2 and with

$$p\left(x\right) = x, \quad q\left(x\right) = -\frac{n^2}{x}, \quad r\left(x\right) = x$$

Since, $r\left(0\right) = 0$, the orthogonality theorem of eigenfunctions of Sturm-Liouville problems implies that the solutions $J_n\left(\lambda x\right)$ dedined on an interval $[0, R]$ that satisfy the boundary condition

$$J_n\left(\lambda R\right) = 0$$

are orthogonal on $[0, R]$ with respect to weight function $p(x) = x$. Since $\lambda R = \alpha_{kn}$ where α_{kn} $(k = 1, 2, \dots)$ is the kth zero of $J_n\left(x\right)$, we have

$$\lambda = \lambda_{kn} = \frac{\alpha_{kn}}{R}, \qquad k = 1, 2, \dots$$

Therefore, we have the following theorem.

Theorem 4.17.1 (Orthogonality of Bessel Functions)**.** *For each fixed* $n = 0, 1, 2, \dots$ *the functions* $J_n\left(\lambda_{1n} R\right),\ J_n\left(\lambda_{2n} R\right),\ J_n\left(\lambda_{3n} R\right), \dots$ *(where* $\lambda_{kn} = \alpha_{kn}/R$*) form an orthogonal set on* $[0, R]$ *with respect to weight* $p(x) = x$*. Thus,*

$$\int_0^R x J_n\left(\lambda_{kn} R\right) J_n\left(\lambda_{jn} R\right)\ dx = 0, \qquad k \neq j$$

Let $f(x)$ be a function defined on $[0, R]$. For fixed n it may be possible to write $f(x)$ as a series in $J_n(\lambda_{kn}x)$, $(k = 1, 2, \dots)$.

$$f(x) = a_1 J_n(\lambda_{1n}x) + a_2 J_n(\lambda_{2n}x) + \cdots + a_k J_n(\lambda_{kn}x) + \cdots \tag{4.59}$$

Such a representation of $f(x)$ is called a **Fourier-Bessel series** of $f(x)$. The coefficients a_k are called the **Fourier-Bessel coefficients** of $f(x)$. These coeffiecients can be computed by the usual general formula.

$$a_k = \frac{\langle f(x), J_n(\lambda_{kn}x)\rangle}{\langle J_n(\lambda_{kn}x), J_n(\lambda_{kn}x)\rangle} = \frac{\int_0^R xf(x) J_n(\lambda_{kn}x)\ dx}{\int_0^R xJ_n^2(\lambda_{kn}x)\ dx}$$

The denominator can be computed once for all as follows. We mulitply equation (4.58) by $2xJ_n'(\lambda x)$ (with $\lambda = \lambda_{kn}$) to get

$$2x\left[xJ_n'(\lambda x)\right]' J_n'(\lambda x) + 2\left(\lambda^2 x - n^2\right) J_n(\lambda x) J_n'(\lambda x) = 0$$

or

$$\left[\left(xJ_n'(\lambda x)\right)^2\right]' + \left(\lambda^2 x - n^2\right)\left[J_n^2(\lambda x)\right]' = 0$$

Nexe, we integrate over $[0, R]$ to get

$$\left(xJ_n'(\lambda x)\right)^2\Big|_0^R = -\int_0^R \left(\lambda^2 x - n^2\right)\left[J_n^2(\lambda x)\right]'\ dx$$

First we compute the left hand side: $\left(xJ_n'(\lambda x)\right)^2\Big|_0^R = \left[RJ_n'(\lambda R)\right]^2$. Now we subtract $J_{n-1} - J_{n+1} = 2J_n'$ from $J_{n-1} + J_{n+1} = \frac{2n}{x}J_n$ to get $2J_{n+1} = \frac{2n}{x}J_n - 2J_n'$ or $xJ_n' = nJ_n - xJ_{n+1}$. Therefore, the left hand side becomes

$$\left[RJ_n'(\lambda R)\right]^2 = \left[nJ_n(\lambda R) - \lambda J_{n+1}(\lambda R)\right]^2 = \lambda^2 R^2 J_{n+1}^2(\lambda R)$$

because $J_n(\lambda R) = J_n(\lambda_{kn}R) = J_n(\alpha_{kn}) = 0$.

For the right hand side we integrate by parts using $u = \lambda^2 x^2 - n^2$ and $dv = \left[J_n^2(\lambda x)\right]'\ dx$. Thus, $du = 2\lambda^2 x dx$ and $v = J_n^2(\lambda x)$. Therefore, we get

$$-\left(\lambda^2 x - n^2\right) J_n^2(\lambda x)\Big|_0^R + 2\lambda^2 \int_0^R xJ_n^2(\lambda x)\ dx = 2\lambda^2 \int_0^R xJ_n^2(\lambda x)\ dx$$

We conclude that

$$\lambda^2 R^2 J_{n+1}^2(\lambda R) = 2\lambda^2 \int_0^R xJ_n^2(\lambda x)\ dx$$

Hence,

$$\int_0^R xJ_n^2(\lambda x)\ dx = \frac{R^2}{2}J_{n+1}^2(\lambda R)$$

Therefore, the Fourier-Bessel coefficient a_k can be computed as $a_k = \frac{2}{R^2J_{n+1}^2(\lambda R)}\int_0^R xf(x)J_n(\lambda_{kn}x)\ dx$. So we have the following theorem.

Theorem 4.17.2 (Fourier-Bessel Series)**.** *The Fourier-Bessel series of $f(x)$ defined on $[0,R]$ is computed by*

$$\begin{aligned} f(x) &= a_1J_n(\lambda_{1n}x) + a_2J_n(\lambda_{2n}x) + \cdots + a_kJ_n(\lambda_{kn}x) + \cdots \\ \lambda_{kn} &= \frac{\alpha_{kn}}{R}, \quad k = 1,2,\ldots, \quad n = 0,1,2,\ldots \\ a_k &= \frac{2}{R^2J_{n+1}^2(\lambda_{kn}R)}\int_0^R xf(x)J_n(\lambda_{kn}x)\ dx \end{aligned}$$

Example 4.17.1. Find the Fourier-Bessel series of $f(x) = 1$ defined on $[0,R]$ using $J_0(\lambda_{k0}x)$ $(k = 1,2,\ldots)$.

Solution The Fourier-Bessel series corefficients are computed by the formulas of Theorem 4.17.2.

$$\begin{aligned} a_k &= \frac{2}{R^2J_1^2(\lambda_{k0}R)}\int_0^R x\cdot 1\cdot J_0(\lambda_{k0}x)\ dx \qquad \text{(change of variables } w = \lambda_{k0}x\text{)} \\ &= \frac{2}{\lambda_{k0}^2R^2J_1^2(\lambda_{k0}R)}\int_0^{\lambda_{k0}R} wJ_0(w)\ dw \\ &= \frac{2}{\alpha_{k0}^2J_1^2(\alpha_{k0})}\int_0^{\alpha_{k0}} wJ_0(w)\ dw \\ &= \frac{2}{\alpha_{k0}^2J_1^2(\alpha_{k0})}\int_0^{\alpha_{k0}} [wJ_1(w)]'\ dw \\ &= \frac{2}{\alpha_{k0}^2J_1^2(\alpha_{k0})}\ wJ_1(w)|_0^{\alpha_{k0}} \\ &= \frac{2}{\alpha_{k0}^2J_1^2(\alpha_{k0})}\alpha_{k0}J_1(\alpha_{k0}) \\ &= \frac{2}{\alpha_{k0}J_1(\alpha_{k0})} \end{aligned}$$

Therefore,

$$1 = 2\left(\frac{J_0(\lambda_{10}x)}{\alpha_{10}J_1(\alpha_{10})} + \frac{J_0(\lambda_{20}x)}{\alpha_{20}J_1(\alpha_{20})} + \frac{J_0(\lambda_{30}x)}{\alpha_{30}J_1(\alpha_{30})} + \cdots\right)$$

Example 4.17.2. Find the Fourier-Bessel series of $f(x) = x^2$ defined on $[0, R]$ using $J_0(\lambda_{k0}x)$ $(k = 1, 2, \dots)$.

Solution Again, the Fourier-Bessel series coefficients are computed by the formulas of Theorem 4.17.2.

$$\begin{aligned}
a_k &= \frac{2}{R^2 J_1^2(\lambda_{k0}R)} \int_0^R x \cdot x^2 J_0(\lambda_{k0}x)\ dx \qquad \text{(change of variables } w = \lambda_{k0}x\text{)} \\
&= \frac{2}{\lambda_{k0}^4 R^2 J_1^2(\alpha_{k0})} \int_0^{\alpha_{k0}} w^2 \left(w J_0(w)\right)\ dw \\
&= \frac{2}{\lambda_{k0}^4 R^2 J_1^2(\alpha_{k0})} \int_0^{\alpha_{k0}} w^2\ \left[w J_1(w)\right]'\ dw \qquad \text{(int. by parts with } u = w^2\text{)} \\
&= \frac{2}{\lambda_{k0}^4 R^2 J_1^2(\alpha_{k0})} \left(w^3 J_1(w)\big|_0^{\alpha_{k0}} - 2\int_0^{\alpha_{k0}} w^2 J_1(w)\ dw \right) \\
&= \frac{2}{\lambda_{k0}^4 R^2 J_1^2(\alpha_{k0})} \left(\alpha_{k0}^3 J_1(\alpha_{k0}) - 2\int_0^{\alpha_{k0}} \left[w^2 J_2(w)\right]'\ dw \right) \\
&= \frac{2}{\lambda_{k0}^4 R^2 J_1^2(\alpha_{k0})} \left(\alpha_{k0}^3 J_1(\alpha_{k0}) - 2\, w^2 J_2(w)\big|_0^{\alpha_{k0}} \right) \\
&= \frac{2}{\lambda_{k0}^4 R^2 J_1^2(\alpha_{k0})} \left(\alpha_{k0}^3 J_1(\alpha_{k0}) - 2\alpha_{k0}^2 J_2(\alpha_{k0}) \right) \\
&= \frac{2R^2}{\alpha_{k0}^4 J_1^2(\alpha_{k0})} \left(\alpha_{k0}^3 J_1(\alpha_{k0}) - 2\alpha_{k0}^2 J_2(\alpha_{k0}) \right)
\end{aligned}$$

Hence,

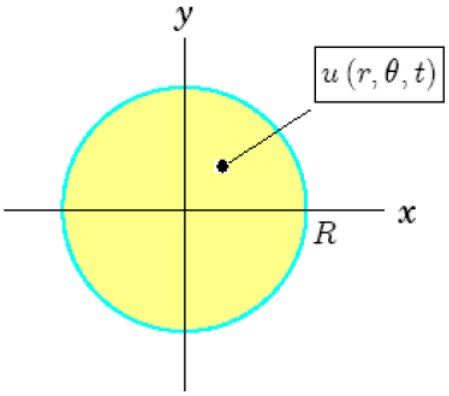
$$a_k = \frac{2R^2}{\alpha_{k0} J_1(\alpha_{k0})} \left[1 - \frac{2J_2(\alpha_{k0})}{\alpha_{k0} J_1(\alpha_{k0})} \right]$$

Therefore,

$$x^2 = 2R^2 \left(\frac{J_0(\lambda_{10}x)}{\alpha_{10} J_1(\alpha_{10})} \left[1 - \frac{2J_2(\alpha_{10})}{\alpha_{10} J_1(\alpha_{10})} \right] + \frac{J_0(\lambda_{20}x)}{\alpha_{20} J_1(\alpha_{20})} \left[1 - \frac{2J_2(\alpha_{20})}{\alpha_{20} J_1(\alpha_{20})} \right] + \cdots \right)$$

4.18 Circular Membrane

In this section we study special vibrations of the circular membrane of radius R. Let $u(r, \theta, t)$ be the deflection at the point with polar coordinates (r, θ) at time t.

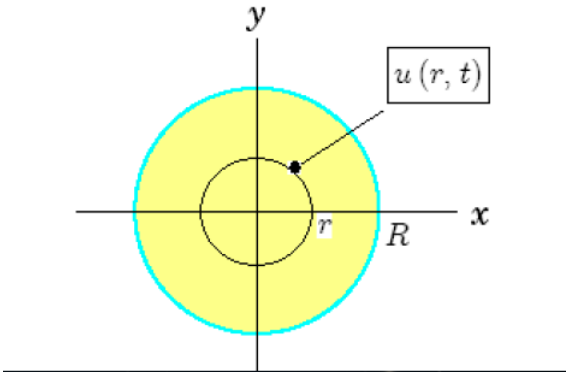


Then the wave equation in polar coordinates is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u = c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} \right)$$

We study the special case where the solutions are "radially symmetric", i.e., the solutions depend on the distance from the origin r but not on the angle θ. So, in this case the deflection is a function of r and t only: $u = (r, t)$. Thus, $u_\theta = 0$ and $u_{\theta\theta} = 0$. Therefore, the wave equation simplifies to

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u = c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r} \frac{\partial u}{\partial r} \right) \tag{4.60}$$



We assume that the boundary is fixed. Hence, we have the boundary condition

$$u(R, t) = 0, \quad t \geq 0 \tag{4.61}$$

We also assume that we are given an initial deflection $f(r)$ and an initial velocity $g(r)$. Thus,

$$\begin{aligned} u(r,0) &= f(r) \\ \frac{\partial u}{\partial t}(r,0) &= g(r) \end{aligned} \tag{4.62}$$

Equations (4.60). (4.61), and (4.62) define an initial boundary value problem which we solve by using separation of variables.

Let $u(r,t) = Q(r)T(t) = QT$, where Q is a function in r only and T is a function in t only. Then substitution into (4.60) yields

$$Q''T = c^2\left(Q''T + \frac{1}{r}Q'T\right)$$

Hence,

$$\frac{T''}{c^2T} = \frac{Q''}{Q} + \frac{1}{r}\frac{Q'}{Q} = -v^2$$

where we called the common constant $-v^2$ $(v > 0)$. We want this constant to be positive on physical grounds: we reject solutions that are exponentials (increasing or decreasing) in time and also linear functions in time (increasing or constant). Hence, we get

$$T'' + c^2v^2T = 0 \tag{4.63}$$

and

$$r^2Q'' + rQ' + r^2v^2Q = 0 \tag{4.64}$$

Equation (4.63) is trivial to solve, if we know v. Equation (4.64) is almost Bessel. It can be transformed to Bessel by a familiar change of variables. Let $w = vr$. By Chain Rule, we have

$$\frac{dQ}{dr} = \frac{dQ}{dw}\frac{dw}{dr} = v\frac{dQ}{dw}$$

Hence,

$$\frac{d^2Q}{dr^2} = \frac{d}{dr}\left(v\frac{dQ}{dw}\right) = v\frac{d}{dw}\left(\frac{dQ}{dw}\right)\frac{dw}{dr} = v^2\frac{d^2Q}{dw^2}$$

Thus, equation (4.64) transforms to

$$w^2\frac{d^2Q}{dw^2} + w\frac{dQ}{dw} + w^2Q = 0$$

This equation is Bessel, so the general solution is

$$Q(w) = c_1 J_0(w) + c_2 Y_0(w)$$

Thus, Q as a function in r is given by

$$Q(r) = c_1 J_0(vr) + c_2 Y_0(vr)$$

However, we may assume that $c_2 = 0$, because we want the deflections to be finite and we know that $Y_0(r)$ approaches $-\infty$ as r approaches 0. So, up to a constant coefficient c_1, we have

$$Q(r) = J_0(vr)$$

On the boundary we must have

$$u(R,t) = Q(R)\,T(t) = 0, \quad \text{for all} \quad t > 0$$

and since $T(t)$ is not identically zero, it must be that $Q(R) = 0$. Thus,

$$J_0(vR) = 0$$

Now J_0 has infinitely many zeros $\alpha_{10} < \alpha_{20} < \cdots$. Hence, $vR = \alpha_{k0}$ $(k = 1, 2, \dots)$. Therefore,

$$v = v_k = \frac{\alpha_{k0}}{R}, \quad k = 1, 2, \dots$$

Hence,

$$Q_k(r) = J_0\left(\frac{\alpha_{k0}}{R} r\right)$$

Now equation (4.63) becomes $T'' + \left(\frac{c\alpha_{k0}}{R}\right)^2 T = 0$. Hence,

$$T_k(t) = a_k \cos\left(\frac{c\alpha_{k0}}{R} t\right) + b_k \sin\left(\frac{c\alpha_{k0}}{R} t\right)$$

Therefore,

$$u_k(r,t) = \left(a_k \cos\left(\frac{c\alpha_{k0}}{R} t\right) + b_k \sin\left(\frac{c\alpha_{k0}}{R} t\right)\right) J_0\left(\frac{\alpha_{k0}}{R} r\right), \quad k = 1, 2, \dots$$

The $u_k(r,t)$ are the **eigenfunctions** corresponding to the **eigenvalues** λ_k. In physics, $u_k(r,t)$ is called the kth **normal mode**. It has frequency $\frac{c\alpha_{k0}}{2\pi R}$.

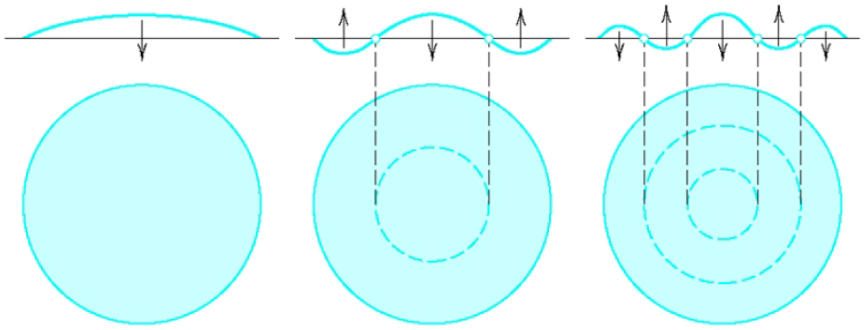
Note The zeros of J_0 are not evenly spaced. This is the reason that the sound of a drum is quite different than that of a string instrument.

Let us compute the stationary points of $u_k(r,t)$. These are the points on the disk where the deflection is 0 for all $t > 0$. I.e., $u_k(r,t) = 0$ for all $t > 0$. This forces $J_0\left(\frac{\alpha_{k0}}{R}r\right) = 0$. Therefore, $\alpha_{n0} = \frac{\alpha_{k0}}{R}r$. Thus, $r = \frac{\alpha_{n0}}{\alpha_{k0}}R$. Now $0 \le r \le R$, hence, we must have $0 \le \frac{\alpha_{n0}}{\alpha_{k0}} \le R$. Therefore, $n = 1, 2, \ldots, k$. So, r must be

$$r = \frac{\alpha_{n0}}{\alpha_{k0}}R, \quad n = 1, 2, \ldots, k$$

For example:

1. If $k = 1$, then $r = R$, so only the boundary remains fixed.
2. If $k = 2$, then $r = \frac{\alpha_{10}}{\alpha_{20}}R$ or $r = R$. So two circles remain fixed: The circle of radius $\frac{\alpha_{10}}{\alpha_{20}}R$ and the boundary circle.
3. If $k = 3$, then $r = \frac{\alpha_{10}}{\alpha_{30}}R$, or $r = \frac{\alpha_{20}}{\alpha_{30}}R$ or $r = R$. So three circles remain fixed.



Finally, let us consider all u_k in an infinite series: $u = \sum_{k=1}^{\infty} u_k$. We have

$$u(r,t) = \sum_{k=1}^{\infty}\left(a_k \cos\left(\frac{c\alpha_{k0}}{R}t\right) + b_k \sin\left(\frac{c\alpha_{k0}}{R}t\right)\right) J_0\left(\frac{\alpha_{k0}}{R}r\right) \tag{4.65}$$

Setting $t = 0$ and using the first initial condition yields

$$u(r,0) = f(r) = \sum_{k=1}^{\infty} a_k J_0\left(\frac{\alpha_{k0}}{R}r\right)$$

This is the Fourier-Bessel series for $f(r)$. Theorem 4.17.2 provides us with a formula for the coefficients.

$$a_k = \frac{2}{R^2 J_1^2(\alpha_{k0})} \int_0^R r f(r) J_0\left(\frac{\alpha_{k0}}{R} r\right) dr, \quad k = 1, 2, \ldots \tag{4.66}$$

Differentiating (4.65) and using the second initial condition yields

$$\frac{\partial u}{\partial t}(r, 0) = g(r) = \sum_{k=1}^{\infty} \frac{c\alpha_{k0}}{R} b_k J_0\left(\frac{\alpha_{k0}}{R} r\right)$$

Therefore,

$$b_k = \frac{2}{c\alpha_{k0} R J_1^2(\alpha_{k0})} \int_0^R r g(r) J_0\left(\frac{\alpha_{k0}}{R} r\right) dr, \quad k = 1, 2, \ldots \tag{4.67}$$

Equations (4.65), (4.66), and (4.67) completely solve the deflection problem for the radially symmetric solutions of a circular drum of radius R. So, we have the following theorem.

Theorem 4.18.1 (Radially Symmetric Solutions of Circular Membrane)**.** *A Fourier series type of solution for the initial boundary value problem*

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= c^2 \left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r}\right) \\ u(R, t) &= 0, \quad t \geq 0 \\ u(r, 0) &= f(r), \quad 0 \leq r \leq R \\ \frac{\partial u}{\partial t}(r, 0) &= g(r), \quad 0 \leq r \leq R \end{aligned}$$

is given by

$$u(r, t) = \sum_{k=1}^{\infty} \left(a_k \cos\left(\frac{c\alpha_{k0}}{R} t\right) + b_k \sin\left(\frac{c\alpha_{k0}}{R} t\right)\right) J_0\left(\frac{\alpha_{k0}}{R} r\right) \tag{4.68}$$

$$a_k = \frac{2}{R^2 J_1^2(\alpha_{kn})} \int_0^R r f(r) J_0\left(\frac{\alpha_{k0}}{R} r\right) dr$$

$$b_k = \frac{2}{c\alpha_{k0} R J_1^2(\alpha_{kn})} \int_0^R r g(r) J_0\left(\frac{\alpha_{k0}}{R} r\right) dr$$

Example 4.18.1. Find the radially symmetric solution for the deflection $u(r,t)$ of a circular drum, if $R=1$, $c^2=4$, the initial velocity is zero and the inital displacement is $f(r)=1-r^2$.

The initial velocity is zero, so we have $b_k=0$. By using the computations of Examples 4.17.1 and 4.17.2 we have

$$\begin{aligned} a_k &= \frac{2}{1^2 J_1^2(\alpha_{k0})}\int_0^1 r\left(1-r^2\right) J_0\left(\frac{\alpha_{k0}}{1}r\right)dr \\ &= \frac{2}{J_1^2(\alpha_{k0})}\left[\int_0^1 r\cdot 1\cdot J_0(\alpha_{k0}r)\,dr - \int_0^1 r\cdot r^2 J_0(\alpha_{k0}r)\,dr\right] \\ &= \frac{2}{\alpha_{k0}J_1(\alpha_{k0})} - \frac{2}{\alpha_{k0}J_1(\alpha_{k0})}\left[1-\frac{2J_2(\alpha_{k0})}{\alpha_{k0}J_1(\alpha_{k0})}\right] \\ &= \frac{4J_2(\alpha_{k0})}{\alpha_{k0}^2 J_1^2(\alpha_{k0})} \end{aligned}$$

Therefore,

$$u(r,t)=4\sum_{k=1}^{\infty}\frac{J_2(\alpha_{k0})}{\alpha_{k0}^2 J_1^2(\alpha_{k0})}\cos(2\alpha_{k0}t)\,J_0(\alpha_{k0}r)$$

Chapter 5

Complex Variables

5.1 Complex Numbers

Today most areas of mathematics, physics, and engineering use complex numbers. Complex numbers were discovered by Cardano and first mentioned in his book *Ars magna* (1545 AD.). It was Gauss, however, who according to G. H. Hardy, "was the first mathematician to use complex numbers in a really confident and scientific way".

Arithmetic with Complex Numbers

The **Imaginary unit**, i or $\sqrt{-1}$ is defined by the property

$$i^2 = -1$$

Hence,

$$i^3 = -i, \quad i^4 = 1, \quad i^5 = i$$

Example 5.1.1. Compute i^{1246}.

Solution:

$$i^{1246} = i^{311 \cdot 4 + 2} = (i^4)^{311}\, i^2 = 1^{311}\,(-i) = -i$$

A **complex number** z is an expression of the form $z = a + bi$, where both a and b are real numbers. The set of all complex numbers is denoted by $\mathbf{C}$. The **real part**, $\mathrm{Re}(z)$, of z is a. The **imaginary part**, $\mathrm{Im}(z)$, of z is b. If

$b = 0$, then z is a real number. If $a = 0$, then z is **pure imaginary**. The **complex conjugate** of z is $\overline{z}$

$$\overline{z} = a - ib$$

Example 5.1.2. We have

$$\text{Re}(1 - 2i) = 1, \;\; \text{Im}(5 - 2i) = -2, \;\; \overline{1 - i} = 1 + i, \;\; \overline{-3} = -3$$

Two complex numbers are **equal** if their respective real and imaginary parts are equal. For example, $5 + xi = y - 4i$ if and only if $y = 5$ and $x = -4$.

The **absolute value** $|z|$ of a complex number $z = a+ib$ is the nonnegative real number

$$|z| = \sqrt{a^2 + b^2}$$

Example 5.1.3. We have

$$|-2 + 3i| = \sqrt{(-2)^2 + 3^2} \; = \sqrt{13}$$

Complex numbers are into one-to-one correspondence with the points of the Euclidean plane. The complex number $z = a+ib$ corresponds to the plane point (a, b). This is a very useful representation of comples numbers. Under this representation the x-axis is called the **real axis** and the y-axis is called the **imaginary axis**. The absolute value $|z|$ of $z = a + ib$ is geometrically interpreted as the distance of the point (a, b) from the origin.
The **sum, difference** and **product** of complex numbers is as in real numbers, with the following provisions: All powers of i are calculated. Terms are collected so that the final result in the form $a + ib$ for real a and b.

Example 5.1.4. We have

$$\begin{aligned}(1 - 2i) - (2 + 3i)(-1 + i) &= (1 - 2i) - \left(-2 + 2i - 3i + 3i^2\right) \\ &= (1 - 2i) - (-5 - i) \\ &= 6 - i\end{aligned}$$

The **quotient**, $\dfrac{z}{w}$, with $w \neq 0$ of two complex numbers $z = a + bi$ and $w = c + di$ is the number

$$\frac{z}{w} = \frac{z\overline{w}}{w\overline{w}} = \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2}i \, , \qquad w \neq 0$$

It is easy to verify that $w(\frac{z}{w}) = z$.

Example 5.1.5. We have

$$\frac{2+3i}{1+i} = \frac{(2+3i)(1-i)}{(1+i)(1-i)} = \frac{5+i}{2} = \frac{5}{2} + \frac{1}{2}i$$

Theorem 5.1.1 (Properties of the Absolute Value)**.** *The absolute value satisfies the following basic properties.*

1. $|zw| = |z|\,|w|$
2. $\left|\dfrac{z}{w}\right| = \dfrac{|z|}{|w|}$, $(w \neq 0)$
3. *The* ***Triangle Inequality***

$$|z+w| \leq |z| + |w|$$

4. *The following consequence of the Triangle Inequality*

$$||z| - |w|| \leq |z+w|$$

The next theorem summarizes the basic properties of complex conjugation.

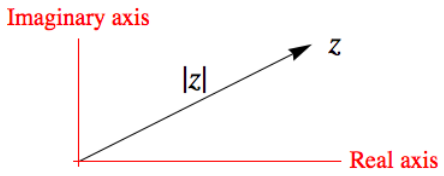
Theorem 5.1.2 (Properties of Conjugation)**.** *Let z and w be complex numbers. Then*

1. $\overline{z+w} = \overline{z} + \overline{w}$
2. $\overline{z-w} = \overline{z} - \overline{w}$
3. $\overline{zw} = \overline{z}\,\overline{w}$
4. $\overline{\left(\dfrac{z}{w}\right)} = \dfrac{\overline{z}}{\overline{w}}$
5. $\mathrm{Re}(z) = \dfrac{z+\overline{z}}{2}$
6. $\mathrm{Im}(z) = \dfrac{z-\overline{z}}{2i}$
7. $z\overline{z} = |z|^2$
8. *z is real if and only if $\overline{z} = z$.*
9. *z is pure imaginary or zero if and only if $\overline{z} = -z$.*

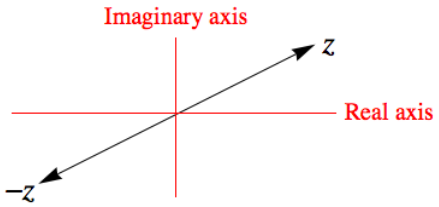
Geometric Interpretation Of Complex Numbers

Every complex number $z = a + ib$ can be represented by the vector (or point) (a, b) in the plane. The x- and the y-axis in this context is called the **real** and the **imaginary** axis, respectively. We have the following geometric interpretations.

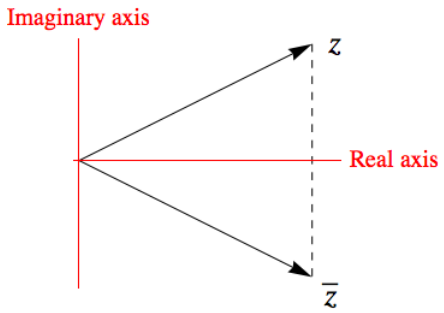
1. The absolute value $|z|$ is the length of vector z.



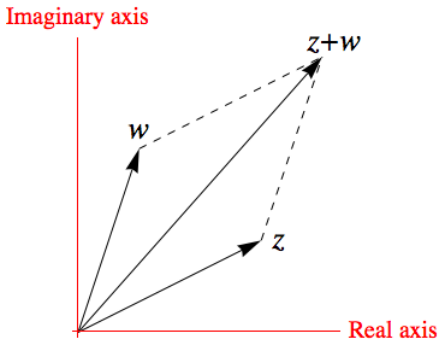
2. The opposite $-z$ of z is the reflection of z with respect to the origin.



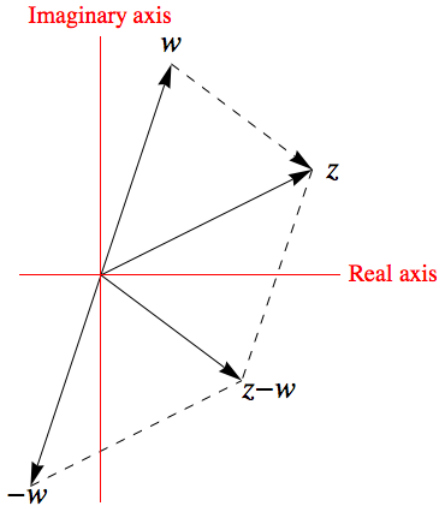
3. The conjugate $\overline{z}$ is the reflection with respect to the real axis.



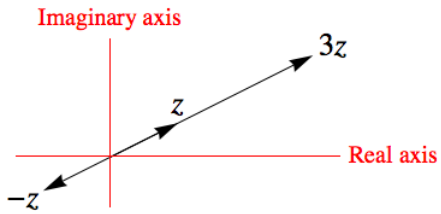
4. Addition of two complex numbers corresponds to vector addition in $\mathbf{R}^2$.



5. Subtraction $z - w$ of two complex numbers corresponds to the vector addition $z + (-w)$ in $\mathbf{R}^2$.



6. Scalar multiplication by a **real** number corresponds to scalar multiplication in $\mathbf{R}^2$.



5.2 Polar Form

Let (r, θ) be the polar coordinates of a point (x, y) of the complex plane. If $z = x + iy$, then we may write $z = r\cos(\theta) + ir\sin(\theta)$. Hence,

$$z = r(\cos(\theta) + i\sin(\theta)) \tag{5.1}$$

Equation (5.1) is called a **polar representation** of z.

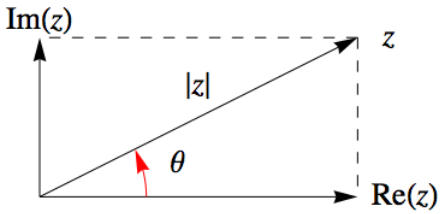
The non-negative number r is the distance of (x, y) from the origin. So

$$r = |z| = \sqrt{x^2 + y^2}$$

The angle θ is called the **argument** of z. We write

$$\theta = \arg(z)$$

See figure below.



The argument of z is multi-valued function. It is determined up to an integer multiple of 2π. Note that

$$\tan(\theta) = \frac{y}{x}$$

Note. The complex number 0 has no polar representation, because its argument is not well-defined.

If we restrict the values of the argument on the interval $(-\pi, \pi]$, then we get a single-valued function which is called the **principal value of the argument** of z. We write for this function $\mathrm{Arg}(z)$ and we have

$$-\pi < \mathrm{Arg}(z) \leq \pi$$

Example 5.2.1. Find the polar representation of $-1 + i$.

