POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER III

SELF LEARNING MATERIAL

PAPER : MATC 3.1
(Pure & Applied Streams)

Block - I : Linear Algebra
Block - II : Special Functions
Block - III : Integral Equations & Integral Transforms

Directorate of Open and Distance Learning
University of Kalyani
Kalyani, Nadia
West Bengal, India

Course Preparation Team

1. Mr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani	2. Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

November, 2019

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

Director's Massage

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Sankar Kumar Ghosh, Hon'ble Vice-Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as weel as each and every member of PGBOS (DODL), University of Kalyani, Heartfelt thanks is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

CONTENTS

Core Paper

MATC 3.1
Block - I
Marks : 40 (SSE : 30; IA : 10)

Linear Algebra

Syllabus

• Unit 1 •

Matrices over a field: Matric polynomial, characteristic polynomial, eigen values and eigen vectors, minimal polynomial.

• Unit 2 •

Linear Transformation (L.T.): Definition and the algebra of L.T., Rank and Nullity of L.T., Dual space, dual basis, Representation of L.T. by matrices, Change of basis.

• Unit 3 •

Normal forms of matrices: Diagonalization of matrices, Smith's normal form.

• Unit 4 •

Invariant factors and elementary divisors, Jordan canonical form.

• Unit 5 •

Rational (or Natural Normal) form, triangular forms.

• Unit 6 •

Bilinear and Quadratic forms: Bilinear forms, quadratic forms, reduction and classification of quadratic forms.

Unit 1

Course Structure

- Matrix polynomial, characteristic polynomial
- Eigen values and eigen vectors
- Minimal polynomial.

---

1 Introduction

You are already aware of matrices and its various properties such as determinants, characteristic polynomials, eigen values and eigen vectors. We will revisit them in this unit and learn about the minimal polynomial of matrices and read about the characteristic polynomial, the eigen values and eigen vectors using the information of the minimal polynomial.

Objectives

After reading this unit, you will be able to

- find the characteristic polynomial of a matrix
- find the eigen values and eigen vectors of a matrix
- learn the various properties of a matrix associated with its eigen vectors and eigen values and also its characteristic polynomial
- find the minimal polynomial of a matrix
- learn the relationship between minimal and characteristic polynomials of a matrix.

1.1 Matrix Polynomials

Let F be a field and A be a matrix with entries from the field F. In this chapter, we are concerned mainly with the matrix polynomials, viz., the characteristic and minimal polynomials. Here, we will consider the underlying field to be either $\mathbb{R}$ or $\mathbb{C}$. Let A be an $n \times n$ matrix over the field $\mathbb{R}$. Then, a matrix polynomial for the matrix A is a polynomial with real coefficients and the variables as the matrix A, that is, if

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

is a real polynomial, then the matrix polynomial evaluated at A is given as

$$p(A) = a_0 I + a_1 A + \cdots + a_n A^n$$

where, I is the n-th order identity matrix. Next we will move on to the definition of the characteristic polynomials.

1.1.1 Characteristic Polynomials

Before stating the definition of characteristic polynomials, we will first define the eigen values and eigen vectors of a matrix.

Definition 1.1. Let A be an $n \times n$ matrix over the field $\mathbb{R}$. Then, a real number λ is said to be an eigen value of the matrix if there exists a non-zero vector $v \in \mathbb{R}^n$ such that

$$Av = \lambda v \tag{1.1.1}$$

holds. Then the non-zero vector v is said to be the eigen vector corresponding to the eigen value λ.

The equation (1.1.1) reduces to

$$(A - \lambda I)v = 0$$

which is an n-th order linear equation in n variables. This equation has non-trivial solution if

$$\det(A - \lambda I) = 0$$

The above equation is called the characteristic equation (polynomial) for the matrix A. The roots of the characteristic polynomials give us the eigen values of the matrix. It should be noted that the characteristic polynomial is a monic polynomial which has exactly degree n.

Consider the matrix

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

For any real number λ, the equation $\det(A - \lambda I) = 0$ gives

$$\begin{aligned} \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} &= 0, \\ \text{or, } \lambda^2 + 1 &= 0, \\ \text{or, } \lambda &= \pm i. \end{aligned}$$

So, the characteristic equation is $\lambda^2 + 1 = 0$ which has no roots in the real field, but has roots $\pm i$ in the complex field. So, A has eigen values in the complex field but no eigen value in the real field.

Eigen values can also be defined as

Definition 1.2. If A is an $n \times n$ matrix over a field F, then $c \in F$ is called an eigen value of A in F if the matrix $(A - cI)$ is singular.

Eigen values are often called characteristic roots, latent roots, eigenvalues, proper values, or spectral values in several roots. We shall call them eigen values throughout. We will now discuss certain properties of characteristic polynomials.

Definition 1.3. Let A and B be two $n \times n$ matrices. Then A and B are said to be similar if there exists an invertible matrix P of order n such that

$$A = P^{-1}BP.$$

Theorem 1.4. Similar matrices have the same characteristic polynomial.

Proof. Let A and B be two $n \times n$ similar matrices. Then there exists an invertible matrix P such that

$$A = P^{-1}BP.$$

Then,

$$\begin{aligned} \det(A - \lambda I) &= \det(P^{-1}BP - \lambda I) \\ &= \det(P^{-1}BP - \lambda P^{-1}IP) \\ &= \det(P^{-1}(B - \lambda I)P) \\ &= \det P^{-1}.\det(B - \lambda I).\det P \\ &= \det(B - \lambda I). \end{aligned}$$

□

We will now move on to define the minimal polynomial of a matrix. Let us start with the following example.

Consider the following matrix

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

Then the characteristic polynomial for A is

$$\begin{vmatrix} 3-\lambda & 1 & -1 \\ 2 & 2-\lambda & -1 \\ 2 & 2 & -\lambda \end{vmatrix} = 0$$

which gives $\lambda^3 - 5\lambda^2 + 8\lambda - 4 = (\lambda - 1)(\lambda - 2)^2 = 0$. Thus, 1 and 2 are the eigen values of A. Find the corresponding eigen vectors!

So the characteristic polynomial for A is $f(\lambda) = \lambda^3 - 5\lambda^2 + 8\lambda - 4 = (\lambda - 1)(\lambda - 2)^2 = 0$. It is obvious that for any other polynomial $g(x)$ in $\mathbb{R}[x]$ (since in this case the underlying field is $\mathbb{R}$. Otherwise we would have taken $F[x]$.), we would have

$$h(x) = g(x)f(x) = 0,$$

or, writing it as

$$h(A) = g(A)f(A) = 0,$$

we can say that the polynomial $h(x)$ annihilates A. All such polynomials $h(x) \in \mathbb{R}[x]$ for which $h(A) = 0$ are called the annihilating polynomial of A. We formally define annihilating polynomial as follows.

Definition 1.5. Let A be an $n \times n$ matrix over a field F. Then a polynomial $f(x) \in F[x]$ is called an **Annihilating Polynomial** of A if $f(A) = 0$. By the definition, we can at once say that the characteristic polynomial of A is an annihilating polynomial of A.

We can check a simple fact that the set of all annihilating polynomials of a matrix A forms an ideal I of the polynomial ring $F[x]$ (verify). Now, since F is a field, so the ideal I is necessarily a principal ideal of $F[x]$. It means that there exists a polynomial $m(x) \in I$ such that $I = \langle m(x) \rangle$, that is I is generated by $m(x)$, that is, each element $f(x)$ of I can be written in the form $f(x) = p(x)m(x)$, where, $p(x) \in F[x]$. This $m(x)$ is called the **minimal polynomial** of the matrix A. We formally define the minimal polynomial of a matrix as follows.

Definition 1.6. Let A be an $n \times n$ matrix over a field F. Then the minimal polynomial $m(x)$ of A is the unique monic generator of the ideal of all polynomials over F which annihilate A.

Thus, we arrive at the following theorem.

Theorem 1.7. Let A be an $n \times n$ matrix over a field F and $m(x)$ be the minimal polynomial of A. Then, $m(x)$ divides each of the annihilating polynomial of A.

Theorem 1.8. Let A be an $n \times n$ matrix over a field F. Then the characteristic and minimal polynomials for A have the same roots, except for multiplicities.

Proof. Let m be the minimal polynomial for A. Let c be a scalar. We want to show that $m(c) = 0$ if and only if c is an eigen value. First suppose that $m(c) = 0$. Then

$$p(x) = (x - c)q(x),$$

where, q is a polynomial in F such that $\deg q < \deg p$. By the definition of minimal polynomial, we can say that $q(A) \neq 0$. Now, choose a vector β such that $q(A)\beta \neq 0$. Let $\alpha = q(A)\beta$. Then,

$$\begin{aligned} 0 &= m(A)\beta \\ &= (A - cI)q(A)\beta \\ &= (A - cI)\alpha \end{aligned}$$

and thus, α is an eigen value of A.

Now, suppose that c is an eigen value of A, say $A\alpha = c\alpha$ for some $\alpha \neq 0$. So, by the properties of matrices, we can say that

$$m(A)\alpha = m(c)\alpha.$$

Since $m(A) = 0$ and $\alpha \neq 0$, we have, $m(c) = 0$. Hence c is a root of the minimal polynomial of A. Thus the theorem. □

Example 1.9. Consider the matrix of the previous example.

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

We have seen that the characteristic polynomial of the matrix is

$$f(x) = (x - 1)(x - 2)^2$$

Now, since minimal polynomial divides characteristic polynomial and both have same roots (excepting multiplicities), so the most probable candidates for the minimal polynomial are

1. $m(x) = (x - 1)(x - 2)^2$, or,
2. $m(x) = (x - 1)(x - 2)$.

One may check whether $(A - I)(A - 2I) = 0$. If yes, then the second option is our required minimal polynomial. If not, then the characteristic polynomial and minimal polynomials coincide in this case.

There are various ways to find the minimal polynomial of a matrix (by finding the eigen vectors, rank, etc. of the matrix). We will deal with it in details in the upcoming units.

Exercise 1.10. 1. Find a 3×3 matrix whose minimal polynomial is x^2.

2. Find the minimal polynomial and eigen values of the following matrix.

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

3. Let a, b, c be elements of a filed F, and let A be the following 3×3 matrix over F:

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

Prove that the characteristic polynomial for A is $x^3 - ax^2 - bx - c$ and that this is also the minimal polynomial for A.

Unit 2

Course Structure

- Linear Transformation (L.T.): Definition and the algebra of L.T.
- Rank and Nullity of L.T., Dual space, dual basis,
- Representation of L.T. by matrices, Change of basis.

---

2 Introduction

We are already familiar with the idea of linear transformations from our undergraduate times. This unit helps to recapitulate those earlier notions and introduces certain new ideas on the algebra of linear transformations and the ideas of dual spaces of a vector space. We will learn of these things in detail. We will start with formally defining linear transformations, giving a few examples and stating the old theorems with their applications and then start on to develop the new ideas about dual and double dual spaces thereon.

Objectives

After reading this unit, you will be able to

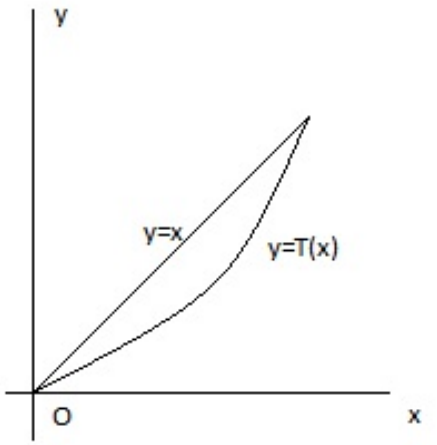
- recapitulate the basic notions of a linear transformation on a vector space
- solve the basic problems related to the representation of a linear transformation (LT) by matrices and change them with basis changes
- solve sums based on the Rank-Nullity theorem
- form an idea about the linear functionals on a vector space V
- define the dual basis on a vector space V
- find the dual basis for the corresponding dual space
- define double dual for a vector space and form the corresponding basis

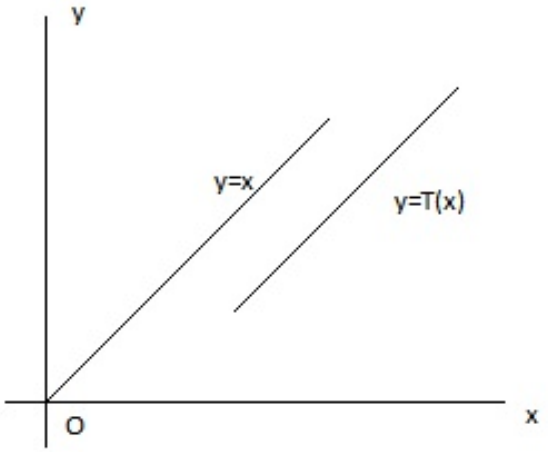
2.1 Transformations

Definition 2.1. Let V and W be two vector spaces over the same field F. A linear transformation from V to W is a function that satisfies the following condition

$$T(ca + db) = cT(a) + dT(b)$$

for all c and $d \in F$ and a, b in V.





A simple calculation yields that $T(0) = 0$ always (can you show it?). Thus, for a simple intuitive example, if we consider the vector space $\mathbb{R}^2$ over the field $\mathbb{R}$, then we can say that any function T from $\mathbb{R}$ to itself is a LT if it takes a line passing through the origin to a line passing through the origin. Let us see the following examples.

Example 2.2. 1. Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be a function defined as $T(v) = v^2$. Then clearly, T takes the line $y = x$ onto the curve $y = x^2$. Hence, T is not a linear transformation on $\mathbb{R}^2$.

2. Consider another example of T on the same vector space $\mathbb{R}^2$ where T is defined as

$$T(v) = v + \alpha$$

where α is a non-zero element of $\mathbb{R}^2$. Thus, we can see that T takes straight lines onto straight lines but does not take origin to itself. Hence, T is not a LT in this case too.

The above example illustrates a few examples of functions which are not LT. Below given are certain standard examples of a LT which are frequently used.

Example 2.3. 1. If V is any vector space, the identity transformation I, defined as $I(v) = v$, is a linear transformation from V into V.

2. The zero transformation 0 on a vector space V, defined as $0(v) = 0$ is also a linear transformation.

Certain other examples include

Example 2.4. 1. Let V be the vector space consisting of all continuous functions on the set of real numbers, over the field of reals. Then the integral operator defined as

$$(T(f))(x) = \int_0^x f(t)dt, \qquad f \in V,$$

is a LT on V.

2. Let V be the vector space consisting of all polynomials on the set of real numbers, over the field of reals. Then the differential operator defined as

$$\begin{aligned} (Df)(x) &= c_1 + 2c_2x + \cdots + kc_kx^{k-1} \\ \text{where,} \quad f(x) &= c_0 + c_1x + \cdots + c_kx^k \in V \end{aligned}$$

is a LT on V.

3. Let V be the vector space consisting of all convergent real sequences over the field of reals. Then the limit operator defined as

$$L(x) = \lim_{n\to\infty} x_n, \qquad x = \{x_n\} \in V,$$

is a LT on V.

Theorem 2.5. Let V be a finite dimensional vector space and $\{a_1, a_2, \ldots a_n\}$ be a basis of V and $\{b_1, b_2, \ldots b_n\}$ be any set of vectors (not necessarily distinct) in another vector space W under the same field F. Then, there exists a unique LT T from V into W such that

$$T(a_i) = b_i, \quad i = 1(1)n.$$

Proof. Since $\{a_1, a_2, \ldots a_n\}$ is a basis of V, so for any $v \in V$, there exists unique scalars $c_1, c_2, \ldots c_n$ of F such that

$$v = c_1a_1 + \cdots + c_na_n.$$

Then we define T as

$$T(v) = c_1b_1 + \cdots + c_nb_n.$$

Then T is a well-defined rule for associating with each vector v of V to a vector $T(v)$ in W. From the definition, we easily get

$$T(a_i) = b_i, \quad i = 1(1)n.$$

To see that T is linear, let us consider another vector w of V as

$$w = d_1a_1 + \cdots + d_na_n$$

and two other scalars x and y in F. Now,

$$\begin{aligned} xv + yw &= xc_1a_1 + \cdots + xc_na_n + yd_1a_1 + \cdots + yd_na_n \\ &= (xc_1 + yd_1)a_1 + \cdots (xc_n + yd_n)a_n. \end{aligned}$$

Then,

$$\begin{aligned} T(xv + yw) &= (xc_1 + yd_1)b_1 + \cdots (xc_n + yd_n)b_n \\ &= xc_1b_1 + \cdots + xc_nb_n + yd_1b_1 + \cdots + yd_nb_n \\ &= x(c_1b_1 + \cdots + c_nb_n) + y(d_1b_1 + \cdots + d_nb_n) \\ &= xT(v) + yT(w). \end{aligned}$$

Hence, T is linear. Now, let U be another LT from V into W such that $U(a_i) = b_i$, $i = 1(1)n$, then for any vector $v = \sum_{i=1}^{n} x_ia_i$, we have

$$\begin{aligned} U(v) &= U\left(\sum_{i=1}^{n} x_ia_i\right) \\ &= \sum_{i=1}^{n} x_iU(a_i) \\ &= \sum_{i=1}^{n} x_ib_i. \end{aligned}$$

so that U is exactly the same as the rule as T is defined. Hence, T is unique. □

Example 2.6. The vectors $u = (1, 2)$, $v = (3, 4)$ are linearly independent and therefore form a basis for $\mathbb{R}^2$. Then, by the previous theorem, there exists a LT T from $\mathbb{R}^2$ to $\mathbb{R}^2$ such that $T(u) = (3, 2, 1)$ and $T(v) = (6, 5, 4)$. Then, we must be able to find $T(1, 0)$ such that

$$(1, 0) = cu + dv = c(1, 2) + d(3, 4)$$

which gives $c = -2$ and $d = 1$. Thus,

$$\begin{aligned} T(1,0) &= -2(3,2,1) + (6,5,4) \\ &= (0,1,2). \end{aligned}$$

There are other interesting subspaces associated with a LT as we will define now.

Definition 2.7. Let V and W be vector spaces over the field and let T be a LT from V into W. Then the **Null Space** of T is the set of all vectors v in V such that $T(v) = 0$. This is clearly a subset of V because

1. $T(0) = 0$, so that N is non-empty;
2. if $T(v) = T(w) = 0$, then

$$T(cv + dw) = cT(v) + dT(w) = c0 + 0 = 0$$

so that $cv + dw$ also belongs to the null space. The dimension of the null space of T is called the **Nullity** of T.

Definition 2.8. The range of T is a subspace of the space W because if a, b in the range of T, then there exists vectors u and v in V such that $T(u) = a$ and $T(v) = b$. Then for the scalars x and y, $T(xu + yv) = xT(u) + yT(v) = xa + yb$. Hence, $xa + yb$ is also in the range T. The dimension of the range of T is called the **Rank** of T.

Theorem 2.9. A LT T is injective if and only if $N = \{0\}$.

Proof. The proof is trivial and has been left as an exercise. □

We have the celebrated Rank-Nullity Theorem for Linear Transformations as follows:

Theorem 2.10. Let V and W be vector spaces over the field F and let T be a LT from V into W. Suppose that V is finite-dimensional. Then

$$\text{Rank}(T) + \text{Nullity}(T) = \text{Dim}V.$$

Proof. Let $\{v_1, v_2, \ldots, v_k\}$ be a basis of N, the null space of T. Then, the above basis can be extended to a basis $\{v_1, v_2, \ldots, v_n\}$ of V. We shall now prove that $\{T(v_{k+1}), \ldots, T(v_n)\}$ is a basis for the range of T. The vectors $T(v_1), T(v_2), \ldots, T(v_n)$ certainly span the range of T, and since $T(v_j) = 0$ for $j \leq k$, we see that $T(v_{k+1}), \ldots, T(v_n)$ span the range of T. To check their independence, suppose that there are scalars c_i such that

$$\sum_{i=k+1}^{n} c_i T(v_i) = 0,$$

which gives

$$T\left(\sum_{i=k+1}^{n} c_i v_i\right) = 0$$

and hence, the vector $v = \sum_{i=k+1}^{n} c_i v_i$ is in the null space of T. Since $\{v_1, v_2, \ldots, v_k\}$ is a basis of N, so v can be represented as a finite linear combination of them, that is,

$$\sum_{i=k+1}^{n} c_i v_i = \sum_{i=1}^{k} b_i v_i$$

and hence

$$\sum_{i=1}^{k} b_i v_i - \sum_{i=k+1}^{n} c_i v_i = 0.$$

Since $\{v_1, v_2, \ldots, v_n\}$ is linearly independent, so we have, $b_1 = b_2 = \cdots = b_k = c_{k+1} = \cdots = c_n = 0$. Thus, we have proved the linear independence of $T(v_{k+1}), \ldots, T(v_n)$ and hence it is a basis of the range of T. Thus, when nullity is k, the rank of T is $n - k$, thus giving us the required result. □

Note 2.11. We know that any set of vectors with the zero element is always linearly dependent. So, the basis of the null space of T never contains the zero element. Thus, if N does not contain any element other than the zero element, then the nullity of T is zero.

The above theorem has huge applications.

Corollary 2.12. A LT T is surjective if and only if Rank$T = \dim V$.

Proof. Left as exercise. □

Exercise 2.13. 1. Find the rank and nullity of the following linear transformations:

   a. $T(x, y, z) = (x - y, y - z, z - x)$.
   b. $T(x, y, z) = (2x, y, 0)$.
   c. $T(x, y, z) = (2x + 3z, 4z, 5y - z)$.

2. Let T be a vector space and T a linear transformation from V to V. Prove that the following two statements are equivalent.

   a. The intersection of the range of T and the null space of T is the zero subspace of V.
   b. If $T(T(v)) = 0$, then $T(v) = 0$.

3. Describe explicitly a LT from $\mathbb{R}^3$ to $\mathbb{R}^2$ for which the range space is spanned by the vectors $(1, 0, -1)$ and $(1, 2, 2)$.

2.1.1 Matrix Representation of Linear Transformations

We have seen that a LT can be represented by matrices earlier depending upon the bases of the vector spaces. Same linear transformation can give rise to different matrices and they are in fact similar. To each matrix, there is a linear transformation, but there may be many matrices corresponding to a single linear transformation, varying with the change in basis. Let us have an illustration.

Illustration 2.14. Let T be a linear transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$ defined as

$$T(x, y) = (x - y, y).$$

Consider the standard **ordered** basis $\mathcal{B} = \{(1, 0), (0, 1)\}$ of $\mathbb{R}^2$. Suppose we are to represent T with respect to the basis $\mathcal{B}$ on both sides. Then the resulting matrix is represented as $[T]_{\mathcal{B}}$. We find it as follows:

$$\begin{aligned} T(1,0) &= (1,0) = \mathbf{1}(1,0) + \mathbf{0}(0,1) \\ T(0,1) &= (-1,1) = -\mathbf{1}(1,0) + \mathbf{1}(0,1) \end{aligned}$$

and the resulting matrix becomes

$$[T]_{\mathcal{B}} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

Again, if we consider another **ordered** basis $\mathcal{C} = \{(1, 1), (1, 0)\}$ of $\mathbb{R}^2$ as the domain set and the basis $\mathcal{B}$ of the range set. Then we have

$$\begin{aligned} T(1,1) &= (0,1) = \mathbf{0}(1,0) + \mathbf{1}(0,1) \\ T(1,0) &= (1,0) = \mathbf{1}(1,0) + \mathbf{0}(0,1) \end{aligned}$$

and the resulting matrix $[T]_{\mathcal{C}}^{\mathcal{B}}$ is given by

$$[T]_{\mathcal{C}}^{\mathcal{B}} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

We have certain theorems in connection to these.

Theorem 2.15. Let V and W be finite-dimensional vector spaces with ordered bases $\mathcal{B}$ and $\mathcal{C}$ respectively, and let $T : V \to W$ a be linear transformation. Then for each $v \in V$, we have

$$[T(v)]_{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}}[v]_{\mathcal{B}}.$$

Theorem 2.16. Let V and W be finite-dimensional vector spaces with ordered bases $\mathcal{B}$ and $\mathcal{C}$ respectively, and let $T, U : V \to W$ be linear transformations. Then

1. $[T + U]_{\mathcal{B}}^{\mathcal{C}} = [T]_{\mathcal{B}}^{\mathcal{C}} + [U]_{\mathcal{B}}^{\mathcal{C}}$.
2. $[aT]_{\mathcal{B}}^{\mathcal{C}} = a[T]_{\mathcal{B}}^{\mathcal{C}}$ for all scalars a.

Theorem 2.17. Let U, V, W be finite-dimensional vector spaces with ordered bases $\mathcal{A}, \mathcal{B}, \mathcal{C}$ respectively. Let $T : U \to V$ and $S : V \to W$ be linear transformations. Then

$$[ST]_{\mathcal{A}}^{\mathcal{C}} = [S]_{\mathcal{B}}^{\mathcal{C}}[T]_{\mathcal{A}}^{\mathcal{B}}.$$

The purpose of matrix representation for a linear transformation T is to enable us to analyse T by working with the matrix, say M. If M is easy to work with, we have gained an advantage; if not, we have no advantage. Since different bases lead to different matrices, the "right" choice of basis to obtain a simple matrix M, such as a diagonal matrix, is important. Diagonal matrices are the easiest to work with. For now, we will restrict our attention to the cases when $v = W$. But, before going into details, let us check the following.

Let $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ and $\mathcal{C} = \{w_1, w_2, \ldots, w_n\}$ be two bases of a vector space V. Then, for each i, we have certain scalars p_{ij} such that

$$\begin{aligned} v_1 &= p_{11}w_1 + p_{12}w_2 + \cdots + p_{1n}w_n \\ v_2 &= p_{21}w_1 + p_{22}w_2 + \cdots + p_{2n}w_n \\ &\vdots \\ v_n &= p_{n1}w_1 + p_{n2}w_2 + \cdots + p_{nn}w_n \end{aligned}$$

which gives

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

Let

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

Then P is called the Transition matrix from the basis $\mathcal{B}$ to $\mathcal{C}$. This transition matrix is invertible. In fact, if Q is the transition matrix from the basis $\mathcal{C}$ to $\mathcal{B}$, then

$$Q = P^{-1}.$$

Now, let us come back to our discussion. We have seen that a linear transformation can have various matrix representations depending upon the choice of basis. Now, what strikes us is that whether there is certain relationship between these matrices. We have the following theorem in this direction.

Theorem 2.18. Let $T : V \to V$ be a linear transformation. Then, any two matrices representing T are similar.

Exercise 2.19.

1. Find the matrix representation of the following linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined as $T(x, y) = (x + 6y, 3x + 4y)$. Also find the matrix representation of T with respect to the basis $\{(2, -1), (1, 1)\}$.
2. Find the matrix representation of the rotation transformation by an angle $\pi/4$ radians counter-clockwise with respect to the standard basis and the basis $\{(1, 1), (1, 2)\}$.
3. Find the matrix representation of $T(x, y, z) = (x + 2y, x + y + z, z)$ with respect to the standard basis and the basis $\{(1, 1, 0), (0, 1, 1), (1, 0, 1)\}$.

2.2 Algebra Of Linear Transformations

In the study of linear transformations from V to W, it is of fundamental importance that the set of these transformations inherits a natural vector space structure. The set of linear transformations from a space V into itself has even more algebraic structure, because ordinary composition of functions provide a "multiplication" of such transformations. Let us see.

Theorem 2.20. Let V and W be vector spaces over the field F. Let T and U be linear transformations from V into W. The function $T + U$ defined by

$$(T + U)(v) = T(v) + U(v)$$

is a linear transformation from V into W. If c is any scalar, then the function cT defined by

$$(cT)(v) = cT(v)$$

is a linear transformation from V into W. The set of all linear transformations from V into W, together with the addition and scalar multiplication defined above, is a vector space over the field F.

Proof. Suppose T and U are linear transformations from V into W and $T + U$ is defined as given. Then we first show that $T + U$ is linear. Let $c, d \in F$. Then

$$\begin{aligned}(T + U)(cu + dv) &= T(cu + dv) + U(cu + dv)\\ &= cT(u) + dT(v) + cU(u) + dU(v)\\ &= c(T(u) + U(u)) + d(T(v) + U(v))\\ &= c(T + U)(u) + d(T + U)(v).\end{aligned}$$

Similarly, we can show that for scalar $c \in F$ and some additional scalars x, y, we have

$$\begin{aligned}(cT)(xu + yv) &= c(T(xu + yv))\\ &= c(xT(u) + yT(v))\\ &= cxT(u) + cyT(v)\\ &= x(cT(u)) + y(cT(v))\\ &= x((cT)(u)) + y((cT)(v))\end{aligned}$$

This shows that cT is linear. The zero transformation from V into W is also linear. It is a routine exercise to check that the other properties of vector space are satisfied similarly. Hence the result. □

The vector space thus formed, is denoted by the symbol $L(V, W)$. We note that $L(V, W)$ is defined only when V and W are defined over the same field.

Theorem 2.21. Let V be an n-dimensional vector space over the field F, and let W be an m-dimensional vector space over the field F. Then the space $L(V, W)$ is finite-dimensional and has dimension mn.

Proof. Let

$$\mathcal{B} = \{v_1, v_2, \ldots, v_n\} \qquad \mathcal{C} = \{w_1, w_2, \ldots, w_m\}$$

be ordered bases for V and W, respectively. For each integers (p, q) with $1 \leq p \leq m$ and $1 \leq q \leq n$, we define a linear transformation $E^{p,q}$ from V into W by

$$\begin{aligned}E^{p,q}(v_i) &= 0, \quad \text{when } i \neq q\\ &= w_p, \quad \text{when } i = q\end{aligned}$$

or,

$$E^{p,q}(v_i) = \delta_{iq} w_p.$$

According to our first theorem, there exists a unique linear transformation from V into W satisfying these conditions. The claim is that, these mn transformations $E^{p,q}$ form a basis for $L(V, W)$. Let T be a linear

transformation from V into W. For each j, $1 \leq j \leq n$, let $A_{ij}, \ldots, A_{mj}$ be the coordinates of the vector $T(v_j)$ in the ordered basis $\mathcal{C}$, that is,

$$T(v_j) = \sum_{p=1}^{m} A_{pj} w_p.$$

We wish to show that

$$T = \sum_{p=1}^{m} \sum_{q=1}^{n} A_{pq} E^{p,q}. \tag{2.2.1}$$

Let U be the linear transformation in the right hand member of the above equation. Then for each j,

$$\begin{aligned} U(v_j) &= \sum_{p} \sum_{q} A_{pq} E^{p,q}(v_j) \\ &= \sum_{p} \sum_{q} A_{pq} \delta_{jp} w_p \\ &= \sum_{p=1}^{m} A_{pj} w_p \\ &= T(v_j). \end{aligned}$$

and consequently $U = T$. Now, (2.2.1) shows that $E^{p,q}$ spans $L(V, W)$. We must prove that they are independent. But this is clear from what we did above; for, if the transformation

$$U = \sum_{p} \sum_{q} A_{pq} E^{p,q}$$

is the zero transformation, then $U(v_j) = 0$ for each j, so

$$\sum_{p=1}^{m} A_{pj} w_p = 0$$

and the independence if w_p implies that $A_{pj} = 0$ for every p and j. Hence the proof. □

Theorem 2.22. Let V, W and Z be vector spaces over the field F. Let $T : V \to W$ and $U : W \to Z$ be linear transformations. Then the composition function UT defined by $UT(v) = U(T(v))$ is a linear transformation from V into Z.

Proof. Left as exercise. □

Definition 2.23. If V is a vector space over a field F, then a linear operator on V is a linear transformation from V into V.

In the previous theorem, when $V = W = Z$, and U and T are linear operators on the space V, we see that the composition UT is again a linear operator on V. The space $L(V, V)$ "has a multiplication" defined on it by composition. In this case the operator TU is also defined, and one should note that in general $UT \neq TU$, that is, $UT - TU \neq 0$. We should take special note of the fact that if T is a linear operator on V then we can compose T with T. We shall use the notation $T^2 = TT$, and in general, $T^n = TT \cdots T(n \text{ factors})$ for $n = 1, 2, \ldots$. We define $T^0 = I$ if $T \neq 0$.

Theorem 2.24. Let V be a vector space over the field F; let U and T_1 and T_2 be linear operators on V and let $c \in F$. Then

1. $IU = UI = U$;
2. $U(T_1 + T_2) = UT_1 + UT_2$; $(T_1 + T_2)U = T_1U + T_2U$;
3. $c(UT_1) = (cU)T_1 = U(cT_1)$.

In everything we have so far discussed, we have left out the invertibility of linear operators. Under what conditions, does a linear operator admit of an inverse, that is, there exists a linear operator T^{-1} for which $TT^{-1} = T^{-1}T = I$?

Definition 2.25. A LT T from a space V to another space W is said to be invertible if there exists a LT U such that $TU = UT = I$. Such function U, if it exists, is unique.

We note that the by the theory of functions, we know that a function is invertible if it is bijective. Thus, by the rank-nullity theorem, we can say that the dimensions of both the spaces V and W must be the same. Let us see the following theorem.

Theorem 2.26. Let V and W be vector spaces over the field F and let T be a LT from V into W. If T is invertible, then the function T^{-1} is also a LT from W onto V.

Proof. When T is bijective, there exists a uniquely determined function T^{-1} which maps W onto V. To prove the linearity of T^{-1}, let us take two vectors b_1 and b_2 in W and two scalars x and y. Let $a_i = T^{-1}(b_i)$, $i = 1, 2$. Then, we have $T(a_i) = b_i$ for all i. Now, since T is linear,

$$\begin{aligned} T(xa_1 + ya_2) &= xT(a_1) + yT(a_2) \\ &= xb_1 + yb_2 \end{aligned}$$

Thus, $xa_1 + ya_2$ is the unique vector in V such that $T(xa_1 + ya_2) = xb_1 + yb_2$ which means that

$$T^{-1}(xb_1 + yb_2) = xa_1 + ya_2 = xT^{-1}(b_1) + yT^{-1}(b_2)$$

which shows that T^{-1} is linear. □

Definition 2.27. A linear transformation T is said to be non-singular if $T(v) = 0$ implies that $v = 0$, that is, if the null space comprises of only the singleton set $\{0\}$. Otherwise, T is said to be singular.

Theorem 2.28. Let T be a LT from V into W. Then T is non-singular if and only if T carries each linearly independent subset of V into a linearly independent subset of W.

Proof. First suppose that T is non-singular. Let S be a linearly independent subset of V. If $S = \{v_1, v_2, \ldots, v_k\}$, then $T(v_1), T(v_2), \ldots, T(v_k)$ are linearly independent, for if

$$c_1T(v_1) + c_2T(v_2) + \cdots + c_kT(v_k) = 0$$

and then

$$T(c_1v_1 + c_2v_2 + \cdots + c_kv_k) = 0$$

and since T is non-singular,

$$c_1v_1 + c_2v_2 + \cdots + c_kv_k = 0$$

from which it follows that each $c_i = 0$ because S is linearly independent set. This shows that the image of S under T is independent.

Suppose that T carries linearly independent set into linearly independent set. Let a be a non-zero vector in V. Then the set S consisting of the one vector a is independent. The image of S is the set consisting of the one vector $T(a)$, and this set is independent. Thus, $T(a) \neq 0$, because the set consisting of the zero vector alone is independent. This shows that the null space of T is the zero subspace, that is, T is non-singular. □

Theorem 2.29. Let V and W be finite-dimensional vector spaces over the field F such that $\dim V = \dim W$. If T is a LT from V into W, the following are equivalent:

1. T is invertible.
2. T is non-singular.
3. T is onto.
4. If $\{v_1, v_2, \ldots, v_n\}$ is a basis for V, then $\{T(v_1), T(v_2), \ldots, T(v_n)\}$ is a basis for W.
5. There is some basis $\{v_1, v_2, \ldots, v_n\}$ for V such that $\{T(v_1), T(v_2), \ldots, T(v_n)\}$ is a basis for W.

2.3 Dual Spaces

Definition 2.30. If V is a vector space over the field F, a linear transformation f from V into the scalar field F is called a linear functional on V.

The concept of linear functional is important in the study of finite-dimensional spaces because it helps to organize and clarify the discussion of subspaces, linear equations, and coordinates.

Example 2.31. Let n be a positive integer and F a field. If A is an $n \times n$ matrix with entries in F, then the trace of A is a scalar

$$\text{tr}A = A_{11} + A_{22} + \cdots + A_{nn}.$$

Then it is a linear functional on the matrix space $F^{n\times n}$ (verify!)

Example 2.32. Let $[a, b]$ be a closed interval on the real line and let $C([a, b])$ be the space of continuous real-valued functions on $[a, b]$. Then

$$L(g) = \int_a^b g(t)dt$$

defines a linear functional on $C([a, b])$.

Definition 2.33. If V is a vector space, then the collection of all linear functionals on V forms a vector space $L(V, F)$ and it is called the **Dual Space** of V. It is also denoted by V^*.

From the knowledge of the dimension of the space $L(V, W)$, we can say that

$$\dim V = \dim V^*.$$

Let $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ be a basis for V. Then, by the first theorem of this unit, there exists a unique linear functional f_i on V such that

$$f_i(v_j) = \delta_{ij}.$$

In this way, we can obtain from $\mathcal{B}$, a set of n distinct linear functionals $f_1, f_2, \ldots, f_n$ on V. These functionals are also linearly independent. For, suppose

$$f = \sum_{i=1}^{n} c_i f_i.$$

Then,

$$\begin{aligned} f(v_j) &= \sum_{i=1}^{n} c_i f_i(v_j) \\ &= \sum_{i=1}^{n} c_i \delta_{ij} \\ &= c_j. \end{aligned}$$

In particular, if f is the zero functional, $f(v_j) = 0$ for each j and hence the scalars c_j are all 0. Now, $f_1, f_2, \ldots, f_n$ are n linearly independent functionals, and since we know that V^* has dimension n, it must be that $\mathcal{B}^* = \{f_1, f_2, \ldots, f_n\}$ is a basis for V^*. This is called the dual basis of $\mathcal{B}$.

Theorem 2.34. Let V be a finite-dimensional vector space over the field F, and let $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ be a basis for V. Then there is a unique dual basis $\mathcal{B}^* = \{f_1, f_2, \ldots, f_n\}$ for V^* such that $f_i(v_j) = \delta_{ij}$. For each linear functional f on V we have

$$f = \sum_{i=1}^{n} f(v_i) f_i$$

and for each vector v in V we have

$$v = \sum_{i=1}^{n} f_i(v) v_i.$$

Proof. The above discussion shows that there i s a unique basis which is dual to the basis $\mathcal{B}$. If f is a linear functional on V, then f is some linear combination of f_i as

$$f = \sum_{i=1}^{n} c_i f_i.$$

Also we have observed that the scalars c_j must be given by $c_j = f(v_j)$. Similarly, if

$$v = \sum_{i=1}^{n} x_i v_i$$

is a vector in V, then

$$\begin{aligned} f_j(v) &= \sum_{i=1}^{n} x_i f_j(v_i) \\ &= \sum_{i=1}^{n} x_i \delta_{ij} \\ &= x_j. \end{aligned}$$

So that the unique expression for v as a linear combination of the v_i is

$$v = \sum_{i=1}^{n} f_i(v) v_i.$$

□

2.4 Few Probable Questions

1. Show that there exists a unique linear transformation from a finite-dimensional vector space V into another vector space W over the same field sending the basis elements $\{v_1, v_2, \ldots, v_n\}$ to another set of arbitrary vectors $\{w_1, w_2, \ldots, w_n\}$, not necessarily distinct.

2. State and prove the Rank-Nullity Theorem.

3. Show that the space of linear transformations $L(V, W)$ from an n-dimensional space V into an m-dimensional space W is of dimension mn.

4. Define non-singular linear transformations. Show that the inverse of a non-singular linear transformation is also so.

5. Find a basis for the dual space of an n-dimensional space V.

6. Show that a non-singular linear transformation takes a basis to a basis.

Unit 3

Course Structure

- Normal forms of matrices: Diagonalization of matrices,
- Smith's normal form.

---

3 Introduction

As we have already mentioned in the previous unit, diagonal matrices are the easiest to deal with. And we have also seen that different bases give rise to different matrices for a linear transformation, so our main aim is to find a particular basis $\mathbb{B}$ for a vector space V, for which a particular linear transformation (or rather, a linear operator) T, defined on V can be represented as a diagonal matrix. It is not always the case that there always exists such a basis for which T can be represented as a diagonal matrix. We will study mainly the cases and circumstances, under which this is possible. And if such basis does not exist, then what are the simplest possible type of matrix by which we can represent T. These are the various issues that will be addressed in this unit.

Objectives

After reading this unit, you will be able to:

- define the characteristic values and vectors of a linear transformation
- recapitulate the basic notions about minimal and characteristic polynomials of a transformation
- define algebraic and geometric multiplicities of a particular eigen value
- define the eigen spaces of a transformation
- determine the cases when a transformation is diagonalizable
- determine the cases when a transformation is not diagonalizable
- find the necessary and sufficient condition for diagonalizability of a transformation
- learn about the Smith's Normal form

3.1 Diagonalizability

As we have already mentioned before, diagonalizability is something related to the matrix of a LT being diagonal. But, before going into the definition of diagonalizability, let us recollect the general notions of eigen values and eigen vectors of a matrix.

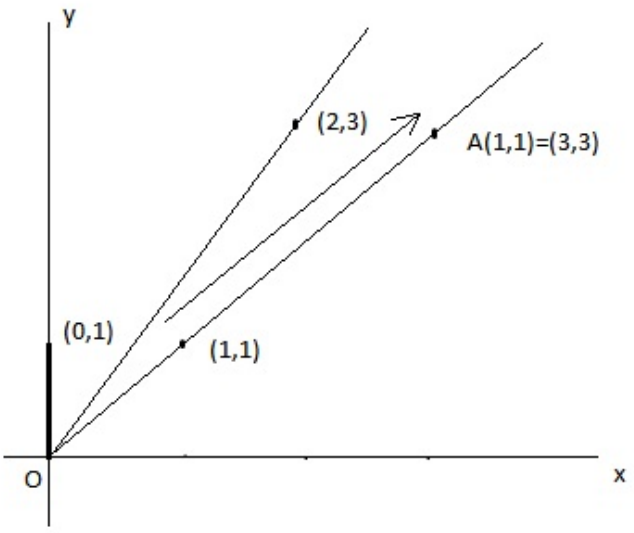


Figure 1: Eigen Values and Eigen Vectors Geometrically

When we operate the matrix over a vector (v_1, v_2) of $\mathbb{R}^2$, and equate it to a constant multiple of (v_1, v_2), we get the system

$$\begin{aligned} v_1 + 2v_2 &= cv_1 \\ 3v_2 &= cv_2 \end{aligned}$$

Geometrically speaking, when we take a particular vector (v_1, v_2) of the xy-plane and operate the matrix on it, we the resulting vector is a scalar multiple of the original one. That is, the resulting vector is either a contracted or expanded form of the original vector depending on the value of c. For example, if we take the vector $(1, 1)$, then the resulting vector will be $(3, 3) = 3(1, 1)$. That is, the particular vector is expanding to thrice its original value.

On the other hand, if we operate the matrix over the vector $(0, 1)$, then the resulting vector $(2, 3)$ is not on the line joining $(0, 1)$ and $(2, 3)$. The vector $(1, 1)$ is called an eigen vector and 3 is the corresponding eigen value. $(0, 1)$ is not an eigen vector.

To summerize, we say that any matrix corresponds to a particular LT and those vectors which do not change their direction on the application of the LT are called its eigen vectors and the factor by which it contracts or expands, is called the corresponding eigen value. We are now in a position to formally define eigen values and eigen vectors of a matrix.

Definition 3.1. Let $T : V \to V$ be a LT over vector spaces on the field F. Then a **non-zero** vector $v \in V$ is said to be an eigen value of T if $T(v) = cv$ for some $c \in F$. This c is called the corresponding eigen value of T.

To find eigen value and eigen vectors of a LT, we generally find so for the corresponding matrix representations of T. It is independent of the bases since similar matrices have same eigen values.

It is important to note that T may not have any eigen value in the first place. And if V is finite-dimensional, say having dimension n, then T can have atmost n eigen values. And the eigen vectors can also be seen as the

null space of the transformation $T - cI$ (of course ignoring the zero vector).

Now, our main concern is to check whether a given LT can be represented as a diagonal matrix or not. So, we are in search of that particular basis of V for which it can be done. If there exists certain basis for which T can be represented as a diagonal matrix, then T is said to be diagonalizable, otherwise T is non-diagonalizable.

Definition 3.2. A linear transformation $T : V \to V$, where V is a **finite-dimensional** vector space, is said to be diagonalizable if there exists a basis $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ for which the corresponding matrix is a diagonal matrix.

A diagonal matrix is of the form

$$\begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_n \end{bmatrix}.$$

An identity matrix is the most common example of a diagonal matrix. So, if we consider the eigen values and vectors of a LT T, that is the vectors v_i satisfying $Tv_i = c_i I v_i$, or the non-zero vectors of the null space $T - c_i I$. The intuitive idea is to break the matrix into diagonal blocks of the form

$$\begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_1 \end{bmatrix},$$

the above block being the diagonal block corresponding to the eigen value c_1. Thus, if the sum of the size of the blocks equals the dimension of V, then T stands diagonalized and the corresponding diagonal matrix is

$$\begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 \\ 0 & c_1 & 0 & \cdots & 0 \\ 0 & 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_n \end{bmatrix}.$$

The size of each block is determined by the "size" of the null spaces, that is, dimension of the null spaces, that is the number of linearly independent eigen vectors spanning each null space. In this way, we come to another equivalent definition of diagonalizability.

Definition 3.3. A linear transformation $T : V \to V$, where V is a **finite-dimensional** vector space, is said to be diagonalizable if there exists a basis $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ comprising of the eigen vectors of T.

Let us illustrate the process.

Illustration 3.4. Let A be an $n \times n$ matrix over a filed F. We first find the eigen values using the "traditional" ways by finding the characteristic polynomial. Let $c_1, c_2, \ldots, c_k \in F$ be the eigen values of A. We find the rank of each of the matrices $A - c_i I$ and then find out the nullity, that is, dimension of the null space of $A - c_i I$ using the Rank-Nullity theorem, for each i, $1 \leq i \leq k$. If $\sum_{i=1}^{k} \dim(A - c_i I) = n$, then the matrix A is diagonalizable otherwise, if $\sum_{i=1}^{k} \dim(A - c_i I) < n$, A is non-diagonalizable. For A, there exists a corresponding linear operator from the vector space F^n to F^n.

Example 3.5. Let A be a real 3×3 matrix

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

Then the characteristic polynomial of A is

$$\begin{vmatrix} x-3 & -1 & 1 \\ -2 & x-2 & 1 \\ -2 & -2 & x \end{vmatrix} = (x-1)(x-2)^2.$$

Then the eigen values of A are 1 and 2. Suppose that T is the linear operator on $\mathbb{R}^3$ which is represented by A in the standard basis. We will find the rank of the matrices $A - I$ and $A - 2I$. Now,

$$A - I = \begin{bmatrix} 2 & 1 & -1 \\ 2 & 1 & -1 \\ 2 & 2 & -1 \end{bmatrix}$$

has clearly rank equals to 2 and hence nullity equals to $3 - 2 = 1$. Also, the matrix

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

has rank 2 and hence nullity $3 - 2 = 1$. When we sum up the nullities of these two matrices, we get $1 + 1 = 2 \neq 3$. Thus, A is not diagonalizable. The nullities of the matrices $A - I$ and $A - 2I$ together tell us that the null space of the above matrices are spanned by one vector space each, that is, there are a maximum of two distinct linearly independent eigen vectors of A and hence we are unable to find a basis of A containing the eigen vectors.

Definition 3.6. Let T be a linear operator over a finite dimensional vector space V and let $c \in F$ be an eigen value of T. Then the null space of the linear operator $T - cI$ is called the **eigen space** of the corresponding eigen value c and the dimension of the eigen space, that is, the nullity of the operator $T - cI$ is called the **geometric multiplicity** of c.

It is a routine exercise to check that the eigen spaces form vector subspaces of V and has been left as an exercise.

Definition 3.7. For an eigen value c of a particular operator T, the power to which the factor $(x - c)$ is raised in the corresponding characteristic polynomial of the matrix representation of T is called the **algebraic multiplicity** of c.

Thus, in the previous example, the algebraic multiplicity of 1 and 2 are 1 and 2 respectively and their corresponding geometric multiplicities are equal to 1 each. We can say that the algebraic multiplicity of an eigen value if always greater than or equals to its geometric multiplicity. Also, the algebraic multiplicities of all the eigen values add up to the dimension of the parent vector space and is less than or equal to the dimension if we consider the geometric multiplicities. When the sum of the geometric multiplicities add up to the dimension of the vector space, we call the operator to be diagonalizable.

Example 3.8. Let T be a linear operator on $\mathbb{R}^3$ which is represented in the standard ordered basis by the matrix

$$A = \begin{bmatrix} 5 & -6 & -6 \\ -1 & 4 & 2 \\ 3 & -6 & -4 \end{bmatrix}.$$

Let us find the characteristic polynomial of A as

$$\begin{vmatrix} x-5 & 6 & 6 \\ 1 & x-4 & -2 \\ -3 & 6 & x+4 \end{vmatrix} = (x-2)^2(x-1).$$

So, 2 and 1 are the eigen values of A with algebraic multiplicities 2 and 1 respectively. We will now find the algebraic and geometric multiplicities of the eigen values. The two matrices

$$A - I = \begin{bmatrix} 4 & -6 & -6 \\ -1 & 3 & 2 \\ 3 & -6 & -5 \end{bmatrix}$$

and

$$A - 2I = \begin{bmatrix} 3 & -6 & -6 \\ -1 & 2 & 2 \\ 3 & -6 & -6 \end{bmatrix}.$$

We know that $A - I$ is singular and obviously rank$(A - I) \geq 2$ (by Rank-Nullity theorem). Therefore, rank$(A - I) = 2$. It is evident that rank$(A - 2I) = 1$. So, the nullity of the matrices are 1 and 2 respectively which sum up to 3. Hence, A is diagonalizable and the corresponding diagonal matrix is

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

Lemma 3.9. Suppose that $T(v) = cv$. If f is any polynomial, then $f(T)(v) = f(c)v$.

Proof. The proof of the lemma is based on the fact that

$$T^2(v) = T(T(v)) = T(cv) = cT(v) = c^2T(v).$$

We can prove by the principle of mathematical induction that

$$T^n(v) = c^n v.$$

Hence $f(T)(v) = f(c)v$, for any polynomial in T. □

Lemma 3.10. Let T be a linear operator on the finite-dimensional space V. Let $c_1, c_2, \ldots, c_k$ be the distinct characteristic values of T and let W_i be the corresponding eigen spaces. If $W = W_1 + W_2 + \cdots + W_k$, then

$$\dim W = \dim W_1 + \dim W_2 + \cdots + \dim W_k.$$

In fact, if $\mathcal{B}_i$ is an ordered basis of W_i, then $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ is an ordered basis for W.

Proof. The space $W = W_1 + W_2 + \cdots + W_k$ is the subspace spanned by all of the eigen vectors of T. Usually when one forms the sum W of subspaces W_i, one expects that $\dim W < \dim W_1 + \dim W_2 + \cdots + \dim W_k$ because of linear relations which may exist between vectors in the various spaces. This lemma states that the characteristic spaces associated with different characteristic values are independent of one another.

Suppose that (for each i) we have a vector b_i in W_i, and assume that

$$b_1 + b_2 + \cdots + b_k = 0.$$

We shall show that $b_i = 0$ for each i. Let f be any polynomial. Since $T(b_i) = c_i b_i$, the preceding lemma tells us that

$$\begin{aligned} 0 &= f(T)(0) \\ &= f(T)(b_1) + f(T)b_2 + \cdots + f(T)b_k \\ &= f(c_1)b_1 + f(c_2)b_2 + \cdots + f(c_k)b_k. \end{aligned}$$

Choose the polynomials $f_1, f_2, \ldots, f_k$ such that

$$\begin{aligned} f_i(c_j) = \delta_{ij} &= 1, \ i = j \\ &= 0, \ i \neq j. \end{aligned}$$

Then

$$0 = f_i(T)(0) = \sum_j \delta_{ij} b_j = b_i.$$

Now, let $\mathcal{B}_i$ be an ordered basis for W_i, and let $\mathcal{B}$ be the sequence $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$. Then $\mathcal{B}$ spans the subspace $W = W_1 + W_2 + \cdots + W_k$. Also, $\mathcal{B}$ is a linearly independent sequence of vectors, for the following reason. Any linear relation between the vectors in $\mathcal{B}$ will have the form $b_1 + b_2 + \cdots + b_k = 0$, where b_i is some linear combination of the vectors in $\mathcal{B}_i$. From what we just did, we know that $b_i = 0$ for each i. Since each $\mathcal{B}_i$ is linearly independent, we see that we have only the trivial linear relation between the vectors in $\mathcal{B}$. □

In the course of proving the above lemma, we have proved the following theorem.

Theorem 3.11. Eigen vectors corresponding to distinct eigen values are linearly independent.

Can you prove the theorem independently?

Thus, we arrive at the following theorem.

Theorem 3.12. Let T be a linear operator on a finite-dimensional space V. Let $c_1, c_2, \ldots, c_k$ be distinct eigen values of T and let W_i be the eigen space of c_i. Then the following are equivalent:

1. T is diagonalizable.
2. The characteristic polynomial for T is

   $$f(x) = (x - c_1)^{d_1} \ldots (x - c_k)^{d_k},$$

   where $\dim W_i = d_i$, $i = 1(1)k$.
3. $\dim W_1 + \dim W_2 + \cdots + \dim W_k = \dim V$.

Proof. We have observed that 1 implies 2. If the characteristic polynomial f is the product of linear factors, as in 2, then $d_1 + d_2 + \cdots + d_k = \dim V$. For, the sum of the $d_i's$ is the degree of the characteristic polynomial, and that degree is $\dim V$. Thus, 2 implies 3. Now suppose that 3 holds. Then by the previous lemma, we must have $V = W_1 + W_2 + \cdots + W_k$, that is, the eien vectors of T span V. □

Let us summerize whatever we have learnt so far.

Let T be a linear operator on an n-dimensional vector space V. If T has n distinct eigen values then it has n linearly independent eigen vectors which form a basis of V and in that case, T is diagonalizable. If it has

less number of eigen values, then we have to check that whether they fulfil the deficiency by having multiple eigen vector for a single eigen value so that the number of linearly independent eigen vectors are still n. In either case, we need to check whether the given operator has n linearly independent eigen vectors or not. We can also say that T is diagonalizable if and only if the geometric multiplicity and algebraic multiplicity for a given eigen value coincides.

Exercise 3.13. 1. Check whether the following matrices are diagonalizable. If yes, then find its diagonal form.

i.
$$\begin{bmatrix} -9 & 4 & 4 \\ -8 & 3 & 4 \\ -16 & 8 & 7 \end{bmatrix}$$

ii.
$$\begin{bmatrix} 6 & -3 & -2 \\ 4 & -1 & -2 \\ 10 & -5 & -3 \end{bmatrix}$$

2. Let T be a linear operator on the n-dimensional vector space V, and suppose that T has n distinct eigen values. Prove that T is diagonalizable.

3. Let V be the vector space of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ and let T be the linear operator on V defined as
$$T(f(x)) = \int_0^x f(t)dt.$$
Prove that T has no eigen values.

4. Let $\mathbb{P}_2$ denote the vector space of all polynomials of degree 2 or less, and let $T : \mathbb{P}_2 \to \mathbb{P}_2$ be a linear operator defined by
$$T(ax^2 + bx + c) = 2ax + b.$$
Check whether T is diagonalizable. If so, find the diagonal matrix.

5. Consider the matrix
$$A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix},$$
where a and b are real numbers and $b \neq 0$. Fince all eigen values of A and determine the corresponding eigen spaces. Hence check whether A is diagonalizable.

6. Check whether the given matrix is diagonalizable. If yes, find the diagonalized matrix.
$$A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

3.1.1 Minimal Polynomials and Diagonalizability

We have seen in the previous units that minimal polynomials and characteristic polynomials of a matrix (or, linear operator) has same roots.

So, if T is a diagonalizable linear operator and $c_1, c_2, \ldots c_k$ are the distinct eigen values of T. Then it is easy to see that the minimal polynomial for T is the polynomial

$$m(x) = (x - c_1)(x - c_2) \ldots (x - c_k).$$

If v is an eigen vector, then one of the operators $T - c_1 I, \ldots, T - c_k I$ sends v into 0. Hence

$$(T - c_1 I) \ldots (T - c_k I)(v) = 0,$$

for every eigen vector v. There is a basis for the underlying space which consists of eigen vectors of T; hence

$$m(T) = (T - c_1 I) \ldots (T - c_k I) = 0.$$

What we have concluded is this. If T is a diagonalizable linear operator, then the minimal polynomial for T is a product of distinct linear factors. As we shall soon see, that property characterizes diagonalizable operators.

Theorem 3.14. Let V be a finite dimensional vector space over the field F and let T be a linear operator on V. Then T is diagonalizable if and only if the minimal polynomial of T is the product of distinct linear factors, that is, of the form

$$m(x) = (x - c_1)(x - c_2) \ldots (x - c_k),$$

where, $c_1, c_2, \ldots, c_k \in F$ are distinct.

Proof. We have noted earlier that, if T is diagonalizable, its minimal polynomial is a product of distinct linear factors. To prove the converse, let W be the subspace spanned by all of the eigen vectors of T, and suppose that $W \neq V$. By a previous lemma, there is a vector v not in W and an eigen value c_j of T such that the vector

$$b = (T - c_j I)(v)$$

lies in W. Since $b \in W$,

$$b = b_1 + b_2 + \cdots + b_k$$

where $T(b_i) = c_i b_i$, $1 \leq i \leq k$, and therefore the vector

$$h(T)(b) = h(c_1)(b_1) + \cdots + h(c_k)(b_k)$$

is in W, for every polynomial h. Now,

$$m(x) = (x - c_j)q(x),$$

for some polynomial q. Also,

$$q - q(c_j) = (x - c_j)h.$$

But we have

$$q(T)(v) - q(c_j)(v) = h(T)(T - c_j I)(v) = h(T)(b).$$

But, $h(T)(b) \in W$ and since

$$0 = m(T)(v) = (T - c_j I)q(T)(v)$$

the vector $q(T)(v)$ is in W. Hence $q(c_j)(v)$ is in W. Since v is not in W, we have $q(c_j) = 0$. This contradicts the fact that m has distinct roots. Hence the theorem. □

Example 3.15. Let A be a 4×4 matrix

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

The powers of A are easy to compute

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

$$A^3 = \begin{bmatrix} 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \\ 4 & 0 & 4 & 0 \end{bmatrix}$$

Thus, $A^3 = 4A$, that is, $f(x) = x^3 - 4x = x(x+2)(x-2)$, then $m(A) = 0$. The minimal polynomial of A must divide f. Minimal polynomial is not of degree 1 since in that case, A would have been a scalar multiple of I, which is not true. Hence the candidates of minimal polynomial polynomial are f, $x(x+2)$, $x(x-2)$, $x^2 - 4$. The three quadratic polynomials can be eliminated since at a glance, we can see that $A^2 \neq 2A$, $A^2 \neq -2A$, and $A^2 \neq 4I$. Hence f is the minimal polynomial for A and since f is the product of distinct linear factors, so A is diagonalizable. Now, we can clearly see that the rank of A is 2 and hence its nullity is also $4-2=2$, which means that the eigen space of $A-0I$ has dimension 2 and thus its algebraic multiplicity will be 2. Thus, the characteristic polynomial is $x^2(x^2-4)$. And the matrix A is similar to the diagonal form

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix}.$$

Exercise 3.16.

1. Every matrix A such that $A^2 = A$ is similar to a diagonal matrix.
2. Using diagonalizability, compute A^n, $n \in \mathbb{N}$ for
$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}.$$
3. Is every diagonalizable matrix invertible? Justify.
4. Let A be an $n \times n$ diagonalizable matrix whose characteristic polynomial is given by
$$f(x) = x^3(x-1)^2(x-2)^5(x+2)^4.$$
   i. Find the size of the matrix A.
   ii. Find the minimal polynomial of A.
   iii. Find the dimension of the eigen space for the eigen value 2.
   iv. Find the rank of the matrix.

3.2 Smith's Normal Form

The Smith normal form is a normal form that can be defined for any matrix (not necessarily square) with entries in a principal ideal domain (PID). The Smith normal form of a matrix is diagonal, and can be obtained from the original matrix by multiplying on the left and right by invertible square matrices. In particular, the integers are a PID, so one can always calculate the Smith normal form of an integer matrix. We will talk particularly about the PID $\mathcal{Z}$.

Definition 3.17. Let A be an $m \times n$ matrix over $\mathbb{Z}$. We say that A is in Smith Normal form if there are non-zero $a_1, a_2, \ldots, a_k \in \mathbb{Z}$ such that a_i divides a_{i+1} for $i < k$ such that

$$A = \begin{bmatrix} a_1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & a_2 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & \cdots & a_k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Theorem 3.18. If A is a matrix with entries in $\mathbb{Z}$, then there are invertible matrices P and Q such that PAQ is in Smith normal form.

Theorem 3.19. Every matrix over $\mathbb{Z}$ has Smith Normal form.

In order to find the Smith Normal form of a matrix, we are allowed to use the following operations

1. interchange two rows and columns,
2. multiply a row or column by ± 1(which are the invertible elements in $\mathbb{Z}$)
3. add an integer multiple of a row (or column) to another row (or column)

Exercise 3.20. Obtain the Smith normal form and rank for

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

over $\mathbb{Z}$.

3.3 Few Probable Questions

1. Show that the eigen vectors corresponding to distinct eigen values are linearly independent.
2. State a necessary and sufficient condition for diagonalizability. Check the diagonalizability of the following matrix

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

3. Let f be the characteristic polynomial of a matrix A over the field $\mathbb{R}$ as

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

Also, let A be diagonalizable. Then

(a) Find the minimal polynomial of A.

(b) Find the eigen values along with their algebraic and geometric multiplicities.

(c) Find the diagonalized form of A.

4. Let A be a matrix over the field $\mathbb{R}$ whose minimal polynomial is of the form

$$f(x) = (x^2-1)(x^2+1).$$

(a) Is A diagonalizable over $\mathbb{R}$? Justify.

(b) Is A diagonalizable over the field $\mathbb{C}$? Justify.

Find the eigen values in each case.

5. Let P be a linear operator over $\mathbb{R}^2$ defined as

$$P((x,y)) = (x,0).$$

Show that P is linear. Find the matrix representation of P with respect to the standard basis of $\mathbb{R}^2$. What is the minimal polynomial of P? Is P diagonalizable?

Unit 4

Course Structure

- Primary Decomposition theorem
- Jordan Canonical forms

---

4 Introduction

There are certain subspaces which remain invariant under a linear operator, that is, the linear operator sends each element of the subspace to itself. Such subspaces are of primary importance as we shall see that we can analyse many properties of the linear operator by finding out the various invariant subspaces of the operator. Also, we have seen in the preceding unit that we want to write the matrix of a linear operator in its simplest possible form, which is possible since the matrix representation of a single linear operator under various bases are similar. And we have also seen that the diagonal matrix is the simplest possible matrix to work with. We are always in search of a basis of the underlying vector space for which the corresponding matrix of the linear operator is diagonal. If such a basis exists, then we are happy and the operator is said to be diagonalizable. We have seen various circumstances under which an operator is diagonalizable. We are okay with them. But, what happens if a given operator is not diagonalizable. Can't we express the operator in a simpler form then? That is where the other canonical forms come into play. We can certainly express the operators in a simpler form, which is "almost" a diagonal matrix. One of them is the Jordan Canonical forms, which we shall come through in this unit.

Objectives

After reading this unit, you will be able to

- define the invariant subspaces and see certain examples
- learn about the independent subspaces of a vector space
- learn about the direct-sum decomposition of a vector space into independent subspaces of it
- learn about the invariant direct sum decomposition of a vector space
- define the cyclic vectors of a vector space
- define the smallest invariant subspace containing a vector
- learn about the Jordan forms and find those for any given matrix or linear operator

4.1 Invariant Subspaces

Definition 4.1. Let V be a vector space and T, a linear operator on V. If W is a subspace of V, we say that W is invariant under T if for each $w \in W$, the vector $T(w)$ is also in W.

Example 4.2. If T is any linear operator on V, then V is invariant under T as is the zero subspace. The range of T and the null space of T are also invariant under T.

Example 4.3. Let F be a field and D be the differentiation operator on the space $F[x]$ of polynomials over F. Let n be a positive integer and W be a subspace of polynomials of degree not greater than n. Then W is invariant under T.

Example 4.4. Let T be the linear operator on $\mathbb{R}^2$ which is represented in the standard basis by the matrix

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

Then the only subspaces of $\mathbb{R}^2$ which are invariant under T are $\mathbb{R}^2$ and the zero subspace. Any other invariant subspace would necessarily have dimension 1. But, if W is the subspace spanned by some non-zero vector v, the fact that W is invariant under T means that v is an eigen vector, but A has no eigen value.

When the subspace W is invariant under the operator T, then T induces a linear operator T_W on the space W. The linear operator T_W is defined by $T_W(v) = T(v)$, for $v \in W$. Now we turn to an investigation of the simplest possible nontrivial invariant subspaces : invariant subspaces with dimension 1. How does an operator behave on an invariant subspace of dimension 1? Subspaces of a vector space V of dimension 1 are easy to describe. Take any non-zero vector $u \in V$ and let U equals the set of all scalar multiples of u, that is

$$U = \{au : \ a \in F\}.$$

where, F is the underlying field. The U is a one-dimensional subspace of V, and every one-dimensional subspace of V is of this form. If $u \in V$ and the subspace defined as above is invariant under T, then $T(u)$ must be in U, which means that there must exist a scalar $c \in F$ such that $T(u) = cu \in U$. Conversely, if u is a non-zero vector in V such that $T(u) = cu$ for some scalar c, then the subspace U defined above is a one-dimensional subspace of V invariant under T. The equation $T(u) = cu$ is same as $(T - cI)u = 0$, so that c is an eigen value and u is an eigen vector of T. Thus, we can see that the one dimensional invariant subspace of an operator T is precisely the eigen space of the operator. But the converse is not true always, that is, any eigen space of T need not be one-dimensional though it is invariant under T (can you think of such an example?).

When V is finite-dimensional, the invariance of a subspace W under the linear operator T has a simple matrix interpretation. Suppose we choose an ordered basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ be an ordered basis of V and $\mathcal{B}' = \{v_1, \ldots, v_r\}$ of $W(r = \dim W)$. Let $A = [T]_{\mathcal{B}}$ so that

$$T(v_j) = \sum_{i=1}^{n} A_{ij} v_i.$$

Since W is invariant under T, the vector $T(v_j)$ belongs to W for $j \leq r$. This means that

$$T(v_j) = \sum_{i=1}^{r} A_{ij} v_i, \quad j \leq r.$$

In other words, $A_{ij} = 0$ if $j \leq r$ and $i > r$. Schematically A has the block form

$$A = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$$

where B is an $r \times r$ matrix, C is an $r \times (n - r)$ matrix, and D is an $(n - r) \times (n - r)$ matrix.

4.1.1 Direct-Sum Decompositions

Definition 4.5. The subspaces $W_1, W_2, \ldots, W_k$ of a vector space V are said to be independent if

$$w_1 + w_2 + \cdots + w_k = 0, \quad w_i \in W_i$$

implies that each w_i is zero.

For $k = 2$, we can say that independence means that $W_1 \cap W_2 = \{0\}$. If $k > 2$, it says that each W_j intersects the sum of the other subspaces only at the zero vector.

The independence can be understood as this: If $W = W_1 + W_2 + \cdots + W_k$ be the subspace spanned by $W_1, W_2, \ldots, W_k$, then each vector $w \in W$ can be uniquely expressed as the sum of the vectors in W_j, that is,

$$w = w_1 + w_2 + \cdots + w_k, \quad w_i \in W_i$$

If w has another representation as

$$w = u_1 + u_2 + \cdots + u_k, \quad u_i \in W_i$$

then subtracting, we get

$$0 = (w_1 - u_1) + \cdots + (w_k - u_k), \quad w_k - u_k = 0$$

and the definition of independence implies that $w_j - u_j = 0$ for $1 \leq j \leq k$. Thus, when $W_1, W_2, \ldots, W_k$ are independent, we can operate with the vectors in W as k-tuples.

Lemma 4.6. Let V be a finite-dimensional vector space and let $W_1, W_2, \ldots, W_k$ be subspaces of V and let $W = W_1 + W_2 + \cdots + W_k$. Then the following are equivalent

1. $W_1, W_2, \ldots, W_k$ are independent.
2. For each j, $2 \leq j \leq k$, we have
$$W_j \cap (W_1 + \cdots + W_{j-1}) = \{0\}.$$
3. If $\mathcal{B}_i$ is an ordered basis for W_i, for each i, then the sequence $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k\}$ is an ordered basis for W.

If the above conditions hold, we say that the sum $W = W_1 + W_2 + \cdots + W_k$ is direct or that W is the direct sum of $W_1, W_2, \ldots, W_k$ and we write it as

$$W = W_1 \oplus W_2 \oplus \cdots \oplus W_k.$$

Example 4.7. Let V be a finite-dimensional vector space over the field F and let $\{v_1, v_2, \ldots, l_n\}$ be a basis for V. If W_i be the one-dimensional subspace spanned by v_i, then

$$V = W_1 \oplus W_2 \oplus \cdots \oplus W_n.$$

Example 4.8. Let T be any linear operator on a finite-dimensional space V. Let $c_1, c_2, \ldots, c_k$ be the distinct eigen values of T, and let W_i be the space of eigen vectors associated with the eigen value c_i. Then $W_1, W_2, \ldots, W_k$. And if T is diagonalizable, then $V = W_1 \oplus W_2 \oplus \cdots \oplus W_n$.

Definition 4.9. If V is a vector space, a projection of V is a linear operator E on V such that $E^2 = E$.

Suppose E is a projection. Let R be the range of E and let N be the null space of E. We establish that $V = R \oplus N$. Because $w \in R$ if and only if $w = E(w)$, since $w = E(v)$ implies $E(w) = E(E(v)) = E^2(v) = E(v) = w$. Conversely, if $w = E(w)$, the obviously $w \in R$. The unique representation of v as the sum of vectors in R and N is $v = E(v) + (v - E(v))$.

If R and N are subspaces of V such that $V = R \oplus N$, there is a unique projection operator E which has range R and null space N. The operator is called the projection on R along N.

Projections are clearly diagonalizable since for any projection E, we always have $E^2 = E$ and since the minimal polynomial divides any annihilating polynomial of an operator, so the minimal polynomial can be either $x = 0$, or $x - 1 = 0$ or $x(x - 1) = 0$ which is the product of distinct linear factors in all the cases.

Projections can be used to describe direct-sum decompositions of the space V.

Theorem 4.10. Let $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$, then there exist k linear operators $E_1, E_2, \ldots, E_k$ on V such that

1. each E_i is a projection,
2. $E_i E_j = 0$, if $i \neq j$,
3. $I = E_1 + E_2 + \cdots + E_k$,
4. the range of E_i is W_i

Conversely, if $E_1, E_2, \ldots, E_k$ are k linear operators on V satisfying conditions 1-3, and if W_i is the range of E_i, then $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$

Proof. Suppose $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$. Then for each j, we define an operator E_j on V. Let $v \in V$ and let $v = v_1 + v_2 + \cdots + c_k$ with $v_i \in W_i$. Then we define E_j as $E_j(v) = v_j$. Then E_j is well-defined and it is easy to check that it is linear and that, the range of E_j is W_j and that $E_j^2 = E_j$. The null space of E_j is the subspace

$$W_1 + W_2 + \cdots + W_{j-1} + W_{j+1} + \cdots + W_k$$

for, the statement that $E_j(v) = 0$ simply means $v_j = 0$, that is, v is actually a sum of vectors from the spaces W_i, with $i \neq j$. In terms of the projections E_j, we have

$$v = E_1(v) + \cdots + E_k(v)$$

for each $v \in V$. So, the identity operator on V can be written as

$$I = E_1 + E_2 + \cdots + E_k.$$

Also, if $i \neq j$, then we see that $E_i E_j = 0$ since the range of E_j is the subspace W_j which lies in the null space of E_j.

Conversely, suppose $E_1, E_2, \ldots, E_k$ are k linear operators on V satisfying conditions 1-4. Then certainly we must have

$$V = W_1 + W_2 + \cdots + W_k.$$

since by condition 3, we have

$$v = E_1(v) + \cdots + E_k(v)$$

for every $v \in V$, and $E_i(v) \in W_i$. This expression for v is unique, because if

$$v = v_1 + \cdots v_k, \quad v_i \in W_i,$$

say $v_i = E_i(w_i)$, then using 1 and 2, we have

$$E_j(v) = \sum_{i=1}^{k} E_j v_i = \sum_{i=1}^{k} E_j E_i w_i = E_j^2(w_j) = E_j(w_j) = v_j.$$

This shows that V is the direct sum of the W_i. □

4.1.2 Invariant Direct Sums

We are primarily interested in direct-sum decompositions of V where each subspace if invariant under some linear operator T. Given such a decomposition of V, T induces a linear operator T_i on each W_i by restriction. Thus, if $v \in V$, then we have the unique representation

$$v = v_1 + \cdots + v_k, \quad v_i \in W_i$$

where, each W_i is an invariant subspace of V into which V decomposes. Then

$$T(v) = T_1(v_1) + \cdots + T_k(v_k)$$

We can say that T is the direct-sum of the operators $T_1, \cdots, T_k$. The fact that $V = W_1 \oplus \cdots \oplus W_k$, enables us to associate a unique k-tuple for each $v \in V$ (which is $(v_1, \ldots, v_k)$), in such a way that we can carry out the linear operations in V by working in the individual subspaces W_i. The fact, that each W_i is invariant under T enables us to view T as independent action of T_i on the subspaces W_i.

The above situation can be interpreted in terms of matrices. Suppose we select an ordered basis $\mathcal{B}_i$ of W_i and let $\mathcal{B}$ be the ordered basis for V consisting of the union of the $\mathcal{B}_i$, arranged in the order $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k$. Let $A = [T]_{\mathcal{B}}$ and let $A_i = [T]_{\mathcal{B}_\rangle}$, then A has the block form

$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & A_k \end{bmatrix}$$

Each A_i is a $d_i \times d_i$ matrix, where $d_i = \dim W_i$, and 0's are symbols for rectangular blocks of scalars 0's of various sizes.

Theorem 4.11. Let T be a linear operator on the space V, and let $W_1, \ldots, W_k$ and $E_1, \ldots, E_k$ be the projections as in the previous theorem. Then a necessary and sufficient condition that each subspace W_i be invariant under T is that T commute with each of the projections E_i, that is

$$TE_i = E_i T, \quad i = 1(1)k.$$

We shall now describe a diagonalizable operator T in the language of invariant direct sum decompositions (projections which commute with T). This will be a great help to us in understanding some deeper decomposition theorems later.

Theorem 4.12. Let T be a linear operator on a finite-dimensional space V. If T is diagonalizable and $c_1, \ldots, c_k$ are the distinct eigen values of T, then there exist linear operators $E_1, \ldots, E_k$ on V such that

1. $T = c_1E_1 + \cdots + c_kE_k$;
2. $I = E_1 + \cdots + E_k$;
3. $E_iE_j = 0$, $i \neq j$;
4. $E_i^2 = E_i$;
5. the range of E_i is the eigen space for T associated with c_i.

Conversely, if there exist k distinct scalars $c_1, \ldots, c_k$ and k non-zero linear operators $E_1, \ldots, E_k$ satisfying conditions 1-3, then T is diagonalizable and conditions 4 and 5 are also satisfied.

Proof. Suppose that T is diagonalizable, with distinct eigen values $c_1, \ldots, c_k$. Let W_i be the eigen spaces of V. We know that,

$$V = W_1 \oplus \cdots \oplus W_k$$

Let $E_1, \ldots, E_k$ be the projections associated with this decomposition, as we have done before. Then 2-5 are satisfied. To verify 1, let $v \in V$ and we have

$$v = E_1(v) + \cdots + E_k(v)$$

So,

$$T(v) = TE_1(v) + \cdots + TE_k(v) = c_1E_1(v) + \cdots + c_kE_k(v).$$

Thus,

$$T = c_1E_1 + \cdots + c_kE_k.$$

Now suppose that we are given a linear operator T along with distinct scalars c_i and non-zero operators E_i which satisfy 1-3. Since $E_iE_j = 0$, for $i \neq j$, we multiply both sides of $I = E_1 + \cdots + E_k$ by E_i, and obtain immediately $E_i^2 = E_i$. Multiplying $T = c_1E_1 + \cdots + c_kE_k$ by E_i, we get $TE_i = c_iE_i$, which shows that any vector in the range of E_i, is in the null space of $T - c_iI$. Since we have assumed that $E_i \neq 0$, this proves that there is a non-zero vector in the null space of $T - c_iI$, that is, c_i is an eigen value of T. Furthermore, c_i are all of the eigen values of T; for if c is any scalar, then

$$T - cI = (c_1 - c)E_i + \cdots + (c_k - c)E_k$$

so that, if $(T - cI)(v) = 0$, we must have $(c_i - c)E_i(v) = 0$. If v is not the zero vector, then $E_i(v) \neq 0$ for some i, so that for this i, we have $c_i - c = 0$.

Certainly T is diagonalizable, since we have shown that every non-zero vector in the range of E_i is an eigen vector of T, and the fact that $I = E_1 + \cdots + E_k$ shows that these characteristic vectors span V. All that remains to be demonstrated is that the null space of $T - c_iI$ is exactly the range of E_i. But this is clear since if $T(v) = c_iv$, then

$$\sum_{j=1}^{k}(c_j - c_i)E_j(v) = 0, \quad \text{for each } j$$

and then

$$E_j(v) = 0, \quad i \neq i.$$

Since $v = E_1(v) + \cdots + E_k(v)$, and $E_j(v) = 0$ for $j \neq i$, we have $v = E_j(v)$, which proves that v is in the range of E_i. □

4.1.3 Primary Decomposition Theorem

We studying a linear operator T on the finite-dimensional space V, by decomposing it into a direct sum of operators which are in some sense elementary. We can do this through the eigen values and vectors of T in certain special cases, i.e.,when T is diagonalizable, or, when the minimal polynomial for T factors over the scalar field F into a product of distinct monic polynomials of degree 1 . What can we do with the general T? While studying T using eigen values, we are confronted with two problems. First, T may not have a single eigen value ; this is really a deficiency in the scalar field, namely, that it is not algebraically closed, and we have nothing to do in that case. Second, even if the characteristic polynomial factors completely over F into a product of polynomials of degree 1, there may not be enough eigen vectors for T to span the space V; this is clearly a deficiency in T. The second situation is illustrated by the operator T on F^3, where F is any field represented in the standard basis by

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

The characteristic polynomial for A is $(x-2)^2(x+1)$ and this is also the minimal polynomial for A, and thus, for T. Hence, T is not diagonalizable and this happens since the nullity of $T-2I$ is 1. On the other hand, the null space of $T+I$ and $(T-2I)^2$ span V. From here, we get the motivation for our further work. Suppose we are given that

$$m = (x-c_1)^{r_1} \dots (x-c_k)^{r_k}$$

where $c_1, \dots, c_k \in F$, then we will show that V is the direct sum of the null spaces of $(T-c_iI)^{r_i}$, $i = 1(1)k$.

Theorem 4.13. Let T be a linear operator on the finite-dimensional vector space V over the field F. Let m be the minimal polynomial for T as

$$m = m_1^{r_1} \dots m_k^{r_k}$$

where the m_i are distinct irreducible monic polynomials over F and r_i are positive integers. Let W_i be the null space of $m_i(T)^{r_i}$, $i = 1(1)k$. Then

1. $V = W_1 \oplus \dots \oplus W_k$;
2. each W_i in invariant under T;
3. if T_i is the operator induced on W_i by T, then the minimal polynomial for T_i is $m_i^{r_i}$.

Proof. Let

$$f_i = \frac{m}{m_i^{r_i}} = \Pi_{j \neq i} m_j^{r_j}.$$

Since m_i are distinct polynomials, the polynomials f_i are relatively prime which implies that there are polynomials $g_1, \dots, g_k$ such that

$$\sum_{i=1}^{n} f_i g_i = 1.$$

Also, if $i \neq j$, then $f_i f_j$ is divisible by the polynomial m, since $f_i f_j$ contains each $m_l^{r_l}$ as factor. We shall show that the polynomials $h_i = f_i g_i$ such that $h_i(T)$ is the identity on W_i and is zero on the other W_j such that $h_1(T) + \dots + h_k(T) = I$.

Let $E_i = h_i(T) = f_i(T) g_i(T)$. Since $h_1 + \dots + h_k = 1$ and p divides $f_i f_j$ for $i \neq j$, we have

$$E_1 + \dots + E_k = I, \quad E_i E_j = 0, \quad \text{if } i \neq j.$$

Thus, E_i are the projections which correspond to some direct-sum decomposition V. We will show that the range of E_i is exactly W_i. It is clear that each vector in the range of E_i is in W_i, since if $v \in E_i$, then $v = E_i(v)$, and so

$$m_i(T)(v) = m_i(T)^{r_i} E_i(v) = m_i(T)^{r_i} f_i(T) g_i(T)(v) = 0$$

since m divides $m_i^{r_i} f_i g_i$. Conversely, suppose that v is in the null space of $m_i(T)^{r_i}$. If $j \neq i$, then $f_j g_j$ is divisible by $m_i^{r_i}$ and so $f_j(T)g_j(T)(v) = 0$, that us $E_j(v) = 0$ for $j \neq i$. But this is immediate that $E_i(v) = v$, that is v is in the range of E_i. This completes the proof of 1.

Also, it is evident that W_i are invariant under T. If T_i is the operator induced on W_i by T, then obviously $m_i(T)^{r_i} = 0$, because by definition, $m_i(T)^{r_i}$ is zero on W_i. This shows that the minimal polynomial for T_i divides $m_i^{r_i}$. Conversely, let g be any polynomial such that $g(T_i) = 0$. Then $g(T)f_i(T) = 0$. Thus gf_i is divisible by the minimal polynomial of T, that is, $m_i^{r_i}$ divides gf_i. It is easily seen that $m_i^{r_i}$ divides g. Hence the minimal polynomial for T_i is $m_i^{r_i}$. □

Exercise 4.14.

1. Let T be a linear operator on a finite-dimensional vector space V. Let R be the range of T and let N be the null space of T. Prove that R and N are independent if and only if $V = R \oplus N$.

2. Let T be a linear operator on V. Suppose $V = W_1 \oplus \cdots \oplus W_k$, where each W_i is invariant under T. Let T_i be the induced operator on W_i. Then show that the characteristic polynomial f of T is the product of those of T_i.

3. Let T be a linear operator on V which commutes with every projection operator on V. What can you say about T?

4. Let T be a linear operator on the finite-dimensional space V with characteristic polynomial

   $$f = (x - c_1)^{d_1} \ldots (x - c_k)^{d_k}$$

   and minimal polynomial

   $$m = (x - c_1)^{r_1} \ldots (x - c_k)^{r_k}.$$

   Let W_i be the null space of $(T - c_i I)^{r_i}$. Then show that W_i is the set of all vectors $v \in V$ such that $(T - c_i I)^m(v) = 0$ for some positive integer m (which may depend on v).

4.1.4 Cyclic Subspaces and Annihilators

If V is a finite-dimensional vector space over a field F and T is a fixed linear operator on V. If v is any vector in V, there is a smallest subspace of V which is invariant under T and contains v. This subspace can be defined as the intersection of all T-invariant subspaces which contain v. If W is any subspace of V which is invariant under T and contains v, then W must also contain $T(v)$ and hence must contain $T^2(v)$, $T^3(v)$, and so on. In other words, W must contain $g(T)(v)$ for every polynomial g over F. This is clearly the smallest subspace which contains the vector v and invariant under T.

Definition 4.15. If v is any vector in V, the T-cyclic subspace generated by v is the subspace $Z(v;T)$ of all vectors of the form $g(T)(v)$, g in $F[x]$. If $Z(v;T) = V$, then v is called a cyclic vector for T.

In other words, $Z(v;T)$ is the subspace $\{v, T(v), T^2(v), \ldots\}$ and v is a cyclic vector if and only if these vectors span V. Every arbitrary operator need not have cyclic vectors.

Example 4.16. For any operator T, the T-cyclic subspace generated by the zero vector is the zero subspace. The space $Z(v;T)$ is one-dimensional if and only if v is an eigen vector for T. For the identity operator, every non-zero vector generates a one-dimensional cyclic subspace; thus, if $\dim V > 1$, the identity operator has no cyclic vector.

For any operator T and vector v, we are interested in the linear relations

$$c_0 + c_1 T(v) + \cdots + c_k T^k(v) = 0$$

between the vectors $T^i(v)$, or, we shall be interested in the polynomials $g = c_0 + c_1 x + \cdots + c_k x^k$ such that $g(T)(v) = 0$. The set of all g satisfying the property in $F[x]$ is clearly a non-zero ideal since it contains the minimal polynomial m of the operator T.

Definition 4.17. If v is any vector in V, the T-annihilator of v is the ideal $M(v;T)$ in $F[x]$ consisting of all polynomials g over F such that $g(T)(v) = 0$. Then the unique monic polynomial m_v which generates this ideal will also be called the T-annihilator of v.

We note that the degree of m_v should be greater than zero unless v is the zero vector.

Theorem 4.18. Let v be any non-zero vector in V and m_v be the T-annihilator of v. Then

1. the degree of m_v is equal to the dimension of the cyclic subspace $Z(v;T)$;
2. if the degree of m_v is k, then the vectors $v, T(v), T^2(v), \ldots, T^{k-1}(v)$ form a basis for $Z(v;T)$
3. if U is the linear operator on $Z(v;T)$ induced by T, then the minimal polynomial for U is m_v.

If v is a cyclic vector for T, then the minimal polynomial for T must have degree equal to the dimension of the space V ; hence, the Cayley-Hamilton theorem tells us that the minimal polynomial for T is the characteristic polynomial for T.

Our plan is to study the general T by using operators which have a cyclic vector. So, let us take a look at a linear operator U on a space W of dimension k which has a cyclic vector v. By the above theorem, the vectors $v, \ldots, U^{k-1}(v)$ forms a basis for the space W, and the annihilator m_v of v is the minimal polynomial for U (and hence also the characteristic polynomial for U). If we let $v_i = U^{i-1}(v)$, $i = 1(1)k$, then the action of U on the ordered basis $\mathcal{B} = \{v_1, \ldots, v_k\}$ is

$$\begin{aligned} U(v_i) &= v_{i+1}, \quad i = 1(1)k-1 \\ U(v_k) &= -c_0 v_1 - c_1 v_2 - \cdots - c_{k-1} v_k \end{aligned}$$

where, $m_v = c_0 + c_1 x + \cdots + x^k$. The expression for $U(v_k)$ follows from the fact that $m_v(U)(v) = 0$, that is

$$U^k(v) + c_{k+1} U^{k-1}(v) + \cdots + c_1 U(v) + c_0 v = 0.$$

This says that the matrix of U in the ordered basis $\mathcal{B}$ is

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & 0 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -c_{k-1} \end{bmatrix}.$$

The matrix is called the companion matrix of the monic polynomial m_v.

Theorem 4.19. If U is a linear operator on the finite-dimensional space W, then U has a cyclic vector if and only if there is some ordered basis for W in which U is represented by the companion matrix of the minimal polynomial for U.

Proof. If U has a cyclic vector, then there is such an ordered basis for W. Conversely, if we have some ordered basis $\{v_1, \ldots, v_k\}$ for W in which U is represented by the companion matrix of its polynomial, it is obvious that v_1 is a cyclic vector for U. □

Corollary 4.20. If A is the companion matrix of a monic polynomial m, then m is both the minimal and the characteristic polynomial of A.

If T is any linear operator on the space V and v is any vector in V, then the operator U which T induces on the cyclic subspace $Z(v;T)$ has a cyclic vector, namely v. Thus, $Z(v;T)$ has an ordered basis in which U is represented by the companion matrix of m_v, the T-annihilator of v.

Exercise 4.21.

1. Show that $Z(v;T)$ is one dimensional if and only if v is an eigen vector of T.
2. Let T be the linear operator on $\mathbb{R}^3$ which is represented in the standard ordered basis by the matrix
$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$
Prove that T has no cyclic vector. What is the T-cyclic subspace generated by the vector $(1, -1, 3)$?
3. Let V be an n-dimensional vector space, and let T be a linear operator on V. Suppose that T is diagonalizable. If T has a cyclic vector, show that T has n distinct eigen values.

4.2 Jordan Canonical Forms

We have seen that the diagonal matrices are "easiest" matrix to handle. So we are always in search of a basis for which a particular linear operator is diagonalizable. But this is not always possible. So we are in search of the next simplest matrix in which the operator can be represented. And the next "easiest" matrix to deal with are the triangular matrices. So we come to the Jordan canonical forms, or simply the Jordan forms. The Jordan Canonical Form is an upper triangular matrix of a particular form called a Jordan matrix representing a linear operator on a finite-dimensional vector space with respect to some basis. Such a matrix has each non-zero off-diagonal entry equal to 1, immediately above the main diagonal (on the superdiagonal), and with identical diagonal entries to the left and below them. Let us check for ourselves. Let A be a matrix as given

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

The eigen values of A are $1, 2, 4, 4$ and the dimensions of the eigen space corresponding to each eigen values are $1, 1, 1$ which does not sum up to 4, so A is not-diagonalizable. But A is similar to the matrix below

$$J = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}.$$

The matrix J is "almost" diagonal and is called the Jordan form of A.

Definition 4.22. Let A be an $n \times n$ matrix and c be an eigen value of A of algebraic multiplicity, say k. Then the elementary Jordan block of A corresponding to c, of size k is given by

$$\begin{bmatrix} c & 1 & 0 & \cdots & 0 \\ 0 & c & 1 & \cdots & 0 \\ 0 & 0 & c & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & c \end{bmatrix}.$$

Then the parent matrix is composed of the elementary Jordan blocks

$$A = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_k \end{bmatrix}.$$

The Jordan form of a matrix has the following properties:

1. Given an eigen value c_j, the number of elementary Jordan blocks corresponding to c_j is equal to the geometric multiplicity of c_j.
2. The sum of the sizes of the Jordan blocks corresponding to an eigen value c_j is equal to its algebraic multiplicity.
3. The maximum size of a Jordan block corresponding to an eigen value c_j is equal to its multiplicity in the minimal polynomial of the parent matrix and there has to be a Jordan block with the maximum size for c_j.

Illustration 4.23. 1. Let us be given a matrix

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

First of all, we calculate the eigen values of A which are 5 and 3. Then find the rank of the matrices $A - 5I$ and $A - 3I$ which happen to be 2 and 1 respectively and hence the nullity of the corresponding matrices are 1 and 2 respectively summing up to 3, the dimension of $\mathbb{R}^3$. Hence the minimal polynomial of A is $(x-3)(x-5)$ and the Jordan form for A is

$$J = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Here there are precisely three Jordan blocks, [5], [3], [3].

2.

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

Then A has only one eigen value, which is 1 and the rank of $A - I$ is 1, which means that it has nullity equal to 2 which does not sum up to 3. Since the nullity, that is the geometric multiplicity of 1 is 2, so

there will be two Jordan blocks for 1 and also the maximum size of the Jordan block should be 2. Thus, the Jordan form for A is

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

Exercise 4.24. 1. Put the matrix

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

into Jordan form.

2. Let A be a 5×5 matrix with characteristic polynomial $f(x) = (x-2)^3(x+7)^2$ and minimal polynomial $m = (x-2)^2(x+7)$. What is the Jordan form for A?

3. How many possible ,Jordan forms are there for a 6×6 complex matrix with characteristic polynomial $(x+2)^4(x-1)^2$?

4. The differentiation operator on the space of polynomials of degree less than or equal to 3 is represented in the 'natural' ordered basis by the matrix

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

What is the Jordan form of this matrix?

4.3 Few Probable Questions

1. Show that for a direct-sum decomposition of a finite-dimensional vector space V, $V = W_1 \oplus W_2 \oplus \cdots \oplus W_k$, there exists k projection operators E_i such that the range of each E_i is W_i and $I = E_1 + \cdots + E_k$.

2. State and prove the primary decomposition theorem.

3. Find the Jordan form of the matrix

$$A = \begin{bmatrix} 5 & 4 & 2 \\ 4 & 5 & 2 \\ 2 & 2 & 2 \end{bmatrix}.$$

Show detailed steps.

Unit 5

Course Structure

- Invariant factors and elementary divisors
- Rational forms

---

5 Introduction

The primary purpose of this section is to prove that if T is any linear operator on a finite-dimensional space V, then there exist vectors $v_1, \ldots, v_k$ in V such that

$$V = Z(v_1; T) \oplus \cdots \oplus Z(v_k; T).$$

This will show that T is the direct sum of a finite number of linear operators, each of which has a cyclic vector. The cyclic decomposition theorem is closely related to the following question. Which T-invariant subspaces W have the property that there exists a T-invariant subspace W' such that $V = W \oplus W'$? In fact, there are many subspaces W' for which $V = W \oplus W'$ but we can't say whether they are invariant or not. This unit is dedicated to the study of the invariant factors and elementary divisors of a linear operator and certain canonical forms of it.

Objectives

After reading this unit, you will be able to

- define T-admissible subspaces of a vector space
- learn the cyclic decomposition theorem for a finite-dimensional vector space with respect to a linear operator T
- learn the generalized Cayley-Hamilton theorem for a linear operator on a finite-dimensional vector space
- define the invariant factors of a matrix
- learn to find the rational canonical form for a matrix

5.1 Invariant Factors

Definition 5.1. Let T be a linear operator on a vector space V. A subspace W of V is said to be T -admissible if

1. W is T-invariant;
2. if $f(T)(v)$ is in W, there exists a vector w in W such that $f(T)(v) = f(T)(w)$.

Note that, from the discussion we had done in the introduction of this unit, if V is decomposed as $V = W \oplus W'$, where both W and W' are invariant, then any vector $v \in V$ has a unique representation $v = w + w'$, where $w \in W$ and $w' \in W'$. If f is any polynomial over the scalar field, then $f(T)(v) = f(T)(w) + f(T)(w')$. Since W and W' are T-invariant, the vectors $f(T)(w)$ and $f(T)(w')$ lies in W and W' respectively. Thus, $f(T)(v)$ is in W if and only if $f(T)(w') = 0$. Hence, we can say that for such a case, W is admissible.

Let W be a proper T-invariant subspace. Let us try to find a non-zero vector v such that

$$W \cap Z(v; T) = \{0\}.$$

We can choose a vector w' which is not in W. Consider the T-conductor $S(w'; W)$, which consists of all polynomials g such that $g(T)(w')$ is in W. Recall that the monic polynomial f which generates the ideal $S(w'; W)$ is also called the T-conductor of w' into W. The vector $f(T)(w')$ is in W. Now, if W is T-admissible, there is a w'' in W with $f(T)(w') = f(T)(w'')$. Let $w = w' - w''$ and let g be any polynomial. Since $w' - w$ is in W, $g(T)(w')$ will be in W if and only if $g(T)(w)$ is in W; in other words, $S(w; W) = S(w'; W)$. Thus, the polynomial f is also the T-conductor of w into W. But $f(T)(w) = 0$ which tells us that $g(T)(w)$ is in W if and only if $g(T)(w) = 0$, that is, the subspaces $Z(v; T)$ and W are independent and f is the T-annihilator of v.

Theorem 5.2. (Cyclic Decomposition Theorem) Let T be a linear operator on a finite-dimensional vector space V and let W_0 be a proper T-admissible subspace of V. There exist non-zero vectors $v_1, \ldots, v_k$ in V with respective T-annihilators $m_1, \ldots, m_k$ such that

1. $V = W_0 \oplus Z(v_1; T) \oplus \cdots \oplus Z(v_k; T)$;
2. m_r divides m_{r-1}, $r = 2, \ldots, k$.

Furthermore, the integer k and the annihilators $m_1, \ldots, m_k$ are uniquely determined by 1 and 2 and the fact that no v_r is 0.

The proof is rather lengthy and has been omitted for general good.

Our next corollary gives us the answer to our primary question which we asked at the beginning of this unit regarding the existence of a T-invariant subspace W' which forms a **complementary** for a T-invariant subspace W of V.

Corollary 5.3. If T is a linear operator on a finite-dimensional vector space, every T-admissible subspace has a complementary subspace which is also invariant under T.

Proof. Let W be an admissible subspace of V. If $W = V$, the required complement is $\{0\}$. If W is proper, then we apply the Cyclic decomposition theorem and let

$$W' = Z(v_1; T) \oplus \cdots \oplus Z(v_k; T).$$

Then W' is invariant under T and $V = W \oplus W'$. □

Corollary 5.4. Let T be a linear operator on a finite-dimensional vector space V.

1. There exists a vector v in V such that the T-annihilator of v is the minimal polynomial for T.
2. T has a cyclic vector if and only if the characteristic and minimal polynomials for T are identical.

Proof. If $V = \{0\}$, the results are trivially true. If $V \neq \{0\}$, let

$$V = Z(v_1; T) \oplus \cdots \oplus Z(v_k; T)$$

where the T-annihilators $m_1, \ldots, m_k$ are such that m_{r+1} divides m_r, $1 \leq r \leq k-1$. As we noted in the previous theorem, it follows easily that m_1 is the minimal polynomial for T, that is, the T-conductor of V into $\{0\}$.

We saw in the previous unit that if T has a cyclic vector, the minimal polynomial for T coincides with the characteristic polynomial. Choose any vector v as in 1. If the degree of the minimal polynomial is $\dim V$, then $V = Z(v; T)$. □

Theorem 5.5. (Generalized Cayley-Hamilton Theorem) Let T be a linear operator on a finite-dimensional vector space V. Let m and f be the minimal and characteristic polynomials for T, respectively. Then

1. m divides f;
2. m and f have the same prime factors, except for multiplicities;
3. If $m = f_1^{r_1} \ldots f_k^{r_k}$ is a prime factorization of m, then $f = f_1^{d_1} \ldots f_k^{d_k}$, where d_i is the nullity of $f_i(T)^{r_i}$ divided by the degree of f_i.

Proof. If $V = \{0\}$, then the case is trivial. To prove 1 and 2, consider a cyclic decomposition of V. As in the proof of the above corollary, $m_1 = m$. Let U_i be the restriction of T to $Z(v_i; T)$. Then U_i has a cyclic vector and so m_i is both the minimal as well as characteristic polynomial for U_i. Hence, the characteristic polynomial f is the product $f = m_1 \ldots m_r$. Clearly, $m_1 = m$ divides f and this proves 1. Obviously any prime divisor of m is a prime divisor of f. Conversely, a prime divisor of $f = m_1 \ldots m_r$ must divide one of the factors m_i, which is turn divides m_1.

Let the given factorization in the statement of the theorem be the prime factorization of m. We use the primary decomposition theorem which tells us that, if V is the null space of $f_i(T)^{r_i}$, then

$$V = V_1 \oplus \cdots \oplus V_k$$

and $f_i^{r_i}$ is the minimal polynomial of the operator T_i, obtained by restricting T to the subspace V_i. Apply part 2 of the present theorem to the operator T_i. Since its minimal polynomial is a power of the prime f_i, the characteristic polynomial for T_i has the form $f_i^{d_i}$, where $d_i \geq r_i$. Obviously

$$d_i = \frac{\dim V}{\deg f_i}$$

and (almost by definition) $\dim V_i = \text{nullity} f_i(T)^{r_i}$. Since T is the direct sum of the operators $T_1, \ldots, T_k$, the characteristic polynomial f is the product

$$f = f_1^{d_1} \ldots f_k^{d_k}.$$

□

The polynomials $m_1, \ldots, m_r$ are called the invariant factors for a matrix B.

5.1.1 Rational Forms

Let us try to understand the cyclic-decomposition theorem for matrices. If we have the operator T and the direct-sum decomposition and $\mathcal{B}_i$ be the cyclic ordered basis $\{v_i, T(v_i), \ldots, T^{k_i-1}(v_i)\}$ for $Z(v_i; T)$. Here, k_i denotes the dimension of $Z(v_i; T)$, that is, the degree of the annihilator m_i. The matrix of the induced operator T_i in the ordered basis $\mathcal{B}_i$ is the companion matrix of the polynomial m_i. Thus, if we let $\mathcal{B}$ be the ordered basis for V which is the union of the $\mathcal{B}_i$ arranged in the order $\mathcal{B}_1, \ldots, \mathcal{B}_r$, then the matrix of T in the ordered basis $\mathcal{B}$ will be

$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_r \end{bmatrix}$$

where A_i is the $k_i \times k_i$ companion matrix of m_i. An $n \times n$ matrix A, which is the direct-sum of companion matrices of non-scalar monic polynomials $m_1, \ldots, m_r$ such that m_{i+1} divides m_i for $i = 1, \ldots, r-1$, will be said to be in rational form.

Theorem 5.6. Let F be a field and let B be an $n \times n$ matrix over F. Then B is similar over the field F to unique matrix which is in rational form.

We have seen a simpler form for non-diagonalizable matrices, that is the Jordan form. We have a theorem for triangular matrices which states that

Theorem 5.7. An $n \times n$ is triangulable, that is, similar to a triangular matrix if and only if its minimal polynomial is the product of linear factors (not necessarily distinct).

Now, the Jordan form is a triangular matrix and we know that the triangular matrices are the next "simplest" matrices to deal with, right after diagonal ones and we have also seen with certain examples that the Jordan form was deducible for a matrix when its minimal polynomial, or we can also say that its characteristic polynomial was the product of linear factors. But this is not always the case. For example, consider the matrix over the real field

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

The characteristic polynomial of the above matrix is $f(x) = x^2 + 1$. Since the minimal polynomial of a matrix divides its characteristic polynomial, and since the characteristic polynomial is irreducible, so the minimal polynomial of the matrix is also $m(x) = x^2 + 1$. These are the cases when the rational forms come into play. We will illustrate how we find the rational form for a matrix.

Illustration 5.8. 1. Consider the real matrix

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

Then the characteristic polynomial of the matrix can be calculated and is equal to $f(x) = x^3 + 6x^2 + 12x + 8 = (x+2)^3$. We have, $A + 2I \neq 0$, but $(A + 2I)^2 = 0$. Thus, the minimal polynomial of the matrix is $(x+2)^2$. We know that the largest invariant factor is simply the minimal polynomial. Furthermore, we know that the size of our canonical form matrix must be 3×3, and that our invariant factors must divide the minimal polynomial. Thus, there are two invariant factors $(x+2)^2 = x^2+4x+4$ and $x + 2$. Therefore, the rational canonical form of the matrix is

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

Note that the minimal polynomial of A is the product of linear factors and hence we can find the Jordan form for A. (Find it)

Exercise 5.9. 1. Find the minimal polynomials and the rational form for the following matrices

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

2. Find the rational form of the matrix

$$\begin{bmatrix} 1 & 3 & 3 \\ 3 & 1 & 3 \\ -3 & -3 & -5 \end{bmatrix}.$$

5.2 Few Probable Questions

1. State and prove the Generalized Cayley-Hamilton theorem.
2. Find the minimal polynomial, invariant factors and the rational form of the following matrix

$$\begin{bmatrix} 2 & -2 & 14 \\ 0 & 3 & -7 \\ 0 & 0 & 2 \end{bmatrix}.$$

Unit 6

Course Structure

- Bilinear and Quadratic forms
- Classification of Quadratic forms

---

6 Introduction

A bilinear form on a real vector space V is a function f which assigns a number to each pair of elements of V, a scalar from the underlying field, satisfying certain properties. We can begin with an example of a map from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, where $\mathbb{R}$ is the underlying field, defined by

$$\langle X, Y \rangle = X^T.Y = x_1y_1 + \cdots + x_ny_n.$$

This is the most common dot product, that we are familiar with. The property of the dot product which we will use to generalize to bilinear forms is bilinearity: the dot product is a linear function from V to F, where F is the underlying field, if one of the elements is fixed. Bilinear forms are meant to be a generalization of the dot product on $\mathbb{R}^n$.

Objectives

After reading this unit, you will be able to

- define bilinear forms and see certain examples of it
- learn properties related to them
- define quadratic forms and associated matrices
- define definiteness of a form and its associated matrices
- learn about the equivalent definitions of definiteness of a matrix and form
- solve problems related to the definiteness of matrices

6.1 Bilinear Forms

Definition 6.1. Let V be a vector space over F. We define a bilinear form to be a function $f : V \times V \to F$ such that

$$\begin{aligned} f(v_1 + v_2, w) &= f(v_1, w) + f(v_2, w), \;\; v_1, v_2, w \in V \\ f(v, w_1 + w_2) &= f(v, w_1) + f(v, w_2), \;\; v, w_1, w_2 \in V \\ f(cv, w) &= cf(v, w) = f(v, cw), \;\; v, w \in W, \; c \in F \end{aligned}$$

We will often use the notation $\langle v, w \rangle$ for $f(v, w)$.

The zero function from $V \times V$ into F is clearly a bilinear form. It is also true that any linear combination of bilinear forms on V is again a bilinear form(check it). All this may be summarized by saying that the set of all bilinear forms on V is a subspace of the space of all functions from $V \times V$ into F. We denote the space of bilinear forms on V by $L(V, V, F)$.

Example 6.2. Let V be a vector space over the field F and let L_1 and L_2 be linear functions on V. Define f by

$$f(u, v) = L_1(u)L_2(v).$$

If we fix v and regard f as a function of u, then we simply have a scalar multiple of the functional L_1. And fixing u, f is a scalar multiple of L_2. Hence f is a bilinear form on V.

Example 6.3. Let m and n be positive integers and F a field. Let V be the vector space of $m \times n$ matrices over F. Let A be a fixed $m \times n$ over F. Define

$$f_A(X, Y) = \text{tr}(X^T AY).$$

Then f_A is a bilinear form on V. If X, Y, Z are $m \times n$ matrices over F, then

$$\begin{aligned} f_A(cX + Z, Y) &= \text{tr}[(cX + Z)^T AY] \\ &= \text{tr}(cX^T AY) + \text{tr}(Z^T AY) = cf_A(X, Y) + f_A(Z, Y). \end{aligned}$$

Of course, we have used the fact that the transpose operation and the trace function are linear. It is even easier to show that f_A is linear as a function of its second argument. In the special case, $n = 1$, the matrix $X^T AY$ is 1×1 matrix, that is, a scalar, and the bilinear form is simply

$$f_A(X, Y) = \sum_{i,j} A_{ij} x_i y_j.$$

Example 6.4. Let F be a field. Let us find all bilinear forms on the space F^2. Suppose f is such a bilinear form. If $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are in F^2, then

$$\begin{aligned} f(x, y) &= f(x_1 e_1 + x_2 e_2, y) \\ &= x_1 f(e_1, y) + x_2 f(e_2, y) \\ &= x_1 f(e_1, y_1 e_1 + y_2 e_2) + x_2 f(e_2, y_1 e_1 + y_2 e_2) \\ &= x_1 y_1 f(e_1, e_1) + x_1 y_2 f(e_1, e_2) + x_2 y_1 f(e_2, e_1) + x_2 y_2 f(e_2, e_2). \end{aligned}$$

Hence, f is completely determined by the four scalars $A_{ij} = f(e_i, e_j) = \langle e_i, e_j \rangle$ by

$$\begin{aligned} f(x, y) &= A_{11} x_1 y_1 + A_{12} x_1 y_2 + A_{21} x_2 y_1 + A_{22} x_2 y_2 \\ &= \sum_{i,j} A_{ij} x_i y_j. \end{aligned}$$

Thus, if X and Y are the coordinate matrices of x and y, and if A is the above matrix, then

$$f(x, y) = X^T AY.$$

This can be generalized for any finite-dimensional vector spaces.

Definition 6.5. (**Bilinear forms on** $\mathbb{R}^n$) Every bilinear form on $\mathbb{R}^n$ has the form

$$\langle x, y \rangle = x^T Ay = \sum_{i,j} a_{ij} x_i y_j, \quad x, y \in \mathbb{R}^n$$

for some $n \times n$ matrix A and we also have $a_{ij} = \langle e_i, e_j \rangle$ for all i, j. e_i is the n tuple of real numbers whose ith entry is 1 and all other entries are 0.

Definition 6.6. Let V be a finite-dimensional vector space, and let $\mathcal{B} = \{v_1, \ldots, v_n\}$ be an ordered basis for V. If f is a bilinear form on V, the matrix of f in the ordered basis $\mathcal{B}$ is the $n \times n$ matrix A with entries $A_{ij} = f(v_i, v_j)$. We shall denote this matrix by $[f]_{\mathcal{B}}$.

Theorem 6.7. Let V be a finite-dimensional vector space over the field F. For each ordered basis $\mathcal{B}$ of V, the function which associates with each bilinear form on V, its matrix in the ordered basis $\mathcal{B}$ is an isomorphism of the space $L(V, V, F)$ onto the space of $n \times n$ matrices over the field F.

Proof. We have seen that $f \to [f]_{\mathcal{B}}$ is a one-one correspondence between the set of bilinear forms on V and the set of all $n \times n$ matrices over F. That this is a linear transformation is easy to see, because

$$(cf + g)(v_i, v_j) = cf(v_i, v_j) + g(v_i, v_j)$$

for each i and j. This simply says that

$$[cf + g]_{\mathcal{B}} = c[f]_{\mathcal{B}} + [g]_{\mathcal{B}}.$$

□

Corollary 6.8. If $\mathcal{B} = \{v_1, \ldots, v_n\}$ is an ordered basis for V, and $\mathcal{B}^* = \{L_1, \ldots, L_n\}$ be an ordered basis for V^*, then the n^2 bilinear forms

$$f_{ij}(x, y) = L_i(x)L_j(y), \quad 1 \neq i \neq n, \ 1 \leq j \leq n$$

form a basis for $L(V, V, F)$. In particular, the dimension of $L(V, V, F)$ is n^2.

The concept of the matrix of a bilinear form in an ordered basis is similar to that of the matrix of a lineal' operator in an ordered basis. Just as for linear operators, we shall be interested in what happens to the matrix representing a bilinear form, as we change from one ordered basis to another. So, suppose $\mathcal{B} = \{v_1, \ldots, v_n\}$ and $\mathcal{B}' = \{v_1', \ldots, v_n'\}$ two ordered bases for V and that f is a bilinear form on V. How are the matrices $[f]_{\mathcal{B}}$ and $[f]_{\mathcal{B}'}$ related? Well, let P be the (invertible) $n \times n$ matrix such that

$$[v]_{\mathcal{B}} = P[v]_{\mathcal{B}'}$$

for all $v \in V$. In other words, define P by

$$v_j' = \sum_{i=1}^{n} P_{ij} v_i.$$

For any vectors $v, w \in V$,

$$\begin{aligned} f(v, w) &= [v]_{\mathcal{B}}^T [f]_{\mathcal{B}} [v]_{\mathcal{B}} \\ &= (P[v]_{\mathcal{B}'})^T [f]_{\mathcal{B}} P[w]_{\mathcal{B}'} \\ &= [v]_{\mathcal{B}'}^T (P^T [f]_{\mathcal{B}} P)[w]_{\mathcal{B}'}. \end{aligned}$$

By the definition and uniqueness of the matrix representing f in the ordered basis $\mathcal{B}'$, we must have

$$[f]_{\mathcal{B}'} = P^T [f]_{\mathcal{B}} P.$$

One consequence of the change of basis formula is the following: If A and B are $n \times n$ matrices which represent the same bilinear form on V in (possibly) different ordered bases, then A and B have the same rank. For, if P is an invertible $n \times n$ matrix and $B = P^T AP$, it is evident that A and B have the same rank. This makes

it possible to define the rank of a bilinear form on V as the rank of any matrix which represents the form in an ordered basis for V.

It is desirable to give a more intrinsic definition of the rank of a bilinear form. This can be done as follows : Suppose f is a bilinear form on the vector space V. If we fix a vector v in V, then $f(v, w)$ is linear as a function of w. If we fix a vector $v \in V$, then $f(v, w)$ is linear as a function of w. In this way, each fixed v determines a linear functional on V; let us denote this linear functional by $L_f(v)$. To repeat, if v is a vector in V, then $L_f(v)$ is the linear functional on V whose value on any vector w is $f(v, w)$. This gives us a transformation $v \to L_f(v)$ from V into the dual space V^*. Since

$$f(cv_1 + v_2, w) = cf(v_1, w) + f(v_2, w)$$

we see that

$$L_f(cv_1 + v_2) = cL_f(v_1) + L_f(v_2)$$

that is, L_f is a linear transformation from V into V^*.

In a similar manner, f determines a linear transformation R_f from V into V^*. For each fixed $w \in V$, $f(v, w)$ is linear as a function of v. We define $R_f(w)$ to be the linear functional on V whose value on the vector v is $f(v, w)$.

Theorem 6.9. Let f be a bilinear form on the finite-dimensional vector space V. Let L_f and R_f be the linear transformations from V into V^* defined by $(L_f(v))(w) = f(v, w) = (R_f(w))(v)$. Then rank$(L_f)$ =rank(R_f).

Definition 6.10. If f is a bilinear form on the finite-dimensional space V, the rank of f is the integer r =rank(L_f) =rank(R_f).

Corollary 6.11. The rank of a bilinear form is equal to the rank of the matrix of the form in any ordered basis.

Corollary 6.12. If f is a bilinear form on the n-dimensional vector space V, the following are equivalent:

1. rank$(f) = n$;
2. For each non-zero $v \in V$, there is a vector $w \in V$ such that $f(v, w) \neq 0$;
3. For each non-zero $w \in V$, there is a vector $v \in V$ such that $f(v, w) \neq 0$.

Definition 6.13. A bilinear form f on a vector space V is called non-degenerate (or non-singular) if it satisfies conditions 2 and 3 of the above corollary.

If V is finite-dimensional, then f is non-degenerate provided f satisfies any one of the three conditions of the above corollary. In particular, f is non-degenerate (non-singular) if and only if its matrix in some (every) ordered basis for V is a non-singular matrix.

Example 6.14. Let $V = \mathbb{R}^n$, and let f be the bilinear form defined on $v = (x_1, \ldots, x_n)$ and $w = (y_1, \ldots, y_n)$ by

$$f(v, w) = x_1y_1 + \cdots + x_ny_n.$$

Then f is a non-degenerate bilinear form on $\mathbb{R}^n$. The matrix of f in the standard ordered basis is the $n \times n$ identity matrix

$$f(X, Y) = X^TY.$$

Example 6.15. Let $V = \mathbb{P}_2$ denote the space of real polynomials of degree at most 2. We can define a bilinear form on V by

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx, \quad f, g \in V.$$

By definition, the matrix of the form is given by

$$a_{ij} = \langle x^{i-1}, x^{j-1}\rangle = \int_0^1 x^{i+j-2}dx = \frac{1}{i+j+2}.$$

Thus, the matrix of the form with respect to the standard basis is

$$A = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}.$$

6.1.1 Symmetric Bilinear Forms

The main purpose of this section is to answer the following question : If f is a bilinear form on the finite-dimensional vector space V, when is there an ordered basis $\mathcal{B}$ for V in which f is represented by a diagonal matrix? We prove that this is possible if and only if f is a symmetric bilinear form, that is, $f(v, w) = f(w, v)$. The theorem is proved only when the scalar field has characteristic zero, that is, that if n is a positive integer the sum $1 + \cdots + 1$ (n times) in F is not 0.

Definition 6.16. Let f be a bilinear form on the vector space V. We say that f is symmetric if $f(v, w) = f(w, v)$ for all $v, w \in V$.

If V is a finite-dimensional, the bilinear form f is symmetric if and only if its matrix A in some (or every) ordered basis is symmetric, $A^T = A$. To see this, one inquires when the bilinear form

$$f(X, Y) = X^T AY$$

is symmetric. This happens if and only if $X^T AY = Y^T AX$, for all column matrices X and Y. Since $X^T AY$ is a 1×1 matrix, we have $X^T AY = Y^T A^T X$. Thus f is symmetric if and only if $Y^T A^T X = Y^T AX$ for all X, Y. Clearly this just means that $A^T = A$. In particular, one should note that if there is an ordered basis for V in which f is represented by a diagonal matrix, then f is symmetric, for any diagonal matrix is a symmetric matrix.

Definition 6.17. If f is a symmetric bilinear form, the quadratic form associated with f is the function q from V into F defined by

$$q(v) = f(v, v).$$

Theorem 6.18. Any quadratic form can be represented by symmetric matrix.

Indeed, if $a_{ij} \neq a_{ji}$, we replace them by new $a'_{ij} = a'_{ji} = \frac{a_{ij}+a_{ji}}{2}$, this does not change the corresponding quadratic form.

Definition 6.19. 1. (**Positive definite**) A bilinear form f on a real vector space V is positive definite, if

$$\langle v, v\rangle = f(v, v) > 0, \quad v \neq 0.$$

A real $n \times n$ matrix A is positive definite if $x^T Ax > 0$ for all $x \neq 0$.

2. (**Negative definite**) A bilinear form f on a real vector space V is negative definite, if

$$\langle v, v\rangle = f(v, v) < 0, \quad v \neq 0.$$

A real $n \times n$ matrix A is positive definite if $x^T Ax < 0$ for all $x \neq 0$.

3. (**Positive Semi-definite**) A bilinear form f on a real vector space V is positive semi-definite, if

$$\langle v, v\rangle = f(v, v) \geq 0, \quad v \in V.$$

A real $n \times n$ matrix A is positive semi-definite if $x^T Ax \geq 0$ for all x.

4. (**Negative Semi-definite**) A bilinear form f on a real vector space V is negative semi-definite, if

$$\langle v, v\rangle = f(v, v) \leq 0, \quad v \in V.$$

A real $n \times n$ matrix A is negative semi-definite if $x^T Ax \leq 0$ for all x.

5. (**Indefinite**) A bilinear form f on a real vector space V is indefinite, if

$$\langle v, v\rangle = f(v, v) > 0, \quad \text{for some } \ v \in V$$

and

$$\langle v, v\rangle = f(v, v) < 0, \quad \text{for some } \ v \in V.$$

Example 6.20. 1. The quadratic form $f(x, y) = x^2 + y^2$ is positive for all nonzero (x, y). Hence f is positive definite.

2. The quadratic form $f(x, y) = -x^2 - y^2$ is negative for all nonzero (x, y). Hence f is negative definite.

3. The quadratic form $f(x, y) = (x - y)^2$ is non-negative. This means that f is either zero or positive for all (x, y). Hence f is positive semi-definite.

4. The quadratic form $f(x, y) = -(x - y)^2$ is non-positive. This means that f is either zero or negative for all (x, y). Hence f is negative semi-definite.

5. The quadratic form $f(x, y) = x^2 - y^2$ is indefinite aince it can take both positive as well as negative for example, $f(3, 1) = 9 - 1 = 8 > 0$ and $f(1, 3) = 1 - 9 = -8 < 0$.

6.1.2 Definiteness of a 2 Variable Quadratic Form

Let $f(x, y) = ax^2 + 2bxy + cy^2$ which is equal to

$$f(x, y) = \begin{bmatrix} x & y \end{bmatrix} \cdot \begin{bmatrix} a & b \\ b & a \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}.$$

Here,

$$A = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

is the symmetric matrix of the quadratic form. The determinant

$$\begin{vmatrix} a & b \\ b & a \end{vmatrix} = ac - b^2$$

is called the discriminant of f. It can be easily seen that

$$ax^2 + 2bxy + cy^2 = a\left(ax + \frac{b}{a}y\right)^2 + \frac{ac - b^2}{a}y^2.$$

Let us use the notation $D_1 = a$, $D_2 = ac - b^2$. Actually D_1 and D_2 are leading principal minors of A. Note that there exists one more principal (non leading) minor (of degree 1) $D_1^{'} = c$. Then

$$f(x, y) = D_1\left(ax + \frac{b}{a}y\right)^2 + \frac{D_2}{D_1}y^2.$$

From this expression we obtain:

1. If $D_1 > 0$ and $D_2 > 0$, then the form $x^2 + y^2$ type, so it is positive definite;
2. If $D_1 < 0$ and $D_2 > 0$, then the form $-x^2 - y^2$ type, so it is negative definite;
3. If $D_1 > 0$ and $D_2 < 0$, then the form $x^2 - y^2$ type, so it is indefinite; If $D_1 < 0$ and $D_2 > 0$, then the form $-x^2 + y^2$ type, so it is also indefinite.

Thus, if $D_2 < 0$, then the form is indefinite.

Semidefiniteness depends not only on leading principal minors D_1, D_2 but also on all principal minors, in this case on $D_1^{'} = c$ too.

4. If $D_1 \geq 0$, $D_1^{'} \geq 0$ and $D_2 \geq 0$, then the form is positive semidefinite.

   Note that the condition $D_1^{'} \geq 0$ is necessary since the form $f(x, y) = -y^2$ with $a = 0$, $b = 0$ and $c = -1$ for which $D_1 = a \geq 0$, $D_2 = ac - b^2 \geq 0$, nevertheless the form is not positive semidiefinite.

5. If $D_1 \leq 0$, $D_1^{'} \leq 0$ and $D_2 \geq 0$, then the form is negative semidefinite.

   Note that the condition $D_1^{'} \leq 0$ is necessary since the form $f(x, y) = y^2$ with $a = 0$, $b = 0$ and $c = 1$ for which $D_1 = a \leq 0$, $D_2 = ac - b^2 \geq 0$, nevertheless the form is not negative semidiefinite.

6.1.3 Definiteness of a 3 Variable Quadratic Form

Let us start with the following example.

Example 6.21. Let $f(x, y, z) = x^2 + 2y^2 - 7z^2 - 4xy + 8xz$. The symmetric matrix of this quadratic form is

$$\begin{bmatrix} 1 & -2 & 4 \\ -2 & 2 & 0 \\ 4 & 0 & -7 \end{bmatrix}.$$

The leading principal minors of this matrix are

$$\left|D_1\right| = 1, \quad \left|D_2\right| = \begin{vmatrix} 1 & -2 \\ -2 & 2 \end{vmatrix} = -2, \quad \left|D_3\right| = \begin{vmatrix} 1 & -2 & 4 \\ -2 & 2 & 0 \\ 4 & 0 & -7 \end{vmatrix} = -18.$$

Also, on simplification, we get

$$f(x, y, z) = x^2 + 2y^2 - 7z^2 - 4xy + 8xz = |D_1|l_1^2 + \frac{D_2}{D_1}l_2^2 + \frac{D_3}{D_3}l_3^2,$$

where

$$\begin{aligned} l_1 &= x - 2y + 4z, \\ l_2 &= y - 4x, \\ l_3 &= z \end{aligned}$$

That is, (l_1, l_2, l_3) are linear combinations of (x, y, z). More precisely,

$$\begin{bmatrix} l_1 \\ l_2 \\ l_3 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 4 \\ 0 & 1 & -4 \\ 0 & 0 & 1 \end{bmatrix} . \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

where

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

is a nonsingular matrix (changing variables).

In general if

$$f(x, y, z) = \begin{bmatrix} x & y & z \end{bmatrix} . \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} . \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

The following three determinants

$$\left|D_1\right| = \left|a_{11}\right|, \quad \left|D_1\right| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \left|D_3\right| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

are leading principal minors. It is possible to show that, if $|D_1| \neq 0$, $|D_2| \neq 0$, then

$$f(x, y, z) = |D_1| l_1^2 + \frac{|D_2|}{|D_1|} l_2^2 + \frac{|D_3|}{|D_2|} l_3^2,$$

where l_1, l_2, l_3 are some linear combinations of x, y, z. This is called Lagrange's Reduction. This implies the following

1. The form is positive definite iff $|D_1| > 0$, $|D_2| > 0$, $|D_3| > 0$, that is all principal minors are positive.
2. The form is negative definite iff $|D_1| < 0$, $|D_2| > 0$, $|D_3| < 0$, that is all principal minors alternate in sign starting with negative one.

Example 6.22. Determine the definiteness of the form $f(x, y, z) = 3x^2 + 2y^2 + 3z^2 - 2xy - 2yz$.

The matrix of our form is

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

The leading principal minors are

$$\left|D_1\right| = 3 > 0, \quad \left|D_1\right| = \begin{vmatrix} 3 & -1 \\ -1 & 2 \end{vmatrix} = 5 > 0, \quad \left|D_3\right| = \begin{vmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{vmatrix} = 18 > 0,$$

thus the form is positive definite.

The above process can be generalized for n variable, which we omit here. We arrive at the following theorems.

Theorem 6.23. 1. A quadratic form is positive definite if and only if

$$|D_1| > 0, \ \ |D_2| > 0, \ \ \cdots \ \ , |D_n| > 0,$$

that is all principal minors are positive;

2. A quadratic form is negative definite if and only if

$$|D_1| < 0, \ \ |D_2| > 0, \ \ |D_3| < 0, \ \ , |D_4| > 0, \ \ \cdots ,$$

that is principal minors alternate in sign starting with negative one.

3. If some kth order leading principal minor is nonzero but does not fit either of the above two sign patterns, then the form is indefinite.

Theorem 6.24. 1. A quadratic form is positive semidefinite if and only if all principal minors are ≥ 0;

2. A quadratic form is negative semidefinite if and only if all principal minors of odd degree are ≤ 0, and all principal minors of even degree are ≥ 0.

6.1.4 Definiteness and Eigen Values

As we know a symmetric $n \times n$ matrix has n real eigenvalues (maybe some multiple).

Theorem 6.25. Given a quadratic form $f(x) = x^T Ax$ and let $c_1, \ldots, c_n$ be eigen values of A. Then f is

1. positive definite iff $c_i > 0$, $i = 1, \ldots, n$;
2. negative definite iff $c_i < 0$, $i = 1, \ldots, n$;
3. positive semidefinite iff $c_i \geq 0$, $i = 1, \ldots, n$;
4. negative semidefinite iff $c_i \leq 0$, $i = 1, \ldots, n$;

6.2 Few Probable Questions

1. Define bilinear forms. Determine the definiteness of the form $f(x, y) = x^2 + 2xy + y^2$.
2. Define quadratic forms. For which real numbers k is the quadratic form $f(x, y) = kx^2 - 6xy + ky^2$ positive-definite?

References

1. Linear Algebra, Friedberg, Insel, Spence
2. Linear Algebra Done Right, Axler
3. Linear Algebra: A Geometric Approach: S. Kumaresan

Core Paper

MATC 3.1
Block - II
Marks : 25 (SSE : 20; IA : 5)

Special Functions

Syllabus

• Unit 7 •

Legendre polynomial : Generating relation, Recurrence relations, Rodrigue's formula, Schlafli's and Laplace's integral formulae, Orthogonal property, Reconstruction of the Legendre differential equations.

• Unit 8 •

Hermite and Laguerre polynomials : Generating relations, Recurrence relations, Rodrigue's formulae, Orthogonal properties, Reconstructions of the respective differential equations.

• Unit 9 •

Chebyshev polynomial : Definition, Series representation, Recurrence relations, Deduction of Chebyshev differential equation, Orthogonal property.

• Unit 10 •

Bessel's functions : Generating relation for integral index, Recurrence relations, Representations for the indices $\frac{1}{2}$ and - $\frac{1}{2}$, Bessel's integral Formulae, Bessel's function of second kind.

Unit 7

Course Structure

- Legendre polynomial : Generating relation, Recurrence relations,
- Rodrigue's formula, Schlafli's and Laplace's integral formulae,
- Orthogonal property, Reconstruction of the Legendre differential equations.

---

7 Introduction

We are familiar with the method of solving ordinary differential equations via series solutions. In particular, we have learnt to find solutions of ODE around a regular point and a regular singular point for the given ODE. We used to employ Frobenius Method to calculate the solution in the latter case. Here, we will study the solutions of certain standard and "difficult" ODE which have applications in various fields using the same method. We will start with Legendre polynomials and explore certain properties of them.

Objectives

After reading this unit, you will be able to

- find the solution of Legendre equations
- define Legendre polynomials
- represent the solutions in a standard manner for further use
- learn the orthogonal properties and Rodrigue's formula for Legendre polynomials

7.1 Legendre Equations

The differential equation of the form

$$(1-x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + n(n+1)y = 0 \tag{7.1.1}$$

where n is a constant is called Legendre's equation. $x = \pm 1$ are the singular points of this equation. Let us see whether $x = \infty$ is a regular singular point of (7.1.1). Let $x = \frac{1}{t}$. Then

$$\frac{dx}{dt} = -\frac{1}{t^2}$$

and hence

$$\frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx} = -t^2\frac{dy}{dt}.$$

Also,

$$\frac{d^2y}{dx^2} = \frac{d}{dt}\left(\frac{dy}{dt}.\frac{dt}{dx}\right)\frac{dt}{dx} = t^4\frac{d^y}{dt^2} + 2t^3\frac{dy}{dt}.$$

Hence equation (7.1.1) becomes

$$t^2(t^2-1)\frac{d^2y}{dt^2}+2t^3\frac{dy}{dt}+n(n+1)y=0. \tag{7.1.2}$$

$t=0$ is clearly a singular point of (7.1.2) which implies that $x=\infty$ is a singular point of (7.1.1). Now, check that

$$\lim_{t\to 0}\frac{2t^4}{t^2(t^2-1)}=0 \quad \& \quad \lim_{t\to 0}t^2\frac{n(n+1)}{t^2(t^2-1)}=-n(n+1).$$

Hence $t=0$ is a regular singular point of (7.1.2).

Assume that

$$y=t^s\sum_{m=0}^{\infty}a_mt^m$$

be a solution of (7.1.2) such that $a_0\neq 0$. Then

$$\frac{dy}{dt}=\sum_{m=0}^{\infty}(m+s)a_mt^{s+m-1} \quad \& \quad \frac{d^2y}{dt^2}=\sum_{m=0}^{\infty}(m+s)(m+s-1)a_mt^{s+m-2}.$$

Then (7.1.2) becomes

$$\sum_{m=0}^{\infty}\{(m+s-2)(m+s-1)a_{m-2}-(m+s+n)(m+s-n-1)a_m\}t^m-(s+n)(s-n-1)a_0-(s+n+1)(s-n)a_1t=0.$$

Then the indicial equation is

$$-(s+n)(s-n-1)a_0=0 \implies s=-n,n+1, \text{ since } a_0\neq 0.$$

When $s=-n$, $a_1=0$ and when $s=n+1$, $a_1=0$. Hence $a_1=0$ in all case and the general recurrence relation is

$$a_m=\frac{(m+s-2)(m+s-1)}{(m+s-n)(m+s-n-1)}, \quad m\geq 2.$$

Since $a_1=0$, so $a_3=a_5=\cdots=a_{2m+1}=\cdots=0$.

Now,

$$\begin{aligned} a_2 &= \frac{s(s+1)}{(s+n+2)(s-n+1)}a_0 \\ a_4 &= \frac{s(s+1)(s+2)(s+3)}{(s+n+2)(s+n+4)(s-n+1)(s-n+3)}a_0 \\ &\vdots \end{aligned}$$

Let n be a positive integer. Taking $m=n+1$, we have

$$a_{n+1}=\frac{(n+s)(n+s-1)}{(2n+s+1)s}a_{n-1}, \quad a_{n+2}=\frac{(n+s)(n+s+1)}{(2n+s+2)(s+1)}a_n.$$

When $s=-n$,

$$a_2=-\frac{n(n-1)}{2(2n-1)}a_0, \quad a_4=\frac{n(n-1)(n-2)(n-3)}{2.4.(2n-1)(2n-3)}a_0,\ldots$$

and $a_{n+1} = a_{n+2} = 0$. Then

$$y = a_0 \left(x^n - \frac{n(n-1)}{2(2n-1)} x^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2.4.(2n-1)(2n-3)} x^{n-4} + \cdots \right). \tag{7.1.3}$$

Taking $s = n + 1$, we have

$$y = a_0 \left(x^{-n-1} - \frac{(n+1)(n+2)}{2(2n+3)} x^{-n-3} + \frac{(n+1)(n+2)(n+3)(n+4)}{2.4.(2n+3)(2n+5)} x^{-n-5} + \cdots \right). \tag{7.1.4}$$

When n is a positive integer, the roots of the indicial equation differ by $2n + 1$, which is an integer. There could be problem in evaluating a_{2n+1} for $s = -n$. But, $a_{n+1} = a_{n+2} = \cdots = 0$, and hence we don't face that problem.

When $n = 1$, $y_1 = a_0 x$.
When $n = 2$, $y_1 = a_0 \left(x^2 - \frac{1}{3} \right)$.
When $n = 3$, $y_1 = \left(x^3 - \frac{3}{5} x \right)$.

If we take

$$a_0 = \frac{1.3.5 \ldots (2n-1)}{n!},$$

then the solution of (7.1.2) is called the **Legendre function of first kind** or **Legendre Polynomial of degree** n and is denoted by $P_n(x)$. Thus, $P_n(x)$ is a solution of (7.1.1). But even if n is a positive integer, solution (7.1.3) is an infinite series. In this case if we take

$$a_0 = \frac{n!}{1.3.5 \ldots (2n+1)},$$

then solution (7.1.3) is denoted by $Q_n(x)$ and is called the Legendre function of second kind. $Q_n(x)$ is not a polynomial and it is linearly independent from $P_n(x)$ and we get the general solution of (7.1.1) as

$$y = AP_n(x) + BQ_n(x).$$

Definition 7.1. Legendre Polynomial of degree n is defined as

$$P_n(x) = \frac{1.3.5 \ldots (2n-1)}{n!} \left(x^n - \frac{n(n-1)}{2(2n-1)} x^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2.4.(2n-1)(2n-3)} x^{n-4} + \cdots \right) \tag{7.1.5}$$

The general term of this polynomial is

$$(-1)^r \frac{n(n-1)(n-2) \ldots (n-2r+1)}{2.4 \ldots 2r(2n-1)(2n-3) \ldots (2n-2r+1)} \frac{1.3.5 \ldots (2n-1)}{n!} x^{n-2r} \tag{7.1.6}$$

Now,

$$1.3.5 \ldots (2n-1) = \frac{1.2.3. \ldots (2n)}{2.4. \ldots (2n)} = \frac{(2n)!}{2^n.n!}.$$

Also,

$$n(n-1)(n-2) \ldots (n-2r+1) = \frac{n!}{(n-2r)!}.$$

$$2.4. \ldots (2r) = 2^r.r!.$$

And

$$(2n-1)(2n-3) \ldots (2n-2r+1) = \frac{(2n)!(n-r)!}{2^r.n!(2n-2r)!}.$$

So, using these things, (7.1.6) becomes

$$(-1)^r \frac{(2n-2r)!}{2^n r!(n-2r)!(n-r)!} x^{n-2r}.$$

(7.1.5) is a polynomial of degree n. Hence $n-2r \geq 0$ or 1 according as n is even or odd, that is, $r \leq \left[\frac{n}{2}\right]$. Hence, Legendre polynomial of degree n is given by

$$P_n(x) = \sum_{r=0}^{\left[\frac{n}{2}\right]} (-1)^r \frac{(2n-2r)!}{2^n r!(n-2r)!(n-r)!} x^{n-2r}.$$

7.1.1 Determination of few Legendre Polynomials

For $n = 0$, we have

$$P_0(x) = (-1)^0 \frac{(2.0-2.0)!}{2^0 0!0!0!} = 1.$$

Similarly, putting $n = 1, 2, 3, 4$ we get

$$\begin{aligned} P_1(x) &= x \\ P_2(x) &= \frac{3}{2}x^2 - \frac{1}{2}. \\ P_3(x) &= \frac{5}{3}x^3 - \frac{3}{2}x. \\ P_4(x) &= \frac{35}{8}x^4 - \frac{15}{4}x^2 + \frac{3}{8}. \end{aligned}$$

7.1.2 Generating Function for Legendre Polynomial

Theorem 7.2. The function

$$w(x,z) = (1-2xz+z^2)^{-1/2}$$

is the generating function for Legendre polynomials, that is,

$$w(x,z) = \sum_{n=0}^{\infty} P_n(x).z^n,$$

holds for sufficiently small values of $|z|$.

Proof. Expanding $(1-2xz+z^2)^{-1/2}$, we get,

$$\begin{aligned} w(x,z) &= (1-a)^{-1/2} \quad \text{taking } a = 2xz - z^2 \\ &= 1 + \frac{a}{2} + \frac{(-1/2)(-1/2-1)}{2!}a^2 + \frac{(-1/2)(-1/2-1)(-1/2-2)}{3!}a^3 + \cdots \\ &= 1 - \frac{2xz-z^2}{2!} + \frac{3}{8}(4x^2z^2 + z^4 - 4xz^3) + \frac{15}{48}(8x^3z^3 - z^6 - 12x^2z^4 + 6xz^5) + \cdots \\ &= 1 - x.z + \left(\frac{3}{2}x^2 - \frac{1}{2}\right)z^2 + \left(\frac{5}{2}x^3 - \frac{3}{2}x\right)z^3 + \cdots \end{aligned}$$

Now, we know that,

$$P_n(x) = \sum_{r=0}^{\left[\frac{n}{2}\right]} (-1)^r \frac{(2n-2r)!}{2^n r!(n-2r)!(n-r)!} x^{n-2r}.$$

Also,

$$\begin{aligned}(1-a)^{-1/2} &= 1+\frac{a}{2}+\frac{(-1/2)(-1/2-1)}{2!}a^2+\frac{(-1/2)(-1/2-1)(-1/2-2)}{3!}a^3+\cdots \\ &= 1+\frac{a}{2}+\frac{1.3}{2^2.2!}a^2+\frac{1.3.5}{2^3.3!}a^3+\cdots\end{aligned}$$

Thus, the kth term is

$$\frac{1.3.5\ldots(2k-1)}{k!.2^k}a^k.$$

Thus, we get

$$w(x,z)=\sum_{k=0}^{\infty}\frac{1.3.5\ldots(2k-1)}{k!.2^k}(2xz-z^2)^k.$$

Now,

$$1.3.5\ldots(2k-1)=\frac{(2k)!}{2^k k!}.$$

Thus,

$$\begin{aligned}w(x,z) &= \sum_{k=0}^{\infty}\frac{(2k)!}{2^{2k}(k!)^2}(2xz-z^2)^k \\ &= \sum_{k=0}^{\infty}\frac{(2k)!}{2^{2k}(k!)^2}\sum_{s=0}^{k}\binom{k}{s}(2xz)^k(-z^2)^{k-s} \\ &= \sum_{k=0}^{\infty}\frac{(2k)!}{2^{2k}(k!)^2}\sum_{s=0}^{k}\binom{k}{s}(2x)^k(-1)^{k-s}z^{2k-s} \\ &= \sum_{k=0}^{\infty}\sum_{s=0}^{\infty}(-1)^{k-s}\frac{(2k)!}{2^{2k}(k!)^2}\frac{k!}{s!(k-s)!}(2x)^s z^{2k-s}.\end{aligned}$$

Consider the portion $(k-s)!$, where s varies from 0 to k. If $s=k+1$, $(k-s)!=(-1)!=\infty$. Similarly, for other $s>k$, $(k-s)!\to\infty$ and so, the terms for $s>k$ becomes zero and the summation can be extended from k to ∞. Interchanging the summations, we get

$$w(x,z)=\sum_{s=0}^{\infty}\sum_{k=0}^{\infty}(-1)^{k-s}\frac{(2k)!}{2^{2k}(k!)^2}\frac{k!}{s!(k-s)!}(2x)^s z^{2k-s}.$$

When $k=0,1,\ldots,(s-1)$, we get $(k-s)!=\infty$. And when $k=s$, $(k-2)!=0!=1$. So, we can effectively start the summation from $k=s$ instead of $k=0$ and the equation becomes

$$w(x,z)=\sum_{s=0}^{\infty}\sum_{k=s}^{\infty}(-1)^{k-s}\frac{(2k)!}{2^{2k}(k!)^2}\frac{k!}{s!(k-s)!}(2x)^s z^{2k-s}.$$

Putting $k-s=p$, and eliminating k, we get

$$w(x,z)=\sum_{s=0}^{\infty}\sum_{p=0}^{\infty}(-1)^p\frac{(2p+2s)!}{2^{2p+s}(s+p)!}\frac{1}{s!p!}x^s z^{2p+s}.$$

Put $2p+s=n$ and eliminate s. Then since p varies from 0 to ∞, s varies from 0 to ∞, n varies from 0 to ∞. Now, $s \geq 0$. So, $n-2p \geq 0$ which implies that $p \leq \left[\frac{n}{2}\right]$. Since p is an integer, $p \leq \left[\frac{n}{2}\right]$. So,

$$\begin{aligned} w(x,z) &= \sum_{n=0}^{\infty}\sum_{p=0}^{\left[\frac{n}{2}\right]}(-1)^p \frac{(2n-2p)!}{2^n(n-p)!}\frac{1}{(n-2p)!p!}x^{n-2p}z^n \\ &= \sum_{n=0}^{\infty} z^n \sum_{p=0}^{\left[\frac{n}{2}\right]}(-1)^p \frac{(2n-2p)!}{2^n(n-p)!}\frac{1}{(n-2p)!p!}x^{n-2p} \\ &= \sum_{n=0}^{\infty} P_n(x)z^n. \end{aligned}$$

□

7.1.3 Recurrence Relations for Legendre Polynomials

Here, we will do certain recurrence relations related to Legendre polynomials.

1. We have, for $n = 0, 1, 2, \ldots,$

$$P_{n+1}^{'}(x) - 2xP_n(x) + P_{n-1}(x) - P_n(x) = 0.$$

Proof. We have

$$w(x,z) = (1-2xz+z^2)^{-1/2} = \sum_{n=0}^{\infty} z^n P_n(x).$$

Taking logarithm on both sides and then differentiating with respect to x, we get

$$\begin{aligned} \frac{d}{dx}[\ln((1-2xz+z^2)^{-1/2})] &= \frac{d}{dx}\left[\ln\left(\sum_{n=0}^{\infty} z^n P_n(x)\right)\right] \\ \text{or, } \frac{1}{2}\frac{2z}{1-2xz+z^2} &= \frac{\sum_{n=0}^{\infty} z^n P_n^{'}(x)}{\sum_{n=0}^{\infty} z^n P_n(x)} \\ \text{or, } (1-2xz+z^2)\sum_{n=0}^{\infty} z^n P_n^{'}(x) &= z\sum_{n=0}^{\infty} z^n P_n(x) \end{aligned}$$

Equating the coefficients of z^n on both sides, we get

$$P_n^{'}(x) - 2xP_{n-1}^{'}(x) + P_{n-2}^{'}(x) = P_{n-1}(x).$$

Replacing n by $n+1$, we get,

$$P_{n+1}^{'}(x) - 2xP_n(x) + P_{n-1}(x) - P_n(x) = 0.$$

□

2. $(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0$, for $n = 0, 1, 2, \ldots$

Proof. We have

$$w(x,z) = (1-2xz+z^2)^{-1/2} = \sum_{n=0}^{\infty} z^n P_n(x).$$

Taking logarithm on both sides and then differentiating with respect to z, we get

$$\begin{aligned} \frac{x-z}{1-2xz+z^2} &= \frac{\sum_{n=1}^{\infty} nP_n(x)z^{n-1}}{\sum_{n=0}^{\infty} z^n P_n(x)} \\ \text{or, } (x-z)\sum_{n=0}^{\infty} z^n P_n(x) &= (1-2xz+z^2)\sum_{n=1}^{\infty} nP_n(x)z^{n-1} \\ &= (1-2xz+z^2)\sum_{n=0}^{\infty}(n+1)P_{n+1}(x)z^n. \end{aligned}$$

Equating coefficients of z^n on both sides, we get,

$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0.$$

□

3. $nP_n(x) = xP_n^{'}(x) - P_{n-1}^{'}(x)$, for $n = 0, 1, 2, \ldots$

Proof. We have

$$(1-2xz+z^2)^{-1/2} = \sum_{n=0}^{\infty} z^n P_n(x). \tag{7.1.7}$$

Differentiating with respect to z, we get,

$$\frac{x-z}{(1-2xz+z^2)^{3/2}} = \sum_{n=1}^{\infty} nP_n(x)z^{n-1} \tag{7.1.8}$$

Again, differentiating (7.1.7) with respect to x, we get,

$$\frac{z}{(1-2xz+z^2)^{3/2}} = \sum_{n=0}^{\infty} P_n^{'}(x)z^n \tag{7.1.9}$$

By (7.1.8)$\times z-$ (7.1.9)$\times(x-z)$, we get

$$(x-z)\sum_{n=0}^{\infty} P_n^{'}(x)z^n = \sum_{n=1}^{\infty} nP_n(x)z^n.$$

Equating the coefficients of z^n on both sides, we get the required result. □

4. $(2n+1)P_n(x) = P_{n+1}^{'}(x) - P_{n-1}^{'}(x)$.

Proof. We have, $(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0$ which gives

$$(2n+1)xP_n(x) = (n+1)P_{n+1}(x) + nP_{n-1}(x).$$

Differentiating both sides with respect to x, we get,

$$(2n+1)P_n(x) + (2n+1)xP_n^{'}(x) = (n+1)P_{n+1}^{'}(x) + nP_{n-1}^{'}(x).$$

From the previous relation 3, we get $xP_n^{'}(x) = nP_n(x) + P_{n+1}^{'}$. Hence the previous equation gives the desired result. □

Exercise 7.3. 1. Prove that $\int_{-1}^{1} P_n(x)dx = 2$, if $n = 0$ and $\int_{-1}^{1} P_n(x)dx = 0$, if $n \geq 1$.

2. Prove the following:

   (a) $(n+1)P_n(x) = P_{n+1}^{'}(x) - xP_n^{'}(x)$.

   (b) $(1-x^2)P_n^{'}(x) = n(P_{n-1}(x) - xP_n(x))$.

   (c) $(1-x^2)P_n^{'}(x) = (n+1)(xP_n(x) - P_{n+1}(x))$.

3. Show that $P_n(x)$ is a solution of Legendre equation of order n.

7.1.4 Rodrigue's Formula

Instead of using the Recurrence relations for the coefficients in the Legendre polynomial, it is easier to use the Rodrigue's Formula.

Legendre Polynomials satisfy the following Rodrigue's formula

$$\frac{1}{2^n n!}\frac{d^n y}{dx^n}(x^2-1)^n = P_n(x).$$

To prove the above result, we find

$$(x^2-1)^n = \sum_{r=0}^{n} \binom{n}{r}(-1)^r (x^2)^{n-r}.$$

Now,

$$\text{RHS} = \frac{1}{2^n n!}\frac{d^n y}{dx^n}(x^2-1)^n = \frac{1}{2^n n!}\sum_{r=0}^{n} \binom{n}{r}(-1)^r \frac{d^n y}{dx^n}(x^2)^{n-r}. \tag{7.1.10}$$

Now,

$$\begin{aligned}\frac{d^n y}{dx^n}(x^m) &= 0;\ m < n \\ &= \frac{m!}{(m-n)!}x^{m-n};\ m \geq n\end{aligned}$$

So, $\frac{d^n y}{dx^n}[x^{2n-2r}]$ will be non-zero if $2n - 2r \geq n$, that is, if $n \geq 2r$, or, $r \leq \frac{n}{2}$. But, r is an integer. So, $r \leq \left[\frac{n}{2}\right]$. Now, from (7.1.10), we get

$$\begin{aligned}\frac{1}{2^n n!}\frac{d^n y}{dx^n}(x^2-1)^n &= \frac{1}{2^n n!}\sum_{r=0}^{\left[\frac{n}{2}\right]} \binom{n}{r}(-1)^r \frac{(2n-2r)!}{(n-2r)!}x^{n-2r} \\ &= \sum_{r=0}^{\left[\frac{n}{2}\right]}(-1)^r \frac{(2n-2r)!}{2^n (n-2r)!(n-r)!r!}x^{n-2r} \\ &= P_n(x) = \text{LHS}\end{aligned}$$

7.2 Orthogonal Property

The Legendre polynomials are orthogonal in the interval $[-1, 1]$ which gives

$$\begin{aligned}\int_{-1}^{1} P_m(x)P_n(x)dx &= 0, \ m \neq n \\ &= \frac{2}{2m+1}, \ m = n\end{aligned}$$

To prove the orthogonality of $P_n(x)$, we will consider two cases, viz., $m = n$ and $m \neq n$. Let us start with the case $m \neq n$.

CaseI: Legendre equation of order m is

$$(1-x^2)\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + m(m+1)y = 0.$$

$P_m(x)$ is a solution of the above equation. So,

$$(1-x^2)P_m^{''}(x) - 2xP_m^{'}(x) + m(m+1)P_m(x) = 0. \tag{7.2.1}$$

Also, $P_n(x)$ is a solution of Legendre equation of order n. So,

$$(1-x^2)P_n^{''}(x) - 2xP_n^{'}(x) + m(m+1)P_n(x) = 0. \tag{7.2.2}$$

Multiplying (7.2.1) by $P_n(x)$ and (7.2.2) by $P_m(x)$ and subtracting, we get

$$\begin{aligned}(1-x^2)[P_m^{''}(x)P_n(x) - P_n^{''}(x)P_m(x)] - \\ 2x[P_m^{'}(x)P_n(x) - P_n^{'}(x)P_m(x)] + [m(m+1) - n(n+1)]P_m(x)P_n(x) &= 0 \\ \text{or } \frac{dy}{dx}[(1-x^2)\{P_m^{'}(x)P_n(x) - P_n^{'}(x)P_m(x)\}] + [m(m+1) - n(n+1)]P_m(x)P_n(x) &= 0\end{aligned}$$

Integrating both sides with respect to x from -1 to 1, we get

$$\begin{aligned}\int_{-1}^{1} \frac{d}{dx}[(1-x^2)\{P_m^{'}(x)P_n(x) - P_n^{'}(x)P_m(x)\}]dx &= (n-m)(n+m+1)\int_{-1}^{1} P_m(x)P_n(x)dx \\ \text{or, } (n-m)(n+m+1)\int_{-1}^{1} P_m(x)P_n(x)dx &= [(1-x^2)\{P_m^{'}(x)P_n(x) - P_n^{'}(x)P_m(x)\}]_{-1}^{1} \\ &= 0 \\ \text{or, } \int_{-1}^{1} P_m(x)P_n(x)dx &= 0.\end{aligned}$$

CaseII: When $m = n$, we have

$$(1-2xz+z^2)^{-1/2} = \sum_{n=0}^{\infty} z^n P_n(x). \tag{7.2.3}$$

Replacing n by m in (7.2.3), we have

$$(1-2xz+z^2)^{-1/2} = \sum_{m=0}^{\infty} z^m P_m(x). \tag{7.2.4}$$

Multiplying (7.2.3) and (7.2.4), we get

$$(1-2xz+z^2)^{-1} = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} z^{n+m}P_n(x)P_m(x).$$

Integrating both sides with respect to x from -1 to 1, we get

$$\begin{aligned}\int_{-1}^{1}(1-2xz+z^2)^{-1}dx &= \int_{-1}^{1}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} z^{n+m}P_n(x)P_m(x)dx \\ &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\int_{-1}^{1} P_n(x)P_m(x)z^{n+m}dx \\ &= \sum_{n=0}^{\infty}\int_{-1}^{1} P_n(x)P_n(x)z^{2n}dx\end{aligned}$$

Now,

$$\int_{-1}^{1}(1-2xz+z^2)^{-1}dx = \frac{1}{z}\ln\left(\frac{1+z}{1-z}\right).$$

Hence,

$$\sum_{n=0}^{\infty}\int_{-1}^{1} P_n(x)P_n(x)z^{2n}dx = \frac{2}{z}\left\{z+\frac{z^3}{3}+\frac{z^5}{5}+\cdots\right\} = \sum_{n=0}^{\infty}\frac{2z^{2n}}{2n+1}.$$

Hence, Equating the like coefficients on both sides, we get

$$\int_{-1}^{1} P_n(x)P_n(x)z^{2n}dx = \frac{2}{2n+1}.$$

7.3 Few Probable Questions

1. Prove that the Legendre polynomials are orthogonal.
2. State and prove the Rodrigue's formula.
3. Prove that for any non-negative integer n, we have $P_{n+1}^{'}(x) - 2xP_n(x) + P_{n-1}^{'}(x) - P_n(x) = 0$.
4. Prove the following:
   (a) $P_n(1) = 1$, $P_n(-1) = (-1)^n$.
   (b) $P_n^{'}(1) = \frac{n(n+1)}{2}$, and $P_n^{'}(-1) = (-1)^{n-1}\frac{n(n+1)}{2}$
   (c) $P_n(-x) = (-1)^n P_n(x)$. Hence deduce that $P_n(-1) = (-1)^n$.

Unit 8

Course Structure

- Hermite and Laguerre polynomials : Generating relations, Recurrence relations,
- Rodrigue's formulae, Orthogonal properties,
- Reconstructions of the respective differential equations.

---

8 Introduction

In mathematics, the Hermite polynomials are a classical orthogonal polynomial sequence. These arise in probability, combinatorics, numerical analysis, systems theory, random matrix theory and many more. Hermite polynomials were defined by Pierre-Simon Laplace in 1810, though in scarcely recognizable form, and studied in detail by Pafnuty Chebyshev in 1859. Chebyshev's work was overlooked, and they were named later after Charles Hermite, who wrote on the polynomials in 1864, describing them as new. They were consequently not new, although Hermite was the first to define the multidimensional polynomials in his later 1865 publications. And the Laguerre polynomials arise in quantum mechanics, in the radial part of the solution of the Schrödinger equation for a one-electron atom. They also describe the static Wigner functions of oscillator systems in quantum mechanics in phase space. They further enter in the quantum mechanics of the Morse potential and of the 3D isotropic harmonic oscillator. The generalized Laguerre polynomials are related to the Hermite polynomials. This unit is dedicated to the study of Hermite as well as Laguerre polynomials.

Objectives

After reading this unit, you will be able to

- solve the Hermite's equation and find the general structure of Hermite's polynomial
- define a general Laguerre polynomial
- derive the Rodrigue's formula for both Hermite and Laguerre polynomials
- establish the orthogonality of Hermite and Laguerre polynomials
- find a generating function for Laguerre and Hermite's polynomials
- learn some recurrence relations relating to both
- solve certain problems relating to both

8.1 Solution of Hermite's Equations

The Hermite's equation is

$$\frac{d^2y}{dx^2} - 2x\frac{dy}{dx} + 2ny = 0 \tag{8.1.1}$$

where, n is a constant. We solve it by Frobenius Method, about $x = 0$. Assume that

$$y = \sum_{m=0}^{\infty} a_m x^{s+m}$$

be the solution of (8.1.1), where $a_0 \neq 0$ and s is to be determined. Then

$$\begin{aligned} \frac{dy}{dx} &= \sum_{m=0}^{\infty}(s+m)a_m x^{s+m-1} \\ \frac{d^2}{dx^2} &= \sum_{m=0}^{\infty}(s+m)(s+m-1)a_m x^{s+m-2} \end{aligned}$$

Thus, equation (8.1.1) becomes

$$\sum_{m=0}^{\infty}(s+m)(s+m-1)a_m x^{s+m-2} - 2\sum_{m=0}^{\infty}(s+m)a_m x^{s+m} + 2n\sum_{m=0}^{\infty}a_m x^{s+m} = 0$$

$$\text{or, } \sum_{m=0}^{\infty}(s+m)(s+m-1)a_m x^{s+m-2} - 2\sum_{m=0}^{\infty}(s+m-n)a_m x^{s+m} = 0$$

$$\text{or, } \sum_{m=0}^{\infty}(s+m)(s+m-1)a_m x^{m} - 2\sum_{m=0}^{\infty}(s+m-n)a_m x^{m+2} = 0$$

$$\text{or, } \sum_{m=0}^{\infty}(s+m)(s+m-1)a_m x^{m} - 2\sum_{m=2}^{\infty}(s+m-n-2)a_{m-2} x^{m} = 0$$

$$\text{or, } \sum_{m=2}^{\infty}\{(s+m)(s+m-1)a_m - 2(s+m-n-2)a_{m-2}\}x^m + s(s-1)a_0 + s(s+1)a_1 x = 0$$

The indicial equation is

$$s(s-1)a_0 = 0 \implies s = 0, 1.$$

When $s = 0$, a_1 is indeterminate. When $s = 1$, $a_1 = 0$. The general recurrance relation is

$$a_m = \frac{2(s+m-n-2)}{(s+m)(s+m-1)}a_{m-2}, \quad m \geq 2$$

For $s = 0$, we have

$$a_m = 2\frac{m-n-2}{m(m-1)}a_{m-2}, \quad m \geq 2.$$

Putting $m = 2, 4, \ldots, 2m, \ldots$, we get

$$\begin{aligned} a_2 &= \frac{(-1)^1 2^1 . n a_0}{2!} \\ a_4 &= \frac{(-1)^2 . 2^2 n(n-2) a_0}{4!} \\ a_6 &= \frac{(-1)^3 . 2^3 n(n-2)(n-4) a_0}{6!} \\ &\vdots \\ a_{2m} &= \frac{(-1)^m . 2^m n(n-2)(n-4)\ldots(n-2m+2)}{(2m)!} a_0 \end{aligned}$$

Next, put $m = 3, 5, \ldots, 2m+1, \ldots$, we get

$$\begin{aligned} a_3 &= \frac{(-1).2.(n-1)}{3!} a_1 \\ a_5 &= \frac{(-1)^2 . 2^2 (n-1)(n-3)}{5!} a_1 \\ &\vdots \\ a_{2m+1} &= \frac{(-1)^m . 2^m (n-1)(n-3)\ldots(n-2m+1)}{(2m+1)!} a_1 \end{aligned}$$

Hence, the series solution gives

$$y = a_0 \left[1 + \frac{(-1)^1 2^1 . n}{2!} x^2 + \cdots + \frac{(-1)^m . 2^m n(n-2)(n-4)\ldots(n-2m+2)}{(2m)!} x^{2m} + \cdots\right] + a_1 \left[x + \frac{(-1).2.(n-1)}{3!} x^3 + \cdots + \frac{(-1)^m . 2^m (n-1)(n-3)\ldots(n-2m+1)}{(2m+1)!} x^{2m+1} + \cdots\right]$$

which is of the form $a_0 y_1(x) + a_1 y_2(x)$. It is observed that in the case when the constant represents a positive integer, then one of the solutions y_1 or y_2 reduces to a polynomial according as n is even or odd.

When $n = 2$, $y_1 = 1 - 2x^2$.
When $n = 4$, $y_1 = 1 - 4x^2 + 4/3x^4$. and so on.
If $n = 1$, $y_2 = x$.
If $n = 3$, $y_2 = x - 2/3x^3$ and so on.

Thus, when n is a positive integer, one solution of Hermite's equation will be a polynomial and the other solution will be an infinite power series. We try to find the form of the polynomial solution which is as follows.

$$\begin{aligned} y &= a_n x^n + a_{n-2} x^{n-2} + \cdots + a_1 x, \quad \text{when } n \text{ is odd} \\ &= a_n x^n + a_{n-2} x^{n-2} + \cdots + a_0, \quad \text{when } n \text{ is even} \end{aligned} \tag{8.1.2}$$

Here, we are going to have series solution of Hermite's equation in decreasing powers of x. While solving Hermite's equation by Frobenius method, we got the recurrence relation when $s = 0$, as

$$a_m = 2\frac{m-n-2}{m(m-1)} a_{m-2}, \quad m \geq 2.$$

Here, we would express all the coefficients in terms of a_n instead of a_1 or a_0. We have from the previous equation,

$$a_{m-2} = \frac{m(m-1)}{2(m-n-2)}a_m.$$

Replacing m by $m+2$, we get

$$a_m = \frac{(m+1)(m+2)}{2(m-n)}a_{m+2}. \tag{8.1.3}$$

Put $m = n-2$. Then we get

$$a_{n-2} = (-1)\frac{n(n-1)}{2.2}a_n.$$

and putting $m = n-4$, we get

$$a_{n-4} = (-1)^2\frac{n(n-1)(n-2)(n-3)}{2.2.2.4}a_n.$$

Putting these in (8.1.2), we get,

$$y = a_n\left[x^n - \frac{n(n-1)}{2.2}x^{n-2} + \cdots + (-1)^r\frac{n(n-1)\ldots(n-2r+1)}{2^r.2.4\ldots 2r}x^{n-2r} + \cdots\right]$$

When n is even, $n-2r \geq 0$ which gives $r \leq n/2$. And when n is odd, $n-2r \geq 1$ which gives $r \leq (n-1)/2$. Thus, r vanishes from 0 to $n/2$ or $(n-1)/2$ according as n is even or odd, which implies that r varies from 0 to $[n/2]$. Hence, the general form of the polynomial solution to (8.1.1) is

$$y = a_n\sum_{r=0}^{\left[\frac{n}{2}\right]}(-1)^r\frac{n(n-1)\ldots(n-2r+1)}{2^{2r}.r!}x^{n-2r}.$$

Taking $a_n = 2^n$, and denoting the solution by $H_n(x)$, we obtain a standard solution to (8.1.1) known as the Hermite's polynomial of order n.

Definition 8.1. Hermite's polynomial of order n is denoted and defined by

$$H_n(x) = \sum_{r=0}^{\left[\frac{n}{2}\right]}(-1)^r\frac{n!}{r!(n-2r)!}(2x)^{n-2r}.$$

Exercise 8.2. Compute some of the first Hermite's polynomials.

8.1.1 Generating Function for Hermite's Polynomial

We have,

$$\mathrm{e}^{2tx-t^2} = \sum_{n=0}^{\infty}\frac{t^n}{n!}H_n(x).$$

Each term in the expansion of e^{2tx-t^2} gives a Hermite's polynomial. In fact, coefficient of $t^n/n!$ gives a Hermite's polynomial of order n. That is why e^{2tx-t^2} is called the generating function for Hermite's polynomial.

We have,

$$\begin{aligned} \mathrm{e}^{2tx-t^2} = \mathrm{e}^{2tx}\,\mathrm{e}^{-t^2} &= \sum_{s=0}^{\infty}\frac{(2tx)^s}{s!}\sum_{r=0}^{\infty}\frac{(-t^2)^r}{r!} \\ &= \sum_{s=0}^{\infty}\sum_{r=0}^{\infty}(-1)^r\frac{t^{s+2r}}{r!s!}(2x)^s. \end{aligned}$$

Putting $s+2r=n$ and eliminating s we get

$$\mathrm{e}^{2tx-t^2} = \sum_{n=0}^{\infty}\sum_{r=0}^{\left[\frac{n}{2}\right]}(-1)^r\frac{t^n}{r!(n-2r)!}(2x)^{n-2r}$$

since both r and s varies from 0 to ∞, so n varies from 0 to ∞ and $n-2r\geq 0$ which gives $r\leq n/2$ and we arrive at the same conclusion as we had arrived in case of Legendre's polynomial. Thus, we have

$$\mathrm{e}^{2tx-t^2} = \sum_{n=0}^{\infty}\left\{\sum_{r=0}^{\left[\frac{n}{2}\right]}(-1)^r\frac{n!}{r!(n-2r)!}(2x)^{n-2r}\right\}\frac{t^n}{n!} = \sum_{n=0}^{\infty}\frac{t^n}{n!}H_n(x).$$

Exercise 8.3. Verify the fact that e^{2tx-t^2} is indeed the generating function for $H_n(x)$ by expanding the exponential function and showing that the coefficients of the individual terms $t^n/n!$ are indeed the Hermite's polynomials.

8.1.2 Rodrogue's Formula for Hermite's Polynomial

Hermite's polynomials satisfy the following formula which is known as the Rodrigue's formula for Hermite's polynomials

$$H_n(x) = (-1)^n\,\mathrm{e}^{x^2}\,\frac{d^n}{dx^n}(\mathrm{e}^{-x^2}).$$

The function e^{2tx-t^2} is analytic in any neighbourhood of the point $t=0$. Thus, for any fixed value of x it has a Taylor series expansion of the form

$$\mathrm{e}^{2tx-t^2} = \sum_{n=0}^{\infty}\frac{t^n}{n!}\left[\frac{\partial^n}{\partial t^n}\left(\mathrm{e}^{2tx-t^2}\right)\right]_{t=0}. \tag{8.1.4}$$

Now, we have

$$\frac{\partial^n}{\partial t^n}\left(\mathrm{e}^{2tx-t^2}\right) = \mathrm{e}^{x^2}\,\frac{\partial^n}{\partial t^n}\left(\mathrm{e}^{-(x-t)^2}\right).$$

On calculation, we get

$$\frac{\partial}{\partial t}\left(\mathrm{e}^{-(x-t)^2}\right) = -2(x-t)\,\mathrm{e}^{-(x-t)^2} = -\frac{\partial}{\partial x}\left(\mathrm{e}^{-(x-t)^2}\right).$$

By repeated use of this, we get

$$\frac{\partial^n}{\partial t^n}\left(\mathrm{e}^{-(x-t)^2}\right) = (-1)^n\frac{\partial^n}{\partial x^n}\left(\mathrm{e}^{-(x-t)^2}\right)$$

Thus, we get

$$\left[\frac{\partial^n}{\partial t^n}\left(\mathrm{e}^{2tx-t^2}\right)\right]_{t=0} = \mathrm{e}^{x^2}(-1)^n\frac{d^n}{dx^n}\left(\mathrm{e}^{-x^2}\right).$$

Using this result in (8.1.4), we get

$$\mathrm{e}^{2tx-t^2} = \mathrm{e}^{x^2}\sum_{n=0}^{\infty}(-1)^n\frac{d^n}{dx^n}\left(\mathrm{e}^{-x^2}\right)\frac{t^n}{n!}.$$

From the formula of generating function of Hermite's polynomial and equating the coefficients of $t^n/n!$ on both sides, we get the required result.

Exercise 8.4. 1. Verify Rodrigue's formula for first three Hermite's polynomials.

2. Find Hermite's polynomials upto order 6 by using Rodrigue's formula.

3. Prove that $H_{2n}(0) = (-1)^n\frac{(2n)!}{n!}$ and $H_{2n+1}(0) = 0$.

8.1.3 Recurrence Relations for Hermite's Polynomials

We have

1. $H_n^{'}(x) = 2nH_{n-1}(x)$, for $n \geq 1$ and $H_0^{'}(0) = 0$.

   Proof. We have

   $$\mathrm{e}^{2tx-t^2} = \sum_{n=0}^{\infty}H_n(x)\frac{t^n}{n!}.$$

   Differentiating both sides with respect to x, we have

   $$\begin{aligned}\sum_{n=0}^{\infty}H_n^{'}(x)\frac{t^n}{n!} = 2t.\,\mathrm{e}^{2tx-t^2} &= 2t\sum_{n=0}^{\infty}H_n(x)\frac{t^n}{n!}\\ &= \sum_{n=0}^{\infty}2H_n(x)\frac{t^{n+1}}{n!}\\ &= \sum_{n=1}^{\infty}2nH_{n-1}(x)\frac{t^n}{n!}\end{aligned}$$

   Equating the power of t^0 on both sides, we get $H_0^{'}(0) = 0$ and equating the coefficients of $t^n/n!$ on both sides for $n \geq 1$, we get the desired result. □

2. $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$, for $n \geq 1$ and $H_1(x) = 2xH_0(x)$.

   Proof. We have

   $$\sum_{n=0}^{\infty}H_n(x)\frac{t^n}{n!} = \mathrm{e}^{2tx-t^2}.$$

Differentiating both sides with respect to t, we get

$$\begin{aligned} 2(x-t).\,\mathrm{e}^{2tx-t^2} &= \sum_{n=0}^{\infty} nH_n(x)\frac{t^{n-1}}{n!} \\ \text{or, } 2(x-t)\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} &= \sum_{n=1}^{\infty} H_n(x)\frac{t^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} H_{n+1}(x)\frac{t^n}{n!} \\ \text{or, } 2x\sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!} &= 2\sum_{n=0}^{\infty} H_n(x)\frac{t^{n+1}}{n!} + \sum_{n=0}^{\infty} H_{n+1}(x)\frac{t^n}{n!} \end{aligned}$$

Equating the coefficients of t^0 and $t^n/n!$ on both sides, we get the desired results. □

3. $H_n^{'}(x) = 2xH_n(x) - H_{n+1}(x)$.

   Proof. Left as an exercise. □

4. $H_n^{''}(x) - 2xH_n^{'}(x) + 2nH_n(x) = 0$.

   Proof. Left as an exercise. □

8.1.4 Orthogonality Properties

We have

$$\begin{aligned} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2} H_m(x)H_n(x)dx &= 0, \;\; m \neq n \\ &= \sqrt{\pi}2^n.n!, \;\; m = n \end{aligned}$$

This shows that the Hermite's polynomials are orthogonal in the interval $(-\infty, \infty)$.

We have

$$\mathrm{e}^{2tx-t^2} = \sum_{n=0}^{\infty} H_n(x)\frac{t^n}{n!}.$$

Replacing n by m and t by s and multiplying the resulting equation with the above equation we get,

$$\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{H_n(x)H_m(x)}{n!.m!}t^n s^m = \mathrm{e}^{2tx-t^2=2sx-s^2}.$$

Multiplying both sides by e^{-x^2} and integrating with respect to x from $-\infty$ to ∞, we get

$$\begin{aligned} \int_{-\infty}^{\infty}\left(\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{\mathrm{e}^{-x^2} H_n(x)H_m(x)}{n!.m!}\right)dx.t^n s^m &= \int_{-\infty}^{\infty} \mathrm{e}^{-x^2+2x(t+s)-(t^2+s^2)} \\ \text{or, } \sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left(\int_{-\infty}^{\infty} \mathrm{e}^{-x^2} H_n(x)H_m(x)dx\right)\frac{t^n.s^m}{n!.m!} &= \mathrm{e}^{2ts}\int_{-\infty}^{\infty} \mathrm{e}^{-(x-(t+s))^2}\,dx \end{aligned} \tag{8.1.5}$$

Putting $x-(t+s)=y$, we get, $dx=dy$. Thus,

$$\begin{aligned} \mathrm{e}^{2ts}\int_{-\infty}^{\infty}\mathrm{e}^{-y^2} &= \mathrm{e}^{2ts}\sqrt{\pi}, \quad \text{using gamma integral} \\ &= \sum_{n=0}^{\infty}\frac{2^n.t^n.s^n}{n!}\sqrt{\pi}. \end{aligned}$$

Thus, (8.1.5) becomes

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left(\int_{-\infty}^{\infty}\mathrm{e}^{-x^2}\,H_n(x)H_m(x)dx\right)\frac{t^n.s^m}{n!.m!}=\sqrt{\pi}\sum_{n=0}^{\infty}\frac{2^n.t^n.s^n}{n!} \tag{8.1.6}$$

We note that the powers of t and s are always equal in each term of the RHS of (8.1.6). So when $m\neq n$, equating the coefficients of $t^n s^m$ on both sides of (8.1.6), we have

$$\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-x^2}\,H_n(x)H_m(x)}{n!m!}dx=0,$$

and when $m=n$, we have

$$\int_{-\infty}^{\infty}\frac{\mathrm{e}^{-x^2}\,H_n(x)H_n(x)}{n!}dx=2^n.n!\sqrt{\pi}.$$

Hence, we are done.

Exercise 8.5. 1. Prove that

$$\int_{-\infty}^{\infty}x^2\,\mathrm{e}^{-x^2}\,H_n(x)H_n(x)dx=\sqrt{\pi}2^n.n!\left(n+\frac{1}{2}\right).$$

8.2 Laguerre Polynomials

The Laguerre equation is of the form

$$x\frac{d^2y}{dx^2}+(1-x)\frac{dy}{dx}+ny=0 \tag{8.2.1}$$

where n is a constant. When n is a positive integer, then the solution of (8.2.1) is called the Laguerre polynomial which is of the form

$$L_n(x)=\sum_{r=0}^{n}\frac{(-1)^r}{r!}\binom{n}{r}x^r.$$

Exercise 8.6. Compute the first few Laguerre polynomials using the summation formula.

8.2.1 Generating Function for Laguerre Polynomials

The generating function for the Laguerre polynomials is

$$g(x,t) = \frac{\mathrm{e}^{-\frac{xt}{1-t}}}{1-t} = \sum_{n=0}^{\infty} t^n L_n(x)$$

since each term of the summation contains a Laguerre polynomial.

We have,

$$\begin{aligned}
\frac{\mathrm{e}^{-\frac{xt}{1-t}}}{1-t} = \frac{1}{1-t}\sum_{r=0}^{\infty}\left(\frac{-xt}{1-t}\right)^r \frac{1}{r!} &= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!}x^r t^r (1-t)^{-r-1} \\
&= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!}x^r t^r \sum_{s=0}^{\infty}\frac{(r+s)!}{r!s!}t^s \\
&= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}(-1)^r\frac{(r+s)!}{(r!)^2 s!}x^r t^{r+s}.
\end{aligned}$$

We put $r+s=n$, that is, $s=n-r$, where r is fixed. Then the coefficients of t^n is

$$(-1)^r\frac{n!}{(r!)^2(n-r)!}x^r$$

Now, $s \geq 0$ implies $r \leq n$, giving all possible values of r. Hence all the coefficients of t^n is given by

$$\sum_{r=0}^{n}(-1)^r\frac{n!}{(r!)^2(n-r)!}x^r = L_n(x).$$

Hence the result.