Solution: First, $|-1+i| = \sqrt{2}$. The argument of $-1+i$ can be computed from

$$\sqrt{2}\cos(\theta) = -1, \qquad \sqrt{2}\sin(\theta) = 1$$

which imply that $\theta = \frac{3\pi}{4} \pm 2\pi k$, $(k = 0, 1, 2, \dots)$. Hence,

$$-1+i = \sqrt{2}\left(\cos\left(\frac{3\pi}{4} \pm 2\pi k\right) + i\sin\left(\frac{3\pi}{4} \pm 2\pi k\right)\right)$$

If we use the principal argument where θ is restricted to values in $(-\pi, \pi]$, then we may write

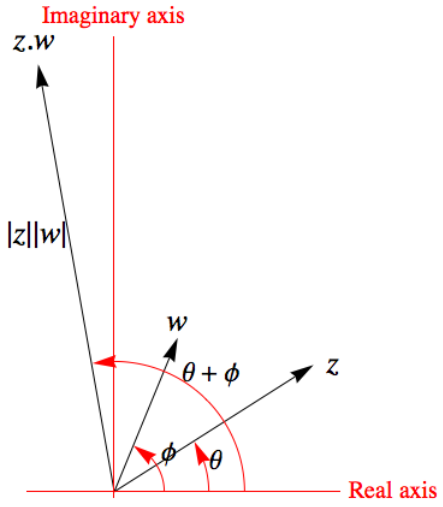
$$-1+i = \sqrt{2}\left(\cos\left(\frac{3\pi}{4}\right) + i\sin\left(\frac{3\pi}{4}\right)\right)$$

Multiplication and Division in Polar Form

If $z = r(\cos\theta + i\sin\theta)$ and $w = s(\cos\phi + i\sin\phi)$, then the product zw has polar representation

$$zw = rs(\cos(\theta+\phi) + i\sin(\theta+\phi)) \tag{5.2}$$

We see that the new length is the product of the lengths and the new argument is the sum of the arguments of z and w. (See figure below.)



Up to any integer multiple of 2π we have

$$\arg(zw) = \arg(z) + \arg(w)$$

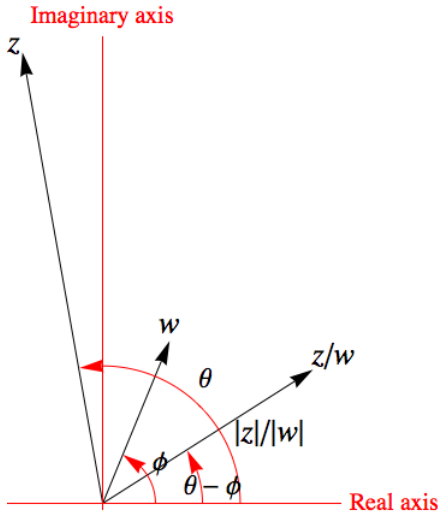
Equation (5.2) is proved by using the standard trigonometric identities as follows

$$\begin{aligned} zw &= rs(\cos\theta + i\sin\theta)(\cos\phi + i\sin\phi) \\ &= rs[(\cos\theta\cos\phi - \sin\theta\sin\phi) + i(\cos\theta\sin\phi + \cos\phi\sin\theta)] \\ &= rs(\cos(\theta+\phi) + i\sin(\theta+\phi)) \end{aligned}$$

By a similar calculation, we find that the polar representation of the quotient z/w is given by

$$\frac{z}{w} = \frac{r}{s}(\cos(\theta-\phi) + i\sin(\theta-\phi)), \qquad w \neq 0 \tag{5.3}$$

The length of the quotient is the quotient of the lengths and the new argument is the difference of the arguments of z and w.



A special case of equation (5.3) is the polar perpesentation of z^{-1}, $z \neq 0$. We have for $z = r(\cos(\theta) + i\sin(\theta))$

$$\frac{1}{z} = \frac{1}{r}(\cos(\theta) - i\sin(\theta)), \qquad z \neq 0 \tag{5.4}$$

The polar representation of the nth power is found by iterating (5.2) with $z = z_1 = z_2 = r\left(\cos\left(\theta\right) + i\sin\left(\theta\right)\right)$. We get

$$z^n = r^n(\cos\left(n\theta\right) + i\sin\left(n\theta\right)) \tag{5.5}$$

An interesting special case of equation (5.5) occurs when $z = \cos\left(\theta\right) + i\sin\left(\theta\right)$ in which case $|z| = 1$. Then we get **DeMoivre's Law**

$$\left(\cos\left(\theta\right) + i\sin\left(\theta\right)\right)^n = \cos\left(n\theta\right) + i\sin\left(n\theta\right), \qquad n = 1, 2, 3, \ldots$$

Example 5.2.2. Write each number in the form $a + ib$.

1. $(-1+i)^{10}$
2. $(-1+i)^{20}$

Solution: Recall from Example 5.2.1 that $-1+i = \sqrt{2}\left(\cos\left(\frac{3\pi}{4}\right) + i\sin\left(\frac{3\pi}{4}\right)\right)$. By equation (5.5), we have

1.
$$\begin{aligned}(-1+i)^{10} &= \left(\sqrt{2}\right)^{10}\left(\cos\left(10\cdot\frac{3\pi}{4}\right) + i\sin\left(10\cdot\frac{3\pi}{4}\right)\right)\\ &= 32\left(\cos\left(\frac{3\pi}{2}\right) + i\sin\left(\frac{3\pi}{2}\right)\right)\\ &= -32i\end{aligned}$$

2.
$$\begin{aligned}(-1+i)^{20} &= \left(\sqrt{2}\right)^{20}\left(\cos\left(20\cdot\frac{3\pi}{4}\right) + i\sin\left(20\cdot\frac{3\pi}{4}\right)\right)\\ &= 1024\left(\cos\left(\pi\right) + i\sin\left(\pi\right)\right)\\ &= -1024\end{aligned}$$

5.3 Roots

General nth Roots

Let z be a complex number. We seek all complex numbers w such that for a non-negative integer n we have

$$z = w^n \tag{5.6}$$

Such w is called an nth **root of** z and it is denoted by

$$w = \sqrt[n]{z}$$

To find such w we start with the polar representation of $z = r(\cos\theta + i\sin\theta)$. Let w have unknown polar representation $w = \rho(\cos\phi + i\sin\phi)$. Then by DeMoivre' Law (5.6) yields

$$\begin{aligned} r(\cos\theta + i\sin\theta) &= \rho^n(\cos\phi + i\sin\phi)^n \\ &= \rho^n(\cos(n\phi) + i\sin(n\phi)) \end{aligned}$$

By taking absolute values and noting that $|\cos\alpha + i\sin\alpha| = 1$, we have

$$r = \rho^n \tag{5.7}$$

Hence,

$$\cos\theta + i\sin\theta = \cos(n\phi) + i\sin(n\phi) \tag{5.8}$$

By (5.7) we have

$$\rho = \sqrt[n]{r}$$

This is the unique non-negative nth root of the non-negative number r. Also, (5.8) implies

$$n\phi = \theta + 2k\pi, \quad k = 0, \pm1, \pm2, \ldots$$

Hence,

$$\phi = \frac{\theta + 2k\pi}{n}, \quad k = 0, 1, \ldots, n-1$$

Note that only the values of k in $\{0, 1, 2, \ldots, n-1\}$ matter because these values give distinct values of ϕ that do not differ by $2\pi j$ where j is an integer.

We conclude that each complex number z has n distinct nth roots $\sqrt[n]{z}$ given by

$$w_k = \sqrt[n]{r}\left(\cos\left(\frac{\theta + 2k\pi}{n}\right) + i\sin\left(\frac{\theta + 2k\pi}{n}\right)\right), \quad k = 0, 1, \ldots, n-1$$

or

$$\sqrt[n]{z} = \sqrt[n]{|z|}\left(\cos\left(\frac{\arg(z) + 2k\pi}{n}\right) + i\sin\left(\frac{\arg(z) + 2k\pi}{n}\right)\right), \quad k = 0, \ldots, n-1$$

Note. All the nth roots of a complex number z have the same absolute value, namely, $\sqrt[n]{|z|}$. So all the roots lie on the circle centered at the origin of radius $\sqrt[n]{|z|}$ and they define a regular n-gon inscribed in that circle.

Example 5.3.1. Find the fourth roots of $1 - i$.

Solution: We have

$$1 - i = \sqrt{2}\left(\cos\frac{7\pi}{4} + i\sin\frac{7\pi}{4}\right)$$

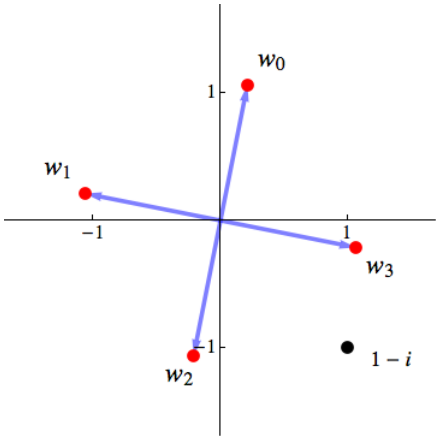
Hence,

$$\sqrt[4]{1-i} = \sqrt[8]{2}\left(\cos\left(\frac{7\pi/4 + 2k\pi}{4}\right) + i\sin\left(\frac{7\pi/4 + 2k\pi}{4}\right)\right), \quad k = 0, 1, 2, 3$$

We get the following 4th roots.

$$\begin{aligned}
w_0 &= \sqrt[8]{2}\left(\cos\left(\frac{7\pi}{16}\right) + i\sin\left(\frac{7\pi}{16}\right)\right) \simeq 0.212\,75 + 1.069\,6i \\
w_1 &= \sqrt[8]{2}\left(\cos\left(\frac{15\pi}{16}\right) + i\sin\left(\frac{15\pi}{16}\right)\right) \simeq -1.069\,6 + 0.212\,75i \\
w_2 &= \sqrt[8]{2}\left(\cos\left(\frac{23\pi}{16}\right) + i\sin\left(\frac{23\pi}{16}\right)\right) \simeq -0.212\,75 - 1.069\,6i \\
w_3 &= \sqrt[8]{2}\left(\cos\left(\frac{31\pi}{16}\right) + i\sin\left(\frac{31\pi}{16}\right)\right) \simeq 1.069\,6 - 0.212\,75i
\end{aligned}$$

These four 4th roots of $1 - i$ are plotted in the graph below.



The Roots of Unity

In the special case where $z = 1$ we have the nth roots of 1 or the nth **roots of unity**. Since as a real number $\sqrt[n]{|1|} = 1$, the nth roots of unity are the n complex numbers given by the formulas

$$\sqrt[n]{1} = \left(\cos\left(\frac{2k\pi}{n}\right) + i\sin\left(\frac{2k\pi}{n}\right)\right), \quad k = 0, 1, 2, \ldots, \ n-1$$

The nth roots unity form a regular n-gon inscribed on the unit circle.

Example 5.3.2. Find the sixth roots of unity.

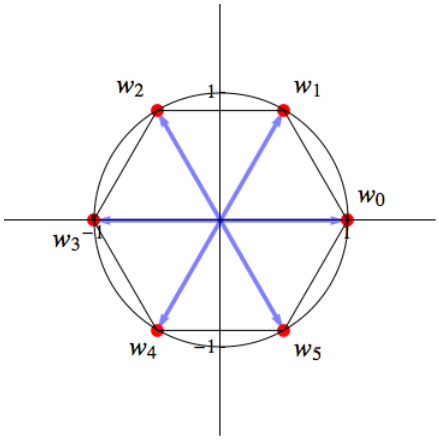
Solution: We have

$$\sqrt[6]{1} = \cos\left(\frac{2k\pi}{6}\right) + i\sin\left(\frac{2k\pi}{6}\right), \quad k = 0, 1, 2, 3, 4, 5$$

We get the following six 6th roots of unity.

$$\begin{aligned}
w_0 &= \cos(0) + i\sin(0) \simeq 1 \\
w_1 &= \cos\left(\frac{\pi}{3}\right) + i\sin\left(\frac{\pi}{3}\right) = \frac{1}{2} + \frac{\sqrt{3}}{2}i \\
w_2 &= \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right) = -\frac{1}{2} + \frac{\sqrt{3}}{2}i \\
w_3 &= \cos(\pi) + i\sin(\pi) = -1 \\
w_4 &= \cos\left(\frac{4\pi}{3}\right) + i\sin\left(\frac{4\pi}{3}\right) = -\frac{1}{2} - \frac{\sqrt{3}}{2}i \\
w_5 &= \cos\left(\frac{5\pi}{3}\right) + i\sin\left(\frac{5\pi}{3}\right) = \frac{1}{2} - \frac{\sqrt{3}}{2}i
\end{aligned}$$

The graph below shows that these roots form a regular hexagon inscribed on the unit cicrle.



5.4 Basic Regions in the Complex Plane

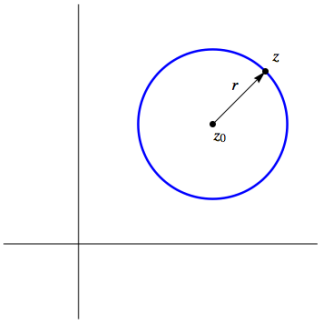
In this section we consider various subsets of the complex plane that will play an important role in what follows.

Let $S \subseteq \mathbf{C}$. The **complement**, S^c, of S in $\mathbf{C}$ consists of all complex numbers not in S.

Circles, Disks, Annuli

A **circle**, $C(z_0, r)$, centered at the complex number z_0 of radius r $(r > 0)$, is the complex numbers z whose distance from z_0 is r. Hence, we have

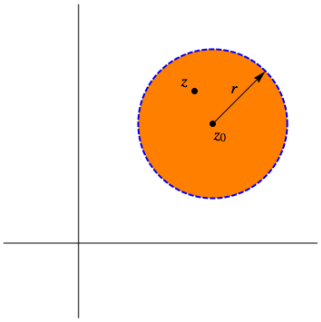
$$C(z_0, r) = \{z \in \mathbf{C}, \quad |z - z_0| = r\}$$



The **open disk**, $D(z_0, r)$, centered at z_0 and of radius r, is all z whose

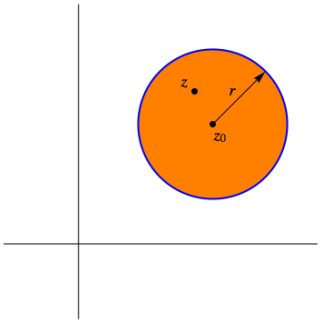
distance from z_0 is less than r. So, we have

$$D\,(z_0, r) = \{z \in \mathbf{C}, \quad |z - z_0| < r\}$$



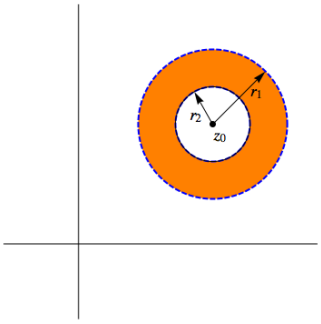
Likewise, we define the closed disk $\overline{D}\,(z_0, r)$ centered at z_0 and of radius r, by

$$\overline{D}\,(z_0, r) = \{z \in \mathbf{C}, \quad |z - z_0| \leq r\}$$

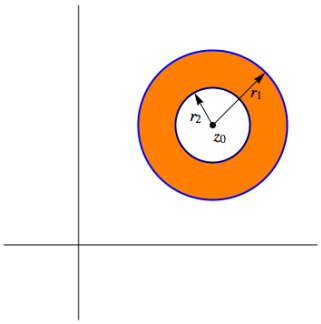


An **annulus** is the region in $\mathbf{C}$ between two concentric circles. An **open annulus**, $A\,(z_0, r_1, r_2)$, is defined by

$$A\,(z_0, r_1, r_2) = \{z \in \mathbf{C}, \quad r_1 < |z - z_0| < r_2\} \qquad (r_1 < r_2)$$



Similarly, we may define a closed or a half-open annulus by using $\leq$, instead of $<$.



We also have **punctured disks** where the center of the disk is missing. For example, the following set defines an open punctured disk.

$$\{z \in \mathbf{C}, \quad 0 < |z - z_0| < r\}$$

Vertical and Horizontal Half Planes and Strips

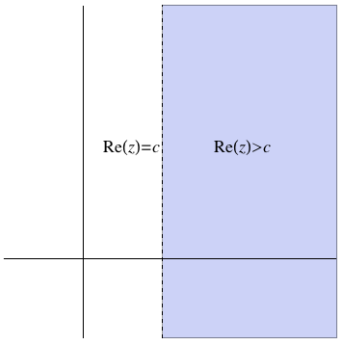
A vertical half-plane is the set of complex numbers to the left or to the right of a vertical line. A vertical line is the set of complex numbers with the same real part, say $\mathrm{Re}\,(z) = c$, where c is a real constant. So, an open left vertical half-plane may be described by

$$\{z \in \mathbf{C}, \quad \mathrm{Re}\,(z) < c\}$$

for some real constant c.

Likewise, we have an open right vertical half-plane

$$\{z \in \mathbf{C}, \quad \mathrm{Re}\,(z) > c\}$$

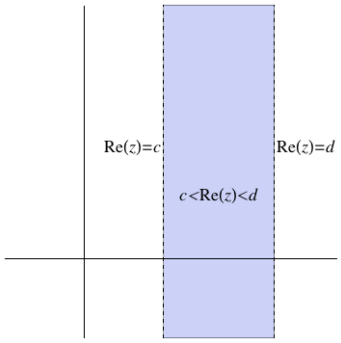


and also the corresponding closed half-planes left and right, respectively:

$$\{z \in \mathbf{C}, \quad \mathrm{Re}\,(z) \leq c\} \qquad \text{and} \qquad \{z \in \mathbf{C}, \quad \mathrm{Re}\,(z) \geq c\}$$

We may also have vertical strips defined as areas between two vertical half-planes. For example,

$$\{z \in \mathbf{C}, \quad c < \mathrm{Re}\,(z) < d\} \qquad c < d)$$

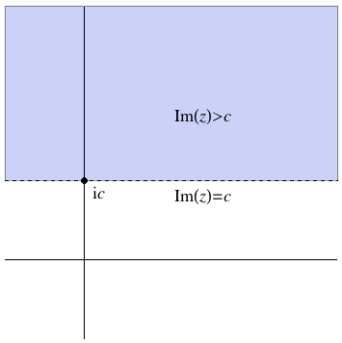


In addition to the vertical half-planes, we have horizontal upper and lower half-planes defined by horizontal staight lines. Such lines have defining equation $\mathrm{Im}\,(z) = c$. So, these planes are given by

$$\{z \in \mathbf{C}, \quad \mathrm{Im}\,(z) < c\} \qquad \text{and} \qquad \{z \in \mathbf{C}, \quad \mathrm{Im}\,(z) > c\}$$

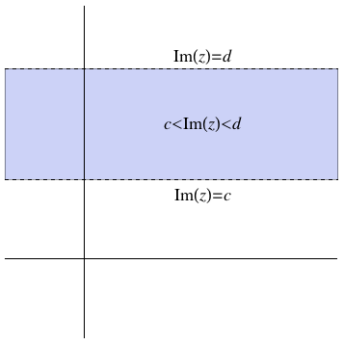
and also by

$$\{z \in \mathbf{C}, \quad \mathrm{Im}\,(z) \leq c\} \qquad \text{and} \qquad \{z \in \mathbf{C}, \quad \mathrm{Im}\,(z) \geq c\}$$



Finally, we may consider horizontal strips defined as areas between two horizontal half-planes. For example,

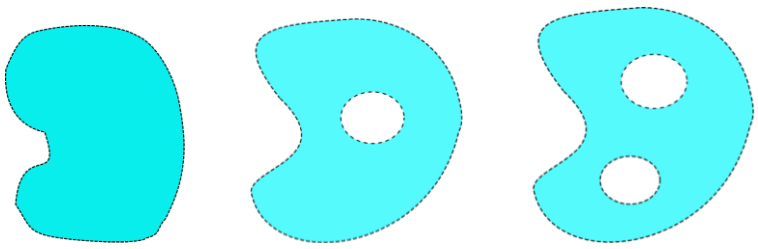
$$\{z \in \mathbf{C}, \quad c < \operatorname{Im}(z) < d\} \qquad c < d)$$



Open, Closed, and Connected Sets

Let z be a complex number. A **neighborhood** of z is a disk centered at z. We have an **open neighborhood**, if the disk is an open disk.[1]

Definition A subset S of $\mathbf{C}$ is called **open**, if each point in S has an open neighborhood that lies entirely in S.



Examples of open sets: open disk, open half-plane, open strip, the complement of closed disk, and all of $\mathbf{C}$. In addition, an arbitrary union of open sets is open and a finite intersection of open sets is open. The empty set, $\emptyset$, is considered to be open.

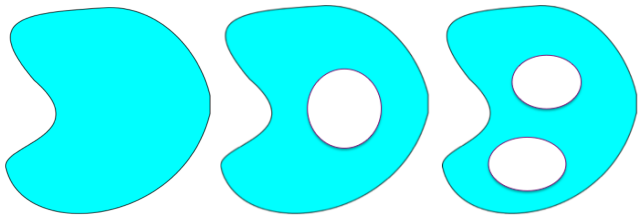
Non-examples of open sets: The following sets are not open: point, straight line, circle, curve in general, closed disk, the complement of open disk,

[1] There is a more general definition of a neighborhood. S is a neighborhood of a complex number z, if S contains a disk centered at z.

and closed half-plane. Visually, any set that contains a form of boundary is not an open set. The se $D(0,1) \cup \{z \in \mathbf{C}, \ \ |z| = 1, \ \ \mathrm{Im}(z) \geq 0\}$ that is the union of the unit open disk with the upper unit semicircle is not open.

Definition A subset S of $\mathbf{C}$ is called **closed**, if its complement S^c is open.

Examples of closed sets: point, closed disk, closed half-plane, closed strip, the complement of open disk, straight line, circle, curve in general, the empty set, $\emptyset$, and all of $\mathbf{C}$. In addition, an arbitrary intersection of closed sets is closed and a finite union of closed sets is closed.



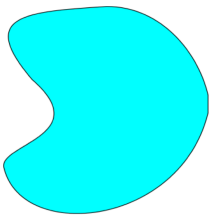
Non-examples of closed sets: The following sets are not closed: open disk, open half-plane, open strip, the complement of closed disk. The set $D(0,1) \cup \{z \in \mathbf{C}, \ \ |z| = 1, \ \ \mathrm{Im}(z) \geq 0\}$ is not closed.

Note It is a common misconception that if a set is not open, then it must be closed. This is incorrect: The set $D(0,1) \cup \{z \in \mathbf{C}, \ \ |z| = 1, \ \ \mathrm{Im}(z) \geq 0\}$ is neither open nor closed. Also, $\emptyset$ and $\mathbf{C}$ are both open and closed.

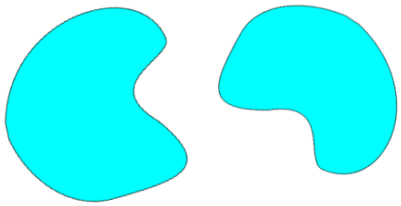
Definition A subset S of $\mathbf{C}$ is called **(path) connected**, if any points in S

can be connected by a path that lies entirely in S. If a set is not connected, then it is the disjoint union of subsets that are connected. These are called the **(path) connected components**.

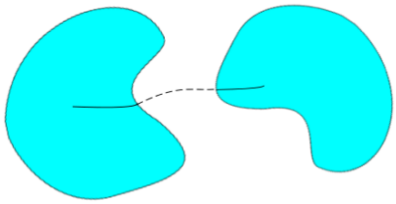
The following set is connected.



The next set is not connected. It has two connected components.



The connected components cannot be connected by a path that lies entirely in the set.



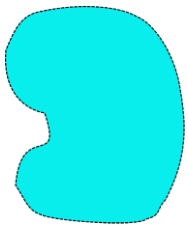
Definition A subset S of $\mathbf{C}$ is called a **domain**, if it is open and connected.

Note The above definition of domain should not be confused with the concept of the domain of a function.

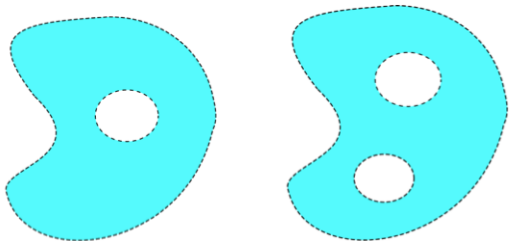
A domain D in $\mathbf{C}$ is called **simply connected** or **1-connected** , if for every simple closed curve that is in D, its interior is also in D.

Intuitively, a simply connected domain is a domain without any "holes". For example, an open disk is simply connected. A punctured open disk is not simply connected. An open annulus is not simply connected.

The domain below is simply connected



while the following two are is not simply connected.



The first of the above is a 2-connected or **doubly-connected** domain (one hole) and the second is 3-connected or **triply-connected** (two holes). In general, an ***n*-connected domain** has $n-1$ holes.

5.5 Limits and Continuity

Definition of Limit; Examples

Let $S \subseteq \mathbf{C}$. The complex number z_0 is called a **limit point** of S, if each neighborhood of z_0 contains points of S other than z_0. So, points of S come arbitrarily close to z_0. Note that z_0 may, or may not be in S.

Let $f(z)$ be a complex valued function of one complex variable z. So f is of the form $f : S \to \mathbf{C}$, where $S \subseteq \mathbf{C}$. Let z_0 be a limit point of S. The complex number L is said to be the **limit** of the function $f(z)$ as z approaches a complex number z_0, if for each neighborhood $D(L, \varepsilon)$ of L, there is a neighborhood $D(z_0, \delta)$ of z_0 that, with the possible exception of z_0, $f(z)$ maps $D(z_0, \delta)$ into $D(L, \varepsilon)$. That is, for each $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that

$$\text{for} \quad z \in S \quad \text{and} \quad 0 < |z - z_0| < \delta \Longrightarrow |f(z) - L| < \varepsilon$$

If such L exists, we write

$$\lim_{z \to z_0} f(z) = L$$

Sometimes we may write

$$f(z) \to L \quad \text{as} \quad z \to z_0$$

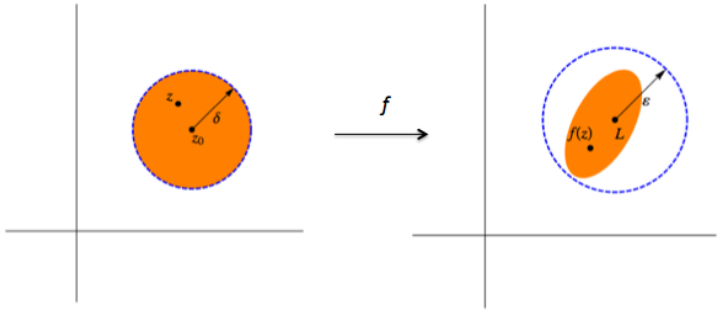
The above definition expresses, precisely, the following intuitive statement: As z approaches z_0 (without actually becoming z_0), then the values $f(z)$ approach L.

Notes

1. The limit of $f(z)$ as $z \to z_0$ may or may not exist. However, if it exists, it is unique. In particular, the limit is independent of the direction along which z approaches z_0.

2. In the definition of limit, we do not require that $f(z)$ is defined at z_0.

3. We want z_0 to be a limit point of S so that the punctured disk $0 < |z - z_0| < \delta$ always contains points z of the domain S of f, thus, $f(z)$ is defined in the inequality $|f(z) - L| < \varepsilon$.



Example 5.5.1. Use the definition of limit to prove that

$$\lim_{z \to 2i} z^2 = -4$$

Solution: For a given $\varepsilon > 0$ we need $\delta > 0$ such that

$$\left|z^2 - (-4)\right| < \varepsilon \quad \text{whenever} \quad 0 < |z - 2i| < \delta$$

Note that if we rewrite

$$\left|z^2 - (-4)\right| = |(z - 2i)(z + 2i)| = |(z - 2i)((z - 2i) + 4i)|$$

By the triangle inequality we have

$$|((z - 2i) + 4i)| \leq |z - 2i| + |4i| = |z - 2i| + 4$$

Hence,

$$\left|z^2 - (-4)\right| \leq |z - 2i|(|z - 2i| + 4)$$

Now if we choose $\delta = \min(1, \varepsilon/5)$, then

$$\left|z^2 - (-4)\right| \leq \frac{\varepsilon}{5}(1 + 4) = \varepsilon$$

Example 5.5.2. We have

$$\lim_{z\to i} f(z) = \lim_{z\to i} g(z) = \lim_{z\to i} h(z) = 2i$$

where,

1. $f(z) = 3i - z$
2. $g(z) = \begin{cases} 3i - z, & z \neq i \\ 1 + 2i, & z = i \end{cases}$
3. $h(z) = \dfrac{z^2 + 1}{z - i}$

Solution Exercise.

Properties of Limits

Theorem 5.5.1. *Let $f(z) = u(x,y) + iv(x,y)$, where $z = x + iy$ and let $z_0 = x_0 + iy_0$. If*

$$\lim_{z\to z_0} f(z) = L$$

then

$$\lim_{(x,y)\to(x_0,y_0)} u(x,y) = \operatorname{Re}(L) \quad \text{and} \quad \lim_{(x,y)\to(x_0,y_0)} v(x,y) = \operatorname{Im}(L)$$

Conversely, if

$$\lim_{(x,y)\to(x_0,y_0)} u(x,y) = a \quad \text{and} \quad \lim_{(x,y)\to(x_0,y_0)} v(x,y) = b$$

then

$$\lim_{z\to x_0+iy_0} f(z) = a + ib$$

Theorem 5.5.2. *If the following limits exist*

$$\lim_{z\to z_0} f_1(z) = L_1 \qquad \text{and} \qquad \lim_{z\to z_0} f_2(z) = L_2$$

then the following limits exist and

1. $\lim_{z\to z_0} [f_1(z) + f_2(z)] = L_1 + L_2$

2. $\lim_{z\to z_0} [f_1(z) - f_2(z)] = L_1 - L_2$

3. $\lim_{z\to z_0} f_1(z) f_2(z) = L_1 L_2$

4. $\lim_{z\to z_0} \dfrac{f_1(z)}{f_2(z)} = \dfrac{L_1}{L_2}$ (if $L_2 \neq 0$)

Limits with Infinity

The concept of $\lim_{z\to z_0} f(z)$ can be extended to the case where z_0 is the point at infinity. We define

$$\lim_{z\to\infty} f(z) = L$$

as the number L, if it exists, with the following property: For each positive each $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that

$$\text{for } \ z \in \text{Domain}(f) \quad \text{and} \quad |z| > \frac{1}{\delta} \Longrightarrow |f(z) - L| < \varepsilon$$

In other words, we basically have the same definition as before but we replace the corresponding neighborhood of z_0 by the neighboorhood of ∞ (i.e., all z such that $|z| > M$, where $M > 0$).

We may also give meaning to the expression $\lim_{z\to z_0} f(z) = \infty$. We write

$$\lim_{z\to z_0} f(z) = \infty$$

if for every $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that

$$\text{for } \ z \in \text{Domain}(f) \quad \text{and} \quad 0 < |z - z_0| < \delta \Longrightarrow |f(z)| > \frac{1}{\varepsilon}$$

Theorem 5.5.3. *We have*

1. $\lim_{z\to\infty} f(z) = L$ *if and only if* $\lim_{z\to 0} f\left(\dfrac{1}{z}\right) = L$.

2. $\lim_{z\to z_0} f(z) = \infty$ *if and only if* $\lim_{z\to z_0} \dfrac{1}{f(z)} = 0$.

Continuous Functions

The complex function $f(z)$ is **continuous** at z_0, if

$$\lim_{z \to z_0} f(z) = f(z_0)$$

In this definition two things are implicit: that the limit exists and that z_0 is in the domain of $f(z)$.

A complex function $f(z)$ is **continuous** in a complex region R, if it continuous at every point of R.

Remark Note that

1. $f(z) = u(x,y) + iv(x,y)$ $(z = x + iy)$ is continuous at $z_0 = x_0 + iy_0$ if and only if $u(x,y)$ and $v(x,y)$ are each continuous at (x_0, y_0).

2. Sums, differences, products, quotients, and compositions of continuous functions are also continuous.

Examples of continuous functions: polynomials, rational functions in their domain of definition (i.e., quotients of polynomials $f(z) = \dfrac{g(z)}{h(z)}$ for all z such that $h(z) \neq 0$), exponential functions, and trigonometric functions.[2]

5.6 Differentiable Functions

The Derivative

Let $f(z)$ be a complex function defined in a neighborhood of a point z_0. The **derivative** of $f(z)$ at z_0 , denoted by $f'(z_0)$ or by $\dfrac{df}{dz}(z_0)$, is

$$\begin{aligned} f'(z_0) &= \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \\ &= \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}, \qquad \Delta z = z - z_0 \end{aligned}$$

if this limit exists. In this case we say that $f(z)$ is **differentiable** at z_0. Note that value of the limit must be independent of the path of z approaching z_0.