Exercise 8.7. Prove that $\int_0^\infty \mathrm{e}^{-st}\, L_n(t)dt = 1/s(1-1/s)^n$.

8.2.2 Rodrigue's Formula for Laguerre polynomial

The Rodrigue's representation for Laguerre polynomials is

$$L_n(x) = \frac{\mathrm{e}^x}{n!}\frac{d^n}{dx^n}(x^n\,\mathrm{e}^{-x}).$$

By the Leibnitz's theorem, we have

$$\begin{aligned}
D(uv) &= \frac{d^n}{dx^n}(uv) \\
&= D^n u.v + \binom{n}{1}D^{n-1}u.Dv + \cdots + \binom{n}{r}D^{n-r}u.D^r v + \cdots + uD^n v.
\end{aligned}$$

Using this, we get

$$\frac{\mathrm{e}^x}{n!}\frac{d^n}{dx^n}(x^n\,\mathrm{e}^{-x}) = \frac{\mathrm{e}^x}{n!}\sum_{r=0}^{n}\binom{n}{r}D^{n-r}x^n D^r\,\mathrm{e}^{-x}\,.$$

We have

$$D^n x^m = \frac{m!}{(m-n)!}x^{m-n} \quad \text{and} \quad D^n\,\mathrm{e}^{ax} = a^n\,\mathrm{e}^{ax}\,.$$

Thus, using these in the previous equation, we get

$$\begin{aligned}\frac{\mathrm{e}^x}{n!}\frac{d^n}{dx^n}(x^n\,\mathrm{e}^{-x}) &= \frac{\mathrm{e}^x}{n!}\sum_{r=0}^{n}\binom{n}{r}\frac{n!}{(n-(n-r))!}x^{n-(n-r)}(-1)^r\,\mathrm{e}^{-x}\\ &= \sum_{r=0}^{n}\frac{\mathrm{e}^x}{n!}\frac{n!}{r!(n-r)!}.\frac{n!}{r!}.x^r.(-1)^r\,\mathrm{e}^{-x} = L_n(x).\end{aligned}$$

Exercise 8.8. 1. Verify the Rodrigue's formula for first four positive integers.

8.2.3 Recurrence Relations

The Laguerre polynomials satisfy the recurrence relations

1. $(n+1)L_{n+1}(x) = (2n+1-x)L_n(x) - nL_{n-1}(x)$.

   Proof. We have, $g(x,t) = \frac{\mathrm{e}^{-\frac{xt}{1-t}}}{1-t} = \sum_{n=0}^{\infty} t^n L_n(x)$. Differentiating both sides with respect to t, we get

   $$\begin{aligned}\sum_{n=0}^{\infty} t^{n-1}nL_n(x) &= \frac{1}{(1-t)^2}\,\mathrm{e}^{-\frac{xt}{1-t}} - \frac{1}{(1-t)}\,\mathrm{e}^{-\frac{xt}{1-t}}.\frac{x}{(1-t)^2}\\ &= \frac{1}{1-t}\sum_{n=0}^{\infty} t^n L_n(x) - \frac{x}{(1-t)^2}\sum_{n=0}^{\infty} t^n L_n(x).\end{aligned}$$

   Multiplying both sides by $(1-t^2)$ and simplifying, we obtain

   $$\sum_{n=0}^{\infty} t^{n-1}nL_n(x) - 2\sum_{n=0}^{\infty} t^n nL_n(x) + \sum_{n=0}^{\infty} t^{n+1}nL_n(x) = \sum_{n=0}^{\infty} t^n L_n(x) - \sum_{n=0}^{\infty} t^{n+1}L_n(x) - x\sum_{n=0}^{\infty} t^n L_n(x).$$

   Equating the coefficients of t^n on both sides, we get the desired result. □

2. $xL_n'(x) = nL_n(x) - nL_{n-1}(x)$.

   Proof. We have,

   $$g(x,t) = \frac{\mathrm{e}^{-\frac{xt}{1-t}}}{1-t} = \sum_{n=0}^{\infty} t^n L_n(x).$$

   Differentiating both sides with respect to x, and using the generating function, we get

   $$\sum_{n=0}^{\infty} t^n L_n'(x) = \frac{1}{1-t}\,\mathrm{e}^{-\frac{xt}{1-t}}.\frac{-t}{1-t} = \frac{-t}{1-t}\sum_{n=0}^{\infty} t^n L_n(x).$$

   Multiplying both sides by $(1-t)$ and simplifying, we get

   $$\sum_{n=0}^{\infty} t^n L_n'(x) - \sum_{n=0}^{\infty} t^{n+1}L_n'(x) = \sum_{n=0}^{\infty} t^{n+1}L_n(x).$$

   Equating the coefficients of t^n on both sides, we get the desired result. □

3. $L_n^{'}(x) = -\sum_{r=0}^{n-1} L_r(x)$.

Proof. We have,

$$g(x,t) = \frac{\mathrm{e}^{-\frac{xt}{1-t}}}{1-t} = \sum_{n=0}^{\infty} t^n L_n(x).$$

Differentiating both sides with respect to x, and using the generating function, we get

$$\sum_{n=0}^{\infty} t^n L_n^{'}(x) = \frac{1}{1-t}\,\mathrm{e}^{-\frac{xt}{1-t}}.\frac{-t}{1-t} = \frac{-t}{1-t}\sum_{r=0}^{\infty} t^r L_r(x) = -t\sum_{s=0}^{\infty} t^s \sum_{r=0}^{\infty} t^r L_r(x).$$

Thus,

$$\sum_{n=0}^{\infty} t^n L_n^{'}(x) = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} t^{r+s+1} L_r(x).$$

The coefficients of t^n on the LHS is clearly $L_n^{'}(x)$. We will find the coefficients of t^n on the RHS. Let $r+s+1=n$, so that $s=n-r-1$. Hence, for a fixed value of r, the coefficient of t^n on the RHS of the above equation is $-L_r(x)$. But, $s \geq 0$, which implies that $n-r-1 \geq 0 \implies r \leq n-1$, which gives all the values of r for which $-L_r(x)$ is the coefficient of t^n. Hence the total coefficients of t^n on the RHS is given by $-\sum_{r=0}^{n-1} L_r(x)$ and equating the coefficients on both sides, we get the desired result. □

Exercise 8.9. 1. Deduce the second recurrence relation from the first.

2. Prove that
   (a) $L_n^{'}(x) = n[L_{n-1}^{'}(x) - L_{n-1}(x)]$.
   (b) $xL_n(x) = nL_n(x) - n^2 L_{n-1}(x)$.

8.2.4 Orthogonality Properties

If $L_m(x)$ and $L_n(x)$ are Laguerre polynomials (m, n being positive integers), then

$$\int_0^{\infty} \mathrm{e}^{-x} L_n(x) L_m(x) dx = 0, \quad m \neq n$$
$$= 1, \quad m = n.$$

The generating function for Laguerre polynomial gives

$$\frac{\mathrm{e}^{-\frac{xt}{1-t}}}{1-t} = \sum_{n=0}^{\infty} t^n L_n(x)$$
$$\&, \quad \frac{\mathrm{e}^{-\frac{xs}{1-s}}}{1-s} = \sum_{m=0}^{\infty} s^m L_m(x).$$

Multiplying both the equations we get

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} s^m.t^n.L_m(x)L_n(x) = \frac{\mathrm{e}^{-\frac{xs}{1-s}-\frac{xt}{1-t}}}{(1-s)(1-t)}$$

Multiplying both sides by e^{-x} and integrating with respect to x from 0 to ∞, we get

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left\{\int_0^{\infty}\mathrm{e}^{-x}\,L_m(x)L_n(x)dx\right\}s^m.t^n &= \frac{1}{(1-t)(1-s)}\int_0^{\infty}\mathrm{e}^{-x(1+t/(1-t)+s(1-s)}\,dx\\ &= \frac{1}{(1-t)(1-s)}\left|\frac{\mathrm{e}^{-x(1+t/(1-t)+s(1-s)}}{-(1+t/(1-t)+s(1-s)}\right|_0^{\infty}\\ &= \frac{1}{1-st}.\end{aligned}$$

Now, we have

$$(1-st)^{-1} = 1+st+s^2t^2+\cdots = \sum_{n=0}^{\infty}s^nt^n.$$

Using this in the previous equation, we get

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left\{\int_0^{\infty}\mathrm{e}^{-x}\,L_m(x)L_n(x)dx\right\}s^m.t^n = \sum_{n=0}^{\infty}s^nt^n.$$

Equating the coefficients of t^ns^m on both sides, we get the desired result.

Exercise 8.10. 1. Prove that $\int_x^{\infty}\mathrm{e}^{-y}\,L_n(y)dy = \mathrm{e}^{-x}[L_n(x)-L_{n-1}(x)]$.

8.3 Few Probable Questions

1. Solve the Hermite's differential equation and deduce the structure of the Hermite's polynomial.
2. Establish the Rodrige's polynomial for Hermite's polynomial.
3. Establish the Rodrige's polynomial for Laguerre polynomial.
4. Show that e^{2xt-t^2} is the generating function for Hermite's polynomial.
5. Show that $\mathrm{e}^{-xt/(1-t)}/(1-t)$ is the generating function for Laguerre polynomial.
6. Establish the orthogonality of Hermite's polynomials.
7. Establish the orthogonality of Laguerre polynomials.
8. Show that e^{2tx-t^2} is the generating function for the Hermite's polynomial. Hence show that for $m<n$,

$$\frac{d^m}{dx^m}H_m(x) = \frac{2^m n!}{(n-m)!}H_{n-m}(x).$$

Unit 9

Course Structure

- Chebyshev polynomial : Definition, Series representation,
- Recurrence relations, Deduction of Chebyshev differential equation,
- Orthogonal property.

---

9 Introduction

In mathematics the Chebyshev polynomials, named after Pafnuty Chebyshev, are a sequence of orthogonal polynomials which are related to de Moivre's formula and which can be defined recursively. Chebyshev polynomials are important in approximation theory because the roots of the Chebyshev polynomials of the first kind, which are also called Chebyshev nodes, are used as nodes in polynomial interpolation. The resulting interpolation polynomial minimizes the problem of Runge's phenomenon and provides an approximation that is close to the polynomial of best approximation to a continuous function under the maximum norm. This approximation leads directly to the method of Clenshaw–Curtis quadrature. We will study about the Chebyshev polynomials and its properties in this unit.

Objectives

After reading this section, you will be able to

- know the Chebyshev's equations
- define Chebyshev's polynomials
- learn the Rodrigue's formula for Chebyshev's polynomials
- deduce a generating function for Chebyshev's polynomials
- learn the orthogonality condition for Chebyshev's polynomials
- learn the recurrence relations concerning Chebyshev polynomials
- solve various problems related to the above topics

9.1 Chebyshev Polynomials

The Chebyshev differential equation is written as

$$(1-x^2)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + n^2y = 0, \tag{9.1.1}$$

where $|x| < 1$ and n is any real number. This equation can be converted to a simpler form using the substitution $x = \cos t$. Then we have

$$dx = -\sin t dt \implies \frac{dt}{dx} = -\frac{1}{\sin t}.$$

Hence,

$$\frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx} = -\frac{1}{\sin t}\frac{dy}{dt},$$

and

$$\frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dt}\frac{dt}{dx}\left(-\frac{1}{\sin t}\frac{dy}{dt}\right) = -\frac{1}{\sin t}\frac{d}{dt}\left(-\frac{1}{\sin t}\frac{dy}{dt}\right) = \frac{1}{\sin^2 t}\left[\left(-\frac{\cos t}{\sin t}\right)\frac{dy}{dt} + \frac{d^2y}{dt^2}\right].$$

Substituting these in (9.1.1), and simplifying, we get

$$\frac{d^2y}{dt^2} + n^2 y = 0,$$

whose general solution is given by

$$y(t) = C\cos(nt + a).$$

For simplicity, we set $a = 0$. Thus, the general solution of the equation (9.1.1) is given by

$$y(x) = C\cos(n \arccos x).$$

Now, if n is an integer, then the above function is the Chebyshev polynomial of first kind.

Definition 9.1. The Chebyshev polynomial of the first kind is called the function

$$T_n(x) = \cos(n \arccos x) = \frac{n}{2}\sum_{k=0}^{[n/2]}(-1)^k \frac{(n-k-1)!}{k!(n-2k)!}(2x)^{n-2k},$$

where $|x| < 1$ and $n = 0, 1, 2, \ldots$

Exercise 9.2. Compute few Chebyshev's polynomials using the formula given above.

9.2 Recurrence Relations

The Chebyshev's polynomial follows the following recurrence relations

1. $2xT_n(x) = T_{n+1}(x) + T_{n-1}(x)$.

   Proof. Putting $x = \cos t$, we have

   $$T_{n-1}(t) = \cos((n-1)\arccos x) = \cos((n-1)t)$$
   $$T_{n+1}(t) = \cos((n+1)\arccos x) = \cos((n+1)t).$$

   Also, we have,

   $$T_1(x) = \cos(\arccos x) = \cos(\arccos(\cos t)) = \cos t = x$$
   $$T_n(x) = \cos(n\arccos x) = \cos(n\arccos(\cos t)) = \cos(nt).$$

Further,

$$\begin{aligned}\cos((n-1)t)+\cos((n+1)t) &= 2\cos\frac{(n-1)t+(n+1)t}{2}\\ &= 2\cos\frac{2nt}{2}\cos\frac{-2t}{2} = 2\cos(nt)\cos t.\end{aligned}$$

Thus, we get

$$T_{n-1}(x)+T_{n+1} = 2T_n(x)T_1(x) = 2xT_n(x).$$

□

2. $(1-x^2)T_n'(x) = -nxT_n(x) + nT_{n-1}(x)$.
3. For $-1 < x < 1$, we have $T_n^2(x) - T_{n-1}(x)T_{n+1}(x) = 1-x^2$.

Exercise 9.3.
1. Compute few higher Chebyshev's polynomials using the first Recurrence relation and $T_1(x) = x$.
2. Prove that $T_n(-x) = (-1)^n T_n(x)$. Thus show that $T_n(-1) = (-1)^n$.
3. Show that

$$\begin{aligned}T_n(0) &= (-1)^n, \text{ when } n \text{ is even}\\ &= 0, \text{ when } n \text{ is odd.}\end{aligned}$$

9.3 Rodrigue's Formula

The Chebyshev's polynomial satisfy the following Rodrigue's formula

$$T_n(x) = \frac{(-2)^n n!}{(2n)!}\sqrt{1-x^2}\frac{d^n}{dx^n}(1-x^2)^{n-1/2}.$$

Exercise 9.4. Verify Rodrigue's formula for first few Chebyshev's polynomials.

9.4 Generating Functions

The function

$$w(x,t) = \frac{2-xt}{1-xt+t^2},$$

is the generating function for Chebyshev polynomials, that is,

$$\frac{2-xt}{1-xt+t^2} = \sum_{n=0}^{\infty} T_n(x)t^n.$$

We have,

$$\frac{1}{1-xt+t^2} = \sum_{k=0}^{\infty}\sum_{l=0}^{k}(-1)^{k-l}\binom{k}{k-l}x^l t^{2k-l} = \sum_{k,l}(-1)^{k-l}\binom{k}{k-l}x^l t^{2k-l}.$$

Put $m = 2k - l$ and $n = k - l$. Then $k = m - n$ and $l = m - 2n$. Then the above equation changes to

$$\sum_{m,n}(-1)^n\binom{m-n}{n}x^{m-2n}t^m.$$

Now,

$$\begin{aligned}
(2-xt)\sum_{m,n}(-1)^n\binom{m-n}{n}x^{m-2n}t^m &= 2\sum_{m,n}(-1)^n\binom{m-n}{n}x^{m-2n}t^m + \sum_{m,n}(-1)^{n+1}\binom{m-n}{n}x^{m-2n+1}t^{m+1} \\
&= 2\sum_{m,n}(-1)^n\binom{m-n}{n}x^{m-2n}t^m + \sum_{m,n}(-1)^n\binom{m-n-1}{n}x^{m-2n}t^m \\
&= \sum_{m,n}(-1)^n\left\{2\binom{m-n}{n} - \binom{m-n-1}{n}\right\}x^{m-2n}t^m \\
&= \sum_{m,n}(-1)^n\frac{m}{m-n}\binom{m-n}{n}x^{m-2n}t^m = \sum_m T_m(x)t^m.
\end{aligned}$$

9.5 Orthogonality Property

The Chebyshev's polynomials satisfy the following orthogonal property

$$\begin{aligned}
\int_{-1}^{1}\frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}dx &= 0, \ \ m \neq n \\
&= \pi/2, \ \ m = n \neq 0 \\
&= \pi, \ \ m = n = 0.
\end{aligned}$$

9.6 Few Probable Questions

1. Define Chebyshev polynomials. Solve Chebyshev's equation.
2. Deduce a generating function for Chebyshev polynomials.

Unit 10

Course Structure

- Bessel's functions : Solutions of Bessel's equations, Generating relation for integral index,
- Recurrence relations, Representations for the indices $\frac{1}{2}$ and - $\frac{1}{2}$,
- Bessel's integral Formulae, Bessel's function of second kind.

---

10 Introduction

The differential equation

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - n^2)y = 0 \tag{10.0.1}$$

where n is a constant, is called a Bessels's equation whose solution gives the Bessel's functions. These are an orthogonal sequence of functions that have many closely related definitions. This unit is dedicated to the study of Bessel's functions and its properties.

Objectives

After reading this section, you will be able to

- solve Bessel's equations by Frobenius method
- deduce the generating functions for integral index
- deduce the recurrence relations for Bessel's functions
- deduce the orthogonality condition for Bessel's functions
- solve related problems

10.1 Solution of Bessel's Equations

To solve the differential equation (10.0.1), we see that $x = 0$ is a regular singular point of (10.0.1) (Verify!). We will thus attempt to solve it about $x = 0$ by Frobenius method. The resulting series solution is valid in a neighbourhood of $x = 0$. Let us assume the solution to be

$$y = \sum_{m=0}^{\infty} c_m x^{m+s}, \quad \text{where } c_0 \neq 0$$

and s is to be determined. Thus

$$\frac{dy}{dx} = \sum_{m=0}^{\infty}(m+s)c_m x^{m+s-1}, \quad \& \quad \frac{d^2y}{dx^2} = \sum_{m=0}^{\infty}(m+s)(m+s-1)c_m x^{m+s-2}$$

Thus, (10.0.1) becomes

$$\sum_{m=0}^{\infty}(m+s)(m+s-1)c_m x^{m+s}+\sum_{m=0}^{\infty}(m+s)c_m x^{m+s}+\sum_{m=0}^{\infty}c_m x^{m+s+2}-n^2\sum_{m=0}^{\infty}c_m x^{m+s} = 0$$

or,
$$\sum_{m=0}^{\infty}\left\{(m+s)(m+s-1)+(m+s)-n^2\right\}c_m x^{m+s}+\sum_{m=0}^{\infty}c_m x^{m+s+2} = 0$$

or,
$$\sum_{m=0}^{\infty}(m+s+n)(m+s-n)c_m x^m+\sum_{m=2}^{\infty}c_{m-2}x^m = 0$$

or,
$$\sum_{m=2}^{\infty}\left\{(m+s+n)(m+s-n)c_m+c_{m-2}\right\}x^m+(s+n)(s-n)c_0+(1+s+n)(1+s-n)c_1 x = 0$$

The indicial equation is

$$(s+n)(s-n)c_0=0 \implies s=-n,n, \quad \text{since } c_0\neq 0$$

and the general recurrence relation is

$$c_m=-\frac{1}{(m+s+n)(m+s-n)}c_{m-2}, \quad m\geq 2.$$

When $s=n$, we have $(1+n-n)(1+n+n)c_1=0$ which implies $c_1=0$ if $n\neq -1/2$.
When $s=-n$, we have $(1-n-n)(1-n+n)c_1=0$ which implies $c_1=0$ if $n\neq 1/2$.

CaseI: When $n\neq 1/2$, then $c_1=0$. Then

$$\begin{aligned}
c_2 &= -\frac{1}{(2+s+n)(2+s-n)}c_0\\
c_4 &= \frac{1}{(2+s+n)(4+s+n)(2+s-n)(4+s-n)}c_0\\
c_6 &= -\frac{1}{(2+s+n)(4+s+n)(6+s+n)(2+s-n)(4+s-n)(6+s-n)}c_0\\
&\vdots\\
c_{2m} &= (-1)^m\frac{1}{(2+s+n)(4+s+n)\ldots(2m+s+n)(2+s-n)(4+s-n)\ldots(2m+s-n)}c_0
\end{aligned}$$

and $c_1=c_3=c_5=\cdots=c_{2m+1}=\cdots=0$. Thus, we get the solution as

$$y=c_0x^s\left[1-\frac{1}{(2+s+n)(2+s-n)}x^2+\cdots\right]$$

Putting $s=n$, we get

$$y=y_1=c_0x^n\left[1-\frac{x^2}{4(n+1)}+\frac{x^4}{4.8.(n+1)(n+2)}-\cdots\right]. \tag{10.1.1}$$

Putting $s=-n$, we get

$$y=y_2=c_0'x^{-n}\left[1-\frac{x^2}{4(1-n)}+\frac{x^4}{4.8.(1-n)(2-n)}-\cdots\right]. \tag{10.1.2}$$

The particular solution (10.1.1) of (10.0.1), taking

$$c_0 = \frac{1}{2^n \Gamma(n+1)}$$

is called the Bessel's function of first kind of order n and denoted by $J_n(x)$, that is,

$$J_n(x) = \frac{x^n}{2^n \Gamma(n+1)} \left[1 - \frac{x^2}{4(1-n)} + \frac{x^4}{4.8.(1-n)(2-n)} - \cdots \right].$$

The general term of $J_n(x)$ is, on simplification,

$$(-1)^r . \frac{1}{r!(n+1)\ldots(n+r).\Gamma(n+1)} \frac{x^{n+2r}}{2^{n+2r}}.$$

Now,

$$\Gamma(n+1) = n\Gamma(n).$$

Thus,

$$\Gamma(n+r+1) = (n+r)(n+r-1)\ldots(n+1)\Gamma(n+1).$$

Hence, the general term of the summation in $J_n(x)$ is

$$(-1)^r . \frac{x^{2r+n}}{2^{2r+n}\Gamma(r+1).\Gamma(n+r+1)}.$$

Thus, we can now formally define the Bessel's function as

Definition 10.1. The Bessel's function of first kind, of order n is defined as

$$J_n(x) = \sum_{r=0}^{\infty} (-1)^r . \frac{1}{\Gamma(r+1)\Gamma(n+r+1)} \left(\frac{x}{2}\right)^{2r+n}.$$

Similarly, the other solution is obtained by putting $-n$ for n, that is, the other solution, when n is not an integer is given by

$$J_{-n}(x) = \sum_{r=0}^{\infty} (-1)^r . \frac{1}{\Gamma(r+1)\Gamma(r-n+1)} \left(\frac{x}{2}\right)^{2r-n}.$$

CaseII: When $n = \pm 1/2$, then the Bessel's equation becomes

$$x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} + \left(x^2 - \frac{1}{4}\right) y = 0.$$

Let

$$y = \sum_{m=0}^{\infty} c_m x^{m+s}$$

be a solution of the above equation, where $c_0 \neq 0$ and s is to be determined. Then the above equation becomes

$$\begin{aligned}
\sum_{m=0}^{\infty}(m+s)(m+s-1)c_m x^{m+s} + \sum_{m=0}^{\infty}(m+s)c_m x^{m+s} + \sum_{m=0}^{\infty} c_m x^{m+s+2} - \frac{1}{4}\sum_{m=0}^{\infty} c_m x^{m+s} &= 0 \\
\text{or,} \quad \sum_{m=0}^{\infty}\left(m+s+\frac{1}{2}\right)\left(m+s-\frac{1}{2}\right)c_m x^m + \sum_{m=0}^{\infty} c_m x^{m+2} &= 0 \\
\text{or,} \quad \sum_{m=0}^{\infty}\left\{\left(m+s+\frac{1}{2}\right)\left(m+s-\frac{1}{2}\right)c_m + c_{m-2}\right\}x^m + \left(s+\frac{1}{2}\right)\left(s-\frac{1}{2}\right)c_0 + & \\
\left(s+\frac{3}{2}\right)\left(s+\frac{1}{2}\right)c_1 x &= 0
\end{aligned}$$

Thus, the indicial equation is

$$\left(s+\frac{1}{2}\right)\left(s-\frac{1}{2}\right)c_0 = 0 \implies s = \pm\frac{1}{2}, \text{ since } c_0 \neq 0.$$

When $s = 1/2$, $c_1 = 0$ and when $s = -1/2$, c_1 is indeterminate. Take c_1 as constant. Now, the general recurrence relation is

$$c_m = -\frac{1}{(m+s+1/2)(m+s-1/2)}c_{m-2}, \quad m \geq 2.$$

When $s = -1/2$, we have

$$c_m = -\frac{1}{m(m-1)}c_{m-2}, \quad m \geq 2.$$

Thus

$$\begin{aligned}
c_2 = -\frac{1}{2!}c_0, \quad c_4 = \frac{1}{4!}c_0, \quad c_6 = -\frac{1}{6!}c_0, \ldots, \\
c_3 = -\frac{1}{3!}c_1, \quad c_5 = \frac{1}{5!}c_1, \quad c_7 = -\frac{1}{7!}c_1, \ldots.
\end{aligned}$$

Thus, the solution becomes

$$y = c_0\left[x^{-1/2} - \frac{x^{3/2}}{2!} + \frac{x^{7/2}}{4!} - \cdots\right] + c_1\left[x^{1/2} - \frac{x^{5/2}}{3!} + \frac{x^{9/2}}{5!} - \cdots\right] = 0.$$

We have,

$$J_n(x) = \sum_{r=0}^{\infty}(-1)^r . \frac{1}{\Gamma(r+1)\Gamma(n+r+1)}\left(\frac{x}{2}\right)^{2r+n}.$$

Thus,

$$J_{1/2}(x) = \sum_{r=0}^{\infty}(-1)^r . \frac{1}{\Gamma(r+1)\Gamma(1/2+r+1)}\left(\frac{x}{2}\right)^{2r+1/2}.$$

Simplification of the above equation yields the same result as we have got on solving the Bessel's equation for $n = \pm 1/2$.

Note 10.2. The function $J_n(x)$ is one of the solutions of the Bessel's equation. Also, $J_{-n}(x)$ represents a solution of Bessel's equation which may or may not be independent of $J_n(x)$ always.

Theorem 10.3. If n is an integer, then

$$J_{-n}(x) = (-1)^n J_n(x).$$

This shows that $J_n(x)$ and $J_{-n}(x)$ do not provide with two independent solutions of Bessel's equations when n is an integer.

*Proof.*CaseI: When n is a positive integer, then

$$\begin{aligned} J_n(x) &= \sum_{r=0}^{\infty}(-1)^r.\frac{1}{\Gamma(r+1)\Gamma(n+r+1)}\left(\frac{x}{2}\right)^{2r+n}, \\ \text{and} \quad J_{-n}(x) &= \sum_{r=0}^{\infty}(-1)^r.\frac{1}{\Gamma(r+1)\Gamma(r-n+1)}\left(\frac{x}{2}\right)^{2r-n}. \end{aligned}$$

We know that $\Gamma(m) = \infty$ if $m = 0$ or a negative integer. Thus, $-n+r+1$ should be greater than zero, that is, $-n+r+1 \geq 1 \implies r \neq n$. So,

$$J_{-n}(x) = \sum_{r=n}^{\infty}(-1)^r.\frac{1}{r!\Gamma(r-n+1)}\left(\frac{x}{2}\right)^{2r-n}.$$

Put $m = r - n$ and eliminate r. Thus,

$$\begin{aligned} J_{-n}(x) &= \sum_{m=0}^{\infty}(-1)^{m+n}.\frac{1}{(m+n)!\Gamma(m+1)}\left(\frac{x}{2}\right)^{2m+n} \\ &= \sum_{m=0}^{\infty}(-1)^{m+n}.\frac{1}{\Gamma(m+n+1)\Gamma(m+1)}\left(\frac{x}{2}\right)^{2m+n} \\ &= (-1)^n\sum_{m=0}^{\infty}(-1)^m.\frac{1}{\Gamma(m+n+1)\Gamma(m+1)}\left(\frac{x}{2}\right)^{2m+n} = (-1)^n J_n(x). \end{aligned}$$

CaseII: When n is a negative integer. Let $n = -p$, where p is a positive integer. Then,

$$\begin{aligned} J_{-p}(x) &= (-1)^p J_p(x), \quad \text{[by CaseI]} \\ \text{or,} \quad J_n(x) &= (-1)^{-n} J_{-n}(x) \\ \text{or,} \quad J_{-n}(x) &= (-1)^n J_n(x). \end{aligned}$$

□

Note 10.4. When n is an integer, $J_{-n}(x)$ is not independent of $J_n(x)$. Hence $y = AJ_n(x) + BJ_{-n}(x)$ is not a general solution of (10.0.1) when n is an integer. But when n is non-integral, then the general solution is given by $y = AJ_n(x) + BJ_{-n}(x)$, for arbitrary constants A and B.

We will investigate the general solution for Bessel's equation for integral n. We have the theorem below in this direction.

Theorem 10.5. The two linearly independent solutions of (10.0.1) may be taken to be two functions taken as $y_1(x) = J_n(x)$ and

$$y_2(x) = \lim_{\nu \to n}\frac{\cos(\nu\pi)J_\nu(x) - J_{-\nu}(x)}{\sin(\nu\pi)} = Y_n(x).$$

*Proof.*CaseI: When n is not an integer. Since n is not an integer, so $\sin n\pi \neq 0$. Hence

$$Y_n(x) = \cot(n\pi)J_n(x) - \text{cosec}(n\pi)J_{-n}(x),$$

that is, Y_n is a linear combination of $J_n(x)$ and $J_{-n}(x)$. But we know that J_n and J_{-n} are independent solutions of Bessel's equation, when n is not an integer, that is,

$$W(J_n(x), J_{-n}(x)) = \begin{vmatrix} J_n & J_{-n} \\ J_n' & J_{-n}' \end{vmatrix} \neq 0.$$

Now, on simplifying, we get

$$W(J_n, Y_n) = \begin{vmatrix} J_n & Y_n \\ J_n' & Y_n' \end{vmatrix} = -\text{cosec}(n\pi)W(J_n(x), J_{-n}(x)) \neq 0.$$

Thus, J_n and Y_n are two independent solutions of Bessel's equation of order n.

CaseII: Let n be an integer. Then $\sin(n\pi) = 0$ and $\cos(n\pi) = (-1)^n$ and also $J_{-n}(x) = (-1)^n J_n(x)$. First, we deduce a simplified form of $Y_n(x)$.

$$\begin{aligned} Y_n(x) &= \lim_{\nu \to n} \frac{\cos(\nu\pi)J_\nu(x) - J_{-\nu}(x)}{\sin(\nu\pi)} \left(\frac{0}{0}\right) \\ &= \frac{-\pi \sin(n\pi)J_n(x) + \cos(n\pi)\left[\frac{\partial}{\partial \nu}J_\nu(x)\right]_{\nu=n} - \left[\frac{\partial}{\partial \nu}J_{-\nu}(x)\right]_{\nu=n}}{\pi\cos(n\pi)} \\ &= \frac{1}{\pi}\left[\frac{\partial}{\partial \nu}J_\nu(x) - (-1)^n \frac{\partial}{\partial \nu}J_{-\nu}(x)\right]_{\nu=n}. \end{aligned} \tag{10.1.3}$$

We now establish the following two results for $Y_n(x)$.

1. $Y_n(x)$ is a solution of Bessel's equation.

   Proof. $J_\nu(x)$ and $J_{-\nu}(x)$ are solutions of Bessel's equation of order ν. Thus,

   $$x^2\frac{d^2}{dx^2}J_\nu + x\frac{d}{dx}J_\nu + (x^2 - \nu^2)J_\nu = 0 \tag{10.1.4}$$

   $$x^2\frac{d^2}{dx^2}J_{-\nu} + x\frac{d}{dx}J_{-\nu} + (x^2 - \nu^2)J_{-\nu} = 0 \tag{10.1.5}$$

   Differentiating (10.1.4) and (10.1.5), with respect to ν we get,

   $$x^2\frac{d^2}{dx^2}\left(\frac{\partial}{\partial \nu}J_\nu\right) + x\frac{d}{dx}\left(\frac{\partial}{\partial \nu}J_\nu\right) + (x^2 - \nu^2)\left(\frac{\partial}{\partial \nu}J_\nu\right) - 2\nu J_\nu = 0 \tag{10.1.6}$$

   $$x^2\frac{d^2}{dx^2}\left(\frac{\partial}{\partial \nu}J_{-\nu}\right) + x\frac{d}{dx}\left(\frac{\partial}{\partial \nu}J_{-\nu}\right) + (x^2 - \nu^2)\left(\frac{\partial}{\partial \nu}J_{-\nu}\right) - 2\nu J_{-\nu} = 0 \tag{10.1.7}$$

   By (10.1.6) $-(-1)^\nu$(10.1.7), we get

   $$x^2\frac{d^2}{dx^2}\left[\frac{\partial}{\partial \nu}J_\nu - (-1)^\nu\frac{\partial}{\partial \nu}J_{-\nu}\right] + x\frac{d}{dx}\left[\frac{\partial}{\partial \nu}J_\nu - (-1)^\nu\frac{\partial}{\partial \nu}J_{-\nu}\right] + (x^2 - \nu^2)\left[\frac{\partial}{\partial \nu}J_\nu - (-1)^\nu\frac{\partial}{\partial \nu}J_{-\nu}\right] - 2\nu[J_\nu - (-1)^\nu J_{-\nu}] = 0.$$

At $\nu = n$, the above equation becomes

$$\left[x^2\frac{d^2}{dx^2} + x\frac{d}{dx} + (x^2 - n^2)\right]\left[\frac{\partial}{\partial\nu}J_\nu - (-1)^\nu\frac{\partial}{\partial\nu}J_{-\nu}\right]_{\nu=n} - 2n(J_n - (-1)^n J_{-n}) = 0$$

$$\text{or,} \quad \left[x^2\frac{d^2}{dx^2} + x\frac{d}{dx} + (x^2 - n^2)\right]\left[\frac{\partial}{\partial\nu}J_\nu - (-1)^\nu\frac{\partial}{\partial\nu}J_{-\nu}\right]_{\nu=n} = 0$$

$$\text{or,} \quad \left[x^2\frac{d^2}{dx^2} + x\frac{d}{dx} + (x^2 - n^2)\right]Y_n(x) = 0.$$

Thus, Y_n is a solution of Bessel's equation of order n. □

2. $Y_n(x)$ is independent of $J_n(x)$.

   This is evident from the structure of Y_n and J_n.

□

Definition 10.6. Bessel's function of second kind of order n, denoted by $Y_n(x)$ is defined as

$$\begin{aligned} Y_n(x) &= \lim_{\nu\to n}\frac{\cos(\nu\pi)J_\nu(x) - J_{-\nu}(x)}{\sin(\nu\pi)}, \text{ when n is an integer} \\ &= \frac{J_n(x)\cos(n\pi) - J_{-n}(x)}{\sin(n\pi)}, \text{ when n is not an integer.} \end{aligned}$$

Thus, the general solution of Bessel's equation is

$$y(x) = C_1 J_n(x) + C_2 Y_n(x)$$

where C_1 and C_2 are independent constants.

10.2 Recurrence Relations for Bessel's Equations

For integral n, the Bessel's function satisfies the following recurrence relations:

1. $xJ_n'(x) = nJ_n(x) - xJ_{n+1}(x)$.

   Proof. We have

   $$J_n(x) = \sum_{r=0}^{\infty}(-1)^r\frac{1}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r}.$$

   Differentiating with respect to x, we get

   $$\begin{aligned} J_n'(x) &= \frac{1}{2}\sum_{r=0}^{\infty}(-1)^r\frac{(n+2r)}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r-1} \\ \text{or,} \quad xJ_n'(x) &= \sum_{r=0}^{\infty}(-1)^r\frac{(n+2r)}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r} \\ &= n\sum_{r=0}^{\infty}(-1)^r\frac{1}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r} + 2\sum_{r=0}^{\infty}(-1)^r\frac{r}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r} \\ &= nJ_n(x) - 2\sum_{s=0}^{\infty}(-1)^s\frac{1}{s!\Gamma(n+s+2)}\left(\frac{x}{2}\right)^{n+2s+2} \quad \text{[Putting r-1=s and eliminating r]} \\ &= nJ_n(x) - x\frac{1}{\Gamma(s+1)\Gamma(n+s+2)}\left(\frac{x}{2}\right)^{n+1+2s} = nJ_n(x) - xJ_{n+1}(x). \end{aligned}$$

□

2. $xJ_n^{'}(x) = xJ_{n-1}(x) - nJ_n(x)$.

Proof. We have

$$J_n(x) = \sum_{r=0}^{\infty}(-1)^r \frac{1}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r}.$$

Differentiating with respect to x, we get

$$\begin{aligned} J_n^{'}(x) &= \frac{1}{2}\sum_{r=0}^{\infty}(-1)^r \frac{(n+2r)}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r-1} \\ \text{or, } 2J_n^{'}(x) &= \sum_{r=0}^{\infty}(-1)^r \frac{2(n+2r)}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r-1} - \sum_{r=0}^{\infty}(-1)^r \frac{n}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r-1} \\ &= 2\sum_{r=0}^{\infty}(-1)^r \frac{1}{r!\Gamma(r+n)}\left(\frac{x}{2}\right)^{n-1+2r} - n\sum_{r=0}^{\infty}(-1)^r \frac{1}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r-1} \\ &= 2J_{n-1}(x) - \frac{2n}{x}\sum_{r=0}^{\infty}(-1)^r \frac{1}{r!\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r} = 2J_{n-1}(x) - \frac{2n}{x}J_n(x). \end{aligned}$$

Simplifying, we get the desired result. □

3. $\frac{2n}{x}J_n(x) = J_{n-1}(x) + J_{n+1}(x)$.

Proof. From 1, we have

$$xJ_n^{'}(x) = nJ_n(x) - xJ_{n+1}(x).$$

From 2, we have

$$xJ_n^{'}(x) = xJ_{n-1}(x) - nJ_n(x).$$

Subtracting and simplifying, we get the desired result. □

4. $2J_n^{'}(x) = J_{n-1}(x) - J_{n+1}(x)$.

Proof. From 1, we have

$$xJ_n^{'}(x) = nJ_n(x) - xJ_{n+1}(x).$$

From 2, we have

$$xJ_n^{'}(x) = xJ_{n-1}(x) - nJ_n(x).$$

Adding and simplifying, we get the desired result. □

Exercise 10.7. 1. Prove that for integral n, $4J_n^{"} = J_{n-2} - 2J_n + J_{n+2}$.

2. Prove the following:

(a)

$$\frac{d}{dx}\left(x^n J_n(x)\right) = x^n J_{n-1}(x).$$

(b)

$$\frac{d}{dx}\left(x^{-n} J_n(x)\right) = -x^{-n} J_{n+1}(x).$$

10.3 Generating Function for Bessel's Functions

For all values of x and for all values of z such that $0 < |z| < \infty$, the function

$$w(x,z) = \mathrm{e}^{\frac{x}{2}\left\{z-\frac{1}{z}\right\}},$$

generates the Bessel's function of **integral** order n, that is,

$$\mathrm{e}^{\frac{x}{2}\left\{z-\frac{1}{z}\right\}} = \sum_{n=-\infty}^{\infty} J_n(x) z^n.$$

Proof. We have,

$$\begin{aligned}\mathrm{e}^{\frac{x}{2}\left\{z-\frac{1}{z}\right\}} = \mathrm{e}^{xz/2}\,.\,\mathrm{e}^{-x/(2z)} &= \left\{\sum_{r=0}^{\infty}\left(\frac{xz}{2}\right)^r \frac{1}{r!}\right\}\left\{\sum_{s=0}^{\infty}(-1)^s\left(\frac{x}{2z}\right)^s \frac{1}{s!}\right\} \\ &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}(-1)^s \frac{1}{\Gamma(r+1)\Gamma(s+1)}\left(\frac{x}{2}\right)^{r+s} z^{r-s}. \qquad (10.3.1)\end{aligned}$$

CaseI: When $r-s \geq 0$, let $r-s=n$. Then $n \geq 0$, that is, n is an integer varying from 0 to ∞. Eliminating s, we get

$$\sum_{n=0}^{\infty}\sum_{r=n}^{\infty}(-1)^{r-n}\frac{1}{\Gamma(r+1)\Gamma(r-n+1)}\left(\frac{x}{2}\right)^{2r-n} z^n.$$

[For real values of $\Gamma(r+1)$ we must have $r+1 \geq 1$, that is, $r \geq 0$ and for real values of $\Gamma(r-n+1)$ we must similarly have $r \geq n$.] Now, let

$$w(x,z) = \sum_{n=0}^{\infty} f_n(x) z^n, \quad \text{where } f_n(x) = \sum_{r=n}^{\infty}(-1)^{r-n}\frac{1}{\Gamma(r+1)\Gamma(r-n+1)}\left(\frac{x}{2}\right)^{2r-n}.$$

Putting $r-n=p$ and eliminating r, we get

$$f_n(x) = \sum_{p=0}^{\infty}(-1)^p \frac{1}{\Gamma(p+1)\Gamma(p+n+1)}\left(\frac{x}{2}\right)^{2p+n} = J_n(x). \qquad (10.3.2)$$

CaseII: When $r-s<0$, let $r-s=-n_1$, where n_1 is a positive integer. So n_1 takes values from 1 to ∞. Eliminating s, we get,

$$w(x,z) = \sum_{n_1=1}^{\infty}\sum_{r=0}^{\infty}(-1)^{r+n_1}\frac{1}{\Gamma(r+1)\Gamma(r+n_1+1)}\left(\frac{x}{2}\right)^{2r+n_1} z^{-n_1}.$$

[For real values of $\Gamma(r+1)$ we must have $r+1 \geq 1$, that is, $r \geq 0$ and for real values of $\Gamma(r+n_1+1)$ we must similarly have $r \geq -n_1$. Both inequalities simultaneously give $r \geq 0$.] We assume

$$w(x,z) = \sum_{n_1=1}^{\infty} f_{n_1}(x) z^{-n_1},$$

where

$$f_{n_1}(x) = \sum_{r=0}^{\infty}(-1)^{r+n_1}\frac{1}{\Gamma(r+1)\Gamma(r+n_1+1)}\left(\frac{x}{2}\right)^{2r+n_1} = (-1)^{n_1} J_{n_1}(x).$$

Thus,

$$\begin{aligned} w(x,z) &= \sum_{n_1=1}^{\infty}(-1)^{n_1}J_{n_1}(x)z^{-n_1} \\ &= \sum_{n=-\infty}^{-1}(-1)^{-n}J_{-n}(x)z^{n} \\ &= \sum_{n=-\infty}^{-1}(-1)^{-n}(-1)^{n}J_{n}(x)z^{n} = \sum_{n=-\infty}^{}J_{n}(x)z^{n}. \end{aligned} \tag{10.3.3}$$

Combining (10.3.2) and (10.3.3), we get the desired result.

□

10.4 Orthogonality Conditions

If c_i and c_j are the roots of the equation $J_n(ca) = 0$, then

$$\begin{aligned} \int_0^a xJ_n(c_ix)J_n(c_jx)dx &= 0, \quad i \neq j \\ &= \frac{a^2}{a}J_{n+1}^2(c_ia), \quad i = j. \end{aligned}$$

To prove the above, we first write the Bessel's equation of order n.

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - n^2)y = 0.$$

The general solution is

$$y(x) = AJ_n(x) + BY_n(x).$$

We first show that $J_n(cx)$ satisfies the following equation, known as the **modified Bessel's equation**,

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} + (c^2x^2 - n^2)y = 0.$$

Put $z = cx$. Then we get

$$\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx} = c\frac{dy}{dz}, \qquad \frac{d^2y}{dx^2} = \frac{d}{dx}\left(\frac{dy}{dz}.c\right) = \frac{d}{dz}\left(\frac{dy}{dz}.c\right)\frac{dz}{dx} = c^2\frac{d^2y}{dz^2}.$$

Using these in the modified Bessel's equation, we get

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} + (z^2 - n^2)y = 0.$$

This is Bessel's equation in the variable z whose general solution is

$$y(z) = AJ_n(z) + BY_n(z).$$

Thus $J_n(z)$, that is, $J_n(cx)$ is the solution of the modified Bessel's equation. We now move on to prove the orthogonality condition.

*Proof.*CaseI: When $i \neq j$. c_i and c_j are the roots of $J_n(ca) = 0$, that is,

$$J_n(c_i a) = 0, \qquad J_n(c_j a) = 0.$$

We know that $J_n(cx)$ satisfies the modified Bessel's equation. Thus,

$$x^2 \frac{d^2}{dx^2} J_n(cx) + x \frac{d}{dx} J_n(cx) + (c^2 x^2 - n^2) J_n(cx) = 0.$$

Thus,

$$x^2 \frac{d^2}{dx^2} J_n(c_i x) + x \frac{d}{dx} J_n(c_i x) + (c_i^2 x^2 - n^2) J_n(c_i x) = 0 \tag{10.4.1}$$

$$x^2 \frac{d^2}{dx^2} J_n(c_j x) + x \frac{d}{dx} J_n(c_j x) + (c_j^2 x^2 - n^2) J_n(c_j x) = 0 \tag{10.4.2}$$

Putting $u = J_n(c_i x)$ and $v = J_n(c_j x)$, we get from (10.4.1) and (10.4.2),

$$x^2 \frac{d^2 u}{dx^2} + x \frac{du}{dx} + (c_i^2 x^2 - n^2) u = 0 \tag{10.4.3}$$

$$x^2 \frac{d^2 v}{dx^2} + x \frac{dv}{dx} + (c_j^2 x^2 - n^2) v = 0 \tag{10.4.4}$$

Now, (10.4.3)$\times v -$(10.4.4)$\times u$ gives on simplification

$$\frac{d}{dx}\left\{ x \left(\frac{du}{dx} v - \frac{dv}{dx} u \right) \right\} + (c_i^2 - c_j^2) xuv = 0.$$

Integrating the above equation with respect to x from 0 to a, we get

$$\begin{aligned}
(c_j^2 - c_i^2) \int_0^a xuv\,dx &= \left[x \left(\frac{du}{dx} v - \frac{dv}{dx} u \right) \right]_0^a \\
&= \left[x \left(v \frac{d}{dx} u - u \frac{d}{dx} v \right) \right]_0^a \\
&= a J_n(c_j a) \left[\frac{d}{dx} J_n(c_i x) \right]_{x=a} - a J_n(c_i a) \left[\frac{d}{dx} J_n(c_j x) \right]_{x=a} = 0
\end{aligned}$$

[since c_i and c_j are the roots of the equation $J_n(ca) = 0$]. Hence the result.

CaseII: When $i = j$. Multiplying (10.4.3) by $2\frac{du}{dx}$, we get on simplifying,

$$\frac{d}{dx}\left[\left(x^2 \left(\frac{du}{dx} \right)^2 \right) - n^2 u^2 + c_i^2 x^2 u^2 \right] - 2 c_i^2 x u^2 = 0.$$

Integrating the above equation with respect to x from 0 to a, we get

$$\begin{aligned}
2c_i^2 \int_0^a x u^2\,dx &= \left[\left(x^2 \left(\frac{du}{dx} \right)^2 \right) - n^2 u^2 + c_i^2 x^2 u^2 \right]_0^a \\
&= \left[\left(x^2 \left(\frac{d}{dx} J_n(c_i x) \right)^2 \right) - n^2 [J_n(c_i x)]^2 + c_i^2 x^2 [J_n(c_i x)]^2 \right]_0^a \\
&= a^2 \left[\frac{d}{dx} J_n(c_i x) \right]_{x=a} - n^2 [J_n(c_i a)]^2 + c_i^2 a^2 [J_n(c_i a)]^2 + n^2 (J_n(0))^2 \\
&= a^2 \left[\frac{d}{dx} J_n(c_i x) \right]_{x=a} + n^2 (J_n(0))^2 = a^2 \left[\frac{d}{dx} J_n(c_i x) \right]_{x=a}.
\end{aligned} \tag{10.4.5}$$

Recurrence relation 1 gives

$$xJ_n^{'}(x) = nJ_n(x) - xJ_{n+1}(x).$$

Now, put $x = c_i y$. Then $dx = c_i dy$. Thus,

$$\begin{aligned} c_i y \frac{1}{c_i}\frac{d}{dy}J_n(c_i y) &= nJ_n(c_i y) - c_i y J_{n+1}(c_i y) \\ \text{or,}\quad \frac{d}{dy}J_n(c_i y) &= \frac{n}{y}J_n(c_i y) - c_i J_{n+1}(c_i y) \\ \text{or,}\quad \frac{d}{dx}J_n(c_i x) &= \frac{n}{x}J_n(c_i x) - c_i J_{n+1}(c_i x). \end{aligned}$$

Putting $x = a$ on both sides of the last equation, we get

$$\left[\frac{d}{dx}J_n(c_i x)\right]_{x=a} = \frac{n}{a}J_n(c_i a) - c_i J_{n+1}(c_i a) = -c_i J_{n+1}(c_i a).$$

Thus, (10.4.5) gives

$$2c_i^2 \int_0^a x[J_n(c_i x)]^2 dx = a^2 c_i^2 [J_{n+1}(c_i a)]^2.$$

Simplifying, we get the desired result.

□

Example 10.8. Show that

$$J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x \quad \text{and} \quad J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x.$$

We have,

$$J_n(x) = \sum_{r=0}^{\infty}(-1)^r \frac{1}{\Gamma(r+1)\Gamma(r+n+1)}\left(\frac{x}{2}\right)^{n+2r}.$$

Thus,

$$\begin{aligned} J_{-1/2}(x) &= \sum_{r=0}^{\infty}(-1)^r \frac{1}{\Gamma(r+1)\Gamma(r+1/2)}\left(\frac{x}{2}\right)^{-1/2+2r} \\ &= \sqrt{\frac{2}{\pi x}}\left[1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots\right] = \sqrt{\frac{2}{\pi x}}\cos x. \end{aligned}$$

We can similarly prove the other part by expanding the Bessel's function $J_{1/2}(x)$.

Example 10.9. Show that

$$\int_0^{\pi/2} \sqrt{\pi x} J_{1/2}(2x) dx = 1.$$

We have,

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x.$$

Thus,

$$J_{1/2}(2x) = \sqrt{\frac{1}{\pi x}}\sin 2x.$$

Integrating the above with respect to x from 0 to $\pi/2$, we get the required result.

Exercise 10.10. 1. Prove that

$$J_{-3/2}(x) = \sqrt{\frac{2}{\pi x}}\left(-\frac{\cos x}{x} - \sin x\right) \quad \text{and} \quad J_{3/2}(x) = \sqrt{\frac{2}{\pi x}}\left(\frac{\sin x}{x} - \cos x\right).$$

2. Express J_3 and $J_4(x)$ in terms of J_0 and J_1.

3. Prove that $J_0^{'} = -J_1$.

4. For $n > 1$ show that

$$\int_0^x x^{n+1} J_n(x)dx = x^{n+1} J_{n+1}(x)dx.$$

5. Prove that

   (a)
   $$\frac{d}{dx}(xJ_1(x)) = xJ_0(x).$$

   (b)
   $$\int_0^b xJ_0(ax)dx = \frac{b}{a}J_1(ab).$$

10.5 Few Probable Questions

1. Deduce a generating function for Bessel's functions.
2. Deduce the orthogonality of Bessel's functions.
3. Define Bessel's function of second kind. Show that the Bessel's function of second kind is a solution of Bessel's equations.
4. Prove that $xJ_n^{'}(x) = nJ_n(x) - xJ_{n+1}(x)$. Hence show that

   $$\frac{d}{dx}J_0(x) = -J_1(x)$$

   and
   $$\int_a^b J_0(x)J_1(x)dx = \frac{1}{2}[(J_0(a))^2 - (J_0(b))^2].$$

References

1. Special Functions of Mathematical Physics and Chemistry; I. N. Sneddon.
2. Special Functions & Their Applications; Lebedev & Silverman
3. Special Functions; E. D. Rainville.

Core Paper

MATC 3.1

Block - III

Marks : 35 (SSE : 30; IA : 05)

Integral Equations & Integral Transforms

Syllabus

• Unit 11 •

Integral Equations. Definitions of integral equations and their classification. Fredholm integral equations of second kind : Resolvent kernel, solution in terms of resolvent kernel, solution with separable kernels, Method of successive approximations, iterative scheme for Fredholm integral equations. Volterra integral equations of second kind : Solution by successive approximations. Resolvent kernel and solutions of Volterra integral equations.

• Unit 12 •

Classical Fredholm theory : Fredholm theorems, Fredholm Alternative Principles. Hilbert-Schmidt theory : Symmetric kernels, Orthogonal system of functions, Fundamental properties of eigenvalues and eigenfunctions for symmetric kernels, Hilbert-Schmidt theorem.

• Unit 13 •

Integral Transforms. Laplace Transform : Definition and basic properties. Laplace integral. Lerch's theorem (statement only). Laplace transforms of elementary functions, of derivatives and Dirac-delta function. Differentiation and integration. Convolution. Inverse transform. Applications to solve ordinary differential equations.

• Unit 14 •

Fourier Transform : Definition and basic properties. Fourier transform of some elementary functions, of derivatives. Inverse Fourier transform. Convolution theorem and Parseval's relation. Applications of Fourier inversion and convolution theorems. Fourier sine and cosine transforms.

• Unit 15 •

Hankel Transform : Definition and inversion formula. Hankel transform of derivatives. Finite Hankel transform.

• Unit 16 •

Applications : Applications of integral transforms to solve two-dimensional Laplace and one dimensional diffusion and wave equations.

Unit 11

Course Structure

Integral Equations: Definitions of integral equations and their classification. Fredholm integral equations of second kind : Resolvent kernel, solution in terms of resolvent kernel, solution with separable kernels, Method of successive approximations, iterative scheme for Fredholm integral equations. Volterra integral equations of second kind : Solution by successive approximations. Resolvent kernel and solutions of Volterra integral equations.

---

11 Introduction

Many physical problems of science and technology which were solved with the help of theory of ordinary and partial differential equations can be solved by better methods of theory of integral equations. For example, while searching for the representation formula for the solution of linear differential equation in such a manner so as to include boundary conditions or intitial conditions explicitly, we arrive at an integral equation. The solution of the integral equation is much easier than the orginal boundary value or initial value problem. The theory of integral equations is very useful tool to deal with problems in applied mathematics, theoretical mechanis, and mathematical physics. Several situations of science lead to integral equations, e.g., neutron diffusion problem and radiation transfer problem etc.

Objectives

The objective of this course is to learn the students all of the above topics and by the end of it students should be able to
(1) know different kinds of kernels and techniques for solving each kind.
(2) know number of numerical methods for solving integral equations.
(3) know the relation between differential and integral equations, and how to change from one to another.
(4) know basic theory of calculus of variations and see some applications.

11.1 Integral Equation

Definition 11.1. An integral equation is an equation is which an unknown function appears under one or more integral signs. For example, for $a \leq x \leq b,\ a \leq t \leq b$, the equations

$$\int_a^b K(x,t)y(t)dt = f(x) \tag{11.1.1}$$

$$y(x) - \lambda \int_a^b K(x,t)y(t)dt = f(x) \tag{11.1.2}$$

$$\text{and} \quad y(x) = \int_a^b K(x,t)[y(t)]^2 dt, \tag{11.1.3}$$

where the function $y(x)$, is the unknown function while the functions $f(x)$ and $K(x,t)$ are known functions and λ, a and b are constants, are all integral equations. The above mentioned functions may be complex-valued functions of the real variables x and t.

Definition 11.2. Linear and Non-linear Integral Equation : An integral equation is called *linear* if only linear operations are performed in it upon the unknown function. An integral equation which is not linear is known as a *non-linear integral equation*. By writing either

$$L(y) = \int_a^b K(x,t)\, y(t)\, dt \quad \text{or} \quad L(y) = y(x) - \lambda \int_a^b K(x,t)y(t)dt, \tag{11.1.4}$$

we can easily verify that L is a linear integral operator. In fact, for any constants c_1 and c_2, we have

$$L\{c_1y_1(x) + c_2y_2(x)\} = c_1L\{y_1(x)\} + c_2L\{y_2(x)\} \tag{11.1.5}$$

which is well known general criterion for a *linear operator*. In this block, we shall study only linear integral equations. The most general type of linear integral equation is of the form

$$g(x)y(x) = f(x) + \lambda \int_a K(x,t)y(t)dt, \tag{11.1.6}$$

where the upper limit may be either variable x or fixed. The functions f, g and K are known functions while y is to be determined; λ is a non-zero real or complex, *parameter*. The function $K(x,t)$ is known as the kernel of the integral equation.

Remark 11.3. The constant λ can be incorporated into the kernel $K(x,t)$ in Eq.(11.1.6). However, in many applications λ represents a significant parameter which may take on various values in a discussion being considered.

Remark 11.4. If $g(x) \neq 0$, Eq.(11.1.6) is known as linear integral equation of the third kind. When $g(x) = 0$, Eq.(11.1.6) reduces to

$$f(x) + \lambda \int_a K(x,t)y(t)dt, = 0 \tag{11.1.7}$$

which is known as *linear integral equation of the first kind*. Again, when $g(x) = 1$, Eq.(11.1.6) reduces to

$$y(x) = f(x) + \lambda \int_a K(x,t)y(t)dt. \tag{11.1.8}$$

which is known as linear integral equation of the second kind. In the present block, we shall study in details equations of the form (11.1.7) and (11.1.8) only.

Definition 11.5. Fredholm Integral Equation : A linear integral equation of the form

$$g(x)y(x) = f(x) + \lambda \int_a^b K(x,t)y(t)dt, \tag{11.1.9}$$

where a, b are both constants, $f(x)g(x)$ and $K(x,t)$ are known functions while $y(x)$ is unknown function and λ is a non-zero real or complex parameter, is called *Fredholm integral equation of third kind*. The function $K(x,t)$ is known as the kernel of the integral equation.

• Setting $g(x) = 0$ in Eq.(11.1.9), we have the *Fredholm integral equation of the first kind*.

• Setting $g(x) = 1$ in Eq.(11.1.9), we have the *Fredholm integral equation of the second kind.*

• Setting $g(x) = 1$ and $f(x) = 0$ in Eq.(11.1.9), we have the *Homogeneous Fredholm integral equation of the second kind.*

11.2 Solution of Fredholm integral equations

11.2.1 Method of Successive approximations :

Consider the Fredholm integral equation of the second kind given by Eq.(11.2.4). As a zero-order approximation to the required solution $y(x)$, let us take $y_0(x) = f(x)$. Further, if $y_n(x)$ and $y_{n-1}(x)$ are the n-th and $(n-1)$-th order approximations respectively, then these are connected by

$$y_n(x) = f(x) + \lambda \int_a^b K(x,t) y_{n-1}(t) dt. \tag{11.2.1}$$

We know that the iterated kernels (or iterated functions) $K_n(x,t), (n = 1, 2, 3, \ldots)$ are defined by

$$\begin{aligned} K_1(x,t) &= K(x,t) \\ \text{and} \quad K_n(x,t) &= \int_a^b K(x,z) K_{n-1}(z,t) dz. \end{aligned}$$

Putting $n = 1$ in Eq.(11.2.1), the first-order approximation $y_1(x)$ is given by

$$\begin{aligned} y_1(x) &= f(x) + \lambda \int_a^b K(x,t) y_0(t) dt. \\ \Rightarrow y_1(x) &= f(x) + \lambda \int_a^b K(x,t) f(t) dt. \end{aligned} \tag{11.2.2}$$

Putting $n = 2$ in Eq.(11.2.1), the second-order approximation $y_2(x)$ is given by

$$\begin{aligned} y_2(x) &= f(x) + \lambda \int_a^b K(x,t) y_1(t)\, dt. \\ \Rightarrow y_2(x) &= f(x) + \lambda \int_a^b K(x,z) y_1(z)\, dz \\ \Rightarrow y_2(x) &= f(x) + \lambda \int_a^b K(x,z) \left[f(z) + \lambda \int_a^b K(z,t)\, f(t)\, dt \right] dz \\ \Rightarrow y_2(x) &= f(x) + \lambda \int_a^b K(x,t)\, f(t)\, dt + \lambda^2 \int_a^b f(t) \left[\int_a^b K(x,z) K(z,t)\, dz \right] dt \\ \Rightarrow y_2(x) &= f(x) + \lambda \int_a^b K_1(x,t)\, f(t)\, dt + \lambda^2 \int_a^b K_2(x,t) f(t)\, dt \end{aligned}$$

$$\Rightarrow y_2(x) = f(x) + \sum_{m=1}^{2} \lambda^m \int_a^b K_m(x,t) f(t)\, dt.$$

Proceeding likewise, we easily obtain by Mathematical induction the n-th approximate solution $y_n(x)$ as

$$y_n(x) = f(x) + \sum_{m=1}^{n} \lambda^m \int_a^b K_m(x,t)\, f(t)\, dt. \tag{11.2.3}$$

11.2.2 Resolvent kernel :

Suppose solution of Fredholm integral equation of the second kind

$$y(x) = f(x) + \lambda \int_a^b K(x,t)y(t)dt \tag{11.2.4}$$

takes the form

$$y(x) = f(x) + \lambda \int_a^b R(x,t;\lambda)f(t)dt, \tag{11.2.5}$$

then $R(x,t;\lambda)$ is known as the resolvent kernal of (11.2.4). If $K_n(x,t)$ be iterated kernals then

$$R(x,t;\lambda) = \sum_{m=1}^{\infty} \lambda^{m-1} K_m(x,t)$$

Example 11.6. Find the iterated kernels for the following kernels

$$K(x,t) = \sin(x-t),\ 0 \le x \le 2\pi,\ 0 \le t \le 2\pi$$

Solution : Iterated kernel $K_n(x,t)$ are given by

$$K_1(x,t) \quad = \quad K(x,t) \tag{11.2.6}$$

$$\text{and} \quad K_n(x,t) \quad = \quad \int_0^{2\pi} K(x,z)K_{n-1}(z,t)\,dz, \quad (n = 2,3,\ldots) \tag{11.2.7}$$

From Eq.(11.2.6) $K_1(x,t) = K(x,t) = \sin(x-2t)$. Putting $n = 2$ in Eq.(11.2.7), we have

$$\begin{aligned} K_2(x,t) &= \int_0^{2\pi} K(x,z)K_1(z,t)\,dz = \int_0^{2\pi} \sin(x-2z)\sin(z-2t)\,dz \\ &= \frac{1}{2}\int_0^{2\pi} [\cos(x+2t-3z) - \cos(x-2t-z)]\,dz = \frac{1}{2}\left[-\frac{1}{3}\sin(x+2t-3z) + \sin(x-2t-z\right] \\ &= 0, \quad \text{on simplification.} \end{aligned}$$

Putting $n = 3$ in Eq.(11.2.7), we have

$$K_3(x,t) = \int_0^{2\pi} K(x,z)K_2(z,t)\,dz = 0 \quad [\because K_2(z,t) = 0]$$

Thus, $K_1(x,t) = \sin(x-2t)$ and $K_n(x,t) = 0$ for $n = 2,3,4,\ldots$.

Example 11.7. Determine the resolvent kernels for the Fredholm integral equation having kernels

$$K(x,t) = e^{x+t};\ a = 0,\ b = 1.$$

Solution : Iterated kernels $K_m(x,t)$ are given by

$$K_1(x,t) \quad = \quad K(x,t) \tag{11.2.8}$$

$$K_m(x,t) \quad = \quad \int_0^1 K(x,z)K_{m-1}(z,t)\,dz \tag{11.2.9}$$

From Eq.(11.2.8) $K_1(x,t) = K(x,t) = e^{x+t}$.

Putting $n = 2$ in Eq.(11.2.9), we have

$$\begin{aligned} K_2(x,t) &= \int_0^1 K(x,z)K_1(z,t)\,dz = \int_0^1 e^{x+z}e^{z+t}\,dz \\ &= e^{x+t}\int_0^1 e^{2z}\,dz \;=\; e^{x+t}\left[\frac{1}{2}e^{2z}\right] = e^{x+t}\left(\frac{e^2-1}{2}\right), \end{aligned}$$

Putting $n = 3$ in Eq.(11.2.9), we have

$$\begin{aligned} K_3(x,t) &= \int_0^1 K(x,z)K_2(z,t)\,dz = \int_0^1 e^{x+z}e^{z+t}\left(\frac{e^2-1}{2}\right)\,dz \\ &= e^{x+t}\left(\frac{e^2-1}{2}\right)\int_0^1 e^{2z}\,dz \;=\; e^{x+t}\left[\frac{1}{2}e^{2z}\right] = e^{x+t}\left(\frac{e^2-1}{2}\right)^2 \quad \text{and so on,} \end{aligned}$$

Observing above, we may write

$$K_m(x,t) = e^{x+t}\left(\frac{e^2-1}{2}\right)^{m-1}, \quad m = 1,2,3,\ldots$$

Now, the required resolvent kernel is given by

$$R(x,t;\lambda) = \sum_{m=1}^{\infty}\lambda^{m-1}K_m(x,t) = \sum_{m=1}^{\infty}\lambda^{m-1}e^{x+t}\left(\frac{e^2-1}{2}\right)^{m-1} = e^{x+t}\sum_{m=1}^{\infty}\left\{\frac{\lambda(e^2-1)}{2}\right\}^{m-1}$$

But $\displaystyle\sum_{m=1}^{\infty}\left\{\frac{\lambda(e^2-1)}{2}\right\}^{m-1} = 1 + \frac{\lambda(e^2-1)}{2} + \left\{\frac{\lambda(e^2-1)}{2}\right\}^2 + \cdots$

which is an infinite geometric series with common ratio $\{\lambda(e^2-1)\}/2$.

$$\therefore \sum_{m=1}^{\infty}\left\{\frac{\lambda(e^2-1)}{2}\right\}^{m-1} = \frac{1}{1-\{\lambda(e^2-1)\}/2} = \frac{2}{2-\lambda(e^2-1)},$$

provided $\left|\dfrac{\lambda(e^2-1)}{2}\right| < 1$ or $|\lambda| < \dfrac{2}{e^2-1}$

Therefore $R(x,t;\lambda) = \dfrac{2e^{x+t}}{2-\lambda(e^2-1)}$, provided $|\lambda| < \dfrac{2}{e^2-1}$

11.2.3 Solution in terms of resolvent kernel :

Let the Fredholm integral is given by Eq.(11.2.4). Let $K_m(x,t)$ be the m-th iterated kernel and let $R(x,t;\lambda)$ be the resolvent kernel of Eq.(11.2.4). Then we have

$$R(x,t;\lambda) = \sum_{m=1}^{\infty}\lambda^{m-1}K_m(x,t) \tag{11.2.10}$$

Suppose the sum of the infinite series (11.2.10) exists and so $R(x,t;\lambda)$ can be obtained in the closed form. Then, the required solution of Eq.(11.2.4) is given by

$$y(x) = f(x) + \lambda\int_a^b R(x,t;\lambda)\,f(t)\,dt \tag{11.2.11}$$

Example 11.8. Solve

$$y(x) = x + \int_0^{1/2} y(t)\, dt$$

Solution : Comparing the given equation with

$$y(x) = f(x) + \lambda \int_0^{1/2} K(x,t)\, y(t)\, dt,$$

$$\text{we have} \quad f(x) = x, \quad \lambda = 1, \quad K(x,t) = 1 \tag{11.2.12}$$

Let $K_m(x,t)$ be the m-th iterated kernel. Then, we have

$$K_1(x,t) \;=\; K(x,t) \tag{11.2.13}$$

$$K_m(x,t) \;=\; \int_0^1 K(x,z) K_{m-1}(z,t)\, dz \tag{11.2.14}$$

From (11.2.12), $K_1(x,t) = K(x,t) = 1$.

Putting $m = 2$ in (11.2.14), we have

$$K_2(x,t) = \int_0^{1/2} K(x,z) K_1(z,t)\, dz = \int_0^{1/2} dz = [z]_0^{1/2} = \frac{1}{2}.$$

Putting $m = 3$ in (11.2.14), we have

$$K_3(x,t) = \int_0^{1/2} K(x,z) K_2(z,t)\, dz = \int_0^{1/2} \frac{1}{2}\, dz = \left(\frac{1}{2}\right)^2.$$

Observing above we find

$$K_m(x,t) = \left(\frac{1}{2}\right)^{m-1}$$

Now, the resolvent kernel $R(x,t;\lambda)$ is given by

$$R(x,t;\lambda) = \sum_{m=1}^{\infty} \lambda^{m-1} K_m(x,t) = \sum_{m=1}^{\infty} \left(\frac{1}{2}\right)^{m-1}$$

$$\text{But} \quad \sum_{m=1}^{\infty} \left(\frac{1}{2}\right)^{m-1} = 1 + \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \cdots$$

which is an infinite geometric series with common ratio $1/2$.

$$\therefore \quad \sum_{m=1}^{\infty} \left(\frac{1}{2}\right)^{m-1} = \frac{1}{1-(1/2)} = 2 \quad \text{and hence} \quad R(x,t;\lambda) = 2$$

Finally, the required solution of the given equation is given by

$$\begin{aligned} y(x) &= f(x) + \lambda \int_0^{1/2} R(x,t;\lambda)\, f(t)\, dt \\ \Rightarrow y(x) &= x + \int_0^{1/2} (2t)\, dt \\ \Rightarrow y(x) &= x + 2\left[\frac{t^2}{2}\right]_0^{1/2} = x + \frac{1}{4} \end{aligned}$$

Exercise 11.9. Solve

$$i) \quad y(x) = e^x - \frac{1}{2}e + \frac{1}{2} + \frac{1}{2}\int_0^1 y(t)\,dt \quad ii) \quad y(x) = \frac{5x}{6} + \frac{1}{2}\int_0^1 xt\,y(t)\,dt$$

$$iii) \quad y(x) = \sin x - \frac{x}{4} + \frac{1}{4}\int_0^{\pi/2} xt\,y(t)\,dt \quad iv) \quad y(x) = \frac{3}{2}e^x - \frac{1}{2}xe^x - \frac{1}{2} + \frac{1}{2}\int_0^1 t\,y(t)\,dt$$

Answers :

$$i) \quad y(x) = e^x, \quad (ii) \quad y(x) = x, \quad (iii) \quad y(x) = \sin x, \quad (iv) \quad y(x) = \frac{3e^x}{2} - \frac{xe^x}{2} - \frac{e}{3} + 1$$

11.2.4 Iterative scheme for Fredholm integral equations

When the resolvent kernel cannot be obtained in closed form i.e., the sum of infinite series occurring in the formula of the resolvent kernel can not be determined, we use the method of successive approximations to find solutions upto third order.

Let the given Fredholm integral equation of the second kind be

$$y(x) = f(x) + \lambda \int_a^b K(x,t)\,y(t)\,dt \tag{11.2.15}$$

As zero-order approximation, we take $y_0(x) = f(x)$. If n-th order approximation be $y_n(x)$, then

$$y_n(x) = f(x) + \lambda \int_a^b K(x,t) y_{n-1}(x) \tag{11.2.16}$$

Sometimes the zero-order approximation is mentioned in the problem. In that case, we will modify the scheme accordingly.

Example 11.10. Solve the following integral equation

$$y(x) = 1 + \lambda \int_0^1 (x+t)\,y(t)\,dt,$$

by the method of successive approximation to third order.

Solution : Given

$$y(x) = 1 + \lambda \int_0^1 (x+t)\,y(t)\,dt \tag{11.2.17}$$

Let $y_0(x)$ denote the zero-order approximation. Then we may take

$$y_0 = 1.$$

If $y_n(x)$ denotes the n-th order approximation, then we know that

$$y_n(x) = 1 + \lambda \int_0^1 (x+t)\,y_{n-1}(t)\,dt, \tag{11.2.18}$$

Putting $n = 1$ in (11.2.18),

$$y_1(x) = 1 + \lambda \int_0^1 (x+t)\, y_0(t)\, dt = 1 + \lambda \int_0^1 (x+t)\, dt,$$

$$\Rightarrow y_1(x) = 1 + \lambda \left[xt + \frac{1}{2}t^2 \right]_0^1 = 1 + \lambda \left(x + \frac{1}{2} \right). \tag{11.2.19}$$

Next, putting $n = 2$ in (11.2.18), we have

$$\begin{aligned}
y_2(x) &= 1 + \lambda \int_0^1 (x+t)\, y_1(t)\, dt = 1 + \lambda \int_0^1 (x+t) \left\{ 1 + \lambda \left(t + \frac{1}{2} \right) \right\}, \\
&= 1 + \lambda \int_0^1 (x+t) \left\{ \left(1 + \frac{\lambda}{2} \right) + \lambda t \right\} dt = 1 + \lambda \int_0^1 \left[x \left(1 + \frac{\lambda}{2} \right) + t \left(1 + \frac{\lambda}{2} + \lambda x \right) + \lambda t^2 \right] dt \\
&= 1 + \lambda \left[x \left(1 + \frac{\lambda}{2} \right) t + \frac{t^2}{2} \left(1 + \frac{\lambda}{2} + \lambda x \right) + \frac{\lambda t^3}{3} \right]_0^1 \\
&= 1 + \lambda \left[x \left(1 + \frac{\lambda}{2} \right) t + \frac{1}{2} \left(1 + \frac{\lambda}{2} + \lambda x \right) + \frac{\lambda}{3} \right] = 1 + \lambda \left(x + \frac{1}{2} \right) + \lambda^2 \left(x + \frac{7}{12} \right)
\end{aligned}$$

Finally, putting $n = 3$ in (11.2.18), we have

$$\begin{aligned}
y_3(x) &= 1 + \lambda \int_0^1 (x+t)\, y_2(t)\, dt = 1 + \lambda \int_0^1 (x+t) \left\{ 1 + \lambda \left(t + \frac{1}{2} \right) + \lambda^2 \left(t + \frac{7}{2} \right) \right\}, \\
&= 1 + \lambda \int_0^1 (x+t) \left\{ \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} \right) + \lambda t (1+\lambda) \right\} dt \\
&= 1 + \lambda \int_0^1 \left[x \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} \right) + t \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} + \lambda x + \lambda^2 x \right) + \lambda t^2 (1+\lambda) \right] dt \\
&= 1 + \lambda \left[x \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} \right) t + \frac{t^2}{2} \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} + \lambda x + \lambda^2 x \right) + \frac{\lambda 1}{3} \lambda t^3 (1+\lambda) \right]_0^1 \\
&= 1 + \lambda x \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} \right) t + \frac{\lambda}{2} \left(1 + \frac{\lambda}{2} + \frac{7\lambda^2}{12} + \lambda x + \lambda^2 x \right) + \frac{1}{3} \lambda^2 (1+\lambda)
\end{aligned}$$

$$\text{Therefore,} \quad y_3(x) = 1 + \lambda \left(x + \frac{1}{2} \right) + \lambda^2 \left(x + \frac{7}{12} \right) + \lambda^3 \left(\frac{13}{12} x + \frac{5}{8} \right)$$

Exercise 11.11. Exercise : Solve the inhomogeneous Fredholm integral equation of the second kind

$$y(x) = 2x + \lambda \int_0^1 (x+t)\, y(t)\, dt,$$

by the method of successive approximations to the third order by taking $y_0(x) = 1$.

Answers :

$$y_3(x) = 2x + \lambda \left(x + \frac{2}{3} \right) + \lambda^2 \left(\frac{7}{6} x + \frac{2}{3} \right) + \lambda^3 \left(\frac{13}{12} x + \frac{5}{8} \right)$$

Definition 11.12. Volterra Integral Equation : A linear integral equation of the form

$$g(x) y(x) = f(x) + \lambda \int_a^x K(x,t) y(t) dt, \tag{11.2.20}$$

where a, b are both constants, $f(x)$, $g(x)$ and $K(x,t)$ are known functions while $y(x)$ is unknown function; λ is a non-zero real or complex parameter is called *Volterra integral equation of third kind*. The function $K(x,t)$ is known as the kernel of the integral equation.

• Setting $g(x) = 0$ in Eq.(11.2.20), we have the *Volterra integral equation of the first kind*.

• Setting $g(x) = 1$ in Eq.(11.2.20), we have the *Volterra integral equation of the second kind*.

• Setting $g(x) = 1$ and $f(x) = 0$ in Eq.(11.2.20), we have the *Homogeneous Volterra integral equation of the second kind*.

11.3 Solution of Volterra integral equations

11.3.1 Determination of Resolvent kernel for Volterra integral equations

Example 11.13. Find the resolvent kernel of the Volterra integral equation with the kernel $K(x,t) = 1$.

Solution : Iterated kernels $K_n(x,t)$ are given by

$$K_1(x,t) = K(x,t) \tag{11.3.1}$$

$$\text{and} \quad K_n(x,t) = \int_t^x K(x,z)K_{n-1}(z,t)\,dz, \;\; n = 1,2,3,\ldots \tag{11.3.2}$$

Given $K(x,t) = 1$. Thus we have

$$K_1(x,t) = K(x,t) = 1$$

Putting $n = 2$ in Eq.(11.3.2) we have

$$K_2(x,t) = \int_t^x K(x,z)K_1(z,t)\,dz = \int_t^x dz = \big[z\big]_t^x = x - t$$

Putting $n = 3$ in Eq.(11.3.2) we have

$$K_3(x,t) = \int_t^x K(x,z)K_2(z,t)\,dz = \int_t^x 1\cdot(z-t)\,dz = \left[\frac{(z-t)^2)}{2}\right]_t^x = \frac{(x-t)^2}{2!}$$

Putting $n = 4$ in Eq.(11.3.2) we have

$$K_4(x,t) = \int_t^x K(x,z)K_3(z,t)\,dz = \int_t^x 1\cdot\frac{(z-t)^2}{2!}\,dz = \frac{1}{2!}\left[\frac{(z-t)^3)}{3}\right]_t^x = \frac{(x-t)^3}{3!}$$

Observing above, we find by mathematical induction, that

$$K_n(x,t) = \frac{(x-t)^{n-1}}{(n-1)!}, \;\; n = 1,2,3,\ldots$$

Now by the definition of the resolvent kernel, we have

$$\begin{aligned} R(x,t;\lambda) &= \sum_{m=1}^{\infty} K_m(x,t) = K_1(x,t) + \lambda K_2(x,t) + \lambda^2 K_3(x,t) + \cdots \\ &= 1 + \frac{\lambda(x-t)}{1!} + \frac{[\lambda(x-t)]^2}{2!} + \frac{[\lambda(x-t)]^3}{3!} + \cdots \\ &= e^{\lambda(x-t)} \end{aligned}$$

Exercise 11.14. Find the resolvent kernel of the Volterra integral equation with the kernel

$$i) \quad K(x,t) = e^{x-t} \qquad ii) \quad K(x,t) = (2+\cos x)/(2+\cos t)$$

Answers :

$$i) \quad R(x,t;\lambda) = e^{(x-t)(1+\lambda)}, \qquad (ii) \quad R(x,t;\lambda) = \frac{2+\cos x}{2+\cos t} e^{\lambda(x-t)}$$

11.3.2 Solution of Volterra integral equation in terms of resolvent kernel

Working Rule : Let

$$y(x) = f(x) + \lambda \int_a^x K(x,t)\, y(t)\, dt \tag{11.3.3}$$

be given Volterra integral equation. Let $K_m(x,t)$ be the m-th iterated kernel and let $R(x,t;\lambda)$ be the resolvent kernel of (11.3.3). Then we have

$$R(x,t;\lambda) = \sum_{m=1}^{\infty} \lambda^{m-1} K_m(x,t). \tag{11.3.4}$$

Suppose the sum of infinite series (11.3.4) exists and so $R(x,t;\lambda)$ can be obtained in the closed form. Then the required solution of (11.3.3) is given by

$$y(x) = f(x) + \lambda \int_a^x R(x,t;\lambda)\, f(t)\, dt. \tag{11.3.5}$$

Example 11.15. With the aid of the resolvent kernel, find the solution of the integral equation

$$y(x) = e^{x^2} + \int_0^x e^{x^2-t^2} y(t)\, dt.$$

Solution : Comparing the given equation with

$$y(x) = f(x) + \lambda \int_0^x K(x,t)\, y(t)\, dt$$

we have

$$f(x) = e^{x^2}, \qquad \lambda = 1, \qquad K(x,t) = e^{x^2-t^2}$$

Let $K_m(x,t)$ be the m-th iterated kernel. Then we have

$$K_1(x,t) = K(x,t)$$

$$\text{and} \quad K_m(x,t) = \int_t^x K(x,z) K_{m-1}(z,t)\, dz \tag{11.3.6}$$

Thus we have

$$K_1(x,t) = K(x,t) = e^{x^2-t^2} \tag{11.3.7}$$

Putting $m = 2$ in (11.3.6), we have

$$K_2(x,t) = \int_t^x K(x,z) K_1(z,t)\, dz = \int_t^x e^{x^2-z^2} e^{z^2-t^2} dz = e^{x^2-t^2} \int_t^x dz = e^{x^2-t^2}(x-t)$$

Putting $m = 3$ in (11.3.6), we have

$$\begin{aligned} K_3(x,t) &= \int_t^x K(x,z)K_2(z,t)\,dz = \int_t^x e^{x^2-z^2}e^{z^2-t^2}(z-t)\,dz = e^{x^2-t^2}\int_t^x (z-t)\,dz \\ &= e^{x^2-t^2}\left[\frac{(z-t)^2}{2}\right]_t^x = e^{x^2-t^2}\frac{(x-t)^2}{2!} \end{aligned}$$

Putting $m = 4$ in (11.3.6), we have

$$\begin{aligned} K_4(x,t) &= \int_t^x K(x,z)K_3(z,t)\,dz = \int_t^x e^{x^2-z^2}e^{z^2-t^2}\frac{(z-t)^2}{2!}\,dz \\ &= \frac{e^{x^2-t^2}}{2!}\left[\frac{(z-t)^3}{3}\right]_t^x = e^{x^2-t^2}\frac{(x-t)^3}{3!} \end{aligned}$$

Observing above by mathematical induction we may write

$$K_m(x,t) = e^{x^2-t^2}\frac{(x-t)^{m-1}}{(m-1)!}, \quad m = 1,2,3,\ldots \tag{11.3.8}$$

Now, by the definition of the resolvent kernel, we have

$$\begin{aligned} R(x,t;\lambda) &= \sum_{m=1}^{\infty} K_m(x,t) = K_1(x,t) + \lambda K_2(x,t) + \lambda^2 K_3(x,t) + \cdots \\ &= e^{x^2-t^2} + e^{x^2-t^2}\frac{(x-t)}{1!} + e^{x^2-t^2}\frac{(x-t)^2}{2!} + \cdots \\ &= e^{x^2-t^2}\left[1 + \frac{(x-t)}{1!} + \frac{(x-t)^2}{2!} + \cdots\right] \\ &= e^{x^2-t^2}e^{x-t} \end{aligned}$$

Finally, the required solution of the given equation is given by

$$\begin{aligned} y(x) &= f(x) + \lambda\int_0^x R(x,t;\lambda)\,f(t)\,dt = e^{x^2} + \int_0^x e^{x^2-t^2}e^{x-t}e^{t^2}\,dt \\ &= e^{x^2} + e^{x^2+x}\int_0^x e^{-t}\,dt = e^{x^2} + e^{x^2+x}\big[-e^{-t}\big]_0^x \\ &= e^{x^2} + e^{x^2+x}[-e^{-x}+1] = e^{x^2} - e^{x^2} + e^{x^2+x} = e^{x^2+x} \end{aligned}$$

Exercise 11.16. Solve the following integral equation by means of resolvent kernel

$$i) \quad y(x) = e^x\sin x + \int_0^x \frac{2+\cos x}{2+\cos t}y(t)\,dt \qquad ii) \quad y(x) = \cos x - x - 2 + \int_0^x (t-x)\,y(t)\,dt$$

Answers :

$$i) \quad y(x) = e^x\sin x - e^x(2+\cos x)\log\left(\frac{2+\cos x}{3}\right), \qquad (ii) \quad y(x) = -\cos x - \sin x - \frac{x}{2}\sin x$$

11.3.3 Method of Successive approximations for solving Volterra integral equation

Working Rule : Let $f(x)$ be continuous in $[0, a]$ and $K(x, t)$ be continuous for $0 \leq x \leq a,\ \ 0 \leq t \leq x$. We start with some function $y_0(x)$ continuous in $[0, a]$. Replacing $y(t)$ on R.H.S of (11.2.20) by $y_0(x)$, we obtain

$$y_1(x) = f(x) + \lambda \int_0^x K(x, t)\, y_0(t)\, dt. \tag{11.3.9}$$

$y_1(x)$ given by (11.3.9) is itself continuous in $[0, a]$. Proceeding likewise we arrive at a sequence of functions $y_0(x),\ y_1(x), \ldots, y_n(x), \ldots$, where

$$y_n(x) = f(x) + \lambda \int_0^x K(x, t)\, y_{n-1}(t)\, dt. \tag{11.3.10}$$

In view of continuity of $f(x)$ and $K(x, t)$, the sequence $\{y_n(x)\}$ converges, as $n \to \infty$ to obtain the solution of $y(x)$ of given integral equation (11.2.20). It should be note that when $y_0(x) = f(x)$, we obtain the so called *Neumann series*.