[2] Complex exponential and trigonometric functions will be introduced later in the notes.

Example 5.6.1. Show that $f(z) = z^2 - 3z + 1$ is differentiable at any point z_0 and compute $f'(z_0)$.

Solution

$$\begin{aligned} f'(z_0) &= \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \\ &= \lim_{\Delta z \to 0} \frac{\left((z_0 + \Delta z)^2 - 3(z_0 + \Delta z) + 1\right) - (z_0^2 - 3z_0 + 1)}{\Delta z} \\ &= \lim_{\Delta z \to 0} \frac{(2z_0 + \Delta z - 3)\Delta z}{\Delta z} \\ &= \lim_{\Delta z \to 0} (2z_0 + \Delta z - 3) \\ &= 2z_0 - 3 \end{aligned}$$

The limit exists for all z_0, so $f(z)$ is differential everywhere and $f'(z_0) = 2z_0 - 3$.

Examples of differentiable functions: polynomials $p(z)$ for all z, rational functions for all z in their domain of definition, exponential functions for all z, and trigonometric functions in their domains of definition.

Example 5.6.2. Show that $f(z) = \overline{z}$ is nowhere differentiable.

Solution:

$$\begin{aligned} f'(z_0) &= \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \\ &= \lim_{\Delta z \to 0} \frac{\overline{z_0 + \Delta z} - \overline{z_0}}{\Delta z} \\ &= \lim_{\Delta z \to 0} \frac{\overline{\Delta z}}{\Delta z} \end{aligned}$$

The last limit does not exist because the values of the quotient depend on the path approaching z_0. If $\Delta z = \Delta x + i\Delta y$ approaches zero via a path in the direction parallel to the y-axis, i.e., that such that $\Delta y = 0$. Then as $\Delta x \to 0$, the quotient $\frac{\overline{\Delta z}}{\Delta z}$ approaches the value 1. However, if a path is chosen such that $\Delta x = 0$, then as $\Delta y \to 0$, the quotient $\frac{\overline{\Delta z}}{\Delta z}$ approaches the value -1. We conlude that the function $f(z) = \overline{z}$ is nowhere differentiable.

Example 5.6.3. Show that $f(z) = |z|^2$ is nowhere differentiable except at the origin, where $f'(0) = 0$.

Solution: Exercise.

Remark If $f(z)$ is differentiable at z_0, then $f(z)$ is continuous at z_0. The converse is not true.

The basic rules of differentiation of real functions are also valid for complex functions.

Theorem 5.6.1. *Sums, differences, products, quotients, and compositions of differentiable functions are also differentiable. More precisely, let $f(z)$ and $g(z)$ be differentiable at z_0. Then*

1. $(f+g)'(z_0) = f'(z_0) + g'(z_0)$ *(Sum Rule)*
2. $(f-g)'(z_0) = f'(z_0) - g'(z_0)$ *(Difference Rule)*
3. $(fg)'(z_0) = f'(z_0)\,g(z_0) + f(z_0)\,g'(z_0)$ *(Product Rule)*
4. $\left(\dfrac{f}{g}\right)'(z_0) = \dfrac{f'(z_0)\,g(z_0) - f(z_0)\,g'(z_0)}{g(z_0)^2}$ *(Quotient Rule)*
5. *If $h(z)$ is differentiable at $f(z_0)$, then*

$$(h \circ f)'(z_0) = h'(f(z_0))\,f'(z_0)$$

(Chain Rule)

The Cauchy-Riemann Equations

If $f(z)$ is differentiable at z_0, then the limit defining $f'(z_0)$ is independent of the path along which Δz approaches 0. So, we

1. first, take $\Delta z \to 0$ in a direction parallel to the x-axis. Hence, we may assume that $\Delta z = \Delta x$. Then

   $$f(z_0+\Delta z) - f(z_0) = u(x_0+\Delta x, y_0) + iv(x_0+\Delta x, y_0) - u(x_0,y_0) - iv(x_0,y_0)$$

   Therefore,

$$\begin{aligned} f'(z_0) &= \lim_{\Delta x \to 0} \frac{u(x_0+\Delta x, y_0) - u(x_0,y_0)}{\Delta x} \\ &\quad + i \lim_{\Delta x \to 0} \frac{v(x_0+\Delta x, y_0) - v(x_0,y_0)}{\Delta x} \\ &= \frac{\partial u}{\partial x}(x_0,y_0) + i\frac{\partial v}{\partial x}(x_0,y_0) \end{aligned}$$

2. next, take $\Delta z \to 0$ in a direction parallel to the y-axis. So, we may assume that $\Delta z = i\Delta y$. Then

   $$f(z_0 + \Delta z) - f(z_0) = u(x_0, y_0 + \Delta y) + iv(x_0, y_0 + \Delta y) - u(x_0, y_0) - iv(x_0, y_0)$$

   Therefore,

   $$\begin{aligned} f'(z_0) &= \lim_{\Delta x \to 0} \frac{u(x_0, y_0 + \Delta y) - u(x_0, y_0)}{i\Delta y} \\ &\quad + i \lim_{\Delta x \to 0} \frac{v(x_0, y_0 + \Delta y) - v(x_0, y_0)}{i\Delta y} \\ &= \frac{1}{i}\frac{\partial u}{\partial y}(x_0, y_0) + \frac{\partial v}{\partial y}(x_0, y_0) \end{aligned}$$

If we combime the two different forms of $f'(z_0)$, we conclude that

$$f' = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{1}{i}\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}$$

Equating real and imaginary parts yields

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} \tag{5.9}$$

These two equations are called the **Cauchy-Riemann equations** of $f(z)$. They are necessary conditions for the existence of the derivative of a complex function. Are they also necessary conditions? In other words, if the first partial derivatives of u and v exist and satisfy the Cauchy-Riemann equations, does this imply that $f(z)$ is differentiable? The answer is "almost". We need an extra assumption: that the first partial derivatives are also continuous.

Theorem 5.6.2. *Let $f(z) = u(x, y) + iv(x, y)$ be a complex function.*

1. *If $f(z)$ is differentiable, then the first partials of $u(x, y)$ and $v(x, y)$ exist and satisfy the Cauchy-Riemann equations*

   $$u_x = v_y \quad \text{and} \quad v_x = -u_y$$

2. *If the partials u_x, u_v, v_x, v_y exist and are continuous and satisfy the Cauchy-Riemann equations*

   $$u_x = v_y \quad \text{and} \quad v_x = -u_y$$

   then $f(z)$ is differentiable.

5.7 Analytic Functions

Definition and Examples

A complex valued function $f(z)$ is called **analytic** at the point z_0, if $f'(z)$ exists in a neighborhood of z_0. Another word for analytic is **holomorphic** .

Analytic functions are the most important class of functions in complex variables.

Remark If the function $f(z)$ is analytic at z_0, then it is differentiable at z_0. This is because z_0 is in each of its neighborhoods. However, if a function is differentible at z_0, it is not necessarily analytic at z_0. For example, $f(z) = |z|^2$ is differentiable at 0, but not analytic at 0, because the derivative $f'(z)$ does not exist for $z \neq 0$. (See Example 5.6.3.)

A function $f(z)$ is called **analytic in an open set** D of the complex plane, if it is analytic at all $z \in D$. If $f(z)$ is analytic in all of $\mathbf{C}$, then it is called **entire**.

Note that since every point in an open set D has a neighborhood that lies enrtirely in D, then $f(z)$ is analytic in the open set D, if $f'(z)$ exists for all $z \in D$.

Examples of analytic functions: polynomials $p(z)$ for all z, rational functions for all z in their domain of definition, exponential functions for all z, and trigonometric functions in their domains of definition.

The Cauchy-Riemann equations can be used as a very important test for analyticity. We have the following theorem.

Theorem 5.7.1 (Test for Analyticity)**.** *Let $u(x,y)$, and $v(x,y)$ be real-valued functions in two real variables x and y defined in a domain D and let $f(z)$ be the complex function $f(z) = u(x,y) + iv(x,y)$ (where $z = x + iy$).*

1. *If $f(z)$ is analytic in D, then the first partials of $u(x,y)$ and $v(x,y)$ exist in D and satisfy the Cauchy-Riemann equations*

$$u_x = v_y \quad \text{and} \quad v_x = -u_y$$

2. *If the partials u_x, u_v, v_x, v_y exist in D and are continuous in D and also satisfy the Cauchy-Riemann equations in D*

$$u_x = v_y \quad \text{and} \quad v_x = -u_y$$

then $f(z)$ is analytic in D.

Example 5.7.1. Use the Cauchy-Riemann equations to check the functions for analyticity.

1. $f(z) = z^3$
2. $g(z) = \overline{z}$
3. $h(z) = \mathrm{Re}(z)$
4. $k(z) = \mathrm{Im}(z)$
5. $q(z) = \dfrac{1}{z}$

Solution:

1. We have $f(z) = z^3 = (x+iy)^3 = (x^3 - 3xy^2) + i(3x^2y - y^3)$. Hence, $u(x,y) = x^3 - 3xy^2$ and $v(x,y) = 3x^2y - y^3$. Now

$$u_x = 3x^2 - 3y^2 = v_y \qquad \text{and} \quad v_x = 6xy = -u_y$$

for all x and y so $f(z)$ is analytic everywhere (thus, entire).

2. We have $g(z) = \overline{z} = x - iy$. Hence, $u(x,y) = x$ and $v(x,y) = -y$. Now

$$u_x = 1 \neq v_y = -1$$

So $g(z)$ is not analytic anywhere.

3. We have $h(z) = x$. Hence, $u(x,y) = x$ and $v(x,y) = 0$. Now

$$u_x = 1 \neq v_y = 0$$

So $h(z)$ is not analytic anywhere.

4. It is left as exercise to show that $k(z)$ is not analytic anywhere.

5. We have $q(z) = \dfrac{1}{z} = \dfrac{1}{x+iy} = \dfrac{x}{x^2+y^2} + i\dfrac{-y}{x^2+y^2}$. Hence, $u(x,y) = \dfrac{x}{x^2+y^2}$ and $v(x,y) = \dfrac{-y}{x^2+y^2}$. We leave it to the reader to show that the Cauchy-Riemann equations are satisfied everywhere except at the origin. Hence $q(z)$ is analytic in the domain $\mathbf{C} - \{0\}$.

Singular Points

The complex number z_0 is called a **singular point** or a **singularity** of a complex function $f(z)$, if f is not analytic at z_0, but every neighborhood of z_0 contains at least one point at which $f(z)$ is analytic.

Example 5.7.2. We have

1. $\dfrac{1}{1-z}$ has one singularity at $z = 1$.
2. $\dfrac{1}{z^2+1}$ has two singularities, at $z = i$ and $z = -i$.
3. $\tan z$ has infinitely many singularities, at $z = \dfrac{(2n-1)\pi}{2}$ $(n \in \mathbf{Z})$.

Laplace's Equation

Theorem 5.7.2. *If the complex function $f(z) = u(x,y) + iv(x,y)$ is analytic in a domain D, then $u(x,y)$ and $v(x,y)$ satisfy **Laplace's equation**.*

$$\nabla^2 u = u_{xx} + u_{yy} = 0 \qquad \text{and} \qquad \nabla^2 v = v_{xx} + v_{yy} = 0$$

in D and have continuous second partial derivatives in D.

Proof. (Sketch of proof) Since $f(z)$ is analytic, the Cauchy-Riemann equations are satisfied. Hence, $u_x = v_y$, which implies $u_{xx} = v_{yx}$. Likewise, $u_y = -v_x$ implies $u_{yy} = -v_{xy}$. Adding these yields $u_{xx} + u_{yy} = 0$. The equation $v_{xx} + v_{yy} = 0$ is proved similarly. □

5.8 Exponential Function

Definition

Let $z = x + iy$ be any complex number. We define the **complex exponential function** e^z, also denoted by $\exp(z)$, as

$$e^z = e^x(\cos y + i \sin y)$$

Note

1. If $z = x$ is real, then e^z reduces to the real exponetial e^x.

2. If $z = iy$ is pure imaginary, then e^z reduces to **Euler's formula**

$$e^{iy} = \cos y + i \sin y$$

The last formula shows that the polar form of $z = r(\cos\theta + i\sin\theta)$ can be aslo written as

$$z = re^{i\theta} \qquad (r = |z|,\ \theta = \arg(z))$$

Example 5.8.1. Prove each of the following.

1. $e^{2\pi i} = 1$
2. $e^{\pi i} = -1$
3. $e^{\pi i/2} = i$
4. $e^{3\pi i/2} = -i$
5. $e^{1.3-0.5i} \simeq 3.2201 - 1.7592i$

Solution: We have

1. $$e^{2\pi i} = \cos(2\pi) + i\sin(2\pi) = 1 + 0i = 1$$
2. $$e^{\pi i} = \cos(\pi) + i\sin(\pi) = -1 + 0i = -1$$
3. $$e^{\pi i/2} = \cos(\pi/2) + i\sin(\pi/2) = 0 + 1i = i$$
4. $$e^{3\pi i/2} = \cos(3\pi/2) + i\sin(3\pi/2) = 0 - 1i = -i$$
5. $$e^{1.3-0.5i} = e^{1.3}(\cos(-0.5) + i\sin(-0.5)) \simeq 3.2201 - 1.7592i$$

Properties of $\exp(z)$

Theorem 5.8.1. *The complex exponential e^z is entire and*

$$(e^z)' = e^z$$

Proof. We have

$$\begin{aligned}(e^z)' &= (e^x \cos y)_x + i\,(e^x \sin y)_x \\ &= e^x \cos y + ie^x \sin y \\ &= e^z\end{aligned}$$

□

Theorem 5.8.2. *We have the following properties.*

1. $e^{z+w} = e^z e^w$ *for all* z *and* w *in* $\mathbf{C}$.
2. $e^{z-w} = \dfrac{e^z}{e^w}$ *for all* z *and* w *in* $\mathbf{C}$.

Proof. We have

1. For $z = x_1 + iy_1$ and $w = x_2 + iy_2$ we have

$$\begin{aligned}e^{z+w} &= e^{x_1+x_2}\left[\cos(y_1+y_2) + i\sin(y_1+y_2)\right] \\ &= e^{x_1}e^{x_2}[\cos y_1 \cos y_2 - \sin y_1 \sin y_2 \\ &+ i\sin y_1 \cos y_2 + i \sin y_2 \cos y_1] \\ &= e^{x_1}\left(\cos y_1 + i \sin y_1\right) e^{x_2}\left(\cos y_2 + i \sin y_2\right) \\ &= e^z e^w\end{aligned}$$

2. The proof is similar to that of part 1.

□

Theorem 5.8.3. *We have the following properties.*

1. $|e^z| = e^{\mathrm{Re}(z)}$ *for all* z *in* $\mathbf{C}$.
2. $e^z \neq 0$ *for all* z *in* $\mathbf{C}$.

3. $|e^{it}| = 1$ *for all real* t.

4. $e^{2\pi ni} = 1$ *for all integers* n.

5. $e^{z+2\pi ni} = e^z$ *for all* z *in* $\mathbf{C}$ *and all integers* n.

6. *If* $e^z = 1$, *then* $z = 2\pi ni$ *for some integer* n.

Proof. Let $z = x + iy$.

1. We have

$$\begin{aligned} |e^z| &= |e^x|\,|\cos y + i \sin y| \\ &= e^x \sqrt{\cos^2 y + \sin^2 y} \\ &= e^x \end{aligned}$$

2. Since $|e^z| = e^x \neq 0$, we have that $e^z \neq 0$.

3. By part 1, $|e^{it}| = e^0 = 1$.

4. $$e^{2\pi ni} = \cos(2\pi n) + i \sin(2\pi n) = 1 + 0i = 1$$

5. By part 4, we have

$$e^{z+2\pi ni} = e^z e^{2\pi ni} = e^z 1 = e^z$$

6. Let $e^z = 1$. Then $|e^z| = e^x = 1$. Hence, $x = 0$. Therefore, $e^z = \cos y + i \sin y = 1$. Hence, $\sin y = 0$ and $\cos y = 1$. So it must be that $y = 2\pi n$ for some integer n. Thus, $z = 0 + i(2\pi n) = 2\pi ni$ as claimed.

□

NOTE The complex exponential can take complex values or real negative values. This is in cotrast with the real exponential that is always strictly positive. However, we always have $e^z \neq 0$.

5.9 Trigonometric and Hyperbolic Functions

Trigonometric Functions

For real x we have by Euler's formula

$$\left.\begin{aligned} e^{ix} &= \cos x + i \sin x \\ e^{-ix} &= \cos x - i \sin x \end{aligned}\right\} \Longrightarrow \begin{aligned} \cos x &= \frac{1}{2}\left(e^{ix} + e^{-ix}\right) \\ \sin x &= \frac{1}{2i}\left(e^{ix} - e^{-ix}\right) \end{aligned}$$

This is true for real variables. Euler extended the definitions of sin and cos over complex variables using these identities as defining relations.

For any complex z we define

$$\cos z = \frac{1}{2}\left(e^{iz} + e^{-iz}\right)$$

and

$$\sin z = \frac{1}{2i}\left(e^{iz} - e^{-iz}\right)$$

Then we define

$$\tan z = \frac{\sin z}{\cos z}, \quad \cot z = \frac{\cos z}{\sin z}, \quad \sec z = \frac{1}{\cos z}, \quad \csc z = \frac{1}{\sin z}$$

These functions are defined if the denominators are not zero.

Properties of Trigonometric Functions

The differentiation formulas for the complex trigonometric functions are the same as in the case of the real such functions. Also, the various identities of the real trig functions are still valid. We collect these properties here. The proofs are easy deductions of the definitions and are omitted.

Theorem 5.9.1. *We have the following differentiation formulas*

$$\begin{aligned} (\sin z)' &= \cos z, \quad (\cos z)' = -\sin z, \\ (\tan z)' &= \sec^2 z, \quad (\cot z)' = -\csc^2 z, \\ (\sec z)' &= \tan z \sec z, \quad (\csc z)' = -\cot z \csc z \end{aligned}$$

Proof. Exercise. □

Theorem 5.9.2. *We have the following identities.*

1. *For all complex z we have*

$$\cos z^2 + \sin^2 z = 1$$

2. *For all complex z we have*

$$e^{iz} = \cos z + i \sin z$$

(Euler's formula is also valid for complex numbers.)

3. *For all complex z in the domain of $\tan(z)$ we have*

$$\sec^2 z = 1 + \tan^2 z$$

4. *For all complex z in the domain of $\cot(z)$ we have*

$$\csc^2 z = 1 + \cot^2 z$$

Proof. Exercise. □

Theorem 5.9.3. *We have the following addition formulas for all complex z_1 and z_2.*

$$\begin{aligned}
\sin(z_1 + z_2) &= \sin z_1 \cos z_2 + \cos z_1 \sin z_2 \\
\sin(z_1 - z_2) &= \sin z_1 \cos z_2 - \cos z_1 \sin z_2 \\
\cos(z_1 + z_2) &= \cos z_1 \cos z_2 - \sin z_1 \sin z_2 \\
\cos(z_1 - z_2) &= \cos z_1 \cos z_2 + \sin z_1 \sin z_2
\end{aligned}$$

Proof. Exercise. □

Example 5.9.1. Solve for z the equation

$$\cos z = 2$$

Solution: Note that this has no real solutions since $-1 \leq \cos x \leq 1$ for all real x. For complex z we have

$$\cos z = \frac{1}{2}\left(e^{iz} + e^{-iz}\right) = 2$$

which implies

$$e^{2iz} - 4e^{iz} + 1 = 0$$

This equation is qudratic in e^{iz}. Hence,

$$e^{iz} = 2 \pm \sqrt{3} \Rightarrow e^{ix-y} = 2 \pm \sqrt{3}$$

Taking absolute values yields

$$e^{-y} = 2 \pm \sqrt{3}$$

Hence,

$$y = -\ln\left(2 \pm \sqrt{3}\right)$$

Also,

$$e^{ix} = 1$$

Thus

$$x = 2n\pi, \qquad (n \in \mathbf{Z})$$

Therefore, we get the infinitely many solutions

$$z = 2n\pi - i\ln\left(2 \pm \sqrt{3}\right), \qquad (n \in \mathbf{Z})$$

Hyperbolic Functions

For any complex z we define the **hyperbolic cosine** by

$$\cosh z = \frac{1}{2}\left(e^{z} + e^{-z}\right)$$

and the **hyperbolic sine** by

$$\sinh z = \frac{1}{2}\left(e^{z} - e^{-z}\right)$$

The differentialtion formulas are the same as in the real case.

Theorem 5.9.4. *We have the following differentiation formulas.*

$$(\sinh z)' = \cosh z, \quad (\cosh z)' = -\sinh z$$

We also define as in the real case

$$\tanh z = \frac{\sinh z}{\cosh z}, \quad \coth z = \frac{\cosh z}{\sinh z},$$
$$\operatorname{sech} z = \frac{1}{\cosh z}, \quad \operatorname{csch} z = \frac{1}{\sinh z}$$

Most of the known identities of the real hyperbolic functions are also valid in the case of the complex generalizations.

5.10 Logarithm and General Powers

The Logarithm

Let $z = x + iy$ be any nonzero complex number. The **natural logarithm**, $\ln z$, of z is any complex number w such that $e^w = z$.

$$w = \ln z \Leftrightarrow e^w = z$$

If $z = 0$, then $\ln z$ is not defined because always $e^w \neq 0$.

We find a formula for the number(s) $\ln z$. Let $w = u + iv$ and $z = re^{i\theta}$ (for $r > 0$). Then $e^w = z$ implies that

$$re^{i\theta} = z = e^w = e^{u+iv} = e^u e^{iv}$$

Hence, $\left|re^{i\theta}\right| = |e^u e^{iv}|$ or $r = e^u$. Therefore, $u = \ln r$ where ln is the real logarithm of a positive real. Thus, $e^{i\theta} = e^{iv}$ which implies that $v = \theta = \arg z$ up to $\pm 2\pi n$ $(n = 0, 1, 2, \dots)$. Hence, we have

$$\ln z = \ln r + i\theta, \quad (r = |z| > 0,\ \theta = \arg z)$$

Since θ is defined up to an integer multiple of 2π we see that $\ln(z)$ has *infinitely many values* for a given $z \neq 0$. This is an example of a **multiple-valued function**. The value of $\ln z$ that corresponds to the principal value

$\operatorname{Arg} z$ is denoted by $\operatorname{Ln} z$ and it called the **principal value** of $\ln z$. So, we have

$$\operatorname{Ln} z = \ln|z| + i \operatorname{Arg} z$$

and we may write

$$\ln z = \operatorname{Ln} z \pm 2n\pi i \quad (n = 0, 1, 2, \dots)$$

Example 5.10.1. Prove each of the following.

1. $\ln 1 = 0,\ \pm 2\pi i,\ \pm 4\pi i,\ \dots$, and $\operatorname{Ln} 1 = 0$
2. $\ln(-1) = \pm\pi i,\ \pm 3\pi i,\ \pm 5\pi i,\ \dots$, and $\operatorname{Ln}(-1) = \pi i$
3. $\ln i = \pi i/2,\ -3\pi i/2,\ \ 5\pi i/2,\ -7\pi i/2, \dots$, and $\operatorname{Ln}(i) = \pi i/2$
4. $\ln e = 1,\ 1 \pm 2\pi i,\ 1 \pm 4\pi i,\ \dots$, and $\operatorname{Ln} e = 1$
5. $\ln(1+i) = \dfrac{\ln 2}{2} + \dfrac{i\pi}{4} + 2\pi n i, \quad (n \text{ integer})$, and $\operatorname{Ln}(1+i) = \dfrac{\ln 2}{2}$
6. $\ln(3+4i) = \ln 5 + i \arctan(4/3) + 2\pi n i \quad (n \in \mathbf{Z})$, and $\operatorname{Ln}(3+4i) = \ln 5 + i \arctan(4/3)$

Solution: We have the following computations, where the left hand side logarithms are complex and the ones in the right hand sides are real.

1. $$\ln 1 = \ln 1 + i \arg 1 = 2\pi n i, \quad (n \in \mathbf{Z})$$
2. $$\ln(-1) = \ln|-1| + i \arg(-1) = \pi i + 2\pi n i, \quad (n \in \mathbf{Z})$$
3. $$\ln(i) = \ln|i| + i \arg(i) = \frac{\pi i}{2} + 2\pi n i, \quad (n \in \mathbf{Z})$$
4. $$\begin{aligned} \ln e &= \ln e + i \arg e \\ &= 1 + 2\pi n i, \quad (n \in \mathbf{Z}) \end{aligned}$$

5.

$$\begin{aligned}\ln(1+i) &= \ln\sqrt{2} + i\arctan(1/1) + 2\pi ni \\ &= \frac{\ln 2}{2} + \frac{i\pi}{4} + 2\pi ni, \qquad (n \in \mathbf{Z})\end{aligned}$$

6.

$$\begin{aligned}\ln(3+4i) &= \ln|3+4i| + i\arctan(4/3) + 2\pi ni \\ &= \ln 5 + i\arctan(4/3) + 2\pi ni \\ &\simeq 1.6094 + 0.9273i + 2\pi ni, \qquad (n \in \mathbf{Z})\end{aligned}$$

Theorem 5.10.1. *We have the following properties.*

1. $\ln(zw) = \ln(z) + \ln(w)$
2. $\ln\left(\frac{z}{w}\right) = \ln(z) - \ln(w)$
3. $e^{\ln z} = z$
4. $\ln(e^z) = z + 2\pi ni, \qquad (n \in \mathbf{Z})$
5. $(\ln z)' = \frac{1}{z}$

Proof. Exercise. □

The General Power z^c

For any complex number c and any nonzero complex number z we define the power z^c by

$$z^c = e^{c\ln z}, \qquad (z \neq 0)$$

This is a multi-valued correspondence, because of $\ln z$. The principal part of of z^c is defined by $e^{c\,\mathrm{Ln}\,z}$.

Example 5.10.2 (i^i is real)**.** Compute i^i.

Solution: By the definition of z^c and by part 3 of Example 5.10.1 we have

$$i^i = e^{i\ln(i)} = e^{i\left(\frac{\pi i}{2} + 2\pi ni\right)} = e^{-\frac{\pi}{2} - 2\pi n}, \qquad (n \in \mathbf{Z})$$

So i^i consists of an infinite set of real numbers.

Example 5.10.3. Compute $(1+i)^i$.

Solution: By the definition of z^c and by part 5 of Example 5.10.1 we have

$$\begin{aligned}(1+i)^i &= e^{i\ln(1+i)} \\ &= e^{i\left(\frac{\ln 2}{2}+\frac{i\pi}{4}+2\pi ni\right)} \\ &= e^{-\frac{\pi}{4}-2\pi n}\left[\cos\left(\left(\ln 2\right)/2\right)+i\sin\left(\left(\ln 2\right)/2\right)\right], \quad (n \in \mathbf{Z})\end{aligned}$$

5.11 Complex Integration

Smooth Curves

Definitions

A **curve**, C, in the complex plane is a continuous function $z(t)$ from a closed interval of real numbers $[a,b]$ $(a<b)$ to the set of complex numbers $\mathbf{C}$.

$$C: \ z(t) \in \mathbf{C}, \quad t \in [a,b]$$

The image of $z(t)$ traces a continuous curve in the complex plane in the intuitive sense. The points $z(a)$ and $z(b)$ are called the **endpoints** of the curve. If $z(a)=z(b)$, then the curve is called **closed**. A curve is called **simple**, if it has no self intersections other than possibly the end points. So, for a simple curve we have $z(t_1) \neq z(t_2)$ for all t_1, t_2 in (a,b).

The curve $z(t)$ is called **smooth**, if it has continuous nonzero derivative $\dfrac{dz}{dt} \neq 0$ at each point of its domain of definition. Geometrically, C has a unique and continuously turning tangent vector. Since $z(t)$ is complex we may write

$$C: \ z(t) = x(t) + iy(t)$$

where $x(t)$ and $y(t)$ are continuous real functions in t.

A curve $z(t)$ is called **piecewise smooth**, if it can be written as the union of finitely many smooth curves $z_1(t), z_2(t), \ldots, z_i(t), \ldots$ so that the right end of each $z_i(t)$ is the same as the left end of the next $z_{i+1}(t)$.

Our main focus is on simple piecewise smooth curves. Such a curve is called a **contour**.

We think of curves as being **oriented**, by which we mean that the image points are *ordered* according to the increasing values of the parameter t.

The **length**, $L_a^b(z)$, of the curve $C:\ z(t),\ t\in[a,b]$ is the integral[3]

$$L_a^b(z)=\int_a^b |z'(t)|\,dt$$

If $L_a^b(z)$ is finite we say that $z(t)$ has **finite length**.

Often we are interested in the *image set* of a curve and how this image set is traced. A curve and its image set are not the same. For example, each of the following curves has the unit circle as its image.

1. $z_1(t)=e^{it}=\cos t+i\sin t,\quad t\in[0,2\pi]$
2. $z_2(t)=e^{2ti}=\cos(2t)+i\sin(2t),\quad t\in[0,\pi]$
3. $z_3(t)=e^{(\pi/2-t)i}=\sin t+i\cos t,\quad t\in[0,2\pi]$
4. $z_4(t)=e^{-it}=\cos t-i\sin t,\quad t\in[0,2\pi]$
5. $z_5(t)=e^{2ti}=\cos(2t)+i\sin(2t),\quad t\in[0,2\pi]$

Curves $z_1(t)$ and $z_2(t)$ trace out the unit circle once in the positive (counterclockwise) direction, but $z_2(t)$ traces it at "double speed". Curves $z_3(t)$ and $z_4(t)$ trace out the unit circle once in the negative (clockwise) direction, but they have different endpoints. Curve $z_5(t)$ traces out the unit circle twice in the positive direction.

It is common (and it may lead to confusion) that the image set of a curve, along with the orientation traced, to be called a curve. In such case any function $z(t)$ that represents the same set of points (oriented the same way) is called a **parametrization** of the curve and t is called the **parameter**.

If $z(t),\ t\in[a,b]$ is a parametrization of a curve C, then the parametrization

$$g(t)=z(a+b-t),\quad t\in[a,b]$$

defines a curve C^* which has the same image set as C but the points $g(t)$ are traced in the opposite direction than the that of C. So g starts at $g(a)=z(b)$ and ends at $g(b)=z(a)$. This curve C^* is called the **opposite** of C.

We say that a simple closed curve has **positive orientation**, if when we travel on it in the increasing values of the parameter the interior of the curve is on the left. This is the same as counterclockwise orientation. The opposite orientation is called **negative orientation**.

[3]This is the familiar formula of arclength from vector calculus: $\int_a^b |z'(t)|\,dt=\int_a^b \sqrt{x'(t)^2+y'(t)^2}dt$.

Examples of Curve Parametrizations

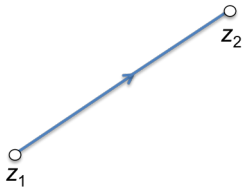
In general, there are many ways of parametrizing a curve. Knowing a parametrization $z(t)$ of a curve can be useful. For example, by using a formula for $z(t)$ for various values of t we may compute and sketch the corresponding images $z(t)$. However, finding parametrizations of curves can be difficult.

We are mainly interested in parametrizing straight line segments and circles. Here we discuss the "standard" parametizations of these curves.

Parametrization of a Directed Straight Line Segment We consider the directed straight line segment going from the complex number z_1 to the complex number z_2. A parametrization is given by

$$z(t) = (1-t)\, z_1 + t z_2, \quad t \in [0,1]$$

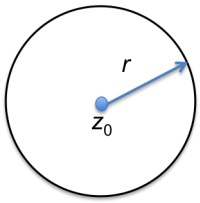
For $t = 0$ we are at $z(0) = z_1$ and at $t = 1$ we are at $z(1) = z_2$. The midpoint is at $t = 1/2$ and it is $(z_1 + z_2)/2$.



Parametrization of Circle A parametrization of the circle centered at the complex number z_0 and radius r $(r > 0)$ traced in the *positive direction* is given by

$$z(t) = z_0 + re^{it}, \quad t \in [0, 2\pi]$$

This is because all points z whose distance from z_0 is r satisfy $|z - z_0| = r$. So, $z - z_0$ has polar representation $z - z_0 = re^{it}$ for this fixed r, where t is the varying $\arg(z - z_0)$.