Example 11.17. Using the method of successive approximations, solve the integral equation

$$y(x) = 1 + \int_0^x y(t)\, dt, \quad \text{taking } \ y_0(x) = 0.$$

Solution : Comparing the given equation with

$$y(x) = f(x) + \lambda \int_0^x K(x, t)\, y(t)\, dt,$$

we find

$$f(x) = 1, \qquad \lambda = 1, \qquad K(x, t) = 1$$

The n-th order approximation is given by

$$y_n(x) = 1 + \int_0^x y_{n-1}(t)\, dt, \tag{11.3.11}$$

Putting $n = 1$ in (11.3.11), we have

$$y_1(x) = 1 + \int_0^x y_0(t)\, dt = 1 + \int_0^x (0)\, dt = 1.$$

Putting $n = 2$ in (11.3.11), we have

$$y_2(x) = 1 + \int_0^x y_1(t)\, dt = 1 + \int_0^x dt = 1 + \big[t\big]_0^x = 1 + x.$$

Putting $n = 3$ in (11.3.11), we have

$$y_3(x) = 1 + \int_0^x y_2(t)\, dt = 1 + \int_0^x (1 + t)\, dt = 1 + \left[t + \frac{t^2}{2}\right]_0^x = 1 + x + \frac{x^2}{2!}.$$

Putting $n = 4$ in (11.3.11), we have

$$y_4(x) = 1 + \int_0^x y_3(t)\, dt = 1 + \int_0^x \left(1 + t + \frac{t^2}{2!}\right)\, dt = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}.$$

Observing the above trend, we find

$$y_n(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^{n-1}}{(n-1)!}$$

Making $n \to \infty$, we find the required solution is given by

$$\begin{aligned} y(x) &= \lim_{n\to\infty} y_n(x) \\ &= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots\cdots = e^x \end{aligned}$$

Exercise 11.18. Using the method of successive approximations, solve the integral following integral equations.

$i)\ y(x) = 1 + x - \int_0^x y(t)\,dt,$ taking $y_0(x) = 1$ $ii)\ y(x) = x - \int_0^x (x-t)\,y(t)\,dt,$ taking $y_0(x) = 0$

$iii)\ y(x) = 1 + \int_0^x (x-t)\,y(t)\,dt,$ taking $y_0(x) = 1$ $iv)\ y(x) = \frac{1}{2}x^3 - 2x - \int_0^x y(t)\,dt,$ taking $y_0(x) = x^2$

Answers :

$i)\quad y(x) = 1, \quad (ii)\quad y(x) = \sin x, (ii)\quad y(x) = \cosh x, \quad (iv)\quad y(x) = x^2 - 2x$

Unit 12

Course Structure

Classical Fredholm theory : Fredholm theorems, Fredholm Alternative Principles. Hilbert-Schmidt theory : Symmetric kernels, Orthogonal system of functions, Fundamental properties of eigenvalues and eigenfunctions for symmetric kernels, Hilbert-Schmidt theorem.

---

12 Introduction

In Unit 11, we obtained the solution of the Fredholm integral equation of the second kind

$$y(x) = f(x) + \lambda \int_a^b K(x,t)\, y(t)\, dt \tag{12.0.1}$$

as a uniformly convergent power series in the parameter λ for $|\lambda|$ suitably small. Fredholm derived the solution of (12.0.1) in general form which is valid for all values of the parameter λ. He gave three important results which are known as Fredholm's first, second and third fundamental theorems. In the present unit we propose to discuss these theorems.

Objective

The objective of this course is to learn the students all of the above topics and by the end of it students should be able to

- know different fundamental theorems of Fredholm integral equation.
- know the method of solution of Fredholm integral equation using fundamental theorems.
- know various aspects of Hilbert-Schmidt theory
- know fundamental properties of eigenvalues and eigenfunctions for symmetric kernels

12.1 Fredholm's First Fundamental Theorem

The non-homogeneous Fredholm integral equation of second kind

$$y(x) = f(x) + \lambda \int_a^b K(x,t)\, y(t)\, dt \tag{12.1.1}$$

where the functions $f(x)$ and $y(t)$ are integrable, has a unique solution

$$y(x) = f(x) + \lambda \int_a^b R(x,t;\lambda)\, f(t)\, dt \tag{12.1.2}$$

where the resolvent kernel $R(x,t;\lambda)$ is given by

$$R(x,t;\lambda) = \frac{D(x,t;\lambda)}{D(\lambda)} \tag{12.1.3}$$

with $D(\lambda) \neq 0$, is a meromorphic function of the complex variable λ, being the ratio of two entire functions defined by the series

$$\begin{aligned} D(x,t;\lambda) &= K(x,t) + \sum_{p=1}^{\infty} \frac{(-\lambda)^p}{p!} \int \cdots \int K \begin{pmatrix} x, & z_1, & \cdots, & z_p \\ t, & z_1, & \cdots, & z_p \end{pmatrix} dz_1 \cdots dz_p \\ \text{and } D(\lambda) &= 1 + \sum_{p=1}^{\infty} \frac{(-\lambda)^p}{p!} \int \cdots \int K \begin{pmatrix} z_1, & \cdots, & z_p \\ z_1, & \cdots, & z_p \end{pmatrix} dz_1 \cdots dz_p, \end{aligned} \tag{12.1.4}$$

both of which converge for all values of λ. Also, note the following symbol for the determinant formed by the values of the values of the kernel at all points (x_i, t_i)

$$\begin{vmatrix} K(x_1,t_1) & K(x_1,t_2) & \cdots & K(x_1,t_n) \\ K(x_2,t_1) & K(x_2,t_2) & \cdots & K(x_2,t_n) \\ \vdots & \vdots & \ldots & \vdots \\ K(x_n,t_1) & K(x_n,t_2) & \cdots & K(x_n,t_n) \end{vmatrix} = K \begin{pmatrix} x_1, & x_2, & \cdots, & x_n \\ t_1, & t_2, & \cdots, & t_n \end{pmatrix} \tag{12.1.5}$$

which is known as the *Fredholm determinant.* In particular, the solution of the Fredholm homogeneous equation

$$y(x) = \lambda \int_a^b K(x,t)\, y(t)\, dt \tag{12.1.6}$$

is identically zero.

Result 12.1. For Fredholm integral equation

$$y(x) = f(x) + \lambda \int_a^b K(x,t)\, y(t)\, dt \tag{12.1.7}$$

the resolvent kernel is given by

$$R(x,t;\lambda) = \frac{D(x,t;\lambda)}{D(\lambda)} \tag{12.1.8}$$

where

$$D(x,t;\lambda) = K(x,t) + \sum_{m=1}^{\infty} \frac{(-\lambda)^m}{m!} B_m(x,t) \tag{12.1.9}$$

$$\text{and } D(\lambda) = 1 + \sum_{m=1}^{\infty} \frac{(-\lambda)^m}{m!} C_m \tag{12.1.10}$$

where

$$B_n(x,t) = \underbrace{\int_a^b \cdots \int_a^b}_{n} \begin{vmatrix} K(x,t) & K(x,z_1) & \cdots & K(x,z_n) \\ K(z_1,t) & K(z_1,z_1) & \cdots & K(z_1,z_n) \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ K(z_n,t) & K(z_n,z_1) & \cdots & K(z_n,z_n) \end{vmatrix} dz_1\, dz_2\, \cdots\, dz_n, \tag{12.1.11}$$

$$\text{and} \quad C_n = \underbrace{\int_a^b \cdots \int_a^b}_{n} \begin{vmatrix} K(z_1,z_1) & K(z_1,z_2) & \cdots & K(z_1,z_n) \\ K(z_2,z_1) & K(z_2,z_2) & \cdots & K(z_2,z_n) \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ K(z_n,z_1) & K(z_n,z_2) & \cdots & K(z_n,z_n) \end{vmatrix} dz_1\, dz_2\, \cdots\, dz_n, \tag{12.1.12}$$

The function $D(x,t;\lambda)$ is called the *Fredholm minor* and $D(\lambda)$ is called the *Fredhom determinant*.

12.2 Alternative Procedure of calculating $\mathbf{B_m(x,t)}$ and $\mathbf{C_m}$

The following results will be used

$$\begin{aligned} C_0 &= 1 \\ C_p &= \int_a^b B_{p-1}(s,s)\, ds, \;\; p \geq 1 \\ B_0(x,t) &= K(x,t) \\ B_p(x,t) &= C_p\, K(x,t) - p \int_a^b K(x,z)\, B_{p-1}(z,t)\, dz, \;\; p \geq 1. \end{aligned} \tag{12.2.1}$$

After getting $R(x,t;\lambda)$, the required solution is given by

$$y(x) = f(x) + \lambda \int_a^b R(x,t;\lambda)\, f(t)\, dt \tag{12.2.2}$$

Example 12.2. Using Fredholm determinants, find the resolvent kernel and hence solve the following integral equation

$$y(x) = f(x) + \lambda \int_0^1 xe^t\, y(t)\, dt, \;\; (\lambda \neq 1)$$

Solution : Here

$$K(x,t) = xe^t$$

From Eq.(12.1.11)

$$B_1(x,t) = \int_0^1 \begin{vmatrix} K(x,t) & K(x,z_1) \\ K(z_1,t) & K(z_1,z_1) \end{vmatrix} dz_1 = \int_0^1 \begin{vmatrix} xe^t & xe^{z_1} \\ z_1e^t & z_1e^{z_1} \end{vmatrix} dz_1 = 0$$

$$B_2(x,t) = \int_0^1 \int_0^1 \begin{vmatrix} K(x,t) & K(x,z_1) & K(x,z_2) \\ K(z_1,t) & K(z_1,z_1) & K(z_1,z_2) \\ K(z_2,t) & K(z_2,z_1) & K(z_2,z_2) \end{vmatrix} dz_1\, dz_2 = \int_0^1 \int_0^1 \begin{vmatrix} xe^t & xe^{z_1} & xe^{z_2} \\ z_1e^t & z_1e^{z_1} & z_1e^{z_2} \\ z_2e^t & z_2e^{z_1} & z_2e^{z_2} \end{vmatrix} dz_1\, dz_2 = 0$$

Since $B_1(x,t) = B_2(x,t) = 0$, it follows that $B_n(x,t) = 0$, for $n \geq 1$. Now from Eq.(12.1.12), we have

$$C_1 = \int_0^1 K(z_1,z_1)\, dz_1 = \int_0^1 z_1e^{z_1}\, dz_1 = \left[z_1e^{z_1}\right]_0^1 - \int_0^1 e^{z_1} dz_1 = e - \left[e^{z_1}\right]_0^1 = e - (e-1) = 1.$$

$$C_2 = \int_0^1 \int_0^1 \begin{vmatrix} K(z_1,z_1) & K(z_1,z_2) \\ K(z_2,z_1) & K(z_2,z_2) \end{vmatrix} dz_1\, dz_2 = \int_0^1 \int_0^1 \begin{vmatrix} z_1e^{z_1} & z_1e^{z_2} \\ z_2e^{z_1} & z_2e^{z_2} \end{vmatrix} dz_1\, dz_2 = 0$$

It follows that $C_m = 0$ for all $m \geq 2$. Now Eq.(12.1.9) and Eq.(12.1.10) respectively gives

$$\begin{aligned} D(x,t;\lambda) &= K(x,t) - \lambda B_1(x,t) + \frac{\lambda^2}{2!} B_2(x,t) - \cdots = xe^t \\ D(\lambda) &= 1 - \lambda C_1 + \frac{\lambda^2}{2!} C_2 - \ldots = 1 - \lambda \end{aligned}$$

Hence, Eq.(12.1.8) yields

$$R(x,t;\lambda) = \frac{D(x,t;\lambda)}{D(\lambda)} = \frac{xe^t}{1-\lambda}$$

Hence the required solution is

$$\begin{aligned} y(x) &= f(x) + \lambda \int_0^1 R(x,t;\lambda)\, f(t)\, dt \\ \Rightarrow y(x) &= f(x) + \lambda \int_0^1 \frac{xe^t}{1-\lambda} f(t)\, dt \\ \Rightarrow y(x) &= f(x) + \frac{\lambda x}{1-\lambda} \int_0^1 e^t\, f(t)\, dt. \end{aligned}$$

Alternative Method : We shall use the results of Eqs.(12.2.1) to compute $R(x,t;\lambda)$ as follows. First write down these results for complete solution. Here

$$\begin{aligned} &C_0 = 1 \\ &B_0(x,t) = K(x,t) = xe^t \\ &C_1 = \int_0^1 B_0(s,s)\, ds = \int_0^1 se^s\, ds = \left[se^s\right]_0^1 - \int_0^1 e^s\, ds = e - \left[e^s\right]_0^1 = e - (e-1) = 1 \\ &B_1 = C_1 K(x,t) - \int_0^1 K(x,z) B_0(z,t)\, dz = xe^t - \int_0^1 xe^z ze^t\, dz == xe^t - xe^t \int_0^1 ze^z\, dz = 0 \\ &C_2 = \int_0^1 B_1(s,s)\, ds = 0 \\ &B_2(x,t) = C_2 K(x,t) - 2\int_0^1 K(x,z)\, B_1(z,t)\, dz = 0 \\ &\therefore B_m(x,t) = 0 \quad \text{for all } m \geq 1 \quad \text{and} \quad C_m = 0 \quad \text{for all } m \geq 2. \end{aligned}$$

Now we proceed as before to determine $R(x,t;\lambda)$ and can solve the given Fredholm integral equation.

Important Observation : The reader will find that the above alternative method is a short cut. However, he should find the required quantities strictly in the following order :

$$C_0, \quad B_0(x,t), \quad C_1, \quad B_1(x,t), \quad C_2, \quad B_2(x,t) \quad \text{and so on.}$$

Exercise 12.3. Using Fredholm determinants, find the resolvent kernel and hence solve the following integral equation

$$\begin{aligned} &i)\ y(x) = e^{-x} + \lambda \int_0^1 xe^t\, y(t)\, dt, && ii)\ y(x) = 1 + \int_0^1 (1-3xt)\, y(t)\, dt, \\ &iii)\ y(x) = \sin x + \lambda \int_4^{10} x\, y(t)\, dt, && iv)\ y(x) = 1 + \int_0^{\pi} \sin(x+t)\, y(t)\, dt \end{aligned}$$

Answers :

$$\begin{aligned} &i)\ y(x) = e^{-x} + \frac{\lambda x}{1-\lambda}, \text{ if } \lambda \neq 1 && (ii)\ y(x) = \frac{8-6x}{3}, \\ &(iii)\ y(x) = \sin x + \frac{2\lambda x \sin 7 \sin 3}{1-42\lambda}, && (iv)\ y(x) = 1 + \frac{4}{4-\pi^2}(2\cos x + \pi \sin x) \end{aligned}$$

12.3 Fredholm Second Fundamental Theorem

If λ_0 is a zero of multiplicity m of the function $D(\lambda)$, then the homogeneous integral equation

$$y(x) = \lambda_0 \int_a^b K(x,t)\, y(t)\, dt \tag{12.3.1}$$

possesses at least one, and the most m, linearly independent solutions

$$y_i(x) = D_r \left(\begin{array}{ccccccc} x_1, & \cdots, & x_{i-1}, & x, & x_{i+1}, & \cdots, & x_r \\ t_1, & \cdots, & t_{i-1}, & t, & t_{i+1}, & \cdots, & t_r \end{array} \middle| \, \lambda_0 \right), \quad i = 1, 2, \ldots, r; \;\; 1 \le r \le m. \tag{12.3.2}$$

not identically zero. Any other solution of this equation is a linear combination of these solutions. Here, we have to remember the following definition of the Fredholm minor

$$\begin{aligned} D_n \begin{pmatrix} x_1, & x_2, & \cdots, & x_n \\ t_1, & t_2, & \cdots, & t_n \end{pmatrix} &= K \begin{pmatrix} x_1, & x_2, & \cdots, & x_n \\ t_1, & t_2, & \cdots, & t_n \end{pmatrix} \\ &+ \sum_{p=1}^{\infty} \frac{(-\lambda)^p}{p!} \int_a^b \cdots \int_a^b K \begin{pmatrix} x_1, & \cdots, & x_n, & z_1 & \cdots & z_p \\ t_1, & \cdots, & t_n, & z_1 & \cdots & z_p \end{pmatrix} dz_1 \cdots dz_p, \end{aligned} \tag{12.3.3}$$

where $\{x_i\}$ and $\{t_i\},\ i = 1, 2, \ldots, n$, are two sequences of arbitrary variables. Series (12.3.3) converges for all values of λ and hence it is an entire function of λ.

12.4 Fredholm Third Fundamental Theorem

For an inhomogeneous integral equation

$$y(x) = f(x) + \lambda_0 \int_a^b K(x,t)\, y(t)\, dt \tag{12.4.1}$$

to possesses a solution in the case $D(\lambda_0) = 0$, it is necessary and sufficient that the given function $f(x)$ be orthogonal to all the eigenfunctions $z_i(x), i = 1, 2, \ldots, \nu$, of the transposed homogeneous equation corresponding to the eigenvalue λ_0. The general solution has the form

$$y(x) = f(x) + \lambda \int_a^b \frac{D_{r+1} \left(\begin{array}{ccccc} x, & x_1, & x_2, & \cdots, & x_r \\ t, & t_1, & t_2, & \cdots, & t_r \end{array} \middle| \, \lambda_0 \right)}{D_r \left(\begin{array}{ccccc} x_1, & x_2, & \cdots, & x_r \\ t_1, & t_2, & \cdots, & t_r \end{array} \middle| \, \lambda_0 \right)} f(t)\, dt + \sum_{h=1}^{r} C_h \Phi_h(x), \tag{12.4.2}$$

where $\Phi_i(x)$ are given by

$$\Phi_i(x) = \frac{D_r \left(\begin{array}{ccccccc} x_1, & \cdots, & x_{i-1}, & x, & x_{i+1}, & \cdots, & x_r \\ t_1, & \cdots & \cdots & & \cdots & \cdots & t_r \end{array} \middle| \, \lambda_0 \right)}{D_r \left(\begin{array}{ccccccc} x_1, & \cdots, & x_{i-1}, & x_i, & x_{i+1}, & \cdots, & x_r \\ t_1, & \cdots & \cdots & & \cdots & \cdots & t_r \end{array} \middle| \, \lambda_0 \right)}, \qquad i = 1, 2, \ldots, r \tag{12.4.3}$$

12.5 Hilbert-Schmidt Theory

12.5.1 Symmetric Kernels

A kernel is called *symmetric* if it coincides with its own complex conjugate. Such a kernel is characterized by the identity

$$K(x,t) = \overline{K}(t,x),$$

where the bar denotes the complex conjugate. If the kernel is real, then its *symmetry* is defined by the identity $K(x,t) = K(t,x)$. An integral equation with a symmetric kernel is called a symmetric equation.

Remark 12.4. For a symmetric kernel that is not identically zero, at least one eigenvalue will always exist. This is an important characteristic of symmetric kernel. An eigenvalue is *simple* if there is only one corresponding eigenfunction, otherwise the eigenvalues are degenerate. The *spectrum* of the kernel $K(x,t)$ is the set of all its eigenvalues. Thus the *spectrum of a symmetric kernel is never empty.*

12.5.2 Orthogonal system of functions

A finite or an infinite set $\{\phi_k(x)\}$ defined on an interval $a \le n \le b$ is said to be an orthogonal set if

$$(\phi_i, \phi_j) = 0 \quad \text{or} \quad \int_a^b \phi_i(x)\phi_j(x) = 0, \quad i \neq j. \tag{12.5.1}$$

If none of the elements of this set is a zero vector, then it is called a *proper orthogonal set.* The set $\{\phi_i(x)\}$ is *orthonormal* if

$$(\phi_i, \phi_j) = \int_a^b \phi_i(x)\phi_j(x)\, dx = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \tag{12.5.2}$$

Any function $\phi(x)$ for which $||\phi(x)|| = 1$ is said to be *normalized.*

Some examples of the complete orthogonal and orthonormal systems.

(i) The system $\phi_n(x) = (2\pi)^{-1/2} e^{inx}$, where n takes every integer value from $-\infty$ to ∞, is orthonormal in the interval $(-\pi, \pi)$.

(ii) The functions $1,\ \cos x,\ \cos 2x,\ \cos 3x, \ldots$ form an orthogonal system in the interval $(0, \pi)$. Again the functions $\sin x,\ \sin 2x,\ \sin 3x, \ldots$ also form an orthogonal system in $(0, \pi)$.

(iii) The Legendre polynomials given by

$$P_0(x) = 1, \quad P_n(x) = \frac{1}{2^n n!} \frac{d^n (x^2 - 1)^n)}{dx^n}, \quad n = 1, 2, 3, \ldots$$

are orthogonal in the interval $(-1, 1)$. It can be shown that

$$\int_{-1}^{1} P_m(x) P_n(x)\, dx = \begin{cases} 0, & \text{if } m \neq n, \\ 2/(2n+1), & \text{if } m = n. \end{cases}$$

(iv) The Chebychev polynomials $T_n(x) = 2^{1-n} \cos(n\ \cos^{-1} x),\ n = 0, 1, 2, 3, \ldots$ are orthogonal with weight $r(x) = 1/(1 - x^2)^{1/2}$ in the interval $(-1, 1)$. They can be normalized by multiplying $T_n(x)$ by the quality $(2^{2n-1}/\pi)^{1/2}$.

12.5.3 Fundamental properties of eigenvalues and eigenfunctions of symmetric kernels

Theorem 12.5. *If a kernel is symmetric then all its iterated kernels are also symmetric.*

Proof. Let kernel $K(x,t)$ be symmetric. Then by definition

$$K(x,t) = \overline{K}(t,x). \tag{12.5.3}$$

By definition, the iterated kernels $K_n(x,t),\ n = 1,2,3,\ldots$ are defined as follows:

$$K_1(x,t) = K(x,t) \tag{12.5.4}$$

$$K_n(x,t) = \int_a^b K(x,z)K_{n-1}(z,t)\,dz,\ n = 2,3,\ldots \tag{12.5.5}$$

$$\text{and } K_n(x,t) = \int_a^b K_{n-1}(x,z)K(z,t)\,dz,\ n = 2,3,\ldots \tag{12.5.6}$$

We shall use mathematical induction to prove the required result. Now

$$\begin{aligned} K_2(x,t) &= \int_a^b K(x,z)K_1(z,t)\,dz = \int_a^b K(x,z)K(z,t)\,dz \\ &= \int_a^b \overline{K}(z,x)\overline{K}_1(t,z)\,dz = \int_a^b \overline{K}(t,z)\overline{K}_1(z,x)\,dz = \overline{K}_2(t,x) \end{aligned}$$

Thus,

$$K_2(x,t) = \overline{K}_2(t,x) \tag{12.5.7}$$

showing that $K_2(x,t)$ is symmetric. Hence the required result is true for $n = 1$ and $n = 2$.

Let $K_n(x,t)$ be symmetric for $n = m$. Then by definition, we have

$$K_m(x,t) = \overline{K}_m(t,x). \tag{12.5.8}$$

We shall now prove that $K_n(x,t)$ is also symmetric for $n = m+1$, i.e.,

$$K_{m+1}(x,t) = \overline{K}_{m+1}(t,x). \tag{12.5.9}$$

$$\begin{aligned} \text{L.H.S of (12.5.9)} = K_{m+1}(x,t) &= \int_a^b K(x,z)K_m(z,t)\,dz = \int_a^b \overline{K}(z,x)\overline{K}_m(t,z)\,dz \\ &= \int_a^b \overline{K}_m(t,z)\overline{K}(z,x)\,dz = \overline{K}_{m+1}(t,x) = \text{R.H.S of (12.5.9)} \end{aligned}$$

Thus iterated Kernel $K_n(x,t)$ is symmetric for $n = 1$ and $n = 2$. Moreover, $K_n(x,t)$ is symmetric for $n = m+1$ whenever it is symmetric for $n = m$. Hence, by the mathematical induction, $K_n(x,t)$ is symmetric for $n = 1,2,3,\ldots$.

□

Theorem 12.6. Hilbert Theorem

Every symmetric kernel with a norm not equal to zero has at least one eigenvalue.

OR, *If the kernel $K(x,t)$ is symmetrical and not identically equal to zero, then it has at least one eigenvalue.*

Theorem 12.7. *The eigenvalues of a symmetric kernel are real.* OR, *If K(x,t) is real, symmetric, continuous and identically not equal to zero, then all the characteristic constants (eigenvalues) are real.*

Proof. Let λ an $\phi(x)$ be an eigenvalue and a corresponding eigenfunction of the kernel $K(x,t)$. Then by definition

$$\phi(x) = \lambda \int_a^b K(x,t)\phi(t)\, dt \tag{12.5.10}$$

Multiplying (12.5.10) by $\overline{\phi}(x)$ and integrating with respect x from to x from a to b.

$$\int_a^b \phi(x)\overline{\phi}(x)\, dx = \lambda \int_a^b \left\{ \int_a^b K(x,t)\phi(t)\, dt \right\} \overline{\phi}(x)\, dx \tag{12.5.11}$$

By definition of Fredholm operator K, we have

$$K\phi = \int_a^b K(x,t)\phi(t)\, dt \tag{12.5.12}$$

$$\text{Also,} \quad ||\phi(x)|| = \int_a^b \left\{ \phi(x)\overline{\phi}(x)\, dx \right\}^{1/2} \tag{12.5.13}$$

Using (12.5.12) and (12.5.13) and the definition of inner product, (12.5.11) reduces to

$$||\phi(x)||^2 = \lambda(K\phi, \phi) \quad \text{so that} \quad \lambda = ||\phi(x)||^2/(K\phi, \phi)$$

Since bothe the numerator and denominator are real, it follows that λ is also real and thus the required result is proved. □

Theorem 12.8. *The eigenfunctions of a symmetric kernel, corresponding to different eigenvalues are orthogonal.* OR *The fundamental functions (i.e. eigenfunctions) $\phi_m(x)$ and $\phi_n(x)$ of the symmetric kernel $K(x,t)$ for corresponding eigenvalues λ_m and λ_n ($\lambda_m \neq \lambda_n$) are orthogonal in the domain (a,b).*

Proof. Since $\phi_m(x)$ and $\phi_n(x)$ are eigenfunctions corresponding to eigenvalues λ_m and λ_n respectively, where $\lambda_m \neq \lambda_n$. Then, by definition, we have

$$\phi_m(x) = \lambda_m \int_a^b K(x,t)\phi_m(t)\, dt \tag{12.5.14}$$

$$\text{and} \quad \phi_n(x) = \lambda_n \int_a^b K(x,t)\phi_n(t)\, dt \tag{12.5.15}$$

$$\text{Since } \lambda_n \text{ is real, (12.5.15) may be re-written as} \qquad \overline{\phi}(x) = \lambda_n \int_a^b \overline{K}(x,t)\overline{\phi}_n(t)\, dt \tag{12.5.16}$$

$$\text{Since } K(x,t) \text{ is symmetric, we have} \qquad \overline{K}(x,t) = K(x,t) \tag{12.5.17}$$

$$\text{Using (12.5.17), (12.5.16) may be re-written as} \qquad \overline{\phi}(x) = \lambda_n \int_a^b K(t,x)\overline{\phi}_n(t)\, dt \tag{12.5.18}$$

$$\text{Interchanging } x \text{ and } t \text{ in (12.5.18)}, wehave \qquad \overline{\phi}_n(t) = \lambda_n \int_a^b K(x,t)\overline{\phi}_n(x)\, dx \tag{12.5.19}$$

Multiplying both sides of (12.5.14) by $\overline{\phi}_n(x)$ and then integrating the both sides w.r.t. 'x' from a to b, we have

$$\begin{aligned}\int_a^b \phi_m(x)\overline{\phi}_n(x)\,dx &= \lambda_m \int_a^b \left\{\int_a^b K(x,t)\phi_m(t)\,dt\right\}\overline{\phi}_n(x)\,dx \\ &= \lambda_m \int_a^b \left\{\int_a^b K(x,t)\phi_n(x)\,dx\right\}\overline{\phi}_m(t)\,dt \quad \text{[on changing the order of integration]} \\ &= (\lambda_m/\lambda_n)\int_a^b \phi_m(t)\overline{\phi}_n(t)\,dt, \quad \text{by Eq.(12.5.19)}\end{aligned}$$

$$\therefore \quad \int_a^b \phi_m(x)\overline{\phi}_n(x)\,dx = \lambda_m \int_a^b \phi_m(x)\overline{\phi}_n(x)\,dx$$

$$\Rightarrow (\lambda_n - \lambda_m)\int_a^b \phi_m(x)\overline{\phi}_n(x)\,dx = 0,$$

$$\Rightarrow (\lambda_n - \lambda_m)(\phi_m, \phi_n) = 0$$

Since $\lambda_n \neq \lambda_m$, $(\lambda_n - \lambda_m) \neq 0$ and so we have $(\phi_m, \phi_n) = 0$, showing that the eigenfunctions ϕ_m and ϕ_n are orthogonal. □

12.5.4 Hilbert-Schmidt Theorem

Theorem 12.9. Let $F(x)$ be generated from a continuous function $y(x)$ y the operator

$$\lambda \int_a^b K(x,t)\,y(t)\,dt$$

where $K(x,t)$ is continuous, real and symmetric, so that

$$F(x) = \lambda \int_a^b K(x,t)\,y(t)\,dt.$$

Then $F(x)$ can be represented over interval (a,b) by a linear combination of the normalized eigenfunctions of homogeneous integral equation

$$y(x) = \lambda \int_a^b K(x,t)\,y(t)\,dt,$$

having $K(x,t)$ as its kernel.

Result 12.10. Schmidt's Solution of non-homogeneous fredholm integral equation of second kind

Let

$$y(x) = f(x) + \lambda \int_a^b K(x,t)\,y(t)\,dt \tag{12.5.20}$$

be a non-homogeneous Fredholm integral equation of the second kind in which $K(x,t)$ is continuous, real and symmetric and λ is not an eigenvalue. Then the solution of Eq.(12.5.20) may be expressed as

$$y(x) = f(x) + \lambda \sum_m \frac{f_m}{\lambda_m - \lambda}\phi_m(x) \tag{12.5.21}$$

where $f_m = \int_a^b f(t)\phi_m(t)\,dt$. Solution exists *uniquely* if and only if λ does not take on an eigenvalue. If $\lambda = \lambda_k$, where λ_k is the k-th eigenvalue and eigenfunction $\phi_k(x)$ is not orthogonal to $f(x)$, then *no solution* exists. Finally, if $\lambda = \lambda_k$ and eigenfunction $\phi_k(x)$ is orthogonal to $f(x)$, then we have *infinitely many solutions of Eq.*(12.5.20).

Example 12.11. Solve the symmetric integral equation

$$y(x) = (x+1)^2 + \int_{-1}^{1} (xt + x^2t^2) y(t)\, dt,$$

by using Hilbert-Schmidt theorem.

Solution:

$$\text{Given} \quad y(x) = (x+1)^2 + \int_{-1}^{1} (xt + x^2t^2) y(t)\, dt, \tag{12.5.22}$$

$$\text{Comparing (12.5.22) with} \quad y(x) = f(x) + \lambda \int_{-1}^{1} (xt + x^2t^2) y(t)\, dt, \tag{12.5.23}$$

$$\text{here} \quad f(x) = (x+1)^2 \qquad \text{and} \qquad \lambda = 1 \tag{12.5.24}$$

We begin with determining eigenvalues and the corresponding normalized eigenfunctions of

$$y(x) = \lambda \int_{-1}^{1} (xt + x^2t^2) y(t)\, dt \tag{12.5.25}$$

$$\Rightarrow y(x) = \lambda x \int_{-1}^{1} t y(t)\, dt + \lambda x^2 \int_{-1}^{1} t^2 y(t)\, dt.$$

$$\text{Let} \quad C_1 = \int_{-1}^{1} t y(t)\, dt \quad \text{and} \quad C_2 = \int_{-1}^{1} t^2 y(t)\, dt$$

Thus we have from Eq.(12.5.25)

$$y(x) = \lambda C_1 x + \lambda C_2 x^2 \quad \Rightarrow \quad y(t) = \lambda C_1 t + \lambda C_2 t^2 \tag{12.5.26}$$

$$\begin{aligned} \text{Hence} \quad C_1 &= \int_{-1}^{1} t(\lambda C_1 + \lambda C_2 t^2)\, dt \quad \Rightarrow \quad C_1 = C_1\lambda \left[\frac{t^3}{3}\right]_{-1}^{1} + C_2\lambda \left[\frac{t^4}{4}\right]_{-1}^{1} \\ \Rightarrow C_1 &= \frac{2C_1\lambda}{3} + 0 \quad \Rightarrow \quad C_1\left(1 - \frac{2\lambda}{3}\right) + 0.C_2 = 0. \end{aligned} \tag{12.5.27}$$

$$\begin{aligned} \text{Again} \quad C_2 &= \int_{-1}^{1} t^2(\lambda C_1 + \lambda C_2 t^2)\, dt \quad \Rightarrow \quad C_2 = C_1\lambda \left[\frac{t^4}{4}\right]_{-1}^{1} + C_2\lambda \left[\frac{t^5}{5}\right]_{-1}^{1} \\ \Rightarrow C_2 &= 0 + \frac{2C_2\lambda}{5} \quad \Rightarrow \quad 0.C_1 + \left(1 - \frac{2\lambda}{5}\right) C_2 = 0. \end{aligned} \tag{12.5.28}$$

Equations (12.5.27) and (12.5.28) have a nontrivial solution only if

$$D(\lambda) = \begin{vmatrix} 1 - (2\lambda/3) & 0 \\ 0 & 1 - (2\lambda/5) \end{vmatrix} = 0$$

$$\Rightarrow \{1 - (2\lambda/3)\}\{1 - (2\lambda/5)\} = 0 \quad \text{giving} \quad \lambda = 3/2 \quad \text{and} \quad \lambda = 5/2$$

Hence the required eigenvalues are $\lambda_1 = 3/2$ and $\lambda_2 = 5/2$.

Determination of eigenfunction corresponding to $\lambda_1 = 3/2$

Putting $\lambda = \lambda_1 = 3/2$ in (12.5.27) and (12.5.28), we obtain

$$C_1 \cdot 0 + 0 \cdot C_2 = 0 \quad \text{and} \quad 0 \cdot C_1 + \left[1 - \left(\frac{2}{5} \times \frac{3}{2}\right)\right] C_2 = 0$$

Hence $C_2 = 0$ and C_1 is arbitrary. Putting these values in (12.5.26) and noting that $\lambda = 3/2$, we have the required eigenfunction $y_1(x)$ is given by

$$y(x) = (3/2) \times C_1 x$$

Setting $(3/2) \times C_1 = 1$, we may take $y_1(x) = x$. Now, the corresponding normalized eigenfunction $\phi_1(x)$ is given by

$$\phi_1(x) = \frac{y_1(x)}{\left[\int_{-1}^{1} \{y_1(x)\}^2\right]^{1/2}} = \frac{x}{\left[\int_{-1}^{1} x^2 dx\right]^{1/2}} = \frac{x}{\left\{[x^3/3]_{-1}^{1}\right\}^2} = \frac{x}{\sqrt{(2/3)}} = x \times \left(\frac{3}{2}\right)^{1/2} = \frac{x\sqrt{6}}{2}.$$

Determination of eigenfunction corresponding to $\lambda_1 = 5/2$

Putting $\lambda = \lambda_1 = 5/2$ in (12.5.27) and (12.5.28), we obtain

$$\left[1 - \left(\frac{2}{3} \times \frac{5}{2}\right)\right] C_1 + 0 \cdot C_2 = 0 \quad \text{and} \quad 0 \cdot C_1 + 0 \cdot C_2 = 0$$

Hence $C_1 = 0$ and C_2 is arbitrary. Putting these values in (12.5.26) and noting that $\lambda = 5/2$, we have the required eigenfunction $y_2(x)$ is given by

$$y(x) = (5/2) \times C_2 x^2$$

Setting $(5/2) \times C_2 = 1$, we may take $y_2(x) = x^2$. Now, the corresponding normalized eigenfunction $\phi_2(x)$ is given by

$$\phi_2(x) = \frac{y_2(x)}{\left[\int_{-1}^{1} \{y_2(x)\}^2\right]^{1/2}} = \frac{x^2}{\left[\int_{-1}^{1} x^4 dx\right]^{1/2}} = \frac{\sqrt{10}}{2} x^2.$$

Also,

$$\begin{aligned} f_1 &= \int_{-1}^{1} f(x)\phi_1(x)\, dx = \int_{-1}^{1} (x+1)^2 \left(\frac{\sqrt{6}}{2} x\right) dx \\ &= \frac{\sqrt{6}}{2} \int_{-1}^{1} (x^2 + 2x + 1)x\, dx = \frac{2\sqrt{6}}{3} \\ f_2 &= \int_{-1}^{1} f(x)\phi_2(x)\, dx = \int_{-1}^{1} (x+1)^2 \left(\frac{\sqrt{10}}{2} x^2\right) dx = \frac{8}{15}\sqrt{10} \end{aligned} \tag{12.5.29}$$

Now we have $\lambda = 1$. Also $\lambda_1 = 3/2$ and $\lambda_2 = 5/2$. Hence $\lambda \neq \lambda_1$ and $\lambda \neq \lambda_2$. Therefore the unique solution

given by

$$\begin{aligned} y(x) &= f(x) + \lambda \sum_{m=1}^{2} \frac{f_m}{\lambda_m - \lambda} \phi_m(x) \\ \Rightarrow y(x) &= (x+1)^2 + \frac{f_1\phi_1(x)}{\lambda_1 - 1} + \frac{f_2\phi_2(x)}{\lambda_2 - 1} \\ \Rightarrow y(x) &= (x+1)^2 + \frac{(2\sqrt{6}/3) \times (x\sqrt{6}/2)}{(3/2) - 1} + + \frac{(8\sqrt{10}/15) \times (x^2\sqrt{10}/2)}{(5/2) - 1} \\ \Rightarrow y(x) &= (x+1)^2 + 4x + (16/9) \times x^2 = x^2 + 2x + 1 + 4x + (16/9) \times x^2 \\ \Rightarrow y(x) &= \frac{25}{9}x^2 + 6x + 1. \end{aligned}$$

Exercise 12.12. Using Hilbert-Schmidt theorem, find the solution of the symmetric integral equation

$$i)\ y(x) = x^2 + 1 + \frac{3}{2}\int_{-1}^{1} (xt + x^2t^2)\ y(t)\ dt, \qquad ii)\ y(x) = 1 + \int_0^{\pi} \cos(x+t)\ y(t)\ dt,$$

Answers :

$$i)\ y(x) = 5x^2 + Cx + 1, \text{ where } C \text{ is constant} \qquad (ii)\ y(x) = 1 + C\ \cos x - (2/\pi)\sin x,$$

Unit 13

Course Structure

Integral Transforms. Laplace Transform : Definition and basic properties. Laplace integral. Lerch's theorem (statement only). Laplace transforms of elementary functions, of derivatives and Dirac-delta function. Differentiation and integration. Convolution. Inverse transform. Applications to solve ordinary differential equations.

---

13 Introduction

The integral transform methods are very convenient in solving integral equations of some special forms. Suppose that a relationship of the form

$$y(x) = \int_a^b \int_a^b \Gamma[x,z]K(z,t)y(t)\,dt\,dz \tag{13.0.1}$$

be known to be valid and that this double integral can be evaluated as an iterated integral. Then from (13.0.1), it follows that if

$$F(x) = \int_a^b K(x,t)\,y(t)\,dt \tag{13.0.2}$$

we also have

$$y(x) = \int_a^b \Gamma(x,t)\,F(t)\,dt. \tag{13.0.3}$$

Thus, if Eq.(13.0.2) is an integral equation in y, a solution is given by Eq., whereas if Eq.(13) is regarded as an integral equation in F a solution is given by Eq.(13.0.2). It is conventional to refer to one of the function as the *transform* of the second function, and to the second function as an *inverse transform* of the first.

Definition 13.1. Function of exponential order: A function $f(x)$ is said to be of exponential order a as $x \to \infty$ if

$$\lim_{x\to\infty} e^{-ax} f(x) = \text{finite quantity}$$

i.e., if given a positive integer n_0, there exists a real number $M > 0$ s.t.

$$|e^{-ax} f(x)| < M \quad \forall x \geq n_0 \quad \text{or} \quad |f(x)| < Me^{ax} \quad \forall x \geq n_0$$

Example 13.2. Show that x^n is of exponential order as $x \to \infty$, n being any positive integer.

Solution:

$$\lim_{x\to\infty} e^{-ax}x^n = \lim_{x\to\infty} \frac{x^n}{e^{ax}} = \lim_{x\to\infty} \frac{n!}{a^n e^{ax}} = \frac{n!}{\infty} = 0$$

Example 13.3. Show that $F(t) = e^{t^2}$ is not of exponential order as $t \to \infty$.

Solution:

$$\lim_{t\to\infty} e^{-at}F(t) = \lim_{t\to\infty} e^{t(t-a)} = \infty$$

Hence $F(t)$ is not of exponential order.

Definition 13.4. Laplace transform: Suppose $F(t)$ is a real valued function defined over the interval $(-\infty, \infty)$ such that $F(t) = 0$. The *Laplace transform* of $F(t)$, denoted by $L\{F(t)\}$, is defined as

$$L\{F(t)\} = f(s) = \int_0^\infty e^{-st} F(t)\, dt \tag{13.0.4}$$

Sometimes we use symbol p for the parameter s. The Laplace transform is said to exist if the integral (13.0.4) is convergent for some value of s.

Table of Laplace transform of some elementary functions

Serial Number	F(t)	L{F(t)} or $\overline{F}(p)$ or $f(p)$
1	1	$1/p,\ p > 0$
2	$t^n,\ n > -1$	$\Gamma(n+1)/p^{n+1},\ p > 0$
3	t^n (n is positive integer)	$n!/p^{n+1},\ p > 0$
4	e^{at}	$1/(p-a),\ p > a$
5	$\sin at$	$a/(p^2+a^2),\ p > 0$
6	$\cos at$	$p/(p^2+a^2),\ p > 0$
7	$\sinh at$	$a/(p^2-a^2),\ p > \lvert a\rvert$
8	$\cosh at$	$p/(p^2-a^2),\ p > \lvert a\rvert$

Theorem 13.5. Linear Property: Suppose $f_1(s)$ and $f_2(s)$ are Laplace forms of $F_1(t)$ and $F_2(t)$ respectively. Then

$$L\{c_1F_1(t) + c_2F_2(t)\} = c_1L\{F_1(t)\} + c_2L\{F_2(t)\}$$

where c_1 and c_2 are constants.

Proof. Let

$$L\{F_1(t)\} = f_1(s) = \int_0^\infty e^{-st}F_1(t)\, dt \quad \text{and} \quad L\{F_2(t)\} = f_2(s) = \int_0^\infty e^{-st}F_2(t)\, dt$$

Also let c_1 and c_2 be arbitrary constants. Now

$$\begin{aligned} L\{c_1F_1(t) + c_2F_2(t)\} &= \int_0^\infty e^{-st}\Big[c_1F_1(t) + c_2F_2(t)\Big]\, dt \\ &= c_1\int_0^\infty e^{-st}F_1(t)\, dt \ + \ c_2\int_0^\infty e^{-st}F_2(t)\, dt \\ &= c_1L\{f_1(t)\} + c_2L\{f_2(t)\} \end{aligned}$$

□

Theorem 13.6. First Shifting Theorem (First Translation): If $L\{F(t)\} = f(s)$, then $L\{e^{at}F(t)\} = f(s-a)$.

Proof. We know that

$$\begin{aligned} L\{e^{at}F(t)\} &= \int_0^\infty e^{-st}e^{at}F(t)\, dt = \int_0^\infty e^{-(s-a)t}F(t)\, dt = \int_0^\infty e^{-ut}F(t)\, dt, \text{ where } u = s-a > 0 \\ &= f(u) = f(s-a) \end{aligned}$$

□

Example 13.7. Find $L\{e^{-t}(3\sinh 2t - 5\cosh 2t)\}$

Solution: We know that

$$L\{\sinh 2t\} = \frac{2}{s^2 - 2^2}, \qquad L\{\cosh 2t\} = \frac{s}{s^2 - 2^2}$$

Therefore

$$L\{e^{-t}(3\sinh 2t - 5\cosh 2t)\} = 3\frac{2}{(s+1)^2 - 2^2} - 5\frac{s+1}{(s+1)^2 - 2^2} = \frac{1 - 5s}{s^2 + 2s - 3}$$

Theorem 13.8. Second Shifting Theorem (Second Translation)

$$\text{If } L\{F(t)\} = f(s) \text{ and } G(t) = \begin{cases} F(t-a) & t > a \\ 0 & t < a \end{cases} \quad \text{Then } L\{G(t)\} = e^{-as} f(s)$$

Proof.

$$\text{Let } L\{F(t)\} = f(s) \text{ and } G(t) = \begin{cases} F(t-a) & t > a \\ 0 & t < a \end{cases}$$

Now

$$\begin{aligned} L\{G(t)\} &= \int_0^\infty e^{-st} G(t)\, dt = \int_0^a e^{-st} G(t)\, dt + \int_a^\infty e^{-st} G(t)\, dt \\ &= \int_0^a e^{-st} \cdot 0\, dt + \int_0^\infty e^{-st} F(t-a)\, dt = 0 + \int_a^\infty e^{-st} F(t-a)\, dt \end{aligned}$$

Now putting $t - a = p$ so that $dt = dp$. If $t = a$, then $p = t - a = a - a = 0$ and if $t = \infty$, then $p = \infty - a = \infty$.

$$\therefore L\{G(t)\} = \int_0^\infty e^{-s(p+a)} F(p)\, dp = e^{-sa} \int_0^\infty e^{-sp} F(p)\, dp = e^{-sa} f(s)$$

□

Example 13.9. Find the Laplace transform of $F(t)$, where $F(t) = \begin{cases} \cos\left(t - \frac{2\pi}{3}\right) & \text{if } t > \frac{2\pi}{3} \\ 0 & \text{if } t > \frac{2\pi}{3} \end{cases}$

Solution: Let $a = \frac{2\pi}{3}$, and $G(t) = \cos t$, then

$$L\{G(t)\} = \frac{p}{p^2 + 1} = g(p), \quad \text{as } L\{\cos at\} = \frac{p}{p^2 + a^2}. \text{ Also } F(t) = \begin{cases} G(t-a) & t > a \\ 0 & t < a \end{cases}$$

By second shifting theorem,

$$L\{F(t)\} = e^{-ap} g(p) = e^{-\frac{2\pi p}{3}} \frac{p}{p^2 + 1}$$

Theorem 13.10. Change of Scale Property

$$\text{If } L\{F(t)\} = f(s), \quad \text{then } L\{F(at)\} = \frac{1}{a} f\left(\frac{s}{a}\right)$$

Proof.

$$\begin{aligned}\text{Let } L\{F(t)\} = f(s), \text{ then } L\{F(at)\} &= \int_0^\infty e^{-st}F(at)\,dt = \int_0^\infty e^{-\frac{sx}{a}}F(x)\frac{dx}{a}, \text{ Putting } x = at \\ &= \frac{1}{a}\int_0^\infty e^{-\frac{sx}{a}}F(x)\,dx = \frac{1}{a}\int_0^\infty e^{-\frac{st}{a}}F(t)\,dt \\ &= \frac{1}{a}\int_0^\infty e^{-pt}F(t)\,dt, \text{ where } p = \frac{s}{a} \;= \frac{1}{a}f(p) = \frac{1}{a}f\left(\frac{s}{a}\right)\end{aligned}$$

□

Example 13.11. If $L\{\cos^2 t\} = \frac{s^2+2}{s(s^2+4)}$, then find $L\{cos^2(at)\}$. **Answer:** $\frac{(s^2+2a^2)}{s(s^2+4a^2)}$

13.1 Laplace Transform of Derivatives

Theorem 13.12. If $L\{F(t)\} = f(s)$, then $L\{F^{(n)}(t)\} = s^n f(s) - s^{n-1}F(0) - s^{n-2}F'(0) - \ldots - sF^{(n-2)}(0) - F^{(n-1)}(0)$ where $F^{(n)}(t)$ stands for $\frac{d^nF(t)}{dt^n}$.

13.2 Laplace Transform of Integral

Theorem 13.13. If $L\{F(t)\} = f(s)$, then $\frac{1}{s}f(s) = L\left\{\int_0^t F(u)\,du\right\}$.

13.3 Multiplication by Powers of t

Theorem 13.14. If $L\{F(t)\} = f(s)$, then $L\{t^nF(t)\} = (-1)^n\frac{d^n}{ds^n}f(s)$ for $n = 1, 2, 3, \ldots$.

13.4 Division by t

Theorem 13.15. If $L\{F(t)\} = f(s)$, then $L\{\frac{F(t)}{t}\} = \int_s^\infty f(x)\,dx$.

Note: Proofs of the above theorems are left for the readers.

Example 13.16. Find $L\{\cos^2 t\}$ and $L\{\sin^2 t\}$.

Solution:

$$L\{\cos^2 t\} = \frac{1}{2}L\{1+\cos 2t\} = \frac{1}{2}\left[\frac{1}{s} + \frac{s}{s^2+2^2}\right] = \frac{(s^2+2)}{s(s^2+4)}$$
$$L\{\sin^2 t\} = \frac{1}{2}L\{1-\cos 2t\} = \frac{1}{2}\left[\frac{1}{s} - \frac{s}{s^2+2^2}\right] = \frac{2}{s(s^2+4)}$$

Example 13.17. Using Laplace transform prove that (i) $\int_0^\infty \left(\frac{\sin t}{t}\right)dt = \frac{\pi}{2}$ and (ii) $\int_0^\infty te^{-3t}\sin t\,dt = \frac{3}{50}$.

Solution: (i) We know, $L\{\sin t\} = \frac{1}{p^2+1}$, then

$$L\left\{\frac{\sin t}{t}\right\} = \int_p^\infty \frac{dp}{1+p^2} = \Big(\tan^{-1}p\Big)_p^\infty = \frac{\pi}{2} - \tan^{-1}(p) = \cot^{-1}(p) = \tan^{-1}\left(\frac{1}{p}\right)$$

Putting $p = 0$, we get $\int_0^\infty \left(\frac{\sin t}{t}\right)\,dt = \tan^{-1}(\infty) = \frac{\pi}{2}$.

(ii) Since $L\{\sin t\} = \frac{1}{p^2+1}$ and so $L\{t\sin t\} = (-1)^1\frac{d}{dp}\left(\frac{1}{p^2+1}\right) = \frac{2p}{(p^2+1)^2}$

$$\Rightarrow \int_0^\infty e^{-pt}(t\,\sin t)\,dt = \frac{2p}{(p^2+1)^2} \quad \text{Putting } p = 3, \text{ we have } \int_0^\infty e^{-3t}(t\,\sin t)\,dt = \frac{3}{50}$$

Example 13.18. Find $L\{F_\varepsilon(t)\}$ where $F_\varepsilon(t)$ is dirac delta function.

Solution:

$$\begin{aligned} F_\varepsilon(t) &= \begin{cases} 1/\varepsilon & \text{if } 0 \le t \le \varepsilon \\ 0 & \text{if } t > \varepsilon \end{cases} \\ L\{F_\varepsilon(t)\} &= \int_0^\infty e^{-st} F_\varepsilon(t)\, dt = \int_0^\varepsilon e^{-st} F_\varepsilon(t)\, dt + \int_\varepsilon^\infty e^{-st} F_\varepsilon(t)\, dt \\ &= \int_0^\varepsilon \frac{e^{-st}}{\varepsilon} dt + \int_\varepsilon^\infty e^{-st} \cdot 0 \cdot dt = \frac{1}{\varepsilon}\left[\frac{e^{-st}}{-s}\right]_{t=0}^{\varepsilon} + 0 \\ &= \frac{1}{\varepsilon s}(1 - e^{-\varepsilon s}). \end{aligned} \tag{13.4.1}$$

13.5 Inverse Laplace transform

Let $L\{F(t)\} = \overline{F}(p)$. Then $F(t)$ is called an *inverse Laplace transform* of $\overline{F}(p)$, and we write $F(t) = L^{-1}\{\overline{F}(p)\}$, in which L^{-1} is known as the *inverse Laplace transformation operator.*

Theorem 13.19. Lerch's Theorem: Let $L\{F(t)\} = f(s)$. Let $F(t)$ be piecewise continuous in every finite interval $0 \le t \le a$ and of exponential order for $t > a$, then the inverse Laplace transform of $F(t)$ is unique $f(s)$.

Table of inverse Laplace transform of some elementary functions

Serial Number	$\overline{F}(p)$	$L^{-1}\{\overline{F}(p)\}$
1	$1/p$	1
2	$1/p^{n+1}, n > -1$	$t^n/\Gamma(n+1)$
3	$1/p^{n+1}$ (n is positive integer)	$t^n/n!$
4	$1/(p-a)$	e^{at}
5	$1/(p^2+a^2)$	$(\sin at)/a$
6	$p/(p^2+a^2)$	$\cos at$
7	$1/(p^2-a^2)$	$(\sin at)/a$
8	$p/(p^2-a^2)$	$\cos at$

Theorem 13.20. Inverse Laplace transform of derivatives: If $L^{-1}\{f(s)\} = F(t)$, then $L^{-1}\{f^{(n)}(s)\} = (-1)^n t^n F(t)$

Theorem 13.21. First Shifting theorem: If $L^{-1}\{f(s)\} = F(t)$, then $L^{-1}\{f(s-a)\} = e^{at}F(t)$.

Theorem 13.22. Second Shifting theorem: If $L^{-1}\{f(s)\} = F(t)$, then $L^{-1}\{e^{-as}f(s)\} = G(t)$, where $G(t) = \begin{cases} F(t-a) & \text{if } t > a \\ 0 & \text{if } t < a \end{cases}$

Example 13.23.

$$\text{Find} \quad L^{-1}\left\{\frac{s-2}{(s-2)^2+5^2} + \frac{s+4}{(s+4)^2+9^2} + \frac{1}{(s+2)^2+3^2}\right\}$$

Solution:

$$
\begin{aligned}
&L^{-1}\left\{\frac{s-2}{(s-2)^2+5^2}+\frac{s+4}{(s+4)^2+9^2}+\frac{1}{(s+2)^2+3^2}\right\}\\
&=L^{-1}\left\{\frac{s-2}{(s-2)^2+5^2}\right\}+L^{-1}\left\{\frac{s+4}{(s+4)^2+9^2}\right\}+L^{-1}\left\{\frac{1}{(s+2)^2+3^2}\right\}\\
&=e^{2t}L^{-1}\left\{\frac{s}{s^2+5^2}\right\}+e^{-4t}L^{-1}\left\{\frac{s}{s^2+9^2}\right\}+e^{-2t}L^{-1}\left\{\frac{1}{s^2+3^2}\right\}\\
&=e^{2t}\cos 5t+e^{-4t}\cos 9t+\frac{e^{-2t}}{3}\sin 3t
\end{aligned}
$$

Example 13.24. *Find* $L^{-1}\left\{\frac{e^{4-3p}}{(p+4)^{5/2}}\right\}$

Solution:

$$L^{-1}\left[\frac{1}{(p+4)^{5/2}}\right]=e^{-4t}L^{-1}\{\frac{1}{(p-4+4)^{5/2}}\}=e^{-4t}L^{-1}\left\{\frac{1}{p^{(3/2)+1}}\right\}=e^{-4t}\frac{t^{3/2}}{\Gamma(5/2)}$$

But $\Gamma\left(\frac{5}{2}\right)=\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi}=\frac{3}{4}\sqrt{\pi}$,

Therefore, $L^{-1}\left[\frac{1}{(p+4)^{5/2}}\right]=\frac{4}{3\sqrt{\pi}}e^{-4t}t^{3/2}\ \Rightarrow\ L^{-1}\left[\frac{e^{4-3p}}{(p+4)^{5/2}}\right]=e^4L^{-1}\left[\frac{e^{-3p}}{(p+5)^{5/2}}\right]$

Using second shifting theorem

$$L^{-1}\left[\frac{e^{4-3p}}{(p+4)^{5/2}}\right]=\begin{cases}\frac{4e^4}{3\sqrt{\pi}}e^{-4(t-3)}(t-3)^{3/2} & \text{if } t>3\\ 0 & \text{if } t<4\end{cases}=\frac{4e^4}{3\sqrt{\pi}}e^{-4(t-3)}(t-3)^{3/2}H(t-3).$$

Note 13.25. In the above example $H(t)=\begin{cases}1 & \text{if } t>0\\ 0 & \text{if } t<0\end{cases}$ is *Heaviside unit step function.*

Definition 13.26. Convolution (or Faltung): The convolution of $F(t)$ and $G(t)$ is denoted and defined as

$$F*G=\int_0^t F(x)G(t-x)\,dx \quad \text{or} \quad F*G=\int_0^t F(t-x)G(x)\,dx$$

Theorem 13.27. Convolution theorem or Convolution property: If $L^{-1}\{\overline{F}(p)\}=F(t)$ and $L^{-1}\{\overline{G}(p)\}=G(t)$, then $L^{-1}\{\overline{F}(p)\overline{G}(p)\}=\int_0^t F(x)G(t-x)\,dx=F*G$ or $L^{-1}\{\overline{F}(p)\overline{G}(p)\}=\int_0^t F(t-x)G(x)\,dx=F*G$.

Example 13.28. Find $L^{-1}\left\{\frac{1}{(s+a)(s+b)}\right\}$.

Solution: Let $f(s)=\frac{1}{s-a}$, $g(s)=\frac{1}{s+b}$. Then $F(t)=e^{-at}$, $G(t)=e^{-bt}$. Hence

$$
\begin{aligned}
L^{-1}\{f(s)\cdot g(s)\} &= \int_0^t F(u)G(t-u)\,du\\
&= \int_0^t e^{-au}e^{-b(t-u)}\,du=e^{-bt}\int_0^t e^{u(b-a)}du\\
&= e^{-bt}\left[\frac{e^{u(b-a)}}{b-a}\right]_{u=0}^{t}=\frac{e^{-bt}}{b-a}[e^{t(b-a)}-1]=\frac{e^{-at}-e^{=bt}}{b-a}
\end{aligned}
$$

Example 13.29. Apply the convolution theorem to show that

$$\int_0^t \sin u \cos(t-u)\, du = \frac{1}{2}(t \sin t).$$

Solution: Let $F(t) = \int_0^t \sin u \cos(t-u)\, du$. Then, by convolution theorem,

$$L\{F(t)\} = L\{\sin t\}L\{\cos t\} = \frac{1}{s^2+1} \cdot \frac{s}{s^2+1} = \frac{s}{(s^2+1)^2}$$
$$\therefore F(t) = L^{-1}\left\{\frac{s}{(s^2+1)^2}\right\}$$

In order to calculate the above inverse Laplace transform, let $F(t) = \sin t$. Therefore,

$$\begin{aligned} f'(s) = -\frac{2s}{(s^2+1)^2} &\Rightarrow L^{-1}\{f'(s)\} = L^{-1}\left\{-\frac{2s}{(s^2+1)^2}\right\} \\ &\Rightarrow (-1)^1 t^1 L^{-1}\{f(s)\} = L^{-1}\left\{-\frac{2s}{(s^2+1)^2}\right\} \\ &\Rightarrow \frac{t \sin t}{2} = L^{-1}\left\{\frac{s}{(s^2+1)^2}\right\} \end{aligned}$$

Hence $F(t) = \frac{t \sin t}{2}$.

Exercise 13.30. (i) Evaluate $L^{-1}\left\{\frac{1}{(s+1)(s+2)}\right\}$ **Answer:** $\frac{e^{2t}-e^{-t}}{3}$.

(ii) Evaluate $L^{-1}\left\{\frac{1}{(s-3)(s+4)}\right\}$ **Answer:** $\frac{1}{7}(e^{3t} - e^{-4t})$.

13.6 Some special types of integral equations

Definition 13.31. (i) Integro-differential equation: An integral equation in which various derivatives of the unknown function $y(t)$ can also be present is said to be an *integro-differential equation.* For example, the following integral equation is an integro-differential equation.

$$y'(t) = y(t) + f(t) + \int_0^t \sin(t-x)y(x)\, dx.$$

Definition 13.32. (i) Integral equation of convolution type: The integral equation

$$y(t) = f(t) + \int_0^t K(t-x)y(x)\, dx,$$

in which the kernel $K(t-x)$ is a function of the difference $(t-x)$ only, is known as integral equation of the convolution type. Using the definition of convolution, we may re-write it as

$$y(t) = f(t) + K(t) * y(t).$$

Example 13.33. Solve the integro-differential equation

$$y'(t) = \sin t + \int_0^t y(t-x)\cos x\, dx, \quad \text{where} \quad y(0) = 0$$

Solution: Rewriting the given equation, we have

$$y'(t) = \sin t + y(t) * \cos t, \qquad y(0) = 0$$

Applying the Laplace transform on both sides, we obtain

$$\begin{aligned}
&L\{y't)\} = L\{\sin t\} + L\{y(t) * \cos t\} \\
&\Rightarrow pL\{y(t)\} - y(0) = \frac{1}{p^2+1} + L\{y(t)\}L\{\cos t\} \\
&\Rightarrow \left(1 - \frac{1}{p^2+1}\right) pL\{y(t)\} = \frac{1}{p^2+1} \\
&\Rightarrow L\{y(t)\} = \frac{1}{p^3}, \quad \text{Inverting we have} \quad y(t) = L^{-1}\{\frac{1}{p^3}\} = \frac{t^2}{2!} = \frac{t^2}{2}.
\end{aligned}$$

Exercise 13.34. (i) Solve $y'(t) = t + \int_0^t y(t-x)\cos x\, dx, \quad y(0) = 4$ **Answer:** $y(t) = 4 + \frac{5}{2}t^2 + \frac{1}{24}t^4$.

Example 13.35. Solve the integral equation $y(t) = 1 + \int_0^t y(x)\sin(t-x)\, dx, \quad .$

Solution: The given integral equation can be re-written as $y(t) = 1 + y(t) * \sin t$. Applying Laplace transform, we obtain

$$\begin{aligned}
&L\{y(t)\} = L\{1\} + L\{y(t)\} \cdot L\{\sin t\} \\
&\Rightarrow L\{y(t)\} = \frac{1}{p} + L\{y(t)\} \times \frac{1}{p^2+1} \Rightarrow \left(1 - \frac{1}{p^2+1}\right) L\{y(t)\} = \frac{1}{p} \\
&\Rightarrow L\{y(t)\} = \frac{p^2+1}{p^3} = \frac{1}{p} + \frac{1}{p^3}
\end{aligned}$$

Inverting the above equation, we have

$$y(t) = L^{-1}\{y(t)\} = L^{-1}\{\frac{1}{p}\} + L^{-1}\{\frac{1}{p^3}\} = 1 + \frac{t^2}{2!} = 1 + \frac{t^2}{2}.$$

Exercise 13.36. (i) Solve $y(t) = a\, \sin t - \int_0^t y(x)\cos(t-x)\, dx$, **Answer:** $y(t) = a\, t\, e^{-t}$.
(ii) Solve $y(t) = e^{-t} - 2\int_0^t \cos(t-x)\, y(x)\, dx$ **Answer:** $y(t) = e^{-t}(1-t)^2$.
(iii) Solve $y(t) = t + 2\int_0^t \cos(t-x) y(x)\, dx$ **Answer:** $y(t) = 2e^t(t-1) + 2 + t$.
(iv) Solve $t = \int_0^t e^{t-x} y(x)\, dx$ **Answer:** $y(t) = 1 - t$.

13.7 Application to solve ordinary differential equations

Consider the differential equation

$$a\frac{d^2x}{dt^2} + b\frac{dx}{dt} + x = F(t) \tag{13.7.1}$$

Case I: When a, b are constants. Taking Laplace transform on the both sides of Eq. (13.7.1), we have

$$aL\{x''\} + bL\{x'\} + L(x) = L\{F(t)\}. \tag{13.7.2}$$

Letting $L\{x(t)\} = \overline{x}(s)$, we have from Eq. (13.7.2)

$$a\{s^2\overline{x} - s\,x(0) - x'(0)\} + b\{s\overline{x} - x(0)\} + \overline{x} = f(s)$$

Required solution is obtained by taking the inverse Laplace transform of $\overline{x}(s)$.

Example 13.37. Solve by using Laplace transformation:

$$(D^2 + 9)y = \cos(2t) \quad \text{if} \quad y(0) = 1,\ y(\pi/2) = -1$$

Solution: Taking Laplace transform, we get

$$\begin{aligned}
&p^2\overline{y} - py(0) - y'(0) + 9\overline{y} = \frac{p}{p^2 + 2^2} \\
&\Rightarrow (p^2 + 9)\overline{y} = p + a + \frac{p}{p^2 + 2^2} \quad \text{where } y'(0) = a \\
&\Rightarrow \overline{y} = \frac{p}{p^2 + 3^2} + \frac{a}{p^2 + 3^2} + \frac{p}{(p^2 + 2^2)(p^2 + 3^2)} \\
&\Rightarrow \overline{y} = \frac{p}{p^2 + 3^2} + \frac{a}{p^2 + 3^2} + \frac{1}{5}\left[\frac{p}{p^2 + 2^2} - \frac{p}{p^2 + 3^2}\right] \\
&\Rightarrow \overline{y} = \frac{4}{5}\frac{p}{p^2 + 3^2} + \frac{a}{p^2 + 3^2} + \frac{1}{5}\frac{p}{p^2 + 2^2}
\end{aligned}$$

Taking inverse Laplace transform, we get

$$\begin{aligned}
y = \frac{4}{5}\cos 3t + \frac{a}{3}\sin 3t + \frac{1}{5}\cos 2t \qquad &\Rightarrow -1 = y(\pi/2) = 0 - \frac{a}{3} - \frac{1}{5} \Rightarrow \frac{a}{3} = \frac{4}{5} \\
\therefore y = \frac{4}{5}\cos 3t + \frac{4}{5}\sin 3t + \frac{1}{5}\cos 2t.&
\end{aligned}$$

Example 13.38.

$$\text{Solve } 2\frac{d^2y}{dt^2} + 5\frac{dy}{dt} + 2y = e^{-2t},\ \ y(0) = 1,\ y'(0) = 1$$

Solution: Taking Laplace transform of the equation

$$\begin{aligned}
&(2D^2 + 5D + 2)y = e^{-2t}, \\
&\text{we get} \quad 2[s^2\overline{y} - sy(0) - y'(0)] + 5[s\overline{y} - y(0)] + 2\overline{y} = \frac{1}{s - 2} \\
&\text{Putting } \ y(0) = 1 = y'(0), \ \text{ we get} \\
&\qquad 2[s^2\overline{y} - s - 1] + 5[s\overline{y} - 1] + 2\overline{y} = \frac{1}{s - 2} \\
&\Rightarrow (2s^2 + 5s + 2)\overline{y} - 2s - 7 = \frac{1}{s - 2} \\
&\Rightarrow \overline{y} = \frac{1}{(s + 2)(2s^2 + 5s + 2)} + \frac{2s + 7}{2s^2 + 5s + 2} \\
&\Rightarrow \overline{y} = \frac{1}{(s + 2)^2(2s + 1)} + \frac{2s + 7}{(s + 2)(2s + 1)}
\end{aligned}$$

Now $L^{-1}\left\{\frac{1}{(s+2)^2}(2s+1)\right\} = e^{-2t}L^{-1}\left\{\frac{1}{s^2\{2(s-2)+1\}}\right\} = e^{-2t}L^{-1}\left\{\frac{1}{s^2(2s-3)}\right\}$

But $\frac{1}{s(2s-3)} = \frac{1}{3}\left[\frac{2}{2s-3} - \frac{1}{s}\right] = \frac{1}{3}\left[\frac{1}{s-\frac{3}{2}} - \frac{1}{s}\right]$

$\Rightarrow L^{-1}\left\{\frac{1}{s(2s-3)}\right\} = \frac{1}{3}\left[e^{3t/2}-1\right] \Rightarrow L^{-1}\left\{\frac{1}{s^2(2s-3)}\right\} = \frac{1}{3}\int_0^t \left[e^{3t/2}-1\right]dx$

$\Rightarrow L^{-1}\left\{\frac{1}{s^2(2s-3)}\right\} = \frac{2}{3}\left[e^{3t/2}-1\right] - \frac{1}{3}t$

and $\frac{2s+7}{(s+2)(2s+1)} = \frac{4}{2s+1} - \frac{1}{s+2} = \frac{2}{s+\frac{1}{2}} - \frac{1}{s+2}$

$\Rightarrow L^{-1}\left\{\frac{2s+7}{(s+2)(2s+1)}\right\} = 2e^{-t/2} - e^{-2t}.$

Using these inverse Laplace transforms we get

$$y(t) = (2e^{-t/2} - e6{-}2t) + e^{-2t}\left[\frac{2}{9}(e^{3t/2}-1) - \frac{t}{3}\right] = \frac{20}{9}e^{-t/2} - e^{-2t}\left(\frac{11}{9} + \frac{t}{3}\right)$$

Exercise 13.39. (i) Solve $y'' + 25y = 10\ \cos(5t),\ \ y(0) = 2,\ y'(0) = 0,$ **Answer:** $y(t) = t\sin 5t + 2\cos(5t)$.
(ii) Solve $(D + D^2)x = 2$ when $x(0) = 3,\ x'(0) = 1$ **Answer:** $y(t) = e^{-t} + 2t + 2$.
(iii) Solve $(D^2{+}D)y = t^2{+}2t$ when $y(0) = 4,\ y'(0) = -2$ **Answer:** $2{+}2e^{-t}{+}\frac{t^3}{3}$.

Case II: When a, b are functions of t, i.e., of the form

$$t^2\frac{d^2x}{dt^2} + t\frac{dx}{dt} + x = F(t) \tag{13.7.3}$$

In this case, we use the theorem

$$L\left\{t^m\frac{d^nx}{dt^n}\right\} = L\{t^m x^{(n)}(t)\} = (-1)^m\frac{d^m}{ds^m}L\{x^{(n)}\} \tag{13.7.4}$$

Taking the Laplace transform of Eq. (13.7.3),

$$(-1)^2\frac{d^2}{ds^2}\Big[s^2\overline{x} - s\,x(0) - x'(0)\Big] - \frac{d}{ds}\Big\{s\overline{x} - x(0)\Big\} + \overline{x} = f(s)$$

The required solution is obtained by taking inverse Laplace transform of $\overline{x}(s)$.

Example 13.40. Using Laplace transform solve the following differential equation.

$$y'' + ty' - y = 0 \quad \text{if} \quad y(0) = 0,\ y'(0) = 1$$

Solution: Taking Laplace transform of $y'' + ty' - y = 0$, we get

$$p^2\overline{y} - py(0) - y'(0) + (-1)^1 \frac{d}{dp}\Big[p\overline{y} - y(0)\Big] - \overline{y} = 0$$

Putting $y(0) = 0, y'(0) = 1$, we get

$$\begin{aligned} & p^2\overline{y} - 1 - \frac{d}{dp}\Big[p\,\overline{y}\Big] - \overline{y} = 0 \\ \Rightarrow\ & (p^2-1)\overline{y} - \Big(\overline{y} + p\frac{d\overline{y}}{dp}\Big) = 1 \\ \Rightarrow\ & (p^2-2)\overline{y} - p\frac{d\overline{y}}{dp} = 1 \\ \Rightarrow\ & \frac{d\overline{y}}{dp} + \left(\frac{2}{p} - p\right)\overline{y} = -\frac{1}{p} \end{aligned}$$

Therefore the integrating factor

$$I.F = e^{\int\left(\frac{2}{p}-p\right)dp} = e^{2\log p - p^2/2} = e^{\log p^2}\cdot e^{-p^2/2} = p^2e^{-p^2/2}$$

Therefore

$$\overline{y}p^2e^{-p^2/2} = c + \int (p^2e^{-p^2/2})\left(-\frac{1}{p}dp\right) = c - \int pe^{p^2/2}dp.$$

Put $p^2/2 = z$, then $p\,dp = dz$. Therefore,

$$\begin{aligned} & \overline{y}p^2e^{-p^2/2} = c - \int e^{-z}dz = c + e^{-z} = c + e^{-p^2/2} \\ \Rightarrow\ & \overline{y} = \frac{c}{p^2}e^{p^2/2} + \frac{1}{p^2} \end{aligned}$$

Taking inverse transform,

$$y(t) = t + cL^{-1}\left\{\frac{1}{p^2}e^{p^2/2}\right\}$$

Subjecting this to the condition $y = 0$ when $t = 0$, we get $c = 0$. Therefore, the solution is

$$y(t) = t$$

Exercise 13.41. (i) Solve $ty'' + (1-2t)y' - 2y = 0,\ \ y(0) = 1,\ y'(0) = -2$ **Answer:** $y(t) = e^{2t}$.

Unit 14

Course Structure

Fourier Transform : Definition and basic properties. Fourier transform of some elementary functions, of derivatives. Inverse Fourier transform. Convolution theorem and Parseval's relation. Applications of Fourier inversion and convolution theorems. Fourier sine and cosine transforms.

---

14 Introduction

The Fourier transform is a generalization of the Fourier series representation of functions. The Fourier series is limited to periodic functions, while the Fourier transform can be used for a larger class of functions which are not necessarily periodic. Since the transform is essential to the understanding of several exercises, we briefly explain some basic Fourier transform concepts in this unit.

Objective

After reading this unit readers will able to know fundamental mathematical properties of the Fourier transform including linearity, shift, symmetry, scaling, modulation and convolution. Further, the reader will be able to calculate the Fourier transform or inverse transform of common functions.

14.1 The Infinite Fourier Transform

Definition 14.1. Infinite Fourier sine transform: The Fourier sine transform of $F(x)$ on $0 < x < \infty$ is denoted by $f_s(s)$ or $F_s\{F(x)\}$ and is defined as

$$f_s(s) = F_s\{F(x)\} = \int_0^\infty F(x) \sin sx \; dx.$$

The inverse formula for infinite Fourier sine transform is given by

$$F(x) = F_s^{-1}\{f_s(s)\} = \frac{2}{\pi} \int_0^\infty f_s(s) \; \sin sx \; ds.$$

Definition 14.2. Infinite Fourier cosine transform: The Fourier sine transform of $F(x)$ on $0 < x < \infty$ is denoted by $f_c(s)$ or $F_c\{F(x)\}$ and is defined as

$$f_c(s) = F_c\{F(x)\} = \int_0^\infty F(x) \cos sx \; dx.$$

The inverse formula for infinite Fourier cosine transform is given by

$$F(x) = F_c^{-1}\{f_c(s)\} = \frac{2}{\pi} \int_0^\infty f_c(s) \; \cos sx \; ds.$$

Definition 14.3. Infinite Fourier transform: The infinite Fourier transform of $F(x)$ on $0 < x < \infty$ is denoted by $f(s)$ or $F\{F(x)\}$ and is defined as

$$f(s) = F\{F(x)\} = \int_{-\infty}^{\infty} F(x)e^{-isx}\, dx.$$

The inverse formula for infinite Fourier sine transform is given by

$$F(x) = F^{-1}\{f(s)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(s)\, e^{isx}\, ds.$$

14.2 Relationship between Fourier transform and Laplace transform

We define a function as follows:

$$F(t) = \begin{cases} e^{-xt}\phi(t), & , \quad t > 0 \\ 0 & , \quad t < 0 \end{cases}$$

Taking infinite Fourier transform of this function we obtain

$$\begin{aligned} F\{F(t)\} &= \int_{-\infty}^{\infty} e^{-iyt}F(t)\, dt = \int_{-\infty}^{0} e^{-iyt}F(t)\, dt + \int_{0}^{\infty} e^{-iyt}F(t)\, dt \\ &= \int_{-\infty}^{0} e^{-iyt}\cdot 0\, dt + \int_{0}^{\infty} e^{-iyt}e^{-xt}\phi(t)\, dt = \int_{0}^{\infty} e^{-(x+iy)t}\phi(t)\, dt \\ &= e^{-st}\phi(t)\, dt = L\{\phi(t)\}, \quad \text{where } s = x + iy \end{aligned}$$

Therefore $F\{F(t)\} = L\{\phi(t)\}$. This is the required relation between Fourier transform and Laplace transform.

14.3 Some theorems

Theorem 14.4. Linear Property: If c_1 and c_2 are arbitrary constants, then

$$F\{c_1F(x) + c_2G(x)\} = c_1F\{F(x)\} + c_2F\{G(x)\}$$

Theorem 14.5. Change of Scale Property: If $f(s)$ is the Fourier transform of $F(x)$, then $\frac{1}{a}f\left(\frac{s}{a}\right)$ is the Fourier transform of $F(ax)$.

Theorem 14.6. Shifting Property: If $f(s)$ is the Fourier transform of $F(x)$, then $e^{-ias}f(s)$ is the Fourier transform of $F(x-a)$.

Theorem 14.7. Modulation Theorem: If $F(x)$ has the Fourier transform $f(s)$, then $F(x)\cos ax$ has the Fourier transform

$$\frac{1}{2}f(s-a) + \frac{1}{2}f(s+a).$$

Theorem 14.8. Derivative Theorem: The Fourier transform of $F'(x)$, the derivative of $F(x)$, is $is\ f(s)$, where $f(s)$ is the Fourier transform of $F(x)$. Moreover,

$$F\left\{\frac{d^nF}{dx^n}\right\} = (is)^n f(s), \text{ where } F\{F(x)\} = f(s)$$

if the first $(n-1)$ derivative of $F(x)$ vanish identically as $x \to \pm\infty$.

• Proofs are left as exercise.

Theorem 14.9. Convolution Theorem: The convolution for the Fourier transform is defined as

$$F * G = \int_{-\infty}^{\infty} F(u)\, G(x-u)\, du.$$

If $F\{f(x)\}$ and $F\{g(x)\}$ are the Fourier transforms of functions $f(x)$ and $g(x)$ respectively, then Fourier transform of the convolution of $f(x)$ and $g(x)$ is the product of the their Fourier transforms, i.e.,

$$F\{f(x) * g(x)\} = F\{f(x)\} \cdot F\{g(x)\}$$

Proof.

We have
$$\begin{aligned}
F\{f(x) * g(x)\} &= \int_{-\infty}^{\infty} \{f(x) * g(x)\} e^{-isx} dx \\
&= \int_{x=-\infty}^{\infty} \left[\int_{u=-\infty}^{\infty} f(u)\, g(x-u)\, du \right] e^{-isx} dx \\
&= \int_{u=-\infty}^{\infty} f(u) \left[\int_{x=-\infty}^{\infty} g(x-u) e^{-isx} dx \right] du \quad \text{(Changing order of integration)} \\
&= \int_{u=-\infty}^{\infty} f(u) e^{-isu} F\{g(x)\} du \quad \text{(Using Shifting Property)} \\
&= F\{g(x)\} \int_{u=-\infty}^{\infty} f(u) e^{-isu} du \\
&= F\{g(x)\} \cdot F\{f(x)\} = F\{f(x)\} \cdot F\{g(x)\}
\end{aligned}$$

□

Example 14.10. Find the Fourier transform of $f(x) = \begin{cases} x & |x| \le a \\ 0 & |x| > a. \end{cases}$

Solution: Given that $f(t) = \begin{cases} t & -a \le t \le a \\ 0 & |t| > a. \end{cases}$ Now

$$\begin{aligned}
F\{f(t)\} &= \int_{-\infty}^{\infty} e^{-ist} f(t)\, dt \\
&= \int_{-\infty}^{-a} e^{-ist} f(t)\, dt + \int_{-a}^{a} e^{-ist} f(t)\, dt + \int_{a}^{\infty} e^{-ist} f(t)\, dt \\
&= \int_{\infty}^{a} e^{isy} f(-y)\, (-dy) + \int_{-a}^{a} e^{-ist}\, t\, dt + \int_{a}^{\infty} e^{-ist} \cdot 0 \cdot dt \\
&= \int_{a}^{\infty} e^{isy} \cdot 0 \cdot dy + \int_{-a}^{a} e^{-ist}\, t\, dt \\
&= \left(-\frac{a}{is}\right) \{e^{-isa} + e^{isa}\} + \frac{1}{s^2} \{e^{-isa} - e^{isa}\} \\
&= \left(\frac{2ai}{s}\right) \cos(sa) - \frac{2}{s^2} \sin(sa)
\end{aligned}$$

Example 14.11. Find the Fourier transform of $f(x) = \begin{cases} 1 & |x| < a \\ 0 & |x| > a. \end{cases}$ and hence evaluate

$$(i) \int_{-\infty}^{\infty} \frac{\sin sa \cdot \cos sx}{s} dx, \qquad (ii) \int_{0}^{\infty} \frac{\sin s}{s} ds.$$

Solution: For the first part, proceed as the above example and find the answer is $\frac{2}{s}\sin sa$. For the second part, let $F\{f(x)\} = \overline{f}(s)$. We know that if

$$\overline{f}(s) = F\{f(x)\} = \int_{-\infty}^{\infty} f(x)e^{-isx}dx \quad \text{then} \quad f(x) = F^{-1}\{\overline{f}(s)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{f}(s)e^{isx}ds.$$

$$\therefore \int_{-\infty}^{\infty}\overline{f}(s)e^{isx}ds = 2\pi f(x) = \begin{cases} 2\pi & \text{if} |x| < a \\ 0 & \text{if} |x| > a. \end{cases}$$

But $\overline{f}(s) = \frac{2\sin sa}{s}$, by first part.

$$\therefore \int_{-\infty}^{\infty}\frac{2\sin sa}{s}(\cos sx + i\sin sx)\,ds = \begin{cases} 2\pi & \text{if} \quad |x| < a \\ 0 & \text{if} \quad |x| > a. \end{cases}$$

$$\Rightarrow \int_{-\infty}^{\infty}\frac{\sin sa\cos sx}{s}\,ds + i\int_{-\infty}^{\infty}\frac{\sin sa\sin sx}{s}\,ds = \begin{cases} \pi & \text{if} \quad |x| < a \\ 0 & \text{if} \quad |x| > a. \end{cases}$$

Equating real parts on the both sides, we obtain

$$\int_{-\infty}^{\infty}\frac{\sin sa\sin sx}{s}\,ds = \begin{cases} \pi & \text{if} \quad |x| < a \\ 0 & \text{if} \quad |x| > a. \end{cases}$$

Now if $x = 0$ and $a = 1$, then the second part gives

$$\int_{-\infty}^{\infty}\frac{\sin s}{s}\,ds = \pi \Rightarrow 2\int_{0}^{\infty}\frac{\sin s}{s}\,ds = \pi \Rightarrow \int_{0}^{\infty}\frac{\sin s}{s}\,ds = \frac{\pi}{2}.$$

Exercise 14.12. (i) Find the Fourier transform of $F(x) = \begin{cases} (1-x^2), & |x| < 1 \\ 0 & |x| > 1. \end{cases}$ and hence evaluate $\int_0^\infty \left(\frac{x\cos x - \sin x}{x^3}\right)\cos\left(\frac{x}{2}\right)dx$. **Answer:** $f(s) = \frac{4}{s^3}(\sin s - s\cos s)$; $-\frac{3\pi}{16}$.

(ii) Show that the Fourier transform of $f(x) = e^{-x^2/2}$ is $\sqrt{2\pi}e^{-s^2/2}$

(iii) Find the complex Fourier transform of $f(x) = e^{-a|x|}$ **Answer:** $f(s) = \frac{2a}{s^2+a^2}$.

(iv) Find the inverse Fourier transform of $\overline{f}(s) = e^{-|s|y}$ **Answer:** $F^{-1}\{\overline{f}(s)\} = \frac{y}{\pi(y^2+x^2)}$.

14.4 Problems related to Integral Equations

Example 14.13. Solve the integral equation $\int_0^\infty f(x)\cos sx\,dx = e^{-\lambda}$.

Solution: We have $\int_0^\infty f(x)\cos sx\,dx = e^{-s}$. By definition,

$$F_c\{f(x)\} = \int_0^\infty f(x)\cos sx\,dx = \overline{f}_c(s) \quad \text{and} \quad F_c^{-1}\{\overline{f}_c(s)\} = f(x) = \frac{2}{\pi}\int_0^\infty \overline{f}_c(s)\cos sx\,ds$$

Comparing this with the given equation, we have $\overline{f}_c(s) = e^{-s}$. Using this, we obtain

$$f(x) = \frac{2}{\pi}\int_0^\infty e^{-s}\cos xs\,ds = \frac{2}{\pi}\frac{1}{1+x^2}.$$

Example 14.14. Solve the integral equation $\int_0^\infty F(x)\sin(xt)\,dx = F(x) = \begin{cases} 1 & , \quad 0 \le t < 1 \\ 2 & , \quad 1 \le t < 2 \\ 0 & , \quad t \ge 2 \end{cases}$

Solution: By the definition, we know

$$F_s\{F(x)\} = \int_0^\infty F(x)\sin(sx)\,dx = f_c(s). \tag{14.4.1}$$

Then $f_s(s) = \begin{cases} 1 & , \quad 0 \le t < 1 \\ 2 & , \quad 1 \le t < 2 \\ 0 & , \quad t \ge 2 \end{cases}$

The sine inversion formula relative to (14.5.1) is

$$F_s^{-1}\{f_s(s)\} = F(x) = \frac{2}{\pi}\int_0^\infty f_s(s)\sin sx\,ds$$

From which we have

$$\begin{aligned}
\frac{\pi}{2}F(x) &= \int_0^\infty f_s(s)\ \sin sx\,ds \\
\Rightarrow \frac{\pi}{2}F(x) &= \int_0^1 f_s(s)\sin sx\,dx + \int_1^2 f_s(s)\ \sin sx\,ds + \int_2^\infty f_s(s)\ \sin sx\,ds \\
\Rightarrow \frac{\pi}{2}F(x) &= \int_0^1 1\cdot\sin sx\,dx + \int_1^2 2\cdot\ \sin sx\,ds + \int_2^\infty 0\cdot\ \sin sx\,ds \\
\Rightarrow \frac{\pi}{2}F(x) &= \frac{1}{x}\Big[(1-\cos x) + 2(\cos x - \cos 2x)\Big] = \frac{1+\cos x - 2\cos 2x}{x} \\
\Rightarrow F(x) &= \frac{2}{\pi x}(1+\cos x - 2\cos 2x)
\end{aligned}$$

Exercise 14.15. (i) Show that $\int_0^\infty \frac{\cos\lambda x}{\lambda^2+1}d\lambda = \frac{\pi}{2}e^{-x}$

(ii) Solve the integral equation $\int_0^\infty F(x)\cos\lambda x\,dx = \begin{cases} 1-\lambda, & 0 \le \lambda \le 1 \\ 0, & \lambda > 1. \end{cases}$

Answer: $F(x) = \frac{2}{\pi x^2}(1-\cos x)$

14.5 The finite Fourier Transform

Definition 14.16. The finite Fourier sine transform of $F(x)$: The finite Fourier sine transform of $F(x)$ on $0 < x < l$, is defined by

$$F_s\{F(x)\} = f_s(s) = \int_0^l F(x)\sin\frac{s\pi x}{l}dx$$

where s is a positive integer. The function $F(x)$ is then called the *inverse finite Fourier sine transform* of $f_s(s)$ and is given by

$$F_s^{-1}\{f_s(s)\} = F(x) = \frac{2}{l}\sum_{s=1}^\infty f_s(s)\ \sin\frac{s\pi x}{l}$$

This formula is obtained from Fourier sine series $f(x) = \sum_{n=1}^\infty b_n \sin\frac{n\pi x}{l}$.