For a *negatively oriented* circle, we may use the parametrization

$$z(t) = z_0 + re^{-it}, \quad t \in [0, 2\pi]$$

If we want to parametrize the **arc** of the circle centered at the complex number z_0 and radius r $(r > 0)$ traced in the *positive direction* from the complex number z_1 to the complex number z_2, we may use the parametrization

$$z(t) = z_0 + re^{it}, \quad t \in [\alpha, \beta]$$

where $\alpha = \arg(z_1 - z_0)$ and $\beta = \arg(z_2 - z_0)$.

Complex Line Integral

Definition

Let $f(z)$ be a continuous complex function defined in an open subset U of $\mathbf{C}$. Let C be a smooth curve $z(t)$, $t \in [a, b]$ such that $z(t) \in U$ for all t. We abbreviate all this by saying that $f(z)$ *is defined on* C.

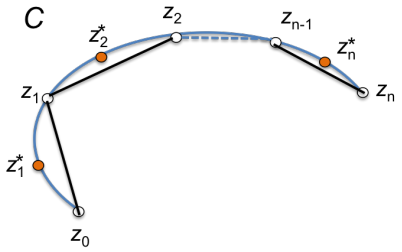
We partition the interval $a \leq t \leq b$ as

$$t_0 (= a) < t_1 < \cdots < t_{n-1} < t_n (= b)$$

and let

$$z_0 = z(t_0), \ z_1 = z(t_1), \ \ldots, \ z_n = z(t_n)$$

This divides the curve C into arcs C_i connecting each z_i with z_{i+1}. Let z_i^* be an arbitrary point in C_i. So $z_i^* = z(t_i^*)$ for some t_i^* in $[t_i, t_{i+1}]$.



We form the Riemann sum

$$S_n = \sum_{i=1}^{n} f(z_i^*) \Delta z_i \qquad \text{where} \qquad \Delta z_i = z_i - z_{i-1}$$

and take the limit of S_n over all partitions so that the largest $\Delta t_i = |t_i - t_{i-1}|$ approaches zero as $n \to \infty$. Since $z(t)$ is continuous, this implies that $|\Delta z_i|$ also approaches zero. The limit is called the **complex line integral** of $f(z)$ over the curve C and its denoted by

$$\int_C f(z)\, dz$$

The function $f(z)$ is called the **integrand** of the integral and the curve C is called the **path of integration**.

If the path of integration C is a closed curve we usually write

$$\oint_C f(z)\, dz$$

for the integral.

NOTE: It can be proved that under the given assumptions that $f(z)$ is continuous and C is smooth, the line integral always exists. Actually, one may extend this definition of integral over curves that are not necessarily smooth but only have finite length and prove that such integral exists, still assuming that $f(z)$ is continuous.

If C is a piecewise smooth curve consisting of k consecutive smooth curves $C_1, \ldots, C_k$, then we define the integral of $f(z)$ over C as the sum

$$\int_C f(z)\, dz = \int_{C_1} f(z)\, dz + \cdots + \int_{C_k} f(z)\, dz$$

The integral over a piecewise smooth curve being a finite sum of complex numbers always exists.

From now on we consider integrals where C is any either smooth or piecewise smooth. This includes the case of C being a contour, i.e., a simple piecewise smooth curve.

Properties

Theorem 5.11.1 (Properties of Line Integrals)**.** *The line integral satisfies the following basic properties. We assume that all functions are continuous defined in an open subset U of $\mathbf{C}$ that contains the curve C.*

1. ***Linearity.*** *For all complex constants c_1 and c_2 the integral of $c_1 f_1(z) + c_2 f_2(z)$ over C exists and*

$$\int_C [c_1 f_1(z) + c_2 f_2(z)]\, dz = c_1 \int_C f_1(z)\, dz + c_2 \int_C f_2(z)\, dz$$

2. ***Path reversal.*** *The integral of $f(z)$ over the opposite curve C^* of C exists and*

$$\int_{C^*} f(z)\, dz = -\int_C f(z)\, dz$$

3. ***Partitioning of path.*** *If a path C is partitioned into paths $C_1, C_2, \ldots, C_k$ then*

$$\int_C f(z)\, dz = \int_{C_1} f(z)\, dz + \cdots + \int_{C_k} f(z)\, dz$$

We study two main methods of integration (a) by parametrization of the path and (b) by indefinite integration. The first method is more general and it applies to continuous $f(z)$. The second applies only to analytic $f(z)$ and it is immediate if we know an **antiderivative** $F(z)$ of $f(z)$ (so that $F'(z) = f(z)$).

Integration by Parametrization of Path

Theorem 5.11.2 (Integration by Path Paramertization)**.** *Let C be a smooth curve defined on a continuous function $f(z)$ and let $z(t)$, $t \in [a, b]$ be a parametrization of C. Then*

$$\int_C f(z)\, dz = \int_a^b f(z(t))\, z'(t)\ dt \qquad \left(z'(t) = \frac{dz}{dt}\right)$$

NOTE: The complex line integral reduces to the real definite integral $\int_a^b f(t)\, dt$ of Calculus. This is seen by using Theorem 5.11.2 and choosing C to be $z(t) = t$, $t \in [a, b]$ and $f(t)$ to be a real-valued function in t that is continuous on $[a, b]$.

Example 5.11.1. Let C be the unit circle traced once in the positive direction. Show that

1. $\displaystyle\int_C \frac{1}{z} dz = 2\pi i$

2. $\int_C z^n dz = 0$, if $n = \pm 1, \pm 2, \ldots$ (So, $n \neq -1$).

Solution: We parametrize C by $z(t) = \cos(t) + i\sin(t) = e^{it}$ with $t \in [0, 2\pi]$. We have $z'(t) = ie^{it}$.

1. $f(z(t)) = 1/z(t) = e^{-it}$. Hence,

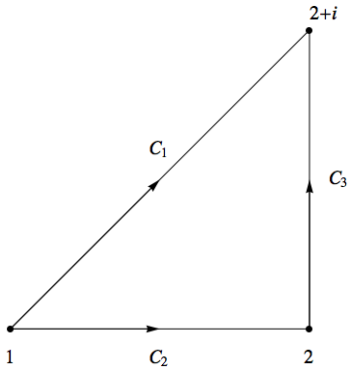
$$\int_C \frac{1}{z} dz = \int_0^{2\pi} e^{-it} i e^{it}\, dt = i\int_0^{2\pi} dt = 2\pi i$$

2. $f(z(t)) = z^n(t) = e^{int}$. Since n is an integer other than -1, we have

$$\begin{aligned}
\int_C z^n dz &= \int_0^{2\pi} e^{int} i e^{it}\, dt \\
&= i\int_0^{2\pi} e^{i(n+1)t} dt \\
&= \left.\frac{e^{i(n+1)t}}{n+1}\right|_0^{2\pi} \\
&= \frac{1}{n+1}\left(e^{2\pi i(n+1)} - 1\right) \\
&= 0
\end{aligned}$$

Example 5.11.2 (Integral of non analytic function; dependence of path)**.** Integrate $f(z) = \mathrm{Re}(z) = x$ from 1 to $2 + i$

1. along C_1 and
2. along C consisting of C_2 and C_3.



Solution: (a) We parametrize C_1 by $z_1(t) = (1-t)\,1 + t\,(2+i) = (t+1) + it$, with $t \in [0,1]$. Hence, $z_1'(t) = 1+i$ and $f(z_1(t)) = \mathrm{Re}(z_1(t)) = t+1$. Therefore,

$$\int_{C_1} \mathrm{Re}(z)\,dz = \int_0^1 (t+1)(1+i)\,dt = (1+i)\int_0^1 (t+1)\,dt = \frac{3}{2} + \frac{3}{2}i$$

(b) The integral over C is the sum of the integrals over C_2 and C_3. We parametrize C_2 by $z_2(t) = t+1$, with $t \in [0,1]$. Hence, $z_2'(t) = 1$ and $f(z_2(t)) = \mathrm{Re}(z_2(t)) = t+1$. Therefore,

$$\int_{C_2} \mathrm{Re}(z)\,dz = \int_0^1 (t+1)\,dt = \frac{3}{2}$$

We parametrize C_3 by $z_3(t) = (1-t)\,2 + t\,(2+a) = 2 + it$, with $t \in [0,1]$. Hence, $z_3'(t) = i$ and $f(z_3(t)) = \mathrm{Re}(z_3(t)) = 2$. Therefore,

$$\int_{C_3} \mathrm{Re}(z)\,dz = \int_0^1 2i\,dt = 2i$$

We conclude that

$$\int_C \mathrm{Re}(z)\,dz = \frac{3}{2} + 2i$$

Integration of Analytic Functions by Indefinite Integration

Theorem 5.11.3 (Indefinite Integration of Analytic Function)**.** *Let $f(z)$ be an analytic function defined in a simply connected domain D . Then $f(z)$ has an indefinite integral (antiderivative) $F(z)$ that is analytic in D. Furthermore, for any piecewise smooth curve C that lies entirely in D joining two points z_0 and z_1 of D we have*

$$\int_C f(z)\ dz = F(z_1) - F(z_0) \qquad\qquad (F' = f)$$

Theorem 5.11.3 is a generalization of the *Fundamental Theorem of Calculus*, if we let C be $z(t) = t$ for $t \in [a,b]$ and $f(z) = f(t)$ be a real-valued function in t that is differentiable on (a,b) and continuous on $[a,b]$.

Theorem 5.11.4. *Let $f(z)$ be an analytic function defined in a simply connected domain D . Let C be a smooth simple closed curve that lies entirely in D. Then*

$$\int_C f(z)\ dz = 0$$

Proof. Since the ends of the curve are the same we have by Theorem 5.11.3 that the integral equals $F(z_1) - F(z_0) = 0$. □

Note that the integral in the theorem only depends on the function f and the end points z_0 and z_1. It is independent of C and also of which antiderivative F of f we use. To indicate the independence of path and the dependence on the endpoints in this case we may (cautiously) write

$$\int_C f(z)\ dz = \int_{z_0}^{z_1} f(z)\ dz$$

Example 5.11.3. Use Theorem 5.11.3 to compute each of the following integrals.

1. $\int_{C_1} z^2 dz$ where C_1 is the line segment from 0 to $2+i$.
2. $\int_{C_2} \sin z\ dz$ where C_2 is a path from $-\pi i$ to πi.
3. $\int_{C_3} e^{2z}\ dz$ where C_3 is a path from $\pi i/2$ to πi.

Solution: All functions are entire so the simply connected domain D is all of $\mathbf{C}$. We have by Theorem 5.11.3:

1. $\int_{C_1} z^2 dz = \left.\frac{z^3}{3}\right|_0^{2+i} = \frac{1}{3}(2+i)^3 = \frac{2}{3} + \frac{11}{3}i$
2. $\int_{C_2} \sin z\ dz = -\cos z|_{-\pi i}^{\pi i} = -\cos(\pi i) + \cos(-\pi i) = 0$
3. $\int_{C_3} e^{2z}\ dz = \left.\frac{e^{2z}}{2}\right|_{\pi i/2}^{\pi i} = \frac{1}{2}(e^{2\pi i} - e^{\pi i}) = \frac{1}{2}(1-(-1)) = 1$

NOTE: **Simple connectedness is important** in Theorem 5.11.3. In part 1 of Example 5.11.1, we may consider as a domain U of definition of the function $f(z) = 1/z$ the punctured open disk of radius 2 with the origin removed. Then $f(z)$ is analytic in U and U is not simply connetcted. The value of integral is $2\pi i$ and not zero as we would expect, if Theorem 5.11.3 were applicable.

Bound for the Absolute Value of the Integral

Theorem 5.11.5. *Let C be a smooth curve of length L and let $f(z)$ be a continuous complex function defined on C. If there is a positive constant M such that $|f(z)| \leq M$ for all z in C, then*

$$\left| \int_C f(z)\ dz \right| \leq ML$$

Example 5.11.4. Estimate the absolute values of the integral $\int_C e^{z^2}\ dz$, where C is the positively oriented upper unit semicircle.

Solution: The length of C is $L = \pi$. Now for $z \in C$ we have $|z| = 1$. Thus, $x^2 + y^2 = 1$. Therefore,

$$\left| e^{z^2} \right| = \left| e^{x^2 - y^2 + i(2xy)} \right| = e^{x^2 - y^2} = e^{1-2y^2}$$

Now e^{1-2y^2} has maximum value $M = e$ for $y \in [0, 1]$. Hence, by Theorem 5.11.5 we have the estimate

$$\left| \int_C e^{z^2}\ dz \right| \leq \pi e \simeq 8.53973$$

5.12 Cauchy's Integral Theorem

The next theorem is one of the pillars of complex function theory. It has far reaching implications in complex integration and its applications.

Theorem 5.12.1 (Cauchy's Integral Theorem)**.** *Let $f(z)$ be analytic in a simply connected domain D. Then for every simple closed curve C in D we have*

$$\oint_C f(z)\ dz = 0$$

Example 5.12.1 (Entire functions)**.** For every simple closed curve C we have

$$\oint_C \sin(z)\ dz = 0, \qquad \oint_C \cos(z)\ dz = 0,$$
$$\oint_C e^z\ dz = 0, \qquad \oint_C z^n\ dz = 0, \quad (n = 0, 1, 2, \dots)$$

Solution: The integrals are zero by Cauchy's Integral Theorem 5.12.1 since all integrands are entire.

Example 5.12.2 (Non analytic outside contour)**.** For every simple closed curve C that does not include any singular points of the integrands of the following integrals, we have

1. $\displaystyle\oint_C \frac{1}{z^2+1}\,dz = 0$ (C does not include $z = i$ or $z = -i$).
2. $\displaystyle\oint_C \frac{z}{z^2+2z-3}\,dz = 0$ (C does not include $z = 1$ or $z = -3$).
3. $\displaystyle\oint_C \tan z\,dz = 0$ (C does not include any of $z = (2n-1)\pi/2$, $n \in \mathbf{Z}$).

REMARK: All assumptions in Cauchy's Integral Theorem are important. If some assumption is missing, then the theorem cannot be applied. Let us examine the following cases, where C is the positively oriented unit circle.

1. Analyticity of $f(z)$ is required. $\displaystyle\oint_C \overline{z}\,dz = 2\pi i \neq 0$ ($f(z) = \overline{z}$ is not analytic.)
2. Simple connectedness of the domain is required. $\displaystyle\oint_C \frac{1}{z}\,dz = 2\pi i \neq 0$. (The domain D here contains C but excludes $z = 0$. So, it is not simply connected.)
3. The integral is zero but Cauchy's Integral Theorem is not applicable. For example, $\displaystyle\oint_C \frac{1}{z^2}\,dz = 0$. However, the theorem does not apply. (State two reasons.)

Theorem 5.12.2 (Independence of Path)**.** *Let $f(z)$ be analytic in a simply connected domain D. For any curve C in D we have that the integral*

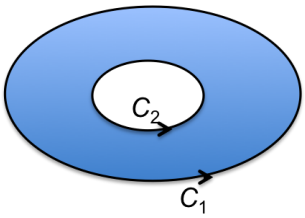
$$\int_C f(z)\,dz$$

is independent of path. This means that the integral only depends on the endpoints of C and not on C itself.

Cauchy's Theorem for Multiply-Connected Domains

Theorem 5.12.3. *Let $f(z)$ be analytic in a doubly-connected domain D bounded by two simple closed curves C_1 and C_2.* [4] *Then we have*

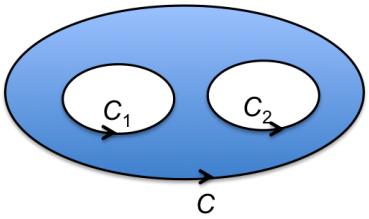
$$\oint_{C_1} f(z)\,dz = \oint_{C_2} f(z)\,dz$$



In the statement of the theorem by "

Theorem 5.12.3 generalizes to multiply-connected domains in an obvious way. For example, for triply-connected domains we have (see Fig. below)

$$\oint_{C} f(z)\,dz = \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz$$



5.13 Cauchy's Integral Formula

Theorem 5.13.1 (Cauchy's Integral Formula)**.** *Let $f(z)$ be analytic in a simply connected domain D. Then for every point z_0 in D and every positively oriented simple closed curve C in D that has z_0 as an interior point, we have*

$$\oint_{C} \frac{f(z)}{z - z_0}\,dz = 2\pi i f(z_0)$$

[4] The domain D is open, so it actually extends passed the two curves.

Example 5.13.1. Compute $\displaystyle\oint_C \frac{z^4 - z^2 + 1}{z + i}\, dz$ for any positively oriented simple closed curve C enclosing $-i$.

Solution: The function $f(z) = z^4 - z^2 + 1$ is entire. Hence, by Cauchy's Integral Formula 5.13.1 with $z_0 = -i$, we have

$$\begin{aligned}\oint_C \frac{z^4 - z^2 + 1}{z + i}\, dz &= 2\pi i \left(z^4 - z^2 + 1\right)\Big|_{z=-i} \\ &= 6\pi i\end{aligned}$$

Example 5.13.2. Compute $\displaystyle\oint_C \frac{z^2 + 1}{2z - i}\, dz$ for any positively oriented simple closed curve C enclosing $i/2$.

Solution: The function $f(z) = z^2 + 1$ is entire. Hence, by Cauchy's Integral Formula 5.13.1 with $z_0 = i/2$, we have

$$\begin{aligned}\oint_C \frac{z^2 + 1}{2z - i}\, dz &= \frac{1}{2}\oint_C \frac{z^2 + 1}{z - i/2}\, dz \\ &= \pi i \left(z^2 + 1\right)\Big|_{z=i/2} \\ &= \frac{3\pi i}{4}\end{aligned}$$

Example 5.13.3. Compute $\displaystyle\oint_C \frac{\sin(2z)}{z - i}\, dz$ for any simple closed curve C enclosing i and oriented *clockwise*.

Solution: By Cauchy's Integral Formula 5.13.1, we have, after we reverse the sign, due to the negative orientation

$$\begin{aligned}\oint_C \frac{\sin(2z)}{z - i}\, dz &= -2\pi i \sin(2z)|_{z=i} \\ &= -2\pi i \sin(2i) \\ &= 2\pi \sinh 2\end{aligned}$$

Example 5.13.4. Compute $\displaystyle\oint_C \frac{e^z}{(z - 1)(z^2 + 1)}\, dz$ for any positively oriented simple closed curve C for which 1 lies inside and $\pm i$ lie outside.

Solution: The function $\dfrac{e^z}{z^2+1}$ is analytic on and inside C. Hence, by Cauchy's Integral Formula 5.13.1 and $f(z) = e^z/(z^2+1)$, we have

$$\begin{aligned}\oint_C \frac{e^z/(z^2+1)}{z-1}\,dz &= 2\pi i \left.\frac{e^z}{z^2+1}\right|_{z=1} \\ &= \pi i e\end{aligned}$$

Example 5.13.5. Compute $\displaystyle\oint_C \frac{2\sin z}{z^2-1}\,dz$ for any positively oriented simple closed curve C that contains the points ± 1.

Solution: The integrand has two singularities: at $z = \pm 1$. Unfortunately, both singularities are inside the integration contour C. However, if we use partial fractions, we only get one singular point at a time. We have

$$\begin{aligned}\oint_C \frac{2\sin z}{z^2-1}\,dz &= \oint_C \frac{\sin z}{z-1}\,dz - \oint_C \frac{\sin z}{z+1}\,dz \\ &= 2\pi i \sin 1 - 2\pi i \sin(-1) \\ &= 4\pi i \sin 1\end{aligned}$$

We used Cauchy's Integral Formula 5.13.1 for each integral. For the first integral we used $f(z) = \sin z$ and $z_0 = 1$ and for the second integral we used $f(z) = \sin z$ and $z_0 = -1$.

5.14 Cauchy's Theorem for Derivatives

Theorem 5.14.1 (Cauchy's Theorem for Derivatives)**.** *Let $f(z)$ be analytic in a simply connected domain D. Then $f(z)$ has derivatives of all orders in D that are analytic. Furthermore, for every point z_0 in D and every positively oriented simple closed curve C in D that has z_0 as an interior point, we have*

$$\oint_C \frac{f(z)}{(z-z_0)^{n+1}}\,dz = \frac{2\pi i}{n!} f^{(n)}(z_0), \quad n = 1, 2, \ldots$$

Example 5.14.1. Compute $\displaystyle\oint_C \frac{z^4 - z^2 + 1}{(z+i)^3}\,dz$ for any positively oriented simple closed curve C enclosing $-i$.

Solution: By Cauchy's Theorem for Derivatives 5.14.1, we have

$$\begin{aligned}\oint_C \frac{z^4 - z^2 + 1}{(z+i)^3}\, dz &= \frac{2\pi i}{2!}\left(z^4 - z^2 + 1\right)''\Big|_{z=-i} \\ &= \pi i\left(12z^2 - 2\right)\Big|_{z=-i} \\ &= -14\pi i\end{aligned}$$

Example 5.14.2. Compute $\oint_C \frac{\sin(2z)}{(z-i)^4}\, dz$ for any simple closed curve C enclosing i and oriented **clockwise**.

Solution: By Cauchy's Theorem for Derivatives 5.14.1, we have, after we reverse the sign, due to the negative orientation

$$\begin{aligned}\oint_C \frac{\sin(2z)}{(z-i)^4}\, dz &= -\frac{2\pi i}{3!}\sin(2z)'''|_{z=i} \\ &= -\frac{\pi i}{3}\left(-8\cos(2i)\right) \\ &= \frac{8\pi i}{3}\cosh 2\end{aligned}$$

Example 5.14.3. Compute $\oint_C \frac{\cos z}{(i-2z)^3}\, dz$ for any simple closed curve C enclosing $i/2$ and oriented counterclockwise.

Solution: We have

$$\begin{aligned}\oint_C \frac{\cos z}{(i-2z)^3}\, dz &= \frac{1}{(-2)^3}\oint_C \frac{\cos z}{(z-i/2)^3}\, dz \\ &= -\frac{\pi i}{8}\cos(z)''|_{z=i/2} \\ &= \frac{\pi i}{8}\cosh\frac{1}{2}\end{aligned}$$

Example 5.14.4. Compute $\oint_C \frac{e^z}{(z-1)^2(z^2+1)}\, dz$ for any positively oriented simple closed curve C for which 1 lies inside and $\pm i$ lie outside.

Solution: We have

$$\begin{aligned}\oint_C \frac{e^z}{(z-1)^2(z^2+1)}\,dz &= 2\pi i \left(\frac{e^z}{z^2+1}\right)'\Big|_{z=1} \\ &= 2\pi i \left.\frac{e^z(z^2+1)-2ze^z}{(z^2+1)^2}\right|_{z=1} \\ &= 0\end{aligned}$$

5.15 Sequences and Series

5.16 Taylor Series

5.17 Laurent Series

5.18 Poles and Zeros

5.19 The Residue Theorem

Chapter 6

Appications

6.1 Application of PDEs: Two-dimensional Fluid Flow

Chapter 7

Probability

7.1 Sample Space and Events

Probability theory creates mathematical models for phenomena that are governed by randomness or chance. Examples include weather forecasting, life insurance, games with dice or cards, quality of goods, etc.

The accuracy of these models is studied by various observations and experiments. This the main goal of **statistics**.

An **experiment** is a process of observation or measurement. The experiments of interest are the ones that involve **chance** or **randomness**. For example, flipping a coin a few times and counting the number of "heads". Or, counting the number of defective pairs of shoes in a random sample.

A single performance of an experiment is called a **trial**. The result of a trial is an **outcome** or a **sample point** . The **sample space** S of an experiment is the set of all possible outcomes. Any subset of S is called an **event**. The outcomes are called **simple events**. Note that S itself is an event.

Example 7.1.1 (Random Experiments and Sample spaces)**.**

1. Flipping a coin. $S = \{\text{heads}, \text{tails}\}$

2. Rolling a die. $S = \{1, 2, 3, 4, 5, 6\}$

3. Checking the room temperature are a certain time. S is some interval of numbers. For example, $S = [60°, 80°]$ in degrees Fahrenheit.

4. Counting the number of pedestrians crossing an intersection in an hour. S is some finite interval of integers.

□

Example 7.1.2 (Events.)**.** Some events for rolling a die.

1. $A = \{1, 3, 5\}$, "an odd number".
2. $B = \{2, 4, 6\}$, "an even number".
3. $C = \{1, 2\}$, "a number less than 3".
4. All the simple events are the outcomes $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$.
5. $S = \{1, 2, 3, 4, 5, 6\}$, the event of all outcomes of rolling a die.

□

Unions, Intersections, and Complements of Events

Next we need the following concepts for events (subsets) A, B, C, A_i of a sample S.

- The **union** $A \cup B$ is the set of all outcomes that are either in A or in B.
- The **intersection** $A \cap B$ is the set of all outcomes that are both in A and B.
- The **complement** A^c is the set of outcomes of S that are not in A.

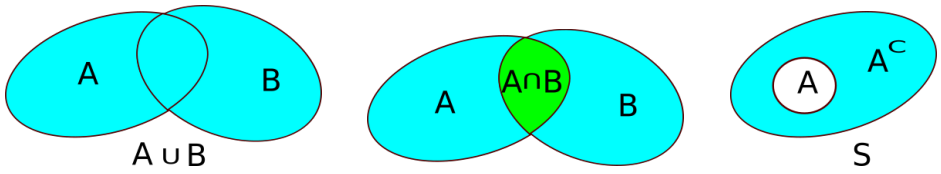


Figure 7.1: Union, intersection, and compelement.

We may also have the intersection or union of more that two sets.

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_n$$

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_n$$

We may also have the union or intersection of an infinite family of subsets of S.

If events A and B have no points in common we call them **mutually exclusive** or **disjoint** and we write

$$A \cap B = \varnothing$$

where $\varnothing$ is the empty set.

Example 7.1.3. Consider the events in rolling a die.

1. $A = \{3, 4, 5, 6\}$ is a number greater than 2.
2. $B = \{1, 2, 3, 4\}$ is a number less than 5.
3. $C = \{1, 3, 5\}$ is an odd number.

The we have $A \cap B = \{3, 4\}$, $B \cap C = \{1, 3\}$, $A \cap B \cap C = \{3\}$. In addition, we have $A \cup B = S = \{1, 2, 3, 4, 5, 6\}$, $B \cup C = \{1, 2, 3, 4, 5\}$, and $A^c = \{1, 2\}$. □

7.2 Probability

The probability of an event A measures how frequently the event is likely to occur, if we make many trials. For example, if we flip a fair coin, we expect heads H and tails T to appear approximately equally often. We say that H and T are **equally likely** events. Similarly, for a fair die each of the simple events $1, \ldots, 6$ are equally likely. These examples suggest the following simple definition of probability.

Definition (Naive Definition of Probability)**.** If A is an event of a finite sample space S, where the outcomes are equally likely, then the probability $P(A)$ of A is defined by

$$P(A) = \frac{\text{Number of elements of } A}{\text{Number of elements of } S}$$

From the definition we see immediately that

$$P(S) = 1, \text{ and } P(\varnothing) = 0$$

Example 7.2.1. In rolling a fair die once, what is the probability $P(A)$ of A of obtaining a 2 or a 6? And what is the probability $P(B)$ of B of obtaining an odd number?

Solution: Since the outcomes are equally likely we have $P(A) = 2/6 = 1/3$ because $A = \{2, 6\}$ has 2 points and $S = \{1, \ldots, 6\}$ has 6 points.

Likewise, $P(B) = 3/6 = 1/2$, since $B = \{1, 3, 5\}$ has 3 points.

Example 7.2.2. Find the probability $P(A)$ of the event A which is that at least one head occurs in a sequence of two tosses of a fair coin.

Solution: The sample space consists of $S = \{HH, HT, TH, TT\}$ and $A = \{HH, HT, TH\}$. Hence, $P(A) = 3/4$.

□

The naive definition of probability is too restrictive to be able to use in all situations, especially when we have large or even infinite numbers of outcomes. Another interesting idea in defining the probability of an event A is to look at the **relative frequency** of A, i.e., how often the event A occurs over a number of trials.

$$f_{rel}(A) = \frac{\text{Number of occurances of } A}{\text{Number of trials}}$$

Relative frequences satisfy some basic properties:

$$0 \leq f_{rel}(A) \leq 1, \quad f_{rel}(S) = 1, \quad f_{rel}(\varnothing) = 0$$

and for mutually exclusive events A and B

$$f_{rel}(A \cup B) = f_{rel}(A) + f_{rel}(B)$$

Now if we want a more accurate definition of probability of an event we need to take the limit of the relative frequency as n goes to infinity.

Although the above definition would serve better than the first one, an even more general definition seems to be necessary for successful use in applications. Such definition was introduced by Kolmogorov in 1965.

Definition (General Definition of Probability)**.** For a sample space S and each event A, we associate a number $P(A)$, the **probability** of A such that the following **axioms** are satisfied.

1. For every event A (subset of S)

$$0 \leq P(A) \leq 1 \tag{7.1}$$

2. For S

$$P(S) = 1 \tag{7.2}$$

3. For mutually exclusive events A and B $(A \cap B = \varnothing)$

$$P(A \cup B) = P(A) + P(B) \qquad (A \cap B = \varnothing) \tag{7.3}$$

If S is infinite Axiom 3 is replaced by: For mutually, exclusive events $A_1, A_2, A_3, \ldots,$

$$P(A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots) = P(A_1) + P(A_2) + \cdots + P(A_n) + \cdots$$

In the case of S being infinite, we restrict S so that its subsets form a σ-algebra, in which the infinite sum above is well defined as a convergent infinite series. This notion, however, is beyond the scope of our notes.

Properties of Probability

Axiom 3 combined with mathematical induction yields the following generalization of the axiom.

Theorem 7.2.1 (Generalization of Axiom 3)**.** *For finitely many mutually exclusive events $A_1, A_2, \ldots A_n$ of a sample space S*

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) \tag{7.4}$$

Example 7.2.3 (Mutually Exclusive Events)**.** A writer of fiction writes $5-10$, $11-15$, $16-20$, and over 20 pages per day with respective probabilities 0.20, 0.40, 0.30, and 0.10. What is the probability that at least 11 pages are written?

Solution: The events are mutually exclusive, so by Theorem 7.2.1 we have that the required probability is $0.40 + 0.30 + 0.10 = 0.80$ or 80%.

□

Theorem 7.2.2 (Probability of the Complement)**.** *For any event A*

$$P(A^c) = 1 - P(A) \tag{7.5}$$

Proof. We have $S = A \cup A^c$ and $A \cap A^c = \varnothing$. Hence, by Axioms 2 and 3,

$$1 = P(S) = P(A) + P(A^c)$$

which proves equation (7.5).

□

Note that Theorem 7.2.2 implies for the special case of $A = \varnothing$

$$P(\varnothing) = 0 \tag{7.6}$$

since in this case $A^c = S$.

Example 7.2.4 (Coin Tosses)**.** What is the probability that at least one head occurs if four fair coins are flipped?

Solution: Each coin turns up either heads or tails. So the sample space has $2^4 = 16$ outcomes. The coins are fair, so we may assign the probability of 1/16 for each outcome. Then the event A^c = "no heads occur" has the probability of one outcome, i.e., 1/16. Hence,

$$P(A) = 1 - P(A^c) = 1 - \frac{1}{16} = \frac{15}{16}$$

□

Theorem 7.2.3 (Addition Property for Any Events)**.** *For any events A and B*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{7.7}$$

Proof. $A \cup B$ may be viewed as the union $E \cup F \cup G$ where E, F, G are as in Fig. 7.2 are mutually exclusive. Hence, by Theorem 7.2.1

$$P(A \cup B) = P(E) + P(F) + P(G)$$

Now on the right $P(E)+P(F)=P(A)$, since E and F are disjoint. Likewise, $P(F)+P(G)=P(B)$. Hence, $P(G)=P(B)-P(F)=P(B)-P(A\cap B)$. Hence,

$$P(E)+P(F)+P(G)=P(A)+P(G)=P(A)+P(B)-P(A\cap B)$$

□

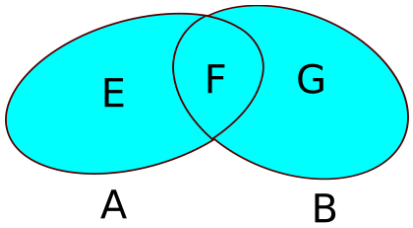


Figure 7.2: Probality Addition Property.

Example 7.2.5 (Union of Events)**.** What is the probability of the event "either an even number or a number less than 5" in rolling a fair die?

Solution: Let A be the event "even number" and let B be the event "a number less than 5". Then by Theorem 7.2.3

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)=\frac{3}{6}+\frac{4}{6}-\frac{2}{6}=\frac{5}{6}$$

□

Conditional Probability and Bayes' Theorem

Often we need the probability of an event B given that an event A occurs. This probability is called the **conditional probability**, $P(B|A)$, of B *given* A. In this case A serves as our new (smaller) sample space and this probability is the fraction of $P(A)$ which corresponds to the event $A\cap B$. So, we have

$$P(B|A)=\frac{P(A\cap B)}{P(A)} \tag{7.8}$$

Equation (7.8) makes sense only if $P(A)\neq 0$.