Definition 14.17. The finite Fourier cosine transform of $F(x)$: The finite Fourier cosine transform of $F(x)$ on $0 < x < l$, is defined by

$$F_c\{F(x)\} = f_c(s) = \int_0^l F(x)\cos\frac{s\pi x}{l}dx$$

where s is a positive integer or zero. The function $F(x)$ is then called the *inverse finite Fourier cosine transform* of $f_c(s)$ and is given by

$$F_c^{-1}\{f_c(s)\} = F(x) = \frac{1}{l}f_c(0) + \frac{2}{l}\sum_{s=1}^{\infty} f_c(s)\ \cos\frac{s\pi x}{l}$$

This formula is obtained from Fourier cosine series $F(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} b_n \cos\frac{n\pi x}{l}$.

Theorem 14.18. Fourier Integral Theorem: It sates that if $f(x)$ satisfies the following conditions:

- $f(x)$ satisfies the Dirichlet conditions in every interval $-l \le x \le l$.
- $\int_{-\infty}^{\infty} |f(x)|\, dx$ converges, i.e., $f(x)$ is absolutely integrable in the interval $-\infty < x < \infty$, then

$$f(x) = \frac{1}{2\pi}\int_{s=-\infty}^{\infty}\int_{t=-\infty}^{\infty} f(t)\cos s(x-t)\ ds\ dt.$$

The integral on R.H.S is called *Fourier integral* or *Fourier integral expansion of* $f(x)$.

Theorem 14.19. Different forms of Fourier integral formula:

$$(i) f(x) = \frac{1}{\pi}\int_{s=0}^{\infty}\int_{t=-\infty}^{\infty} f(t)\cos s(x-t)\ ds\ dt.$$

$$(ii) f(x) = \frac{2}{\pi}\int_{0}^{\infty}\int_{0}^{\infty} f(t)\ \cos st\cos sx\ ds\ dt \qquad \text{(Cosine Form)}$$

$$(iii) f(x) = \frac{2}{\pi}\int_{0}^{\infty}\int_{0}^{\infty} f(t)\ \sin st\sin sx\ ds\ dt \qquad \text{(Sine Form)}$$

$$(iv) f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)\ e^{-ist}e^{isx}\ ds\ dt \qquad \text{(Exponential Form)}$$

Theorem 14.20. Parseval's identity for Fourier series: Suppose the Fourier series corresponding to $f(x)$ converges uniformly to $f(x)$ in the interval $-l < x < l$, then

$$\frac{1}{l}\int_{-l}^{l}\Big[f(x)\Big]^2 dx = \frac{a_0^2}{2} + \sum_{n=1}^{\infty}(a_n^2 + b_n^2),$$

where the integral on L.H.S is supposed to exist.

Proof. Let the Fourier series of $f(x)$ converges uniformly to $f(x)$ at every point of the interval $-l < x < l$, so that

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{l} + b_n\sin\frac{n\pi x}{l}\right) \tag{14.5.1}$$

and that term by term integration of this series is possible. Here

$$a_n = \frac{1}{l}\int_{-l}^{l} f(x)\cos\frac{n\pi x}{l}\ dx\ \ (n = 0, 1, 2, 3, \ldots) \quad \text{and} \quad b_n = \frac{1}{l}\int_{-l}^{l} f(x)\sin\frac{n\pi x}{l}\ dx\ \ (n = 1, 2, 3, \ldots)$$

Multiplying (14.5.1) by $f(x)$ and integrating term by term from $-l$ to l, we get

$$\begin{aligned}
\int_{-l}^{l}\Big[f(x)\Big]^2 dx &= \frac{a_0}{2}\int_{-l}^{l} f(x)\,dx + \sum_{n=1}^{\infty} a_n \int_{-l}^{l} f(x)\cos\frac{n\pi x}{l}dx + \sum_{n=1}^{\infty} b_n \int_{-l}^{l} f(x)\sin\frac{n\pi x}{l}dx \\
\Rightarrow \int_{-l}^{l}\Big[f(x)\Big]^2 dx &= \frac{a_0}{2}\cdot la_0 + \sum_{n=1}^{\infty} l(a_n^2+b_n^2) \\
\Rightarrow \frac{1}{l}\int_{-l}^{l}\Big[f(x)\Big]^2 dx &= \frac{a_0^2}{2} + \sum_{n=1}^{\infty}(a_n^2+b_n^2)
\end{aligned}$$

□

Theorem 14.21. Parseval's identity for Fourier transform. Rayleigh's Theorem: If $f(p)$ and $g(p)$ are complex Fourier transforms of $F(x)$ and $G(x)$ respectively, then

$$(i)\ \frac{1}{2\pi}\int_{-\infty}^{\infty} f(p)\,\overline{g(p)}\,dp = \int_{-\infty}^{\infty} F(x)\,\overline{G(x)}\,dx \quad \text{and} \quad (ii)\ \frac{1}{2\pi}\int_{-\infty}^{\infty}|f(p)|^2\,dp = \int_{-\infty}^{\infty}|F(x)|^2\,dx$$

where bar represents the complex conjugate.

Proof. Using the inversion formula for Fourier transform, we get

$$G(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(p)\,e^{ipx}\,dp \tag{14.5.2}$$

Taking conjugate complex of the both sides in (14.5.2), we obtain

$$\overline{G(x)} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{g(p)}e^{-ipx}\,dp \tag{14.5.3}$$

Therefore,

$$\begin{aligned}
\int_{-\infty}^{\infty} F(x)\,\overline{G(x)}\,dx &= \int_{-\infty}^{\infty} F(x)\,dx\cdot\left\{\frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{g(p)}e^{-ipx}\,dp\right\}, \qquad \text{(Using (14.5.3))} \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x)\,\overline{g(p)}e^{-ipx}\,dx\,dp \qquad \text{(On changing order of integration)} \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(x)\,\overline{g(p)}e^{-ipx}\,dp\,dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{g(p)}\,dp\left[\int_{-\infty}^{\infty} F(x)\,e^{-ipx}\,dx\right] \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\overline{g(p)}\,dp\,\{f(p)\} \qquad \text{(By the def. of Fourier transform)} \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(p)\,\overline{g(p)}\,dp
\end{aligned}$$

Thus, we have proved that

$$\int_{-\infty}^{\infty} F(x)\,\overline{G(x)}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(p)\,\overline{g(p)}\,dp. \tag{14.5.4}$$

This proves the first part. Now putting $G(x) = F(x)$ in (14.5.4), we get

$$\int_{-\infty}^{\infty} F(x)\,\overline{F(x)}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(p)\,\overline{f(p)}\,dp \quad \Rightarrow \quad \int_{-\infty}^{\infty}|F(x)|^2\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}|f(p)|^2\,dp.$$

□

Example 14.22. Use Parseval's identity to prove that

$$(i)\ \int_0^\infty \frac{dt}{(a^2+t^2)(b^2+t^2)} = \frac{\pi}{2ab(a+b)}, \quad (ii)\ \int_0^\infty \frac{\sin(at)\,dt}{t(a^2+t^2)} = \frac{\pi}{2}\left(\frac{1-e^{-a^2}}{a^2}\right).$$

Solution: (i) Let $F(x) = e^{-ax}, G(x) = e^{-bx}$. Now,

$$f_c(p) = \int_0^\infty F(x)\ \cos(px)\,dx = \int_0^\infty e^{-ax}\cos(px)\,dx = \frac{a}{a^2+p^2}$$

Similarly, we can find $g_c(p) = \frac{b}{b^2+p^2}$. By Parseval's identity for Fourier transform, we get

$$\frac{2}{\pi}\int_0^\infty f_c(p)\ g_c(p)\,dp = \int_0^\infty F(x)\ G(x)\,dx.$$

Putting values, we get

$$\begin{aligned}
&\frac{2}{\pi}\int_0^\infty \frac{a}{(a^2+p^2)}\cdot\frac{b}{(b^2+p^2)}dp = \int_0^\infty e^{-ax}\cdot e^{-bx}\,dx \\
\Rightarrow &\int_0^\infty \frac{dp}{(b^2+p^2)(a^2+p^2)} = \frac{\pi}{2ab}\left[\frac{e^{-(a+b)x}}{-(a+b)}\right]_{x=0}^\infty = \frac{\pi}{2ab(a+b)}\{1-0\} \\
\Rightarrow &\int_0^\infty \frac{dt}{(b^2+t^2)(a^2+t^2)} = \frac{\pi}{2ab(a+b)}
\end{aligned}$$

(ii) Let $F(x) = e^{-ax}$, then $f_c(p) = \frac{a}{a^2+p^2}$. Also let $G(x) = \begin{cases} 1 & , \quad 0 < x < a \\ 0 & , \quad x > a. \end{cases}$

Then

$$\begin{aligned}
g_c(p) &= \int_0^\infty G(x)\ \cos(px)\,dx \\
&= \int_0^a G(x)\ \cos(px)\,dx + \int_a^\infty G(x)\ \cos(px)\,dx \\
&= \int_0^a \cos(px)\,dx + \int_a^\infty 0\cdot\ \cos(px)\,dx \\
&= \left[\frac{\sin(px)}{p}\right]_{x=0}^a = \frac{\sin(pa)}{p}
\end{aligned}$$

$$\begin{aligned}
&\text{Since}\quad \frac{2}{\pi}\int_0^\infty f_c(p)\ g_c(p)\,dp = \int_0^\infty F(x)\ G(x)\,dx \\
\Rightarrow &\frac{2}{\pi}\int_0^\infty \frac{a}{a^2+p^2}\frac{\sin(pa)}{p}\,dp = \int_0^\infty e^{-ax}\ G(x)\,dx \\
\Rightarrow &\frac{2a}{\pi}\int_0^\infty \frac{\sin(pa)\,dp}{p(a^2+p^2)}\,dp = \int_0^a e^{-ax}\ G(x)\,dx + \int_a^\infty e^{-ax}\ G(x)\,dx \\
\Rightarrow &\frac{2a}{\pi}\int_0^\infty \frac{\sin(pa)\,dp}{p(a^2+p^2)}\,dp = \int_0^a e^{-ax}\ \cdot 1\,dx + \int_a^\infty e^{-ax}\ \cdot 0\,dx \\
\Rightarrow &\frac{2a}{\pi}\int_0^\infty \frac{\sin(pa)\,dp}{p(a^2+p^2)}\,dp = \frac{1}{a}(1-e^{-a^2}) \\
\Rightarrow &\int_0^\infty \frac{\sin(pa)\,dp}{p(a^2+p^2)}\,dp = \frac{\pi}{2a^2}(1-e^{-a^2}).
\end{aligned}$$

Example 14.23. Find the Fourier transform of $f(x)$ defined by

$$f(x) = \begin{cases} 1 & , \quad |x| < a \\ 0 & , \quad |x| > a. \end{cases}$$

and hence prove that

$$\int_0^\infty \frac{\sin^2(ax)}{x^2} dx = \frac{\pi a}{2}.$$

Solution: First Part:

$$\begin{aligned} F\{f(x)\} &= \int_{-\infty}^{\infty} e^{-ipx} f(x)\, dx \\ &= \int_{-\infty}^{-a} e^{-ipx}\, f(x)\, dx + \int_{-a}^{a} e^{-ipx}\, f(x)\, dx + \int_{a}^{\infty} e^{-ipx}\, f(x)\, dx \\ &= \int_{\infty}^{a} e^{ipy}\, f(-y)\, (-dy) + \int_{-a}^{a} e^{-ipx}\, dx + \int_{a}^{\infty} e^{-ipx}\, \cdot 0\, dx \\ &= \int_{a}^{\infty} e^{ipy} \cdot 0 \cdot dy + \frac{1}{-ip}\left(e^{-ipx}\right)_{-a}^{a} + 0 \\ &= \frac{e^{ipa} - e^{-ipa}}{ip} = \frac{2}{p}\sin pa = \overline{f}(p) \end{aligned}$$

Second Part: Using Parseval's identity for Fourier integral, we get

$$\begin{aligned} \textstyle\int_{-\infty}^{\infty} |f(x)|^2 dx &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\overline{f}(p)|^2 dp \\ \Rightarrow \textstyle\int_{-a}^{a} 1^2 dx &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{4}{p^2}\sin^2 pa\, dp \\ \Rightarrow 2a &= \frac{2}{2\pi}\int_{0}^{\infty} \frac{4}{p^2}\sin^2 pa\, dp \\ \Rightarrow \textstyle\int_0^\infty \frac{\sin^2(ax)}{x^2} dx &= \frac{\pi a}{2}. \end{aligned}$$

14.6 Problems related to finite Fourier Sine and Cosine transform:

Example 14.24. Find finite Fourier $sine$ and $cosine$ transform of

$$f(x) = x^2, \quad 0 < x < 4.$$

Solution: (i)

$$\begin{aligned} F_s\{f(x)\} &= \int_0^c f(x)\, \sin\frac{n\pi x}{c} dx = \int_0^4 x^2 \sin\frac{n\pi x}{4} dx \\ &= \left[-\frac{4}{n\pi}x^2\cos\frac{n\pi x}{4}\right]_{x=0}^{4} + \int_0^4 2x\frac{4}{n\pi}\cos\frac{n\pi x}{4} dx \\ &= -\frac{4^3}{n\pi}\cos n\pi + \frac{8}{n\pi}\left[\frac{4x}{n\pi}\sin\frac{n\pi x}{4} + \frac{4^2}{n^2\pi^2}\cos\frac{n\pi x}{4}\right]_{x=0}^{4} \\ &= -\frac{4^3}{n\pi}\cos n\pi + \frac{8\cdot 4^2}{n\pi\cdot n^2\pi^2}(\cos n\pi - 1) \\ \therefore \overline{f}_s(n) &= -\frac{64}{n\pi}\cos n\pi + \frac{128}{n^3} \end{aligned}$$

(ii)

$$\begin{aligned} F_c\{f(x)\} &= \overline{f}_c(n) = \int_0^4 f(x)\cos\frac{n\pi x}{4} = \int_0^4 x^2\cos\frac{n\pi x}{4}dx \\ &= \left[\frac{4x^2}{n\pi}\sin\frac{n\pi x}{4}\right]_{x=0}^4 - \int_0^4 \frac{4}{n\pi}2x\sin\frac{n\pi x}{4}dx \\ &= 0 - \frac{8}{n\pi}\left[-\frac{4x}{n\pi}\cos\frac{npx}{4} + \frac{4^2}{n^2\pi^2}\sin\frac{n\pi x}{4}\right]_{x=0}^4 \\ &= \frac{128}{n^2\pi^2}\cos n\pi \end{aligned}$$

Example 14.25. Find $f(x)$ if its finite $sine$ transform is given by

$$\overline{f}_s(s) = \frac{1-\cos s\pi}{s^2\pi^2}, \text{ where } 0 < x < \pi,\ s = 1,2,3,\ldots$$

Solution: We know that

$$f(x) = \frac{2}{l}\sum_{n=1}^{\infty}\overline{f}_s(n)\sin\frac{n\pi x}{l}.$$

In our case this becomes

$$\begin{aligned} f(x) &= \frac{2}{\pi}\sum_{s=1}^{\infty}\overline{f}_s(s)\sin\left(\frac{s\pi x}{\pi}\right) = \frac{2}{\pi}\sum_{s=1}^{\infty}\frac{1-\cos s\pi}{s^2\pi^2}\sin sx \\ \Rightarrow f(x) &= \frac{2}{\pi^3}\sum_{s=1}^{\infty}\left(\frac{1-\cos\pi s}{s^2}\right)\sin xs. \end{aligned}$$

Exercise 14.26. (i) Find the finite cosine transform of $\left(1-\frac{x}{\pi}\right)^2$. **Answer:** $f_c(s) = \begin{cases} \frac{\pi}{3}, & s = 0 \\ \frac{2}{\pi s^2}, & s = 1,2,3,\ldots \end{cases}$

(ii) Show that the finite sine transform of $\frac{x}{\pi}$ is $(-1)^{s+1}\frac{1}{s}$

(iii) When $f(x) = \sin mx$, where, m is a positive integer, show that $f_s(p) = \begin{cases} 0, & p \neq m \\ \frac{\pi}{2}, & p = m \end{cases}$

(iv) If $f_s(n) = 2\pi\frac{(-1)^{n-1}}{n^2}, n = 1,2,3,\ldots$ where $0 < x < \pi$, then find $f(x)$. **Answer:** $\frac{2}{s}\sin\frac{s\pi}{2},\ s > 0$

(v) Find the finite cosine transform of $f(x)$ if $f(x) = \begin{cases} 1, & 0 < x < \frac{\pi}{2} \\ -1, & \frac{\pi}{2} < x < \pi \end{cases}$

(vi) Show that $f_c\left\{\frac{x^2}{2\pi} - \frac{\pi}{6}\right\} = \begin{cases} 0, & n = 0 \\ (-1)^n/n^2, & n = 1,2,3,\ldots \end{cases}$

Unit 15

Course Structure

Hankel Transform : Definition and inversion formula. Hankel transform of derivatives. Finite Hankel transform.

15 Introduction

Hankel transforms are integral transformations whose kernels are Bessel functions. They are sometimes referred to as Bessel transforms. When we are dealing with problems that show circular symmetry, Hankel transforms may be very useful. Laplace's partial differential equation in cylindrical coordinates can be transformed into an ordinary differential equation by using the Hankel transform. Because the Hankel transform is the two-dimensional Fourier transform of a circularly symmetric function, it plays an important role in optical data processing.

15.1 Definition: Infinite Hankel Transform

The *infinite Hankel transform* of a function $f(x),\ 0 < x < \infty$, is defined as

$$H\{f(x)\} = \overline{f}(s) = \int_0^\infty f_n(x) \cdot xJ_n(sx)\, dx \tag{15.1.1}$$

where $J_n(sx)$ is the Bessel function of the first kind of order n. Also here $\overline{f}(s)$ is defined as *Hankel transform of order n* of the function $f(x)$.

Remark: In the integral (15.1.1), $xJ_n(sx)$ is called the Kernel of the transformation.

15.2 Definition: Inverse Hankel Transform

If $\overline{f}(s)$ is the infinite Hankel transform of order n of the function $f(x)$, then we write

$$H\{f(x)\} = \overline{f}(s) = \int_0^\infty f(x) \cdot xJ_n(sx)\, dx. \tag{15.2.1}$$

Here $f(x)$ is called the *inverse transform* of the function $\overline{f}(s)$ and we write $f(x) = H^{-1}\{\overline{f}(s)\}$. The inverse formula for inverse Hankel Transform is

$$f(x) = H^{-1}\{\overline{f}(s)\} = \int_0^\infty \overline{f}(s) \cdot sJ_n(sx)\, ds. \tag{15.2.2}$$

15.3 Some Important Results on Bessel functions

I. Bessel function of first kind:$J_n(x) = \sum_{r=0}^{\infty} \frac{(-1)^r}{r!\Gamma(n+r+1)} \left(\frac{x}{2}\right)^{n+2r}$

II. Recurrence formula for $J_n(x)$:

$(i)\ xJ_n'(x) = nJ_n(x) - xJ_{n+1}(x)$ $\qquad$ $(ii)\ xJ_n'(x) = -nJ_n(x) + xJ_{n-1}(x)$

$(iii)\ 2J_n(x) = J_{n-1}(x) - J_{n+1}(x)$ $\qquad$ $(iv)\ 2nJ_n'(x) = x[J_{n-1}(x) + J_{n+1}(x)]$

$(v)\ \dfrac{d}{dx}\left[x^{-n}J_n(x)\right] = -x^{-n}J_{n+1}(x)$ $\qquad$ $(vi)\ \dfrac{d}{dx}\left[x^{n}J_n(x)\right] = x^{n}J_{n-1}(x)$

III. Infinite Integrals Involving Bessel Functions

$(i)\ \displaystyle\int_0^\infty e^{-ax}J_0(sx)\,dx = (a^2+s^2)^{-1/2}$ $\qquad$ $(ii)\ \displaystyle\int_0^\infty e^{-ax}J_1(sx)\,dx = \frac{1}{s} - \frac{a}{s(a^2+s^2)^{1/2}}$

$(iii)\ \displaystyle\int_0^\infty xe^{-ax}J_0(sx)\,dx = a(a^2+s^2)^{-3/2}$ $\qquad$ $(iv)\ \displaystyle\int_0^\infty e^{-ax}J_1(sx)\,dx = s(a^2+s^2)^{-3/2}$

$(v)\ \displaystyle\int_0^\infty \frac{e^{-ax}}{x}J_1(sx)\,dx = \frac{(a^2+s^2)^{1/2}-a}{s}$

Theorem: If $f(x)$ and $g(x)$ are two functions and a, b two constants, then

$$H\{af(x)+bg(x)\} = aH\{f(x)\} + bH\{g(x)\}$$

Proof:

$$\begin{aligned} H\{af(x)+bg(x)\} &= \int_0^\infty x[af(x)+bg(x)]J_n(sx)\,dx \\ &= a\int_0^\infty xf(x)J_n(sx)\,dx \ +\ b\int_0^\infty xg(x)J_n(sx)\,dx \\ &= aH\{f(x)\} + bH\{g(x)\} \end{aligned}$$

Theorem: If $H\{f(x)\} = \overline{f}(s)$, then $H\{f(ax)\} = \frac{1}{a^2}\overline{f}\left(\frac{s}{a}\right)$, a being a constant.

Proof: Let $H\{f(x)\} = \overline{f}(s)$. Then

$$\begin{aligned} H\{f(ax)\} &= \int_0^\infty xf(ax)J_n(sx)\,dx \\ &= \int_0^\infty \frac{y}{a}f(y)J_n\left(\frac{sy}{a}\right)\frac{dy}{a}, \quad \text{where } ax = y \\ &= \frac{1}{a^2}\int_0^\infty yf(y)J_n\left(\frac{s}{a}y\right)\,dy \\ &= \frac{1}{a^2}\overline{f}\left(\frac{s}{a}\right) \end{aligned}$$

Example: Find the Hankel transform of $\frac{e^{-ax}}{x}$, taking $xJ_0(sx)$ as the kernel of the transform.

Solution:

$$H\left\{\frac{e^{-ax}}{x}\right\} = \int_0^\infty \frac{e^{-ax}}{x}xJ_0(sx)\,dx = \int_0^\infty e^{-ax}J_0(sx)\,dx = (a^2+s^2)^{-1/2}$$

Theorem: *Hankel transform of the derivatives of a function*

Let $\overline{f}(s)$ be the Hankel transform of order n of the function $f(x)$ and $\overline{f'}_n(s)$ is the transform of $f'(x)$. Then

$$\overline{f'}_n(s) = -\frac{s}{2n}\left[(n+1)\overline{f}_{n-1}(s) - (n-1)\overline{f}_{n+1}(s)\right].$$

Proof:

$$\begin{aligned} \overline{f}_n(s) &= \int_0^\infty x f(x) J_n(sx)\, dx \\ \text{and } \overline{f'}_n(s) &= \int_0^\infty x \frac{df}{dx} J_n(sx)\, dx \quad \Rightarrow \quad \overline{f'}_n(s) = \int_0^\infty \frac{df}{dx}\left[xJ_n(sx)\right]\, dx \end{aligned}$$

Integrating by parts and assuming that $x\, f(x) \to 0$ as $x \to 0$ and $x\, f(x) \to 0$ as $x \to \infty$, we obtain

$$\begin{aligned} \overline{f'}_n(s) &= \Big[f(x)\cdot xJ_n(sx)\Big]_0^\infty - \int_0^\infty f(x)\cdot \frac{d}{dx}[xJ_n(sx)]\, dx \\ &= 0 - \int_0^\infty f(x)\left[J_n(sx) + sx\, J_n'(sx)\right] dx \end{aligned} \tag{15.3.1}$$

But $xJ_n'(x) = -nJ_n(x) + xJ_{n-1}(x)$. Replacing x by sx, we get

$$\begin{aligned} &sxJ_n'(sx) = -nJ_n(sx) + sxJ_{n-1}(sx) \\ &\Rightarrow sxJ_n'(sx) + J_n(sx) = (1-n)J_n(sx) + sxJ_{n-1}(sx) \end{aligned}$$

Using this in Eq.(15.3.1), we have

$$\begin{aligned} \overline{f'}_n(sx) &= -\int_0^\infty f(x)\Big[(1-n)J_n(sx) + sxJ_{n-1}(sx)\Big]\, dx \\ \Rightarrow \overline{f'}_n(sx) &= -(1-n)\int_0^\infty f(x)J_n(sx)\, dx - s\int_0^\infty xf(x)J_{n-1}(sx)\, dx \\ \Rightarrow \overline{f'}_n(sx) &= -(1-n)\int_0^\infty f(x)J_n(sx)\, dx - s\overline{f}_{n-1}(s) \end{aligned} \tag{15.3.2}$$

We know from Recurrence formula for $J_n(x)$, that

$$2nJ_n(x) = x\Big[J_{n-1}(x) + J_{n+1}(x)\Big]$$

Replacing x by sx, we have

$$2nJ_n(sx) = sx\Big[J_{n-1}(sx) + J_{n+1}(sx)\Big]$$

Multiplying this by $f(x)$ and then integrating, we obtain

$$\begin{aligned} 2n\int_0^\infty f(x)J_{sx}\, dx &= s\left[\int_0^\infty xf(x)J_{n-1}(sx)\, dx + \int_0^\infty xf(x)J_{n+1}(sx)\, dx\right] = s\Big[\overline{f}_{n-1}(s) + \overline{f}_{n+1}(s)\Big] \\ \Rightarrow \int_0^\infty f(x)J_n(sx)\, dx &= \frac{s}{2n}\Big[\overline{f}_{n-1}(s) + \overline{f}_{n+1}(s)\Big] \end{aligned}$$

In this event Eq.(15.3.2) becomes

$$\begin{aligned}
\overline{f'}_n(s) &= -\frac{(1-n)s}{2n}\Big[\overline{f}_{n-1}(s) + f_{n+1}(s)\Big] - s\overline{f}_{n-1}(s) \\
\Rightarrow \overline{f'}_n(s) &= \frac{s}{2n}\Big[(n-1)\overline{f}_{n-1}(s) + (n-1)\overline{f}_{n+1} - 2n\overline{f}_{n+1}(s)\Big] \\
\Rightarrow \overline{f'}_n(s) &= \frac{s}{2n}\Big[-(n+1)\overline{f}_{n-1}(s) + (n-1)\overline{f}_{n+1}\Big] \\
\Rightarrow \overline{f'}_n(s) &= -\frac{s}{2n}\Big[(n+1)\overline{f}_{n-1}(s) - (n-1)\overline{f}_{n+1}(s)\Big] \qquad (15.3.3)
\end{aligned}$$

Remark 1. When $n = 1$, from Eq.(15.3.3) we have

$$\begin{aligned}
\overline{f'}_1(s) &= -s\overline{f}_0(s) \\
\Rightarrow H\{f'(x), n=1\} &= -sH\{f(x), n=0\}
\end{aligned}$$

Remark 2. When $n = 2$, from Eq.(15.3.3) we have

$$\overline{f'}_2(s) = -\frac{s}{4}\Big[3\overline{f}_1(s) - \overline{f}_3(s)\Big]$$

Remark 3. When $n = 3$, from Eq.(15.3.3) we have

$$\overline{f'}_3(s) = -\frac{s}{6}\Big[4\overline{f}_2(s) - 2\overline{f}_4(s)\Big]$$

Result 1. Prove that

$$\overline{f''}_n(s) = \frac{s^2}{4}\left[\left(\frac{n+1}{n-1}\right)\overline{f}_{n-2}(s) - 2\left(\frac{n^2-3}{n^2-1}\right)\overline{f}_n(s) + \left(\frac{n-1}{n+1}\right)\overline{f}_{n+2}(s)\right].$$

Proof: From Eq.(15.3.3) we have

$$\overline{f'}_n(s) = -s\left[\left(\frac{n+1}{2n}\right)\overline{f}_{n-1}(s) - \left(\frac{n-1}{2n}\right)\overline{f}_{n+1}(s)\right] \qquad (15.3.4)$$

Replacing f by f', we get

$$\overline{f''}_n(s) = -s\left[\left(\frac{n+1}{2n}\right)\overline{f'}_{n-1}(s) - \left(\frac{n-1}{2n}\right)\overline{f'}_{n+1}(s)\right] \qquad (15.3.5)$$

Replacing n by $(n-1)$, and $(n+1)$ respectively in Eq.(15.3.4), we get

$$\begin{aligned}
\overline{f'}_{n-1}(s) &= -s\left[\left(\frac{n}{2(n-1)}\right)\overline{f}_{n-2}(s) - \left(\frac{n-2}{2(n-1)}\right)\overline{f}_n(s)\right] \\
\overline{f'}_{n+1}(s) &= -s\left[\left(\frac{n+2}{2(n+1)}\right)\overline{f}_n(s) - \left(\frac{n}{2(n+1)}\right)\overline{f}_{n+2}(s)\right]
\end{aligned}$$

Writing Eq.(15.3.5) with the help of these two equations, we obtain

$$\begin{aligned}
\overline{f''}_n(s) &= \frac{s^2}{4}\left[\left(\frac{n+1}{n}\right)\left\{\frac{n}{n-1}\overline{f}_{n-2}(s) - \frac{n-2}{n-1}\overline{f}_n(s)\right\} - \left(\frac{n-1}{n}\right)\left\{\left(\frac{n+2}{n+1}\right)\overline{f}_n(s) - \left(\frac{n}{n+1}\right)\overline{f}_{n+2}(s)\right\}\right] \\
&= \frac{s^2}{4}\left[\left(\frac{n+1}{n-1}\right)\overline{f}_{n-2}(s) - 2\left(\frac{n^2-3}{n^2-1}\right)\overline{f}_n(s) + \left(\frac{n-1}{n+1}\right)\overline{f}_{n+2}(s)\right].
\end{aligned}$$

Example: Find the Hankel transform of $\frac{df}{dx}$, when $f = \frac{e^{-ax}}{x}$ and $n = 1$.

Solution: Let $f(x) = \frac{e^{-ax}}{x}$. To determine $H\left\{\frac{df}{dx}, n = 1\right\} = \overline{f}_1'(s)$, we know that

$$\begin{aligned}
\overline{f}_1'(s) = -sf_0(s) &= -s\int_0^\infty xf(x)J_0(sx)\,dx \\
&= -s\int_0^\infty x\frac{e^{-ax}}{x}J_0(sx)\,dx \\
&= -s\int_0^\infty e^{-ax}J_0(sx)\,dx \\
&= -s(a^2+s^2)^{-1/2}.
\end{aligned}$$

Example: Find the Hankel transform of $x^{-2}e^{-x}$ of order one.

Solution:

$$\begin{aligned}
H\{x^{-2}e^{-x}, n = 1\} &= \int_0^\infty x^{-2}e^{-x}xJ_1(sx)\,dx \\
&= \int_0^\infty \frac{e^{-x}}{x}J_1(sx)\,dx = \frac{(1+s^2)^{1/2}-1}{s}.
\end{aligned}$$

Example: Evaluate $H^{-1}\{s^{-2}e^{as}\}$ when $n = 1$, that is, find out inverse Hankel transform of $s^{-2}e^{-as}$ of order one.

Solution:

$$\begin{aligned}
H^{-1}\{s^{-2}e^{as}, n = 1\} &= \int_0^\infty s^{-2}e^{-as}sJ_1(sx)\,ds \\
&= \int_0^\infty \frac{e^{-as}}{s}J_1(sx)\,ds \\
&= \frac{(a^2+x^2)^{1/2}-a}{x}
\end{aligned}$$

Example: Find the Hankel transformation of

$$f(x) = \begin{cases} 1 & 0 < x < a, \quad n = 0, \\ 0 & x > a, \qquad n = 0. \end{cases}$$

Solution:

$$\begin{aligned}
H\{f(x), n = 0\} &= \int_0^\infty f(x)\cdot xJ_0(sx)\,dx \\
&= \int_0^a f(x)\cdot xJ_0(sx)\,dx + \int_0^\infty f(x)\cdot xJ_0(sx)\,dx \\
&= \int_0^a 1\cdot xJ_0(sx)\,dx + \int_0^\infty 0\cdot xJ_0(sx)\,dx \\
&= \int_0^a xJ_0(sx)\,dx \qquad (15.3.6)
\end{aligned}$$

By Recurrence formula for Bessel's function, we have

$$\frac{d}{dx}\{x^nJ_n(x)\} = x^nJ_{n-1}(x).$$

Replacing n and x by 1 and sx respectively,

$$\begin{aligned}\frac{d}{s\,dx}\{sxJ_1(sx)\} &= sxJ_0(sx)\\ \Rightarrow \frac{d}{dx}\{xJ_1(sx)\} &= sxJ_0(sx).\end{aligned}$$

Integrating this form $x = 0$ to $x = a$,

$$\Big[xJ_1(sx)\Big]_0^a = s\int_0^a xJ_0(sx)\,dx \quad \Rightarrow \int_0^a xJ_0(sx)\,dx = \frac{a}{s}J_1(sa). \tag{15.3.7}$$

Now using Eq.(15.3.7) in Eq.(15.3.6) we obtain

$$H\{f(x), n = 0\} = \frac{a}{s}J_1(sa)$$

Exercise 15.1. (i) Find the Fourier transform of $F(x) = \begin{cases} (1-x^2), & |x| < 1 \\ 0 & |x| > 1. \end{cases}$ and hence evaluate $\int_0^\infty \left(\frac{x\cos x - \sin x}{x^3}\right)\cos\left(\frac{x}{2}\right)dx$. **Answer:** $f(s) = \frac{4}{s^3}(\sin s - s\cos s); -\frac{3\pi}{16}$.

15.4 Finite Hankel Transform

Definition 15.2. Finite Hankel Transform: If $f(x)$ satisfies Dirichlet's conditions in the closed interval $[0, a]$, then its finite Hankel transform $\overline{f}(s_i)$ of order n is given by

$$\overline{f}(s_i) = \int_0^a f(x)\cdot xJ_n(xs_i)\,dx, \tag{15.4.1}$$

where a is positive root of the transcendental equation

$$J_n(as_i) = 0 \tag{15.4.2}$$

If the function $f(x)$ is continuous at any point of the interval $[0, a]$, then the inversion formula for $\overline{f}(s_i)$ is

$$f(x) = \frac{2}{a^2}\sum_i \overline{f}(s_i)\frac{J_n(xs_i}{[J_n'(ac_i)]^2} \tag{15.4.3}$$

where the sum is taken over all the positive roots of the Eq.(15.4.2). If $f(x)$ is represented by generalised Fourier Bessel series

$$f(x) = \sum_i c_i J_n(xs_i), \quad 0 \le x \le a, \tag{15.4.4}$$

then the coefficient c_i is given by

$$\begin{aligned}c_i &= \frac{2}{a^2J_{n+1}^2(as_i)}\int_0^a f(x)\cdot xJ_n(xs_i)\,dx\\ &= \frac{2\overline{f}}{a^2[J_{n+1}(as_i)]^2} = \frac{2\overline{f}(s_i)}{a^2[J_n'(as_i)]^2}\end{aligned}$$

The recurrence formula,

$$
\begin{aligned}
&xJ_n'(x) = nJ_n(x) - xJ_{n+1}(x).\\
&\therefore as_iJ_n'(as_i) = nJ_n(as_i) - as_iJ_{n+1}(as_i), \quad \text{[Replacing } x \text{ by } as_i\text{]}\\
&\Rightarrow as_iJ_n'(as_i) = -as_iJ_{n+1}(as_i) \quad \text{[Using (15.4.2)]}\\
&\Rightarrow J_n'(as_i) = -J_{n+1}(as_i).
\end{aligned}
$$

Consequently,

$$f(x) = \frac{2}{a^2}\sum_i \overline{f}(s_i)\frac{J_n(xs_i)}{[J_{n+1}(as_i)]^2}$$

Remark 15.3. It has been found in practice that the choice of unity as the upper limit of the integral defining the transform is more convenient. Therefore the definition of finite Hankel transform becomes

$$\overline{f}(s_i) = \int_0^1 f(x)\cdot xJ_n(xs_i)\,dx$$

Theorem 15.4. Finite Hankel transform of $\frac{df}{dx}$, i.e.,

$$H_n\left(\frac{df}{dx}\right) = \int_0^a \frac{df}{dx}xJ_n(sx)\,dx,$$

where s is any root of $J_n(sa) = 0$. To show that

$$H_n\left\{\frac{df}{dx}\right\} = \frac{s}{2n}[(n-1)H_{n+1}\{f(x)\} - (n+1)H_{n-1}\{f(x)\}]$$

Proof. The finite Hankel transform of $\frac{df}{dx}$ of order n is denoted by $H_n\left\{\frac{df}{dx}\right\}$.

$$
\begin{aligned}
H_n\left\{\frac{df}{dx}\right\} &= \int_0^a \frac{df}{dx}\cdot xJ_n(sx)\,dx\\
&= \Big[f(x)\cdot xJ_n(sx)\Big]_{x=0}^{x=a} - \int_0^a f(x)\cdot\frac{d}{dx}\{xJ_n(sx)\}\,dx\\
&= -\int_0^a f(x)\frac{d}{dx}\{xJ_n(sx)\}\,dx \qquad (15.4.5)
\end{aligned}
$$

By Recurrence formula,

$$2J_n'(x) = J_{n-1}(x) - J_{n+1}(x) \quad \text{and} \quad 2nJ_n(x) = x[J_{n-1}(x) + J_{n+1}(x)]$$

Replacing x by sx in both equations,

$$2J_n'(sx) = J_{n-1}(sx) - J_{n+1}(sx) \quad \text{and} \quad 2nJ_n(sx) = sx[J_{n-1}(sx) + J_{n+1}(sx)]$$

Now

$$
\begin{aligned}
\frac{d}{dx}\{xJ_n(sx)\} &= J_n(sx) + sxJ_n'(sx)\\
&= \frac{sx}{2n}[J_{n-1}(sx) + J_{n+1}(sx)] + \frac{sx}{2}[J_{n-1}(sx) - J_{n+1}(sx)]\\
&= \frac{sx}{2n}[J_{n-1}(sx)\cdot(1+n) + (1-n)J_{n+1}(sx)]
\end{aligned}
$$

Now from Eq.(15.4.5) we have

$$\begin{aligned} H_n\left\{\frac{df}{dx}\right\} &= -\int_0^a f(x)\cdot\frac{sx}{2n}[J_{n-1}(sx)\cdot(n-1)+(1-n)J_{n+1}(sx)] \\ &= -\frac{s}{2n}\int_0^a [f(x)\cdot xJ_{n-1}(sx)\cdot(1+n)-(n-1)f(x)xJ_{n+1}(sx)]dx \\ &= \frac{s}{2n}[(n-1)H_{n+1}\{f(x)\}-(1+n)H_{n-1}\{f(x)\}] \end{aligned}$$

□

Corollary 15.5. If $n=1$, the the last gives

$$H_1\left\{\frac{df}{dx}\right\} = \frac{s}{2}[0-2H_0\{f(x)\}] = -sH_0\{f(x)\}$$

Theorem 15.6.

$$H\left\{\frac{d^2f}{dx^2}+\frac{1}{x}\frac{df}{dx}\right\} = \frac{s}{2n}\left[-H_{n-1}\left\{\frac{df}{dx}\right\}+H_{n+1}\left\{\frac{df}{dx}\right\}\right]$$

$$H\left\{\frac{d^2f}{dx^2}+\frac{1}{x}\frac{df}{dx}-\frac{n^2}{x^2}f\right\} = -sa\ f(a)J_n'(sa)-s^2H_n\{f(x)\}$$

Proof. Proof of the above theorems are left as exercise. □

Example 15.7. Show that

$$H_0(c) = \frac{ca}{s}J_1(as)$$

Solution:

$$H_0\{c\} = \int_0^a cxJ_0(sx)\,dx = c\int_0^a xJ_0(sx)\,dx \tag{15.4.6}$$

By recurrence formula No. (vi), we know that

$$\frac{d}{dx}[x^nJ_n(x)] = x^nJ_{n-1}(x).$$

Putting $n=1$, $\quad\frac{d}{dx}[xJ_1(x)] = xJ_0(x).$

Replacing x *by* sx, we have

$$\frac{d}{s\,dx}[sxJ_1(sx)] = sxJ_0(sx)$$

$$\Rightarrow \frac{d}{dx}\{xJ_1(sx)\} = sxJ_0(sx).$$

Using this in (15.4.6), we get

$$H_0\{c\} = \frac{c}{s}\int_0^a \frac{d}{dx}\{xJ_1(sx)\} = \frac{c}{s}\Big[xJ_1(sx)\Big]_0^\infty = \frac{ca}{s}J_1(sa).$$

Example 15.8. Find finite Hankel transform of x^2 if $xJ_0(sx)$ is the Kernel of the transform.

Solution: By recurrence formula No. (iv) and (vi) we have

$$2nJ_n(x) = x[J_{n-1}(x) + J_{n+1}(x)] \quad \text{and} \quad \frac{d}{dx}[x^n J_n(x)] = x^n J_{n-1}(x)$$

Replacing x by sx, we have

$$2nJ_n(sx) = sx[J_{n-1}(sx) + J_{n+1}(sx)] \tag{15.4.7}$$

$$\frac{d}{s\,dx}[x^n J_n(sx)] = x^n J_{n-1}(sx). \tag{15.4.8}$$

Now

$$H_0\{x^2\} = \int_0^a x^2 \cdot xJ_0(sx)\,dx = \int_0^a x^2 \cdot \frac{d}{s\,dx}\{xJ_1(sx)\}dx, \text{ according to (15.4.8)}$$

Integrating by parts, we obtain

$$\begin{aligned} H_0\{x^2\} &= \frac{1}{s}\Big[x^2 \cdot xJ_1(sx)\Big]_0^a - \frac{1}{s}\int_0^a 2x \cdot xJ_1(sx)\,dx \\ &= \frac{a^3}{s}J_1(sa) - \frac{2}{s}\int_0^a x^2 J_1(sx)\,dx \\ &= \frac{a^3}{s}J_1(sa) - \frac{2}{s}\int_0^a \frac{d}{s\,dx}[x^2 J_2(sx)]\,dx, \quad \text{[according to (15.4.8)]} \\ &= \frac{a^3}{s}J_1(sa) - \frac{2}{s^2}\Big[x^2 J_2(sx)\Big]_0^a \\ &= \frac{a^3}{s}J_1(sa) - \frac{2a^2}{s^2}J_2(sa). \end{aligned} \tag{15.4.9}$$

Putting $n = 1,\ x = a$ in (15.4.7), we have

$$2J_1(sa) = sa[J_0(sa) + J_2(sa)] \quad \Rightarrow \quad \frac{2}{sa}J_1(sa) - J_0(sa) = J_2(sa).$$

Putting this in Eq.(15.4.9), we obtain

$$\begin{aligned} H_0\{x^2\} &= \frac{a^2}{s}J_1(sa) - \frac{2a^2}{s^2}\left[\frac{2}{sa}J_1(sa) - J_0(sa)\right] \\ &= \frac{a^3}{s}J_1(sa) - \frac{4a^2}{s^3 a}J_1(sa) + \frac{2a^2}{s^2}J_0(sa) \\ &= \frac{a^2}{s^2}\left[\left(as - \frac{4}{as}\right)J_1(sa) + 2J_0(sa)\right] \end{aligned}$$

Example 15.9. Find the finite Hankel transform of

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial V}{\partial r}\right) - \frac{n^2 V}{r^2}, \quad \text{where} \quad V = \begin{cases} 0 & when\ r = 0 \\ V_1 & when\ r = 1. \end{cases}$$

Solution: From the problem it is clear that we should take the limits $x = 0$ and $x = 1$ of the transform.

$$\text{Let } f(r) = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial V}{\partial r}\right) - \frac{n^2 V}{r^2}$$

Then $f(r)$ is expressible as

$$\begin{aligned} f(r) &= \frac{1}{r}\left(\frac{\partial V}{\partial r} + r\frac{\partial^2 V}{\partial r^2}\right) - \frac{n^2 V}{r^2} \\ &= \frac{\partial^2 V}{\partial r^2} + \frac{1}{r}\frac{\partial V}{\partial r} - \frac{n^2 V}{r^2}. \end{aligned} \tag{15.4.10}$$

By Theorem 15.6, we can write

$$H_n\{f(r)\} = -sf(1)J_n'(s) - s^2 H_n\{f\} = -sV_1 J_n'(s) - s^2 H_n\{f\}$$

Exercise 15.10. (i) Find the finite Hankel transform of $x^n,\ (n > -1)$ if $xJ_n(sx)$ is the Kernel of the transform. **Answer:** $H_n\{x^n\} = \frac{a^{n+1}}{s}J_{n+1}(sa)$.
(ii) Find the finite Hankel transform of $(1 - x^2)$, taking $xJ_0(sx)$ as the kernel.
Answer: $H_0\{1 - x^2\} = \frac{a}{s}J_1(as) - \frac{a^2}{s^2}\left[\left(as - \frac{4}{as}\right)J_1(sa) + 2J_0(sa)\right]$
(iii) Find the Hankel transform of $(a^2 - x^2)$ if $xJ_0(sx)$ is the kernel of the transform.
Answer: $H_0\{a^2 - x^2\} = \frac{4a}{s^3}J_1(sa) - \frac{2a^2}{s^2}J_0(sa)$
(iv) Show that

$$\int_0^a r^3 J_0\{pr\}\, dr = \frac{a^2}{p^2}\left[2J_0(pa) + \left(ap - \frac{4}{ap}J_1(pa)\right)\right]$$

Unit 16

Course Structure

Applications : Applications of integral transforms to solve two-dimensional Laplace and one dimensional diffusion and wave equations.

---

16 Introduction

The given partial differential equations are given along with certain prescribed conditions on the functions which arise from the physical situation. The conditions which are given at $t = 0$ are known as *initial conditions* whereas the conditions given at the boundary of the region or interval are called *boundary conditions.* Most of the well known partial differential equations like Laplace equation, Heat equation and Wave equation can be solved by using the method of integral transform. Readers are suggested to familiar with the following important partial differential equations.

1. One dimensional heat conduction or diffusion equation:

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2}, \quad 0 < x < \infty, t > 0$$

Here $u(x,t)$ is the temperature in a solid at position x at time t. The constant k is called the diffusivity of the material of the solid. Again $k = K/\sigma\rho$, where the thermal conductivity K, the specific heat σ and the density ρ are assumed constant. The amount of heat per unit area per unit time conducted across a plane is given by $-K\ u_x(x,t)$.

2. One dimensional wave equation:

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \quad x > 0, t > 0$$

This equation is applicable to the small transverse vibrations of a taut flexible string initially located on the $x-$axis and set into motion. Here $u(x,t)$ is the transverse displacement of the string at any time t. Again, $c^2 = T/\rho$, where T is constant tension in the string and ρ is constant mass per unit length of the string.

3. Two dimensional Laplace's equation:

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

We can solve one dimensional heat and wave equation by the method of Laplace transform as well as Fourier transform while only Fourier transform method is used to solved boundary value problem governed by Laplace equation.

Objective

The unit is aimed at exposing the students to learn the application of Laplace transforms and Fourier transforms. To make them familiar with the methods of solving differential equations, partial differential equations, IVP and BVP using Laplace transforms and Fourier transforms.

16.1 Solution of two dimensional Laplace Equation using Finite Fourier Transform (FFT)

In this subsection, we will solve the two dimensional Laplace equation over a finite region using the finite Fourier transform. Let us begin with the following example. For infinite or semi infinite range, one may use infinite cosine or sine transform.

Example 16.1. Determine a function $V(x, y)$ which is harmonic in the open square $0 < x < \pi,\ 0 < y < \pi$, takes a constant value V_0 on the edge $y = \pi$ and vanishes on the other edges of the square.

or

Find the steady temperature $V(x, y)$ in a long square bar of side π when one face is kept at constant temperature V_0 and the other faces at zero temperature. Also $V(x, y)$ is bounded.

Solution: The steady temperature $V(x, y)$ is governed by the Laplace equation (since V is harmonic)

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0 \tag{16.1.1}$$

with the conditions (i) $V(x, \pi) = V_0$, (ii) $V(0, y) = 0 = V(\pi, y)$ for every y and (iii) $V(x, y)$ is bounded. Taking finite Fourier sine transform of Eq. (16.1.1), we have

$$\int_0^\pi \frac{\partial^2 V}{\partial x^2} \sin sx\, dx + \int_0^\pi \frac{\partial^2 V}{\partial y^2} \sin sx\, dx = 0$$

$$\Rightarrow \left[\frac{\partial V}{\partial x} \sin sx\right]_0^\pi - s \int \frac{\partial V}{\partial x} \cos sx\, dx + \frac{\partial^2}{\partial y^2} \int_0^\pi V\ \sin sx\, dx = 0$$

$$\Rightarrow \frac{d^2\overline{V_s}}{dy^2} + 0 - s \left[\Big(V \cos sx\Big)_0^\pi + s \int_0^\pi V \sin sx\, dx\right] = 0$$

$$\Rightarrow \frac{d^2\overline{V_s}}{dy^2} - s^2\overline{V_s} - s\Big[V(\pi, y) \cos s\pi - V(0, y) \cos 0\Big] = 0$$

$$\Rightarrow \frac{d^2\overline{V_s}}{dy^2} - s^2\overline{V_s} - s\Big[V(\pi, y) \cos s\pi - V(0, y) \cos 0\Big] = 0$$

$$\Rightarrow \frac{d^2\overline{V_s}}{dy^2} - s^2\overline{V_s} = 0 \qquad \text{[using boundary condition (ii)]}$$

The solution of this equation is

$$V_s = A \cosh sy + B \sinh sy. \tag{16.1.2}$$

$$\text{Now, } V(x,\pi)=V_0 \;\Rightarrow\; F_s\{V(x,\pi)\}=F_s\{V_0\}$$
$$\Rightarrow \int_0^{\pi} V(x,\pi)\sin sx\,dx=\int_0^{\pi} V_0\sin sx\,dx$$
$$\Rightarrow \overline{V_s}(s,\pi)=\frac{V_0}{s}\Big(-\cos sx\Big)_0^{\pi}=V_0\left(\frac{1-\cos s\pi}{s}\right)$$
$$\therefore\; \overline{V_s}(s,\pi)=V_0\frac{1-\cos s\pi}{s}$$

$$\text{Again, } V(x,0)=0 \;\Rightarrow\; \overline{V_s}(s,0)=0$$
$$\Rightarrow A\cdot 1+B\cdot 0=0 \quad [\text{Using Eq.(16.1.2)}]$$
$$\Rightarrow A=0$$

$$\therefore \text{From Eq. (16.1.2), we have } \overline{V_s}=B\sinh sy, \tag{16.1.3}$$

$$\Rightarrow \overline{V_s}(s,\pi)=B\sinh(s\pi)$$
$$\Rightarrow V_0\left(\frac{1-\cos s\pi}{s}\right)=B\sinh s\pi$$
$$\Rightarrow B=V_0\frac{1-\cos s\pi}{s\,\sinh s\pi}$$

Now Eq.(16.1.3) takes the form

$$\overline{V_s}=\frac{V_0}{s}\frac{(1-\cos s\pi)}{\sinh s\pi}\sinh(sy)$$

Taking inverse finite sine transform, we obtain

$$V(x,y) = \frac{2}{\pi}\sum_{s=1}^{\infty}\frac{V_0(1-\cos s\pi)\sinh sy\sin sx}{s\;\sinh s\pi}$$
$$= \frac{2V_0}{\pi}\sum_{s=1}^{\infty}\frac{[1-(-1)^s]\sinh sy\sin sx}{s\sinh s\pi}$$
$$= \frac{4V_0}{\pi}\sum_{n=0}^{\infty}\frac{\sinh(2n+1)y\cdot\sin(2n+1)x}{(2n+1)\sinh(2n+1)\pi}.$$

Example 16.2. Use a cosine transform to show that the steady temperature in the semi-infinite solid $y>0$ when the temperature on the surface $y=0$ is kept at unity over the strip $|x|<a$ and at zero outside the strip, is

$$\frac{1}{\pi}\left[\tan^{-1}\left(\frac{a+x}{y}\right)+\tan^{-1}\left(\frac{a-x}{y}\right)\right]$$

The result $\int_0^{\infty} e^{-sx}x^{-1}\sin rx\,dx=\tan^{-1}\frac{r}{s}$, $r>0, s>0$ may be assumed.

Solution: We know that the steady temperature in the semi-infinite solid is represented by two-dimensional Laplace equation

$$\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}=0 \quad 0<y<\infty, \quad -\infty<x<\infty \tag{16.1.4}$$

subject to the boundary conditions:

$$U(x,0) = 1, \quad |x| < a \quad i.e. \; -a < x < a \tag{16.1.5}$$

$$U(x,0) = 1, \quad x < -a \quad or \quad x > a \tag{16.1.6}$$

Taking the Fourier cosine transform of (16.1.4), we get

$$\int_0^\infty \frac{\partial^2 U}{\partial x^2} \cos sx \, dx + \int_0^\infty \frac{\partial^2 U}{\partial y^2} \cos sx \, dx = 0$$

$$\Rightarrow \left[\frac{\partial U}{\partial x} \cos sx\right]_0^\infty - \int_0^\infty \frac{\partial U}{\partial x}(-s \, \sin sx) \, dx + \frac{d^2}{dy^2} \int_0^\infty U(x,y) \cos sx \, dx = 0$$

$$\Rightarrow s \int_0^\infty \frac{\partial U}{\partial x} \sin sx \, dx + \frac{d^2 \overline{U_c}}{dy^2} = 0, \quad \text{where } \; \overline{U_c}(s,y) = \int_0^\infty U(x,y) \; \cos sx \, dx$$

$$\left[\because \text{due to symmetry, } \; \frac{\partial U}{\partial x} \to 0 \text{ as } x \to \infty \text{ and } \frac{\partial U}{\partial x} \to 0 \text{ as } x \to 0.\right]$$

$$\Rightarrow s \left\{ [U(x,y) \sin sx]_0^\infty - \int_0^\infty U(x,y) \; s \; \cos sx \, dx \right\} + \frac{d^2 \overline{U_c}}{dy^2} = 0$$

$$\Rightarrow -s^2 \overline{U_c} + \frac{d^2 \overline{U_c}}{dy^2} = 0 \qquad \text{if } \; U(x,y) \to 0 \; \text{ as } \; x \to \infty$$

$$\Rightarrow (D^2 - s^2)\overline{U_c} = 0 \qquad \text{where } \; D \equiv \frac{d}{dy}$$

whose general solution is

$$\overline{U_c}(s,y) = C_1 e^{sy} + C_2 e^{-sy}, \quad C_1 \text{ and } C_2 \text{ being arbitrary constants.} \tag{16.1.7}$$

Since $\overline{U_c}(s,y)$ is finite, we must take $C_1 = 0$ in (16.1.7), otherwise $\overline{U_c}(s,y)$ would become infinite as $y \to \infty$. Hence (16.1.7) reduces to

$$\overline{U_c}(s,y) = C_2 e^{-sy} \tag{16.1.8}$$

Again

$$\int_0^\infty U(x,0) \; \cos sx \, dx = \int_0^a U(x,0) \; \cos sx \, dx + \int_a^\infty U(x,0) \; \cos sx \, dx$$

$$\Rightarrow \overline{U_c}(s,0) = \int_0^a \cos sx \, dx = \frac{\sin sa}{s} = C_2 \tag{16.1.9}$$

Hence from (16.1.8), we finally find

$$\overline{U_c}(s,y) = \frac{\sin sa}{s} e^{-sy} \tag{16.1.10}$$

Now taking the inverse Fourier cosine transform, we get

$$\begin{aligned} U(x,y) &= \frac{2}{\pi} \int_0^\infty \frac{\sin sa}{s} e^{-sy} \cos sx \, ds = \frac{1}{\pi} \int_0^\infty \frac{e^{-sy}}{s} \Big[\sin(a+x)s + \sin(a-x)s \Big] \, ds \\ &= \frac{1}{\pi} \left[\tan^{-1}\left(\frac{a+x}{y}\right) + \tan^{-1}\left(\frac{a-x}{y}\right) \right] \end{aligned}$$

Exercise 16.3. (i) Using the finite Fourier transform, solve $\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=0,\ 0<x<\pi,\ 0<y<y_0$ subject to $u(0,y)=0, u(\pi,y)=1, u_y(x,0)=0, u(x,y_0)=0$ **Answer:** $u(x,y)=\dfrac{2}{\pi}+\dfrac{2}{\pi}\displaystyle\sum_{n=1}^{\infty}\dfrac{(-1)^n}{n}\dfrac{\cosh ny}{\cosh ny_0}\sin nx$

(ii) Solve the boundary value problem in the half-plane $y>0$, described by $\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=0,\ -\infty<x<\infty,\ y>0$ subject to $u(x,0)=f(x),\ -\infty<x<\infty$, u is bounded as $y\to\infty$, u and $\dfrac{\partial u}{\partial x}$ both vanish as $|x|\to\infty$. **Answer:** $u(x,y)=\dfrac{y}{\pi}\displaystyle\int_{-\pi}^{\pi}\dfrac{f(\xi)}{(\xi-x)^2+y^2}d\xi$

(iii) Solve the boundary value problem in the half-plane $x>0$, described by $\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}=0,\ -\infty<y<\infty,\ x>0$ subject to $u(0,y)=f(y),\ -\infty<y<\infty$, u is bounded as $x\to\infty$, u and $\dfrac{\partial u}{\partial y}$ both vanish as $|y|\to\infty$. **Answer:** $u(x,y)=\dfrac{x}{\pi}\displaystyle\int_{-\pi}^{\pi}\dfrac{f(\xi)}{x^2+(y-\xi)^2}d\xi$

16.2 Application to Heat Conduction and Wave Equations

16.2.1 Formulae for Laplace transform method

In order to solve heat equation using the method of Laplace transform, the following results will be used frequently

$$
\begin{aligned}
&L\{u(x,t)\}=\overline{u}(x,s)\\
&L\left\{\frac{\partial u}{\partial x}\right\}=\frac{d\overline{u}}{dx},\quad L\left\{\frac{\partial^2 u}{\partial x^2}\right\}=\frac{d^2\overline{u}}{dx^2},\\
&L\left\{\frac{\partial u}{\partial t}\right\}=s\overline{u}-u(x,0)\\
&L\left\{\frac{\partial^2 u}{\partial t^2}\right\}=s^2\overline{u}-s\,u(x,0)-u_t(x,0).
\end{aligned}
$$

Example 16.4. Find the temperature $u(x,t)$ in a slab whose ends $x=0$ and $x=a$ are kept at temperature zero and whose initial temperature is $\sin(\pi x/a)$.

Solution: We have to solve one-dimensional heat conduction equation

$$\frac{\partial u}{\partial t}=k\frac{\partial^2 u}{\partial x^2},\ 0<x<a,\ t>0, \tag{16.2.1}$$

$u(x,t)$ being the temperature in the slab at any point x at any time t and k being the diffusivity of the material of the bar, subject to the boundary conditions $u(0,t)=0$, $u(a,t)=0$ and initial condition $u(x,0)=\sin(\pi x/a)$. Let $L\{u(x,t)\}=\overline{u}(x,s)$. Taking the Laplace transform of both sides of (16.2.1), we have

$$
\begin{aligned}
L\left\{\frac{\partial u}{\partial t}\right\}=kL\left\{\frac{\partial^2 u}{\partial x^2}\right\}\ &\Rightarrow\ s\overline{u}(x,s)-u(x,0)=k\frac{d^2\overline{u}}{dx^2}\\
\Rightarrow s\overline{u}-\sin\left(\frac{\pi x}{a}\right)=k\frac{d^2\overline{u}}{dx^2}\ &\Rightarrow\ \left(D^2-\frac{s}{k}\right)\overline{u}=-\frac{1}{k}\sin\frac{\pi x}{a}
\end{aligned}
$$

Calculating the complementary function corresponding to the homogeneous part and the particular solution using classical methods of ordinary differential equations, we may write the general solution of the aforesaid ODE as

$$\overline{u}(x,s) = c_1 e^{x\sqrt{s/k}} + c_1 e^{-x\sqrt{s/k}} + \frac{1}{s+(\pi^2 k/a^2)} \sin\frac{\pi x}{a}. \tag{16.2.2}$$

Taking the Laplace transform of boundary conditions, we have

$$\overline{u}(0,s) = 0 \quad \text{and} \quad \overline{u}(a,s) = 0 \tag{16.2.3}$$

Now using the conditions (16.2.3) in (16.2.1), we obtain $c_1 = c_2 = 0$ and therefore (16.2.1) reduces to

$$\overline{u}(x,s) = \frac{\sin(\pi x/a)}{s+(\pi^2 k/a^2)} \quad \text{so that} \quad u(x,t) = sin\frac{\pi x}{a} L^{-1}\left\{\frac{1}{s+(\pi^2 k/a^2)}\right\} = \sin\frac{\pi x}{a} e^{-(\pi^2 kt/a^2)}$$

Example 16.5. The faces $x = 0$ and $x = 1$ of a slab of material for which $k = 1$ are kept at temperature 0 and 1 respectively until the temperature distribution becomes $u = x$. After time $t = 0$ both faces are held at temperature 0. Determine the temperature formula. It is given that

$$L^{-1}\left\{\frac{\sinh x\sqrt{s}}{s\ \sinh a\sqrt{s}}\right\} = \frac{x}{a} + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n} e^{-n^2\pi^2 t/a^2} \sin\left(\frac{n\pi x}{a}\right).$$

Solution: The temperature $u(x,t)$ in the slab is governed by the partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{16.2.4}$$

with the boundary conditions (i) $u(0,t) = 0$, (ii) $u(1,t) = 0$, (iii) $u(x,0) = x$. From Eq. (16.2.4) we have

$$\begin{aligned} L\left\{\frac{\partial u}{\partial t}\right\} &= L\left\{\frac{\partial^2 u}{\partial x^2}\right\} \\ \Rightarrow s\,\overline{u} - u(x,0) &= \frac{d^2\overline{u}}{dx^2} \\ \Rightarrow \frac{d^2\overline{u}}{dx^2} - s\,\overline{u} &= -x \qquad [\because u(x,0) = x] \\ \Rightarrow (D^2 - s)\overline{u} &= -x \end{aligned}$$

The solution of it is

$$\begin{aligned} \overline{u} &= a\,e^{-x\sqrt{s}} + b\,e^{x\sqrt{s}} + \frac{1}{D^2-s}(-x) \\ &= a\,e^{-x\sqrt{s}} + b\,e^{x\sqrt{s}} + \frac{1}{s}\left(1 - \frac{D^2}{s}\right)^{-1} x \\ &= a\,e^{-x\sqrt{s}} + b\,e^{x\sqrt{s}} + \frac{x}{s}. \end{aligned}$$

It is also expressed as

$$\overline{u} = a\ \cosh x\sqrt{s} + b\ \sinh x\sqrt{s} + \frac{x}{s}.$$

Now

$$(i) \Rightarrow L\{u(0,t)\} = 0 \Rightarrow \overline{u}(0,s) = 0 \Rightarrow a = 0$$

Hence,

$$\overline{u} = b\ \sinh x\sqrt{s} + \frac{x}{s}.$$

$$(ii) \Rightarrow L\{u(1,t)\} = 0 \Rightarrow \overline{u}(1,s) = 0 \Rightarrow 0 = b\ \sinh\sqrt{s} + \frac{1}{s} \Rightarrow b = -\frac{1}{s\ \sinh\sqrt{s}}$$

Using this we obtain

$$\begin{aligned}
&\overline{u} = \frac{x}{s} - \frac{\sinh x\sqrt{s}}{s\ \sinh\sqrt{s}} \\
&\Rightarrow u = L^{-1}\left\{\frac{x}{s}\right\} - L^{-1}\left\{\frac{\sinh x\sqrt{s}}{s\ \sinh\sqrt{s}}\right\} \\
&\Rightarrow u = x - \left[x + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}e^{-n^2\pi^2 t}\sin(n\pi x)\right] \\
&\Rightarrow u = -\frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}e^{-n^2\pi^2 t}\sin(n\pi x)
\end{aligned}$$

Example 16.6. Solve the wave equation $\dfrac{\partial^2 u}{\partial t^2} + c^2\dfrac{\partial^2 u}{\partial x^2}, x > 0, t > 0$, where $u(x,0) = 0,\ u_t(x,0) = 0,\ x > 0$ and $u(0,t) = F(t)$. $\lim\limits_{x\to\infty} u(x,t) = 0,\ t > 0$.

Solution: We have to solve one-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2} \tag{16.2.5}$$

subject to boundary condition: $u(0,t) = F(t), \quad \lim\limits_{x\to\infty} u(x,t) = 0$
and initial conditions: $u(x,0) = 0, \quad u_t(x,0) = 0.$
Let $L\{u(x,t)\} = \overline{u}(x,s)$. Applying Laplace transform to Eq.(16.2.5), we have

$$s^2\overline{u}(x,s) - su(x,0) - u_t(x,0) = c^2\frac{d^2\overline{u}}{dx^2} \quad \Rightarrow \quad \frac{d^2\overline{u}}{dx^2} - \frac{s^2}{c^2}\overline{u} = 0$$

Its solution is

$$\overline{u}(x,s) = c_1\ e^{sx/c} + c_2\ e^{-sx/c}, \quad c_1, c_2 \text{ being the arbitrary constants}$$

Now using the above boundary conditions we have $\overline{u}(0,s) = f(s)$ where $f(s) = L\{F(t)\}$ and $\overline{u}(x,s) = 0$ as $x \to \infty$. Since $\overline{u}(x,s) = 0$ as $x \to \infty$, we must choose $c_1 = 0$. Hence the solution reduces to

$$\overline{u}(x,s) = c_2\ e^{-sx/c}$$

Putting $x = 0$ in the above equation and using $\overline{u}(0,s) = f(s)$, we get $c_2 = f(s)$. Then the solution reduces to

$$\overline{u}(x,s) = f(s)\ e^{-sx/c}$$

Taking inverse Laplace transform, we obtain

$$u(x,t) = L^{-1}\{f(s)\ e^{-sx/c}\} = F(t - x/c)H(t - x/c)$$

where $H(t - x/c)$ is the Heaviside unit step function.

Exercise 16.7. (i) A string is stretched between two fixed points $(0, 0)$ and $(a, 0)$. If it is displaced into the curve $u = b\ \sin(\pi x/a)$ and released from rest in that position at time $t = 0$, find its displacement at any time $t < 0$ and at any point $0 < x < a$. **Answer:** $u(x,t) = b\ \sin\dfrac{\pi x}{a}\cos\dfrac{\pi ct}{a}$

(ii) Solve the boundary value problem $\dfrac{\partial^2 u}{\partial t^2} = a^2\dfrac{\partial^2 u}{\partial x^2} - g,\ x > 0,\ t > 0$ with the boundary conditions $u(x,0) = 0 = u_t(x,0),\ x > 0;\ u(0,t) = 0,\ \lim\limits_{x\to\infty} u_x(x,t) = 0,\ t \geq 0.$

Answer: $u(x,t) = \dfrac{1}{2}g(t - x/a)^2 H(t - x/a) - \dfrac{1}{2}gt^2$

Reference

(i) M.D. Raisinghania, Integral Equations and Boundary Value Problems.

(ii) Wazwaz and Abdul, A First Course in Integral Equations

(iii) D.C. SHARMA and M. C. GOYAL, INTEGRAL EQUATIONS

(iv) M.D. Raisinghania Advanced Differential Equations

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER III

SELF LEARNING MATERIAL

PAPER : MATC 3.2
(Pure & Applied Streams)

Block - I : Numerical Analysis (Theory)
Block - II : Calculus of $\mathbb{R}^n$

Directorate of Open and Distance Learning
University of Kalyani
Kalyani, Nadia
West Bengal, India

Course Preparation Team

1. Mr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani	2. Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

November, 2019

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

Director's Massage

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Sankar Kumar Ghosh, Hon'ble Vice-Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as weel as each and every member of PGBOS (DODL), University of Kalyani, Heartfelt thanks is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

CONTENTS

Core Paper

MATC 3.3

Block - II

Marks : 50 (SSE : 40; IA : 10)

Numerical Analysis (Theory)
(Pure and Applied Streams)

Syllabus

• Unit 1 • Interpolation : Hermite's interpolation. Interpolation by iteration – Aitken's and Neville's schemes.

• Unit 2 • Approximation of Function : Least square approximation. Weighted least square approximation. Orthogonal polynomials, Gram – Schmidt orthogonalisation process, Chebysev polynomials, Minimax polynomial approximation.

• Unit 3 • Numerical Integration : Gaussian quadrature formula and its existence. Euler- MacLaurin formula. Gregory-Newton quadrature formula. Romberg integration.

• Unit 4 • Systems of Linear Algebraic Equations : Direct methods, Factorization method; Eigen value and Eigenvector Problems : Direct methods, Iterative method – Power method.

• Unit 5 • Nonlinear Equations : Fixed point iteration method, convergence and error estimation. Modified Newton-Raphson method, Muller's method, Inverse inter- polation method, error estimations and convergence analysis.

• Unit 6 • Ordinary Differential Equations: Initial value problems – Picard's successive approximation method, error estimation. Single-step methods – Euler's method and Runge-Kutta method, error estimations and convergence analysis.

• Unit 7 • Ordinary Differential Equations: Multi-step method – Milne's predictor-corrector method, error estimation and convergence analysis.

• Unit 8 • Partial Differential Equations: Finite difference methods for Elliptic and Parabolic differential equations.

Unit 1

Course Structure

Interpolation : Hermite's interpolation; Interpolation by iteration – Aitken's and Neville's schemes.

1 Introduction

The statement $y = f(x),\ \ x_0 \le x \le x_n$ means: corresponding to every value of x in the range $x_0 \le x \le x_n$, there exists one or more values of y. Assuming that $f(x)$ is single-valued and continuous and that it is known explicitly, then the values of $f(x)$ corresponding to certain given values of x, say $x_0, x_1, \ldots, x_n$ can easily be computed and tabulated. The central problem of numerical analysis is the converse one: Given the set of tabular values $(x_0, y_0),\ (x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$ satisfying the relation $y = f(x)$ where the explicit nature of $f(x)$ is not known, it is required to find a simpler function, say $\phi(x)$, such that $f(x)$ and $\phi(x)$ *agree at the set of tabulated points.* Such a process is called *interpolation*. If $\phi(x)$ is a polynomial, the the process is called *polynomial interpolation* and $\phi(x)$ is called the *interpolating polynomial*. In this unit, we shall be concerned with Hermite's interpolation and iterative interpolation by Aitken's and Neville's schemes.

1.1 Hermite's Interpolation Formula

The interpolation formulae so far considered make use of only a certain number of function values. We now derive an interpolation formula in which both the function and its first derivative values are to be assigned at each point of interpolation. This is refereed to as *Hermite's interpolation formula.* The interpolation problem is then defined as follows: Given the set of data points $(x_i, y_i, y_i'),\ i = 0, 1, \ldots, n$, it is required to determine a polynomial of the least degree, say $H_{2n+1}(x)$, such that

$$H_{2n+1}(x_i) = y_i \quad \text{and} \quad H'_{2n+1}(x_i) = y_i'; \quad i = 0, 1, \ldots, n, \tag{1.1.1}$$

where the primes denote differentiation with respect to x. The polynomial $H_{2n+1}(x)$ is called *Hermite's interpolation polynomial.* We have here $(2n + 2)$ conditions and therefore the number of coefficients to be determined is $(2n+2)$ and the degree of the polynomial is $(2n+1)$. In analogy with the Lagrange interpolation formula, we seek a representation of the form

$$H_{2n+1}(x) = \sum_{i=0}^{n} u_i(x) y_i + \sum_{i=0}^{n} v_i(x) y_i', \tag{1.1.2}$$

where $u_i(x)$ and $v_i(x)$ are polynomials in x of degree $(2n + 1)$. Using conditions (1.1.1), we obtain

$$u_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \ne j \end{cases}; \quad v_i(x) = 0, \ \text{for all } \ i$$

$$u_i'(x) = 0, \ \text{for all } i\ ; \qquad v_i'(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \ne j \end{cases} \tag{1.1.3}$$

Since $u_i(x)$ and $v_i(x)$ are polynomials in x of degree $(2n + 1)$, we write

$$u_i(x) = A_i(x)\,[l_i(x)]^2 \quad \text{and} \quad v_i(x) = B_i(x)\,[l_i(x)]^2 \tag{1.1.4}$$

where $l_i(x)$ are given by

$$l_i(x) = \frac{\Pi_{n+1}(x)}{(x - x_i)\Pi'_{n+1}(x_i)} \tag{1.1.5}$$

where

$$\Pi_{n+1}(x) = (x - x_0)(x - x_1)\dots(x - x_{i-1})(x - x_i)(x - x_{i+1})\dots(x - x_n)$$
$$\text{and} \qquad \Pi'_{n+1}(x_i) = (x_i - x_0)(x_i - x_1)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)$$

It is easy to see that $A_i(x)$ and $B_i(x)$ are both linear functions in x. We therefore write

$$u_i(x) = (a_i x + b_i)\,[l_i(x)]^2 \quad \text{and} \quad v_i(x) = (c_i x + d_i)\,[l_i(x)]^2 \tag{1.1.6}$$

Using conditions Eq.(1.1.3) in (1.1.6), we obtain

$$a_i x_i + b_i = 1, \quad c_i x_i + d_i = 0, \quad a_i + 2l'_i(x_i) = 0, \quad c_i = 1 \tag{1.1.7}$$

From Eq.(1.1.7), we deduce

$$a_i = -2l'_i(x_i), \quad b_i = 1 + 2x_i l'_i(x_i), \quad c_i = 1, \quad d_i = -x_i \tag{1.1.8}$$

Hence Eq.(1.1.6) becomes

$$\begin{aligned} u_i(x) &= [-2xl'_i(x_i) + 1 + 2x_i l'_i(x_i)]\,[l_i(x)]^2 \\ &= [1 - 2(x - x_i)l'_i(x_i)]\,[l_i(x)]^2 \end{aligned} \tag{1.1.9}$$

and

$$v_i(x) = (x - x_i)\,[l_i(x)]^2 \tag{1.1.10}$$

Using the above expressions for $u_i(x)$ and $v_i(x)$ in Eq.(1.1.2), we obtain finally

$$H_{2n+1}(x) = \sum_{i=0}^{n} \left[1 - 2(x - x_i)l'_i(x_i)\right]\,[l_i(x)]^2\,y_i + \sum_{i=0}^{n}(x - x_i)\,[l_i(x)]^2\,y'_i, \tag{1.1.11}$$

which is the required *Hermite interpolation formula*.

The following example demonstrate the application of Hermite's formula.

Example 1.1. Find the third-order Hermite polynomial passing through the points $(x_i, y_i, y'_i),\ i = 0, 1.$

Solution : Putting $n = 1$ in Hermite's formula (1.1.11), we obtain

$$\begin{aligned} H_3(x) = \left[1 - 2(x - x_0)l'_0(x_0)\right]\,[l_0(x)]^2 y_0 + \left[1 - 2(x - x_1)l'_1(x_1)\right]\,[l_1(x)]^2 y_1 \\ +(x - x_0)[l_0(x)]^2 y'_0 + (x - x_1)[l_1(x)]^2 y'_1. \end{aligned} \tag{1.1.12}$$

Since

$$l_0(x) = \frac{x - x_1}{x_0 - x_1} = \frac{x_1 - x}{h_1} \quad \text{and} \quad l_1(x) = \frac{x - x_0}{x_1 - x_0} = \frac{x - x_0}{h_1},$$

where $h_1 = x_1 - x_0$. Hence

$$l'_0(x) = -\frac{1}{h_1} \quad \text{and} \quad l'_1(x) = \frac{1}{h_1}. \tag{1.1.13}$$

Then, Eq. (1.1.12) simplifies to

$$H_3(x) = \left[1 + \frac{2(x-x_0)}{h_1}\right]\frac{(x_1-x)^2}{h_1^2}y_0 + \left[1 + \frac{2(x_1-x)}{h_1}\right]\frac{(x-x_0)^2}{h_1^2}y_1 + (x-x_0)\frac{x_1-x)^2}{h_1^2}y_0' + (x-x_1)\frac{x-x_0)^2}{h_1^2}y_1' \tag{1.1.14}$$

which is the required Hermite formula.

Example 1.2. Determine the Hermite polynomial of degree 5, which fits the following data and hence find an approximate value of $\ln 2.7$.

x	$y = \ln x$	$y' = 1/x$
2.0	0.69315	0.5
2.5	0.91629	0.4000
3.0	1.09861	0.33333

Solution : The polynomials $l_i(x)$ are given by

$$l_0(x) = \frac{(x-2.5)(x-3.0)}{(-0.5)(-1.0)} = 2x^2 - 11x + 15.$$

Similarly, we find

$$l_1(x) = -(4x^2 - 20x + 24) \quad \text{and} \quad l_2(x) = 2x^2 - 9x + 10.$$

We therefore obtain

$$l_0'(x) = 4x - 11, \quad l_1'(x) = -8x + 20, \quad l_2'(x) = 4x - 9.$$

Hence

$$l_0'(x_0) = -3, \quad l_1'(x_1) = 0, \quad l_2'(x_2) = 3$$

Equations (1.1.9) and (1.1.10) gives

$$\begin{aligned} u_0(x) &= (6x-11)(2x^2-11x+15)^2, & v_0(x) &= (x-2)(2x^2-11x+15)^2, \\ u_1(x) &= (4x^2-20x+24)^2, & v_1(x) &= (x-2.5)(4x^2-20x+24)^2, \\ u_2(x) &= (19-6x)(2x^2-9x+10)^2, & v_2(x) &= (x-3)(2x^2-9x+10)^2, \end{aligned}$$

Substituting these expressions in Eq.(1.1.11), we obtain the required Hermite polynomial

$$\begin{aligned} H_5(x) = \quad & (6x-11)(2x^2-11x+15)^2(0.69315) + (4x^2-20x+24)(0.91629) \\ & +(19-6x)(2x^2-9x+10)^2(1.09861) + (x-2)(2x^2-11x+15)^2(0.5) \\ & +(x-2.5)(4x^2-20x+24)^2(0.4) + (x-3)(2x^2-9x+10)^2(0.33333). \end{aligned}$$

Putting $x = 2.7$ and simplifying, we obtain

$$\ln(2.7) \approx H_5(2.7) = 0.993252,$$

which is correct to six decimal places. It is worthwhile to note that this result is more accurate than that obtained by using the Lagrange interpolation formula.

1.2 Divided Differences

The Lagrange interpolation formula has the disadvantage that if another interpolation point were added, then the interpolation coefficients $l_i(x)$ will have to be recomputed. We therefore seek an interpolation polynomial which has the property that a polynomial of higher degree may be derived from it by simply adding new terms. Newton's general interpolation formula is one such formula and it employs what are called *divided differences*. It is our principal purpose in this subsection to define such differences and discuss certain of their properties to obtain the basic formula due to Newton.

Let $(x_0, y_0),\ (x_1, y_1), \ldots, (x_n, y_n)$ be the given $(n+1)$ points. Then the divided differences of order 1, 2, ..., n are defined by the relations:

$$\begin{aligned}
[x_0, x_1] &= \frac{y_1 - y_0}{x_1 - x_0}, \\
[x_0, x_1, x_2] &= \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0}, \\
&\vdots \\
[x_0, x_1, \ldots, x_n] &= \frac{[x_1, x_2, \ldots, x_n] - [x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}.
\end{aligned} \tag{1.2.1}$$

1.3 Newton's General Interpolation Formula

By definition, we have

$$[x, x_0] = \frac{y - y_0}{x - x_0},$$

so that

$$y = y_0 + (x - x_0)[x, x_0] \tag{1.3.1}$$

Again

$$[x, x_0, x_1] = \frac{[x, x_0] - [x_0, x_1]}{x - x_1}$$

which gives

$$[x, x_0] = [x_0, x_1] + (x - x_1)[x, x_0, x_1]$$

Substituting this value of $[x, x_0]$ in Eq.(1.3.1), we obtain

$$y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x, x_0, x_1] \tag{1.3.2}$$

But

$$[x, x_0, x_1, x_2] = \frac{[x, x_0, x_1] - [x_0, x_1, x_2]}{x - x_2},$$

and so

$$[x, x_0, x_1] = [x_0, x_1, x_2] + (x - x_2)[x, x_0, x_1, x_2] \tag{1.3.3}$$

Equation (1.3.2) now gives

$$\begin{aligned}
y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\
+(x - x_0)(x - x_1)(x - x_2)[x, x_0, x_1, x_2]
\end{aligned} \tag{1.3.4}$$

Proceeding in this way, we obtain

$$\begin{aligned} y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ +(x - x_0)(x - x_1)(x - x_2)[x_0, x_1, x_2, x_3] + \dots \\ +(x - x_0)(x - x_1)(x - x_2) \cdot (x - x_n)[x, x_0, x_1, \dots, x_n] \end{aligned} \tag{1.3.5}$$

This formula is called *Newton's general interpolation formula with divided differences,* the last term being the remainder term after $(n + 1)$ terms. Hence after generating the divided differences, interpolation can be carried out.

Example 1.3. Certain corresponding values of x and $\log_{10} x$ are $(300,\ 2.4771)$, $(304,\ 2.4829)$, $(305,\ 2.4843)$ $(307,\ 2.4871)$. Find $\log_{10} 301$.

Solution : The divided difference table is

x	$y = \log_{10} x$		
300	2.4771		
		0.00145	
304	2.4829		0.00001
		0.00140	
305	2.4843		0
		0.00140	
307	2.4871		

Hence, Eq.(1.3.5) gives

$$\log_{10} 301 = 2.4771 + 0.00145 + (-3)(-0.00001) = 2.4786$$

1.4 Interpolation by Iteration

Newton's general interpolation formula may be considered as one of a class methods which generate successively higher-order interpolation formulae. We now describe another method of this class, due to A.C. Aitken, which has the advantage of being very easily programmed for a digital computer.

Given the $(n + 1)$ points $(x_0, y_0),\ (x_1, y_1), \dots, (x_n, y_n)$, where the values of x need not necessarily be equally spaced, then to find the value of y corresponding to any given value of x we proceed iteratively as follows:

Obtain a first approximation to y by considering the first-two points only; then obtain its second approximation by considering the first-three points, and so on. We denote the different interpolation polynomials by $\Delta(x)$, with suitable subscripts, so that at the first stage of approximation, we have

$$\Delta_{01}(x) = y_0 + (x - x_0)[x_0, x_1] = \frac{1}{x_1 - x_0} \begin{vmatrix} y_0 & x_0 - x \\ y_1 & x_1 - x \end{vmatrix} \tag{1.4.1}$$

Similarly, we can form $\Delta_{02}(x),\ \Delta_{03}(x), \dots$ Next, we form Δ_{012} by considering the first-three points:

$$\Delta_{012}(x) = \frac{1}{x_2 - x_1} \begin{vmatrix} \Delta_{01}(x) & x_1 - x \\ \Delta_{02}(x) & x_2 - x \end{vmatrix} \tag{1.4.2}$$

Similarly, we obtain $\Delta_{013}(x), \Delta_{014}(x)$, etc. At the n-th stage of approximation, we obtain

$$\Delta_{0123\ldots n}(x) = \frac{1}{x_n - x_{n-1}} \begin{vmatrix} \Delta_{0123\ldots\overline{n-1}}(x) & x_{n-1} - x \\ \Delta_{0123\ldots\overline{n-2n}}(x) & x_n - x \end{vmatrix} \tag{1.4.3}$$

The computations may conveniently be arranged as in Table 1.1 below:

Table 1.1 Aitken's Scheme

x	y				
x_0	y_0				
		$\Delta_{01}(x)$			
x_1	y_1		$\Delta_{012}(x)$		
		$\Delta_{02}(x)$		$\Delta_{0123}(x)$	
x_2	y_2		$\Delta_{013}(x)$		$\Delta_{01234}(x)$
		$\Delta_{03}(x)$		$\Delta_{0124}(x)$	
x_3	y_3		$\Delta_{014}(x)$		
		$\Delta_{04}(x)$			
x_4	y_4				

A modification of this scheme, due to Neville, is given in Table 1.2. Neville's scheme is particularly suited for iterated inverse interpolation.

Table 1.2 Neville's Scheme

x	y				
x_0	y_0				
		$\Delta_{01}(x)$			
x_1	y_1		$\Delta_{012}(x)$		
		$\Delta_{12}(x)$		$\Delta_{0123}(x)$	
x_2	y_2		$\Delta_{123}(x)$		$\Delta_{01234}(x)$
		$\Delta_{23}(x)$		$\Delta_{1234}(x)$	
x_3	y_3		$\Delta_{234}(x)$		
		$\Delta_{34}(x)$			
x_4	y_4				

As an illustration of Aitken's method, we consider, again, Example (1.3).

Example 1.4. Aitken's scheme is

x	$\log_{10} x$			
300	2.4771			
		2.47855		
304	2.4829		2.47858	
		2.47854		2.47860
305	2.4843		2.47857	
		2.47853		
307	2.4871			

Hence $\log_{10} 301 = 2.4786$, as before.

An obvious advantage of AItken's method is that *gives a good idea of the accuracy of the result at any stage.*

Exercise 1.5. (i) Using Hermite's interpolation formula, estimate the value of $\ln 4.2$ from the data (value of x, $\ln x$ and $\frac{1}{x}$):

$$(4.0,\ 1.38629,\ 0.25000),\ (4.5,\ 1.50408,\ 0.22222),\ (5.0,\ 1.60944,\ 0.20000)$$

Answer: 1.435081

(ii) Find the Hermite polynomial of the third degree approximating the function $y(x)$ such that

$$y(0) = 1, \quad y'(0) = 0$$
$$y(1) = 3, \quad y'(1) = 5.$$

Answer: $1 + x^2 + x^3$

(iii) Given $f(x) = \frac{1}{x^2}$. Find the divided differences $[a, b]$, and $[a, b, c]$. **Answer:** $-\dfrac{a+b}{a^2b^2}, \dfrac{ab+bc+ca}{a^2b^2c^2}$

(iv) Given the set of tabulated points $(0, 2),\ (1, 3),\ (2, 12)$ and $(15, 3587)$ satisfying the function $y = f(x)$, compute $f(4)$ using Newton's divided difference formula. **Answer:** 1454

Unit 2

Course Structure

Approximation of Function : Least square approximation. Weighted least square approximation. Orthogonal polynomials, Gram – Schmidt orthogonalisation process, Chebysev polynomials, Minimax polynomial approximation.