In the same way we have the *conditional probability of* A *given* B.

$$P(A|B)=\frac{P(A\cap B)}{P(B)} \tag{7.9}$$

Equations (7.8) and (7.9) compute the probability of $A \cap B$ in two ways.

Theorem 7.2.4 (Multiplication Rule)**.** *For events A and B such that $P(A) \neq 0$ and $P(B) \neq 0$*

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B) \tag{7.10}$$

Example 7.2.6. In manufacturing computer monitors let A be the event "the monitor is too wide" and let B be the event "the monitor is too tall". Suppose that $P(A) = 0.15$ and that the conditional probability that a wide monitor is also tall is $P(B|A) = 0.20$. Find the probability that a randomly picked monitor is both too wide and too tall.

Solution: By Theorem 7.2.4

$$P(A \cap B) = P(A)P(B|A) = 0.15 \cdot 0.20 = 0.03$$

So the probability is 0.03, or 3%.

□

Example 7.2.7 (Sampling Without Replacement)**.** Four defective light bulbs are in a box of 20. Two bulbs are randomly drawn, so that the first is drawn checked and kept out, then the second one is drawn and checked. Find the probability that neither of the two bulbs is defective.

Solution: Define the events:

A: the first drawn bulb is not defective.
B: the second drawn bulb is not defective.

We have $P(A) = \frac{16}{20} = \frac{4}{5}$, because 16 out of 20 bulbs are nondefective. Now if A has occurred, there are 19 bulbs left in the box and 4 of them are defective. Hence, $P(B|A) = \frac{15}{19}$. Therefore, by Theorem 7.2.4 both bulbs are nondefective with probability

$$P(A \cap B) = P(A)P(B|A) = \frac{4}{5} \cdot \frac{15}{19} = \frac{12}{19} \approx 0.6315$$

So the probability is about 63%.

□

The following consequence of Theorem 7.2.4 is of interest.

Theorem 7.2.5 (Bayes' Theorem)**.** *For events A and B such that $P(A) \neq 0$ and $P(B) \neq 0$*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7.11}$$

Example 7.2.8. 15% of the patients in a clinic have fever and 6% of the patients have arthritis. If 10% of the patients with arthritis have fever, what is the probability that a patient with fever has arthritis?

Solution: Define the events:

A: the patient has fever.
B: the patient has arthritis.

Then we are given that $P(A) = 0.15$, $P(B) = 0.06$, and $P(A|B) = 0.10$. We need to find $P(B|A)$. By Theorem 7.2.5

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.10 \cdot 0.06}{0.15} = 0.04$$

So 4% of the patients that have fever, they also have arthritis.

□

Definition. Two events A and B are called **independent**, if

$$P(A \cap B) = P(A)P(B) \tag{7.12}$$

If we assume that $P(A) \neq 0$ and $P(B) \neq 0$, then from (7.10) independence implies

$$P(A|B) = P(A), \qquad P(B|A) = P(B)$$

This means that the probability of A does not depend on whether or not B occurs and similarly for the probability of A.

Similarly, n events $A_1, A_2, \ldots, A_n$ are **independent**, if for every subfamily $A_{j_1}, A_{j_2}, \ldots A_{j_k}$ (for all $k = 2, \ldots, n$) we have

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k}) \tag{7.13}$$

For example, the events A, B, and C are independent if and only if

$$\begin{aligned} P(A \cap B) &= P(A)P(B) \\ P(B \cap C) &= P(B)P(C) \\ P(C \cap A) &= P(C)P(A) \\ P(A \cap B \cap C) &= P(A)P(B)P(C) \end{aligned} \tag{7.14}$$

Example 7.2.9 (Sampling With Replacement)**.** Four defective light bulbs are in a box of 20. Two bulbs are randomly drawn, so that the first is drawn checked and is put back into the box. Then the second bulb is drawn and checked. Find the probability that neither of the two bulbs is defective.

Solution: Define the events:

A: the first drawn bulb is not defective.
B: the second drawn bulb is not defective.

We have $P(A) = \frac{16}{20} = \frac{4}{5}$, because 16 out of 20 bulbs are nondefective. Now if we put the bulb back and draw again, then $P(B) = \frac{16}{20} = \frac{4}{5}$. The events are independent and

$$P(A \cap B) = P(A)P(B) = \frac{4}{5} \cdot \frac{4}{5} = \frac{16}{25} = 0.64$$

So the probability is 64%.

□

7.3 Permutations and Combinations

Permutations

Definition. A **permutation** of different given objects is any arrangement of these objects in some order. For example, for the three letters a, b, c, the possible permutations are the following six.

$$abc,\ bca,\ cab,\ bac,\ acb,\ cba$$

Theorem 7.3.1. *For n different objects we have*

1. *The number of permutations the objects is "n factorial".*

$$n! = 1 \cdot 2 \cdot 3 \cdots n \tag{7.15}$$

2. *The number of ways of selecting k objects in a distinct order is*

$${}_nP_k = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!} \tag{7.16}$$

Proof. There n choices to for choosing the first place. Then $n-1$ objects are still available. So, there are $n-1$ ways to to choose the second place, etc.

The proof of Part 2 is similar. □

For the factorial function, we may also use the special notations

$$1! = 1, \quad 0! = 1$$

Example 7.3.1. The probability of drawing the configuration $2, 3, 4, 8, 5, 1, 6, 7$ out of a box that contains balls numbered $1, \ldots, 8$ is

$$\frac{1}{8!} = \frac{1}{40320} \approx 2.48 \times 10^{-5}$$

This is because only one configuration is the desired one out of the 8! available.

□

Example 7.3.2. Suppose we select 3 individuals from 7 to form a committee consisting of President, Treasurer, and Secretary. In how many different ways can this be done?

Solution:

$${}_7P_3 = 7 \times 6 \times 5 = 210$$

□

Note that $n!$ grows very fast with n. For example,

$$20! = 2432902008176640000 \approx 2.43 \times 10^{18}$$

A convenient way to approximate $n!$ for large n is the following formula.

Theorem 7.3.2 (Sterling's Formula)**.** *The function $n!$ can be approximated for large n by*

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad (e = 2.718\cdots)$$

Example 7.3.3. We have

$$20! \sim \sqrt{40\pi}\left(\frac{20}{e}\right)^{20} \approx 2.422786846761133 \times 10^{18}$$

Combinations

Definition. A **combination** of n different objects taken k at a time is any selection of a set of k objects. Note that in a set the *order of elements does not matter.*

For example, for the letters a, b, c, the possible combinations in selecting two, without caring about the order in each sample, are

$$ab,\ ac,\ bc$$

Theorem 7.3.3. *The number of combinations of n different objects choosing k at a time is "n choose k".*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \tag{7.17}$$

Proof. There are ${}_nP_k$ of ways of selecting k objects if the order matters. However, if the order does not matter since there are $k!$ of the k objects we have

$$\binom{n}{k} = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!}$$

□

Example 7.3.4. The number of all possible poker hands is

$$\binom{52}{5} = \frac{52\times 51\times 50\times 49\times 48}{1\times 2\times 3\times 4\times 5} = 2,598,960$$

because we have 52 different cards and we choose 5 at a time without caring in what order the choosen cards are.

The probability to choose one specific poker hand is then

$$\frac{1}{\binom{52}{5}} = \frac{1}{2,598,960} \approx 3.85 \times 10^{-7}$$

□

The numbers $\binom{n}{k}$ are called the **binomial coefficients** because of the following fact.

Theorem 7.3.4. *For a positive integer n*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

The binomial coefficients can be generalized to the case where in place of n we have any real number a. In this case we define

$$\binom{a}{k} = \frac{a(a-1)(a-2)\cdots(a-k+1)}{k!} \qquad (k \geq 0, \text{ integer})$$

and we define

$$\binom{a}{0} = 1$$

The basic properties of the binomial coefficients are outlined in the following theorem.

Theorem 7.3.5. *We have*

1. *For a real and $k \geq 0$ integer*

$$\binom{a+1}{k+1} = \binom{a}{k} + \binom{a}{k+1} \tag{7.18}$$

2. *For $n \geq 0$ and $k \geq 0$ both integers with $k \leq n$*

$$\binom{n}{k} = \binom{n}{n-k} \tag{7.19}$$

3. *For $k \geq 0$ and $n \geq 1$ both integers*

$$\sum_{j=0}^{n-1} \binom{k+j}{k} = \binom{n+k}{k+1} \tag{7.20}$$

4. For a and b real and $r \geq 0$ integer

$$\sum_{k=0}^{r} \binom{a}{k}\binom{b}{r-k} = \binom{a+b}{r} \tag{7.21}$$

Counting with the Multiplication Rule

Theorem 7.3.6 (Event Multiplication Rule)**.** *If M is an experiment with m possible outcomes and N is an experiment with n possible outcomes that does not depend on M, then the experiment which consists of performing M first and then N has mn possible outcomes.*

Example 7.3.5. How many teams of 2 doctors and 3 nurses can be formed from 4 doctors and 6 nurses?

Solution:

$$\binom{4}{2}\binom{6}{3} = 6 \cdot 20 = 120$$

□

Example 7.3.6. The probability of the event "$H5$" in getting "heads" after tossing a coin and 5 after rolling a die is, by the multiplication rule

$$\frac{1}{2 \cdot 6} = \frac{1}{12}$$

7.4 Probability Distribution

Definition. In an experiment that involves randomness a **random variable** is a function X from the sample space S to the real numbers $\mathbf{R}$ such that for each interval I of $\mathbf{R}$ the set $\{s \in S \,|\, X(s) \in I\}$ is an event in S. The image of X is called the **space** of X and it is denoted by R_X.

The random variable X is called **discrete** if its space is finite or countable, i.e., R_X has at most countably many values $x_1, x_2, x_3, \ldots, x_n, \ldots$. Typically, with a discrete random variable, we **count** something.

The random variable X is called **continuous** if its space is either an interval or a union of intervals. Typically, with a continuous random variable, we **measure** something.

Note that a random variable is not really a variable, it is a *function* that takes random values.

Every random variable X is associated with a **probability distribution** P that is defined on the sample space S and takes real values in $[0, 1]$.

For every real number a the probability

$$P(X = a) \tag{7.22}$$

is the probability that X takes the value a. If we consider a to be a variable x then the function

$$f(x) = P(X = x) \tag{7.23}$$

is called the **probability density function** (**PDF**).

For every interval I the probability

$$P(X \in I) \tag{7.24}$$

is the probability that X takes values in the interval I.

The **cumulative distribution function** (**CDF**) of the random variable X is a function F such that

$$F(x) = P(X \leq x) \tag{7.25}$$

This is the probability that X takes on any values not exceeding the real number x.

Equation (7.25) implies the following useful formula

Theorem 7.4.1. *For $a < b$*

$$P(a < X \leq b) = F(b) - F(a) \tag{7.26}$$

Proof. Since the events $X \leq a$ and $a < X \leq b$ are mutually exclusive. Therefore,

$$\begin{aligned} F(b) = P(X \leq b) &= P(X \leq a) + P(a < X \leq b) \\ &= F(A) + P(a < X \leq b) \end{aligned}$$

□

We have the following easily proved properties of the cumulative distribution function.

Theorem 7.4.2 (Properties of CDF)**.** *Let $F(x)$ be the cumulative distribution function of a random variable X. Then*

1. *$F(x) \geq 0$ for all x.*
2. *$F(-\infty) = 0$*
3. *$F(\infty) = 1$*
4. *F is increasing. I.e., if $x < y$, then $F(x) \leq F(y)$.*

Discrete Random Variables

For a discrete random variable X the possible values are at most countably many, say $x_1, x_2, \ldots, x_n, \ldots$. Let $p_1 = P(X = x_1)$, $p_2 = P(X = x_2)$, $p_3 = P(X = x_3)$, $\ldots$, . Then the probability density function $f(x)$ is

$$f(x) = \begin{cases} p_j & \text{if } x = x_j \\ 0 & \text{otherwise} \end{cases} \tag{7.27}$$

The cumulative distribution function is obtained by taking sums.

$$F(x) = \sum_{x_j \leq x} f(x_j) = \sum_{x_j \leq x} p_j \tag{7.28}$$

Note that the graph of $f(x)$ consists of the discrete points (x_j, p_j) and that $F(x)$ is a **step function**.

Two additional formulas are easily obtained for discrete random variables.

$$P(a < X \leq b) = F(b) - F(a) = \sum_{a < x_j \leq b} p_j \tag{7.29}$$

and

$$\sum_j p_j = 1 \tag{7.30}$$

Example 7.4.1. A fair coin is tossed 3 times. Let X be the *discrete* random variable that counts the number of heads.

1. Find the sample space S.
2. Find the space of X.

3. Find the probability density function of X.

4. Find the cumulative distribution function of X.

Solution:

1. S has $2^3 = 8$ elements, namely all sequences ABC where each letter is either H or T.

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

2. The possible values for x are $0, 1, 2, 3$. Hence, the space of X is $R_X = \{0, 1, 2, 3\}$.

3. The probability density function $f(x) = P(X = x)$ is determined by

$$\begin{aligned} f(0) = P(X = 0) &= \frac{1}{8} \\ f(1) = P(X = 1) &= \frac{3}{8} \\ f(2) = P(X = 2) &= \frac{3}{8} \\ f(3) = P(X = 3) &= \frac{1}{8} \end{aligned}$$

4. The cumulative distribution function is

$$F(y) = P(X \leq y) = \sum_{\{x \in R_X \,|\, x \leq y\}} f(x)$$

So for example,

$$\begin{aligned} F(0) = P(X \leq 0) &= \frac{1}{8} \\ F(1) = P(X \leq 1) &= \frac{1}{8} + \frac{3}{8} = \frac{1}{2} \\ F(2) = P(X \leq 2) &= \frac{1}{2} + \frac{3}{8} = \frac{7}{8} \\ F(3) = P(X \leq 3) &= \frac{7}{8} + \frac{1}{8} = 1 \end{aligned}$$

Also note that

$$\begin{aligned} F(1.5) &= P(X \leq 1.5) \\ &= P(X \leq 1) + P(1 < X \leq 1.5) \\ &= \frac{1}{2} + 0 \\ &= \frac{1}{2} \end{aligned}$$

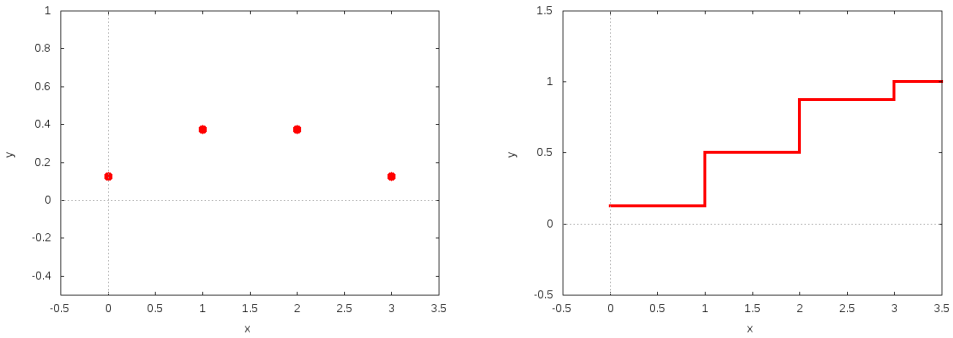


Figure 7.3: Example : Probability density and cumulative distribution.

□

Definition. A **Bernoulli trial** is a random experiment with only two outcomes. The sample space S has two elements. For example, tossing a coin is such experiment with $S = \{H, T\}$. Usually, the two outcomes are designated as "success" (s) and "failure" (f). So $S = \{s, f\}$. In a Bernoulli trial we may define a random variable, X with values

$$X(s) = 1 \qquad \text{and} \qquad X(f) = 0$$

If we denote with p the probability of success, then the PDF $f(x)$ of X is

$$\begin{aligned} f(1) &= P(X = 1) = p \\ f(0) &= P(X = 0) = 1 - p \end{aligned}$$

Hence,

$$f(x) = p^x(1-p)^{1-x} \qquad\qquad (x = 0, 1)$$

Continuous Random Variables

Let X be a continuous random variable. Typically such variable is used in experiments where we measure with uncertainty. For example, measuring the voltage or current of an electrical appliance, or measuring the length of rods used in engines, etc.

For a continuous random variable the **probability density function** is a continuous function $f : \mathbf{R} \to [0, \infty)$ such that for every subset of real numbers A we have

$$P(X \in A) = \int_A f(x)\, dx \tag{7.31}$$

For example,

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx \tag{7.32}$$

It can be seen that

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{7.33}$$

The cumulative distribution function $F(x)$ is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{7.34}$$

The connection between F and f is that

$$F'(x) = f(x) \tag{7.35}$$

Note that we may assume more generally that the density function $f(x)$ is piecewise continuous.

The probability $P(a < X \leq b)$ is computed by

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(t)\, dt \tag{7.36}$$

Note that for a continuous random variable since the cumulative distribution function is given as an integral we may add a point without changing the value of the integral. Hence, we have

$$P(a < X) = P(a \leq X) \tag{7.37}$$

Also, note that $f(x)$ is **not** $P(X = x)$ as it was in the discrete case. The connection between the probability density f and the probability distribution P that the probability that X takes values in the infinitesimal interval $x \leq X \leq x + \Delta x$ is approximately $f(x)\Delta x$.

$$P(x \leq X \leq x + \Delta x) \approx f(x)\Delta x \tag{7.38}$$

Example 7.4.2. Let X be a continuous random variable with density function

$$f(x) = \begin{cases} \frac{3}{5} - \frac{3}{10}x^2 & \text{if } x \in [-1, 1] \\ 0 & \text{otherwise} \end{cases} \tag{7.39}$$

1. Verify equation (7.33)
2. Find the cumulative distribution function.
3. Find the probability $P(-\frac{1}{2} \leq X \leq \frac{1}{2})$.
4. Find the probability $P(\frac{1}{2} \leq X \leq 3)$.
5. Find all x such that $P(X \leq x) = 0.9$.

Solution: We have

1. $$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-1}^{1} \left(\frac{3}{5} - \frac{3}{10}x^2\right)\,dx = \left(\frac{3x}{5} - \frac{x^3}{10}\right)\Big|_{-1}^{1} = 1$$

2. For $-1 < x \leq 1$ we have

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt \\ &= \int_{-1}^{x} \left(\frac{3}{5} - \frac{3}{10}t^2\right)\,dt \\ &= -\frac{x^3}{10} + \frac{3x}{5} + \frac{1}{2} \end{aligned}$$

Notice that we can check that $F'(x) = f(x)$.

3. We have

$$\begin{aligned} P\left(-\frac{1}{2} \leq X \leq \frac{1}{2}\right) &= F\left(\frac{1}{2}\right) - F\left(-\frac{1}{2}\right) \\ &= \int_{-1/2}^{1/2} \left(\frac{3}{5} - \frac{3}{10}x^2\right) dx \\ &= \frac{23}{40} \\ &= 0.575 \\ &= 57.5\% \end{aligned}$$

4. We have

$$\begin{aligned} P\left(\frac{1}{2} \leq X \leq 3\right) &= \int_{1/2}^{3} f(x)\, dx \\ &= \int_{1/2}^{1} \left(\frac{3}{5} - \frac{3}{10}x^2\right) dx + 0 \\ &= \frac{17}{80} \\ &= 0.2125 \\ &= 21.25\% \end{aligned}$$

5. We set $P(X \leq x) = F(x) = 0.9$. Hence, we solve the cubic

$$-\frac{x^3}{10} + \frac{3x}{5} + \frac{1}{2} = \frac{9}{10}$$

for x to get three roots: 2, $\sqrt{3} - 1$, and $-1 - \sqrt{3}$. Out of these roots we keep $x = \sqrt{3} - 1 \approx 0.732$ which the only root in the interval $[-1, 1]$.

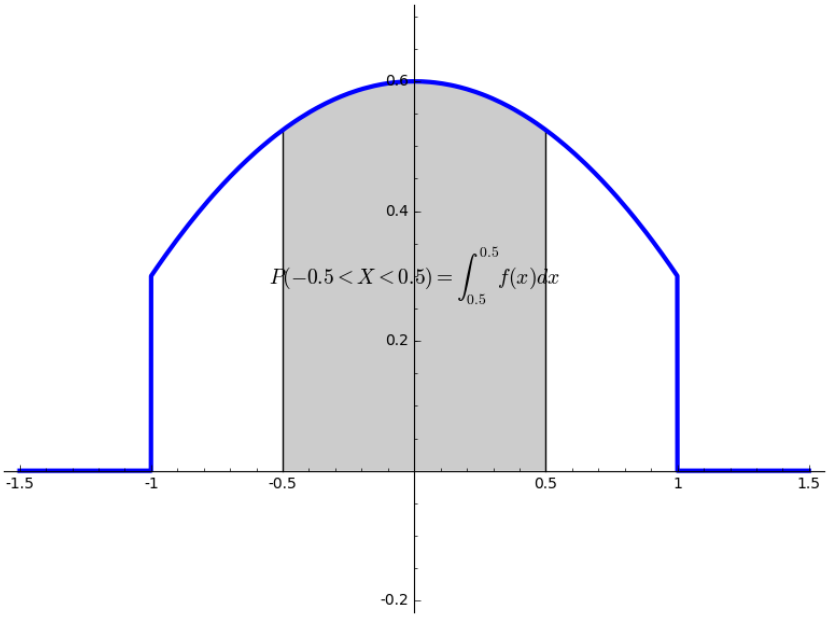


Figure 7.4: Example 7.39: Probability density function.

□

Definition. A continuous random variable is called **uniform on the interval** $[a, b]$, if its probability density function $f(x)$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{7.40}$$

The corresponding probability distribution is called **uniform**.

Note that $f(x) \geq 0$ and that $\int_{-\infty}^{\infty} f(x) = 1$.

The cumulative distribution function for the uniform distribution is easily seen to be

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b \end{cases} \tag{7.41}$$

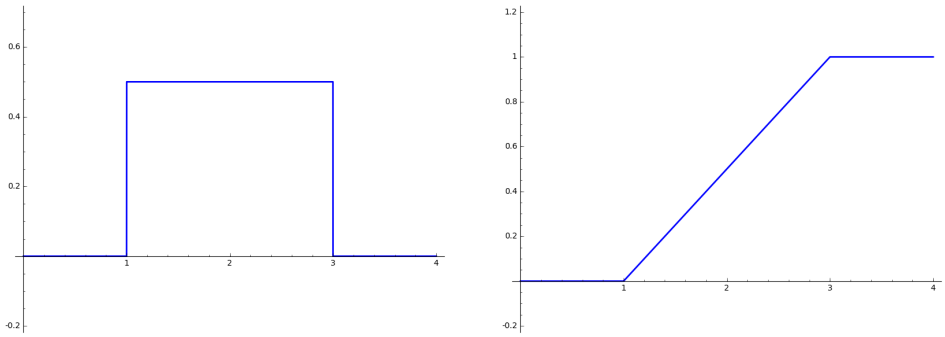


Figure 7.5: The PDF and CDF of the uniform distribution on $[1, 3]$.

7.5 Mean, Variance, and Expectation

Definition. The **mean** μ (mu) of a random variable X with PDF $f(x)$ is defined by

$$\mu = \sum_j x_j f(x_j) \qquad \text{(Discrete distribution)} \tag{7.42}$$

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{(Continuous distribution)} \tag{7.43}$$

The mean is the weighted average of the values of X with weights the probabilities, given by the PDF.

The mean μ is also denoted by $E(X)$ and it is called the **expectation** of X, because it gives the expected average value of X after many trials.

Definition. The **variance** σ^2 (sigma square) of X is defined by

$$\sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \qquad \text{(Discrete distribution)} \tag{7.44}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \qquad \text{(Continuous distribution)} \tag{7.45}$$

The variance is also denoted by $\mathrm{Var}(X)$. Note that

$$\mathrm{Var}(X) = E\left([X - \mu]^2\right) \tag{7.46}$$

The variance measures the *spread* of the values of X. If σ^2 large, then the spread of the values is large and vice versa.

The positive square root σ of the variance is called the **standard deviation** of X.

Note that

$$\sigma^2 > 0 \tag{7.47}$$

with the exception of the discrete distribution that takes on only one value, in which case it is zero.

For a given random variable X, the mean μ and the variance σ^2 may or may not exist. We assume that they exist, which is the case in most applications.

Example 7.5.1. A fair coin is tossed 3 times. Let X be the *discrete* random variable that counts the number of heads. Find the mean and variance of X.

Solution: In Example 7.3 we found the space $R_X = \{0, 1, 2, 3\}$ and the PDF of X

$$f(0) = \frac{1}{8}, \quad f(1) = \frac{3}{8}, \quad f(2) = \frac{3}{8}, \quad f(3) = \frac{1}{8}$$

The mean for X is computed by

$$\mu = \sum_{j=1}^{4} x_j f(x_j) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

This means that after many trials when we toss three coins we expect to get about 1.5 heads. (Does this make sense?)

The variance of X is

$$\begin{aligned}
\sigma^2 &= \sum_{j=1}^{4} (x_j - \mu)^2 f(x_j) \\
&= \left(0 - \frac{3}{2}\right)^2 \frac{1}{8} + \left(1 - \frac{3}{2}\right)^2 \frac{3}{8} + \left(2 - \frac{3}{2}\right)^2 \frac{3}{8} + \left(3 - \frac{3}{2}\right)^2 \frac{1}{8} \\
&= \frac{3}{4}
\end{aligned}$$

□

Example 7.5.2 (Mean and Variance of the Uniform Distribution)**.** Recall that the uniform distribution on the interval $[a, b]$ has density

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Compute the mean and variance.

Solution: The mean is

$$\begin{aligned}\mu &= \int_{-\infty}^{\infty} x f(x)\,dx \\ &= \int_a^b \frac{x}{b-a}\,dx \\ &= \frac{a+b}{2}\end{aligned}$$

The variance is

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \\ &= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\,dx \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

□

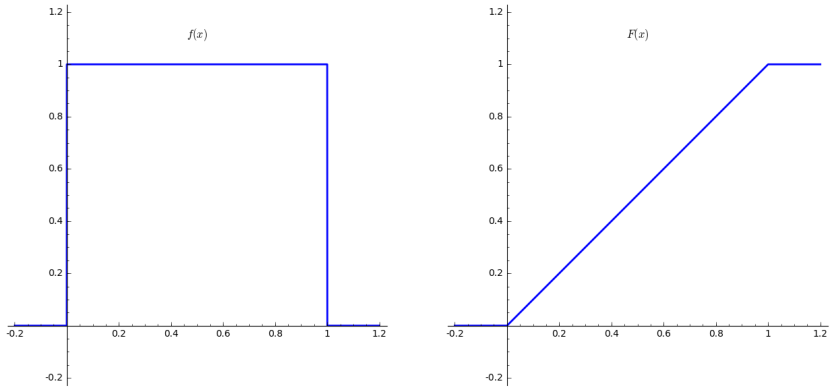
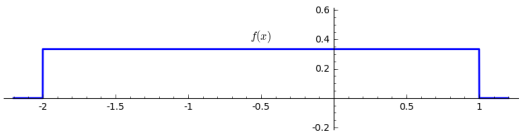


Figure 7.6: The uniform distribution on $[0,1]$ has $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{12}$.



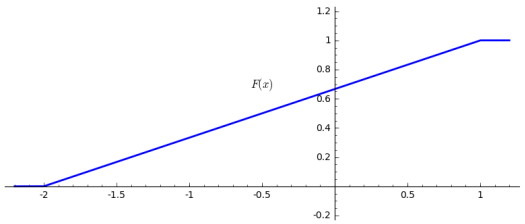


Figure 7.7: The uniform distribution on $[-1, 2]$ has $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{3}{4}$.

Theorem 7.5.1. *For a random variable X and any real numbers a and b we have*

$$E(aX + b) = aE(X) + b \tag{7.48}$$

Proof. The proof follows form the linearity of integrals. □

Theorem 7.5.2. *For a random variable X with mean μ and variance σ^2 we have*

$$\sigma^2 = E(X^2) - \mu^2 \tag{7.49}$$

Equation (7.49) is also written as

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 \tag{7.50}$$

Proof. By Theorem 7.5.1 we have

$$\begin{aligned} \sigma^2 &= E\left([X - \mu]^2\right) \\ &= E\left(X^2 - 2\mu X + \mu^2\right) \\ &= E\left(X^2\right) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

□

Another basic property of the variance that is easily proved is the following.

Theorem 7.5.3. *For a random variable X with mean μ and any real numbers a and b we have*

$$\mathrm{Var}(aX + b) = a^2 \,\mathrm{Var}(X) \tag{7.51}$$

Theorems 7.5.1 and 7.5.3 have the following immediate and useful consequence.

Theorem 7.5.4. *Let X be a random variable with mean μ and variance σ^2. The* ***standardized random variable*** *Z corresponding to X*

$$Z = \frac{X - \mu}{\sigma} \tag{7.52}$$

has mean 0 and variance 1.

Moments

We generalize the notion of expected value to the case that we have a function of the given random variable X. Let $g(x)$ be a nonconstant continuous function. Then $g(X)$ is a new random variable whose expected value is the values of $g(X)$ on the average and it is defined similarly to (7.42) and (7.43) by

$$E(g(X)) = \sum_j g(x_j) f(x_j) \quad \text{or} \quad E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\, dx \tag{7.53}$$

where $f(x)$ is the PDF of X.

An important special case of this generalization is the **kth moment** of X, defined for $k = 1, 2 \ldots$, by

$$E(X^k) = \sum_j x_j^k f(x_j) \quad \text{or} \quad E(X^k) = \int_{-\infty}^{\infty} x^k f(x)\, dx \tag{7.54}$$

Also of interest is the **kth central moment** of X, defined for $k = 1, 2 \ldots$, by

$$E\left([X - \mu]^k\right) = \sum_j (x_j - \mu)^k f(x_j) \quad \text{or} \quad \int_{-\infty}^{\infty} (x - \mu)^k f(x)\, dx \tag{7.55}$$

7.6 Binomial, Poisson, and Hypergeometric Distributions

In this section we study three important *discrete* distributions each having a great number of applications.

Binomial Distribution

We are interested in the number of times an event A occurs after n independent trials. For example, the number of heads in 10 tossings of a fair coin, the number of patients that respond positively after taking a medication 20 times, or the number of defective items after taking 8 samples of the same size.

Let the probability that the event A occurs after one execution of the experiment be $P(A) = p$. So the probability that A does not occur is $q = 1 - P(A) = 1 - p$. Let

$$X = \text{Number of times } A \text{ occurs in } n \text{ independent trials}$$

The possible values of X are $0, 1, 2, \ldots, n$. Our goal to compute the probabilities. If $X = x$, then A occurs in x trials and $B = A^c$ the event that A does not occur in $n - x$ trials. So the result of n trials may look like

$$\underbrace{AA\cdots A}_{x \text{ times}}\ \underbrace{BB\cdots B}_{n-x \text{ times}}$$

or like

$$\underbrace{BB\cdots B}_{n-x \text{ times}}\ \underbrace{AA\cdots A}_{x \text{ times}}$$

The probability for each such occurrence is the product of the probabilities of each of A and B because the trials are independent (see (7.13) on independent events).

$$\underbrace{pp\cdots p}_{x \text{ times}} \cdot \underbrace{qq\cdots q}_{n-x \text{ times}} = p^x q^{n-x}$$

Now we have $\binom{n}{x}$ possible such events because we choose x number of A out of n trials. The Bs are automatic once we have the As. Hence, X has probability density function

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \ldots, n) \tag{7.56}$$

The distribution with probability density function (7.56) is called a **binomial distribution**. The occurrence of A is called **success** and the nonoccurrence of A is called **failure**.

The special case where we have equal chance of success and failure, i.e., the so-called **symmetric case**, where $p = q = \frac{1}{2}$, we get

$$f(x) = \binom{n}{x} \left(\frac{1}{2}\right)^n \qquad (x = 0, 1, \ldots, n)$$

Example 7.6.1. A fair die is rolled 5 times. Find the probability of obtaining

1. exactly three "six",
2. at least three "six".
3. at most three "six".

Solution: Let $A =$"six". Then $P(A) = \frac{1}{6}$, $q = \frac{5}{6}$, and $n = 5$.

1. The probability of having exactly 3 occurrences of A is
$$f(3) = \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 = \frac{125}{3888} \approx 0.0321 = 3.21\%$$
2. The probability of having at least 3 occurrences of A is
$$\begin{aligned} f(3)+f(4)+f(5) &= \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 \\ &\quad + \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1 \\ &\quad + \binom{5}{5}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^0 \\ &= \frac{23}{648} \approx 0.0354 = 3.54\% \end{aligned}$$
3. The probability of having at most 3 occurrences of A is
$$\begin{aligned} f(0)+f(1)+f(2)+f(3) &= \binom{5}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^5 \\ &\quad + \binom{5}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^4 \\ &\quad + \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 \\ &\quad + \binom{5}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^2 \\ &= \frac{3875}{3888} \approx 0.9966 = 99.66\% \end{aligned}$$

□

We have the following formulas for the mean and variance of the binomial distribution.