---

2 Introduction

In experimental work, we often encounter the problem of fitting a curve to data which are subject to errors. The strategy for such cases is to derive an approximating function that *broadly* fits the data without necessarily passing through the given points. The curve drawn is such that the discrepancy between the data points and the curve is least. In the method of least squares, the sum of the squares of the errors is minimized. The problem of approximating a function by means of Chebyshev polynomials is described in this unit.

2.1 Least Squares Curve Fitting Procedures

Let the set of data points be $(x_i, y_i),\ i = 1, 2, \ldots, m$, and let the curve given by $Y = f(x)$ be fitted to this data. At $x = x_i$, the given ordinate is y_i and the corresponding value on the fitting curve is $f(x_i)$. If e_i is the error of approximation at $x = x_i$, then we have

$$e_i = y_i - f(x_i) \tag{2.1.1}$$

If we write

$$\begin{aligned} S &= [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \ldots + [y_m - f(x_m)]^2 \\ &= e_1^2 + e_2^2 + \ldots + e_m^2, \end{aligned} \tag{2.1.2}$$

then the method of least squares consists in minimizing S, i.e., the sum of the squares of the errors. In the following subsection, we shall study the linear least squares fitting to given data $(x_i, y_i),\ i = 1, 2, \ldots, m$.

2.1.1 Fitting a Straight Line

Let $Y = a_0 + a_1 x$ be the straight line to be fitted to the given data, viz. $(x_i, y_i), i = 1, 2, \ldots, m$. Then, corresponding to Eq.(2.1.2), we have

$$S = [y_1 - (a_0 + a_1 x)]^2 + [y_2 - (a_0 + a_1 x)]^2 + \ldots + [y_m - (a_0 + a_1 x_m)]^2 \tag{2.1.3}$$

For S to be minimum, we have

$$\frac{\partial S}{\partial a_0} = 0 = -2[y_1 - (a_0 + a_1 x)] - 2[y_2 - (a_0 + a_1 x_2)] - \ldots - 2[y_m - (a_0 + a_1 x_m)]$$

$$\frac{\partial S}{\partial a_1} = 0 = -2x_1[y_1 - (a_0 + a_1 x)] - 2x_2[y_2 - (a_0 + a_1 x_2)] - \ldots - 2x_m[y_m - (a_0 + a_1 x_m)]$$

The above equations simplify to

$$\begin{aligned} ma_0 + a_1(x_1 + x_2 + \ldots + x_m) &= y_1 + y_2 + \ldots + y_m \\ \text{and } a_0(x_1 + x_2 + \ldots + x_m) + a_1(x_1^2 + x_2^2 + \ldots + x_m^2) &= x_1y_1 + x_2y_2 + \ldots + x_my_m \end{aligned} \quad (2.1.4)$$

or more compactly to

$$ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \quad \text{and} \quad a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i \quad (2.1.5)$$

Equations (2.1.5) are called the *normal equations*, and can be solved for a_0 and a_1, since x_i and y_i are known quantities. We can obtain easily

$$a_1 = \frac{m \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \cdot \sum_{i=1}^{m} y_i}{m \sum x_i^2 - \left(\sum_{i=1}^{m} x_i \right)^2} \quad (2.1.6)$$

and then

$$a_0 = \overline{y} - a_1 \overline{x}. \quad (2.1.7)$$

Since $\dfrac{\partial^2 S}{\partial a_0^2}$ and $\dfrac{\partial^2 S}{\partial a_1^2}$ are both positive at the points a_0 and a_1, it follows that these values provide a *minimum* of S. In Eq.(2.1.7), $\overline{x}$ and $\overline{y}$ are the means of x and y, respectively. Form Eq.(2.1.7), we have

$$\overline{y} = a_0 + a_1 \overline{x},$$

which shows the fitted straight line passes through the centroid of the data points. Sometimes, a goodness of fit is adopted. The correlation coefficient (cc) is defined as

$$cc = \sqrt{\frac{S_y - S}{S_y}}, \qquad \text{where} \qquad S_y = \sum_{i=1}^{m} (y_i - \overline{y})^2 \quad \text{and} \quad S \text{ is defined by Eq.(2.1.3)} \quad (2.1.8)$$

If cc is close to 1, then the fit is considered to be good, although this is not always true.

Example 2.1. Find the best values of a_0 and a_1 if the straight line $Y = a_0 + a_1 x$ is fitted to the data (x_i, y_i):

(1 , 0.6), (2 , 2.4),(3 , 3.5), (4 , 4.8), (5 , 5.7)

Solution:

x_i	y_i	x_i^2	$x_i y_i$	$(y_i - \overline{y})^2$	$(y_i - a_0 - a_1 x_i)^2$
1	0.6	1	0.6	7.84	0.0784
2	2.4	4	4.8	1.00	0.0676
3	3.5	9	10.5	0.01	0.0100
4	4.8	16	19.2	1.96	0.0196
5	5.7	25	28.5	5.29	0.0484
15	17.0	55	63.6	16.10	0.2240

From the given table of values, we find $\overline{x} = 3, \overline{y} = 3.4$, and

$$a_1 = \frac{5(63.6) - (15)(17)}{5(55) - 225} = 1.26 \quad \text{and} \quad a_0 = \overline{y} - a_1 \overline{x} = -0.38$$

The correlation coefficient = $\sqrt{\dfrac{16.10 - 0.2240}{16.10}} = 0.9930$

Exercise 2.2. (i) Certain experimental values of x and y are given below:

(0 , -1), (2 , 5), (5 , 12), (7 , 20)

If the straight line $Y = a_0 + a_1 x$ is fitted to the above data, find the approximate values of a_0 and a_1.
Answer: $a_0 = -1.1381$ and $a_1 = 2.8966$

2.2 Curve Fitting by Polynomials

Let the polynomial of the n-th degree

$$Y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{2.2.1}$$

be fitted to the data points $(x_i, y_i),\ i = 1, 2, \ldots, m$. We then have

$$\begin{aligned} S &= \left[y_1 - (a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n)\right]^2 + \left[y_2 - (a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_n x_2^n)\right]^2 \\ &\quad + \ldots\ldots + \left[y_m - (a_0 + a_1 x_m + a_2 x_m^2 + \ldots + a_n x_m^n)\right]^2 \end{aligned} \tag{2.2.2}$$

Equating to zero the first partial derivatives and simplifying, we obtain the normal equations:

$$\begin{aligned} m a_0 + a_1 \sum x_i + a_2 \sum x_i^2 + \ldots a_n \sum x_i^n &= \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 + \ldots a_n \sum x_i^{n+1} &= \sum x_i y_i, \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots & \\ a_0 \sum x_i^n + a_1 \sum x_i^{n+1} + \ldots a_n \sum x_i^{2n} &= \sum x_i^n y_i, \end{aligned} \tag{2.2.3}$$

where the summations are performed from $i = 1$ to $i = m$. The system (2.2.3) constitutes $(n + 1)$ equations in $(n + 1)$ unknowns, and hence can be solved for $a_0,\ a_1,\ \ldots,\ a_n$. Equation (2.2.1) then gives the required polynomial of the n-th degree.

For larger values of n, system (2.2.3) becomes unstable with the result that round off errors in the data may cause large changes in the solution. Such systems occur quite often in practical problems and are called *ill conditioned* system. Orthogonal polynomials are most suited to solve such systems and one particular form of these polynomials, the Chebyshev polynomial, will be discussed later in this unit.

Example 2.3. Fit a polynomial of the second degree to the data points (x, y) given by

(0, 1), (1, 6), and (2, 17)

Solution: For $n = 2$, Eq.(2.2.3) requires $\sum x_i, \sum x_i^2, \sum x_i^3, \sum x_i^4, \sum y_i, \sum x_i y_i$ and $\sum x_i^2 y_i$. The table of values is as follows:

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	6	1	1	1	6	6
2	17	4	8	16	34	68
3	24	5	9	17	40	74

The normal equations are

$$\begin{aligned} 3a_0 + 3a_1 + 5a_2 &= 24 \\ 3a_0 + 5a_1 + 9a_2 &= 40 \\ 5a_0 + 9a_1 + 17a_2 &= 74 \end{aligned}$$

Solving the above system, we obtain

$$a_0 = 1, \;\; a_1 = 2 \;\text{ and }\; a_2 = 3.$$

The required polynomial is given by $Y = 1 + 2x + 3x^2$, and it can be seen that this fitting is *exact*.

Exercise 2.4. (i) Fit a second degree parabola $y = a_0 + a_1 x + a_2 x^2$ to the data (x_i, y_i):

(1 , 0.63), (3 , 2.05),(4 , 4.08), (6 , 10.78)

Answer: $a_0 = 1.24$, $a_1 = -1.05$ and $a_2 = 0.44$

2.3 Weighted Least Square Approximation

In the previous subsection, we have minimized the sum of squares of the errors. A more general approach is to minimize the weighted sum of the squares of the errors taken over all data points. If this sum is denoted by S, then instead of Eq.(2.1.2), we have

$$\begin{aligned} S &= W_1\Big[y_1 - f(x_1)\Big]^2 + W_2\Big[y_2 - f(x_2)\Big]^2 + \ldots + W_m\Big[y_m - f(x_m)\Big]^2 \\ &= W_1 e_1^2 + W_2 e_2^2 + \ldots + W_m e_m^2. \end{aligned} \tag{2.3.1}$$

In Eq.(2.3.1), the W_i are prescribed positive numbers and are called *weights*. A weight is prescribed according to the relative accuracy of a data points. If all the data points are accurate, we set $W_i = 1$ for all i. We consider again the linear and non-linear cases below.

2.3.1 Linear Weighted Least Squares Approximation

Let $Y = a_0 + a_1 x$ be the straight line to be fitted to the given data points, viz. $(x_1,\ y_1), \ldots, (x_m, y_m)$. Then

$$S(a_0, a_1) = \sum_{i=1}^{m} W_i\Big[y_i - (a_0 + a_1 x_i)\Big]^2. \tag{2.3.2}$$

For maxima or minima, we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = 0, \quad \text{which gives} \tag{2.3.3}$$

$$\frac{\partial S}{\partial a_0} = -2\sum_{i=1}^{m} W_i\Big[y_i - (a_0 + a_1 x_i)\Big] = 0 \quad \text{and} \quad \frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{m} W_i\Big[y_i - (a_0 + a_1 x_i)\Big]x_i = 0.$$

Simplifying yields the system of equations for a_0 and a_1:

$$a_0\sum_{i=1}^{m} W_i + a_1\sum_{i=1}^{m} W_i x_i = \sum_{i=1}^{m} W_i y_i \quad \text{and} \quad a_0\sum_{i=1}^{m} W_i x_i + a_1\sum_{i=1}^{m} W_i x_i^2 = \sum_{i=1}^{m} W_i x_i y_i \tag{2.3.4}$$

which are the *normal equations* in this case and are solved to obtain a_0 and a_1.

Example 2.5. Suppose that in the data of Exercise (2.2), the point $(5, 12)$ is known to be more reliable than the others. Then we prescribe a weight (say, 10) corresponding to this point only and all other weights are taken as unity. Find the new 'linear least squares approximation'.

Solution: Let us calculate the following table.

x	y	W	Wx	Wx^2	Wy	Wxy
0	−1	1	0	0	−1	0
2	5	1	2	4	5	10
5	12	10	50	250	120	600
7	20	1	7	49	20	140
14	36	13	59	303	144	750

The normal Eqs.(2.3.4) then give

$$13a_0 + 59a_1 = 144 \quad \text{and} \quad 59a_0 + 303a_1 = 750$$

Solving the above equations, we obtain

$$a_0 = -1.349345 \quad \text{and} \quad a_1 = 2.73799$$

The 'linear least squares approximation' is, therefore, given by

$$y = -1.349345 + 2.73799x$$

Exercise 2.6. (i) Consider Example (2.5) again with an increased weight, say 100, corresponding to $y(5.0)$ and calculate the new 'linear least squares approximation' and comment the influence of increasing weight to the approximation. **Answer:** $y = -1.41258 + 2.69056x$

2.3.2 Nonlinear Weighted Least Squares Approximation

We now consider the least squares approximation of a set of m data points $(x_i, y_i), i = 1, 2, \ldots, m$, by a polynomial of degree $n < m$. Let

$$y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n \tag{2.3.5}$$

be fitted to the given data points. We then have

$$S(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{m} W_i \Big[y_i - (a_0 + a_1x_i + \ldots + a_nx_i^n) \Big]^2. \tag{2.3.6}$$

If a minimum occurs at $(a_0, a_1, \ldots, a_n)$, then we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \frac{\partial S}{\partial a_2} = \ldots = \frac{\partial S}{\partial a_n} = 0. \tag{2.3.7}$$

These conditions yield the normal equations

$$\begin{aligned}
a_0 \sum_{i=1}^{m} W_i + a_1 \sum_{i=1}^{m} W_ix_i + \ldots + a_n \sum_{i=1}^{m} W_ix_i^n &= \sum_{i=1}^{m} W_iy_i \\
a_0 \sum_{i=1}^{m} W_ix_i + a_1 \sum_{i=1}^{m} W_ix_i^2 + \ldots + a_n \sum_{i=1}^{m} W_ix_i^{n+1} &= \sum_{i=1}^{m} W_ix_iy_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad &\vdots \\
a_0 \sum_{i=1}^{m} W_ix_i^n + a_1 \sum_{i=1}^{m} W_ix_i^{n+1} + \ldots + a_n \sum_{i=1}^{m} W_ix_i^{2n} &= \sum_{i=1}^{m} W_ix_i^ny_i.
\end{aligned} \tag{2.3.8}$$

Equations (2.3.8) are $(n+1)$ equations in $(n+1)$ unknowns $a_0, a_1, \ldots, a_n$. If the x_i are distinct with $n < m$, then the equations possess a 'unique' solution.

2.4 Orthogonal Polynomial approximation method

In the previous subsection, we considered the least squares approximations of discrete data. We shall, in the present subsection, discuss the least squares approximation of a continuous function on $[a, b]$. In this case, the summations in the normal equations are now replaced by definite integrals. However, this method possesses the disadvantage of solving a large linear system of equations. Besides, such a system may exhibit a peculiar tendency called *ill-conditioning*, which means that small change in any of its parameters introduces large errors in the solution - the degree of *ill-conditioning* increasing with the order of the system. Hence, alternative methods of solving the continuous function for least squares problem have gained importance, and of these the method that employs *'orthogonal polynomial'* is currently in use. This method possess the great advantage that it does not require a linear system to be solved and is described below.

We choose the approximation in the form:

$$Y(x) = a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x), \tag{2.4.1}$$

where $f_j(x)$ is a polynomial in x of degree j. Then we write

$$S(a_0, a_1, \ldots, a_n) = \int_0^a W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big]^2 dx. \tag{2.4.2}$$

For S to be minimum, we must have

$$\begin{aligned}
\frac{\partial S}{\partial a_0} &= 0 = -2\int_a^b W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big] f_0(x)\, dx \\
\frac{\partial S}{\partial a_1} &= 0 = -2\int_a^b W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big] f_1(x)\, dx \\
&\vdots \\
\frac{\partial S}{\partial a_n} &= 0 = -2\int_a^b W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big] f_n(x)\, dx
\end{aligned} \tag{2.4.3}$$

The system of normal equations can be written as

$$\begin{aligned}
&a_0 \int_a^b W(x) f_0(x) f_j(x)\, dx \;+\; a_1 \int_a^b W(x) f_1(x) f_j(x)\, dx \;+\; \ldots\ldots \\
&+ a_n \int_a^b W(x) f_n(x) f_j(x)\, dx = \int_a^b W(x) y(x) f_j(x)\, dx, \quad j = 0, 1, 2, \ldots, n.
\end{aligned} \tag{2.4.4}$$

In Eq.(2.4.4), we find products of the type $f_p(x) f_q(x)$ in the integrands, and if we assume that

$$\int_a^b W(x) f_p(x) f_q(x)\, dx = \begin{cases} 0, & p \neq q \\ \int_a^b W(x) f_p^2(x)\, dx, & p = q, \end{cases} \tag{2.4.5}$$

Hence from Eq.(2.4.4), we obtain

$$a_j = \left[\int_a^b W(x)y(x)f_j(x)\,dx\right] \Big/ \left[\int_a^b W(x)f_j^2(x)\,dx\right], \quad j = 0, 1, 2, \ldots, n. \tag{2.4.6}$$

Substitution of $a_0, a_1, \ldots, a_n$ in Eq.(2.4.1) then yields the required least squares approximation, but the functions $f_0(x), f_1(x), \ldots, f_n(x)$ are still not known. The $f_j(x)$, which are polynomials in x satisfying the condition (2.4.5), are called *orthogonal polynomials* and are said to be orthogonal with respect to the weight function $W(x)$. They play an important role in numerical analysis and a few of them are listed below.

Name	$f_j(x)$	*Interval*	$W(x)$
Jacobi	$P_n^{(\alpha,\beta)}(x)$	$[-1, 1]$	$(1-x)^\alpha(1+x)^\beta(\alpha, \beta > -1)$
Chebyshev (first kind)	$T_n(x)$	$[-1, 1]$	$(1-x^2)^{-1/2}$
Chebyshev (second kind)	$U_n(x)$	$[-1, 1]$	$(1-x^2)^{1/2}$
Legendre	$P_n(x)$	$(-1, 1)$	1
Laguerre	$L_n(x)$	$[0, \infty)$	e^{-x}
Hermite	$H_n(x)$	$(-\infty, \infty)$	e^{-x^2}

A brief discussion of some important properties of the Chebyshev polynomials $T_n(x)$ and their usefulness in the approximation of functions will be given in a later subsection in this unit. We now return to our discussion of the problem of determining the least squares approximation. As we noted earlier, the function $f_j(x)$ are yet to be determined. These are obtained by using 'Gram-Schmidt orthogonalization process', which has important applications in numerical analysis.

2.4.1 Gram-Schmidt Orthogonalization Process

Suppose that the orthogonal polynomial $f_i(x)$, valid on the interval $[a, b]$, has the leading term x^i. Then, starting with

$$f_0(x) = 1 \tag{2.4.7}$$

we find that the linear polynomial $f_1(x)$, with leading term x, can be written as

$$f_1(x) = x + k_{1,0}f_0(x), \tag{2.4.8}$$

where $k_{1,0}$ is a constant to be determined. Since $f_1(x)$ and $f_0(x)$ are orthogonal, we have

$$\int_a^b W(x)f_0(x)f_1(x)\,dx = 0 = \int_a^b xW(x)f_0(x)\,dx + k_{1,0}\int_a^b W(x)f_0^2(x)\,dx \quad \text{[using Eqs.(2.4.5) and (2.4.7)]}$$

Now from the above, we obtain

$$k_{1,0} = -\left[\int_a^b xW(x)f_0(x)\,dx\right] \Big/ \left[\int_a^b W(x)f_0^2(x)\,dx\right], \tag{2.4.9}$$

and Eq.(2.4.8) gives

$$f_1(x) = x - \left[\left[\int_a^b xW(x)f_0(x)\,dx\right] \Big/ \left[\int_a^b W(x)f_0^2(x)\,dx\right]\right] f_0(x) \tag{2.4.10}$$

Now, the polynomial $f_2(x)$, of degree 2 in x and with leading term x^2, may be written as

$$f_2(x) = x^2 + k_{2,0}f_0(x) + k_{2,1}f_1(x), \tag{2.4.11}$$

where the constants $k_{0,2}$ and $k_{2,1}$ are to be determined by using the orthogonality conditions in Eq.(2.4.5). Since $f_2(x)$ is orthogonal to $f_0(x)$, we have

$$\int_a^b W(x)f_0(x)\Big[x^2 + k_{2,0}f_0(x) + k_{2,1}f_1(x)\Big]dx = 0. \tag{2.4.12}$$

Since $\int_a^b W(x)f_0(x)f_1(x)\,dx = 0$, the above equation gives

$$k_{2,0} = -\left[\int_a^b x^2W(x)f_0(x)\,dx\right]\Big/\left[\int_a^b W(x)f_0^2(x)\,dx\right] = -\left[\int_a^b x^2W(x)\,dx\right]\Big/\left[\int_a^b W(x)\,dx\right], \tag{2.4.13}$$

Again, since $f_2(x)$ is orthogonal to $f_1(x)$, we have

$$\int_a^b W(x)f_1(x)\Big[x^2 + k_{2,0}f_0(x) + k_{2,1}f_1(x)\Big]dx = 0. \tag{2.4.14}$$

Using the condition that $\int_a^b W(x)f_0(x)f_1(x)\,dx = 0$, the above yields

$$k_{2,1} = -\left[\int_a^b x^2W(x)f_1(x)\,dx\right]\Big/\left[\int_a^b W(x)f_1^2(x)\,dx\right], \tag{2.4.15}$$

Since $k_{2,0}$ and $k_{2,1}$ are known, Eq.(2.4.11) determines $f_2(x)$. Proceeding in this way, the method can be generalized and we write

$$f_j(x) = x^j + k_{j,0}f_0(x) + k_{j,1}f_1(x) + \ldots + k_{j,j-1}f_{j-1}(x), \tag{2.4.16}$$

where the constants $k_{j,i}$ are so chosen that $f_j(x)$ is orthogonal to $f_0(x), f_1(x), \ldots, f_{j-1}(x)$. These conditions yield

$$k_{j,i} = -\left[\int_a^b x^jW(x)f_i(x)\,dx\right]\Big/\left[\int_a^b W(x)f_i^2(x)\,dx\right], \tag{2.4.17}$$

Since the a_i and $f_i(x)$ in Eq.(2.4.1) are known, the approximation $Y(x)$ can now be determined. The following example illustrates the method of procedure.

Example 2.7. Obtain the first-four orthogonal polynomials $f_n(x)$ on $[-1, 1]$ with respect to the weight function $W(x) = 1$.

Solution: Let $f_0(x) = 1$. Then Eq.(2.4.9) gives

$$k_{1,0} = -\left[\int_{-1}^{1} x\, dx\right] \Big/ \left[\int_{-1}^{1} dx\right] = 0,$$

We then obtain from Eq.(2.4.8), $f_1(x) = x$. Equations (2.4.13) and (2.4.15) gives respectively

$$k_{2,0} = -\left[\int_{-1}^{1} x^2\, dx\right] \Big/ \left[\int_{-1}^{1} dx\right] = -\frac{1}{3} \quad \text{and} \quad k_{2,1} = -\left[\int_{-1}^{1} x^2 x\, dx\right] \Big/ \left[\int_{-1}^{1} x^2\, dx\right] = 0.$$

Then Eq.(2.4.11) yields $f_2(x) = x^2 - 1/3$. In a similar manner, we obtain

$$k_{3,0} = -\left[\int_{-1}^{1} x^3\, dx\right] \Big/ \left[\int_{-1}^{1} dx\right] = 0, \quad k_{3,1} = -\left[\int_{-1}^{1} x^3\, x\, dx\right] \Big/ \left[\int_{-1}^{1} x^2\, dx\right] = -\frac{3}{5},$$

$$\text{and} \quad k_{3,2} = -\left[\int_{-1}^{1} x^3(x^2 - 1/3)\, dx\right] \Big/ \left[\int_{-1}^{1} (x^2 - 1/3)^2\, dx\right] = 0,$$

It is easily verified that

$$f_3(x) = x^3 - \frac{3}{5}x.$$

Thus the required orthogonal polynomials are $1, x, x^2 - 1/3$ and $x^3 - (3/5)x$. These polynomials are called *Legendre polynomials* and are usually denoted by $P_n(x)$. It is easy to verify that these polynomials satisfy the orthogonal property given in Eq.(2.4.5).

2.5 Chebyshev Polynomials

The chebyshev polynomial of degree n over the interval $[-1, 1]$ is defined by the relation

$$T_n(x) = \cos(n \cos^{-1} x), \tag{2.5.1}$$

from which follows immediately the relation

$$T_n(x) = T_{-n}(x). \tag{2.5.2}$$

Let $\cos^{-1} x = \theta$ so that $x = \cos\theta$ and Eq.(2.5.1) gives

$$T_n(x) = \cos n\theta. \tag{2.5.3}$$

Hence $T_0(x) = 1$ and $T_1(x) = x$. Using the trigonometric identity

$$\cos(n-1)\theta + \cos(n+1)\theta = 2\cos n\theta \cos\theta, \tag{2.5.4}$$

we obtain easily

$$T_{n-1}(x) + T_{n+1}(x) = 2xT_n(x), \tag{2.5.5}$$

which is the same as

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{2.5.6}$$

This is the *recurrence relation* which can be used to successively compute all $T_n(x)$, since we know $T_0(x)$ and $T_1(x)$. The first seven Chebyshev polynomials are:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x,$$
$$T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \quad T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$$

The graph of the first four Chebyshev polynomials are shown in Fig.2.1.

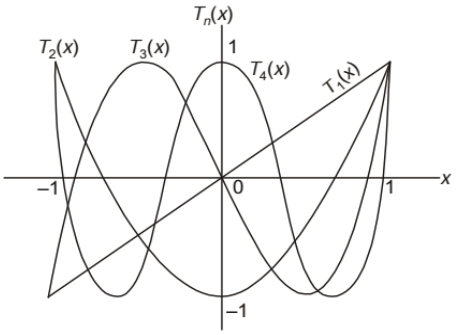


Fig.2.1 Chebyshev polynomials $T_n(x), n = 1, 2, 3, 4$.

It is easy to see that the coefficient of x^n in $T_n(x)$ is always 2^{n-1}. Further, if we set $y = T_n(x) = \cos n\theta$, then we get

$$\frac{dy}{dx} = \frac{n \sin n\theta}{\sin \theta}$$

and

$$\frac{d^y}{dx^2} = \frac{-n^2 \cos n\theta + n \ \sin n\theta \cot \theta}{\sin^2 \theta} = \frac{-n^2 y + x \ (dy/dx)}{1 - x^2} \tag{2.5.7}$$

so that

$$(1 - x^2)\frac{d^y}{dx^2} - x\frac{dy}{dx} + n^2 y = 0, \tag{2.5.8}$$

which is the *differential equation satisfied* by $T_n(x)$. It is also possible to express powers of x in terms of Chebyshev polynomials. We find

$$\begin{aligned} &1 = T_0(x), \quad x = T_1(x), \quad x^2 = \frac{1}{2}[T_0(x) + T_2(x)], \quad x^3 = \frac{1}{4}[3T_1(x) + T_3(x)] \\ &x^4 = \frac{1}{8}[3T_0(x) + 4T_2(x) + T_4(x)], \quad x^5 = \frac{1}{16}[10T_1(x) + 5T_3(x) + T_5(x)], \\ &x^6 = \frac{1}{32}[10T_0(x) + 15T_2(x) + 6T_4(x) + T_6(x)]. \end{aligned} \tag{2.5.9}$$

and so on. These expressions will be useful in the economization of power series which is beyond of our syllabus. An important property of $T_n(x)$ is given by

$$\int_{-1}^{1} \frac{T_m(x)T_n(x) \ dx}{\sqrt{1 - x^2}} = \begin{cases} 0, & m \neq n \\ \pi/2, & m = n \neq 0, \\ \pi, & m = n = 0 \end{cases} \tag{2.5.10}$$

that is, the polynomials $T_n(x)$ are *orthogonal* with the function $1/\sqrt{1 - x^2}$. This property is easily proved since by putting $x = \cos\theta$, the above integral becomes

$$\int_0^{\pi} T_m(\cos\theta)T_n(\cos\theta) \ d\theta = \int_0^{\pi} \cos m\theta \ \cos n\theta \ d\theta = \left[\frac{\sin(m+n)\theta}{2(m+n)} + \frac{\sin(m-n)\theta}{2(m-n)}\right]_0^{\pi},$$

from which follow the values given on the right side of Eq.(2.5.10). We have seen above that $T_n(x)$ is a polynomial of degree n in x and that the coefficient of x^n in $T_n(x)$ is 2^{n-1}. In approximation theory, one use *monic* polynomials, i.e., Chebyshev polynomials in which the coefficient of x^n is unity. If $P_n(x)$ is a monic polynomial, then we can write

$$P_n(x) = 2^{1-n}T_n(x), \quad (n \geq 1). \tag{2.5.11}$$

A remarkable property of Chebyshev polynomials is that *of all monic polynomials, $P_n(x)$, of degree n whose leading coefficient equals unity, the polynomials $2^{1-n}T_n(x)$, has the smallest least upper bound for its absolute value in the range $(-1, 1)$.* Since $|T_n(x)| \leq 1$, the upper bound referred to above is 2^{1-n}. Thus, in Chebyshev approximation, the maximum error is kept down to a minimum. This is often referred to as *minimax principle* and the polynomial in Eq.(2.5.11) is called the *minimax polynomial*. By this process we can obtain the best lower-bound approximation, called the *minimax approximation,* to a given polynomial. This is illustrated in the following example.

Example 2.8. Find the best lower-order approximation to the cubic $2x^3 + 3x^2$.

Solution: Using the relation given in Eq.(2.5.9), we write

$$\begin{aligned} 2x^3 + 3x^2 &= \frac{2}{4}\Big[T_3(x) + 3T_1(x)\Big]^2 + 3x^2 \\ &= 3x^2 + \frac{3}{2}T_1(x) + \frac{1}{2}T_3(x) \\ &= 3x^2 + \frac{3}{2}x + \frac{1}{2}T_3(x), \quad [\because T_1(x) = x]. \end{aligned}$$

The polynomial $3x^2 + (3/2)x$ is the required lower-order approximation to the given cubic with a maximum error $\pm 1/2$ in the range $(-1, 1)$.

Exercise 2.9. (i) If the function $f_1(x) = 1, f_2(x) = x$ are orthogonal on the interval $[-1, 1]$, find the values of a and b so that the function $f_3(x) = 1 + ax + bx^2$ is orthogonal to both f_1 and f_2 on $[-1, 1]$.
(ii) Define an orthogonal set of functions and show that the set $f(x) = \sin\dfrac{n\pi x}{l}$, $n = 1, 2, \ldots$ is orthogonal on $[0, l]$.

Unit 3

Course Structure

Numerical Integration : Gaussian quadrature formula and its existence, Newton's quadrature formula, Romberg integration, Euler- MacLaurin formula.

---

3 Introduction

The general problem of numerical integration may be stated as: Given a set of data points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$ of a function $y = f(x)$, where $f(x)$ is not known explicitly, it is required to compute the value of the definite integral

$$I = \int_a^b y \, dx. \tag{3.0.1}$$

Different integration formulae can be obtained depending upon the type of interpolation formula used.

3.1 Gauss quadrature formula

In Newton's cotes formula for numerical integration we used ordinate in equi-distant point. Gauss observed that if these requirement is removed then the degree of precision can be highly increased. But here it is require that $f(x)$ should be explicitly known so that it can be evaluated at any desired value of x.

3.1.1 Derivation

We first consider a function $y = f(t)$ specified on $[-1, 1]$, we shall show that

$$\int_{-1}^{1} f(t) \, dt = \sum_{i=0}^{n} A_i f(t_i) + R \tag{3.1.1}$$

It is possible to choose the points $t_i (i = 0, 1, 2, \ldots, n)$ and the coefficients $A_i, \ i = 0(1)n$ so that $R = 0$ for $f(t)$, any polynomial of degree $\leq (2n + 1)$. Let $P(t)$ be the interpolating polynomial of degree $\leq (2n + 1)$ which coincides with $f(t)$ at $t = t_0, t_1, \ldots, t_n, t_{n+1}, \ldots, t_{2n+1}$. Then

$$f(t) = P(t) + (t - t_0)(t - t_1) \ldots (t - t_{2n+1}) f(t, t_0, t_1, \ldots, t_{2n+1}) \tag{3.1.2}$$

$$\therefore \int_{-1}^{1} f(t) \, dt = \int_{-1}^{1} P(t) \, dt + R, \text{ where } R = \int_{-1}^{1} (t - t_0)(t - t_1) \ldots (t - t_{2n+1}) f(t, t_0, \ldots, t_{2n+1}) \, dt \tag{3.1.3}$$

Let $L_n(t)$ be the Lagrange's interpolation polynomial of degree n which coincides with $f(t)$ at $t_0, t_1, t_2, \ldots, t_n$ in the interval $[-1, 1]$. Then

$$L_n(t) = \sum_{i=0}^{n} \frac{w(t)}{(t - t_i) w'(t_i)} f(t_i) \tag{3.1.4}$$

where $w(t) = (t - t_0)(t - t_1)\ldots(t - t_n)$. Now $P(t_r) = L_n(t_r),\ r = 0, 1, 2, \ldots, n$. Therefore,

$$P(t) - L_n(t) = c(t - t_0)(t - t_1)\ldots(t - t_n)N(t) \tag{3.1.5}$$

where c is a constant and $N(t)$ is a polynomial of degree n. Therefore

$$P(t) = L_n(t) + c\, w(t)\, N(t) \tag{3.1.6}$$

Using (3.1.6) in (3.1.3), we obtain

$$\begin{aligned} \int_{-1}^{1} f(t)\, dt &= \int_{-1}^{1} L_n(t)\, dt + c\int_{-1}^{1} w(t)\, N(t)\, dt + R \\ &= \sum_{i=0}^{n} A_i f(t_i) + c\int_{-1}^{1} w(t)\, N(t)\, dt + R \end{aligned} \tag{3.1.7}$$

where

$$A_i = \int_{-1}^{1} \frac{w(t)}{(t - t_i)w'(t_i)} dt \tag{3.1.8}$$

Now we see that for a proper choice of the points $t_0, t_1, t_2, \ldots, t_n$ the degree of precision in $(2n + 1)$ if

$$\int_{-1}^{1} w(t)\, N(t)\, dt = 0 \tag{3.1.9}$$

This condition is both necessary and sufficient.

Proof. **Sufficient Part:** To prove the sufficiency we note that assuming (3.1.9), (3.1.7) becomes

$$\int_{-1}^{1} f(t)\, dt = \sum_{i=0}^{n} A_i f(t_i) + R$$

where R is given by (3.1.3). Now $R = 0$ if $f(t)$ is replaced by any polynomial of degree $\leq (2n + 1)$. Hence the degree of precision is $(2n + 1)$. □

Proof. **Necessary Part:** To prove the condition is necessary we assume that

$$\int_{-1}^{1} G_{2n+1}(t)\, dt = \sum_{i=0}^{n} A_i G_{2n+1}(t_i) \tag{3.1.10}$$

for an arbitrary polynomial G_{2n+1} of maximum degree $(2n+1)$. Hence it must be satisfied for the polynomial $G_{2n+1}(t) = w(t)\, N(t)$ giving

$$\int_{-1}^{1} w(t)\, N(t)\, dt = \sum_{i=0}^{n} A_i\, w(t_i)\, N(t_i) = 0 \quad [\because w(t_i) = 0, i = 0, 1, \ldots, n.] \tag{3.1.11}$$

□

Thus we get that formula (3.1.1) is valid if and only if (3.1.9) is satisfied.

Let the polynomial p successive indefinite integration of $w(t)$ be $w_p(t)$. Then after $(n+1)$ times integration by parts we have

$$\int_{-1}^{1} w(t)\, N(t)\, dt = \Big[w_1(t)N(t) - w_2(t)N'(t) + \ldots + (-1)^n w_{n+1}(t)N^n(t)\Big]_{-1}^{1}$$

$$\text{`} + (-1)^{n+1} \int_{-1}^{1} w_{2n+1}(t)N^{n+1}(t)\, dt \tag{3.1.12}$$

Since $N(t)$ is a polynomial of degree n, so $N^n(t)$ =constant and $N^{n+1}(t) = 0$. Therefore,

$$\int_{-1}^{1} w(t)\, N(t)\, dt = \Big[w_1(t)N(t) - w_2(t)N'(t) + \ldots + (-1)^n w_{n+1}(t)N^n(t)\Big]_{-1}^{1}$$

Equation (3.1.9) is satisfied if and only if

$$w_1(\pm 1) = w_2(\pm 1) = \ldots = w_{n+1}(t) = 0$$

This $(2n + 2)$ conditions are satisfied if we have

$$w(t) = c\frac{d^{n+1}}{dt^{n+1}}(t^2 - 1)^{n+1}$$

where c is a constant. Now comparing the coefficient of t^{n+1} from both the side, we obtain

$$c(2n + 2)(2n + 1)\ldots(n + 2) = 1 \Rightarrow c\frac{(2n + 2)!}{(n + 1)!} = 1 \Rightarrow c = \frac{(n + 1)!}{(2n + 2)!}$$

Therefore

$$w(t) = \frac{(n + 1)!}{(2n + 2)!}\frac{d^{n+1}}{dt^{n+1}}(t^2 - 1)^{n+1} \tag{3.1.13}$$

Now we know that the Legendre polynomial $P_{n+1}(x)$ of degree $(n + 1)$ is given by the Rodgrique's formula

$$P_{n+1}(t) = \frac{1}{2^{n+1}(n + 1)!}\frac{d^{n+1}}{dt^{n+1}}(t^2 - 1)^{n+1}$$

Using this we find from (3.1.13) that

$$w(t) = \frac{2^{n+1}[(n + 1)!]^2}{(2n + 2)!}P_{n+1}(t) \tag{3.1.14}$$

Hence t_i are roots of $P_{n+1}(t) = 0$. Now the roots $t_0, t_1, \ldots, t_n$ are all real, distinct and lie in the interval between -1 and 1. Knowing t_i for $i = 0(1)n$, the coefficient $A_i, i = 0(1)n$ are given by (3.1.8). Hence Gauss quadrature formula can be written as

$$\int_{-1}^{1} f(t)\, dt = \sum_{i=0}^{n} A_i\, f(t_i) + R \tag{3.1.15}$$

The error term in Gauss quadrature formula is given by

$$R = \frac{[(n + 1)!]^4}{[(2n + 2)!]^3}\frac{2^{2n+3}}{2n + 3}f^{2n+2}(\xi); \quad -1 < \xi < 1 \tag{3.1.16}$$

Since this is proportional to $(2n + 2)$th derivative of $f(t)$, the formula (3.1.15) is exact for all polynomial of degree $(2n + 1)$ or less.

3.1.2 Modification

Let us consider $\int_a^b f(x)\,dx$. Putting $x = \frac{a+b}{2} + \frac{b-a}{2}t$ so that $dx = \frac{b-a}{2}dt$. When $x = a$, $t = -1$ and for $x = b$, $t = 1$. Let

$$f(x) = f\left\{\frac{a+b}{2} + \frac{b-a}{2}t\right\} \equiv F(t).$$

Therefore

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} F(t)\,dt \tag{3.1.17}$$

Applying Gaussian quadrature formula, we have

$$\begin{aligned}\int_{-1}^{1} F(t)\,dt &= \sum_{i=0}^{n} A_i F(t_i) + R' \\ &= \sum_{i=0}^{n} A_i f(x_i) + R' \end{aligned} \tag{3.1.18}$$

where $x_i = \frac{a+b}{2} + \frac{b-a}{2}t_i$ and $R' = \frac{[(n+1)!]^4}{[(2n+2)!]^3}\frac{2^{2n+3}}{2n+3}F^{2n+2}(\xi'); \quad -1 < \xi' < 1$. Now,

$$\begin{aligned}&\frac{dF}{dt} = \frac{df}{dx}\cdot\frac{dx}{dt} = \frac{b-a}{2}\frac{df}{dx} \\ &\therefore R' = \frac{[(n+1)!]^4}{[(2n+2)!]^3}\frac{2^{2n+3}}{2n+3}\left(\frac{b-a}{2}\right)^{2n+2} f^{2n+2}(\xi'); \quad a < \xi < b. \end{aligned} \tag{3.1.19}$$

Hence Eq.(3.1.17) becomes

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\left[\sum_{i=0}^{n} A_i f(x_i) + R'\right] = \frac{b-a}{2}\sum_{i=0}^{n} A_i f(x_i) + R \tag{3.1.20}$$

where $R = \frac{b-a}{2}R'$

Remark 3.1. (i) The advantage of this formula lies in the fact that by use of $(n+1)$ points only, we are attaining an accuracy which would ordinarily result from the use of $(2n+2)$ points. Hence, this formula is twice as accurate as those base on equally spaced points.
(ii) The disadvantage is that the interpolating points in general corresponds to irrational number and their use may lead to excessive labour in numerical computation.

Illustration: Derive the Gauss quadrature formula for the case of 3 ordinates.

We have

$$\begin{aligned}P_3(t) &= \frac{1}{2^3 3!}\frac{d^3}{dt^3}\{(t^2-1)^3\} \\ &= \frac{1}{48}\frac{d^3}{dt^3}(t^6 - 3t^4 + 3t^2 - 1) \\ &= \frac{1}{48}[5t^3 - 3t]\end{aligned}$$

The points $t_0,\ t_1$ and t_2 are given by

$$t_0 = -\sqrt{3/5}, \quad t_1 = 0, \quad t_2 = +\sqrt{3/5}$$

Here

$$w(t) = \left(t + \sqrt{\frac{3}{5}}\right) t \left(t - \sqrt{\frac{3}{5}}\right)$$

$$\Rightarrow w'(t_0) = t_0\left(t_0 - \sqrt{\frac{3}{5}}\right) = \frac{6}{5},\ w'(t_1) = \left(t_1 + \sqrt{\frac{3}{5}}\right)\left(\sqrt{\frac{3}{4}}\right) = -\frac{3}{5},$$

$$\text{and } w'(t_2) = \left(t_2 + \sqrt{\frac{3}{5}}\right) t_2 = \frac{6}{5}$$

$$\therefore A_0 = \frac{5}{6}\int_{-1}^{1} t\left(t - \sqrt{\frac{3}{5}}\right) dt = \frac{5}{9},\ A_1 = -\frac{5}{3}\int_{-1}^{1}\left(t^2 + \sqrt{\frac{3}{5}}\right) dt = \frac{8}{9}$$

$$\text{and } A_2 = \frac{5}{6}\int_{-1}^{1}\left(t + \sqrt{\frac{3}{5}}\right) t\, dt = \frac{5}{9}$$

$$\text{For } n = 2, R = \frac{[3!]^4}{[6!]^3} f^{vi}(\xi)\frac{2^7}{7} = \frac{f^{vi}}{15750}$$

$$\therefore \int_{-1}^{1} f(t)\, dt = \frac{1}{9}\left[5f\left(-\sqrt{\frac{3}{5}}\right) + 8f(0) + 5f\left(\sqrt{\frac{3}{5}}\right)\right] + \frac{f^{iv}(\xi)}{15750}, \quad -1 < \xi < 1$$

3.2 Newton's quadrature formula

We derive in this section a general formula for numerical integration using Newton's forward difference formula.

Let the interval $[a, b]$ be divided into n equal subintervals such that $a = x_0 < x_1 < x_2 < \ldots < x_n = b$. Clearly, $x_n = x_0 + nh$. Hence the integral becomes

$$I = \int_{x_0}^{x_n} y\, dx. \tag{3.2.1}$$

Approximating y by Newton's forward difference formula, we obtain

$$I = \int_{x_0}^{x_n}\left[y_0 + p\Delta y_0 + \frac{p(p-1)}{2}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{6}\Delta^3 y_0 + \ldots\right] dx. \tag{3.2.2}$$

Since $x = x_0 + ph,\ dx = h\, dp$ and hence the above integral becomes

$$I = h\int_{0}^{n}\left[y_0 + p\Delta y_0 + \frac{p(p-1)}{2}\Delta^2 y_0 + \frac{p(p-1)(p-2)}{6}\Delta^3 y_0 + \ldots\right] dp, \tag{3.2.3}$$

which gives on simplification

$$\int_{x_0}^{x_n} y\,dx = nh\left[y_0 + \frac{n}{2}\Delta y_0 + \frac{n(2n-3)}{12}\Delta^2 y_0 + \frac{n(n-2)^2}{24}\Delta^3 y_0 + \ldots\right]. \tag{3.2.4}$$

From this *general formula*, we can obtain different integration formulae by putting $n = 1, 2, 3, \ldots$ etc. We derive a few of these formulae like *Trapezoidal Rule, Simpson's 1/3 rule, Simpson's 3/8 rule* which you have studied earlier. A short remainder of these formula are given below.

3.3 Numerical Integration Formulae

3.3.1 Trapezoidal Rule

The integral formula for *Trapezoidal rule* is given by

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{2}[y_0 + 2(y_1 + y_2 + \ldots + y_{n-1}) + y_n] \tag{3.3.1}$$

$$\text{Corresponding error is given by } E = -\frac{b-a}{12}h^2 M, \quad \text{where } M = \underbrace{max}_{a\le x\le b}|f''(x)| \tag{3.3.2}$$

3.3.2 Simpson's 1/3 Rule

The integral formula for *Trapezoidal rule* is given by

$$\int_{x_0}^{x_n} y\,dx = \frac{h}{3}[y_0 + 4(y_1 + y_3 + y_5 + \ldots + y_{n-1}) + 2(y_2 + y_4 + y_6 + \ldots + y_{n-2}) + y_n], \tag{3.3.3}$$

$$\text{Corresponding error is given by } \quad E = -\frac{b-a}{180}h^4 M, \quad \text{where } M = \underbrace{max}_{a\le x\le b}|f^{iv}(x)| \tag{3.3.4}$$

3.3.3 Simpson's 3/8 Rule

The integral formula for *Trapezoidal rule* is given by

$$\int_{x_0}^{x_n} y\,dx = \frac{3h}{8}[y_0 + 3y_1 + 3y_2 + 2y_3 + 3y_4 + 3y_5 + 2y_6 + \ldots + 2y_{n-3} + 3y_{n-2} + 3y_{n-1} + y_n], \tag{3.3.5}$$

$$\text{Corresponding error is given by } \quad E = -\frac{3}{80}h^4 M, \quad \text{where } M = \underbrace{max}_{a\le x\le b}|f^{iv}(x)| \tag{3.3.6}$$

3.4 Romberg integration

This method can often used to improve the approximate results obtained by the finite difference methods. Its application to the numerical evaluation of definite integrals, for example in the use of trapezoidal rule, can be described, as follows. We consider the definite integral

$$I = \int_a^b y\,dx \tag{3.4.1}$$

and evaluate it by the trapezoidal rule (3.3.1) with two different subintervals of widths h_1 and h_2 to obtain the approximate values of I_1 and I_2 respectively. Then (3.3.2) gives the errors E_1 and E_2 as

$$E_1 = -\frac{1}{12}(b-a)h_1^2 y''(\xi_1) \quad \text{and} \quad E_2 = -\frac{1}{12}(b-a)h_2^2 y''(\xi_2) \tag{3.4.2}$$

Since the term $y''(\xi_2)$ is also the largest value of $y''(x)$, it is reasonable to assume that the quantities $y''(\xi_1)$ and $y''(\xi_2)$ are very nearly the same. We therefore have

$$\frac{E_1}{E_2} = \frac{h_1^2}{h_2^2} \quad \Rightarrow \frac{E_2}{E_2 - E_1} = \frac{h_2^2}{h_2^2 - h_1^2}$$

Since $E_2 - E_1 = I_2 - I_1$, this gives

$$E_2 = \frac{h_2^2}{h_2^2 - h_1^2}(I_2 - I_1) \tag{3.4.3}$$

We therefore obtain a new approximation I_3 defined by

$$I_3 = I_2 - E_2 = \frac{I_1 h_2^2 - I_2 h_1^2}{h_2^2 - h_1^2}, \tag{3.4.4}$$

which, in general, would be closer to the actual value - provided that the errors decrease monotonically and are of the same sign. If we now set $h_2 = \frac{1}{2}h_1 = \frac{1}{2}h$, Eq.(3.4.4) can be written in the more convenient form

$$I\left(h, \frac{1}{2}h\right) = \frac{1}{3}\left[4I\left(\frac{1}{2}h\right) - I(h)\right] = I\left(\frac{h}{2}\right) + \frac{1}{3}\left[I\left(\frac{1}{2}h\right) - I(h)\right], \tag{3.4.5}$$

where $I(h) = I_1, I\left(\frac{1}{2}h\right) = I_2$ and $I\left(h, \frac{1}{2}h\right) = I_3$. With this notation the following table can be formed

$I(h)$			
	$I\left(h, \frac{1}{2}h\right)$		
$I\left(\frac{1}{2}h\right)$		$I\left(h, \frac{1}{2}h, \frac{1}{4}h\right)$	
	$I\left(\frac{1}{2}h, \frac{1}{4}h\right)$		$I\left(h, \frac{1}{2}h, \frac{1}{4}h, \frac{1}{8}h\right)$
$I\left(\frac{1}{4}h\right)$		$I\left(\frac{1}{2}h, \frac{1}{4}h, \frac{1}{8}h\right)$	
	$I\left(\frac{1}{4}h, \frac{1}{8}h\right)$		
$I\left(\frac{1}{8}h\right)$			

This computations can be stopped when two successive values are sufficiently close to each other. This method, due to L.F. Richardson, is called the *deferred approach to the limit* and the systematic tabulation of this is called *Romberg Integration*.

Illustration: Show that the formula (3.4.5) gives the Simpson's $\frac{1}{3}$ rule of integration.

Proof. Let us divide the interval $[a, b]$ into n equal subintervals by the points $a = x_0, x_2, x_4, \ldots, x_{2n} = b$. Then

$$I_1 = \frac{h}{2}\Big[y_0 + 2(y_2 + y_4 + \ldots + y_{2n-2}) + y_{2n}\Big]$$

□

Again dividing the interval $[a, b]$ into $2n$ equal subintervals, each of length $\frac{1}{2}h$ by the points $a = x_0, x_1, x_2, \ldots, x_{2n-1}, x_{2n} = b$ so that we have

$$I_2 = \frac{h}{4}\Big[y_0 + 2(y_1 + y_2 + y_3 + \ldots + y_{2n-1}) + y_{2n}\Big]$$

Now

$$\begin{aligned}
\frac{1}{3}\Big[4I_2 - I_1\Big] &= \frac{h}{3}\Big[\Big\{y_0 + 2(y_1 + y_2 + y_3 + \ldots + y_{2n-1}) + y_{2n}\Big\} - \frac{1}{2}\Big\{y_0 + 2(y_2 + y_4 + \ldots + y_{2n-2}) + y_{2n}\Big\}\Big] \\
&= \frac{h}{3}\Big[\frac{1}{2}y_0 + 2(y_1 + y_3 + y_5 + \ldots + y_{2n-1}) + (y_2 + y_4 + y_6 + \ldots + y_{2n-2}) + \frac{1}{2}y_{2n}\Big] \\
&= \frac{h/2}{3}\Big[y_0 + 4(y_1 + y_3 + y_5 + \ldots + y_{2n-1}) + 2(y_2 + y_4 + y_6 + \ldots + y_{2n-2}) + y_{2n}\Big]
\end{aligned}$$

This formula gives Simpson's $1/3$ rule, and hence the error is of the order h^4.

Example 3.2. Use Romberg's method to compute $I = \int_0^1 \frac{1}{1+x}dx$, correct to three decimal places.

Solution: We take $h = 0.5,\ 0.25$ and 0.125 successively and obtain

$$I(h) = \frac{1}{4}\Big[1.0000 + 2(0.6667) + 0.5\Big] = 0.7084$$
$$I\left(\frac{1}{2}h\right) = \frac{1}{8}\Big[1.0 + 2(0.8000 + 0.6667 + 0.5714) + 0.5\Big] = 0.6970$$
$$I\left(\frac{1}{4}h\right) = \frac{1}{6}\Big[1.0 + 2(0.8889 + 0.8000 + 0.7273 + 0.6667) + (0.6154 + 0.5714 + 0.5333) + 0.5\Big] = 0.6941$$

Now using the formula (3.4.5), we obtain

$$I\left(h, \frac{1}{2}h\right) = 0.6970 + \frac{1}{3}(0.6970 - 0.7084) = 0.6932$$
$$I\left(\frac{1}{2}h, \frac{1}{4}h\right) = 0.6941 + \frac{1}{3}(0.6941 - 0.6970) = 0.6931$$

Finally, we obtain

$$I\left(h, \frac{1}{2}h, \frac{1}{4}h\right) = 0.6931 + \frac{1}{3}(0.6931 - 0.6932) = 0.6931$$

The table of values are, therefore,

0.7084		
	0.9632	
0.6970		0.6931
	0.6931	
0.6941		

Hence, $I = \int_0^1 \frac{1}{1+x}dx = 0.693$ (correct to three decimal places). An obvious advantage of this method is that the accuracy of the computed value is known at each step.

Exercise 3.3. (i) Use Romberg integration to compute $I = \int_0^1 e^{-x^2} dx$, correct to three decimal places.

Answer: 0.747

(ii) Use Romberg integration to compute $I = \int_0^{2\pi} \sin x \, dx$, correct to three decimal places. **Answer:** 1.999

(iii) Use Romberg integration to compute $I = \int_0^4 x^5 \, dx$, correct to three decimal places. **Answer:** 682.667

3.5 Euler-MacLaurin Formula

Consider the expansion of $1/(e^x - 1)$ in ascending power of x, obtained by writing the MacLaurin expnsion of e^x and simplifying

$$\begin{aligned} \frac{1}{e^x - 1} &= \frac{1}{\left[x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots\right]} = \frac{1}{x}\left[1 + \left(\frac{x}{2!} + \frac{x^2}{3!} + \ldots\right)\right]^{-1} \\ &= \frac{1}{x} - \frac{1}{2} + B_1 x + B_3 x^3 + B_5 x^5 + \ldots, \end{aligned} \tag{3.5.1}$$

where $B_{2r} = 0,\ B_1 = \frac{1}{12},\ B_3 = -\frac{1}{720},\ B_5 = \frac{1}{30240}$, etc. In Eq.(3.5.1), if we set $x = hD$ and use the relation $E \equiv e^{hD}$, we obtain the identity

$$\frac{1}{E-1} = \frac{1}{hD} - \frac{1}{2} + B_1 hD + B_3 h^3 D^3 + B_5 h^5 D^5 + \ldots$$

or equivalently

$$\frac{E^n - 1}{E - 1} = \frac{1}{hD}(E^n - 1) - \frac{1}{2}(E^n - 1) + B_1 hD(E^n - 1) + B_3 h^3 D^3 (E^n - 1) + \ldots \tag{3.5.2}$$

Operating this identity on y_0, we obtain

$$\begin{aligned} \frac{E^n - 1}{E - 1} y_0 &= \frac{1}{hD}(E^n - 1)y_0 - \frac{1}{2}(E^n - 1)y_0 + B_1 hD(E^n - 1)y_0 + \ldots \\ &= \frac{1}{hD}(y_n - y_0) - \frac{1}{2}(y_n - y_0) + B_1 h(y_n' - y_0') + B_3 h^3 (y_n''' - y_0''') + B_5 h^5 (y_n^v - y_0^v) + \ldots \end{aligned}$$

It can be easily shown that the left-hand side denotes the sum $y_0 + y_1 + y_2 + \ldots + y_{n-1}$, whereas the term $\frac{1}{hD}(y_n - y_0)$ on the right side can be written as $\frac{1}{h}\int_{x_0}^{x_n} y \, dx$, since $1/D$ can be interpreted as an integration operator. Hence

$$\begin{aligned} \int_{x_0}^{x_n} y \, dx &= \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \ldots + 2y_{n-1} + y_n) - \frac{h^2}{12}(y_n' - y_0') \\ &+ \frac{h^4}{720}(y_n''' - y_0''') - \frac{h^6}{30240}(y^v - y_0^v) + \ldots \end{aligned} \tag{3.5.3}$$

which is called the *Euler-MacLaurin's formula* for integration. The first expression on the right-side of Eq.(3.5.3) denotes the approximate value of the integral obtained by using trapezoidal rule and the other expressions represent the successive *corrections* to this value. It should be noted that this formula may also be used to find the sum of a series of the form $y_0 + y_1 + y_2 + \ldots + y_n$. The use use of this formula is illustrated by the following example.

Example 3.4. Evaluate $I = \int_0^{\pi/2} \sin x \, dx$ using the Euler-Maclaurin's formula.

Solution: In this case, formula (3.5.3) simplifies to

$$\int_0^{\pi/2} \sin x \, dx = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + \ldots + 2y_{n-1} + y_n) + \frac{h^2}{12} + \frac{h^4}{720} + \frac{h^6}{30240} + \ldots$$

To evaluate the integral, we take $h = \pi/4$. Then we obtain

$$\begin{aligned} \int_0^{\pi/4} \sin x \, dx &= \frac{\pi}{8}(0 + 2 + 0) + \frac{\pi^2}{192} + \frac{\pi^4}{184320} + \ldots \\ &\approx \frac{\pi}{4} + \frac{\pi^2}{192} + \frac{\pi^4}{184320} \\ &= 0.785398 + 0.051404 + 0.000528 = 0.837330 \end{aligned}$$

On the other hand with $h = \pi/8$, we obtain

$$\begin{aligned} \int_0^{\pi/4} \sin x \, dx &= \frac{\pi}{16}\Big[0 + 2(0.382683) + 0.707117 + 0.923879 + 1.000000\Big] \\ &= 0.987119 + 0.012851 + 0.000033 = 1.000003 \end{aligned}$$

Exercise 3.5. (i) Use the Euler-Maclaurin formula to evaluate the integral $I = \int_1^2 (\cos x + \ln x - e^x) \, dx$

Answer: - 4.21667

(ii) Use the Euler-Maclaurin formula to prove $\sum_1^n x^2 = \dfrac{n(n+1)(2n+1)}{6}$

(iii) Use the Euler-Maclaurin formula to find the sum $S = 1^3 + 2^3 + 3^3 + \ldots + n^3$

Unit 4

Course Structure

Systems of Linear Algebraic Equations: Direct methods - Factorization method, Eigen value and Eigenvector Problems : Direct methods, Iterative method – Power method.

---

4 Introduction

Many problems arising from engineering and applied sciences require the solution of systems of linear algebraic equations and computation of eigenvalues and eigenvectors of a matrix. We assume that the readers are familiar with the theory of determinants and elements of matrix algebra since these provide a convenient way to represent linear algebraic equations. For example, the system of equations

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

may be represented as the matrix equation $AX = B$, where

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ and } B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}.$$

The solution of a linear system of equations can be accomplished by a numerical method which falls in one of two categories: (i) direct method, (ii) iterative method. In this unit, we will mainly discuss LU factorization/decomposition method and two types of iterative methods.

4.1 Direct method - LU decomposition or factorization method

Let there be a system of equation

$$AX = B \tag{4.1.1}$$

where A is a $n \times n$ matrix and B is a $n \times 1$ column vector. Now this method is based on the fact that a square matrix A of (4.1.1) can be decomposed or factorized into a product of a lower triangular matrix L and a upper triangular matrix U if all the principal minor in the matrix A are non-singular. Let us write

$$A = LU \tag{4.1.2}$$

where

$$L = \begin{bmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ l_{n1} & l_{n2} & l_{nn} & \dots & l_{nn} \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ 0 & u_{22} & u_{23} & \dots & u_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & u_{nn} \end{bmatrix}$$

Using the matrix multiplication rule to multiply the matrices L and U and comparing the corresponding element of the resulting matrix with those of the matrix A, one obtain

$$l_{i1}u_{1j} + l_{i2}u_{2j} + l_{i3}u_{3j} + \ldots + l_{in}u_{nj} = a_{ij}, \;\; j = 1(1)n, \tag{4.1.3}$$

where $l_{ij} = 0, \;\; j > i$ and $u_{ij} = 0, \;\; j < i$. The system of equation (4.1.3) involves $(n^2 + n)$ unknowns. To find a solution, we either choose $u_{ii} = 1$ or $l_{ii} = 1$ for $i = 1, 2, \ldots, n$. When we choose $l_{ii} = 1$ for all $i = 1, 2, \ldots, n$ the method is called *Do-little's* method and when we choose $u_{ii} = 1$ for all $i = 1, 2, \ldots, n$ is called *Crout's* method. Here we take $u_{ii} = 1$ for all $i = 1, 2, \ldots, n$, the solution of the system (4.1.3) may be written as

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}, \;\; i \geq j \;\; \text{and} \;\; u_{ij} = \frac{1}{l_{ii}}\Big[a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\Big], \;\; i < j, \;\; u_{ii} = 1 \tag{4.1.4}$$

One may note that

$$\begin{aligned} l_{i1} &= a_{i1} \;\; \text{for all} \;\; i = 1(1)n \\ u_{1j} &= a_{1j}/l_{11} \;\; \text{for all} \;\; j = 2(1)n \end{aligned} \tag{4.1.5}$$

Thus the first column of L and first row of U are determined. We now find second column of L and second row of U as follows:

$$\begin{aligned} l_{i2} &= a_{i2} - l_{i1}u_{i2} \;\; \text{for all} \;\; i = 2(1)n \\ u_{2j} &= [a_{2j} - l_{21}u_{1j}]/l_{22} \;\; \text{for all} \;\; j = 2(1)n \end{aligned} \tag{4.1.6}$$

Next we find the third column of L and third row of U, fourth column of L and fourth row of U and so on the $(n-1)$-th column of L and $(n-1)$-th row of U and finally l_{nn} and u_{nn}. After having obtained the matrices L and U the system of equations of (4.1.1) becomes

$$LUX = B, \tag{4.1.7}$$

we write the system (4.1.7) as the following two system of equations

$$UX = Z \quad \text{and} \quad LZ = B \tag{4.1.8}$$

The unknown $z_1, z_2, \ldots, z_n$ can be found by forward substitution, while the unknown $x_1, x_2, \ldots, x_n$ can be determined by backward substitution. Alternatively, one can obtain L^{-1} and U^{-1} in order to find $Z = L^{-1}B$ and $X = U^{-1}Z$. This method is applicable when the matrix A is positive definite (i.e., $X^TAX > 0$ for all non-zero $X \in \mathbb{R}^n$). However this is only the sufficient condition.

Example 4.1. Solve the following system of equation by the method of LU decomposition.

$$\begin{aligned} 2x + 3y + z &= 9 \\ x + 2y + 3z &= 6 \\ 3x + y + 2z &= 8 \end{aligned}$$

Solution: Here

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 9 \\ 6 \\ 8 \end{bmatrix}$$

Let

$$\begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 2 & 3 & 1 \\ 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{bmatrix}$$

Comparing both sides we obtain

$$\begin{array}{lllll} l_{11} = 2 & u_{12} = 3/2 & u_{13} = 1/2 \\ l_{21} = 1 & l_{22} = 1/2 & u_{23} = 5 \\ l_{31} = 3 & l_{32} = -7/2 & l_{33} = 18 \end{array}$$

Therefore

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 1/2 & 0 \\ 3 & -7/2 & 18 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 3/2 & 1/2 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix}$$

Now if $Z = [z_1 \;\; z_2 \;\; z_3]^T$, then the equation $LZ = B$ gives the solution:

$$z_1 = 9/2, \quad z_2 = 3, \quad \text{and} \quad z_3 = 5/18$$

Finally the matrix $UX = Z$ where $X = [x \;\; y \;\; z]^T$, gives the required solution

$$x = 35/18, \quad y = 29/18, \quad \text{and} \quad z = 5/18$$

Check:
Proceed the same problem by *Do-little* method and verify that the computation proceed to the same results.

Exercise 4.2. (i) Decompose the matrix

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

into the form LU where L is unit lower triangular and U an upper triangular matrix. Hence solve the system $AX = B$ where $B = [4 \;\; 8 \;\; 10]^T$ **Answer:** $x_1 = 1.1193,\ x_2 = 0.8685,\ x_3 = 0.1407$
(ii) Design an algorithm to reduce a given system of equations to upper triangular form. Test your algorithm on the system:

$$\begin{array}{rcl} 4x + 3y + 2z &=& 16 \\ 2x + 3y + 4z &=& 20 \\ x + 2y + z &=& 8 \end{array}$$

Answer: $x_1 = 1,\ x_2 = 2,\ x_3 = 3$
(iii) Decompose the matrix

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

into the form LU, where L is a lower triangular matrix and U is unit upper triangular matrix.

4.2 Eigen value and Eigenvector Problems

Let A be a square matrix of order n with elements a_{ij}. We wish to find a column vector X and a constant λ such that

$$AX = \lambda X \tag{4.2.1}$$

In Eq.(4.2.1), λ is called the *eigenvalue* and X is called the corresponding *eigenvector*. The matrix Eq.(4.2.1), when written out in full, represents a set of homogeneous linear equations:

$$\begin{aligned} (a_{11}-\lambda)x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22}-\lambda)x_2 + \ldots + a_{2n}x_n &= 0 \\ \vdots \quad \vdots \quad \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn}-\lambda)x_n &= 0. \end{aligned} \tag{4.2.2}$$

A nontrivial solution exists only when the coefficient determinant in (4.2.2) vanishes. Hence, we have

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn}-\lambda \end{vmatrix} = 0. \tag{4.2.3}$$

This equation, called the *characteristic equation* of the matrix A, is a polynomial equation of degree n in λ, the polynomial being called the *characteristic polynomial* of A. If the roots of Eq.(4.2.3) be given by $\lambda_i (i = 1, 2, \ldots, n)$, then for each value of λ_i, there exist a corresponding X_i such that

$$AX_i = \lambda_i X_i. \tag{4.2.4}$$

The eigenvalues λ_i may be either distinct (i.e. all different) or *repeated*. The evaluation of eigenvectors in the case of the repeated roots is a much involved process and will not be attempted here. The set of all eigenvalues, λ_i, of a matrix A is called the *spectrum* of A and the largest of $|\lambda_i|$ is called the *spectral radius* of A. The eigen values are obtained by solving the algebraic Eq.(4.2.3). This method, which is demonstrated in Example (4.3), is unsuitable for matrices of higher order and better methods must be applied, which is beyond of our syllabus. Readers are suggested to go through any standard book of numerical analysis. In some practical applications only the numerically largest eigenvalue and the corresponding eigenvector are required, and we will describe an iterative method, namely the *Power Method*, to compute the largest eigenvalue. This method is easy of application and also well-suited for machine computations.

4.2.1 Direct Method

In this subsection we will learn, how to calculate eigenvalues and eigenvector a matrix by direct method. Let us consider the following example.

Example 4.3. Find the eigenvalues and eigenvectors of the matrix:

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

Solution: The characteristic equation of this matrix is given by

$$\begin{vmatrix} 5-\lambda & 0 & 1 \\ 0 & -2-\lambda & 0 \\ 1 & 0 & 5-\lambda \end{vmatrix} = 0.$$

which gives $\lambda_1 = -2$, $\lambda_2 = 4$ and $\lambda_3 = 6$. The corresponding eigenvectors are obtained thus

(i) $\lambda_1 = -2$. Let the eigenvector be $X_1 = [x_1 \;\; x_2 \;\; x_3]^T$. Then we have

$$A \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = -2 \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

which gives the equations

$$7x_1 + x_3 = 0 \quad \text{and} \quad x_1 + 7x_3 = 0$$

The solution is $x_1 = x_3 = 0$ with x_2 arbitrary. In particular, we take $x_2 = 1$ and the eigenvector is $X_1 = [0 \;\; 1 \;\; 0]^T$.

(ii) $\lambda_2 = 4$. With $X_2 = [x_1 \;\; x_2 \;\; x_3]^T$ as the eigenvector, the equations are

$$x_1 + x_3 = 0 \quad \text{and} \quad -6x_2 = 0,$$

from which we obtain $x_1 = -x_3$ and $x_2 = 0$. We choose, in particular, $x_1 = 1/\sqrt{2}$ and $x_3 = -1/\sqrt{2}$ so that $x_1^2 + x_2^2 + x_3^2 = 1$. The eigenvector chosen in this way is said to be *normalized*. We, therefore, have $X_2 = [1/\sqrt{2} \;\; 0 \;\; -1/\sqrt{2}]^T$.

(iii) $\lambda_3 = 6$. If $X_3 = [x_1 \;\; x_2 \;\; x_3]^T$ is the required eigenvector, then the equations are

$$\begin{aligned} -x_1 + x_3 &= 0 \\ -8x_2 &= 0 \\ x_1 - x_3 &= 0, \end{aligned}$$

which give $x_1 = x_3$ and $x_2 = 0$. Choosing $x_1 = x_3 = 1/\sqrt{2}$, the normalised eigenvector is given by $X_3 = [1/\sqrt{2} \;\; 0 \;\; 1/\sqrt{2}]^T$.

4.2.2 Iterative method - Power Method

The method for finding the largest eigenvalue in magnitude and the corresponding eigen vector of the eigenvalue problem $AX = \lambda X$, is called the Power method.

We assume that $\lambda_1, \lambda_2, \ldots, \lambda_n$ are distinct eigenvalues such that

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|. \tag{4.2.5}$$

Let $v_1, v_2, \ldots, v_n$ be the eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively. The method is applicable if a complete system of n linearly independent eigenvectors exist, even though some of the eigenvalues may not be distinct. The n linearly independent eigenvectors form an n-dimensional space. Any vector v in this space of eigenvectors $v_1, v_2, \ldots, v_n$ can be written as a linear combination of these vectors. That is,

$$v = c_1 v_1 + c_2 v_2 + \ldots + c_n v_n. \tag{4.2.6}$$

Premultiplying by A and substituting $Av_1 = \lambda_1 v_1, Av_2 = \lambda_2 v_2, \ldots, Av_n = \lambda_n v_n$, we get

$$Av = c_1\lambda_1 v_1 + c_2\lambda_2 v_2 + \ldots + c_n\lambda_n v_n = \lambda_1 \left[c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right) v_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1}\right) v_n \right].$$

Premultiplying repeatedly by A and simplifying, we get

$$\begin{aligned} A^2v &= \lambda_1^2\left[c_1v_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^2 v_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^2 v_n\right]. \\ &\vdots \\ A^kv &= \lambda_1^k\left[c_1v_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^k v_n\right]. \end{aligned} \tag{4.2.7}$$

$$A^{k+1}v = \lambda_1^{k+1}\left[c_1v_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1} v_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1} v_n\right]. \tag{4.2.8}$$

As $k \to \infty$, the right sides of (4.2.7) and (4.2.8) tend to $\lambda_1^k c_1 v_1$ and $\lambda_1^{k+1} c_1 v_1$, since $|\lambda_i/\lambda_1| < 1,\ i = 2, 3, \ldots, n$. Both the right hand side vectors in (4.2.7) and (4.2.8)

$$\begin{gathered} [c_1v_1 + c_2(\lambda_2/\lambda_1)^k v_2 + \ldots + c_n(\lambda_n/\lambda_1)^k v_n], \\ \text{and } [c_1v_1 + c_2(\lambda_2/\lambda_1)^{k+1} v_2 + \ldots + c_n(\lambda_n/\lambda_1)^{k+1} v_n], \end{gathered}$$

tend to c_1v_1, which is the eigenvector corresponding to λ_1. The eigenvalue λ_1 is obtained as the ratio of the corresponding components of $A^{k+1}v$ and A^kv. That is,

$$\lambda_1 = \lim_{k\to\infty} \frac{(A^{k+1}v)_r}{(A^kv)_r}, \quad r = 1, 2, 3, \ldots, n \tag{4.2.9}$$

where the suffix r denotes the r-th component of the vector. Therefore, we obtain n ratio, all of them tending to the same value, which is the largest eigenvalue in magnitude, $|\lambda_1|$. The iterations are stopped when all the magnitudes of the differences of the ratios are less than the given error tolerance.

Remark 4.4. The choice of the initial approximation vector v_0 is important. If no suitable approximation is available, we can choose v_0 with all its components as one unit, that is, $v_0 = [1\ \ 1\ \ 1\ \ \ldots\ \ 1]^T$. However, this initial approximation to the vector should be non-orthogonal to v_1.

Remark 4.5. Faster convergence is obtained when $|\lambda_2| << \lambda_1$. As $k \to \infty$, premultiplication each time by A, may introduce round-off errors. In order to keep the round-off errors under control, we normalize the vector before permultiplying by A. The normalization that we use is to make the largest element in magnitude as unity. If we use the normalization, a simple algorithm for the power method can be written as follows:

$$y_{k+1} = Av_k, \tag{4.2.10}$$

$$v_{k+1} = y_{k+1}/m_{k+1} \tag{4.2.11}$$

where m_{k+1} is the largest element in magnitude of y_{k+1}. Now, the largest element in magnitude of v_{k+1} is one unit. Then (4.2.9) can be written as

$$\lambda_1 = \lim_{k\to\infty} \frac{(y_{k+1})_r}{(v_k)r}, \ \ r = 1, 2, 3, \ldots, n \tag{4.2.12}$$

and v_{k+1} is the required eigenvector.

Remark 4.6. It may be noted that as $k \to \infty$, m_{k+1} also gives $|\lambda_1|$.

Remark 4.7. Power method gives the largest eigenvalue in magnitude. If the sign of the eigenvalue is required, then we substitute this value in the determinant $|A - \lambda_1 I$ and find its value. If this value is approximately zero, then the eigenvalue is of positive sign. Otherwise, it is of negative sign.

Example 4.8. Determine the numerically largest eigenvalue and the corresponding eigenvector of the following matrix, using the power method.

$$A = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix}$$

Solution: Let the initial approximation to the eigenvector be v_0. Then, the power method is given by

$$\begin{aligned} y_{k+1} &= Av_k, \\ v_{k+1} &= y_{k+1}/m_{k+1} \end{aligned}$$

where m_{k+1} is the largest element in magnitude of y_{k+1}. The dominant eigenvalue in magnitude is given by

$$\lambda_1 = \lim_{k\to\infty} \frac{(y_{k+1})_r}{(v_k)_r}, \quad r = 1, 2, 3, \ldots, n$$

and v_{k+1} is the required eigenvector. Let the initial approximation to the eigenvector be $v_0 = [1 \;\; 1 \;\; 1]^T$. We have the following results.

$$y_1 = Av_0 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 28 \\ 4 \\ -2 \end{bmatrix}, \;\; m_1 = 28$$

$$v_1 = \frac{1}{m_1} y_1 = \frac{1}{28} \begin{bmatrix} 28 \\ 4 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.14286 \\ -0.07143 \end{bmatrix}$$

$$y_2 = Av_1 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.14286 \\ -0.07143 \end{bmatrix} = \begin{bmatrix} 25.0000 \\ 1.14286 \\ 2.28572 \end{bmatrix}, \;\; m_2 = 25;$$

$$v_2 = \frac{1}{m_2} y_2 = \frac{1}{25.0} \begin{bmatrix} 25.00000 \\ 1.14286 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.05714 \\ 0.09143 \end{bmatrix}$$

$$y_3 = Av_2 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.05714 \\ 0.09143 \end{bmatrix} = \begin{bmatrix} 25.2400 \\ 1.17142 \\ 1.63428 \end{bmatrix}, \;\; m_3 = 25.24;$$

$$v_3 = \frac{1}{m_3} y_3 = \frac{1}{25.24} \begin{bmatrix} 25.24000 \\ 1.17142 \\ 1.63428 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04641 \\ 0.06475 \end{bmatrix}$$

$$y_4 = Av_3 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.04641 \\ 0.06475 \end{bmatrix} = \begin{bmatrix} 25.17591 \\ 1.13923 \\ 1.74100 \end{bmatrix}, \;\; m_4 = 25.17591;$$

$$v_4 = \frac{1}{m_4} y_4 = \frac{1}{25.17591} \begin{bmatrix} 25.17591 \\ 1.13923 \\ 1.74100 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04525 \\ 0.06915 \end{bmatrix}$$

$$y_5 = Av_4 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.04525 \\ 0.06915 \end{bmatrix} = \begin{bmatrix} 25.18355 \\ 1.13575 \\ 1.72340 \end{bmatrix}, \quad m_5 = 25.18355;$$

$$v_5 = \frac{1}{m_5} y_5 = \frac{1}{25.18355} \begin{bmatrix} 25.18355 \\ 1.13575 \\ 1.72340 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04510 \\ 0.06843 \end{bmatrix}$$

$$y_6 = Av_5 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.04510 \\ 0.06843 \end{bmatrix} = \begin{bmatrix} 25.18196 \\ 1.13530 \\ 1.72628 \end{bmatrix}, \quad m_6 = 25.18196;$$

$$v_6 = \frac{1}{m_6} y_6 = \frac{1}{25.18196} \begin{bmatrix} 25.18196 \\ 1.13530 \\ 1.72628 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04508 \\ 0.06855 \end{bmatrix}$$

$$y_7 = Av_6 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.04508 \\ 0.06855 \end{bmatrix} = \begin{bmatrix} 25.18218 \\ 1.13524 \\ 1.72580 \end{bmatrix}, \quad m_7 = 25.18218;$$

$$v_7 = \frac{1}{m_7} y_7 = \frac{1}{25.18218} \begin{bmatrix} 25.18218 \\ 1.13524 \\ 1.72580 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04508 \\ 0.06853 \end{bmatrix}$$

$$y_8 = Av_7 = \begin{bmatrix} 25 & 1 & 2 \\ 1 & 3 & 0 \\ 2 & 0 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 0.04508 \\ 0.06853 \end{bmatrix} = \begin{bmatrix} 25.18214 \\ 1.13524 \\ 1.72588 \end{bmatrix}, \quad m_8 = 25.18214;$$

$$v_8 = \frac{1}{m_8} y_8 = \frac{1}{25.18214} \begin{bmatrix} 25.18214 \\ 1.13524 \\ 1.72588 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.04508 \\ 0.06854 \end{bmatrix}.$$

Now, we find the ratios

$$\lambda_1 = \lim_{k \to \infty} \frac{(y_{k+1})_r}{(v_k)_r}, \quad r = 1, 2, 3.$$

We obtain the ratios as

$$25.18214, \quad \frac{1.13524}{0.04508} = 25.18279, \quad \frac{1.72588}{0.06853} = 25.18430$$

The magnitude of the errors of the difference of these ratios are $0.00065,\ 0.00216, 0.00151$, which are less than 0.005. Hence, the results are correct to two decimal places. Therefore, the largest eigenvalue in magnitude is $|\lambda_1| = 25.18$. The corresponding eigenvector is $[1 \ \ 0.04508 \ \ 0.06854]^T$.

In remark (4.6), we have noted that as $k \to \infty$, m_{k+1} also gives $|\lambda_1|$. We find that this statement is true since $|m_8 - m_7| = |25.18214 - 25.18220| = 0.00006$.

If we require the sign of the eigenvalue, we substitute λ_1 in the characteristic equation. In the present problem, we find that $|A - 25.15I| = 1.4018$, while $A + 25.18I|$ is very large. Therefore, the required eigenvalue is 25.18.

Exercise 4.9. (i) Determine the dominant eigenvalue of $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ by power method. **Answer:** 5.3722

(ii) Determine the largest eigenvalue in magnitude and corresponding eigenvector of the following matrix by power method.

$$A = \begin{bmatrix} 1 & -3 & 2 \\ 4 & 4 & -1 \\ 6 & 3 & 5 \end{bmatrix} \qquad \textbf{Answer} : |\lambda| = 6.98, \;\; v = [0.29737 \;\; 0.6690 \;\; 1.0]^T$$

Unit 5

Course Structure

Nonlinear Equations : Fixed point iteration method, convergence and error estimation. Modified Newton-Raphson method, Muller's method

5 Introduction

In scientific and engineering studies, a frequently occurring problem is to find the roots of equations of the form

$$f(x) = 0 \tag{5.0.1}$$

If $f(x)$ is quadratic, cubic and a biquadratic expression, then algebraic formulae are available for expressing the roots in terms of the coefficients. On the other hand, when $f(x)$ is a polynomial of higher degree or an expression involving transcendental functions, algebraic methods are not available, and recourse must be taken to find the roots by approximate methods. It is assumed that the readers are already familiar with the bisection method, the method of false position. In these methods, we require an interval in which the root lies. We now describe methods which require one or more approximate values to start the solution.

5.1 Fixed point iteration method

In order to describe the method, we first rewrite the Eq.(5.0.1) in the form

$$x = \phi(x) \tag{5.1.1}$$

Now, let x_0 be an approximate root of Eq.(5.1.1). Then, substituting in Eq.(5.1.1), we get the first approximation as

$$x_1 = \phi(x_0)$$

Successive substitutions give the approximations

$$x_2 = \phi(x_1), \quad x_3 = \phi(x_2), \quad \ldots, \quad x_n = \phi(x_{n-1}).$$

The sequence may not converge to a definite number. But if the sequence converges to a definite number ξ, then ξ will be a root of the equation $x = \phi(x)$. To show this, let

$$x_{n+1} = \phi(x_n) \tag{5.1.2}$$

be the relation between the n-th and $(n+1)$-th approximations. As n increases, $x_{n+1} \to \xi$ and if $\phi(x)$ is a continuous function, then $\phi(x_n) \to \phi(\xi)$. Hence, in the limit, we obtain

$$\xi = \phi(\xi), \tag{5.1.3}$$

which shows that ξ is a root of the equation $x = \phi(x)$.

5.1.1 Condition of Convergence

To establish the condition of convergence of Eq.(5.1.1), we proceed in the following way:

From Eq.(5.1.2), we have

$$x_1 = \phi(x_0) \tag{5.1.4}$$

From Eqs.(5.1.3) and Eq.(5.1.4), we get

$$\xi - x_1 \;=\; \phi(\xi) - \phi(x_0) = (\xi - x_0)\phi'(\xi_0), \;\; x_0 < \xi_0 < \xi, \tag{5.1.5}$$

Similarly, we obtain

$$\begin{aligned}
\xi - x_2 &= (\xi - x_1)\phi'(\xi_1), \quad x_1 < \xi_1 < \xi && (5.1.6)\\
\xi - x_3 &= (\xi - x_2)\phi'(\xi_2), \quad x_2 < \xi_2 < \xi && (5.1.7)\\
&\vdots \\
\xi - x_{n+1} &= (\xi - x_n)\phi'(\xi_n), \quad x_n < \xi_n < \xi && (5.1.8)
\end{aligned}$$

If we assume $|\phi'(\xi_i)| \leq k$ for all i, then the above equation give

$$\begin{aligned}
|\xi - x_1| &\leq k|\xi - x_0| \\
|\xi - x_2| &\leq k|\xi - x_1| \\
|\xi - x_3| &\leq k|\xi - x_2| \\
&\vdots \\
|\xi - x_{n+1}| &\leq k|\xi - x_n|
\end{aligned}$$

Multiplying the corresponding sides of the above equations, we obtain

$$|\xi - x_{n+1}| \leq k^{n+1}|\xi - x_0| \tag{5.1.9}$$

If $k < 1$, i.e., if $|\phi'(\xi_i)| < 1$, then the right side of Eq.(5.1.9) tends to zero and the sequence of approximation $x_0,\ x_1,\ x_2,\ \ldots$ converges to the root ξ. Thus, when we express the equation $f(x) = 0$ in the form $x = \phi(x)$, then $\phi(x)$ must be such that

$$|\phi'(x) < 1|$$

in an immediate neighbourhood of the root. It follows that if *the initial approximation x_0 is chosen in an interval containing the root ξ, then the sequence of approximation converges to the root ξ.*

5.1.2 Error Estimation

We shall fin the error in the root obtained. We have

$$\begin{aligned}
|\xi - x_n| &\leq k|\xi - x_{n-1}| \\
\Rightarrow |\xi - x_n| &= k|\xi - x_n + x_n - x_{n-1}| \\
\Rightarrow |\xi - x_n| &\leq k[|\xi - x_n| + |x_n - x_{n-1}|] \\
\Rightarrow |\xi - x_n| &\leq \frac{k}{1-k}|x_n - x_{n-1}| = \frac{k}{1-k}k^{n-1}|x_1 - x_0| = \frac{k^n}{1-k}|x_1 - x_0|, \qquad (5.1.10)
\end{aligned}$$

which shows that the convergence would be faster for smaller values of k. Now, let ε be the specific accuracy so that

$$|\xi - x_n| \leq \varepsilon$$

Then, Eq.(5.1.10) gives

$$|x_n - x_{n-1}| \leq \frac{1-k}{k}\varepsilon, \tag{5.1.11}$$

which can be used to find the difference between two successive approximation (or iterations) to achieve a prescribed accuracy. From (5.1.11), it is clear that the rate of convergence of the fixed point iteration method is linear.

5.2 Modified Newton-Raphson method

It is known that the Newton-Raphson iterative scheme for finding a simple root $x = \xi$ of the equation $f(x) = 0$ is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{5.2.1}$$

We know that the iterative method converges quadratically for a simple root. Now, if ξ is a root of $f(x) = 0$ with multiplicity m, then by modified Newton-Raphson method the iteration formula corresponding to Eq.(5.2.1) is taken as

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)} \tag{5.2.2}$$

which means that $(1/m)f'(x_n)$ is the slope of the straight line passing through (x_n, y_n) and intersection the x-axis at the point $(x_{n+1}, 0)$. Eq.(5.2.2) is called the *modified Newton's formula*. Since ξ is a root of $f(x) = 0$ with multiplicity m, it follows that ξ is also a root of $f'(x) = 0$ with multiplicity $(p-1)$, of $f''(x) = 0$ with multiplicity $(p-2)$, and so on. Hence the expressions

$$x_0 - m\frac{f(x_0)}{f'(x_0)}, \quad x_0 - (m-1)\frac{f'(x_0)}{f''(x_0)}, \quad x_0 - (m-2)\frac{f''(x_0)}{f'''(x_0}$$

must have the same value if there is a root with multiplicity m, provided that the initial approximation x_0 is chosen sufficiently close to the root.