Theorem 7.6.1. *For the n-trial binomial distribution with probability of success p and $q = 1 - p$, we have*

$$\mu = np, \qquad \sigma^2 = npq \tag{7.57}$$

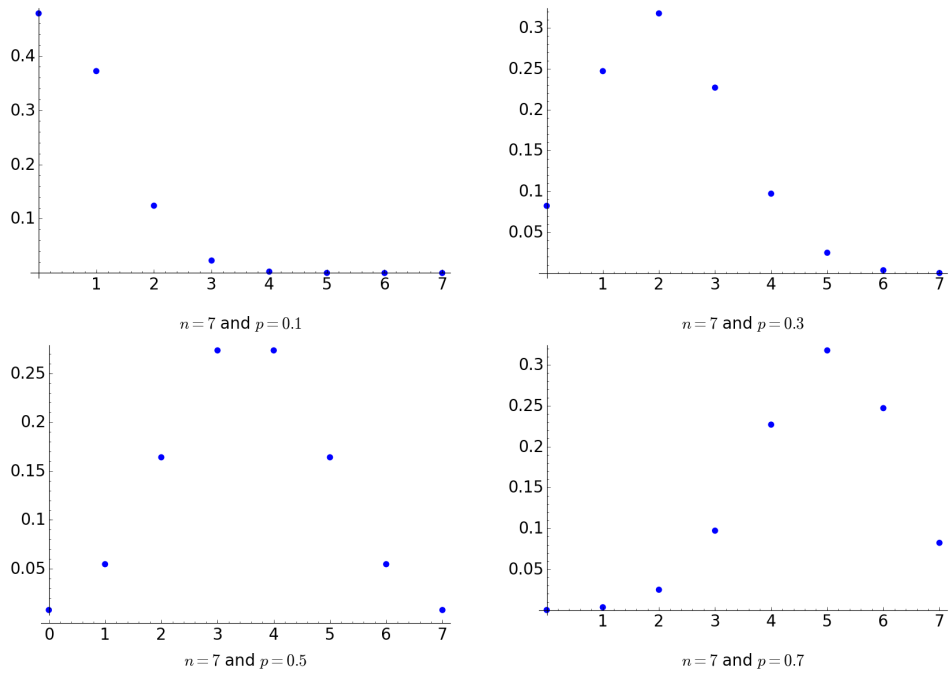


Figure 7.8: The binomial distribution PDF for $n = 7$, $p = 0.1, 0.3, 0.5, 0.7$.

Poisson Distribution

A **Poisson distribution** is a discrete distribution with infinitely (countably) many possible values and probability density function

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \qquad (\mu > 0,\ x = 0, 1, 2, \dots) \tag{7.58}$$

This distribution is named after S. D. Poisson and it can be viewed as a limiting case of the binomial distribution, if we let $p \to 0$ and $n \to \infty$ in such a way that the mean of the binomial $\mu = np$ approaches a finite value or even kept constant. In fact, for the Poisson distribution the mean is μ in formula (7.58) and it equals its variance.

$$\mu = \sigma^2$$

The Poisson distribution is a good model for experiments where events occur in a fixed time interval at a constant rate and do not depend on the time of a previous event. For example, a person keeping track of the amount of daily received emails notices that average number is, say, 50, and the emails arrive independently and from a wide range of individuals.

In formula (7.58), x represents the number of times an event occurs. We assume that events occur independently. The constant μ which is the mean in this case, represents the average number of events per fixed interval.

Example 7.6.2. A person receives on average 4 emails every hour. Assuming the Poisson model is suitable, find the probabilities that

1. 0, 2, 4, and 7 emails are received (each being a separate case).
2. at most 2 emails are received.
3. at least 3 emails are received.

Solution: The mean is $\mu = 4$. Let x be the number of emails received in a hour. We have

$$P(X = x) = f(x) = \frac{4^x e^{-4}}{x!}$$

1. We have

$$P(X = 0) = f(0) = \frac{4^0 e^{-4}}{0!} = e^{-4} \approx 0.018 = 1.8\%$$
$$P(X = 2) = f(1) = \frac{4^2 e^{-4}}{2!} = 8e^{-4} \approx 0.146 = 14.6\%$$
$$P(X = 4) = f(4) = \frac{4^4 e^{-4}}{4!} = \frac{32e^{-4}}{3} \approx 0.195 = 19.5\%$$
$$P(X = 7) = f(7) = \frac{4^7 e^{-4}}{7!} = \frac{1024e^{-4}}{315} \approx 0.059 = 5.9\%$$

2. For at most 2 emails we need $P(X \leq 2) = f(0) + f(1) + f(2)$.

$$f(0)+f(1)+f(2) = \frac{4^0 e^{-4}}{0!}+\frac{4^1 e^{-4}}{1!}+\frac{4^2 e^{-4}}{2!} = 13e^{-4} = 0.2381 = 23.81\%$$

3. We have

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - 0.2381 = 0.7619 = 76.19\%$$

□

Example 7.6.3. The probability of producing a defective light bulb is $p = 0.01$. Suppose we sample 200 bulbs. Find the probability that the sample has more than 2 defectives by using a

1. binomial distribution.

2. Poisson distribution.

Solution: Let A be the event "more than 2 defectives". Then the complementary event A^c is "no more than 2 defectives".

1. With the binomial distribution.

$$P(A^c) = \binom{200}{0} 0.99^{200} + \binom{200}{1} 0.01 \cdot 0.99^{199} + \binom{200}{2} 0.01^2 \cdot 0.99^{198}$$

   Hence, $P(A^c) = 0.676678$ and $P(A) = 0.323321 \approx 32.33\%$.

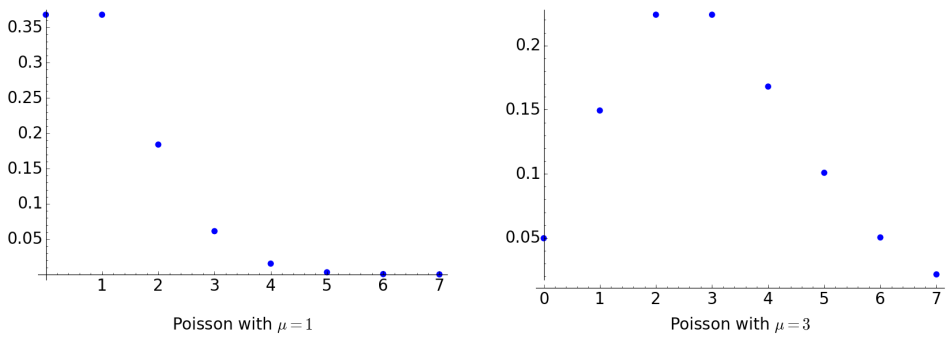
2. With the Poisson distribution: Since $p = 0.01$ is very small, and $n = 200$ large we can approximate this by the much easier to use Poisson distribution with mean $np = 200 \cdot 0.01 = 2$.

$$P(A^c) = e^{-2} \left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} \right) = 0.676676$$

   Hence, $P(A) = 0.676676$ and $P(A^c) = 0.323323 \approx 32.33\%$.

We observe that the Poisson distribution gave us essentially the same answer with much less effort.

□



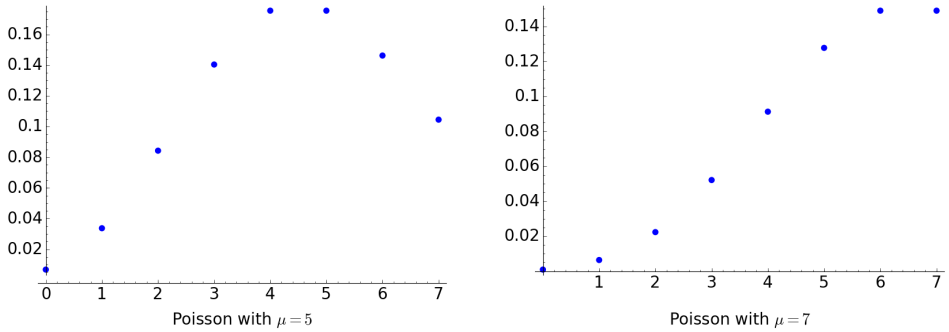


Figure 7.9: The Poisson distribution PDF for $p = 1, 3, 5, 7$.

Sampling With Replacement

An important application of the binomial distribution is in **sampling with replacement**. We draw from a set of objects one by one and after each trial we put it back before we draw again. This makes the trials independent and leads us to the binomial distribution.

Suppose that in a box with m objects there are r defective. So the probability that we draw a defective one is $p = \frac{r}{m}$. Therefore, the probability that we draw a nondefective object is $q = 1 - p = 1 - \frac{r}{m}$. So, in the case of drawing with replacement the probability of x defectives in n trials is

$$f(x) = \binom{n}{x} \left(\frac{r}{m}\right)^x \left(1 - \frac{r}{m}\right)^{n-x} \qquad (x = 0, 1, \ldots, n) \qquad (7.59)$$

by the binomial distribution formula (7.56).

Sampling Without Replacement; Hypergeometric Distribution

In **sampling without replacement**, we draw from a set of objects one by one but we do not put them back. The trials are no longer independent. Instead of (7.59), the probability of drawing without replacement of x defectives in n trials, if there are r defectives in a total of m, is

$$f(x) = \frac{\binom{r}{x}\binom{m-r}{n-x}}{\binom{m}{n}} \qquad (x = 0, \ldots, n,\ x \leq r,\ n \leq m) \qquad (7.60)$$

The distribution with probability density function (7.60) is called a **hypergeometric distribution** .

Formula (7.60) is found as follows. There are

1. $\binom{r}{x}$ ways of picking x defectives out of r.
2. $\binom{m-r}{n-x}$ ways of picking $n-x$ nondefectives out of $m-r$.
3. $\binom{m}{n}$ ways of picking n objects out of m.

Now each way in Part 1 combined with each way in Part 2 gives the total number of distinct ways of getting x defectives in n drawings without replacement. This number is the product $\binom{r}{x}\binom{m-r}{n-x}$. If we divide by the total number of possible outcomes in Part 3 which is $\binom{m}{n}$, we get (7.60).

Theorem 7.6.2. *The hypergeometric distribution has mean*

$$\mu = \frac{nr}{m} \tag{7.61}$$

and variance

$$\sigma^2 = \frac{nr(m-r)(m-n)}{m^2(m-1)} \tag{7.62}$$

Example 7.6.4. We draw randomly and without replacement 4 screws out of a box of 30 screws where we know that 5 screws are defective.

1. What is the probability that 2 screws are defective?
2. What is the probability that at most 2 screws are defective?
3. Compare your answer in Part 1, if the sampling was done with replacement.

Solution: In the notation of (7.60) we have $m = 30$, $r = 5$, $n = 4$. Hence,

$$f(x) = \frac{\binom{5}{x}\binom{25}{4-x}}{\binom{30}{4}}$$

1. $f(2) = \dfrac{\binom{5}{2}\binom{25}{2}}{\binom{30}{4}} = \dfrac{200}{1827} \approx 0.1094 = 10.94\%$.

2. The probabilty is $f(0) + f(1) + f(2)$ or

$$\frac{\binom{5}{0}\binom{25}{4}}{\binom{30}{4}} + \frac{\binom{5}{1}\binom{25}{3}}{\binom{30}{4}} + \frac{\binom{5}{2}\binom{25}{2}}{\binom{30}{4}} = \frac{1810}{1827} \approx 99.06\%$$

3. For replacement we use the binomial distribution (7.59).

$$f(2) = \binom{4}{2}\left(\frac{5}{30}\right)^2\left(1 - \frac{5}{30}\right)^{4-2}\frac{25}{216} \approx 0.1157 = 11.57\%$$

The probability is slightly higher than in Part 1, as expected.

□

7.7 The Normal Distribution

In this section we study the normal distribution, introduced by Abraham DeMoivre. This is the most important continuous distribution. It has numerous applications. Many random variables are either normal, or can be transformed to normal, or can be approximated by normal. Laplace and Gauss applied normal distributions to astronomy and physics.

Definition. A **normal distribution** or *Gauss distribution*, is a distribution with density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (-\infty < x < \infty) \tag{7.63}$$

For a **normal random variable** X with density (7.63) we write

$$X \sim N(\mu, \sigma^2) \tag{7.64}$$

Formula (7.63) leads to some important observations.

1. $f(x) \geq 0$ for all x.

2. The value $f(x)$ approaches zero fast as x goes to $-\infty$ or to ∞.

3. The graph of $f(x)$ is symmetric about the vertical line $x = \mu$.

4. The maximum of $f(x)$ occurs at $x = \mu$ and then the values decrease symmetrically around $x = \mu$, resulting into a *bell-shaped* curve.

5. The constant factor $\frac{1}{\sigma\sqrt{2\pi}}$ makes the area under the graph of $f(x)$ from $-\infty$ to ∞ equal to 1.

6. The constants μ and σ are respectively, the mean and the standard deviation of this distribution.

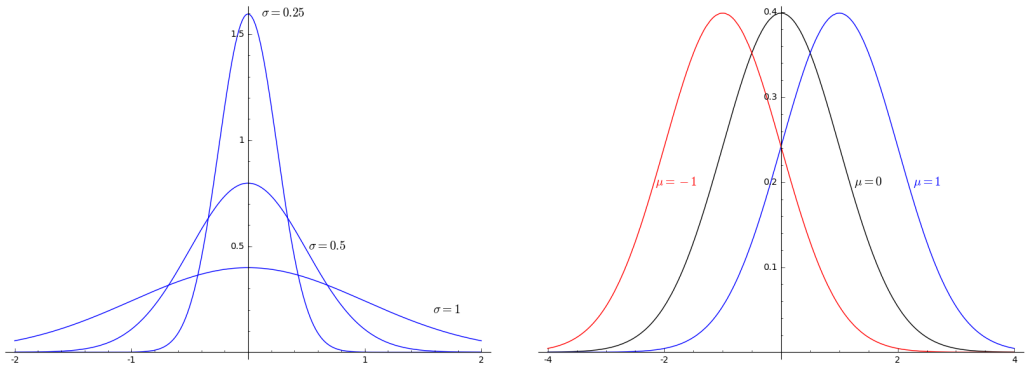


Figure 7.10: PDF of normal distributions: $\mu = 0$ (left), $\sigma = 1$ (right).

Theorem 7.7.1. *If* $X \sim N(\mu, \sigma^2)$, *then*

$$E(X) = \mu, \qquad \text{Var}(X) = \sigma^2 \tag{7.65}$$

Definition. The normal random variable $Z \sim N(0, 1)$ is called **standard normal** . The probability density in this case becomes

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad (-\infty < z < \infty) \tag{7.66}$$

The standard normal variable Z is important, because any normal variable X can be converted to Z by a simple transformation.

Theorem 7.7.2. *If* $X \sim N(\mu, \sigma^2)$, *the random variable* $Z = \frac{X-\mu}{\sigma}$ *is* $N(0, 1)$.

The cumulative distribution function of a normal distribution is given by

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt \tag{7.67}$$

where, $F(x) = P(X \leq x)$.

The integral in (7.67) cannot be evaluated in terms of elementary functions such as exponentials, logarithms, trigonometric functions, etc. However, it can be approximated for any value of x.

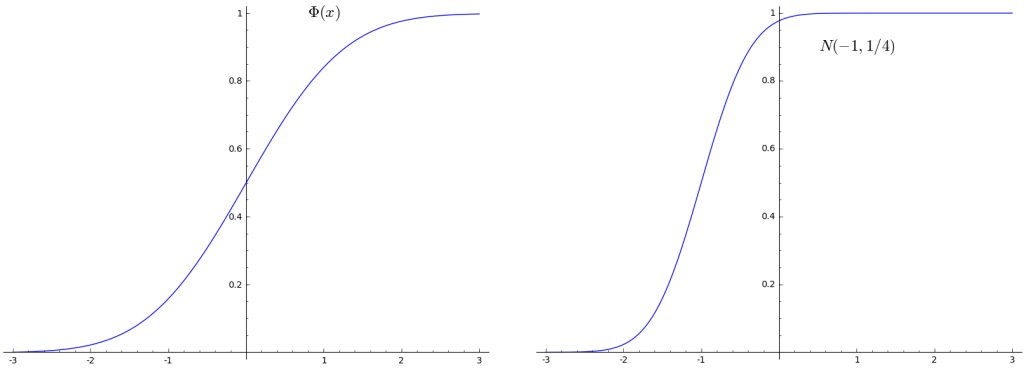


Figure 7.11: CDF of normal distributions: $N(0,1)$ (left), $N(-1,1/4)$ (right).

The special CDF $\Phi(z) = P(Z \leq z)$ of the standard normal distribution is of particular interest.

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt \tag{7.68}$$

This is because by Theorem 7.7.2 we can transform any normal variable X to the standard normal Z, then use precomputed tables for Z to evaluate X.

A useful property of $\Phi(z)$ is its rotational symmetry about the point $\left(0, \frac{1}{2}\right)$. This implies the identity

$$\Phi(-z) = 1 - \Phi(z) \tag{7.69}$$

The connection between the general CDF $F(x)$ and the standard CDF $\Phi(z)$ is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{7.70}$$

To compute $\Phi(z)$ values we use the book Tables A7 and A8 of the Appendix A5.

Example 7.7.1 (Using Table A7)**.** For the standard normal $Z \sim N(0,1)$ we have

$$
\begin{aligned}
&P(X \leq 2.55) \approx 0.9946 = 99.46\% \\
&P(X \leq -1.15) = 1 - \Phi(1.15) = 1 - 0.8749 = 0.1251 \approx 12.51\%(\text{ by (7.69)} \\
&P(X \geq 2) = 1 - P(X \leq 2) = 1 - 0.9772 = 0.0228 \approx 2.3\% \\
&P(1 \leq X \leq 1.55) = \Phi(1.5) - \Phi(1.0) = 0.9332 - 0.8643 = 0.0689 \approx 6.9\%
\end{aligned}
$$

Example 7.7.2. If $X \sim N(2, 64)$, what is $P(4 \leq X \leq 12)$?

Solution: We have

$$
\begin{aligned}
P(4 \leq X \leq 12) &= P\left(\frac{4-2}{8} \leq \frac{X-2}{8} \leq \frac{12-2}{8}\right) \\
&= P\left(\frac{1}{4} \leq Z \leq \frac{5}{4}\right) \\
&= P(Z \leq 1.25) - P(Z \leq 0.25) \\
&= 0.8944 - 0.5987 \qquad \text{(from book Table A7, APP. A5)} \\
&= 0.2957
\end{aligned}
$$

□

The normal distribution $X \sim N(\mu, \sigma^2)$ has an interesting property that is worth remembering. About 68% of its values lie between $\mu \pm \sigma$, about 95% lie between $\mu \pm 2\sigma$, and about 99.7% between $\mu \pm 3\sigma$. So, practically, all the values of X lie between the **three-sigma limits** $\mu \pm 3\sigma$.

$$
\begin{aligned}
&P(\mu - \sigma < X \leq \mu + \sigma) \approx 68\% \\
&P(\mu - 2\sigma < X \leq \mu + 2\sigma) \approx 95.5\% \\
&P(\mu - 3\sigma < X \leq \mu + 3\sigma) \approx 99.7\%
\end{aligned}
$$

7.8 Student's t Distribution; The Chi-Squared Distribution

In this section we introduce Student's t-distribution and the chi-squared distribution. These are very useful in statistics and they are closely related to the the normal distribution.

The Gamma Function

First we recall a few facts about the important gamma function, $\Gamma(\alpha)$, introduced by Euler. This is an extension of the factorial function $n!$ over the real numbers and also over the complex numbers. It appears in probability, in statistics, in combinatorics, and in several applications, such as the vibrations of a circular membrane.

The gamma function is defined for a positive real number α by the indefinite integral

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \tag{7.71}$$

Its basic properties are summerazed here.

1. The functional equation holds.

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \tag{7.72}$$

2. For any positive integer n

$$\Gamma(n+1) = n! \tag{7.73}$$

3. For any positive integer n

$$\Gamma\left(n+\frac{1}{2}\right) = \frac{3\cdot 5\cdot 7\ldots(2n-1)}{2^n}\sqrt{\pi} \tag{7.74}$$

More generally, we have the **upper incomplete gamma function** which is defined by

$$\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1} e^{-t} dt \tag{7.75}$$

and the **lower incomplete gamma function** defined by

$$\gamma(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t} dt \tag{7.76}$$

Student's t-Distribution

Definition. The probability distribution of a continuous random variable T is **Student's t-distribution of ν degrees of freedom**, if its density probability function is given by

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{7.77}$$

for $-\infty < x < \infty$.

The t-distribution is bell-shaped and symmetric about y-axis, just like the **standard normal distribution**. However, its values do not drop off so fast as we move away from the peak. In other words, it is more spread out than the normal distribution.

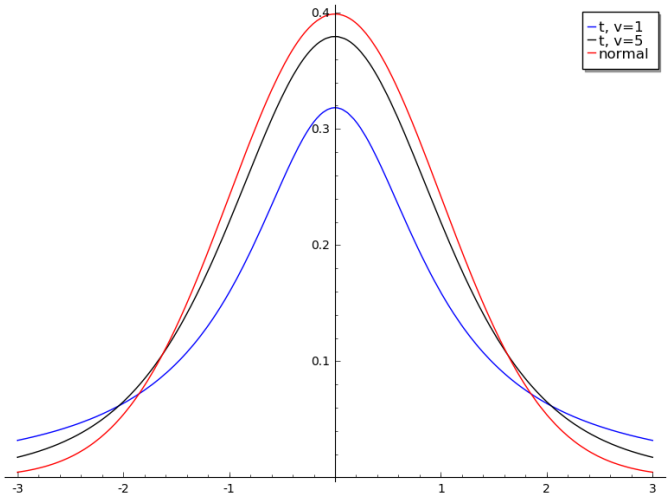


Figure 7.12: t with $\nu = 1$ (blue), $\nu = 5$ (black), and normal (red).

A main property of the t-distribution is that *as the degrees of freedom ν increase, then the t-distribution approaches the standard normal distribution.*
Given these properties the first factor in equation (7.77) can be simplified.

If $\nu > 1$ is even, then

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} = \frac{(\nu-1)(\nu-3)\cdots 5\cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4\cdot 2}$$

If $\nu > 1$ is odd, then

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} = \frac{(\nu-1)(\nu-3)\cdots 4\cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5\cdot 3}$$

The cumulative probability of the t-distribution is

$$F(z;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \int_{-\infty}^{z} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt \tag{7.78}$$

This integral has values that of practical importance. So approximations of several of its values are tabulated.

Appendix 5, Table A9 in the text has values z and $F(z)$ for various degrees of freedom for the t-distribution.

Just as with the normal distribution we have rotational symmetry about the point $\left(0, \frac{1}{2}\right)$ and therefore the following identity holds.

$$F(-z;\nu) = 1 - F(z;\nu) \tag{7.79}$$

Example 7.8.1. The random variable Z follows a t-distribution with 5 degrees of freedom. Find $P(|Z| \leq 1.48)$.

Solution: By using Table A9, Appendix 5 and Equation (7.79) we have

$$\begin{aligned} P(|Z| \leq 1.48) &= P(-1.48 \leq Z \leq 1.48) \\ &= F(1.48;5) - F(-1.48;5) \\ &= F(1.48;5) - (1 - F(1.48;5)) \\ &= 2F(1.48;5) - 1 \\ &= 2(0.9) - 1 \\ &= 0.8 \end{aligned}$$

So $P(|Z| \leq 1.48) = 80\%$.

□

The Chi-Squared Distribution

In this paragraph we introduce the chi-squared distribution which is very useful in the hypothesis testing in statistics.

Definition. The probability distribution of a continuous random variable T is **chi-squared distribution of k degrees of freedom**, or **χ^2-distribution**, if its density probability function is given by

$$f(x;k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{7.80}$$

Fact The distribution of a sum of the squares of k independent standard normal random variables is a chi-squared distribution with k degrees of freedom. In particular, the square of a standard normal distribution is a chi-squared distribution with one degree of freedom.

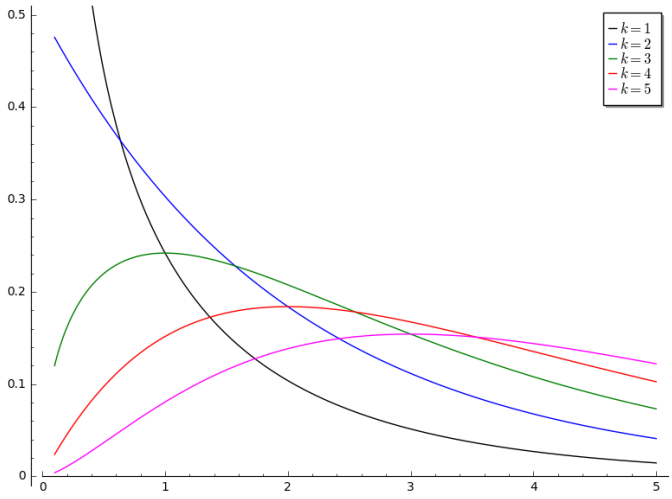


Figure 7.13: χ^2-distribution PDFs with $k = 1, 2, 3, 4, 5$.

The cumulative distribution of chi-squared distribution is given by

$$F(x;k) = \frac{\gamma\left(\frac{k}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \tag{7.81}$$

where γ is the lower incomplete gamma function (7.76).

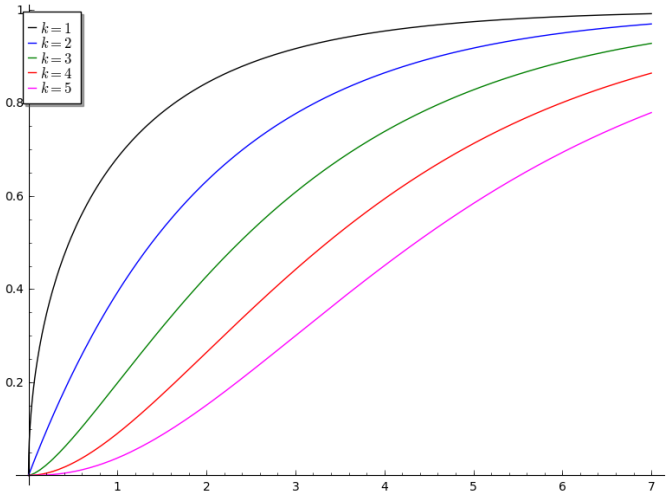


Figure 7.14: χ^2-distribution CDFs with $k = 1, 2, 3, 4, 5$.

Example 7.8.2. The random variable Z follows a χ^2-distribution with 5 degrees of freedom. Find $P(1.15 \leq Z \leq 12.83)$.

Solution: By using Table A10, Appendix 5, we have

$$\begin{aligned} P(1.15 \leq Z \leq 12.83) &= F(12.83; 5) - F(1.15; 5) \\ &= 0.975 - 0.05 \\ &= 0.925 \\ &= 92.5\% \end{aligned}$$

7.9 Two Random Variables

Some random experiments involve distributions of two or more random variables. For example, the joint study the iron content X and hardness Y of steel, or the height X_1, weight X_2, and blood pressure X_3 of a person.

We study such distributions because they

- appear very often, and
- can be used in the mathematical justification of statistical methods.

We concentrate on the study of **two-dimensional random variable** (X, Y). The outcome of a trial of (X, Y) is a pair of numbers x and y such

that $X = x$ and $Y = y$, which we write as $(X, Y) = (x, y)$. So, $(X, Y) = (x, y)$ means the *intersection* of events $X = x$ and $Y = y$.

Definition. The **two-dimensional joint cumulative distribution function** $F(x, y)$ of the random variable (X, Y) is given by the probability distribution function

$$F(x, y) = P(X \leq x, Y \leq y) \tag{7.82}$$

The notation $P(X \leq x, Y \leq y)$ means that in a trial the variable X assumes variables not exceeding x, **and** Y not exceeding y. This is indicated by the green region in Fig 7.9 which is the set $(-\infty, x) \times (-\infty, y)$.

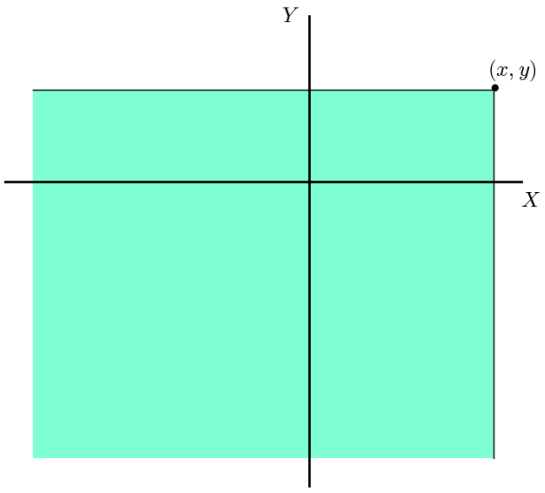


Figure 7.15: The domain of a 2-D CDF.

Recall that in the one-variable case we have $P(a < X \leq b) = F(b) - F(a)$. In the case of two variables it easy to see that for the rectangle $(a_1, a_2] \times (b_1, b_2]$

$$P(a_1 < X \leq a_2, b_1 < X \leq b_2) = F(a_2, b_2) - F(a_1, b_2) - F(a_2, b_1) + F(a_1, b_1) \tag{7.83}$$

Discrete Two-Dimensional Distributions

The variable (X, Y) is discrete if both X and Y are discrete. So (X, Y) can assume at most countably many values $(x_1, y_1), (x_1, y_2), \ldots, (x_i, y_j), \ldots$, with positive probabilities. The probability of any set that does not contain any of these values is zero.

Definition. A function $f : R_X \times R_Y \to \mathbf{R}$ is called **joint probability density function** for X and Y, if

$$f(x, y) = P(X = x, Y = y) \tag{7.84}$$

So, if $R_X = \{x_1, x_2, \dots\}$, $R_Y = \{y_1, y_2, \dots\}$, and $p_{ij} = P(X = x_i, Y = y_j)$ then

$$f(x, y) = \begin{cases} p_{i,j} & \text{if } x = x_i, y = y_j \\ 0 & \text{otherwise} \end{cases} \tag{7.85}$$

Note that

$$f(x, y) \geq 0 \text{ for all } x, y \tag{7.86}$$

The corresponding cumulative distribution function is obtained by taking sums, just was done in (7.28)

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j) = \sum_{x_j \leq x} \sum_{y_j \leq y} p_{ij} \tag{7.87}$$

and we have the condition

$$\sum_i \sum_j f(x_i, y_j) = \sum_i \sum_j p_{i,j} = 1 \tag{7.88}$$

just as we had in (7.30).

Example 7.9.1 (Two-Dimensional Discrete Random Variable)**.** We roll simultaneously a fair four-sided die with variable X whose range is $R_X = \{1, 2, 3, 4\}$ and a six-sided die with variable Y whose range is $R_Y = \{1, 2, 3, 4, 5, 6\}$.

1. Find the joint probability function $f(x, y)$.
2. Find the value of the joint CDF $F(2, 3)$ and interpret your answer in terms of probability of sum event.

Solution: The range of (X, Y) is

$$\begin{aligned} S_{X\times Y} = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\} \end{aligned}$$

1. Since both dies are fair, each of the outcomes occurs with probability $\frac{1}{24}$. Thus,
$$f(x,y) = \begin{cases} \frac{1}{24} & \text{if } (x,y) \in R_{X\times Y} \\ 0 & \text{otherwise} \end{cases}$$

2. We have
$$\begin{aligned} F(2,3) &= f(1,1) + f(1,2) + f(1,3) + f(2,1) + f(2,2) + f(2,3) \\ &= \frac{1}{24} + \frac{1}{24} + \frac{1}{24} + \frac{1}{24} + \frac{1}{24} + \frac{1}{24} \\ &= \frac{1}{6} \end{aligned}$$
$F(2,3)$ is the probability that the four-sided die shows a number up to 2 and the six-sided up to 3.

□

Definition. For a discrete random variable (X,Y) with joint density $f(x,y)$, we may be interested in studying the values of one variable while the remains fixed. The function
$$f_1(x) = P(X = x,\ Y \text{ any}) = \sum_{y\in R_Y} f(x,y) \tag{7.89}$$
is called the **marginal probability density function** of X. Likewise, we have the marginal probability density function of Y.
$$f_2(y) = P(X \text{ any},\ Y = y) = \sum_{x\in R_X} f(x,y) \tag{7.90}$$

We also have the corresponding **marginal cumulative distribution functions**
$$F_1(x) = P(X \le x,\ Y \text{ any}) = \sum_{s\le x} f_1(s) \tag{7.91}$$
and
$$F_2(y) = P(X \text{ any},\ Y \le y) = \sum_{t\le y} f_2(t) \tag{7.92}$$

Example 7.9.2. Using (X,Y) from Example 7.9.1 find

1. the marginal joint density functions $f_1(x)$ and $f_2(y)$.