5.2.1 Order of convergence : Simple Root

Consider the Newton-Raphson method (5.2.1) converges to a root ξ of the equation $f(x) = 0$. Let $\varepsilon_n = \xi - x_n$ be the error in n-th approximation, x_n. Then

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{and} \quad \xi - \varepsilon_{n+1} = \xi - \varepsilon_n - \frac{f(\xi - \varepsilon_n)}{f'(\xi - \varepsilon_n)}$$

Now, on using Taylor Series expansion of the function $f(x)$ about the point $x = \xi$, we have

$$\varepsilon_{n+1} = \varepsilon_n + \frac{f(\xi) - \varepsilon_n f'(\xi) + \frac{1}{2!}\varepsilon_n^2 f''(\xi) - \frac{1}{3!}\varepsilon_n^3 f'''(\xi) + \cdots}{f'(\xi) - \varepsilon_n f''(\xi) + \frac{1}{2!}\varepsilon_n^2 f'''(\xi) - \frac{1}{3!}\varepsilon_n^3 f^{iv}(\xi) + \cdots} \tag{5.2.3}$$

If ξ is the simple root (i.e., multiplicity one), then $f(\xi) = 0$ and $f'(\xi) \neq 0$. On dividing the numerator and denominator in Eq.(5.2.3) with $f'(\xi)$, we get

$$\begin{aligned}
\varepsilon_{n+1} &= \varepsilon_n + \frac{-\varepsilon_n + \frac{1}{2!}\varepsilon_n^2 \frac{f''(\xi)}{f'(\xi)} - \frac{1}{3!}\varepsilon_n^3 \frac{f'''(\xi)}{f'(\xi)} + \cdots}{1 - \left(\varepsilon_n \frac{f''(\xi)}{f'(\xi)} - \frac{1}{2!}\varepsilon_n^2 \frac{f'''(\xi)}{f'(\xi)} + \frac{1}{3!}\varepsilon_n^3 \frac{f^{iv}(\xi)}{f'(\xi)} - \cdots\right)} \\
\Rightarrow \varepsilon_{n+1} &= \varepsilon_n + \left[-\varepsilon_n + \frac{1}{2!}\varepsilon_n^2 \frac{f''(\xi)}{f'(\xi)} - \frac{1}{3!}\varepsilon_n^3 \frac{f'''(\xi)}{f'(\xi)} + \cdots\right] \\
&\qquad \cdot \left[1 - \left(\varepsilon_n \frac{f''(\xi)}{f'(\xi)} - \frac{1}{2!}\varepsilon_n^2 \frac{f'''(\xi)}{f'(\xi)} + \frac{1}{3!}\varepsilon_n^3 \frac{f^{iv}(\xi)}{f'(\xi)} - \cdots\right)\right]^{-1}
\end{aligned} \tag{5.2.4}$$

Let $z = \varepsilon_n \dfrac{f''(\xi)}{f'(\xi)} - \dfrac{1}{2!}\varepsilon_n^2 \dfrac{f'''(\xi)}{f'(\xi)} + \dfrac{1}{3!}\varepsilon_n^3 \dfrac{f^{iv}(\xi)}{f'(\xi)} - \cdots$. Since ε_n is the error term and as $\lim\limits_{n\to\infty} \varepsilon_n \to 0$, so we have $z << 1$. On using the expansion $(1-z)^{-1} = 1 + z + z^2 + \cdot$ in the Eq.(5.2.4), we obtain

$$\varepsilon_{n+1} = \varepsilon_n + \left[-\varepsilon_n + \frac{\varepsilon_n^2}{2!}\frac{f''(\xi)}{f'(\xi)} + O(\varepsilon_n^3)\right]\left[1 + \varepsilon_n \frac{f''(\xi)}{f'(\xi)} + O(\varepsilon_n^2)\right]$$
$$\Rightarrow \varepsilon_{n+1} = -\frac{\varepsilon_n^2}{2}\frac{f''(\xi)}{f'(\xi)} + O(\varepsilon_n^3) \Rightarrow \lim_{n\to\infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon|^2} = \left|\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right|$$

This imply that, the order of convergence of Newton-Raphson method is 2 (quadratic convergence).

5.2.2 Order of convergence : Multiple Root

In the case of multiple roots of order m, the Newton-Raphson method has convergence as follows. Continuing with Eq.(5.2.3), we have

$$\varepsilon_{n+1} = \varepsilon_n + \frac{f(\xi) - \varepsilon_n f'(\xi) + \frac{1}{2!}\varepsilon_n^2 f''(\xi) - \frac{1}{3!}\varepsilon_n^3 f'''(\xi) + \cdots}{f'(\xi) - \varepsilon_n f''(\xi) + \frac{1}{2!}\varepsilon_n^2 f'''(\xi) - \frac{1}{3!}\varepsilon_n^3 f^{iv}(\xi) + \cdots}$$

Consider the equation $f(x) = 0$ has multiple root ξ of order m, then $f'(\xi) = f''(\xi) = \cdots = f^{m-1}(\xi) = 0$ and $f^m(\xi) \neq 0$. So the above equation reduces to the following equation

$$\varepsilon_{n+1} = \varepsilon_n + \frac{\frac{(-1)^m \varepsilon_n^m}{m!} f^{(m)}(\xi) + \frac{(-1)^{m+1}\varepsilon_n^{m+1}}{(m+1)!} f^{(m+1)}(\xi) + \frac{(-1)^{m+2}\varepsilon_n^{m+2}}{(m+2)!} f^{(m+2)}(\xi) + \cdots}{\frac{(-1)^{m-1}\varepsilon_n^{m-1}}{(m-1)!} f^{(m)}(\xi) + \frac{(-1)^m \varepsilon_n^m}{m!} f^{(m+1)}(\xi) + \frac{(-1)^{m+1}\varepsilon_n^{m+1}}{(m+1)!} f^{(m+2)}(\xi) \cdots}$$

On dividing the numerator and denominator by $\dfrac{(-1)^{m-1}\varepsilon_n^{m-1}}{(m-1)!} f^{(m)}(\xi)$, we have

$$\varepsilon_{n+1} = \varepsilon_n + \frac{-\frac{\varepsilon_n^m}{m} + \frac{\varepsilon_n^2}{m(m+1)}\frac{f^{(m+1)}(\xi)}{f^{(m)}(\xi)} - \frac{\varepsilon_n^3}{m(m+1)(m+2)}\frac{f^{(m+2)}(\xi)}{f^{(m)}(\xi)} + \cdots}{1 - \left(\frac{\varepsilon_n}{m}\frac{f^{(m+1)}(\xi)}{f^{(m)}(\xi)} - \frac{\varepsilon_n^2}{m(m+1)}\frac{f^{(m+2)}(\xi)}{f^{(m)}} + \frac{\varepsilon_n^3}{m(m+1)(m+2)}\frac{f^{(m+3)}(\xi)}{f^{(m)}(\xi)} - \cdots\right)}$$

On using the expansion, $(1-z)^{-1} = 1 + z + z^2 + \cdots$, the above expression can be rewritten as

$$\varepsilon_{n+1} = \varepsilon_n \left(1 - \frac{1}{m}\right) - \frac{\varepsilon_n^2}{m^2(m+1)}\frac{f^{m+1}(\xi)}{f^{(m)}(\xi)} + O(\varepsilon_n^3)$$

If $m = 1$ (i.e., ξ is only a simple root) then the coefficient of ε_n is zero and coefficient of ε_n^2 is not equal to zero and hence the scheme is of second order.

If $m \neq= 1$ then the coefficient of ε_n itself is not equal to zero and hence the scheme is only of first order.

Example 5.1. Find a double root of the equation $f(x) = x^3 - x^2 - x + 1 = 0$.

Solution: Choosing $x_0 = 0.8$, we have

$$f'(x) = 3x^3 - 2x - 1, \quad \text{and} \quad f''(x) = 6x - 2.$$

With $x_0 = 0.8$, we obtain

$$x_0 - 2\frac{f(x_0)}{f'(x_0)} = 0.8 - 2\frac{0.072}{-0.68} = 1.012 \quad \text{and} \quad x_0 - \frac{f'(x_0)}{f''(x_0)} = 0.8 - \frac{-0.68}{2.8} = 1.043$$

The closeness of these values indicates that there is a double root near to unity. For the next approximation, we choose $x_1 = 1.01$ and obtain

$$x_1 - 2\frac{f(x_1)}{f'(x_1)} = 1.01 - 0.0099 = 1.0001 \quad \text{and} \quad x_1 - \frac{f'(x_1)}{f''(x_1)} = 1.01 - 0.0099 = 1.0001$$

We conclude, therefore, that there is a double root at $x = 1.0001$ which is sufficiently close to the actual root unity.

5.3 Secant Method

We have seen that the Newton-Raphson method requires the evaluation of derivatives of the function and this is not always possible, particularly in case of functions arising in practical problems. In the secant method, the derivative at x_i is approximated by the formula

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1}}{x_i - x_{i-1}},$$

which can be written as

$$f_i' = \frac{f_i - f_{i-1}}{x_i - x_{i-1}}, \tag{5.3.1}$$

where $f_i = f(x_i)$. Hence, the Newton-Raphson formula becomes

$$x_{i+1} = x_i - \frac{f_i\,(x_i - x_{i-1})}{f_i - f_{i-1}} = \frac{x_{i-1}f_i - x_i f_{i-1}}{f_i - f_{i-1}}. \tag{5.3.2}$$

It should be noted that this formula requires two initial approximation to the root. This method converges super-linearly. The order of converges of this method is $p = \frac{1}{2}(1 + \sqrt{5}) = 1.618$ **(!!! Prove it !!!)**.

Example 5.2. Using the secant method, find a real root of the equation

$$f(x) = xe^x - 1 = 0$$

Solution: We have $f(0) = -1$ and $f(1) = e - 1 = 1.71828$. Therefore, a root lies between 0 and 1. Let $x_0 = 0$ and $x_1 = 1$. Therefore

$$x_2 = \frac{x_0 f_1 - x_1 f_0}{f_1 - f_0} = \frac{1}{2.71828} = 0.36788$$

and

$$f_2 = 0.36788\, e^{0.36788} - 1 = -0.46854$$

Hence

$$x_3 = \frac{x_1 f_2 - x_2 f_1}{f_2 - f_1} = \frac{1(-0.46854) - 0.36788(1.71828)}{-0.46854 - 1.17828} = 0.50332$$

and

$$f_3 = -0.16740$$

Hence

$$x_4 = \frac{x_2 f_3 - x_3 f_2}{f_3 - f_2} = 0.57861 \qquad \text{and} \qquad f_4 = 0.03198$$

Hence

$$x_5 = \frac{x_3 f_4 - x_4 f_3}{f_4 - f_3} = 0.56653 \quad \text{and} \quad f_5 = -0.00169$$

Therefore,

$$x_6 = \frac{x_4 f_5 - x_5 f_4}{f_5 - f_4} = 0.56714 \quad \text{and} \quad f(x_6) = -0.0001196$$

It follows that the required root is 0.5671, correct to four decimal places.

5.4 Muller's Method

In this method, the given function $f(x)$ is approximated by a second degree curve in the vicinity of a root. The roots of the quadratic are then assumed to be the approximations to the roots of the equation $f(x) = 0$. The method is iterative and can be used to compute complex roots. It has quadratic convergence.

Let $(x_{i-2}, y_{i-2}),\ (x_{i-1}, y_{i-1})$ and (x_i, y_i) b three distinct points on the curve $y = f(x)$ where x_{i-2}, x_{i-1} and x_i are approximations to a root of $f(x) = 0$. Now, a second degree curve passing through the three points is given by Lagrange's formula

$$L(x) = \frac{(x - x_{i-1})(x - x_i)}{(x_{i-2} - x_{i-1})(x_{i-2} - x_i)} y_{i-2} + \frac{(x - x_{i-2})(x - x_i)}{(x_{i-1} - x_{i-2})(x_{i-1} - x_i)} y_{i-1} + \frac{(x - x_{i-2})(x - x_{i-1})}{(x_i - x_{i-2})(x_i - x_{i-1})} y_i \tag{5.4.1}$$

Let $h_i = x_i - x_{i-1}$, $h_{i-1} = x_{i-1} - x_{i-2}$. Then

$$\begin{aligned} x - x_{i-1} &= x - x_i + x_i - x_{i-1} = (x - x_i) + h_i, \\ x - x_{i-2} &= x - x_i + x_i - x_{i-2} = (x - x_i) + (h_{i+1} + h_i), \\ x_{i-2} - x_{i-1} &= -h_{i-1} \\ x_{i-2} - x_i &= -(h_{i-1} + h_i) \quad \text{and} \quad \Delta_i = y_i - y_{i-1} \end{aligned} \tag{5.4.2}$$

Hence

$$L(x) = \frac{(x - x_i + h_i)(x - x_i)}{h_{i-1}(h_{i-1} + h_i)} y_{i-2} + \frac{(x - x_i + h_{i-1} + h_i)(x - x_i)}{-h_{i-1} h_i} y_{i-1} + \frac{(x - x_i + h_i + h_{i-1})(x - x_i + h_i)}{h_i(h_{i-1} + h_i)} y_i \tag{5.4.3}$$

After simplification, the preceding equation can be written as

$$L(x) = A(x - x_i)^2 + B(x - x_i) + y_i,$$

where $A = \dfrac{1}{h_{i-1} + h_i} \left(\dfrac{\Delta_i}{h_i} - \dfrac{\Delta_{i-1}}{h_{i-1}} \right)$ and $B = \dfrac{\Delta_i}{h_i} + A h_i$. With these values of A and B, the quadratic Eq.(5.4.1) gives the next approximation x_{i-1}

$$x_{i+1} = x_i + \frac{-B \pm \sqrt{B^2 - 4Ay_i}}{2A} \tag{5.4.4}$$

Since Eq.(5.4.4) leads to inaccurate results, we take the equivalent form

$$x_{i+1} = x_i - \frac{2y_i}{B \pm \sqrt{B^2 - 4Ay_i}} \tag{5.4.5}$$

In Eq.(5.4.5), the sign in the denominator should be chosen so that the denominator will be largest in magnitude. With this choice, Eq.(5.4.5) gives the next approximation to the root.

Example 5.3. Using Muller's method, find the root of the equation

$$f(x) = x^3 - x - 1 = 0$$

with the initial approximations $x_{i-2} = 0, x_{i-1} = 1, x_i = 2$.

Solution: We have $y_{i-2} = -1,\ y_{i-1} = -1,\ y_i = 5$. Also, $h_i = 1, h_{i-1} = 1, \Delta_i = 6, \Delta_{i-1} = 0$. Hence we obtain

$$A = 3 \quad \text{and} \quad B = 9$$

Then $\sqrt{B^2 - 4Ay_i} = \sqrt{21}$. Therefore, Eq.(5.4.5) gives

$$\begin{aligned} x_{i+1} &= 2 - \frac{2(5)}{9 + \sqrt{21}}, \text{ since the sign of B is positive} \\ &= 1.26376 \end{aligned}$$

Error in the above result $= \left| \frac{1.26376 - 2}{1.26376} \right| \times 100 = 58\%$.

For the second approximation, we take

$$x_{i-2} = 1, \;\; x_{i-1} = 2, \;\; x_i = 1.26376.$$

The corresponding values of y are

$$y_{i-2} = -1, \quad y_{i-1} = 5, \quad y_i = -0.24542$$

The computed values of A and B are

$$A = 4.26375 \quad \text{and} \quad B = 3.98546$$

Then

$$x_{i+1} = 1.32174,$$

and the error in the above result is equal to 4.39%.

For the third approximation, we take

$$\begin{aligned} x_{i-2} = 2, \;\; x_{i-1} = 1.26376, \;\; x_i = 1.32174. \\ y_{i-2} = 5, \;\; y_{i-1} = -0.24542, \;\; y_i = -0.01266. \end{aligned}$$

Then $A = 4.58544$, $B = 4.28035$ and $x_{i+1} = 1.32469$. Error in the result is equal to 0.22%. For the next approximation, we have

$$x_{i-2} = 1.26376, \;\; x_{i-1} = 1.32174, \;\; x_i = 1.32469$$

These values gives

$$A = 3.87920, \;\; B = 4.26229, \;\; \text{and} \;\; x_{i+1} = 1.32472.$$

The error in this result is equal to 0.002%. Hence the required root is 1.3247, correct to 4 decimal places.

Exercise 5.4. (i) Find a real root of the equation $x^3 = 1 - x^2$ on the interval $[0, 1]$ with an accuracy of 10^{-4}.

(ii) Describe briefly Muller's method and use it to find (a) the root, between 2 and 3, of the equation $x^3 - 2x - 5 = 0$ and (b) the root, between 0 and 1, of the equation $x = e^{-x} \cos x$.
Answer: (a) 2.09462409 (b) 0.51752

Unit 6

Course Structure

Ordinary Differential Equations: Initial value problems – Picard's successive approximation method, error estimation. Single-step methods – Euler's method and Runge-Kutta method, error estimations and convergence analysis.

6 Introduction

Many problems in science and engineering can be reduced to the problem of solving differential equations satisfying certain given conditions. The analytical methods of solution, with which the reader is assumed to be familiar, can be applied to solve only a selected class of differential equations. Those equations which govern physical systems do not possess, in general closed-form solutions, and hence recourse must be made to numerical methods for solving such differential equations. To describe various numerical methods for the solution of ordinary differential equations, we consider the general first order differential equation

$$\frac{dy}{dx} = f(x, y) \text{ with the initial condition } y(x_0) = y_0 \tag{6.0.1}$$

and illustrate the theory with respect to this equation. This methods so developed can, in general, be applied to the solution of systems of first-order equations.

6.1 Picard's Successive Approximation Method

Integrating the differential equation given in Eq.(6.0.1), we obtain

$$y = y_0 + \int_{x_0}^{x} f(x, y)\, dx. \tag{6.1.1}$$

Equation (6.1.1), in which the unknown function y appears under the integral sign, is called an *integral equation*. Such and equation can be solved by the method of successive approximations in which the first approximation of y is obtained by putting y_0 for y on right side of Eq.(6.1.1), and we write

$$y^{(1)} = y_0 + \int_{x_0}^{x} f(x, x_0)\, dx$$

The integral on the right can now be solved and the resulting $y^{(1)}$ is substituted for y in the integrand of Eq.(6.1.1) to obtain the second approximation $y^{(2)}$:

$$y^{(2)} = y_0 + \int_{x_0}^{x} f(x, y^{(1)})\, dx$$

Proceeding in this way, we obtain $y^{(3)},\ y^{(4)}, \ldots, y^{(n-1)}$ and $y^{(n)}$, where

$$y^{(n)} = y_0 + \int_{x_0}^{x} f(x, y^{(n-1)})\, dx \quad \text{with} \quad y^{(0)} = y_0 \tag{6.1.2}$$

Hence this method yields a sequence of approximations $y^{(1)},\ y^{(2)}, \ldots, y^{(n)}$ and it can be proved that if the function $f(x, y)$ is bounded in some region about the point (x_0, y_0) and if $f(x, y)$ satisfies the *Lipschitz condition*, viz

$$|f(x, y) - f(x, \overline{y})| \leq K|y - \overline{y}|, \quad K \text{ being a constant} \tag{6.1.3}$$

then, the sequence $y^{(1)},\ y^{(2)}, \ldots$ converges to the solution of Eq.(6.0.1).

Example 6.1. Solve the differential equation $\dfrac{dy}{dx} = x + y^2$ with initial condition $y = 1$ when $x = 0$ using Picard's method.

Solution: We start with $y^{(0)} = 1$ and obtain

$$y^{(1)} = 1 + \int_0^x (x + 1)\, dx = 1 + x + \frac{1}{2}x^2.$$

Then the second approximation is

$$\begin{aligned} y^{(2)} &= 1 + \int_0^x \left[x + \left(1 + x + \frac{1}{2}x^2\right)\right] \\ &= 1 + x + \frac{3}{2}x^2 + \frac{2}{3}x^3 + \frac{1}{4}x^4 + \frac{1}{20}x^5. \end{aligned}$$

Proceeding similarly, we can find the higher order approximations. But, it is obvious that the integrations might become more and more difficult as we proceed to higher approximations.

Example 6.2. Given the differential equation $\dfrac{dy}{dx} = \dfrac{x^2}{y^2 + 1}$ with initial condition $y = 0$ when $x = 0$, use Picard's method to obtain y for $x = 0.25,\ 0.5$ and 1.0 correct to three decimal places.

Solution: We have $y = \displaystyle\int_0^x \frac{x^2}{y^2 + 1} dx$. Setting $y^{(0)} = 0$, we obtain

$$\begin{aligned} y^{(1)} &= \int_0^x x^2\, dx = \frac{1}{3}x^3 \\ \text{and } y^{(2)} &= \int_0^x \frac{x^2}{(1/9)x^6 + 1} dx = \tan^{-1}\left(\frac{1}{3}x^3\right) = \frac{1}{3}x^3 - \frac{1}{81}x^9 + \ldots \end{aligned}$$

so that $y^{(1)}$ and $y^{(2)}$ agree to the first term, viz., $(1/3)x^3$. To find the range of values of x so that the series with the term $(1/3)x^3$ alone will give the result correct to three decimal places, we put

$$\frac{1}{81}x^9 \leq 0.0005 \quad \text{which yields} \quad x \leq 0.7$$

Hence

$$y(0.25) = \frac{1}{3}(0.25)^3 = 0.005, \quad y(0.5) = \frac{1}{3}(0.5)^3 = 0.042, \quad y(1.0) = \frac{1}{3} - \frac{1}{81} = 0.321$$

Exercise 6.3. (i) Use Picard's method to obtain a series solution the differential equation $\frac{dy}{dx} = 1+xy,\ y(0) = 1$. **Answer:** $y(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{8} + \ldots$
(ii) Use Picard's method to obtain $y(0.1)$ and $y(0.2)$ of the problem defined by

$$\frac{dy}{dx} = x + yx^4, \quad y(0) = 3 \qquad \textbf{Answer}: \ 3.005, \ \ 3.0202$$

(iii) Using Picard's method, find $y(0.1)$, given that

$$\frac{dy}{dx} = \frac{y-x}{y+x}; \quad y(0) = 1 \qquad \textbf{Answer}: \ 1.0906$$

6.2 Single Step Methods

6.2.1 Euler's Method

We have so far discussed the Picard's method which yield solution of differential equation in the form of a power series. We will now describe the methods which give the solution in the form of a set of tabulated values.

Suppose that we wish to solve Eq.(6.0.1) for values of y at $x = x_r = x_0 + rh$ $(r = 1, 2, \ldots)$. Integrating Eq.(6.0.1), we obtain

$$y_1 = y_0 + \int_{x_0}^{x_1} f(x, y)\, dx. \tag{6.2.1}$$

Assuming that $f(x, y) = f(x_0, y_0)$ in $x_0 \leq x \leq x_1$, this gives Euler's formula

$$y_1 \approx y_0 + h\, f(x_0, y_0). \tag{6.2.2}$$

Similarly for the range $x_1 \leq x \leq x_2$, we have

$$y_2 = y_1 + \int_{x_1}^{x_2} f(x, y)\, dx.$$

Substituting $f(x_1, y_1)$ for $f(x, y)$ in $x_1 \leq x \leq x_2$, we obtain

$$y_2 \approx y_1 + h\, f(x_1, y_1). \tag{6.2.3}$$

Proceeding in this way, we obtain the general formula

$$y_{n+1} = y_n + h\, f(x_n, y_n), \quad n = 0, 1, 2, \ldots \tag{6.2.4}$$

The process is very slow and to obtain reasonable accuracy with Euler's method, we need to take a smaller value of h. Because of this restriction on h, the method is unsuitable for practical use and modification of it, known as the *modified Euler method*, which gives more accurate results, will be described in the following subsection.

Example 6.4. Find the value of $y(x)$ at $x = 0.04$ for the differential equation $y' = -y$ with the condition $y(0) = 1$ using Euler's method.

Solution: Successive application of Eq.(6.2.4) with $h = 0.01$ gives

$$\begin{aligned} y(0.01) &= 1 + 0.001(-1) = 0.99 \\ y(0.02) &= 0.99 + 0.01(-0.99) = 0.9801 \\ y(0.03) &= 0.9801 + 0.01(-0.9801) = 0.9703 \\ y(0.04) &= 0.9703 + 0.01(-0.9703) = 0.9606. \end{aligned}$$

The exact solution is $y = e^{-x}$ and from this the value at $x = 0.04$ is 0.9608.

6.2.2 Error Estimation for the Euler Method

Let the true solution of the differential equation at $x = x_n$ be $y(x_n)$ and also let the approximate solution be y_n. Now, expanding $y(x_{n+1})$ by Taylor's series, we get

$$\begin{aligned} y(x_{n+1}) = y(x_n + h) &= y(x_n) + h\, y'(x_n) + \frac{h^2}{2} y''(x_n) + \ldots \\ &= y(x_n) + hy'(x_n) + \frac{h^2}{2} y''(\tau_n), \quad \text{where} \quad x_n \le \tau_n \le x_{n+1}. \end{aligned} \tag{6.2.5}$$

We usually encounter two type of errors in the solution of differential equations. These are (i) local errors, and (ii) rounding errors. The local error is the result of replacing the given differential equation by means of the equation

$$y_{n+1} = y_n + hy'_n.$$

This error is given by

$$L_{n+1} = -\frac{1}{2} h^2 y''(\tau_n) \tag{6.2.6}$$

The total error is then defined by

$$e_n = y_n - y(x_n) \tag{6.2.7}$$

Since y_0 is exact, it follows that $e_0 = 0$. Neglecting the rounding error, we write the total solution error as

$$\begin{aligned} e_{n+1} &= y_{n+1} - y(x_{n+1}) \\ &= y_n + hy'_n - [y(x_n) + hy'(x_n) - L_{n+1}] \\ &= e_n + h[f(x_n, y_n) - y'(x_n)] + L_{n+1}. \\ &= e_n + h\Big[f(x_n, y_n) - f(x_n, y(x_n))\Big] + L_{n+1} \end{aligned}$$

By mean value theorem, we write

$$f(x_n, y_n) - f(x_n, y(x_n)) = [y_n - y(x_n)] \frac{\partial f}{\partial y}(x_n, \xi_n), \quad y(x_n) \le \xi_n \le y_n.$$

Hence, we have

$$e_{n+1} = e_n \Big[1 + h f_y(x_n, \xi_n)\Big] + L_{n+1} \tag{6.2.8}$$

Since $e_0 = 0$, we obtain successively:

$$\begin{aligned} &e_1 = L_1; \;\; e_2 = \Big[1 + h f_y(x_1, \xi_1)\Big] L_1 + L_2; \\ &e_3 = \Big[1 + h f_y(x_2, \xi_2)\Big] \Big[1 + h f_y(x_1, \xi_1)\Big] (L_1 + L_2) + L_3; \quad \text{etc.} \end{aligned}$$

6.3 Modified Euler's Method

Instead of approximating $f(x, y)$ by $f(x_0, y_0)$ in Eq.(6.2.1), we now approximate the integral given in Eq.(6.2.1) by means of trapezoidal rule to obtain

$$y_1 = y_0 + \frac{h}{2}[f(x_0, y_0) + f(x_1, y_1)] \tag{6.3.1}$$

We thus obtain the iteration formula

$$y_1^{(n+1)} = y_0 + \frac{h}{2}\Big[f(x_0, y_0) + f(x_1, y_1^{(n)})\Big], \;\; n = 0, 1, 2, \ldots \tag{6.3.2}$$

where $y_1^{(n)}$ is the n-th approximation to y_1. The iteration formula (6.3.2) can be started by choosing $y_1^{(0)}$ from Euler's formula:

$$y_1^{(0)} = y_0 + h\, f(x_0, y_0).$$

Example 6.5. Determine the value of y when $x = 0.1$ given that

$$y' = x^2 + y; \quad y(0) = 1$$

Solution: We take $h = 0.05$. With $x_0 = 0$ and $y_0 = 1.0$, we have $f(x_0, y_0) = 1.0$. Hence Euler's formula gives

$$y_1^{(0)} = 1 + 0.05(1) = 1.05$$

Further, $x_1 = 0.05$ and $f(x_1, y_1^{(0)}) = 1.0525$. The average of $f(x_0, y_0)$ and $f(x_1, y_1^{(0)})$ is 1.0262. The value of $y_1^{(1)}$ can therefore be computed by using Eq.(6.3.2) and we obtain

$$y_1^{(1)} = 1.0513$$

Repeating the procedure, we obtain $y_1^{(2)} = 1.0513$. Hence we take $y_1 = 1.0513$, which is correct to four decimal places. Next, with $x_1 = 0.05$, $y_1 = 1.0513$ and $h = 0.05$, we continue the procedure to obtain y_2, i.e., the value of y when $x = 0.1$. The results are

$$y_2^{(0)} = 1.1040, \quad y_2^{(1)} = 1.1055, \quad y_2^{(2)} = 1.1055.$$

Hence, we conclude that the value of y when $x = 0.1$ is 1.1055.

Exercise 6.6. (i) Given the initial value problem $y' = 2x + \cos y,\; y(0) = 1$. Show that it is sufficient to use Euler method with step length $h = 0.2$ to compute $y(0.2)$ with an error less than 0.05.
(ii) Find an approximation to $y(1.6)$ for the initial value problem $y' = x + y^2,\; y(1) = 1$ using the Euler method with $h = 0.1$ and $h = 0.2$. **Answer:** h=0.1: 3.848948; h=0.2: 3.137805
(iii) Given the differential equation $\dfrac{dy}{dx} = x^2 + y,\; y(0) = 1$, compute $y(0.02)$ using Euler's modified method. **Answer:** 1.0202
(iv) Solve, by Euler's modified method, the problem $displaystyle\frac{dy}{dx} = x + y,\; y(0) = 0$ Choose h=0.2 and compute $y(0.2)$ and $y(0.4)$. **Answer:** 0.0222, 0.0938

6.3.1 Runge-Kutta Methods

As already mentioned, Euler's method is less efficient in practical problems since it requires h to be small for obtaining reasonable accuracy. The Runge-Kutta methods are designed to give greater accuracy and they possess the advantage of requiring only the function values at some selected points on the subinterval.

If we substitute $y_1 = y_0 + h\ f(x_0, y_0)$ on the right side of Eq.(6.3.1), we obtain

$$y_1 = y_0 + \frac{h}{2}\Big[f_0 + f(x_0 + h, y_0 + hf_0\Big], \tag{6.3.3}$$

where $f_0 = f(x_0, y_0)$. If we now set

$$k_1 = h\ f_0 \quad \text{and} \quad k_2 = h\ f(x_0 + h, y_0 + k_1)$$

then the above equation becomes

$$y_1 = y_0 + \frac{1}{2}(k_1 + k_2), \tag{6.3.4}$$

which is the *second-order Runge-Kutta* formula. The error in this formula can be shown to be of order h^3 by expanding both sides by Taylor's series. Thus, the left side gives

$$y_0 + hy_0' + \frac{h^2}{2}y_0'' + \frac{h^2}{6}y_0''' + \ldots$$

and on the right side

$$k_2 = h\ f(x_0 + h, y_0 + hf_0) = h\left[f_0 + h\frac{\partial f}{\partial x_0} + hf_0\frac{\partial f}{\partial y_0} + O(h^2)\right].$$

Now, since $\dfrac{df(x,y)}{dx} = \dfrac{\partial f}{\partial x} + f\dfrac{\partial f}{\partial y}$, we obtain $k_2 = h[f_0 + hf_0' + O(h^2)] = hf_0 + h^2 f_0' + O(h^3)$, so that the right side of Eq.(6.3.4) gives

$$y_0 + \frac{1}{2}\Big[hf_0 + hf_0 + h^2 f_0' + O(h^3)\Big] = y_0 + hf_0 + \frac{1}{2}h^2 f_0' + O(h^3) = y_0 + hy_0' + \frac{h^2}{2}y_0'' + O(h^3).$$

It therefore follows that the Taylor series expansions of both sides of Eq.(6.3.4) agree up to terms of order h^2, which means that the error in this formula is of order h^3.

More generally, if we set

$$y_1 = y_0 + W_1k_1 + W_2k_2 \tag{6.3.5}$$

where $k_1 = h\ f_0$, $k_2 = h\ f(x_0 + \alpha_0 h, y_0 + \beta_0 k_1)$, then the Taylor series expansions gives

$$y_0 + hf_0 + \frac{h^2}{2}\left(\frac{\partial f}{\partial x} + f_0\frac{\partial f}{\partial y}\right) + O(h^3) = y_0 + (W_1 + W_2)hf_0 + W_2h^2\left(\alpha_0\frac{\partial f}{\partial x} + \beta_0 f_0\frac{\partial f}{\partial y}\right) + O(h^3).$$

Equating the coefficient of $f(x, y)$ and its derivatives on both the sides, we obtain the relation

$$W_1 + W_2 = 1, \quad W_2\alpha_0 = \frac{1}{2}, \quad W_2\beta_0 = \frac{1}{2}. \tag{6.3.6}$$

Clearly, $\alpha_0 = \beta_0$ and if α_0 is assigned any value arbitrarily, then the remaining parameter can be determined uniquely. If we set, for example $\alpha_0 = \beta_0 = 1$, then we immediately obtain $W_1 = W_2 = 1/2$, which gives formula (6.3.4).

It follows, therefore, that there are several second-order Runge-Kutta formulae and that formulae (6.3.5) and (6.3.6) constitute just one of several such formulae.

Higher-order Runge-Kutta formulae exist, of which we mention only the *fourth-order formula* defined by

$$y_1 = y_0 + W_1 k_1 + W_2 k_2 + W_3 k_3 + W_4 k_4 \tag{6.3.7}$$

where

$$\begin{aligned} k_1 &= hf(x_0, y_0) \\ k_2 &= hf(x_0 + \alpha_0 h, y_0 + \beta_0 + k_1) \\ k_3 &= hf(x_0 + \alpha_1 h, y_0 + \beta_1 k_1 + \nu_1 k_2) \\ k_4 &= hf(x_0 + \alpha_2 h, y_0 + \beta_2 k_1 + \nu_2 k_2 + \delta_1 k_3), \end{aligned} \tag{6.3.8}$$

where the parameters have to be determined by expanding both sides of the Eq.(6.3.7) by Taylor's series and securing agreement of terms up to and including those containing h^4. The choice of the parameters is, again, arbitrary and we have therefore several fourth-order Runge-Kutta formulae. If, for example, we set

$$\begin{aligned} &\alpha_0 = \beta_0 = \frac{1}{2}, && \alpha_1 = \frac{1}{2}, && \alpha_2 = 1, \\ &\beta_1 = \frac{1}{2}(\sqrt{2} - 1), && \beta_2 = 0, \\ &\nu_1 = 1 - \frac{1}{\sqrt{2}}, && \nu_2 = -\frac{1}{\sqrt{2}}, && \delta_1 = 1 + \frac{1}{\sqrt{2}}, \\ &W_1 = W_4 = \frac{1}{6}, && W_2 = \frac{1}{3}\left(1 - \frac{1}{\sqrt{2}}\right), && W_3 = \frac{1}{3}\left(1 + \frac{1}{\sqrt{2}}\right), \end{aligned} \tag{6.3.9}$$

we obtain the method of Gill, Whereas the choice

$$\begin{aligned} &\alpha_0 = \alpha_1 = \frac{1}{2}, && \beta_0 = \nu_1 = \frac{1}{2} \\ &\beta_1 = \beta_2 = \nu_2 = 0, && \alpha_2 = \delta_1 = 1 \\ &W_1 = W_4 = \frac{1}{6}, && W_2 = W_3 = \frac{2}{6} \end{aligned} \tag{6.3.10}$$

leads to the fourth-order Runge-Kutta formula, the most commonly used one in practice:

$$y_1 = y_0 + \frac{1}{6}\Big[k_1 + 2k_2 + 2k_3 + k_4\Big]$$

where

$$\begin{aligned} k_1 &= hf(x_0, y_0), \\ k_2 &= hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right), \\ k_3 &= hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2\right), \\ k_4 &= hf(x_0 + h, y_0 + k_3). \end{aligned}$$

in which the error is of order h^5. Complete derivation of this formula is exceedingly complicated, and the interested reader may referred to the book by Levy and Baggot. We illustrate here the use of the fourth-order formula by means of examples.

Example 6.7. Given $\dfrac{dy}{dx} = y - x$ where $y(0) = 2$, find $y(0.1)$ and $y(0.2)$ correct to four decimal places.

Solution: (i) *Runge-Kutta second order formula*: With $h = 0.1$, we find $k_1 = 0.2$ and $k_2 = 0.21$. Hence

$$y_1 = y(0.1) = 2 + \frac{1}{2}(0.41) = 2.2050.$$

To determine $y_2 = y(0.2)$, we note that $x_0 = 0.1$ and $y_0 = 2.2050$. Hence, $k_1 = 0.1(2.015)$ and $k_2 = 0.1(2.4155 - 0.2) = 0.22155$. It follows that

$$y_2 = 2.2050 + \frac{1}{2}(0.2105 + 0.22155) = 2.4210.$$

Proceeding in a similar way, we obtain

$$y_3 = y(0.3) = 2.6492 \quad \text{and} \quad y_4 = y(0.4) = 2.8990$$

We next choose $h = 0.2$ and compute $y(0.2)$ and $y(0.4)$ directly. With $h = 0.2$, $x_0 = 0$ and $y_0 = 2$, we obtain $k_1 = 0.4$ and $k_2 = 0.44$ and hence $y(0.2) = 2.4200$. Similarly, we obtain $y(0.4) = 2.8880$.

From the analytical solution $y = x + 1 + e^x$, the exact value of $y(0.2)$ and $y(0.4)$ are respectively 2.4214 and 2.8918.

(ii) *Runge-Kutta fourth-order formula:* To determine $y(0.1)$, we have $x_0 = 0,\ y_0 = 2$ and $h = 0.1$. We then obtain

$$k_1 = 0.2, \quad k_2 = 0.205, \quad k_3 = 0.20525, \quad k_4 = 0.21053.$$

Hence

$$y(0.1) = 2 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 2.2052.$$

Proceeding similarly, we obtain $y(0.2) = 2.4214$.

Example 6.8. Given $\dfrac{dy}{dx} = 1 + y^2$ where $y(0) = 0$, find $y(0.4)$ and $y(0.6)$ using fourth order Runge-Kutta method.

Solution: We take $h = 0.2$. With $x_0 = y_0 = 0$, we obtain

$$\begin{aligned} k_1 &= 0.2, \\ k_2 &= 0.2(1.01) = 0.202 \\ k_3 &= 0.2(1 + 0.010201) = 0.20204 \\ k_4 &= 0.2(1 + 0.040820) = 0.20816 \end{aligned}$$

and finally

$$y(0.2) = 0 + \frac{1}{6}\Big[k_1 + 2k_2 + 2k_3 + k_4\Big] = 0.2027$$

which is correct to four decimal places.

To compute $y(0.4)$, we take $x_0 = 0.2$, $y_0 = 0.2027$ and $h = 0.2$. With these values, we can evaluate

$$\begin{aligned} k_1 &= 0.2[1 + (0.2027)^2] = 0.2082, \\ k_2 &= 0.2[1 + (0.3068)^2] = 0.2188, \\ k_3 &= 0.2[1 + (0.3121)^2] = 0.2195, \\ k_4 &= 0.2[1 + (0.4222)^2] = 0.2356, \end{aligned}$$

and finally

$$y(0.4) = 0.2027 + 0.2201 = 04228$$

correct to four decimal places.

Finally, taking $x_0 = 0.4$, $y_0 = 0.4228$ and $h = 0.2$, and proceeding as above, we obtain $y(0.6) = 0.6841$.

Exercise 6.9. (i) Given the problem $\dfrac{dy}{dx} = f(x, y)$ and $y(x_0) = y_0$, an approximate solution at $x = x_0 + h$ is given by third order Runge-Kutta formula

$$y(x_0 + h) = y_0 + \frac{1}{6}\Big[k_1 + 4k_2 + k_3\Big] + R_4,$$

where $k_1 = hf(x_0, y_0)$, $k_2 = hf\left(x_0 + \dfrac{1}{2}h, y_0 + \dfrac{1}{2}k_1\right)$ and $k_3 = hf(x_0 + h, y_0 + 2k_2 - k_1)$. Show that R_4 is of order h^4.
(ii) Use Runge-Kutta fourth order formula to find $y(0.2)$ and $y(0.4)$ given that

$$\frac{dy}{dx} = \frac{y^2 - x^2}{y^2 + x^2}; \;\; y(0) = 1$$

Answer: 0.19598, 1.3751
(iii) Find an approximate value of y when $x = 0.2$ and $x = 0.4$ given that $y' = x + y, \; y(0) = 1$, with $h = 0.2$ **Answer:** 1.2428, 1.583636
(iv) Determine $y(0.2)$ with $h = 0.1$, for the initial value problem $y' = x^2 + y^2, \; y(0) = 1$. **Answer:** 1.253015

Unit 7

Course Structure

Ordinary Differential Equations: Multi-step method – Milne's predictor-corrector method, error estimation and convergence analysis.

7 Introduction

Milne's Predictor-corrector is a multi-step method, i.e., to compute y_{n+1} a knowledge of preceding values of y and y' is essentially required. These values of y to be computed by one of the self starting methods viz. Euler's method, Runge-Kutta Method. W.E. Milne uses two types of quadrature formulae, (i) an open-type quadrature formula to derive the Predictor formula and (ii) a closed-type quadrature formula to derive the corrector formula.

7.1 Milne's Predictor-Corrector Method

Let us assume that the values of y and y' are known (given or computed by the self-starting method) for $x_{n-2},\ x_{n-1},\ x_n$ and the initial value x_{n-3}. We have the Newton's forward formula in terms of $y'[= f(x, y(x))]$ and phase u with starting node point x_{n-3} as:

$$y' = y'_{n-3} + u \cdot \Delta y'_{n-3} + \frac{u(u-1)}{2!} \cdot \Delta^2 y'_{n-3} + \frac{u(u-1)(u-2)}{3!} \cdot \Delta^2 y'_{n-2} + \frac{u(u-1)(u-2)(u-3)}{4!} \cdot \Delta^4 y'_{n-3} + \cdots \tag{7.1.1}$$

where $u = \dfrac{x - x_{n-3}}{h}$ or $x = x_{n-3} + hu$. Therefore $dx = h\ du$. Let the differential equation be

$$\frac{dy}{dx} = f(x, y) \ \text{ with } \ y(x_{n-3}) = y_{n-3}. \tag{7.1.2}$$

Now integrating (7.1.2) over the range x_{n-3} to x_{n+1}, we get

$$\int_{x_{n-3}}^{x_{n+1}} dy = \int_{x_{n-3}}^{x_{n+1}} y'\, dx$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \int_0^4 \left[y'_{n-3} + u \cdot \Delta y'_{n-3} + \frac{u(u-1)}{2!} \cdot \Delta^2 y'_{n-3} + \frac{u(u-1)(u-2)}{6} \cdot \Delta^2 y'_{n-2} + \frac{u(u-1)(u-2)(u-3)}{24} \cdot \Delta^4 y'_{n-3} \right] du$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \left[4y'_{n-3} + 8\Delta y'_{n-3} + \frac{20}{3}\Delta^2 y'_{n-3} + \frac{8}{3}\Delta^3 y'_{n-3} + \frac{14}{45}\Delta^4 y'_{n-3} \right]$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \left[4y'_{n-3} + 8(E-1) y'_{n-3} + \frac{20}{3}(E-1)^2 y'_{n-3} + \frac{8}{3}(E-1)^3 y'_{n-3} \right] + \frac{14}{45} h \Delta^4 y'_{n-3}$$

$$\begin{aligned}
\Rightarrow y_{n+1} - y_{n-3} &= h\Bigg[4y'_{n-3} + 8\{y'_{n-2} - y'_{n-3}\} + \frac{20}{3}\{y'_{n-1} - 2y'_{n-2} + y'_{n-3}\} \\
&\qquad + \frac{8}{3}\{y'_n - 3y'_{n-1} + 3y'_{n-2} - y'_{n-3}\}\Bigg] + \frac{14}{45}h\Delta^4 y'_{n-3} \\
\Rightarrow y_{n+1} - y_{n-3} &= \frac{4h}{3}\Bigg[2y'_{n-2} - y'_{n-1} + 2y'_n\Bigg] + \frac{14}{45}h\Delta^4 y'_{n-3} \\
\Rightarrow y_{n+1} &= y_{n-3} + \frac{4h}{3}\Bigg[2y'_{n-2} - y'_{n-1} + 2y'_n\Bigg] + E_1
\end{aligned}$$

where $E_1 = \dfrac{14}{45}h\Delta^4 y'_{n-3} = \dfrac{14}{45}h^5 y^v(\xi_1),\ \ (x_{n-3} < \xi_1 < x_{n+1})$, assuming that $y^v(x)$ does not vary strongly in the small interval (x_{n-3}, x_{n+1}). Then the formula

$$y^{(p)}_{n+1} = y_{n-3} + \frac{4h}{3}\Big[2y'_{n-2} - y'_{n-1} + 2y'_n\Big] \tag{7.1.3}$$

is called the *Milne's extrapolation formula or Predictor formula* with the error

$$E_1 = \frac{14}{45}h\Delta^4 y'_{n-3} = \frac{14}{45}h^5 y^v(\xi_1),\ \ (x_{n-3} < \xi_1 < x_{n+1}) \tag{7.1.4}$$

To derive the corrector formula, we integrate Eq.(7.1.2) by the Newton's forward formula with starting node x_{n-1}, in terms of y' and u

$$\begin{aligned}
y' = y'_{n-1} + u\cdot\Delta y'_{n-1} + \frac{u(u-1)}{2!}\cdot\Delta^2 y'_{n-1} + \frac{u(u-1)(u-2)}{3!}\cdot\Delta^2 y'_{n-1} \\
+ \frac{u(u-1)(u-2)(u-3)}{4!}\cdot\Delta^4 y'_{n-1} + \cdots
\end{aligned} \tag{7.1.5}$$

where $u = \dfrac{x - x_{n-1}}{h}$ or $x = x_{n-1} + hu$, over the range x_{n-1} to x_{n+1} as follows:

$$\begin{aligned}
\int_{x_{n-1}}^{x_{n+1}} dy &= \int_{x_{n-1}}^{x_{n+1}} y'\,dx \\
\Rightarrow y_{n+1} - y_{n-1} &= h\int_0^2 \Bigg[y'_{n-1} + u\cdot\Delta y'_{n-1} + \frac{u(u-1)}{2!}\cdot\Delta^2 y'_{n-1} + \frac{u(u-1)(u-2)}{6}\cdot\Delta^2 y'_{n-1} \\
&\qquad + \frac{u(u-1)(u-2)(u-3)}{24}\cdot\Delta^4 y'_{n-1}\Bigg]\,du \\
\Rightarrow y_{n+1} - y_{n-1} &= h\Bigg[2y'_{n-1} + 2\Delta y'_{n-1} + \frac{1}{3}\Delta^2 y'_{n-1} - \frac{1}{90}\Delta^4 y'_{n-1}\Bigg] \\
\Rightarrow y_{n+1} - y_{n-1} &= h\Bigg[2y'_{n-1} + 2(E-1)y'_{n-1} + \frac{1}{3}(E-1)^2 y'_{n-1}\Bigg] - \frac{h}{90}\Delta^4 y'_{n-1} \\
\Rightarrow y_{n+1} - y_{n-1} &= h\Bigg[2y'_{n-1} + 2\{y'_n - y'_{n-1}\} + \frac{1}{3}\{y'_{n+1} - 2y'_n + y'_{n-1}\}\Bigg] - \frac{h}{90}\Delta^4 y'_{n-1} \\
\Rightarrow y_{n+1} - y_{n-1} &= \frac{h}{3}\Bigg[y'_{n-1} + 4y'_n + y'_{n+1}\Bigg] - \frac{h}{90}\Delta^4 y'_{n-1}
\end{aligned}$$

$$\Rightarrow y_{n+1} \quad = \quad y_{n-1} + \frac{h}{3}\left[y'_{n-1} + 4y'_n + y'_{n+1}\right] + E_2$$

where $E_2 = -\frac{h}{90}\Delta^4 y'_{n-4} = -\frac{h^5}{90}y^v(\xi_2),\ \ (x_{n-1} < \xi_2 < x_{n+1})$, assuming that $y^v(x)$ does not vary strongly in the small interval (x_{n-1}, x_{n+1}). Then the formula

$$y^{(c)}_{n+1} = y_{n-1} + \frac{h}{3}\left[y'_{n-1} + 4y'_n + y'_{n+1}\right] \tag{7.1.6}$$

is called the *Milne's corrector formula* with the error

$$E_2 = -\frac{h}{90}y^v_{\xi_2},\ \ (x_{n-1} < \xi_2 < x_{n+1}) \tag{7.1.7}$$

The value of y_{n+1} computed by (7.1.3) may be called it predicted value and that computed by (7.1.6) is called the corrected value and are respectively denoted by $y^{(p)}_{n+1}$ and $y^{(c)}_{n+1}$. If $y^v(x)$ does not vary strongly in the small interval (x_{n-3}, x_{n+1}) of length $4h$, in general we may take $y^v(\xi_1) \approx y^v(\xi_2)$. Thus we have $E_1/E_2 \approx -28 \Rightarrow E_1 \approx -28E_2$. If D_{n+1} be the estimation of error, we have

$$D_{n+1} = \text{Corrected value } \ y_{n+1} - \text{Predicted value } \ y_{n+1} = E_1 - E_2 \approx -29E_2 \tag{7.1.8}$$

7.2 Computational Procedure

• Step 1: Compute y'_{n-2}, y'_{n-1}, y'_n by the given differential equation i.e., $y'_r = f(x_r, y_r)$.

• Step 2: Compute $y^{(p)}_{n+1}$ by the predictor formula (7.1.3).

• Step 3: Compute y'_{n+1} by the given differential equation, by using the predicted value $y^{(p)}_{n+1}$ obtained in Step 2.

• Step 4: Using the predicted value y'_{n+1} obtained in Step 3, compute $y^{(c)}_{n+1}$ by the corrector formula (7.1.6).

• Step 5: Compute D_{n+1} = corrected value ($y^{(c)}_{n+1}$ - predicted value $y^{(p)}_{n+1}$. If D_{n+1} is very small then proceed for the next interval and D_{n+1} is not sufficiently small, then reduce, the value of h by taking its half etc.

Example 7.1. Compute $y(2)$, if $y(x)$ satisfies the equation $\frac{dy}{dx} = \frac{1}{2}(x + y)$, given that $y(0) = 2,\ y(0.5) = 2.636,\ y(1.0) = 3.595$ and $y(1.5) = 4.968$, using Milne's Method.

Solution: We take here $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1.0$, $x_3 = 1.5$ and $y(0) = y_0 = 2$, $y(0.5) = y_1 = 2.636$, $y(1) = 3.595$ and $y(1.5) = y_3 = 4.968$. We have to compute $y(2.0) = y_4$.

Putting $n = 3$ in the predictor formula (7.1.3) and in the corrector formula (7.1.6) we get, respectively,

$$y^{(p)}_4 = y_0 + \frac{4h}{3}[2y'_1 - y'_2 + 2y'_3] \tag{7.2.1}$$

$$y^{(p)}_4 = y_2 + \frac{h}{3}[y'_2 + 4y'_3 + y'_4] \tag{7.2.2}$$

From the differential equation $\frac{dy}{dx} = y' = \frac{1}{2}(x+y)$, we get

$$\begin{aligned} y_1' &= \frac{1}{2}(x_1+y_1) = \frac{1}{2}(0.5+2.636) = 1.568 \\ y_2' &= \frac{1}{2}(x_2+y_2) = \frac{1}{2}(1.0+3.595) = 2.2975 \\ y_3' &= \frac{1}{2}(x_3+y_3) = \frac{1}{2}(1.5+4.968) = 3.234 \end{aligned}$$

Thus, from (7.2.1), the predicted value

$$y_4^{(1)_p} = 2 + \frac{4\times 0.5}{2}[2\times 1.569 - 2.2975 + 2\times 3.234] = 6.8710$$

Now by the given differential equation, we have first estimation

$$y_4'^{(0)} = \frac{1}{2}[x_4 + y_4^{(1)_p}] = \frac{1}{2}[2+6.8710] = 4.4355$$

Now by (7.2.2), we get first corrected value as

$$\begin{aligned} y_4^{(1)_c} &= y_2 + \frac{h}{3}[y_2' + 4y_3' + y_4'^{(0)}] \\ &= 3.595 + \frac{0.5}{3}[2.2975 + 4\times 3.234 + 4.4355] = 6.8731667 \approx 6.87317 \end{aligned}$$

Again recomputing y_4' from the differential equation we get,

$$y_4'^{(1)} = \frac{1}{2}[x_4 + y_4^{(1)}] = \frac{1}{2}[2+6.87317] = 4.436585$$

By (7.2.2), we get second corrected value as

$$\begin{aligned} y_4^{(2)_c} &= y_2 + \frac{h}{3}[y_2' + 4y_3' + y_4'^{(1)}] \\ &= 3.595 + \frac{0.5}{3}[2.2975 + 4\times 3.234 + 4.436585] = 6.8733475 \approx 6.873 \end{aligned}$$

As $y_4^{(1)_c}$ = $y_4^{(2)_c}$ = 6.873, therefore $y(2) = 6.873$ correct to 3-decimal places.

Exercise 7.2. (i) Using Milne's predictor-corrector method, find $y(0.4)$ for the initial value problem

$$y' = x^2 + y^2, \;\; y(0) = 1, \;\; \text{with} \;\; h = 0.1$$

Calculate all the required initial values by Euler's method. The result is to accurate to three decimal places.
Answer: 1.63138
(ii) Compute $y(0.5)$, by Milne's predictor-corrector method from $\frac{dy}{dx} = 2e^x - y$ given that

$$y(0.1) = 2.0100, \quad y(0.2) = 2.0401, \quad y(0.3) = 2.0907, \quad y(0.4) = 2.1621$$

Answer: 2.2553

Unit 8

Explicit Method for Solving Parabolic PDE

One of the simplest second order Parabolic Differential Equation in one-dimension is the Heat Conduction Equation, written as:

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} \quad where\ 0 \le x \le L, t \ge 0 \qquad (1.1)$$

which arises in many real problems.

The appropriate, but most simple conditions are:
Initial condition: $u(x,0) = u_0(x)$

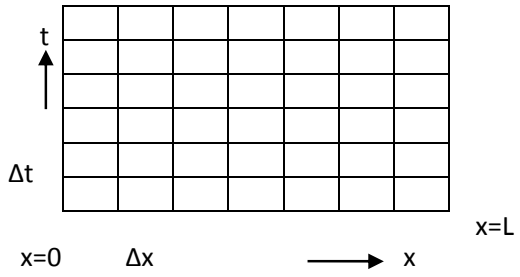
and two Boundary Conditions namely: $u(0,t) = u_1(t)\ and\ u(L,t) = u_2(t).$

Note that the analytical solution of equation (1.1) is usually a trigonometric series, which may create problem in convergence.

The first Finite Difference method is the Explicit Method.

For this, let us discuss **first step**, which is common to all methods i.e. discretization.

The domain of the solution is $0 \le x \le L, t \ge 0$, as shown in fig(1).



fig(1)

It is to be discretized by drawing vertical and horizontal lines at equal distance say $\Delta x\ and\ \Delta t$ respectively. Let i , j be defined dummy variables along x & t axis so that;

$$x_i = i\Delta x,\ t_j = j\Delta t \quad \& \quad u(x_i, t_j) = u_{i,j} = (i\Delta x,\ j\Delta t).$$

Let the domain from x = 0 to x = L be subdivided into N sub-parts so that x = 0 corresponds to i = 0 and x = L corresponds to i = N with t = 0 corresponds to j = 0.

(1.2)

The initial condition then can be written as: $u(x,0)=u_0(x)\Rightarrow u_{i,0}=u_0(i\Delta x)$

The boundary condition will be converted to:
$$\begin{aligned} u(0,t)&=u_1(t)\Rightarrow u_{0,j}=u_1(j\Delta t)\\ u(L,t)&=u_2(t)\Rightarrow u_{N,j}=u_2(j\Delta t)\end{aligned} \qquad (1.3)$$

Step 2: Replacing the derivatives by corresponding Finite Difference representation in equation (1.1) which reduces to:

$$\frac{u_{i,j+1}-u_{i,j}}{\Delta t}=c^2\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{(\Delta x)^2} \qquad (1.4)$$

The truncation error is: $o(\Delta t)+o(\Delta x)^2$

Equation (1.4) thus can be rewritten as:

$$u_{i,j+1}=c^2ru_{i+1,j}+(1-2c^2r)u_{i,j}+c^2ru_{i-1,j} \qquad \text{with, } r=\Delta t/(\Delta x)^2 \qquad (1.5)$$

Equation (1.5) is called the Explicit Scheme.

The computational molecule for scheme (1.5) can be shown as in fig (2)

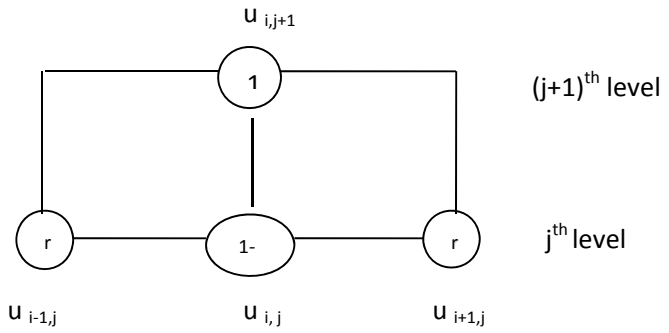


fig (2)

Equation (1.5) is now solved at first time level for j = 0; for all values of i =1,2,---,(N-1). Similarly solution at second,third time level is obtained correspondingly for j =1, 2 ---. It is very important to note that this scheme is not unconditionally stable.

The value of r has to be < 1/2 i.e. $\Delta t < (1/2)(\Delta x)^2$ which makes Δt to be sufficiently small. Thus it requires large no. of computations at intermediate time level ;even for a small time ,as Δx is itself very small (Since the Finite Difference Method for approximating the derivatives is based

on Taylor's expansion hence both Δt and Δx are small) .This is one of the great drawback of this method.

Though the truncation error tends to 0 as $\Delta t \to 0$ and $\Delta x \to 0$ but the detail discussion about it will be discussed in module 3,lecture 1.The main advantage of this scheme is that it is computationally simple as the computations proceed pointwise, thus even manually manageable.

Example 1:- Solve the Heat Conduction Equation $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial t^2}$ $\quad 0 \le x \le 1, t \ge 0$

subject to B.C: u = 0 at x = 0 and $\frac{\partial u}{\partial x} = 0$ at x = 1 ,for all t

and, I.C: $u(x,0) = \sin\frac{3\pi x}{2}$

Using the Explicit Method ,choosing $\Delta x = 0.1$ and $\Delta t = 0.0025$ so that r = 1/4, obtain the solution for one time level and compare with the exact solution.

The exact solution is $u(x,t) = e^{\frac{-9\pi^2 t}{4}} \sin\frac{3\pi x}{2}$

Solution-

At a general point (i,j) the given pde is -

$$\left(\frac{\partial u}{\partial t}\right)_{i,j} = \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j}$$

The Explicit Finite-Difference representation of this equation is:

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{(\Delta x)^2}$$

or $$u_{i,j+1} = ru_{i-1,j} + (1-2r)u_{i,j} + ru_{i+1.j} \qquad (1)$$

where, $r = \frac{\Delta t}{(\Delta x)^2}$

(2)

Initial condition is: $u(x,0) = \sin\frac{3\pi x}{2}$

Boundary conditions are: $u_{0,j} = 0$ and $\left(\frac{\partial u}{\partial x}\right)_{N,j} = 0 \;\; ; N = 10$

Replacing L.H.S of the above boundary condition by Backward Difference,

$$\frac{u_{N-1,j} - u_{N,j}}{\Delta x} = 0 \;\; \Rightarrow u_{N-1,j} = u_{N,j} \Rightarrow u_{10\,j} = u_{9,j} \qquad (3)$$

Substituting r = 1/4 and j = 0 in equation (1)

$$u_{i,1} = \frac{1}{4}\left(u_{i-1,0} + 2u_{i,0} + u_{i+1,0}\right) \qquad (4)$$

Substituting i =1,2------,9 in equation (4) ,we get values at the first time level .These values are used for the solution at the second time level for j=1.These values are shown below:

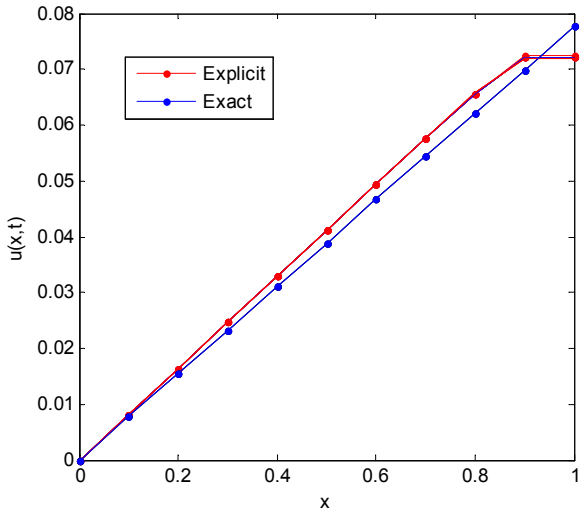
	i = 0	i = 1	i = 2	i = 3	i = 4	i = 5	i = 6	i = 7	i = 8	i = 9	i = 10
	X= 0.	X=0.1	X=0.2	X=0.3	X=0.4	X=0.5	X=0.6	X=0.7	X=0.8	X=0.9	X=1.0
j=0	0	.0082	.0164	.0247	.0329	.0411	.0493	.0575	.0657	.0740	.0740
j=1	0	.0082	.0164	.0247	.0329	.0411	.0493	.0574	.0656	.0739	.0739

The Exact solution is :

$$u(x,t) = e^{\frac{-9\pi^2 t}{4}} \sin\frac{3\pi x}{2}$$ (here t=0.0025)

Comparison between Explicit and Exact solution:

X	0.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Explicit	0	.0082	.0164	.0247	.0329	.0411	.0493	.0575	.0656	.0739	.0739
Exact	0	.0078	.0156	.0233	.0311	.0389	.0467	.0543	.0622	.0699	.0777



Example 2 : Consider the PDE : $\frac{\partial u}{\partial t} = x\frac{\partial^2 u}{\partial x^2}; 0 < x < 1, t > 0$ (1)

B.Cs (i) u=0 at x=0, t>0 (ii) $\frac{\partial u}{\partial x} = \frac{-1}{2}u; x = 1, t > 0$

I.C is u = x(1-x) when t=0 & $0 \le x \le 1$

Solve this equation by an explicit method , employing central-difference for the boundary conditions. Take $\Delta x = h = 0.1$ & r =0.25 and 0.7 and compare the result.

Solution:

The explicit approximation is

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{ih\left(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right)}{h^2}$$

$$\Rightarrow \quad u_{i,j+1} = irhu_{i-1,j} + \left(1 - 2irh\right)u_{i,j} + irhu_{i+1,j}; i = 1, 2, ..., N-1 \qquad (2) \quad \text{with} \quad r = k/h^2$$

Now applying central difference formula to Second B.C. , we get

$$\frac{u_{i+1,j} - u_{i-1,j}}{2h} = \frac{-1}{2}u_{i,j} \Rightarrow u_{i+1,j} - u_{i-1,j} = -hu_{i,j}$$

At x=1 i .e. i=10

$$u_{11,j} - u_{9,j} = -hu_{10,j} \Rightarrow u_{11,j} = u_{9,j} - hu_{10,j} = u_{9,j} - .1u_{10,j} \qquad (3)$$

(i) r=0.25

Putting i=10 , h=0.1 in equation (2);

$$u_{10,j+1} = \frac{1}{4}u_{9,j} + \frac{1}{2}u_{10,j} + \frac{1}{4}u_{11,j} \qquad (4)$$

Eliminating $u_{11,j}$ from eqn.(3) & eqn.(4),we get

$$u_{10,j+1} = \frac{1}{2}u_{9,j} + \frac{19}{40}u_{10,j} \qquad (5)$$

The other B.C. is u=0 at x=0 & t>0

$$\Rightarrow u_{0,1} = u_{0,2} = u_{0,3} = ...u_{0,n} = 0$$

And I.C is: u(x,0)=x(1-x); $0 \le x \le 1$

$u_{0,0} = u(0,0) = 0,$ $u_{1,0} = u(.1,0) = 0.09,$ $u_{2,0} = u(.2,0) = 0.16$

$u_{3,0} = u(.3,0) = 0.21,$ $u_{4,0} = u(.4,0) = 0.24,$ $u_{5,0} = u(.5,0) = 0.25,$ $u_{6,0} = 0.24,$

$u_{7,0} = 0.21,$ $u_{8,0} = 0.16,$ $u_{9,0} = 0.09,$ $u_{10,0} = 0$

Now putting r=0.25 & h=0.1 in (2)

$$\Rightarrow u_{i,j+1} = 0.025iu_{i-1,j} + (1 - 0.05i)u_{i,j} + 0.025iu_{i+1,j} \qquad (6)$$

Ist time level: Putting i=1,2,3…….9 ,j=0 in eqn.(6)

$u_{1,1} = .025u_{0,0} + (1 - .05)u_{1,0} + .025u_{2,0} = 0.0895$

$u_{2,1} = .05u_{1,0} + 0.9u_{2,0} + .05u_{3,0} = 0.1590$

$u_{3,1} = .075u_{2,0} + .85u_{3,0} + .075u_{4,0} = 0.2085$

$u_{4,1} = 0.1u_{3,0} + 0.8u_{4,0} + 0.1u_{5,0} = 0.2380$

$u_{5,1} = 0.2475,$ $u_{6,1} = 0.2370,$ $u_{7,1} = 0.2065$

$u_{8,1} = 0.1560,$ $u_{9,1} = 0.0855$

Putting j=0 in equation (5)

$$u_{10,1} = \frac{1}{2}u_{9,0} + \frac{19}{40}u_{10,0} = \frac{1}{2}u_{9,0} = 0.0450$$

Now**, Second time level**: Putting i=1,2,3…….9 ,j=1 in equation (6)

$$u_{1,2} = 0.025u_{0,1} + (1-0.05)u_{1,1} + 0.025u_{2,1} = .0890$$
$$u_{2,2} = \quad 0.05u_{1,1} + 0.9u_{2,1} + 0.05u_{3,1} = 0.1580$$
$$u_{3,2} = \quad 0.075u_{2,1} + 0.85u_{3,1} + 0.075u_{4,1} = 0.2070$$
$$u_{4,2} = \quad 0.1u_{3,1} + 0.8u_{4,1} + 0.1u_{5,1} = 0.2360$$
$$u_{5,2} = 0.2450, \qquad u_{6,2} = 0.2340, \qquad u_{7,2} = 0.2030$$
$$u_{8,2} = 0.1520, \qquad u_{9,2} = 0.0922$$

Putting j =1 in equation (5)

$$u_{10,1} = \frac{1}{2}u_{9,1} + \frac{19}{40}u_{10,1} = 0.0641$$

(ii) **r=0.7**

Put i =10,r=0.7,h=0.1 in equation(2)

$$u_{10,j+1} = 0.7u_{9,j} - 0.4u_{10,j} + 0.7u_{11,j} \qquad (7)$$

Eliminate $u_{11,j}$ from equation (3) and equation (7); we get

$$u_{10,j+1} = 1.4u_{9,j} - .47u_{10,j} \qquad (8)$$

Put r=0.7 & h=0.1 in equation (2)

$$u_{i,j+1} = 0.07iu_{i-1,j} + (1-0.14i)u_{i,j} + 0.07iu_{i+1,j} \qquad (9)$$

Ist time level

Putting i=1,2,3…….9 , and j=0 in (9)

$$u_{1,1} = .07u_{0,0} + (1-.14)u_{1,0} + .07u_{2,0} = 0.0886$$
$$u_{2,1} = \quad 0.14u_{1,0} + .72u_{2,0} + .14u_{3,0} = 0.1572$$
$$u_{3,1} = \quad 0.21u_{2,0} + .58u_{3,0} + .21u_{4,0} = 0.2058$$
$$u_{4,1} = \quad 0.28u_{3,0} + .44u_{4,0} + .28u_{5,0} = 0.2344$$
$$u_{5,1} = 0.2430, \qquad u_{6,1} = 0.2316, \qquad u_{7,1} = 0.2002$$
$$u_{8,1} = 0.1488, \qquad u_{9,1} = 0.0774$$

Putting j=0 in (8)

$$u_{10,1} = 1.4u_{9,0} - .47u_{10,0} = .126$$

Second time level

Putting i=1,2,3…….9 , and j=1 in (9)

$$u_{1,12} = 0.07u_{0,1} + (1-0.14)u_{1,1} + 0.07u_{2,1} = 0.0872$$

$$u_{2,2} = 0.14u_{1,1} + 0.72u_{2,1} + 0.14u_{3,1} = 0.1544$$

$$u_{3,2} = 0.21u_{2,1} + 0.58u_{3,1} + 0.21u_{4,1} = 0.2016$$

$$u_{4,2} = 0.28u_{3,1} + 0.44u_{4,1} + 0.28u_{5,1} = 0.2288$$

$$u_{5,2} = 0.2360, \quad u_{6,2} = 0.2232, \quad u_{7,2} = 0.1904$$

$$u_{8,2} = 0.1376, \quad u_{9,2} = 0.1530$$

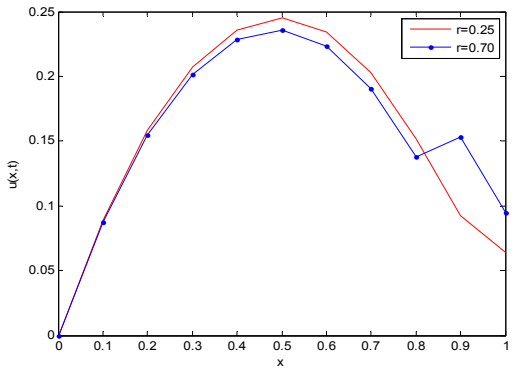
Putting j=1 in (8)

$$u_{10,2} = 1.4u_{9,1} - 47u_{10,1} = 0.0491$$

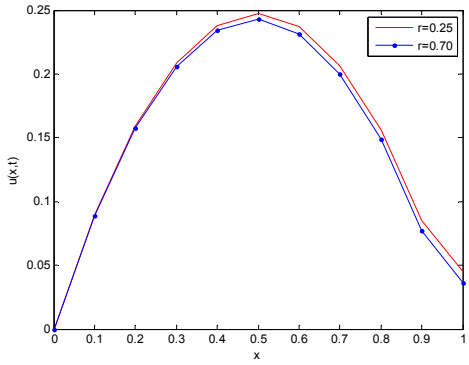
The results are written in Tabular Form:

	For first time level			**For second time level**		
	r=.25	r=.7	Difference	r=.25	r=.7	Difference
I=1	.0895	.0886	.0009	.089	.0872	.0018
I=2	.159	.1572	.00180	.158	.1544	.0036
I=3	.2085	.2058	.0027	.207	.2016	.0054
I=4	.238	.2344	.0036	.236	.2288	.0072
I=5	.2475	.243	.0045	.245	.236	.0090
I=6	.237	.2316	.0054	.234	.2232	.0108
I=7	.2065	.2002	.0063	.203	.1904	.0396
I=8	.156	.1488	.0072	.152	.1376	.0144
I=9	.0855	.0774	.0081	.0923	.153	-.0608
I=10	.045	.036	.009	.064	.095	-.0308

Comparison for different values of 'r' in Explicit Method:-



First Time Level



Second Time Level

Elliptic Partial Differential Equations

(Solution in Cartesian coordinate system)

Other category of second order PDE, which are basically used to characterize steady state systems are called as Elliptic PDE. More prevalent examples are Laplace Equation and Poisson Equation. Every potential function satisfies Laplace Equation. Another simple example is of heat transfer in a rectangular plate under certain boundary conditions,where the temperature is to be determined after a large time under steady state condition.

The Laplace/Poission equation in Cartesian coordinate system is given as

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x,y)\ or\ \nabla^2 u=f(x,y),\quad a\le x\le b, c\le y\le d \qquad (1.1)$$

subject to either Dirichlet conditions or Mixed conditions.

Let the domain be subdivided by drawing horizontal and vertical lines at an equal distance of Δy and Δx respectively. Let i and j be chosen as dummy variables along x and y axis for i =1,2,...,N and j =1,2,...,M.

Replacing both the second order derivatives in eqn.(1.1) by central difference approximations:

$$\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{(\Delta x)^2}+\frac{u_{i,j-1}-2u_{i,j}+u_{i,j+1}}{(\Delta y)^2}=f(i\Delta x, j\Delta y) \qquad (1.2)$$

For $\Delta x=\Delta y$:

$$u_{i-1,j}-4u_{i,j}+u_{i+1,j}+u_{i,j-1}+u_{i,j+1}=(\Delta x)^2 f(i\Delta x, j\Delta x) \qquad (1.3)$$

Equation (1.3) reduces to Laplace equation if f = 0. This equation can be solved iteratively both explicitly as well as implicitly. The boundary conditions can be written as:

As a special case, Let $a=0=c$, then if $\mathrm{i}=1,2,....,\mathrm{N},$ and $\mathrm{j}=1,2,.....,\mathrm{M}$

$x_i=i\Delta x,\ \ y_i=i\Delta y$,

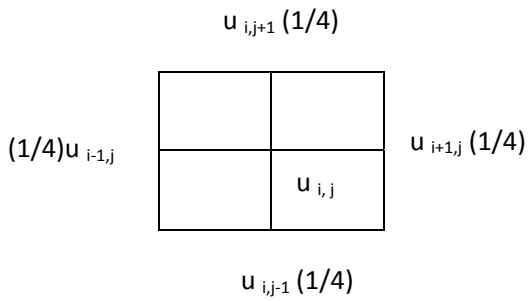
the boundary conditions can be written as $\mathrm{u}_{0,\mathrm{j}}, \mathrm{u}_{\mathrm{N+1,j}}$ and $\mathrm{u}_{\mathrm{i,0}}$ and $\mathrm{u}_{\mathrm{i,M+1}}$.

Explicit scheme:

Rewritting equation (1.3) as:

$$u_{i,j}=\frac{(u_{i-1,j}+u_{i+1,j}+u_{i,j-1}+u_{i,j+1})}{4} \qquad (1.4)$$

The computational molecule for (1.4) can be shown as



We start with here, $j = 1,\ \ i = 1,2,.....,N$. However it involves $u_{i+1,j}$ as well as $u_{i,j+1}$ which are unknowns. Hence one has to start with guessed values. Thus an iterative procedure has to be implemented. The guessed value has to be to chosen carefully in accordance with the boundary conditions. Now R.H.S. can be handled accordingly with Jacobi's or Gauss-Seidel approach. The two possible iterative formulae, thus can be written as:

$$u_{i,j}^{n+1} = \frac{(u_{i-1,j}^{n} + u_{i+1,j}^{n} + u_{i,j-1}^{n} + u_{i,j+1}^{n})}{4} \qquad (1.5)$$

$$u_{i,j}^{n+1} = \frac{(u_{i-1,j}^{n+1} + u_{i,j-1}^{n+1} + u_{i+1,j}^{n} + u_{i,j+1}^{n})}{4} \qquad (1.6)$$

The superscript 'n' denotes the number of iterations. Both the formulae have the truncation error $o(\Delta x)^2 + o(\Delta y)^2$. The iteration may be carried row-wise or column-wise.

It may be noted that Δx and Δy are taken to be small so usually number of nodes are quite large. But sometimes geometrical symmetry may occur depending on the boundary conditions. In that case the computational efforts can be reduced. For example if the boundary conditions are $u(0,y) = u_0,\ u(a,y) = u_0,\ \ u(x,0) = u_1,\ \ u(x,b) = u_1$ then there is symmetry along both x & y axis. This symmetry may be helpful in reducing the computations and one has to find solution only in ¼ of the domain.

Example: $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.5x,$ defined over $0 \le x \le 0.8,\ 0 \le y \le 0.6,$ with subjected to $u = 1$ at $x = 0, y = 0, y = 0.6$ and $\frac{\partial u}{\partial x} = u$ at $x = 0.8,$ Obtain the solution correct to 2d using both equations (1.5) and (1.6) and compare the result.

Solution:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.5x$$

U sin g equation (1.5) we have

$$u_{i,j}^{n+1} = \frac{(u_{i-1,j}^n + u_{i+1,j}^n + u_{i,j-1}^n + u_{i,j+1}^n)}{4} - \frac{0.5(0.2)^3 i}{4} \qquad (1.7)$$

$$\left(\frac{\partial u}{\partial x}\right)_{4,1} = u_{4,1} \quad \Rightarrow \frac{u_{5,1} - u_{3,1}}{2 \times 0.2} = u_{4,1} \qquad (1.8)$$

Starting with assumed values as 0 i.e. $u_{i,j}^0 = 0$ and

Putting $i = 1,\ j = 1$ in equation(1.7)

$$u_{1,1}^{n+1} = \frac{1}{4}\left[u_{1,1}^n + u_{2,1}^n + 1.996\right] \qquad (1.9)$$

Put $i = 2,\ j = 1$ in equation(1.7)

$$u_{2,1}^{n+1} = \frac{1}{4}\left[u_{1,1}^n + u_{2,1}^n + u_{3,1}^n + 0.992\right] \qquad (1.10)$$

Put $i = 3,\ j = 1$ in equation(1.7)

$$u_{3,1}^{n+1} = \frac{1}{4}\left[u_{2,1}^n + u_{3,1}^n + u_{4,1}^n + +0.998\right] \qquad (1.11)$$

Put $i = 4,\ j = 1$ in equation$(1.7)$and by equation(1.8)

$$u_{4,1}^{n+1} = \frac{5}{13}\left[2u_{3,1}^n + 0.984\right] \qquad (1.12)$$

Hence finally we have four equations $(1.9), (1.10), (1.11), (1.12)$ which are solved iteratively. The values as obtained are shown below:

	$u_{1,1}$	$u_{2,1}$	$u_{3,1}$	$u_{4,1}$
n = 0	0.499	0.248	0.2495	0.3785
n = 1	0.6858	0.4971	0.4685	0.5704
n = 2	0.7947	0.6609	0.6335	0.7388
n = 3	0.8629	0.7778	0.7653	0.8658

n = 4	0.9092	0.8495	0.8517	0.9672
n = 5	0.9384	0.9006	0.9166	1.0336
n = 6	0.9587	0.9369	0.9622	1.0835
n = 7	0.9729	0.9625	0.9951	1.1186
n = 8	0.9829	0.9806	1.0186	1.1439
n = 9	0.9899	0.9935	1.0353	1.1620
n = 10	0.9949	1.0027	1.0472	1.1748
n = 11	0.9984	1.0092	1.0557	1.1860
n = 12	1.0009	1.0138	1.0617	1.1905

Hence solution correct to 2d is:

$u_{1,1}$=1 ; $u_{2,1}$=1.01 ; $u_{3,1}$=1.06 ; $u_{4,1}$=1.19

Now **using equation (1.6),** we have four final equations:

$$u_{1,1}^{n+1} = \frac{1}{4}\left[u_{1,1}^{n} + u_{2,1}^{n} + 1.996\right]$$

$$u_{2,1}^{n+1} = \frac{1}{4}\left[u_{1,1}^{n+1} + u_{2,1}^{n} + u_{3,1}^{n} + 0.992\right]$$

$$u_{3,1}^{n+1} = \frac{1}{4}\left[u_{2,1}^{n+1} + u_{3,1}^{n} + u_{4,1}^{n} + +0.998\right]$$

$$u_{4,1}^{n+1} = \frac{5}{13}\left[2u_{3,1}^{n+1} + 0.984\right]$$

These equations are solved iteratively, the results obtained are shown below:

	$u_{1,1}$	$u_{2,1}$	$u_{3,1}$	$u_{4,1}$
n = 0	0.499	0.3728	0.3427	0.6421
n = 1	0.7169	0.6061	0.6472	0.8763
n = 2	0.8297	0.7688	0.8226	1..0112
n = 3	0.8986	0.8715	0.9258	1.0906
n = 4	0.9415	0.9327	0.9868	1.1375
n = 5	0.9676	0.9698	1.0230	1.1654
n = 6	0.9833	0.9920	1.0446	1.1820
n = 7	0.9828	1.0054	1.0575	1.1919
n = 8	1.0020	1.0134	1.0652	1.1948

Hence solution correct to 2d is:

$u_{1,1}$=1 ; $u_{2,1}$=1.01 ; $u_{3,1}$=1.06 ; $u_{4,1}$=1.19

Hence by equation (1.5),we have to work for 13 iterations where as by equation (1.6),we work with only 9 iterations. Therefore, equation (1.6) gives faster convergence.

Core Paper

MATC 3.2

Block - II

Marks : 50 (SSE : 40; IA : 10)

Calculus of $\mathbb{R}^n$ (Pure and Applied Streams)

Syllabus

• Unit 9 •

Differentiation on $\mathbb{R}^n$: Directional derivatives and continuity, the total derivative and continuity, total derivative in terms of partial derivatives, the matrix transformation of $T : \mathbb{R}^n \to \mathbb{R}^n$. The Jacobian matrix.

• Unit 10 •

The chain rule and its matrix form. Mean value theorem for vector valued function. Mean value inequality. A sufficient condition for differentiability. A sufficient condition for mixed partial derivatives.

• Unit 11 •

Functions with non-zero Jacobian determinant, the inverse function theorem, the implicit function theorem as an application of Inverse function theorem.

• Unit 12 •

Extremum problems with side conditions – Lagrange's necessary conditions as an application of Inverse function theorem.

• Unit 13 •

Integration on $\mathbb{R}^n$: Integrals of $f : A \to \mathbb{R}$ where A is subset of $\mathbb{R}^n$, is a closed rectangle. Conditions of integrability.

• Unit 14 •

Integrals of $f : C \to \mathbb{R}$ where C is subset of $\mathbb{R}^n$ is not a rectangle, concept of Jordan measurability of a set in $\mathbb{R}^n$.

• Unit 15 • Fubini's theorem for integral of $f : A \times B \to \mathbb{R}$, where A, B are subsets of $\mathbb{R}^n$, are closed rectangles. Fubini's theorem for $f : C \to \mathbb{R}$, C is a subset of $A \times B$.

• Unit 16 • Formula for change of variables in an integral in $\mathbb{R}^n$.

Unit 9

Course Structure

- Directional derivatives and continuity, the total derivative and continuity.
- Total derivative in terms of partial derivatives, the matrix transformation of $T : \mathbb{R}^n \to \mathbb{R}^n$.
- The Jacobian matrix.

---

8 Introduction

Multivariable calculus (also known as multivariate calculus) is the extension of calculus in one variable to calculus with functions of several variables: the differentiation and integration of functions involving multiple variables, rather than just one. In this unit, we will introduce the basic notions in multivariable calculus.

Objectives

After reading this unit, you will be able to

- define a multivariable function and give some examples of them
- define continuity of a multivariable function and learn certain characteristics in this direction
- learn certain related definitions
- get an introduction to the partial derivatives of a multivariable function
- define the directional derivative of a multivariable function and its relationship with derivatives
- define the Jacobian matrix of a function and its relationship with differentiability
- find the Jacobian matrices of multivariable functions

8.1 Multivarable functions

A function from $\mathbb{R}^n$ to $\mathbb{R}^m$ (which is also sometimes called vector-valued function or, a function of n variables) is a rule which associates to each point in $\mathbb{R}^n$, some point in $\mathbb{R}^m$. The point associated to a point $x \in \mathbb{R}^n$ is denoted by $f(x)$. We write $f : \mathbb{R}^n \to \mathbb{R}^m$ to indicate $f(x)\mathbb{R}^m$ is defined for $x \in \mathbb{R}^n$. The notation $f : A \to \mathbb{R}^m$ indicates that $f(x)$ means that $f(x)$ is defined only for $x \in A$, and A is the domain of f. If $B \subset A$, the we define $f(B)$ as the set $\{f(x) : x \in B\}$ and also, for any $C \subset \mathbb{R}^m$, we define $f^{-1}(C) = \{x \in A : f(x) \in C\}$. The notation $f : A \to B$ indicates that $f(A) \subset B$.

If $f, g : \mathbb{R}^n \to \mathbb{R}$, the functions $f+g, f-g, f.g$ and f/g are defined precisely as in the one-variable case. If $f : A \to \mathbb{R}^m$ and $g : B \to \mathbb{R}^p$, where $B \subset \mathbb{R}^m$, then the composition $g \circ f$ is defined by $g \circ f(x) = g(f(x))$; the domain of $g \circ f$ is $A \cap f^{-1}(B)$. If $f : A \to \mathbb{R}^m$ is one-one, that is, if $f(x) \neq g(x)$, when $x \neq y$, we define $f^{-1} : f(A) \to \mathbb{R}^n$ by the requirement that $f^{-1}(z)$ is the unique $x \in A$ with $f(x) = z$.

A function $f : A \to \mathbb{R}^m$ determines m component functions $f^1, f^2, \ldots, f^m : A \to \mathbb{R}$ by $f(x) = (f^1(x), f^2(x), \ldots, f^m(x))$. If conversely, m functions $g_1, g_2, \ldots, g_m : A \to \mathbb{R}$ are given, there is a unique function $f : A \to \mathbb{R}^m$ such that $f^i = g_i$, namely $f(x) = (g_1(x), g_2(x), \ldots, g_m(x))$. This function f will be denoted by $f = (g_1, g_2, \ldots, g_m)$, so that, we always have $f = (f^1, f^2, \ldots, f^m)$. If $\pi : \mathbb{R}^n \to \mathbb{R}^n$ is the identity function, $\pi(x) = x$, then $\pi^i(x) = x^i$, the function π^i is called the ith projection function.

The notation $\lim_{x \to a} f(x) = b$ means that we can get $f(x)$ as close to b as desired, by choosing x sufficiently close to, but not equal to, a. In mathematical terms this means that for every number $\epsilon > 0$, there is a number $\delta > 0$ such that $|f(x) - b| < \epsilon$ for all x in the domain of f which satisfy the relation $0 < |x - a| < \delta$. A function $f : A \to \mathbb{R}^m$ is continuous at the point $a \in A$ if the limiting value of f at a is equal to the functional value of f at a, that is, $\lim_{x \to a} f(x) = f(x)$, and f is simply continuous if f is continuous at each $a \in A$. The concept of continuity is that it can be defined without using limits.

Theorem 8.1. If $A \subset \mathbb{R}^m$, a function $f : A \to \mathbb{R}^m$ is continuous if and only if for every open set $U \subset \mathbb{R}^m$ there is some open set $V \subset \mathbb{R}^n$ such that $f^{-1}(U) = V \cap A$.

Proof. Suppose f is continuous. If $a \in f^{-1}(U)$, then $f(a) \in U$. Since U is open, there is an open rectangle B with $f(a) \in B \subset U$. Since f is continuous at a, we can ensure that $f(x) \in B$, provided we choose x in some sufficiently small rectangle C containing a. Do this for each $a \in f^{-1}(U)$ and let V be the union of all such C. Clearly $f^{-1}(U) = V \cap A$. The converse is similar. □

Theorem 8.2. If $f : A \to \mathbb{R}^m$ is continuous, where $A \subset \mathbb{R}^n$, and A is compact, then $f(A) \subset \mathbb{R}^m$ is compact.

If $f : A\mathbb{R}$ is bounded, the extent to which f fails to be continuous at $a \in A$ can be measured in a precise way. For $\delta > 0$

$$\begin{aligned} M(a, f, \delta) &= \sup\{f(x) : x \in A \ \& \ |x - a| < \delta\} \\ m(a, f, \delta) &= \inf\{f(x) : x \in A \ \& \ |x - a| < \delta\}. \end{aligned}$$

Definition 8.3. The oscillation $o(f, a)$ of f at a is defined by $o(f, a) = \lim_{\delta \to 0}[M(a, f, \delta) - m(a, f, \delta)]$. This limit always exists, since $M(a, f, \delta) - m(a, f, \delta)$ decreases as δ decreases.