2. the marginal joint cumulative functions $F_1(x)$ and $F_2(y)$.

Solution:

1. We have
$$f_1(x) = \sum_{j=1}^{6} f(x,j) = 6 \times \frac{1}{24} = \frac{1}{4}$$
and
$$f_2(y) = \sum_{j=1}^{4} f(j,y) = 4 \times \frac{1}{24} = \frac{1}{6}$$

2. Next, we have
$$F_1(x) = \sum_{s \le x} f_1(s) = \frac{x}{4}$$
and
$$F_2(y) = \sum_{t \le y} f_2(t) = \frac{y}{6}$$

□

Continuous Two-Dimensional Distributions

The random variable (X, Y) is **continuous** if its joint probability density function is a continuous function $f : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ such that

$$f(x,y) \ge 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\, dx\, dy = 1 \tag{7.93}$$

The joint cumulative probability distribution $F(x,y) = P(X \le x, Y \le y)$ satisfies

$$F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(s,t)\, ds\, dt \tag{7.94}$$

The probability that (X, Y) assumes any value inside a rectangle is

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) = \int_{b_1}^{b_2} \int_{a_1}^{a_2} f(x,y)\, dx\, dy \tag{7.95}$$

More generally, the probability that an event A occurs is

$$P(A) = P((X,Y) \in A) = \iint_A f(x,y)\,dx\,dy \tag{7.96}$$

If A is a small neighborhood of a point (x,y), then we may approximate $P(A)$ by

$$P(A) = P((X,Y) \in A) \approx f(x,y) \times \text{ area of } A$$

Example 7.9.3 (Two-Dimensional Uniform Distribution)**.** Let R be the rectangle $a_1 < X \le a_2, b_1 < Y \le b_2$ and let $r = (a_2 - a_1)(b_2 - b_1)$ be the area of R. The density

$$f(x,y) = \begin{cases} \frac{1}{r} & \text{if } (x,y) \in R \\ 0 & \text{otherwise} \end{cases} \tag{7.97}$$

defines the **uniform distribution** *in the rectangle* R.

For a continuous pair (X,Y), the **marginal joint densities** f_1 of X and f_2 of Y are defined by

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \tag{7.98}$$

and

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \tag{7.99}$$

Their corresponding **marginal cumulative joint densities** are

$$F_1(x) = P(X \le x,\ -\infty < Y < \infty) = \int_{-\infty}^{x} f_1(s)\,ds \tag{7.100}$$

and

$$F_2(y) = P(-\infty < X < \infty,\ Y \le y,) = \int_{-\infty}^{y} f_2(t)\,dt \tag{7.101}$$

Example 7.9.4. Find the (a) marginal density and (b) the marginal cumulative function of Y, for $0 < y < \infty$, if (X,Y) has joint density function

$$f(x,y) = \begin{cases} 2e^{-x-y} & \text{if } 0 < y \le x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Solution: We have

$$\begin{aligned} f_2(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx \\ &= \int_{y}^{\infty} 2e^{-x-y}\,dx \\ &= 2e^{-y}\int_{y}^{\infty} e^{-x}\,dx \\ &= 2e^{-y}\left[-e^{-x}\right]_{y}^{\infty} \\ &= 2e^{-y}e^{-y} \\ &= 2e^{-2y} \qquad (0 < y < \infty) \end{aligned}$$

and

$$\begin{aligned} F_2(y) &= \int_{-\infty}^{y} f_2(t)\,dt \\ &= \int_{0}^{y} 2e^{-2t}\,dt \\ &= \left[-e^{-2t}\right]_{0}^{y} \\ &= 1 - e^{-2y} \qquad (0 < y < \infty) \end{aligned}$$

□

Independence of Random Variables

The random variables X and Y in the pair (X, Y) are called **independent** if

$$F(x,y) = F_1(x)F_2(y) \quad \text{for all } (x,y) \tag{7.102}$$

(X, Y) can be discrete or continuous.

If (7.102) does not hold, then X and Y are called **dependent**.

An equivalent condition of independence is

$$f(x,y) = f_1(x)f_2(y) \quad \text{for all } (x,y) \tag{7.103}$$

Example 7.9.5. Check the discrete variables X and Y of Example 7.9.1 for independence.

Solution: The marginal densities were found to be the constants

$$f_1(x) = \frac{1}{4}, \quad f_2(y) = \frac{1}{6}$$

while the nonzero values of the joint density were $f(x, y) = \frac{1}{24}$ for all $x \in R_X$ and $y \in R_Y$. Hence,

$$f(x, y) = \frac{1}{24} = \frac{1}{4} \times \frac{1}{6} = f_1(x) f_2(y)$$

So, X and Y are independent.

□

Example 7.9.6. Let X and Y be discrete random variables with $R_X = \{1, 2, 3\}$, $R_Y = \{A, B\}$ and joint probability density values given by the table.

$Y \setminus X$	1	2	3
A	0.40	0	0.20
B	0	0.20	0.20

First (a) find $f_1(x)$ and $f_2(y)$, then (b) determine the dependence or independence of X and Y.

Solution: We have

$$f_1(1) = \sum_{y \in \{A,B\}} f(1, y) = 0.40 + 0 = 0.40$$
$$f_1(2) = \sum_{y \in \{A,B\}} f(2, y) = 0 + 0.20 = 0.20$$
$$f_1(3) = \sum_{y \in \{A,B\}} f(3, y) = 0.20 + 0.20 = 0.40$$

$$f_2(A) = \sum_{x \in \{1,2,3\}} f(x, A) = 0.40 + 0 + 0.20 = 0.60$$
$$f_2(B) = \sum_{x \in \{1,2,3\}} f(x, B) = 0 + 0.20 + 0.20 = 0.40$$

Since $f(1,B) = 0 \neq 0.16 = f_1(1)f_2(B)$, X and Y are *dependent*.

□

Example 7.9.7. Are X and Y independent, if their joint density $f(x,y)$?

$$f(x,y) = \begin{cases} e^{-x-y} & \text{if } 0 < x, y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Solution: We compute the marginal densities.

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^{\infty} e^{-x-y}\,dy = e^{-x}$$

Likewise, we find that $f_2(y) = e^{-y}$. Hence,

$$f(x,y) = e^{-x-y} = e^{-x}e^{-y} = f_1(x)f_2(y)$$

So, X and Y are independent.

□

7.10 Several Random Variables

The theory of pairs (X,Y) of random variables generalizes to any finite number of variables $X_1, X_2, \ldots, X_n$ out of which we form the random variable $\mathbf{X}$ as the n-tuple

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) \tag{7.104}$$

The joint cumulative distribution now is

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n,) \tag{7.105}$$

The variables $X_1, X_2, \ldots, X_n$ are called **independent**, if

$$F(x_1, \ldots, x_n) = F_1(x_1)F_2(x_2)\ldots F_n(x_n) \text{ for all } (x_1, \ldots, x_n) \tag{7.106}$$

where the $F_j(x_j)$ is the marginal joint cumulative distribution of X_j. So

$$F_j(x_j) = P(X_j \leq x_j, \text{ all other } X_k \text{ any}) \tag{7.107}$$

If the $X_1, X_2, \ldots, X_n$ are not independent, they are called **dependent**.

We briefly examine functions of several variables with the most important cases being the **sum** and the **product** of the variables.

For the case $n = 2$, let X and Y be random variables. For a nonconstant function $g(x, y)$, we obtain a random variable $Z = g(X, Y)$. For example $Z = X + Y$ or $Z = XY$.

If (X, Y) is **discrete** then the probability density $f(z)$ of $Z = g(X, Y)$ is

$$f(z) = P(Z = z) = \sum \sum_{g(x,y)=z} f(x, y) \tag{7.108}$$

Likewise, the cumulative distribution is

$$F(z) = P(Z \leq z) = \sum \sum_{g(x,y)\leq z} f(x, y) \tag{7.109}$$

For **continuous** (X, Y) we have

$$F(z) = P(Z \leq z) = \iint\limits_{g(x,y)\leq z} f(x, y)\, dx\, dy \tag{7.110}$$

The number

$$E(g(X, Y)) = \begin{cases} \sum_x \sum_y g(x, y) f(x, y) & (X, Y) \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy & (X, Y) \text{ continuous} \end{cases} \tag{7.111}$$

is called the **expectation** of $g(X, Y)$. Here we assume that the double series converges absolutely and that the double integral is finite.

The function $E(g(X, Y))$ is **linear**.

$$E(c_1 g_1(X, Y) + c_2 g_2(X, Y)) = c_1 E(g_1(X, Y)) + c_2 E(g_2(X, Y)) \tag{7.112}$$

The important special case is the sum.

$$E(X + Y) = E(X) + E(Y)$$

Theorem 7.10.1 (Addition of Expectations)**.** *The expectation (mean) for a sum of variables $X_1, \ldots, X_n$ satisfies*

$$E(X_1 + X_2, \cdots + X_n) = E(X_1) + E(X_2), \cdots + E(X_n) \tag{7.113}$$

Theorem 7.10.2 (Product of Independent Expectations)**.** *The expectation (mean) for a product of **independent** variables $X_1, \ldots, X_n$ satisfies*

$$E(X_1 X_2 \cdots X_n) = E(X_1) E(X_2) \cdots E(X_n) \tag{7.114}$$

Note that the **variance** σ^2 of the sum $Z = X + Y$ satisfies

$$\begin{aligned}\sigma^2 &= E([Z-\mu]^2) = E(Z^2) - [E(Z)]^2 \\ &= \sigma_1^2 + \sigma_2^2 + 2\sigma_{XY}\end{aligned}$$

where σ_1^2 is the variance of X and σ_2^2 is the variance of Y and σ_{XY} is the **covariance** of X and Y which is defined by

$$\sigma_{XY} = E(XY) - E(X)E(Y) \tag{7.115}$$

If X and Y are independent then

$$E(XY) = E(X)E(Y)$$

in which case the covariance is zero and we get the addition property of independent variables.

Theorem 7.10.3 (Addition of Variances of Independent Variables)**.** *The variance of the sum of* ***independent*** *variables $X_1, \ldots, X_n$ equals the sum of the variances.*

One the most important theorems in Probability is the Central Limit Theorem, stated next. The theorem says that the finite partial sums of infinitely many variables with the same distribution approach the standard normal distribution.

Theorem 7.10.4 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_n, \ldots$ be* ***independent*** *random variables with the same distribution, hence they have the same mean μ and variance σ^2. Let $Y_n = X_1 + \cdots + X_n$. Then the random variable*

$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}} \tag{7.116}$$

is asymptotically standard normal ($\mu = 0$, $\sigma^2 = 1$). The cumulative distribution $F_n(x)$ of Z_n satisfies the equation

$$\lim_{n\to\infty} F_n(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \tag{7.117}$$

So, we may use the tables of the normal distribution to study nonnormal distributions, provided we **large enough samples**.

Theorem 7.10.5 (Sum and Mean of Normal Random Variables)**.** *Let $X_1, X_2, \ldots, X_n$ be **independent** normal random variables all of the form $N(\mu, \sigma^2)$, i.e., with the same mean μ and variance σ^2. Then*

1. *The sum $X_1 + \cdots + X_n$ is normal with mean $n\mu$ and variance $n\sigma^2$.*

$$X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2)$$

2. *The random variable $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ is normal with mean μ and variance $\frac{\sigma^2}{n}$.*

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

3. *The random variable $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal.*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

4. *The random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, where*

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

   is a Student's t-distribution with $n-1$ degrees of freedom.

5. *With S^2 as in Part 4, the random variable*

$$Y = (n-1)\frac{S^2}{\sigma^2}$$

   has a chi-squared distribution of $n-1$ degrees of freedom.

Chapter 8

Mathematical Statistics

8.1 Random Sampling

Mathematical statistics studies mathematical ways of evaluating random experiments stemming from practical problems. **Random variables** play a vital role in this context. Examples include, the number of defective objects produced by a machine, the tolerance limits of the size of a manufactured object, the temperature fluctuations at a certain place during the day, etc.

In statistics random variables are studied from **samples** selected. For example, we select 30 light bulbs to estimate the average the number of defective ones in an entire batch. Sampling is necessary because, it is impossible to check all light bulbs. The same is true when we study animal populations, performance of cars, etc.

To obtain reliable results from a sample, we must make sure that the sample comes from a **random selection** process. Random selections are often tricky to achieve. There may be factors that bias the sample. For example, a slightly defective (unfair) coin, or a defective machine producing objects that we sample, etc.

A useful tool in helping obtaining random samples is the use of **random numbers**. Producing random numbers by a computer is done by programs that are called **random number generators**. Writing such programs is challenging. All known methods yield **pseudorandom numbers** that eventually cycle through and repeat. The goal is to make the cycles very large so that the numbers *appear to be random*.

Example 8.1.1. We produced 10 pseudorandom real numbers in $[0, 1]$.

1. From Wolfram Alpha (`https://www.wolframalpha.com/`)

```
RandomReal[{0,1},10]
```

```
{0.729468, 0.320247, 0.850759, 0.13616, 0.816587,
0.821304, 0.429888, 0.0120385, 0.345218, 0.526726}
```

2. From sagemath (`http://www.sagemath.org/`)

```
sage: [RR.random_element(0,1) for i in range(10)]
```

```
[0.88, 0.62, 0.12, 0.00, 0.44, 0.44, 0.25, 0.56, 0.94, 0.62]
```

In statistics we study samples and try to infer knowledge about an entire set of objects. Our random variables are studied from samples.

Let $\{x_1, x_2, \ldots, x_n\}$ be a sample **of size** n.

The **sample mean** is defined by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{8.1}$$

The **sample variance** is defined by

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + \cdots (x_n - \bar{x})^2\right] \tag{8.2}$$

The sample mean and variance $\bar{x}$ and s^2 are called **parameters** of the sample.

8.2 Point Estimation

A main task in statistics is the estimation of various parameters of a random variable, such as the mean and the variance.

A **point estimate** of a parameter is a number used as estimate. An **interval estimate** of a parameter is an interval of numbers. In this section we study point estimates.

An estimate $\hat{\mu}$ of the mean μ of a population is the mean $\bar{x}$ of a sample we obtained. So,

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{8.3}$$

Likewise, an estimate $\hat{\sigma}^2$ of the variance σ^2 of a population is the variance s^2 of a sample. So,

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})^2 \tag{8.4}$$

Example 8.2.1. For the binomial distribution, a point estimate $\hat{\mu}$ for the mean $\mu = np$ leads to an estimate of the parameter p which is the probability of success of one trial.

$$\hat{p} = \frac{\bar{x}}{n} \tag{8.5}$$

We are also interested in estimating the **moments** of a distribution (see Section 7.5 and equation (7.54)). This is done by defining the ***k*th moment of a sample** $\{x_1, \ldots, x_n\}$ by

$$m_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k \tag{8.6}$$

The Maximum Likelihood Method

The **maximum likelihood method** was introduced by R. A. Fisher in 1912 to estimate an unknown parameter. The method is based on a simple idea that is also found in the works of Gauss and Bernoulli.

Let X be a random variable with probability density $f(x;\theta)$ which depends on x, as usual, but also depends on an unknown parameter θ. Suppose that we have sample of n **independent** values $x_1, \ldots, x_n$ of X. If X is discrete, then the probability that a sample of size n consists of these values is the joint density

$$l(\mathbf{x};\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) \tag{8.7}$$

where $\mathbf{x}$ is notation for the sample values $x_1, \ldots, x_n$. If X is continuous, then the probability that a sample of size n has values in the infinitesimal intervals $x_j \le x \le x_j + \Delta x$ $(j = 1, 2, \ldots, n)$ is the function

$$f(x_1;\theta)\,\Delta x f(x_2;\theta)\,\Delta x \cdots f(x_n;\theta)\,\Delta x = l(\mathbf{x};\theta)\ (\Delta x)^n \tag{8.8}$$

Now for a fixed sample $x_1, \ldots, x_n$ the function l only depends on θ, so $l = l(\theta)$ and it is called the **likelihood function**. We want to choose such an estimation for θ that **maximizes** the likelihood function $l(\theta)$. Intuitively, this method selects the parameter values that make the data "most probable".

If $l(\theta)$ is differentiable in θ, then a necessary condition that a maximum occurs in an open interval is

$$\frac{\partial l}{\partial \theta} = 0 \tag{8.9}$$

We used partial derivative because the function l also depends on the sample. A solution of equation (8.9) for θ is called a **maximum likelihood estimate** for the parameter θ.

We may solve for θ the equation

$$\frac{\partial \ln l}{\partial \theta} = 0 \tag{8.10}$$

which is often easier to solve because l is a product. We are allowed to replace (8.9) with (8.10), because $f(x_i) > 0$, a maximum of f is nonnegative and $\ln(y)$ is a monotonically increasing function of y. Therefore, the maximum values of l and $\ln l$ occur at the same points.

Example 8.2.2 (Normal Distribution. Estimation of μ and σ)**.** Find the maximum likelihood estimates for the mean μ and standard deviation σ of the normal distribution for any sample of n independent values $x_1, \ldots, x_n$.

Solution: The likelihood function l for a sample with points $x_1, \ldots, x_n$ is

$$l = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-g} \quad \text{where} \quad g = \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2$$

Taking logarithms, we get

$$\ln l = -n \ln \sigma - n \ln \sqrt{2\pi} - g \tag{8.11}$$

To estimate μ we differentiate (8.11) with respect to μ and set to zero

$$\frac{\partial \ln l}{\partial \mu} = -\frac{\partial g}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu) = 0$$

Hence,

$$\sum_{j=1}^{n}(x_j - \mu) = \sum_{j=1}^{n} x_j - n\mu = 0$$

Therefore, the estimate $\hat{\mu}$ for μ is

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j = \bar{x}$$

To estimate σ we differentiate (8.11) with respect to σ and set to zero

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{\partial g}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{n}(x_j - \mu)^2 = 0$$

We solve for σ^2 and replace μ with its estimate $\hat{\mu} = \bar{x}$ to get the estimate $\hat{\sigma}^2$ for σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

□

8.3 Confidence Intervals

A **confidence interval** $[\theta_1, \theta_2]_{\theta,\gamma}$ for estimating a parameter θ is an interval $\theta_1 \leq \theta \leq \theta_2$ that contains θ with some large probability γ. [1]

$$P(\theta_1 \leq \theta \leq \theta_2) = \gamma \tag{8.12}$$

The probability γ is the **confidence level** and θ_1, θ_2 are the **lower** and **upper confidence limits**.

The values $\gamma = 95\%$ and $\gamma = 99\%$ are often used in practice. The value $\gamma = 95\%$ implies that $1 - \gamma = 5\%$ is the probability that θ is not in such interval, i.e., about 1 in 20 such intervals will not contain θ.

The confidence interval depends on any chosen value of γ. Typically, if the value of γ is high, then the length of the confidence interval is small. For a given γ the limits θ_1 and θ_2 are computed from a given sample $x_1, \ldots, x_n$ obtained from n independent observations of a random variable X.

Confidence intervals are more useful than point estimates. Given such interval we may use as a point estimate of a random variable the average of the ends.

[1] Confidence intervals were introduced by J. Neyman in 1935.

The values θ_1 and θ_2 are computed from a sample $x_1, \ldots, x_n$ which comes from n observations of a random variable X. To estimate θ_1 and θ_2 we use the so-called **standard trick**. We think of $x_1, \ldots, x_n$ as **single** observations of the random variables $X_1, \ldots, X_n$ with distributions equal of the distribution of X. So, each one is essentially X. In this case $\theta_1 = \theta_1(x_1, \ldots, x_n)$ and $\theta_2 = \theta_2(x_1, \ldots, x_n)$ are observed values of the two random variables $\Theta_1 = \Theta_1(X_1, \ldots, X_n)$ and $\Theta_2 = \Theta_2(X_1, \ldots, X_n)$. Therefore, (8.13) takes the form

$$P(\Theta_1 \leq \theta \leq \Theta_2) = \gamma \tag{8.13}$$

Normal Distribution: Confidence Interval for μ Given σ^2

We would like to find confidence intervals for the for the mean μ of a normal distribution, given the variance σ^2 for some chosen confidence level γ (e.g., $\gamma = 0.95, 0.99, \ldots$).

We use the standard trick and assume that the mean μ is estimated from n observations of identical but independent normal variables $X_1, \ldots, X_n$ by using the mean variable $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. $\bar{X}$ is also normal by Theorem 7.10.5 and the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{8.14}$$

is standard normal. So, $Z \sim N(0, 1)$.

Now if we find c such that

$$P(-c \leq Z \leq c) = \gamma, \tag{8.15}$$

then since (8.14) implies

$$\bar{X} - k \leq \mu \leq \bar{X} + k, \quad \text{where } k = \frac{c\sigma}{\sqrt{n}} \tag{8.16}$$

we get a confindence interval for μ, if σ is known.

To find such c for a given γ we use

$$P(-c \leq Z \leq c) = \Phi(c) - \Phi(-c) = \gamma \tag{8.17}$$

and also CAS or table values for the cumulative standard normal distribution $\Phi(z)$.

γ	0.90	0.95	0.99
c	1.645	1.960	2.576

Table 8.1: Some values of γ and the corresponding c.

Using a table of values for Φ we get the following c for some common γ.

For example, to see how $c = 1.645$ was obtained, we have $\Phi(c) - \Phi(-c) = 0.90$. Hence, $\Phi(c) - (1 - \Phi(c)) = 0.90$. Therefore, $\Phi(c) = 0.95$. Then we use tables or software to get $c = 1.645$.

Example 8.3.1 (Confidence Interval for μ Given σ^2)**.** Determine the confidence intervals $[\theta_1, \theta_2]_{\mu,0.90}$ with confidence level $\gamma = 0.90$ for the mean μ of a normal distribution with variance $\sigma^2 = 4$, using samples of (a) $n = 50$, (b) $n = 100$, (c) $n = 200$ values, each with sample mean $\bar{x} = 6$.

Solution: By Table 8.1, the value $\gamma = 0.90$ corresponds to $c = 1.645$. Hence,

$$k_1 = c\sigma/\sqrt{n} = 1.645 \cdot 2/\sqrt{50} = 0.4653$$
$$k_2 = c\sigma/\sqrt{n} = 1.645 \cdot 2/\sqrt{100} = 0.3290$$
$$k_3 = c\sigma/\sqrt{n} = 1.645 \cdot 2/\sqrt{200} = 0.2326$$

So the confidence intervals are

$$\bar{x} \pm k_1 = 6 \pm 0.4653 \quad \text{or} \quad [5.5347, 6.4653]_{\mu,0.95}$$
$$\bar{x} \pm k_2 = 6 \pm 0.3290 \quad \text{or} \quad [5.6710, 6.3290]_{\mu,0.95}$$
$$\bar{x} \pm k_2 = 6 \pm 0.2326 \quad \text{or} \quad [5.7674, 6.2326]_{\mu,0.95}$$

□

Example 8.3.2. In Example 8.3.1 how big a sample do we need so that the confidence interval has length 0.2?

Solution: The length of the interval $\bar{x} \pm k$ is $2k$. So $2k = 0.2$. Hence, $k = 0.1 = c\sigma/\sqrt{n} = 1.645 \cdot 2/\sqrt{n}$. So $n = (1.645 \cdot 2/0.1)^2 \approx 1083$.

□

Example 8.3.3. Which of the following is true, if we have a 99% confidence interval for the mean of some population?

1. The interval includes about 99% of the population.

2. The interval has 99% chance of including the sample mean.

Solution: None is correct. The 99% confidence interval for the μ means that the interval has a 99% chance of including the population mean μ.

□

Normal Distribution: Confidence Interval for μ with Unknown σ^2

In the last subsection we assumed that the variance is known and we estimated the mean. However, quite often the variance is not known.

If we want a confidence interval for μ and cannot use σ^2, the next best thing to use the variance, s^2, of the **sample**.

Recall from Theorem 7.10.5 that the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{8.18}$$

where

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$$

is a Student's t-distribution with $n-1$ degrees of freedom.

So the cumulative probability $F(z)$ of the Student t-distribution (7.78) with $\nu = n-1$, now takes the form

$$F(z) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)} \int_{-\infty}^{z} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} dt \tag{8.19}$$

where n is the sample size.

Now if we find c such that

$$P(-c \leq T \leq c) = \gamma, \tag{8.20}$$

then since (8.18) implies

$$\bar{X} - k \leq \mu \leq \bar{X} + k, \quad \text{where } k = \frac{cS}{\sqrt{n}} \tag{8.21}$$

we get a confidence interval for μ, where S was used for the unknown σ.

To find such c for a given γ we set

$$P(-c \leq T \leq c) = F(c) - F(-c) = 2F(c) - 1 = \gamma \tag{8.22}$$

where we substituted $F(-c) = 1 - F(c)$, by the symmetry of the graph of the t-distribution about the y-axis.

Our strategy is to use the tabulated or CAS computed values of $F(z)$ to find c such that

$$F(c) = \frac{\gamma + 1}{2} \tag{8.23}$$

and then find the confidence interval by (8.21).

Example 8.3.4 (Confidence Interval for μ with Unknown σ^2)**.** Six independent measurements are taken from a normal population. The sample has values $x_1 = 85$, $x_2 = 83$, $x_3 = 80$, $x_4 = 87$, $x_5 = 79$, and $x_6 = 81$. Find a 99% confidence interval for the mean.

Solution: The confidence level is $\gamma = 0.99$. Therefore, $F(c) = \frac{1}{2}(1+\gamma) = 0.995$. By tables we get for $n - 1 = 5$ degrees of freedom that $c = 4.03$.

Next, we compute the sample mean $\bar{x} = \frac{1}{6}\sum_{j=1}^{6} x_j = 82.5$ and the sample variance $s^2 = \frac{1}{5}\sum_{j=1}^{6}(x_j - \bar{x})^2 = 9.5$.

Now

$$k = cs/\sqrt{n} = 4.03 \cdot \sqrt{9.5}/\sqrt{6} = 5.0709$$

Hence, the confidence interval is

$$\bar{x} \pm k = 82.5 \pm 5.0709 \quad \text{or} \quad [77.4291, 87.5709]_{\mu,0.999}$$

which means that this interval has a 99% chance of containing the actual mean of the population.

□

8.4 Testing Hypotheses

In statistics we often perform a **testing of a hypothesis** to draw conclusions about a random variable from only a **sample**. The random variable may be associated to a population or the size of a manufactured object, or to the dosage of some medication, etc.

As an example, we may be testing whether a manufacturer's claim that at most 3% of screws sold are defective is true or false.

In this section we develop methods to test a hypothesis by using a sample of a population. The confirmation or rejection of the hypothesis may help us make a **decision**. For example, we may decide to not buy a car if the claim that the car is fuel efficient is found to be incorrect with some probability.

A **(statistical) hypothesis** H is an assumption about the distribution $f(x;\theta)$ of a population X. This assumption is usually either about the parameter θ or about the form of the distribution of X. A hypothesis H is called a **simple hypothesis**, if it completely specifies the density $f(x;\theta)$; otherwise, H is called a **composite hypothesis**.

The hypothesis H_o that is tested is called the **null hypothesis**. The negation H_a of the null hypothesis is called the **alternative hypothesis**.

If θ is a population parameter, then the typical formulation of the null hypothesis and the alternative hypothesis is

$$H_o : \theta \in \Omega_o \quad \text{and} \quad H_a : \theta \in \Omega_a$$

where Ω_o and Ω_a are disjoint subsets of the parameter space Ω of θ.

If Ω_o consists of only one element, then H_o is a simple hypothesis.

A **hypothesis test** is a sequence

$$X_1, x_2, \ldots, X_n \,; H_o, H_a; \, C$$

where $X_1, X_2, \ldots, X_n$ is a random sample from a population X with the probability density function $f(x;\theta)$, H_o and H_a are hypotheses about θ, and C is a set that is called the **critical region** or the **rejection region** in the hypothesis test. The rejection region is the region such that we reject the null hypothesis H_o, if an observed (from the sample) value of the parameter θ is in C.

We always accept or reject a hypothesis with an error probability. The probability threshold α below which we reject the null hypothesis although it may be true is called **significance level** of the test. Typical values for α are 1% and 5%. The basic steps in a hypothesis test for a parameter θ are:

1. Form hypotheses H_o and H_a.

2. Choose a significance level α.

3. Use a random variable $\Theta = h(X_1, \ldots, X_n)$ whose distribution is known and depends on H_o and H_a.

4. Determine the rejection region C by using the assumption of the distribution of Θ and α.

5. Use a sample $x_1, \ldots, x_n$ to find an observed value $\theta = h(x_1, \ldots, x_n)$ of Θ.

6. Accept or reject the null hypothesis H_o, depending on whether θ is outside or inside C.

Example 8.4.1. The null hypothesis H_o is a manufacturer's claim that the breaking load of a 28 mm 6-strand twisted hemp rope is $\mu_0 = 5$ Kg. The alternative hypothesis H_a that the breaking load is $\mu_1 < \mu_0$. After testing a sample of 20 ropes it is found that the sample average breaking load is $\bar{x} = 4.9$ Kg and that the sample standard deviation is $s = 0.3$ Kg. For significance level $\alpha = 5\%$, determine whether or not the manufacturer's claim should be accepted. Assume that the average breaking load μ is uniformly distributed.

Solution: By Theorem 7.10.5, the random variable

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 5}{0.3/\sqrt{20}}$$

with $\mu = \mu_0 = 5$ has a t-distribution with $n - 1 = 19$ degrees of freedom, where $n = 20$ is the sample size.

Next, we determine a critical region for T by setting

$$P(T \leq c) = \alpha = 0.05$$

Hence, $P(T \leq \tilde{c}) = 1 - \alpha = 0.95$. By using Table A9, Appendix 5 of the text, we get $\tilde{c} = 1.73$. Therefore, $c = -\tilde{c} = -1.73$. So, the critical region for T is

$$C = (-\infty, -1.73)$$

Now, if the hypothesis H_o is true, we have a chance of only 5% that we observe a value t of T (computed from the sample) that will fall in $(-\infty, -1.73)$. If we observe such t we reject the hypothesis, otherwise we accept it.

We compute $t = \frac{4.9-5}{0.3/\sqrt{20}} = -1.49$. Since $-1.49 \notin C$, we *accept* the hypothesis that the average breaking load of the rope is 5 Kg. □

One-Sided and Two-Sided Alternatives H_a

Suppose that we test the null simple hypothesis $H_o : \theta = \theta_0$ for a parameter $\theta \in \mathbf{R}$ associated with some random variable. For the alternative hypothesis H_a we have the following options:

(1) $\theta > \theta_0$

(2) $\theta < \theta_0$

(3) $\theta \neq \theta_0$

(1) and (2) are the **one-sided alternatives**. (1) is a **right-sided alternative**, (2) is a **left-sided alternative**. (3) is the **two-sided alternative**.

We observe now that (1) has rejection region of the form

$$C = (c, \infty), \text{ where } c > \theta_0$$

(2) has rejection region of the form

$$C = (-\infty, c), \text{ where } c < \theta_0$$

(3) has rejection region of the form

$$C = (-\infty, c_1) \cup (c_2, \infty), \text{ where } c_1 < \theta_0 < c_2$$

Types of Errors in Tests; The Power Function

When we test a hypothesis we always run the risk of an error, thus, making a wrong decision. There are two common types of errors.

1. **Type I error** where we reject a true hypothesis.

   α is the probability of making a Type I error.

2. **Type II error** where we accept a false hypothesis.

   β is the probability of making a Type II error.

Let us discuss in detail a test of a hypothesis $\theta = \theta_0$ against an alternative that is a single number, say θ_1 ($\theta_1 > \theta_0$). Let us assume that we have a **right-sided test**. For a left-sided or a two-sided test the analysis is similar.

We find a critical number $c > \theta_0$ such that (c, ∞) is the critical region. From a sample $x_1, \ldots, x_n$, we compute a value $\hat{\theta} = h(x_1, \ldots, x_n)$ for the parameter θ. If $\hat{\theta} > c$, we reject the hypothesis and if $\hat{\theta} \leq c$, we accept it.

Type I error The null hypothesis is true but we reject it, because $\hat{\theta} > c$. The probability of making such type of error is

$$P(\hat{\theta} > c)_{\theta=\theta_0} = \alpha \tag{8.24}$$

where α is the significance level.

Type II error The null hypothesis is false but we accept it, because $\hat{\theta} \leq c$. The probability of making such type of error is

$$P(\hat{\theta} \leq c)_{\theta=\theta_1} = \beta \tag{8.25}$$

The quantity

$$\eta = 1 - \beta \tag{8.26}$$

is called the **power** of the test.

	Truth $\theta = \theta_0$	Truth $\theta = \theta_1$
Accepted $\theta = \theta_0$	True decision $P = 1 - \alpha$	Type II error $P = \beta$
Accepted $\theta = \theta_1$	Type I error $P = \alpha$	True decision $P = 1 - \beta$

Table 8.2: Types I and II errors in testing $\theta = \theta_0$ versus $\theta = \theta_1$ ($\theta_1 > \theta_0$).

Equations (8.24) and (8.25) show that both errors α and β depend on c. So we try to choose c such that both errors become as small as possible. In practice we adjust the sample size to make the errors small. Note that when β is small, then the power is large.

If the alternative θ_1 is not a single value but varies, then the power is a function of θ.

$$\eta(\theta) = 1 - \beta(\theta) \tag{8.27}$$

The function $\beta(\theta)$ is called **the operating characteristic** of the test.

Test for the Mean μ of the Normal Distribution with Known Variance σ^2

Example 8.4.2. Let X be a normal random variable with variance $\sigma^2 = 4$. Using a sample of size $n = 20$ with mean $\bar{x}$ test the hypothesis $\mu = \mu_0 = 30$ against the following three kinds of alternatives.

$$(a)\ \mu > \mu_0\,, \qquad (b)\ \mu < \mu_0\,, \qquad (c)\ \mu \neq \mu_0$$

Use significance level $\alpha = 0.05$.

Solution: We estimate the mean μ from

$$\bar{X} = \frac{1}{n}\,(X_1 + \cdots + X_n)$$

If the hypothesis is correct, then $\bar{X}$ is normal with mean $\mu = 30$ and variance $\sigma^2/n = 4/20 = 0.2$, by Theorem 7.10.5.

(a) **Right-Sided Test** By (8.24) we need to find c such that

$$P(\bar{X} > c)_{\mu=30} = 0.05$$

Hence,

$$P(\bar{X} \leq c)_{\mu=30} = \Phi\left(\frac{c-30}{\sqrt{0.2}}\right) = 1 - \alpha = 0.95$$

By Table A8 in Appendix 5 we get $(c-30)/\sqrt{0.2} = 1.645$. Hence, $c = 30.74$ which is greater than $\mu_0 = 30$. So, if $\bar{x} \leq 30.74$, the hypothesis is accepted. If $\bar{x} > 30.74$, the hypothesis is rejected.

The power function is

$$\begin{aligned}\eta(\mu) = P(\bar{X} > 30.74)_\mu &= 1 - P(\bar{X} \leq 30.74)_\mu \\ &= 1 - \Phi\left(\frac{30.74-\mu}{\sqrt{0.2}}\right) \\ &= 1 - \Phi\,(68.74 - 2.24\mu)\end{aligned}$$

(b) **Left-Sided Test** We need to find c such that

$$P(\bar{X} \leq c)_{\mu=30} = \Phi\left(\frac{c-30}{\sqrt{0.2}}\right) = 0.05$$

Hence,

$$\Phi\left(\frac{30-c}{\sqrt{0.2}}\right) = 0.95$$

By Table A8 in Appendix 5 we get $(30-c)/\sqrt{0.2} = 1.645$. Hence, $c = 29.26$ which is less than $\mu_0 = 30$. So, if $\bar{x} \geq 29.26$, the hypothesis is accepted. If $\bar{x} < 29.26$, the hypothesis is rejected.

The power function is

$$\eta(\mu) = P(\bar{X} \leq 29.26)_\mu = \Phi\left(\frac{29.26-\mu}{\sqrt{0.2}}\right) = \Phi\left(65.43 - 2.24\mu\right)$$

(c) **Two-Sided Test** The symmetry of the normal distribution about the line $x = 30$ allows us to use c_1 and c_2 of the form $c_1 = 30-k$ and $c_2 = 30+k$. We need to find k such that

$$P(30-k \leq \bar{X} \leq 30+k)_{\mu=30} = \Phi\left(\frac{k}{\sqrt{0.2}}\right) - \Phi\left(-\frac{k}{\sqrt{0.2}}\right) = 1-\alpha = 0.95$$

Hence,

$$2\Phi\left(\frac{k}{\sqrt{0.2}}\right) - 1 = 0.95$$

Therefore,

$$\Phi\left(\frac{k}{\sqrt{0.2}}\right) = 0.975$$

and by using Table A8, Appendix 5, we get $k/\sqrt{0.2} = 1.960$. Hence, $k = 0.877$. So $c_1 = 30 - 0.877 = 29.123$ and $c_2 = 30 + 0.877 = 30.877$. We conclude that if $29.123 \leq \bar{x} \leq 30.877$ we accept the hypothesis, otherwise we reject it.

The power function is

$$\begin{aligned}
\eta(\mu) &= P(\bar{X} < 29.123) + P(\bar{X} > 30.877) \\
&= P(\bar{X} < 29.123) + 1 - P(\bar{X} \leq 30.877) \\
&= 1 + \Phi\left(\frac{29.123-\mu}{\sqrt{0.2}}\right) - \Phi\left(\frac{30.877-\mu}{\sqrt{0.2}}\right) \\
&= 1 + \Phi\left(65.121 - 2.24\mu\right) - \Phi\left(69.04 - 2.24\mu\right)
\end{aligned}$$

□

Test for the Mean μ of the Normal Distribution with Unknown Variance σ^2 and Test for the Variance σ^2

In Example 8.4.1 we discussed a left-sided test for the mean μ of a normal distribution with unknown variance σ^2.

Next, we test for the variance of a normal distribution of unknown mean.

Example 8.4.3. Test the hypothesis $\sigma^2 = \sigma_0^2 = 5$ against the alternative $\sigma^2 = \sigma_1^2 > \sigma_0^2$ for a normal population. The sample size is $n = 21$ and the sample variance is $s^2 = 7$. Use significance level $\alpha = 5\%$.

Solution: Since the population is normal, the variable Y

$$Y = (n-1)\frac{S^2}{\sigma_0^2} = 20\frac{S^2}{5} = 4S^2$$

is a chi-squared distribution with $n - 1 = 20$ degrees of freedom, by Theorem 7.10.5. Now

$$P(Y > c) = \alpha = 0.05 \quad \text{or} \quad P(Y \leq c) = 0.95$$

Table A10 in Appendix 5 with 20 degrees of freedom gives us $c = 31.41$. Since $Y = 4S^2 = 31.41$, the critical value for S^2 is $c^* = 7.85$. Since $s^2 = 7 < c^* = 7.85$, we accept the null hypothesis.

□

8.5 Quality Control

All production processes have small imperfections and as a result the resulting products are not completely identical. There is always a small variation. In testing the quality of a product we test certain assumptions about it. For example, we may require the property that the mean is $\mu = \mu_0$ and we test this hypothesis against $\mu \neq \mu_0$.

We may use the methods of the last section to accept or reject a hypothesis. However, we have to be cautious about the sampling process. If for example, we test a final product, say the length of $50,000$ screws, it may be too late to do anything to fix the problem, if the sample shows that a hypothesis about the average length should be rejected.

In practice we test samples during the production run. This is done several times, often 5 to 10 times. If after testing a sample we have to reject it, we

stop the production and try to find the cause of the rejection and fix it. The sampling is done in regular time intervals. This process is called **quality control** and it is very important in industry.

If the production process is stopped due to thresholds for randomness although there is no real error we make a Type I error. If the process is not stopped although aomething is wrong we make a Type II error.

Control Chart for the Mean

As an example of quality control process, we test the mean $\mu = 6.08$ cm of the length of certain screws. We use a dozen samples. In each sample we test six screws.

In the following table we display the sample length values, the mean $\bar{x}$, standard deviation s, and range R of values of each sample.

Sample Number	Sample Values						$\bar{x}$	s	R
1	6.14	6.02	6.12	6.01	6.00	6.06	6.058	0.059	0.14
2	6.15	6.07	6.06	6.08	6.11	6.02	6.082	0.044	0.13
3	6.17	6.12	6.09	6.06	6.18	6.05	6.112	0.055	0.13
4	6.09	6.03	6.10	6.15	6.09	6.13	6.098	0.041	0.12
5	6.11	6.05	6.05	6.03	6.07	6.16	6.078	0.048	0.13
6	6.10	6.09	6.00	6.03	6.05	6.10	6.062	0.042	0.10
7	6.01	6.04	6.06	6.14	6.14	6.03	6.070	0.057	0.13
8	6.02	6.10	6.12	6.09	6.11	6.15	6.098	0.044	0.13
9	6.01	6.02	6.01	6.00	6.06	6.06	6.027	0.027	0.06
10	6.09	6.07	6.14	6.17	6.14	6.07	6.113	0.042	0.10
11	6.07	6.18	6.18	6.15	6.15	6.15	6.147	0.040	0.11
12	6.07	6.16	6.18	6.11	6.15	6.11	6.130	0.040	0.11

Table 8.3: Quality Control for 12 Samples.

We choose a **lower control limit** LCL, a **center control line** CL, and an **upper control limit** UCL. We stop the process when during testing a sample the values we check fall outside the **control limits**. When a sample mean falls outside the limits we reject the hypothesis and we assume that the process is "out of control" and we take corrective action.

The control limits should be such that there is a balance. If the limits are too tight we may be stopping the process often without a real problem. If the limits are too loose we may miss some falt in the process.

Let us test the hypothesis of the mean $\mu = \mu_0$ against $\mu \neq \mu_0$ as in Part (c) of Example 8.4.2.

Since $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ is $N(0,1)$ and

$$P(\mu_0 - k \leq \bar{X} \leq \mu_0 + k) = \Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{k}{\sigma/\sqrt{n}}\right) = 1 - \alpha$$

we have

$$2\Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) - 1 = 1 - \alpha$$

or

$$\Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) = \frac{2-\alpha}{2}$$

For $\alpha = 1\%$ we get $\Phi\left(\frac{k}{\sigma/\sqrt{n}}\right) = 0.995$ and by Table A8, Appendix 5, $\frac{k}{\sigma/\sqrt{n}} = 2.58$. We conlude that

$$\text{LCL} = \mu_0 - \frac{2.58\sigma}{\sqrt{n}} \quad \text{and} \quad \text{UCL} = \mu_0 + \frac{2.58\sigma}{\sqrt{n}} \tag{8.28}$$

We assume that the standard deviation σ is known. If it is not known we compute standard deviations from many samples and take their mean as an approximation of σ.

For example, if we use Table 8.3 with $\mu_0 = 6.08$ cm, $\sigma = 0.045$, and $n = 12$, we find that LCL = 6.04648 and ULC = 6.11352.

Given this information we plot a **control chart** (Figure 8.5) which consists of a plot of the sample the means and the lower and upper control limits.

It seems that in Sample 9 we need to stop the process and try to determine whether or not the process has become faulty.

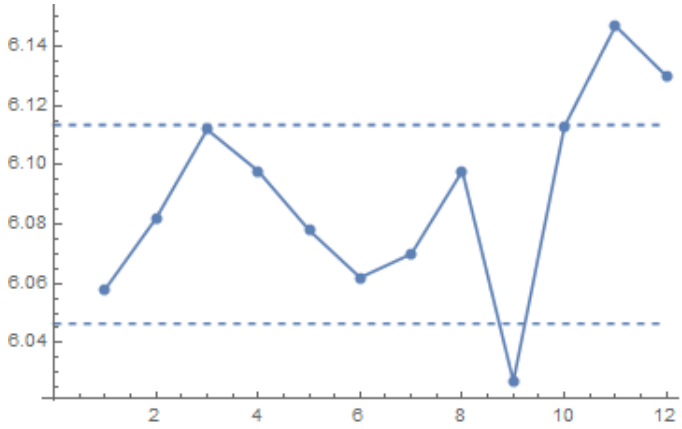


Figure 8.1: Control Chart for the Mean of 12 Samples.

Control Chart for the Variance

We often need to control the variance of a population in a control process. Let us set up a control chart for the variance of a normal distribution by using the method of Example 8.4.3.

For a hypothesis $\sigma^2 = \sigma_0^2$ against the alternative $\sigma^2 > \sigma_0^2$, the variable Y

$$Y = (n-1)\frac{S^2}{\sigma_0^2}$$

is a chi-squared distribution with $n-1$ degrees of freedom, by Theorem 7.10.5. Now we may compute the critical value c from

$$P(Y > c) = \alpha \quad \text{or} \quad P(Y \leq c) = 1 - \alpha$$

Since the alternative is $\sigma^2 > \sigma_0^2$, it makes sense to find an **upper control limit** UCL.

We use Table A10, Appendix 5 with $n-1$ degrees of freedom to find c and we conclude that the critical value for S^2 is

$$\text{UCL} = \frac{\sigma_0^2 c}{n-1} \tag{8.29}$$

So for sample values $s^2 < \frac{\sigma_0^2 c}{n-1}$, we accept the null hypothesis for the sample.

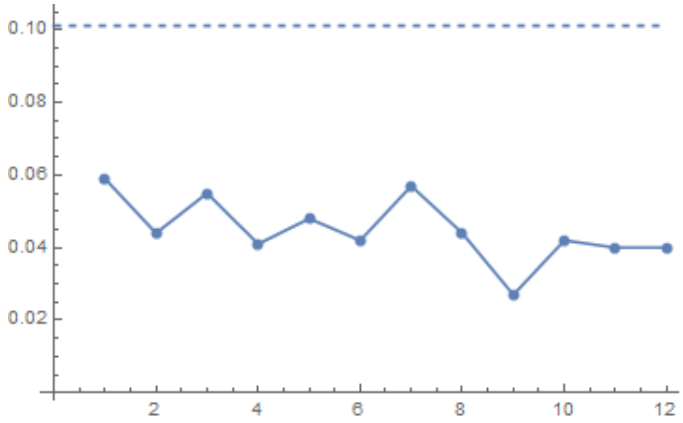


Figure 8.2: Control Chart for the Variance of 12 Samples.

In Figure 8.5 we display a variance control chart and the UCL for the values of Table 8.3 with $\sigma_0^2 = 0.045$, $n = 12$, $\alpha = 0.01$ and $c = 24.72$, which we got from Table A10, Appendix 5 with 11 degrees of freedom. We see that for all samples the hypothesis $\sigma_0^2 = 0.045$ is accepted and the process is allowed to run its course.

We may also need a control chart for the variance with both an upper control limit UCL and a lower control limit LCL in which case we get

$$\text{LCL} = \frac{\sigma_0^2 c_1}{n-1} \quad \text{and} \quad \text{UCL} = \frac{\sigma_0^2 c_2}{n-1} \tag{8.30}$$

where c_1 and c_2 are critical values obtained from Table A10, Appendix 5 with $n-1$ degrees of freedom and the equations

$$P(Y \leq c_1) = \frac{\alpha}{2} \quad \text{and} \quad P(Y \leq c_2) = 1 - \frac{\alpha}{2} \tag{8.31}$$

8.6 Goodness of Fit; The χ^2-Test

A test for **goodness of fit** is a statistical test that confirms or rejects a hypothesis about the unknown distribution of a random variable by using sample values. For example, by using a sample we determine with some probability whether or not an unknown distribution is normal.

Goodness of fit tests are important because often we have some idea about the distribution of a random variable and by using a sample we try to validate or reject it.

A **chi-squared test** or **χ^2-test** is a goodness of fit test for which the null hypothesis is that the unknown distribution is chi-squared.

We recall the useful fact that the distribution of a sum of the squares of k independent standard normal random variables is a chi-squared distribution with k degrees of freedom. In particular, the square of a standard normal distribution is a chi-squared distribution with one degree of freedom.

A chi-squared test is often used to reject the hypotesis that a set of normal random variables is independent.

The χ^2-Test

Let X be a random variable from which we obtain sample values $x_1, x_2, \cdots, x_n$.

Hypothesis: $F(X)$ is the cumulative distribution function of X.

1. We subdivide the real axis into consecutive intervals $I_1, I_2, \cdots, I_k$. Let a_j be the number of the sample values inside I_j.

2. We use $F(x)$ to compute the probability p_j that the random variable X assumes a value in I_j. Let

$$e_j = np_j$$

   (This the theoretically expected number of sample values in I_j, if the hypothesis is true.)

3. We compute

$$\chi_0^2 = \sum_{j=1}^{k} \frac{(a_j - e_j)^2}{e_j} \tag{8.32}$$

   This number measures the total closeness of observed a_j to expected e_j.

   Fact χ_0^2 is approximately a chi-squared distribution with $k-1$ degrees of freedom, provided that $F(X)$ has no unknown parameters. If $F(x)$ has m unknown parameters, then χ_0^2 is approximately a chi-squared distribution with $k-m-1$ degrees of freedom.

4. We choose a significance level α and we determine c in the equation

$$P(\chi^2 \leq c) = 1 - \alpha$$

   by using Table A10, Appendix 5 or CAS.

5. If $\chi_0^2 \leq c$, we accept the hypothesis. If $\chi_0^2 > c$, we reject the hypothesis.

□

Example 8.6.1. A die is rolled 36 times with results shown in the following table.

Die value	1	2	3	4	5	6
frequency x_j	1	4	7	7	8	9

If a chi-squared goodness of fit test is used to test the hypothesis that the die is fair, compute the chi-squared statistic χ_0^2 and test the hypothesis of fairness at significance level $\alpha = 0.05$.

Solution: The null hypothesis is

$$H_o: \ p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$$

The alternative hypothesis is that one of the p_j is not equal to $\frac{1}{6}$. The test is based on 36 trials. Hence, $n = 36$. If we use the intervals $I_1 = (-\infty, 1.5)$, $I_2 = (1.5, 2.5)$, $I_3 = (2.5, 3.5)$, $I_4 = (3.5, 4.5)$, $I_5 = (4.5, 5.5)$, $I_6 = (5.5, \infty)$, then $m = 6$ and we have

$$e_j = np_j = 36 \times \frac{1}{6} = 6$$

So

$$\begin{aligned}
\chi_0^2 &= \sum_{j=1}^{k} \frac{(a_j - e_j)^2}{e_j} = \sum_{j=1}^{6} \frac{(a_j - 6)^2}{6} \\
&= \frac{1}{6}\left(25 + 4 + 1 + 1 + 4 + 9\right) \\
&= \frac{22}{3} \\
&= 7.333
\end{aligned}$$

Now we solve for c the equation

$$P(\chi^2 \leq c) = 1 - \alpha = 1 - 0.05 = 0.95$$

by using Table A10, Appendix 5 with $m-1=5$ degrees of freedom. We get $c=11.07$. Since, $\chi_0^2=7.333\leq c=11.07$, we accept the hypothesis that the die is fair.

□

Example 8.6.2 (χ^2-Test for Normal Distribution)**.** Test whether the random variable X is a normal distribution given the following 100 sample values of X. Use $k=15$ intervals.

400	400	460	490	500	420	400	470	430	520
540	520	420	450	470	430	430	520	500	400
510	510	410	430	410	500	430	520	440	400
530	410	450	510	450	510	420	460	400	540
450	500	440	450	440	480	540	520	400	540
420	470	480	540	470	490	500	500	460	440
540	460	430	530	430	530	530	400	500	530
520	510	450	530	420	450	480	500	450	440
520	440	480	430	410	440	510	420	450	410
510	520	480	510	540	400	510	470	490	510

Table 8.4: 100 Measurements of the Random Variable X.

Solution: The maximum likelihood estimates for μ and σ^2 are the sample mean $\hat{\mu_0}=470.2$, the sample variance $\hat{\sigma}^2=2030.26$, and the sample standard deviation $\hat{\sigma}=45.06$, respectively. Next, we compute the chi-squared statistic χ_0^2 as follows. We check the absolute frequencies of the different values of the sample. These frequencies are shown in Column 2 of Table 8.5.

Next, we perform some computations displayed in Table 8.6. We subdivide $(-\infty,\infty)$ into $k=15$ subintervals $(-\infty,405]$, $(405,415]$, etc., as shown in Column 1. For each subinterval we compute the quantity $((x-470.2)/45.06)$ at each end of the interval. This is done to convert the hypothetically normal distributing into a standard normal. The results are shown in Column 2. We then evaluate the CDF, $\Phi(z)$, of the standard normal at these points as seen in Column 3. The difference $\Phi(x_{j+1})-\Phi(x_j)$ is the theoretical probability p_i that the variable takes a value in the subinterval. Then we compute $e_j=np_i=100p_j$ which is seen in Column 4. Column 5 shows tha a_j. The last column computes the terms of the sum (8.32). Finally we add the terms

of the last column to get

$$\chi_0^2 = 17.182$$

Sample Values	Absolute Frequency	Relative Frequency	Absolute Cumulative Frequency	Relative Cumulative Frequency
400	9	0.09	9	0.09
410	5	0.05	14	0.14
420	6	0.06	20	0.20
430	8	0.08	28	0.28
440	7	0.07	35	0.35
450	9	0.09	44	0.44
460	4	0.04	48	0.48
470	5	0.05	53	0.53
480	5	0.05	58	0.58
490	3	0.03	61	0.61
500	8	0.08	69	0.69
510	10	0.10	79	0.79
520	8	0.08	87	0.87
530	6	0.06	93	0.93
540	7	0.07	100	1.00

Table 8.5: 100 Frequences of X.

Since the normal distribution has two unknown parameters μ and σ, we have that $m = 2$. Now, we solve

$$P(\chi^2 \leq c) = 1 - \alpha = 0.95$$

for $k-m-1 = 15-3 = 12$ degrees of freedom. by using Table A10, Appendix 5 or CAS. We get $c = 21.03$. Since $\chi_0^2 \leq c$, we accept the hypothesis that X has a normal distribution with estimated mean 470.2 and estimated standard deviation 45.06.

x_j	$\dfrac{x_j - 470.2}{45.06}$	$\Phi\left(\dfrac{x_j - 472.41}{40.74}\right)$	e_j	a_j	Term in (8.32)
$(-\infty, 405]$	$-\infty \cdots -1.45$	$0.0000 \ldots 0.0735$	7.35	9	0.370
$(405, 415]$	$-1.45 \cdots -1.23$	$0.0735 \ldots 0.1093$	3.58	5	0.563
$(415, 425]$	$-1.23 \cdots -1.00$	$0.1093 \ldots 0.1587$	4.94	6	0.227
$(425, 435]$	$-1.00 \cdots -0.78$	$0.1587 \ldots 0.2177$	5.90	8	0.747
$(435, 445]$	$-0.78 \cdots -0.56$	$0.2177 \ldots 0.2877$	7.00	7	0.000
$(445, 455]$	$-0.56 \cdots -0.34$	$0.2877 \ldots 0.3669$	7.92	9	0.147
$(455, 465]$	$-0.34 \cdots -0.12$	$0.3669 \ldots 0.4522$	8.53	4	2.406
$(465, 475]$	$-0.12 \ldots 0.11$	$0.4522 \ldots 0.5438$	9.16	5	1.889
$(475, 485]$	$0.11 \ldots 0.33$	$0.5438 \ldots 0.6293$	8.55	5	1.474
$(485, 495]$	$0.33 \ldots 0.55$	$0.6293 \ldots 0.7088$	7.95	3	3.082
$(495, 505]$	$0.55 \ldots 0.77$	$0.7088 \ldots 0.7794$	7.06	8	0.125
$(505, 515]$	$0.775 \ldots 0.99$	$0.7794 \ldots 0.8389$	5.95	10	2.757
$(515, 525]$	$0.99 \ldots 1.22$	$0.8389 \ldots 0.8888$	4.99	8	1.816
$(525, 535]$	$1.22 \ldots 1.44$	$0.8888 \ldots 0.9251$	3.63	6	1.547
$(535, \infty)$	$1.44 \ldots \infty$	$0.9251 \ldots 1.0000$	7.49	7	0.032

Table 8.6: lalal

Appendices

Appendix A

Standard Normal CFD

z	Phi(z)	z	Phi(z)	z	Phi(z)
0	0.5				
0.01	0.504	0.51	0.695	1.01	0.8438
0.02	0.508	0.52	0.6985	1.02	0.8461
0.03	0.512	0.53	0.7019	1.03	0.8485
0.04	0.516	0.54	0.7054	1.04	0.8508
0.05	0.5199	0.55	0.7088	1.05	0.8531
0.06	0.5239	0.56	0.7123	1.06	0.8554
0.07	0.5279	0.57	0.7157	1.07	0.8577
0.08	0.5319	0.58	0.719	1.08	0.8599
0.09	0.5359	0.59	0.7224	1.09	0.8621
0.1	0.5398	0.6	0.7257	1.1	0.8643
0.11	0.5438	0.61	0.7291	1.11	0.8665
0.12	0.5478	0.62	0.7324	1.12	0.8686
0.13	0.5517	0.63	0.7357	1.13	0.8708
0.14	0.5557	0.64	0.7389	1.14	0.8729
0.15	0.5596	0.65	0.7422	1.15	0.8749
0.16	0.5636	0.66	0.7454	1.16	0.877
0.17	0.5675	0.67	0.7486	1.17	0.879
0.18	0.5714	0.68	0.7517	1.18	0.881
0.19	0.5753	0.69	0.7549	1.19	0.883
0.2	0.5793	0.7	0.758	1.2	0.8849
0.21	0.5832	0.71	0.7611	1.21	0.8869
0.22	0.5871	0.72	0.7642	1.22	0.8888
0.23	0.591	0.73	0.7673	1.23	0.8907
0.24	0.5948	0.74	0.7704	1.24	0.8925

0.25	0.5987	0.75	0.7734	1.25	0.8944
0.26	0.6026	0.76	0.7764	1.26	0.8962
0.27	0.6064	0.77	0.7794	1.27	0.898
0.28	0.6103	0.78	0.7823	1.28	0.8997
0.29	0.6141	0.79	0.7852	1.29	0.9015
0.3	0.6179	0.8	0.7881	1.3	0.9032
0.31	0.6217	0.81	0.791	1.31	0.9049
0.32	0.6255	0.82	0.7939	1.32	0.9066
0.33	0.6293	0.83	0.7967	1.33	0.9082
0.34	0.6331	0.84	0.7995	1.34	0.9099
0.35	0.6368	0.85	0.8023	1.35	0.9115
0.36	0.6406	0.86	0.8051	1.36	0.9131
0.37	0.6443	0.87	0.8078	1.37	0.9147
0.38	0.648	0.88	0.8106	1.38	0.9162
0.39	0.6517	0.89	0.8133	1.39	0.9177
0.4	0.6554	0.9	0.8159	1.4	0.9192
0.41	0.6591	0.91	0.8186	1.41	0.9207
0.42	0.6628	0.92	0.8212	1.42	0.9222
0.43	0.6664	0.93	0.8238	1.43	0.9236
0.44	0.67	0.94	0.8264	1.44	0.9251
0.45	0.6736	0.95	0.8289	1.45	0.9265
0.46	0.6772	0.96	0.8315	1.46	0.9279
0.47	0.6808	0.97	0.834	1.47	0.9292
0.48	0.6844	0.98	0.8365	1.48	0.9306
0.49	0.6879	0.99	0.8389	1.49	0.9319
0.5	0.6915	1	0.8413	1.5	0.9332

z	Phi(z)	z	Phi(z)	z	Phi(z)
1.51	0.9345	2.01	0.9778	2.51	0.994
1.52	0.9357	2.02	0.9783	2.52	0.9941
1.53	0.937	2.03	0.9788	2.53	0.9943
1.54	0.9382	2.04	0.9793	2.54	0.9945
1.55	0.9394	2.05	0.9798	2.55	0.9946
1.56	0.9406	2.06	0.9803	2.56	0.9948
1.57	0.9418	2.07	0.9808	2.57	0.9949
1.58	0.9429	2.08	0.9812	2.58	0.9951
1.59	0.9441	2.09	0.9817	2.59	0.9952
1.6	0.9452	2.1	0.9821	2.6	0.9953

1.61	0.9463	2.11	0.9826	2.61	0.9955
1.62	0.9474	2.12	0.983	2.62	0.9956
1.63	0.9484	2.13	0.9834	2.63	0.9957
1.64	0.9495	2.14	0.9838	2.64	0.9959
1.65	0.9505	2.15	0.9842	2.65	0.996
1.66	0.9515	2.16	0.9846	2.66	0.9961
1.67	0.9525	2.17	0.985	2.67	0.9962
1.68	0.9535	2.18	0.9854	2.68	0.9963
1.69	0.9545	2.19	0.9857	2.69	0.9964
1.7	0.9554	2.2	0.9861	2.7	0.9965
1.71	0.9564	2.21	0.9864	2.71	0.9966
1.72	0.9573	2.22	0.9868	2.72	0.9967
1.73	0.9582	2.23	0.9871	2.73	0.9968
1.74	0.9591	2.24	0.9875	2.74	0.9969
1.75	0.9599	2.25	0.9878	2.75	0.997
1.76	0.9608	2.26	0.9881	2.76	0.9971
1.77	0.9616	2.27	0.9884	2.77	0.9972
1.78	0.9625	2.28	0.9887	2.78	0.9973
1.79	0.9633	2.29	0.989	2.79	0.9974
1.8	0.9641	2.3	0.9893	2.8	0.9974
1.81	0.9649	2.31	0.9896	2.81	0.9975
1.82	0.9656	2.32	0.9898	2.82	0.9976
1.83	0.9664	2.33	0.9901	2.83	0.9977
1.84	0.9671	2.34	0.9904	2.84	0.9977
1.85	0.9678	2.35	0.9906	2.85	0.9978
1.86	0.9686	2.36	0.9909	2.86	0.9979
1.87	0.9693	2.37	0.9911	2.87	0.9979
1.88	0.9699	2.38	0.9913	2.88	0.998
1.89	0.9706	2.39	0.9916	2.89	0.9981
1.9	0.9713	2.4	0.9918	2.9	0.9981
1.91	0.9719	2.41	0.992	2.91	0.9982
1.92	0.9726	2.42	0.9922	2.92	0.9982
1.93	0.9732	2.43	0.9925	2.93	0.9983
1.94	0.9738	2.44	0.9927	2.94	0.9984
1.95	0.9744	2.45	0.9929	2.95	0.9984
1.96	0.975	2.46	0.9931	2.96	0.9985
1.97	0.9756	2.47	0.9932	2.97	0.9985
1.98	0.9761	2.48	0.9934	2.98	0.9986
1.99	0.9767	2.49	0.9936	2.99	0.9986
2	0.9772	2.5	0.9938	3	0.9987

Appendix B

Trigonometric Identities

1. $\sin(a+b) = \sin a \cos b + \cos a \sin b$
2. $\cos(a+b) = \cos a \cos b - \sin a \sin b$
3. $\sin(a-b) = \sin a \cos b - \cos a \sin b$
4. $\cos(a-b) = \cos a \cos b + \sin a \sin b$
5. $\sin(2a) = 2 \sin a \cos a$
6. $\cos(2a) = 2\cos^2 a - 1$
7. $\cos^2(a) = \dfrac{1+\cos 2a}{2}$
8. $\sin^2(a) = \dfrac{1-\cos 2a}{2}$
9. $\sin a \cos b = \frac{1}{2}\sin(a+b) + \frac{1}{2}\sin(a-b)$
10. $\sin a \sin b = \frac{1}{2}\cos(a-b) - \frac{1}{2}\cos(a+b)$
11. $\cos a \cos b = \frac{1}{2}\cos(a-b) + \frac{1}{2}\cos(a+b)$
12. $\cos(k\pi) = (-1)^k$, k integer.
13. $\sin(k\pi) = 0$, k integer.
14. $\cos\left((2k-1)\frac{\pi}{2}\right) = 0$, k integer.

Appendix C

Rules of Differentiation

Appendix D

Partial Fractions

Appendix E

Integration by Substitution

1. $\sin(a+b) = \sin a \cos b + \cos a \sin b$
2. $\cos(a+b) = \cos a \cos b - \sin a \sin b$
3. $\sin(a-b) = \sin a \cos b - \cos a \sin b$
4. $\cos(a-b) = \cos a \cos b + \sin a \sin b$
5. $\sin(2a) = 2 \sin a \cos a$
6. $\cos(2a) = 2\cos^2 a - 1$
7. $\cos^2(a) = \dfrac{1+\cos 2a}{2}$
8. $\sin^2(a) = \dfrac{1-\cos 2a}{2}$
9. $\sin a \cos b = \frac{1}{2}\sin(a+b) + \frac{1}{2}\sin(a-b)$
10. $\sin a \sin b = \frac{1}{2}\cos(a-b) - \frac{1}{2}\cos(a+b)$
11. $\cos a \cos b = \frac{1}{2}\cos(a-b) + \frac{1}{2}\cos(a+b)$
12. $\cos(k\pi) = (-1)^k$, k integer.
13. $\sin(k\pi) = 0$, k integer.
14. $\cos\left((2k-1)\frac{\pi}{2}\right) = 0$, k integer.

Appendix F

Integration by Parts

Appendix G

Improper Integrals

Index