Theorem 8.4. A bounded function f is continuous at a if and only if $o(f, a) = 0$.

Proof. Let f be continuous at a. For every number $\epsilon > 0$, we can choose a number $\delta > 0$ such that $|f(x) - f(a)| < \epsilon$ for all $x \in A$ with $|x - a| < \delta$. Thus, $M(a, f, \delta) - m(a, f, \delta) \leq 2\epsilon$. Since $\epsilon > 0$ is arbitrary, so we have, $o(f, a) = 0$. The converse is similar. □

Theorem 8.5. Let $A \subset \mathbb{R}^n$ be closed. If $f : A \to \mathbb{R}$ is any bounded function, and $\epsilon > 0$, then $\{x \in A : o(f, x) \geq \epsilon\}$ is closed.

Exercise 8.6.

1. If $f : A \to \mathbb{R}^m$ and $a \in A$, then show that $\lim_{x \to a} f(x) = b$ if and only if $\lim_{x \to a} f^i(x) = b^i$, for $i = 1, \ldots, m$.
2. Prove that $f : A \to \mathbb{R}^m$ is continuous at a if and only if each f^i is so.

8.2 Directional Derivatives

In order to define the derivative of a multivalued function f mapping a subset $A \subset \mathbb{R}^n$ into $\mathbb{R}^m$, we come to the notion of directional derivatives which have the following

Definition 8.7. Let $A \subset \mathbb{R}^n$ and let $f : A \to \mathbb{R}^m$ and let A contains a neighbourhood of a. Given $u \in \mathbb{R}^n$ with $u \neq 0$, we define

$$f'(a;u) = \lim_{h\to 0} \frac{f(a+hu) - f(a)}{h},$$

provided the limit exists. This limit depends both on a and on u and is called the directional derivative of f at a with respect to u.

Example 8.8. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = xy$. Then the directional derivative of f at $a = (a^1, a^2)$ with respect to $u = (1, 0)$ is

$$f'(a;u) = \lim_{h\to 0} \frac{(a^1+h)a^2 - a^1a^2}{h} = a^2.$$

With respect to $v = (1, 2)$, the directional derivative is

$$f'(a;v) = \lim_{h\to 0} \frac{(a^1+h)(a^2+2h) - a^1a^2}{h} = a^2 + 2a^1.$$

It is tempting to believe that the "directional derivative" is the appropriate generalization of the notion of "derivative," and to say that f is differentiable at a if $f'(a;u)$ exists for every $u \neq 0$. But this is usually not the case. We want to generalize the derivative in such a way that differentiability implies continuity. But this fails in case of the directional derivatives. Also, the composites of differentiable functions are not always differentiable in this case. So, we seek something stronger.

8.3 Differentiation

We know that a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $a \in \mathbb{R}$ if the limit

$$\lim_{h\to 0} \frac{f(a+h) - f(a)}{h}$$

exists in which case, the limit is called the derivative of f at a and denoted by $f'(a)$. We want to generalize it for multivalued case. For this, we first look into the following case. Let $c : \mathbb{R} \to \mathbb{R}$ be a linear transformation defined by $c(h) = f'(a).h$. Then the above equation becomes

$$\lim_{h\to 0} \frac{f(a+h) - f(a) - c(h)}{h} = 0.$$

The above equation is often interpreted as saying that $c + f(a)$ is a good approximation to f at a. Thus, we can define the univariable differentiation as follows.

Definition 8.9. A function $f : \mathbb{R} \to \mathbb{R}$ is differentiable at $a \in \mathbb{R}$ if there exists a linear transformation $c : \mathbb{R} \to \mathbb{R}$ such that

$$\lim_{h\to 0} \frac{f(a+h) - f(a) - c(h)}{h} = 0.$$

We can similarly generalize the derivative for multivariable function as

Definition 8.10. A function $f : \mathbb{R}^n \to \mathbb{R}^m$, where $A \subset \mathbb{R}^n$ is differentiable at $a \in \mathbb{R}^n$ if there exists a linear transformation $c : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{h\to 0} \frac{f(a+h) - f(a) - c(h)}{h} = 0.$$

This is equivalent to saying that there exists a matrix $m \times n$ B such that

$$\lim_{h \to 0} \frac{f(a+h) - f(a) - B.h}{|h|} = 0.$$

Example 8.11. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be defined by the equation

$$f(x) = Bx + b,$$

where, B is an $m \times n$ matrix and $b \in \mathbb{R}^m$. Then f is differentiable and $Df(x) = B$. Indeed, since

$$f(a+h) - f(h) = Bh,$$

the quotient used in defining the derivative vanishes identically.

Theorem 8.12. Let $A \subset \mathbb{R}^n$ and $f : A \to \mathbb{R}^m$. If f is differentiable at a, then all the directional derivatives of f at a exist, and

$$f'(a; u) = Df(a).u.$$

Proof. Let $c = Df(a)$. Set $h = tu$ in the definition of differentiability, where $t \neq 0$. Then by hypothesis

$$\lim_{t \to 0} \frac{f(a+tu) - f(a) - c.tu}{|tu|} = 0.$$

If $t \to 0$ through positive values, we multiply the above equation by $|u|$ to conclude that

$$\lim_{t \to 0} \frac{f(a+tu) - f(a)}{t} - c.u = 0.$$

If t approaches 0 through negative values, then we multiply the first equation by $-|u|$ to reach the same conclusion. Thus, $f'(a; u) = c.u = Df(a)$. □

But the converse is not true in general.

Example 8.13. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by setting $f(0) = 0$ and

$$\begin{aligned} f(x,y) &= \frac{x^2 y}{x^4 + x^2}; \quad (x,y) \neq (0,0) \\ &= 0; \qquad\quad (x,y) = (0,0). \end{aligned}$$

We show all directional derivatives of f exist at 0, but that f is not differentiable at 0. Let $u \neq 0$. Then for $u = (h, k)$ we have

$$\frac{f(0+tu) - f(0)}{t} = \frac{(th)^2(tk)}{(th)^4 + (tk)^2} \frac{1}{t} = \frac{h^2 k}{(t^2 h^4 + k^2},$$

so that

$$\begin{aligned} f'(0; u) &= \frac{h^2}{k}; \quad k \neq 0, \\ &= 0; \qquad k = 0. \end{aligned}$$

Thus $f'(0; u)$ exists for all $u \neq 0$. However, the function f is not differentiable at 0. For if $g : \mathbb{R}^2 \to \mathbb{R}$ is a function that is differentiable at 0, then $Dg(0)$ is a 1×2 matrix of the form $[a \;\; b]$, and

$$g'(0; u) = ah + bk,$$

which is a linear function of u. But $f'(0; u)$ is not a linear function of u.

In this form the definition has a simple generalization to higher dimensions.

Definition 8.14. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $a \in \mathbb{R}^n$ if there is a linear transformation $c : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{h\to 0} \frac{|f(a+h) - f(a) - c(h)|}{|h|} = 0.$$

Since h is a point of $\mathbb{R}^n$ and $f(a+h) - f(a) - c(h)$ is a point of $\mathbb{R}^m$, so the norm sign is necessary. If the above limit exists , then the linear transformation c is called the derivative of f at a and is denoted by $Df(a)$.

We have shown that at least one linear transformation exists for the differentiable function f. But one might get curious about the existence of more than one such functions. This is answered by the following theorem.

Theorem 8.15. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $a \in \mathbb{R}^n$, then there is a unique linear transformation $c : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{h\to 0} \frac{|f(a+h) - f(a) - c(h)|}{|h|} = 0$$

holds.

Proof. Let $d : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation that satisfies

$$\lim_{h\to 0} \frac{|f(a+h) - f(a) - d(h)|}{|h|} = 0.$$

If $m(h) = f(a+h) - f(a)$, then

$$\begin{aligned} \lim_{h\to 0} \frac{|c(h) - d(h)|}{|h|} &= \lim_{h\to 0} \frac{|c(h) - m(h) + m(h) - d(h)|}{|h|} \\ &\leq \lim_{h\to 0} \frac{|c(h) - m(h)|}{|h|} + \lim_{h\to 0} \frac{|m(h) - d(h)|}{|h|} = 0. \end{aligned}$$

If $x \in \mathbb{R}^n$, then $tx \to 0$ as $t \to 0$. Hence for $x \neq 0$, we have

$$0 = \lim_{h\to 0} \frac{|c(tx) - d(tx)|}{|tx|} = \frac{|c(x) - d(x)|}{|x|}.$$

Hence, $c(x) = d(x)$. □

Example 8.16. Define $f : \mathbb{R}^2 \to \mathbb{R}$ as $f(x, y) = \sin x$. Thus, $Df(a, b) = c$ satisfies $c(x, y) = (\cos a).x$. We will prove this.

$$\begin{aligned} &\lim_{(h,k)\to(0,0)} \frac{|f(a+h, b+k) - f(a, b) - c(h, k)|}{|(h, k)|} \\ &= \lim_{(h,k)\to(0,0)} \frac{|\sin(a+h) - \sin a - (\cos a).h|}{|h|}. \end{aligned}$$

Since $\sin'(a) = \cos a$, we have

$$\lim_{h\to 0} \frac{|\sin(a+h) - \sin a - (\cos a).h|}{|h|} = 0.$$

Since $|(h, k)| \geq |h|$, it is also true that

$$\lim_{h\to 0} \frac{|\sin(a+h) - \sin a - (\cos a).h|}{|(h, k)|} = 0.$$

8.4 Jacobian Matrix

It is often convenient to consider the matrix of $Df(a) : \mathbb{R}^n \to \mathbb{R}^m$ with respect to the usual bases of $\mathbb{R}^n$ and $\mathbb{R}^m$. This $m \times n$ matrix is called the Jacobian matrix of f at a, and denoted by $f'(a)$. For the previous example, we have, $f'(a, b) = (\cos a, 0)$. If $f : \mathbb{R} \to \mathbb{R}$, then $f'(a)$ is a 1×1 matrix matrix whose single entry is the number which is denoted by $f'(a)$ in single variable calculus.

The definition of $Df(a)$ could be made iff were defined only in some open set containing a. Considering only functions defined on $\mathbb{R}^n$ streamlines the statement of theorems and produces no real loss of generality. It is convenient to define a function $f : \mathbb{R}^n \to \mathbb{R}^m$ to be differentiable on A iff is differentiable at a for each $a \in A$. If $f : A \to \mathbb{R}^m$, then f is called differentiable if it can be extended to a differentiable function on some open set containing A.

Theorem 8.17. 1. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a constant function, then we have $Df(a) = 0$.

2. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation, then $Df(a) = f$.

3. If $f : \mathbb{R}^n \to \mathbb{R}^m$, then f is differentiable at $a \in \mathbb{R}^n$ if and only if each f^i is differentiable, and

$$Df(a) = (Df^1(a), \ldots, Df^m(a)).$$

   Thus, $f^{'}(a)$ is the $m \times n$ matrix whose ith row is $(f^i)^{'}(a)$.

4. If $s : \mathbb{R}^2 \to \mathbb{R}$ is defined by $s(x, y) = x + y$, then $D_s(a, b) = s$.

5. If $p : \mathbb{R}^2 \to \mathbb{R}$ is defined by $p(x, y) = x.y$, then $D_p(a, b)(x, y) = bx + ay$. Thus $p'(a, b) = (b, a)$.

Proof. 1. Left as exercise.

2. Left as exercise.

3. If each f^i is differentiable at a and $c = (Df^1(a), \ldots, Df^m(a))$, then

$$f(a + h) - f(a) - c(h) = (f^1(a + h) - f^1(a) - Df^1(a)(h), \ldots, f^m(a + h) - f^m(a) - Df^m(a)(h)).$$

   Hence

$$\lim_{h\to 0} \frac{|f(a + h) - f(a) - c(h)|}{|h|} \leq \lim_{h\to 0} \sum_{i=1}^{m} \frac{|f^i(a + h) - f^i(a) - Df^i(a)(h)|}{|h|} = 0.$$

   Also, if f is differentiable at a, then $f^i = \pi^i \circ f$ is differentiable at a by 2.

4. Follows from 2.

5. Let $c(x, y) = bx + ay$. Then

$$\lim_{(h,k)\to 0} \frac{|p(a + h, b + k) - p(a, b) - c(h, k)|}{|(h, k)|} = \lim_{(h,k)\to 0} \frac{|hk|}{|(h, k)|}.$$

   Now,

$$\begin{aligned} |hk| &\leq |h|^2, \ |k| \leq |h|, \\ &\leq |k|^2, \ |h| \leq |k|. \end{aligned}$$

Hence $|hk| \leq |h|^2 + |k|^2$. Thus,

$$\frac{|hk|}{|(h,k)|} \leq \frac{h^2+k^2}{\sqrt{h^2+k^2}} = \sqrt{h^2+k^2},$$

so

$$\lim_{(h,k)\to 0} \frac{|hk|}{|(h,k)|} = 0.$$

□

We can immediately get the following

Corollary 8.18. If $f, g : \mathbb{R}^n \to \mathbb{R}$ are differentiable at a, then

$$\begin{aligned} D(f+g)(a) &= Df(a) + Dg(a), \\ D(f.g)(a) &= g(a)Df(a) + f(a)Dg(a). \end{aligned}$$

Also, if $g(a) \neq 0$, then

$$D(f/g)(a) = \frac{g(a)Df(a) + f(a)Dg(a)}{[g(a)]^2}.$$

Exercise 8.19. 1. Find f' for each of the following functions:

(a) $f(x,y,z) = x^y$.
(b) $f(x,y,z) = (x^y, z)$.
(c) $f(x,y,z) = (x+y)^z$.

2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\begin{aligned} f(x,y) &= \frac{x|y|}{\sqrt{x^2+y^2}}, \quad (x,y) \neq 0, \\ &= (0,0), \quad (x,y) = 0. \end{aligned}$$

Show that f is not differentiable at $(0,0)$.

3. Let $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = \sqrt{|xy|}$. Show that f is not differentiable at $(0,0)$.

8.5 Partial Derivatives

We are somewhat familiar with the idea of partial derivatives when we learnt the two-variable calculus. When we keep a variable constant and change the other variable, then the rate of change of the second variable with respect to the latter is called the partial derivative or simply, derivative with respect to the latter variable. When we generalize this, for any function $f : \mathbb{R}^n \to \mathbb{R}$ and $a = (a^1, \ldots, a^n) \in \mathbb{R}^n$, then the ith partial derivative of f at a is defined as

$$D_i f(a) = \lim_{h\to 0} \frac{f(a^1, \ldots, a^i + h, \ldots, a^n) - f(a^1, \ldots, a^n)}{h},$$

provided the above limit exists. We can also consider $D_i f(a)$ as the ordinary derivative of a certain function; in fact, if $g(x) = f(a^1, \ldots, x, \ldots, a^n)$, then $D_i f(a) = g'(a^i)$. This means that $D_i f(a)$ is the slope of the tangent line at $(a, f(a))$ to the curve obtained by intersecting the graph of f with the plane $x^j = a^j,\ j \neq i$.

If $D_i f(x)$ exists for all $x \in \mathbb{R}^n$, we obtain a function $D_i f : \mathbb{R}^n \to \mathbb{R}$. The jth partial derivative of $D_i f$ at x, that is, $D_j(D_i f)(x)$, often written as $D_{i,j} f(x)$. The order of writing i and j can not be always reversed. We thus come to the following

Theorem 8.20. If $D_{i,j}f$ and $D_{j,i}f$ are continuous in an open set containing a, then $D_{i,j}f(a) = D_{j,i}f(a)$.

We will come back to this proof in the next unit.

Theorem 8.21. If $f : \mathbb{R}^n \to \mathbb{R}^m$, then $Df(a)$ exists if all $D_j f^i(x)$ exist in an open set containing a and if each function $D_j f^i$ is continuous at a.

Proof. It suffices to consider the case $m = 1$, so that $f : \mathbb{R}^n \to \mathbb{R}$. Then

$$\begin{aligned} f(a+h) - f(a) = f(a^1+h^1, a^2, \ldots, a^n) - f(a^1, \ldots, a^n) + f(a^1+h^1, a^2+h^2, a^3, \ldots, a^n) \\ - f(a^1+h^1, \ldots, a^n) + \cdots + \\ f(a^1+h^1, a^2+h^2, a^3+h^3, \ldots, a^n+h^n) - f(a^1+h^1, a^2+h^2, a^3+h^3, \ldots, a^{n-1}+h^{n-1}, a^n). \end{aligned}$$

Recall that $D_1 f$ is the derivative of the function g defined by $g(x) = f(x, a^2, \ldots, a^n)$. Applying the mean-value theorem to g we obtain

$$f(a^1+h^1, a^2, \ldots, a^n) - f(a^1, \ldots, a^n) = h^1.D_1(b_1, a^2, \ldots, a^n),$$

for some b_1 between a^1 and $a^1 + h^1$. Similarly, the ith term in the sum equals

$$h^i.D_i(a^1+h^1, \ldots, a^{i-1}+h^{i-1}, b_i, \ldots, a^n) = h^i D_i f(c_i),$$

for some c_i. Then

$$\begin{aligned} \lim_{h\to 0} \frac{\left| f(a+h) - f(a) - \sum_{i=1}^{n} D_i f(a).h^i \right|}{|h|} &= \lim_{h\to 0} \frac{\left| \sum_{i=1}^{n} [D_i f(c_i) - D_i f(a)].h^i \right|}{|h|} \\ &\leq \lim_{h\to 0} \sum_{i=1}^{n} |D_i f(c_i) - D_i f(a)|.\frac{|h^i|}{|h|} \\ &\leq \lim_{h\to 0} \sum_{i=1}^{n} |D_i f(c_i) - D_i f(a)| = 0, \end{aligned}$$

since $D_i f$ is continuous at a. □

Exercise 8.22. 1. Find the partial derivatives of the following functions:

(a) $f(x, y, z) = x^y$.

(b) $f(x, y, z) = z$.

(c) $f(x, y, z) = (x + y)^z$.

2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by setting $f(0) = 0$ and

$$f(x, y) = \frac{xy}{x^2 + y^2}, \quad (x, y) = 0.$$

(a) For which vectors $u \neq 0$ does $f'(0; u)$ exist? Evaluate it when it exists.

(b) Do $D_1 f$ and $D_2 f$ exist at 0?

(c) Is f differentiable at 0?

(d) Is f is continuous at 0?

3. Let f be defined as

$$\begin{aligned} f(x,y) &= \frac{x^2y^2}{x^2y^2+(y-x)^2}; \quad (x,y)\neq 0 \\ &= 0; \qquad (x,y)=0. \end{aligned}$$

Repeat exercise 2 for this function.

4. Let f be defined as

$$\begin{aligned} f(x,y) &= \frac{x^3}{x^2+y^2}; \quad (x,y)\neq 0 \\ &= 0; \qquad (x,y)=0. \end{aligned}$$

Repeat exercise 2 for this function.

5. Let $f:\mathbb{R}^2\to\mathbb{R}$ be defined by

$$\begin{aligned} f(x,y) &= xy\frac{x^2-y^2}{x^2+y^2}; \quad (x,y)\neq 0 \\ &= 0; \qquad (x,y)=0. \end{aligned}$$

(a) Show that $D_2f(x,0)=x$ for all x and $D_1f(0,y)=-y$ for all y.

(b) Show that $D_{1,2}f(0,0)\neq D_{2,1}f(0,0)$.

8.6 Jacobian Matrix Continued

We have already mentioned the Jacobian matrix for a differentiable function $f:\mathbb{R}^n\to\mathbb{R}^m$. If $A\subset\mathbb{R}^n$, then the derivative of the function $f:A\to\mathbb{R}^m$ at a point $a\in A$, also called the total derivative of f at a is defined as

$$Df(a)=\begin{bmatrix} D_1f^1(a) & D_2f^1(a) & \cdots & D_nf^1(a) \\ D_1f^2(a) & D_2f^2(a) & \cdots & D_nf^2(a) \\ \vdots & \vdots & \ddots & \vdots \\ D_1f^m(a) & D_2f^m(a) & \cdots & D_nf^m(a) \end{bmatrix}$$

We will check certain examples regarding the Jacobian matrix of a function.

Illustration 8.23. Let a function $f:\mathbb{R}^3\to\mathbb{R}^2$ be given by $f(x,y,z)=(xy+2yz,2xy^2z)$. Then we have,

$$\begin{aligned} f^1 &= xy+2yz \\ f^2 &= 2xy^2z. \end{aligned}$$

In order to find the Jacobian of f, we determine each of the following.

$$\begin{aligned} D_xf^1(x,y)=y, \quad D_yf^1(x,y)=x+2z, \quad D_zf^1(x,y)=2y, \\ D_xf^2(x,y)=2y^2z, \quad D_yf^2(x,y)=4xyz, \quad D_zf^2(x,y)=2xy^2. \end{aligned}$$

Thus, the Jacobian matrix function at any point $(x,y,z)\in\mathbb{R}^3$ is given by

$$Df(x,y,z)=\begin{bmatrix} y & x+2z & 2y \\ 2y^2 & 4xyz & 2xy^2 \end{bmatrix}.$$

The determinant of the matrix $Df(x)$, provided $m = n$, is called the Jacobian determinant. The Jacobian determinant at a given point gives important information about the behaviour of f near a point. For instance, the continuously differentiable function f is invertible near a point $a \in \mathbb{R}^n$ if the Jacobian determinant at a is non-zero. Further, if it is positive, then f preserves the orientation near a and if it is negative, then f reverses the orientation near a. Also, the absolute value of the determinant gives us the factor by which f expands or shrinks near a.

Exercise 8.24.

1. Find the Jacobian matrix of the function $f(x, y) = (xy, x + y)$. Also calculate the determinant. Hence find the total derivative of f at $(2, 1)$.
2. Find the Jacobian matrix and determinant of the function $f(x, y) = (x^2y, 5x + \sin y)$.
3. Find the Jacobian matrix and determinant(if possible) of the function $f(x, y, z) = (x, 5z, 4y^2-2z, z \sin x)$.
4. Find the Jacobian determinant of the function $f(x, y, z) = (5y, 4x^2 - 2\sin(yz), yz)$.

8.7 Few Probable Questions

1. Define continuity of a multivariable function $f : A \to \mathbb{R}^m$ where $A \subset \mathbb{R}^n$. Show that f is continuous at a point $a \in A$ if and only if each of the components f^i is so.
2. Define the directional derivative of a function $f : A \to \mathbb{R}^m$ where $A \subset \mathbb{R}^n$. Find the directional derivative of f where f is defined as
$$f(x, y) = \cos\left(\frac{x}{y}\right)$$
in the direction $(3, -4)$.
3. Show that if a function $f : A \to \mathbb{R}^m$ where $A \subset \mathbb{R}^n$ is differentiable at $a \in A$, then all the directional derivatives of f at a exist, and $f'(a; u) = Df(a).u$. Is the converse true? Justify.
4. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by setting $f(0) = 0$ and
$$f(x, y) = \frac{xy}{x^2 + y^2}, \quad (x, y) = 0.$$
   (a) For which vectors $u \neq 0$ does $f'(0; u)$ exist? Evaluate it when it exists.
   (b) Do D_1f and D_2f exist at 0?
   (c) Is f differentiable at 0?
   (d) Is f is continuous at 0?
5. Show that if $f : A \to \mathbb{R}^m$, where $A \subset \mathbb{R}^n$ is differentiable at $a \in A$, then f is continuous there. Is the converse true? Justify.
6. Define the ith partial derivative of a function $f : A \to \mathbb{R}^m$, where $A \subset \mathbb{R}^n$ at a point $a \in A$. Show that if all $D_jf^i(x)$ exist in an open set containing a and if each function D_jf^i is continuous at a.

Unit 10

Course Structure

- The chain rule and its matrix form.
- Mean value theorem for vector valued function. Mean value inequality.
- A sufficient condition for differentiability. A sufficient condition for mixed partial derivatives

---

9 Introduction

We are already familiar with the basic ideas of derivatives. This unit is an extension of the previous one. In single-variable calculus, we found that one of the most useful differentiation rules is the chain rule, which allows us to find the derivative of the composition of two functions. The same thing is true for multivariable calculus, but this time we have to deal with more than one form of the chain rule. In this section, we study extensions of the chain rule and learn how to take derivatives of compositions of functions of more than one variable.

Objectives

After reading this unit, you will be able to

- learn the chain rule for multivariable functions and do certain related problems
- learn a sufficient condition for differentiability
- learn a sufficient condition for mixed partial derivatives

9.1 Chain Rule and its Matrix form

For the univariable functions $f, g : \mathbb{R} \to \mathbb{R}$, then the chain rule gives the derivative for two composite functions $f \circ g$ given as

$$\frac{d}{dx}(f \circ g(x)) = \frac{d}{dx}(f(g(x))) = \frac{df}{dg}\frac{dg}{dx}.$$

We generalize this for multivariable function in the following

Theorem 9.1. (Chain Rule). If $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at a, and $f : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(a)$, then the composition $g \circ f : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at a, and

$$D(g \circ f)(a) = Dg(f(a)) \circ Df(a).$$

For $m = n = p = 1$, then we get our old chain rule.

Proof. Let $b = f(a)$ and let $c = Df(a)$ and let $d = Dg(f(a))$. Let us define

1. $\phi(x) = f(x) - f(a) - c(x - a)$,

2. $\psi(x) = g(y) - g(b) - d(y-b)$,

3. $\rho(x) = g \circ f(x) - g \circ f(a) - d \circ c(x-a)$,

then

$$\lim_{x\to a} \frac{|\phi(x)|}{|x-a|} = 0, \tag{9.1.1}$$

$$\lim_{y\to b} \frac{|\psi(y)|}{|y-b|} = 0, \tag{9.1.2}$$

and we must show that

$$\lim_{x\to a} \frac{|\rho(x)|}{|x-a|} = 0.$$

Now,

$$\begin{aligned}
\rho(x) &= g(f(x)) - g(b) - d(c(x-a)) \\
&= g(f(x)) - g(b) - d(f(x) - f(a) - \psi(x)) \\
&= [g(f(x)) - g(b) - d(f(x) - f(a))] + d(\psi(x)) \\
&= \psi(f(x)) + d(\psi(x)).
\end{aligned}$$

Thus, we must prove

$$\lim_{x\to a} \frac{|\psi(f(x))|}{|x-a|} = 0, \tag{9.1.3}$$

$$\lim_{x\to a} \frac{|d(\phi(x))|}{|x-a|} = 0. \tag{9.1.4}$$

Equation (9.1.4) follows easily from (9.1.1). If $\epsilon > 0$, it follows from (9.1.2) that for some $\delta > 0$, we have

$$|\psi(f(x))| < \epsilon |f(x) - b| \quad \text{if} \quad |f(x) - b| < \delta,$$

which is true if $|x-a| < \delta_1$ for some suitable δ_1. Then

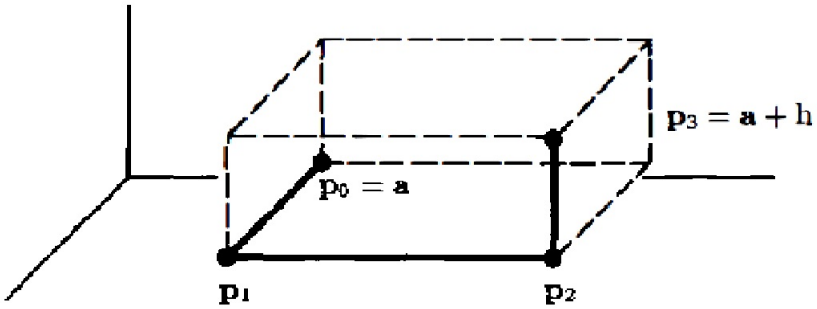
$$\begin{aligned}
|\psi(f(x))| &< \epsilon |f(x) - b| \\
&= \epsilon |\phi(x) + c(x-a)| \\
&\leq \epsilon |\phi(x)| + \epsilon M |x-a|
\end{aligned}$$

for some M. Equation (9.1.3) now follows. □

Exercise 9.2. 1. Find f' for the following functions using Chain Rule

(a) $f(x,y) = \sin(xy)$

(b) $f(x,y) = (\sin(xy), \sin(x \sin y), x^y)$.



9.2 Mean Value Theorems for Vector-Valued Functions

In this section, we obtain a useful criterion for differentiability. We know that mere existence of the partial derivatives does not imply differentiability. If, however, we impose the (comparatively mild) additional condition that these partial derivatives be continuous, then differentiability is assured.

We begin by recalling the mean-value theorem of single-variable analysis.

Theorem 9.3. If $f : [a, b] \to \mathbb{R}$ is continuous at each point of the closed interval $[a, b]$, and differentiable at each point of the open interval (a, b) , then there exists a point c of (a, b) such that

$$f(b) - f(a) = f'(c)(b - a).$$

In practice, we most often apply this theorem when f is differentiable on an open interval containing $[a, b]$. In this case, of course , f is continuous on $[a, b]$.

We know that, if $f : A \to \mathbb{R}^m$, where $A \subset \mathbb{R}^n$, is differentiable at $a \in A$, then

$$Df(a) = \begin{bmatrix} D_1f^1(a) & D_2f^1(a) & \cdots & D_nf^1(a) \\ D_1f^2(a) & D_2f^2(a) & \cdots & D_nf^2(a) \\ \vdots & \vdots & \ddots & \vdots \\ D_1f^m(a) & D_2f^m(a) & \cdots & D_nf^m(a) \end{bmatrix}.$$

Theorem 9.4. Let A be an open subset of $\mathbb{R}^n$. Suppose that the partial derivatives $D_jf^i(x)$ of the component functions of f exist at each point x of A and are continuous on A . Then f is differentiable at each point of A.

Such a function f is called continuously differentiable function, or of class C^1, on A.

Proof. It is sufficient to prove that each component function of f is differentiable. Hence, we may restrict ourselves to the case of a real-valued function $f : A \to \mathbb{R}$.

Let a be a point of A. We are given that, for some ϵ, the partial derivatives $D_jf(x)$ exist and are continuous for $|x - a| < \epsilon$. We wish to show that f is differentiable at a.

Let $h \in \mathbb{R}^n$ with $0 < |h| < \epsilon$; let $h_1, \ldots, h_n$ be the components of h. Consider the following sequence of points of $\mathbb{R}^n$:

$$\begin{aligned} p_0 &= a, \\ p_1 &= a + h_1e_1, \\ p_2 &= a + h_1e_1 + h_2e_2, \\ \cdots & \\ p_n &= a + h_1e_1 + \cdots + h_ne_n = a + h. \end{aligned}$$

The points p_i all belong to the closed cube C of radius $|h|$ centered at a. The figure illustrates the case when $n = 3$ and all h_i are positive.

Since we are concerned with the differentiability of f, we shall need to deal with the difference $f(a+h) - f(a)$. We begin by writing it in the form

$$f(a+h) - f(a) = \sum_{j=1}^{n} [f(p_j) - f(p_{j-1})]. \tag{9.2.1}$$

Consider the general term of this summation. Let j be fixed, and define

$$\phi(t) = f(p_{j-1} + te_j).$$

Assume $h_j \neq 0$ for the moment. As t ranges over the closed interval I with end points 0 and h_j, the point $p_{j-1} + te_j$ ranges over the line segment from p_{j-1} to p_j; this line segment lies in C, and hence in A. Thus, ϕ is defined for t in an open interval about I.

As t varies, only the jth component of the point $p_{j-1} + te_j$ varies. Hence because $D_j f$ exists at each point of A , the function ϕ is differentiable on an open interval containing I . Applying the mean-value theorem to ϕ, we conclude that

$$\phi(h_j) - \phi(0) = \phi'(c_j)h_j$$

for some point c_j between 0 and h_j. (This argument applies whether h_j is positive or negative.) We can rewrite this equation in the form

$$f(p_j) - f(p_{j-1}) = D_j f(q_j)h_j, \tag{9.2.2}$$

where q_j is the point $p_{j-1} + c_j e_j$ of the line segment from p_{j-1} to p_j, which lies in C.

We derived (9.2.2) under the assumption that $h_j \neq 0$. If $h_j = 0$, then (9.2.2) holds automatically, for any point q_j of C.

Using (9.2.2), we rewrite (9.2.1) in the form

$$f(a+h) - f(a) = \sum_{j=1}^{n} D_j f(q_j)h_j, \tag{9.2.3}$$

where each point q_j lies in the cube C of radius $|h|$ centered at a.

We now prove the theorem. Let

$$B = [D_1 f(a) \dots D_n f(a)].$$

Then

$$B.h = \sum_{j=1}^{n} D_j f(a)h_j.$$

Using (9.2.3), we have

$$\frac{f(a+h) - f(a) - B.h}{|h|} = \sum_{j=1}^{n} \frac{[D_j f(q_j) - D_j f(a)]h_j}{|h|};$$

then we let $h \to 0$. Since q_j lies in the cube C of radius $|h|$ centered at a, we have $q_j \to a$. Since the partials of f are continuous at a, the factors in brackets all go to zero. The factors $h_j/|h|$ are of course bounded in absolute value by 1. Hence the entire expression goes to zero, as desired. □

One effect of this theorem is to reassure us that the functions familiar to us from calculus are in fact differentiable. We know how to compute the partial derivatives of such functions as $\sin(xy)$ and $xy^2 + z\,\mathrm{e}^{xy}$, and we know that these partials are continuous. Therefore these functions are differentiable.

In practice, we usually deal only with functions that are of class C^1. While it is interesting to know there are functions that are differentiable but not of class C^1, such functions occur rarely enough that we need not be concerned with them.

Suppose f is a function mapping an open set A of $\mathbb{R}^n$ into $\mathbb{R}^n$, and suppose the partial derivatives $D_j f^i$ of the component functions of f exist on A. These then are functions from A to $\mathbb{R}$, and we may consider their partial derivatives, which have the form $D_k(D_j f_i)$ and are called the second-order partial derivatives of f. Similarly, one defines the third-order partial derivatives of the functions f_i or more generally the partial derivatives of order r for arbitrary r.

If the partial derivatives of the functions f^i of order less than or equal to r are continuous on A, we say f is of class C^r on A. Then the function f is of class C^r on A, if and only if each of the functions $D_j f^i$ is of class C^{r-1} on A. We say f is of class C^∞ on A, if the partials of the functions f^i of all orders are continuous on A.

As you may recall, for most functions the "mixed" partial derivatives

$$D_k D_j f^i \quad \& \quad D_j D_k f^i$$

are equal. This result in fact holds under the hypothesis that the function f is of class C^2, as we now show.

Theorem 9.5. Let A be open in $\mathbb{R}^n$; let $f : A \to \mathbb{R}$ be a function of class C^2. Then for each $a \in A$, we have

$$D_k D_j f(a) = D_j D_k f(a).$$

Proof. Since one calculates the partial derivatives in question by letting all variables other than x_k and x_j remain constant, it suffices to consider the case where f is a function merely of two variables. So we assume that A is open in $\mathbb{R}^2$, and that $f : A \to \mathbb{R}^2$ is of class C^2.

Let

$$Q = [a, a+h] \times [b, b+k]$$

be a rectangle contained in A. Define

$$\lambda(h,k) = f(a,b) - f(a+h,b) - f(a,b+k) + f(a+h,b+k).$$

Then λ is the sum, with appropriate signs, of the values of f at the four vertices of Q. We show that there are points p and q of Q such that

$$\lambda(h,k) = D_2 D_1 f(p).hk, \quad \& \quad \lambda(h,k) = D_1 D_2 f(q).hk.$$

By symmetry, it suffices to prove the first of these equations. To begin, we define

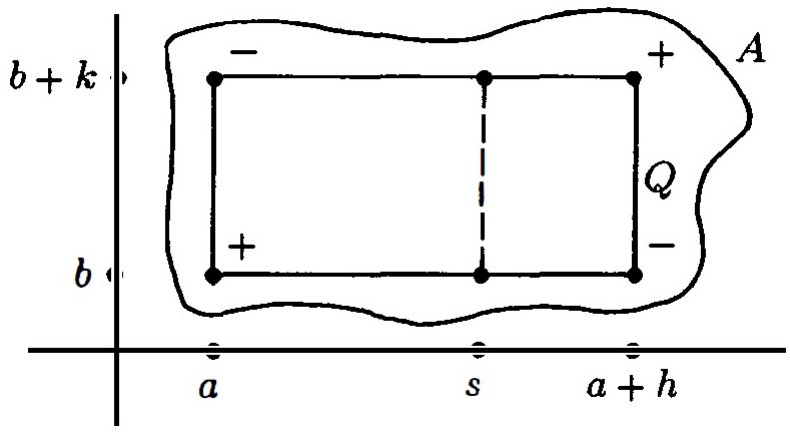
$$\phi(s) = f(s, b+k) - f(s,b).$$

Then $\phi(a+h) - \phi(a) = \lambda(h,k)$. Because $D_1 f$ exists in A, the function ϕ is differentiable in an open interval containing $[a, a+h]$. The mean-value theorem implies that

$$\phi(a+h) - \phi(a) = \phi'(s_0).h$$

for some s_0 between a and $a+h$. This equation can be rewritten in the form

$$\lambda(h,k) = [D_1 f(s_0, b+k) - D_1 f(s_0, b)].h. \tag{9.2.4}$$



Now, s_0 is fixed, and we consider the function $D_1 f(s_0, t)$. Because $D_2 D_1 f$ exists in A, this function is differentiable for t in an open interval about $[b, b+k]$. We apply the mean-value theorem once more to conclude that

$$D_1 f(s_0, b+k) - D_1 f(s_0, b) = D_2 D_1 f(s_0, t_0).k \tag{9.2.5}$$

for some t_0 between b and $b+k$. Combining (9.2.4) and (9.2.5), we get,

$$\lambda(h, k) = D_2 D_1 f(s_0, t_0).hk. \tag{9.2.6}$$

Now, we prove the theorem. Given the point $a = (a, b)$ of A and given $t > 0$, let Q_t be the rectangle

$$Q_t = [a, a+t] \times [b, b+t].$$

If t is sufficiently small, Q_t is contained in A. Then by (9.2.6), we get

$$\lambda(t, t) = D_2 D_1 f(p_t).t^2$$

for some point p_t in Q_t. If we let $t \to 0$, then $p_t \to a$. Because $D_2 D_1 f$ is continuous, it follows that

$$\lambda(t, t)/t^2 \to D_2 D_1 f(a) \quad \text{as} \quad t \to 0.$$

A similar argument, using symmetry, gives

$$\lambda(t, t)/t^2 \to D_1 D_2 f(a) \quad \text{as} \quad t \to 0.$$

Hence the theorem. □

As another application of the chain rule, we generalize the mean-value theorem of single-variable analysis to real-valued functions defined in $\mathbb{R}^n$.

Theorem 9.6. (**Mean Value Theorem:**) Let A be open in $\mathbb{R}^n$; let $f : A \to \mathbb{R}$ be differentiable on A . If A contains the line segment with end points a and $a+h$, then there exists a point $c = a + t_0 h$ with $0 < t_0 < 1$ of this line segment such that

$$f(a+h) - f(a) = Df(c).h.$$

Proof. Set $\phi(t) = f(a + th)$; then ϕ is defined for t in an open interval about $[0, 1]$. Being the composite of differentiable functions, ϕ is differentiable; its derivative is given by the formula

$$\phi'(t) = Df(a + th).h.$$

The ordinary mean-value theorem implies that

$$\phi(1) - \phi(0) = \phi'(t_0).1$$

for some t_0 with $0 < t_0 < 1$. This equation can be rewritten in the form

$$f(a + h) - f(a) = Df(a + t_0h).h.$$

□

Exercise 9.7. 1. Let $f : \mathbb{R}^3 \to \mathbb{R}^2$ satisfy the conditions $f(0) = (1, 2)$ and

$$Df(0) = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the equation

$$g(x, y) = (x + 2y + 1, 3xy).$$

Find $D(g \circ f)(0)$.

2. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the equation

$$f(r, \theta) = (r \cos\theta, r \sin\theta).$$

   (a) Calculate Df and $\det Df$

   (b) Sketch the image under f of the set $S = [1, 2] \times [0, \pi]$.

3. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the equation

$$f(x, y) = (\mathrm{e}^x \cos y, \mathrm{e}^x \sin y).$$

Then

   (a) Calculate Df and $\det Df$

   (b) Sketch the image under f of the set $S = [0, 1] \times [0, \pi]$.

9.3 Few Probable Questions

1. Show that the function $f(x, y) = |xy|$ is differentiable at 0, but is not of class C^1 in any neighborhood of 0.

2. Define $f : \mathbb{R} \to \mathbb{R}$ by setting $f(0) = 0$ and

$$f(t) = t^2 \sin\left(\frac{1}{t}\right) \quad \text{if} \quad t \neq 0.$$

   (a) Show f is differentiable at 0, and calculate $f'(0)$.

   (b) Calculate $f'(t)$ if $t \neq 0$.

   (c) Show f' is not continuous at 0.

   (d) Conclude that f is differentiable on $\mathbb{R}$ but not of class C^1 on $\mathbb{R}$.

3. Show that if $A \subset \mathbb{R}^n$ and $f : A \to \mathbb{R}$, and if the partials $D_j f$ exist and are bounded in a neighborhood of a, then f is continuous at a.

Unit 11

Course Structure

- Functions with non-zero Jacobian determinant, the inverse function theorem
- The implicit function theorem as an application of Inverse function theorem.

---

10 Introduction

In mathematics, specifically differential calculus, the inverse function theorem gives a sufficient condition for a function to be invertible in a neighbourhood of a point in its domain: namely, that its derivative is continuous and non-zero at the point. The theorem also gives a formula for the derivative of the inverse function. In multivariable calculus, this theorem can be generalized to any continuously differentiable, vector-valued function whose Jacobian determinant is nonzero at a point in its domain, giving a formula for the Jacobian matrix of the inverse which we will explore here.

We will explore the implicit function theorem as an application of the inverse function theorem in this unit. In multivariable calculus, the implicit function theorem is a tool that allows relations to be converted to functions of several real variables. It does so by representing the relation as the graph of a function. There may not be a single function whose graph can represent the entire relation, but there may be such a function on a restriction of the domain of the relation. The implicit function theorem gives a sufficient condition to ensure that there is such a function.

Objectives

After reading this unit, you will be able to

- learn about the consequences of non-zero Jacobian determinant of vector valued functions
- learn the inverse function theorem and related theorems and lemmas
- apply the inverse function theorem in various examples
- learn the implicit function theorem as an application of the inverse function theorem
- apply the implicit function theorem in various problems

10.1 Functions with non-zero Jacobian determinant

We have read about the chain rule in the previous unit and the mean value theorem as an application of it. As yet another application of the chain rule, we consider the problem of differentiating an inverse function.

Recall the situation that occurs in single-variable analysis. Suppose $\phi(x)$ is differentiable on an open interval, with $\phi^{''}(x) > 0$ on that interval. Then ϕ is strictly increasing and has an inverse function ψ, which is defined by letting $\psi(y)$ be that unique number x such that $\phi(x) = y$. The function ψ is in fact differentiable, and its derivative satisfies the equation

$$\psi'(y) = \frac{1}{\phi'(x)},$$

where $y = \phi(x)$.

There is a similar formula for differentiating the inverse of a function f of several variables. In the present section, we do not consider the question whether the function f has an inverse, or whether that inverse is differentiable. We consider only the problem of finding the derivative of the inverse function.

Theorem 10.1. Let A be open in $\mathbb{R}^n$ and let $f : A \to \mathbb{R}^n$ such that $f(a) = b$. Suppose that g maps a neighbourhood of b into $\mathbb{R}^n$, such that $g(b) = a$ and $g(f(x)) = x$ for all x in a neighbourhood of a. If f is differentiable at a and if g is differentiable at b, then $Dg(b) = [Df(a)]^{-1}$.

Proof. Let $i : \mathbb{R}^n \to \mathbb{R}^n$ be the identity function; its derivative is the identity matrix I_n. We are given that $g(f(x)) = i(x)$ for all x in a neighbourhood of a. Then the chain rule implies that

$$Dg(b).Df(a) = I_n.$$

Thus, $Dg(b)$ is the inverse matrix to $Df(a)$. □

The preceding theorem implies that if a differentiable function f is to have a differentiable inverse, it is necessary that the matrix Df be non-singular. It is a somewhat surprising fact that this condition is also sufficient for a function f of class C^1 to have an inverse , at least locally.

Let us make a comment on notation. The usefulness of well-chosen notation can hardly be overemphasized. Arguments that are obscure, and formulas that are complicated, sometimes become beautifully simple once the proper notation is chosen. Our use of matrix notation for the derivative is a case in point . The formulas for the derivatives of a composite function and an inverse function could hardly be simpler. Nevertheless, a word may be in order for those who remember the notation used in calculus for partial derivatives, and the version of the chain rule proved there.

In advanced mathematics, it is usual to use either the functional notation ϕ' or the operator $D\phi$ for for the derivative of a real-valued function of a real variable. ($D\phi$ denotes a 1×1 matrix in this case!) In calculus, however, another notation is common. One often denotes the derivative $\phi'(x)$ by the symbol $d\phi/dx$.

10.2 The Inverse Function Theorem

Let A be open in $\mathbb{R}^n$ and let $f : A \to \mathbb{R}^n$ be of class C^1. We know that for f to have a differentiable inverse, it is necessary that the derivative $Df(x)$ of f be non-singular. We now prove that this condition is also sufficient for f to have a differentiable inverse, at least locally. This result is called the inverse function theorem.

We begin by showing that non-singularity of Df implies that f is locally one-to-one.

Lemma 10.2. Let A be open in $\mathbb{R}^n$ and let $f : A \to \mathbb{R}^n$ be of class C^1. If $Df(a)$ is non-singular, then there exists an $a > 0$ such that the inequality

$$|f(x_0) - f(x_1)| \geq a|x_0 - x_1|$$

holds for all x_0, x_1 in some open cube $C(a; \epsilon)$ centered at a. It follows that f is one-to-one on this open cube.

Proof. Let $E = Df(a)$. Then E is non-singular. We first consider the linear transformation that maps x to Ex. We compute

$$|x_0 - x_1| = |E^{-1}.(E.x_0 - E.x_1)| \leq |E^{-1}||E.x_0 - E.x_1|.$$

If we set $2a = 1/n|E^{-1}|$, then for all $x_0, x_1 \in \mathbb{R}^n$,

$$|E.x_0 - E.x_1| \geq 2a|x_0 - x_1|.$$

Now consider the function $H(x) = f(x) - E.x$. Then $DH(x) = Df(x) - E$, so that $DH(a) = 0$. Since H is of class C^1, we can choose $\epsilon > 0$ such that $|DH(x)| < a/n$ for x in the open cube $C = C(a; \epsilon)$. The

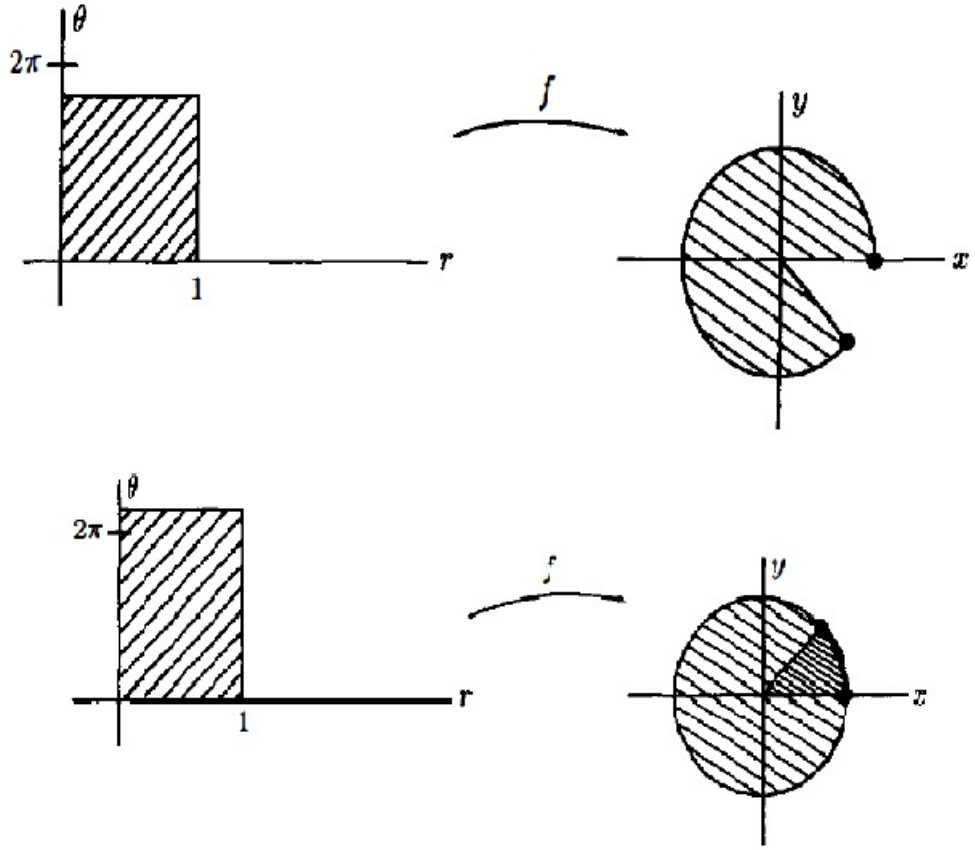


Figure 1: f in example 10.5

mean-value theorem, applied to the ith component function of H, tells us that, given $x_0, x_1 \in C$, there is a $c \in C$ such that

$$|H_i(x_0) - H_i(x_1)| = |DH_i(c).(x_0 - x_1)| \neq n(a/n)|x_0 - x_1|.$$

Thus for $x_0, x_1 \in C$, we have

$$\begin{aligned} a|x_0 - x_1| &\geq |H(x_0) - H(x_1)| \\ &= |f(x_0) - E.x_0 - f(x_1) + E.x_1| \\ &\geq |E.x_1 - E.x_0| - |f(x_1) - f(x_0)| \\ &\geq 2a|x_1 - x_0| - |f(x_1) - f(x_0)|. \end{aligned}$$

Hence the result. □

We will now state a theorem which says that the non-singularity of Df, in the case where f is one-to-one, implies that the inverse function is differentiable.

Theorem 10.3. Let A be open in $\mathbb{R}^n$ and let $f : A \to \mathbb{R}^n$ be of class C^r. Let $B = f(A)$. If f is one-to-one on A and if $Df(x)$ is non-singular for $x \in A$, then the set B is open in $\mathbb{R}^n$ and the inverse function $g : B \to A$ is of class C^r.

We leave the proof of this theorem. We will finally prove the inverse function theorem.

Theorem 10.4. (**Inverse Function Theorem**) Let A be open in $\mathbb{R}^n$ and let $f : A \to \mathbb{R}^n$ be of class C^r. If $Df(x)$ is non-singular at the point $a \in A$, there is a neighbourhood U of the point a such that f carries U in a one-to-one fashion onto an open set V of $\mathbb{R}^n$ and the inverse function is of class C^r.

Proof. By lamma 10.2, there is a neighborhood U_0 of a on which f is one-to-one. Because $\det Df(x)$ is a continuous function of x, and $\det Df(a) \neq 0$, there is a neighbourhood U_1 of a such that $\det Df(x) \neq 0$ on U_1. If $U = U_0 \cap U_1$, then the hypotheses of the preceding theorem are satisfied for $f : U \to \mathbb{R}^n$. The theorem follows. □

This theorem is the strongest one that can be proved in general . While the non-singularity of Df on A implies that f is locally one-to-one at each point of A, it does not imply that f is one-to-one on all of A. Consider the following example:

Example 10.5. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the equation

$$f(r, \theta) = (r \cos\theta, r \sin\theta).$$

Then

$$Df(r, \theta) = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix},$$

so that $\det Df(r, \theta) = r$.

Let A be the open set $(0, 1) \times (0, b)$ in the r-θ plane. Then Df is non-singular at each point of A. However, f is one-to-one on A only if $b \leq 2\pi$.

Exercise 10.6. 1. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the equation

$$f(x, y) = (x^2 - y^2, 2xy).$$

(a) Show that f is one-to-one on the set A consisting of all (x, y) with $x > 0$.

(b) What is the set $B = f(A)$?

(c) If g is the inverse function, find $Dg(0, 1)$.

2. Let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by the equation

$$f(x, y) = (\mathrm{e}^x \cos y, \mathrm{e}^x \sin y).$$

(a) Show that f is one-to-one on the set A consisting of all (x, y) with $0 < y < 2\pi$.

(b) What is the set $B = f(A)$?

(c) If g is the inverse function, find $Dg(0, 1)$.

10.3 Implicit Function Theorem

The topic of implicit differentiation is one that is probably familiar to you from calculus. Here is a typical problem:

Assume that the equation $x^3y + 2\,\mathrm{e}^{xy} = 0$ determines y as a differentiable function of x. Find dy/dx.

One solves this calculus problem by "looking at y as a function of x, " and differentiating with respect to x. One obtains the equation

$$3x^2y + x^3\frac{dy}{dx} + 2\,\mathrm{e}^{xy}\left(y + x\frac{dy}{dx}\right) = 0,$$

which one solves for dy/dx. The derivative dy/dx is of course expressed in terms of x and the unknown function y.

The case of an arbitrary function f is handled similarly. Supposing that the equation $f(x, y) = 0$ determines y as a differentiable function of x , say $y = g(x)$, the equation $f(x, g(x)) = 0$ is an identity. One applies the chain rule to calculate

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}g'(x) = 0,$$

which gives

$$g'(x) = -\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}},$$

where the partial derivatives are evaluated at the point $(x, g(x))$. Note that the solution involves a hypothesis not given in the statement of the problem. In order to find $g'(x)$, it is necessary to assume that $\partial f/\partial y$ is non-zero at the point in question.

It in fact turns out that the non-vanishing of $\partial f/\partial y$ is also sufficient to justify the assumptions we made in solving the problem. That is, if the function $f(x, y)$ has the property that $\partial f/\partial y \neq 0$ at a point (a, b) that is a solution of the equation $f(x, y) = 0$, then this equation does determine y as a function of x , for x near a , and this function of x is differentiable.

This result is a special case of a theorem called the implicit function theorem, which we prove in this section. The general case of the implicit function theorem involves a system of equations rather than a single equation. One seeks to solve this system for some of the unknowns in terms of the others. Specifically, suppose that $f : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is a function of class C^1. Then the vector equation

$$f(x_1, \ldots, x_{k+n}) = 0$$

is equivalent to a system of n scalar equations in $k + n$ unknowns. One would expect to be able to assign arbitrary values to k of the unknowns and to solve for the remaining unknowns in terms of these. One would also expect that the resulting functions would be differentiable, and that one could by implicit differentiation find their derivatives.

There are two separate problems here. The first is the problem of finding the derivatives of these implicitly defined functions, assuming they exist; the solution to this problem generalizes the computation of $g'(x)$ just given. The second involves showing that (under suitable conditions) the implicitly defined functions exist and are differentiable.

In order to state our results in a convenient form, we introduce a new notation for the matrix D f and its submatrices:

Definition 10.7. Let A be open in $\mathbb{R}^m$; let $f : A \to \mathbb{R}^n$ be differentiable and $f_1, \ldots, f_n$ be the component functions of f. We sometimes use the notation

$$Df = \frac{\partial(f_1, \ldots, f_n)}{\partial(x_1, \ldots, x_m)}$$

for the derivative of f. On occasion we shorten this to the notation

$$Df = \frac{\partial f}{\partial x}.$$

More generally, we shall use the notation

$$\frac{\partial(f_{i_1},\ldots,f_{i_k})}{\partial(x_{j_1},\ldots,x_{j_l})}$$

to denote the $k \times l$ matrix that consists of the entries of Df lying in rows $i_1,\ldots,i_k$ and columns $j_1,\ldots,j_l$. The general entry of this matrix, in row p and column q, is the partial derivative $\partial f_{i_p}/\partial x_{j_q}$.

Now we deal with the problem of fin ding the derivative of an implicitly defined function, assuming it exists and is differentiable. For simplicity, we shall assume that we have solved a system of n equations in $k+n$ unknowns for the last n unknowns in terms of the first k unknowns.

Theorem 10.8. Let A be open in $\mathbb{R}^{k+n}$; let $f : A \to \mathbb{R}^n$ be differentiable. Write f in the form $f(x,y)$, for $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^n$; then Df has the form

$$Df = \begin{bmatrix} \dfrac{\partial f}{\partial x} & \dfrac{\partial f}{\partial y} \end{bmatrix}.$$

Suppose there is a differentiable function $g : B \to \mathbb{R}^n$ defined on an open set B in $\mathbb{R}^k$, such that

$$f(x, g(x)) = 0$$

for all $x \in B$. Then for $x \in B$,

$$\frac{\partial f}{\partial x}(x,g(x)) + \frac{\partial f}{\partial y}(x,g(x)).Dg(x) = 0.$$

This equation implies that if the $n \times n$ matrix $\partial f/\partial y$ is non-singular at the point $(x, g(x))$, then

$$Dg(x) = -\left[\frac{\partial f}{\partial y}(x,g(x))\right]^{-1}.\frac{\partial f}{\partial x}(x,g(x)).$$

Note that in the case $n = k = 1$, this is the same formula for the derivative that was derived earlier; the matrices involved are 1×1 matrices in that case.

Proof. Given g, let us define $h : B \to \mathbb{R}^{k+n}$ by the equation

$$h(x) = (x, g(x)).$$

The hypotheses of the theorem imply that the composite function

$$H(x) = f(h(x)) = f(x, g(x))$$

is defined and equals zero for all $x \in B$. The chain rule then implies that

$$\begin{aligned} 0 &= DH(x) = Df(h(x)).Dh(x) \\ &= \begin{bmatrix} \dfrac{\partial f}{\partial x}(h(x)) & \dfrac{\partial f}{\partial y}(h(x)) \end{bmatrix} \cdot \begin{bmatrix} I_k \\ Dg(x) \end{bmatrix} \\ &= \frac{\partial f}{\partial x}(h(x)) + \frac{\partial f}{\partial y}(h(x)).Dg(x), \end{aligned}$$

as desired. □

The preceding theorem tells us that in order to compute Dg, we must assume that the matrix $\partial f/\partial y$ is non-singular. Now we prove that the non-singularity of $\partial f/\partial y$ suffices to guarantee that the function g exists and is differentiable.

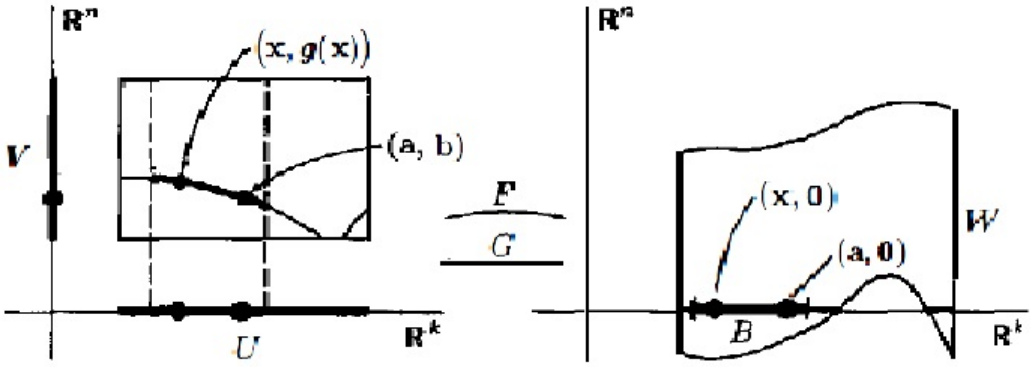


Figure 2: Implicit Function Theorem

Theorem 10.9. (Implicit function theorem). Let A be open in $\mathbb{R}^{k+n}$; let $f : A \to \mathbb{R}^n$ be of class C^r. Write f in the form $f(x, y)$, for $x \in \mathbb{R}^k$ and $y \in \mathbb{R}^n$. Suppose that (a, b) is a point of A such that $f(a, b) = 0$ and

$$\det \frac{\partial f}{\partial y}(a, b) \neq 0.$$

Then there is a neighbourhood B of a in $\mathbb{R}^k$ and a unique continuous function $g : B \to \mathbb{R}^n$ such that $g(a) = b$ and

$$f(x, g(x)) = 0, \ \ \forall x \in B.$$

The function g is in fact of class C^r.

Proof. We construct a function F to which we can apply the inverse function theorem. Define $F : A \to \mathbb{R}^{k+n}$ by the equation

$$F(x, y) = (x, f(x, y)).$$

Then F maps the open set A of $\mathbb{R}^{k+n}$ into $\mathbb{R}^k \times \mathbb{R}^n = \mathbb{R}^{k+n}$. Furthermore,

$$DF = \begin{bmatrix} I_k & 0 \\ \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}.$$

Computing $\det DF$, we have

$$\det DF = \det \frac{\partial f}{\partial y}.$$

Thus DF is non-singular at the point (a, b). Now $F(a, b) = (a, 0)$. Applying the inverse function theorem to the map F, we conclude that there exists an open set $U \times V$ of $\mathbb{R}^{k+n}$ about (a, b) (where U is open in $\mathbb{R}^k$ and V is open in $\mathbb{R}^n$) such that

1. F maps $U \times V$ in a one-to-one fashion onto an open set W in $\mathbb{R}^{k+n}$ containing $(a, 0)$.
2. The inverse function $G : W \to U \times V$ is of class C^r.

Note that since $F(x, y) = (x, f(x, y))$, we have

$$(x, y) = G(x, f(x, y)).$$

Thus G preserves the first k coordinates, as F does . Then we can write G in the form

$$G(x,z) = (x, h(x,z)), \text{ for } x \in \mathbb{R}^k \text{ and } x \in \mathbb{R}^n.$$

Here h is a function of class C^r mapping W into $\mathbb{R}^n$.

Let B be a connected neighbourhood of a in $\mathbb{R}^k$, chosen small enough that $B \times 0$ is contained in W. We prove existence of the function $g : B \to \mathbb{R}^n$. If $x \in B$, then $(x, 0) \in W$, so that we have

$$\begin{aligned} G(x,0) &= (x, h(x,0)), \\ (x,0) &= F(x, h(x,0)) = (x, f(x, h(x,0))), \\ 0 &= f(x, h(x,0)). \end{aligned}$$

We set $g(x) = h(x,0)$, for $x \in B$; then g satisfies the equation $f(x, g(x)) = 0$, as desired. Further

$$(a,b) = G(a,0) = (a, h(a,0));$$

then $b = g(a)$, as desired.

Now we prove the uniqueness of g. Let $g_0 : B \to \mathbb{R}^n$ be a continuous function satisfying the conditions in the conclusion of our theorem. Then in particular, g_0 agrees with g at the point a. We show that if g_0 agrees with g at the point $a_0 \in B$, then g_0 agrees with g in a neighbourhood B_0 of a_0. This is easy. The map g carries a_0 into V. Since g_0 is continuous, there is a neighbourhood B_0 of a_0 contained in B such that g_0 also maps B_0 into V. The fact that $f(x, g_0(x)) = 0$ for $x \in B_0$ implies that

$$\begin{aligned} F(x, g_0(x)) &= (x, 0), \quad \text{so} \\ (x, g_0(x)) &= G(x,0) = (x, h(x,0)). \end{aligned}$$

Thus, g_0 and g agrees on B_0. It follows that g_0 and g agrees on all of B: The set of points of B for which $|g(x) - g_0(x)| = 0$ is open in B and so is the set of points of B for which $|g(x) - g_0(x)| > 0$ (by continuity of g and g_0). Since B is connected, the latter set must be empty. □

In our proof of the implicit function theorem, there was of course nothing special about solving for the last n coordinates ; that choice was made simply for convenience. The same argument applies to the problem of solving for any n coordinates in terms of the others.

For example, suppose A is open in $\mathbb{R}^5$ and $f : A \to \mathbb{R}^2$ is a function of class C^r. Suppose one wishes to "solve" the equation $f(x,y,z,u,v) = 0$ for the two unknowns y and u in terms of the other three. In this case, the implicit function theorem tells us that if a is a point of A such that $f(a) = 0$ and

$$\det \frac{\partial f}{\partial (y,u)}(a) \neq 0,$$

then one can solve for y and u locally near that point, say $y = \phi(x,z,v)$ and $u = \psi(x,z,v)$. Furthermore, the derivatives of ϕ and ψ satisfy the formula

$$\frac{\partial(\phi,\psi)}{\partial(x,z,v)} = -\left[\frac{\partial f}{\partial(y,u)}\right]^{-1} \cdot \left[\frac{\partial f}{\partial(x,z,v)}\right].$$

Example 10.10. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by the equation $f(x,y) = x^2 + y^2 - 5$. Then the point $(x,y) = (1,2)$ satisfies the equation $f(x,y) = 0$. Both $\partial f/\partial x$ and $\partial f/\partial y$ are non-zero at $(1,2)$, so we can solve this equation locally for either variable in terms of the other. In particular, we can solve for y in terms of x, obtaining the function

$$y = g(x) = [5 - x^2]^{1/2}.$$

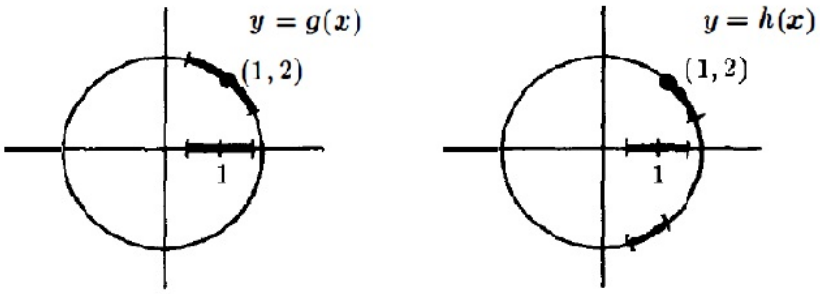


Figure 3: Example 10.10

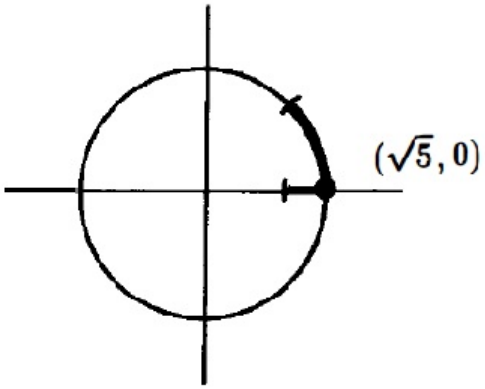


Figure 4: Example 10.11

Note that this solution is not unique in a neighbourhood of $x = 1$ unless we specify that g is continuous. For instance, the function

$$\begin{aligned} h(x) &= [5 - x^2]^{1/2}, \quad \text{for } x \geq 1, \\ &= -[5 - x^2]^{1/2}, \quad \text{for } x < 1. \end{aligned}$$

satisfies the same conditions, but is not continuous.

Example 10.11. The point $(x, y) = (\sqrt{5}, 0)$ also satisfies the equation $f(x, y) = 0$ for the function in example 10.10. The derivative $\partial f/\partial y$ vanishes at $(\sqrt{5}, 0)$, so we do not expect to be able to solve for y in terms of x near this point. And, in fact, there is no neighbourhood B of $\sqrt{5}$ on which we can solve for y in terms of x.

Example 10.12. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by the equation $f(x, y) = x^2 - y^3$. Then $(0, 0)$ is a solution of the equation $f(x, y) = 0$. Because $\partial f/\partial y$ vanishes at $(0, 0)$, we do not expect to be able to solve this equation for y in terms of x near $(0, 0)$. But in fact, we can; and furthermore, the solution is unique! However, the function we obtain is not differentiable at $x = 0$.

Example 10.13. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by the equation $f(x, y) = y^2 - x^4$. Then $(0, 0)$ is a solution of the equation $f(x, y) = 0$. Because $\partial f/\partial y$ vanishes at $(0, 0)$, we do not expect to be able to solve for y in terms of x near $(0, 0)$. In fact , however, we can do so, and we can do so in such a way that the resulting function is differentiable. However, the solution is not unique. Now the point $(1, 2)$ also satisfies the equation

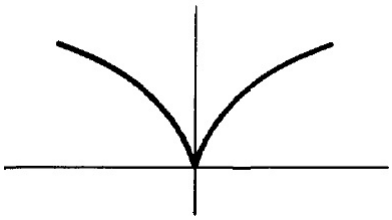


Figure 5: Example 10.12

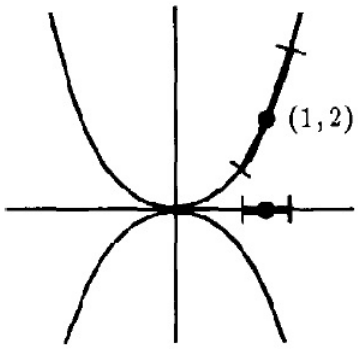


Figure 6: Example 10.13

$f(x, y) = 0$. Because $\partial f/\partial y$ is non-zero at $(1, 2)$, one can solve this equation for y as a continuous function of x in a neighbourhood of $x = 1$. One can in fact express y as a continuous function of x on a larger neighbourhood than the one pictured, but if the neighbourhood is large enough that it contains 0, then the solution is not unique on that larger neighbourhood.

10.4 Few Probable Questions

1. State and prove the inverse function theorem.
2. State and prove the implicit function theorem.
3. Let $f : \mathbb{R}^3 \to \mathbb{R}^2$ be of class C^1; write f in the form $f(x, y_1, y_2)$. Assume that $f(3, -1, 2) = 0$ and

$$Df(3, -1, 2) = \begin{bmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \end{bmatrix}.$$

   (a) Show that there exists a function $g : B \to \mathbb{R}^2$ of class C^1 defined on an open set B in $\mathbb{R}$ such that

$$f(x, g_1(x), g_2(x)) = 0, \text{ for } x \in B, \text{ and } g(3) = (-1, 2).$$

   (b) Find $Dg(3)$.

   (c) Discuss the problem of solving the equation $f(x, y_1, y_2) = 0$ for an arbitrary pair of the unknowns in terms of the third, near the point $(3, -1, 2)$.

4. Let $f : \mathbb{R}^5 \to \mathbb{R}^2$ be of class C^1. Let $a = (1, 2, -1, 3, 0)$. Suppose that $f(a) = 0$ and

$$Df(a) = \begin{bmatrix} 1 & 3 & 1 & -1 & 2 \\ 0 & 0 & 1 & 2 & -4 \end{bmatrix}.$$

(a) Show that there exists a function $g : B \to \mathbb{R}^2$ of class C^1 defined on an open set B in $\mathbb{R}^3$ such that

$$f(x_1, g_1(x), g_2(x), x_2, x_3) = 0, \text{ for } x = (x_1, x_2, x_3) \in B, \text{ and } g(1, 3, 0) = (2, -1).$$

(b) Find $Dg(1, 3, 0)$.

(c) Discuss the problem of solving the equation $f(x) = 0$ for an arbitrary pair of the unknowns in terms of the others, near the point a.

Unit 12

Structure

7.1 INTRODUCTION

In this unit, we discuss maxima and minima for real-valued functions of vector variables. You are already familiar with this concept from your undergraduate Real Analysis and Calculus courses. There you have seen that these concepts are studied locally also and they are called local maxima or local minima; together they are called local extrema. The extension of these concepts to the vector variable case helps to solve many real-life problems arising in economics, finance and other fields.

In Sec. 7.2, we shall discuss a necessary condition for a function to have local extrema in terms its partial derivatives. Then we shall discuss a sufficient condition for the existence of local extrema by using Taylor's theorem for real valued functions of vector variables.

One of the main applications of the concept of maxima and minima is to solve optimization problems arising in economics such as expenditure minimization problem, profit maximization problem, utility maximization problem. Most of these problems are concerned with maximizing and minimizing real-valued n-variable function called objective function and there are some constraints also attached with the problem which are again represented as a functional relationship. Such problems can be solved by a method called Lagrange Multiplier method. In Sec. 7.3, we discuss this method. We shall briefly explain the utility of this method by giving a practical problem in optimization. The understanding of the method requires some techniques in linear algebra such as quadratic forms and the related matrix theory. You are advised to look into any standard book on Linear algebra that are available at your programme study centre or the Block 4 of IGNOU course on Linear Algebra with the Code MTE-02 titled Inner Products and Quadrics which is also available at your programme centre.

Objectives

After studying this unit, you should be able to

- define critical points, stationary points, saddle points, local maxima and local minima;
- state a necessary condition for functions to have local extrema and apply it;
- state and prove the theorem known as "second derivative test" which gives a sufficient condition for finding local maxima and minima;

- use Hessian for classifying local maxima and local minima; and
- apply Lagrange's multiplier method for finding the stationary points when the variables are subject to some constraints.

7.2 LOCAL MAXIMA AND LOCAL MINIMA

This section deals with the concept of maxima and minima (or extrema) for real-valued functions of vector variables. You are already familiar with this concept for the one-variable case. In the case of one-variable you might have studied that there are functions which do not have an extrema at a point with repsect to the whole domain whereas the functions have extrema at that point locally. These points are called local extrema. In this section we shall take up the study of local extrema for functions from $\mathbf{R}^n$ to $\mathbf{R}$.

Definition 1: Let $f : E \subseteq \mathbf{R}^n \to \mathbf{R}$ be a function. A point $\mathbf{a} \in E$ is called a maximum point w.r.t. E, if $f(\mathbf{x}) \leq f(\mathbf{a})$ $\forall \mathbf{x} \in E$. A point $\mathbf{a} \in E$ is called a minimum point w.r.t. E if $f(\mathbf{x}) \geq f(\mathbf{a})$ $\forall x \in E$.

If a point $\mathbf{a} \in E$ is either maximum or minimum point w.r.t. E, then that point is called an extreme point or point of extrema.

Now we define local extrema.

Definition 2: Let $f : E \subseteq \mathbf{R}^n \to \mathbf{R}$ be a function where E is an open subset of $\mathbf{R}^n$. A point $\mathbf{a} \in E$ is said to be a local maximum for f if there exists a neighbourhood $E_\mathbf{a}$ of $\mathbf{a}$ such that $f(\mathbf{x}) \leq f(\mathbf{a})$ for all $\mathbf{x} \in E_\mathbf{a}$.

A local minima is similarly defined.

Example 1: Let us consider the function given by

$$f(x, y) = (x + 1)^2 + (y - 3)^2 - 1.$$

We first note that $f(-1, 3) = -1$.

Also $f(x, y) \geq f(-1, 3)$ for all $(x, y) \in \mathbf{R}^2$.

This show that the function has a minimum at $(-1, 3)$ and the minimum value is $f(-1, 3) = -1$. (See Fig. 1)

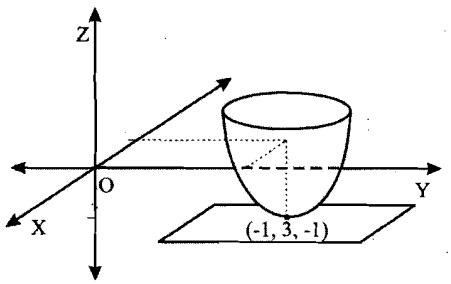


Fig. 1

* * *

Example 2: Let us consider the function

$$f(x, y) = \frac{1}{2} - \sin(x^2 + y^2)$$

Here $f(0, 0) = \frac{1}{2}$. Let us consider the neighbourhood U of (0, 0) given by

$$U = \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 < \frac{\pi}{6}\}$$

Then for any $(x, y) \in U$, we have

$$\sin(x^2 + y^2) > 0$$

and therefore

$$f(x, y) = \frac{1}{2} - \sin(x^2 + y^2) < \frac{1}{2} = f(0, 0).$$

Thus $f(x, y) \leq f(0, 0)$ for all $(x, y) \in U$ in the disc. Note that $f(x, y)$ can be greater than $\frac{1}{2}$ for $(x, y) \notin U$. Hence f has a local minimum at (0, 0).

* * *

Example 3: Let us consider another function given by

$$f(x, y) = 1 + \sqrt{x^2 + y^2}$$

If we look at the graph of the function given below, then we can see that f has a minimum at (0, 0).

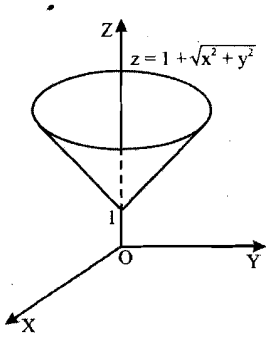


Fig. 2

* * *

If you closely look at the above examples, then you can notice that in the case of Examples 1 and 2 we have

$$\frac{\partial f}{\partial x}(x_0, y_0) = 0 = \frac{\partial f}{\partial y}(x_0, y_0)$$

where (x_0, y_0) is a point of extrema. Whereas we notice that in the case of Example 3, this is not the case as function does not have any first order partial derivatives at (0, 0).

Now we state a result which shows that if all the first order partial derivatives of $f : E \subset \mathbf{R}^n \to \mathbf{R}$ exists at a point $a \in E$ where E is an open set, then they necessarily vanish at the points of extrema.

Theorem 1: Let $f : E \subseteq \mathbf{R}^n \to \mathbf{R}$ be a function where E is an open subset of $\mathbf{R}^n$. Suppose that all the first order partial derivatives of the function f exists at a point $a \in E$. Then a necessary condition for the function to have a local extremum at the point a is that $\frac{\partial f}{\partial x_i}(a) = 0$ for $i = 1, \ldots, n$.

Proof: Suppose that f has a local extrema at the point $\mathbf{a} = (a_1, a_2, \ldots, a_n)$.

Let us consider the real-valued function ϕ defined by

$$\phi(t) = f(t, a_2, \ldots, a_n).$$

Since a an extreme point of f, we get that a_1 is extreme point for ϕ. Then from the one-variable case you know that

$$\phi'(a_1) = \frac{\partial f}{\partial x_1}(a_1, a_2, \ldots, a_n) = 0$$

In this way we can show that $\frac{\partial f}{\partial x_i}(a_1, \ldots, a_n) = 0$ for each $i = 1, \ldots, n$.

Hence the result. □

Now we make the following definition:

Definition 3: Let $f : E \subset \mathbf{R}^n \to \mathbf{R}$ be a function. A point $\mathbf{a} \in E$ is called a critical point of f if either

i) the partial derivatives of f do not exist at a, or

ii) $\frac{\partial f}{\partial x_i}(\mathbf{a}) = 0$ for $i = 1, \ldots, n$.

The points for which the condition (ii) is satisfied are called **stationary points**.

You may recall that all stationary points of a function need not be its point of local extrema. Such points are called saddle points. Note that a point $\mathbf{a} \in E$ is called a **saddle point** if every neighbourhood E_a of a contains points $\mathbf{x} \in E$ such that $f(x) > \mathbf{a}$ and other points $y \in E_a$ such that $y < f(\mathbf{a})$.

Let us consider an example.

Example 4: Let us consider the function $f : \mathbf{R}^2 \to \mathbf{R}$ given by

$$f(x, y) = (y - x^2)(y - 2x^2).$$

Here we have $\frac{\partial f}{\partial x}(0, 0) = \frac{\partial f}{\partial y}(0, 0) = 0$. Thus, (0, 0) is a stationary point. Now the graph of the function f given below shows that (0, 0) is not a point of local extrema. Note that the function f assumes both positive and negative values in every neighbourhood of (0, 0). Therefore (0, 0) is a saddle point for the function f.

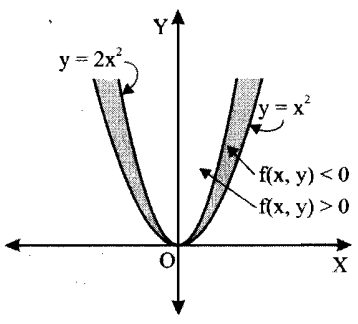


Fig. 3

* * *

Next we discuss a sufficient condition in terms of second order partial derivatives to check whether a point is an local extremum point.

Theorem 2: (Second-derivative test for extrema): Let $f : E \to \mathbf{R}$ be a function define on an open set $E \subset \mathbf{R}^n$. Assume that the second-order partial derivatives $D_{i,j}f$ exist in an n-ball $B(\mathbf{a})$ and are continuous at $\mathbf{a} \in \mathbf{R}^n$, where $\mathbf{a}$ is a stationary point of f. Let

$$Q(x) = \frac{1}{2}f''(a;x) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{i,j}f(\mathbf{a})x_i x_j \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$. Then

a) If $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq 0$, f has a relative minimum at $\mathbf{a}$.

b) If $Q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq 0$, f has a relative maximum at $\mathbf{a}$.

c) If $Q(\mathbf{x})$ takes both positive and negative values, then f has a saddle point at $\mathbf{a}$.

Proof: We first apply Taylor's theorem to the function f. Taking $m = 2$ and $b = a + x$ in Taylor's theorem (see Sec. 5.4, Unit 5), we get that there exists a z which lies on the line segment joining $\mathbf{a}$ and $a + x$ such that

$$f(\mathbf{a}+\mathbf{x}) - f(\mathbf{a}) = f'(\mathbf{a})\mathbf{x} + \frac{1}{2}f''(\mathbf{z},\mathbf{x}) \tag{2}$$

and

$$f''(\mathbf{z},\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij}f(\mathbf{z})x_i x_i$$

Since $\mathbf{a}$ is a stationary point, we have $f'(a) = 0$. Therefore Equation (2) becomes

$$f(a+x) - f(a) = \frac{1}{2}f''(z,x)$$

Therefore as $a + x$ ranges over $B(\mathbf{a})$, the algebraic sign of $f(\mathbf{a}+\mathbf{x}) - f(\mathbf{a})$ is determined by that of $f''(\mathbf{z};\mathbf{x})$. We can write Equation (2) in the form

$$f(\mathbf{a}+\mathbf{x}) - f(\mathbf{a}) = \frac{1}{2}f''(\mathbf{a},\mathbf{x}) + \|\mathbf{x}\|^2 E(\mathbf{x}) \tag{3}$$

where

$$\|\mathbf{x}\|^2 E(\mathbf{x}) = \frac{1}{2}f''(\mathbf{z},\mathbf{x}) - \frac{1}{2}f''(\mathbf{a};\mathbf{x})$$

Substituting for $f''(\mathbf{z},\mathbf{x})$ and $f''(\mathbf{a};\mathbf{x})$, we get that

$$\|\mathbf{x}\|^2|E(\mathbf{x})| \leq \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|D_{ij}f(\mathbf{z}) - D_{ij}f(\mathbf{a})|\|\mathbf{x}\|^2 \tag{4}$$

Since the second order partial derivatives of f are continuous at $\mathbf{a}$ we get that $E(\mathbf{x}) \to 0$ as $\mathbf{x} \to 0$.

Now we rewrite Equation (3) in the form

$$f(\mathbf{a}+\mathbf{x}) - f(\mathbf{a}) = Q(\mathbf{x}) + \|\mathbf{x}\|^2 E(\mathbf{x}) \tag{5}$$

where Q(x) is as given in Equation(1).

The function Q is continuous at each point x in $\mathbf{R}^n$. Let $S = \{x : \|x\| = 1\}$ denotes the boundary of the n-ball $B(0; 1)$. (Recall that we defined the norm function $\| \ \|$ in Unit 5). If $Q(x) > 0$ for all $x \neq 0$, then Q(t) is positive on S. Since S is compact, Q has a minimum on S. Let us call it m. Then $m > 0$.

Now $Q(cx) = c^2Q(x)$ for every real number c. Taking $c = 1/\|x\|$ where $x \neq 0$ we see that $cx \in S$ and hence $c^2Q(x) \geq m$, so $Q(x) \geq m\|x\|^2$. Using this in (A_0) we find

$$f(a + x) - f(a) = Q(x) + \|x\|^2E(x) \geq m\|x\|^2 + \|x\|^2E(x)$$

Since $E(x) \to 0$ as $x \to 0$, there is a positive number r such that $|E(x)| < \frac{1}{2}m$ whenever $0 < \|x\| < r$. For such x we have $0 \leq \|x\|^2|E(x)| < \frac{1}{2}m\|x\|^2$, so

$$f(a + x) - f(a) > m\|x\|^2 - \frac{1}{2}m\|x\|^2 = \frac{1}{2}m\|x\|^2 > 0.$$

Therefore f has a relative minimum at a, which proves (a).

To prove (b) we use a similar argument, or apply part (a) to the function $-f$.

Finally, we prove (c). For each $\lambda > 0$ we have, from (A_0).

$$f(a + \lambda x) - f(a) = Q(\lambda x) + \lambda^2\|x\|^2E(\lambda x) = \lambda^2\{Q(x) + \|x\|^2E(\lambda x)\}$$

Suppose $Q(x) \neq 0$ for some x. Since $E(y) \to 0$ as $y \to 0$, there is a positive r such that

$$\|x\|^2E(\lambda x) < \frac{1}{2}|Q(x)| \quad \text{if } 0 < \lambda < r.$$

Therefore, for each such λ the quantity $\lambda^2\{Q(x) + \|x\|^2E(\lambda x)\}$ has the same sign as Q(x). Therefore, if $0 < \lambda < r$, the difference $f(a + \lambda x) - f(a)$ has the same sign as Q(x). Hence, if Q(x) takes both positive and negative values, it follows that f has a saddle point at a. □

Note: A real-valued function Q defined on $\mathbf{R}^n$ by an equation of the type

$$Q(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j,$$

where $x = (x_1, \ldots, x_n)$ and the a_{ij} are real is called a **quadratic form**. The form is called symmetric if $a_{ij} = a_{ji}$ for all i and j and is called positive definite if $x \neq 0$ implies $Q(x) > 0$, and negative definite if $x \neq 0$ implies $Q(x) < 0$.

You might be already familiar with quadratic forms from your undergraduate Linear algebra course. (You can refer to IGNOU course MTE-02, Block 4)

In general, it is not easy to determine whether a quadratic form is positive or negative definite. One criterion, involving determinants, can be described as follows. Let Δ = determinant of the matrix $[a_{ij}]$ and let Δ_k denote the determinant of the $k \times k$ matrix obtained by deleting the last $(n - k)$ rows and columns of the matrix $[a_{ij}]$.

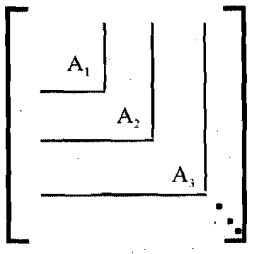


Fig.4

Also, put $\Delta_0 = 1$. From the theory of quadratic forms it is known that a necessary and sufficient condition for a symmetric form to be positive definite is that the $n+1$ numbers $\Delta_0, \Delta_1, \ldots, \Delta_n$ be positive. The form is negative definite if and only if, the same $n+1$ numbers are alternately positive and negative. The quadratic form which appears in Equation (1) is symmetric because the mixed partials $D_{i,j}f(a)$ and $D_{j,i}f(a)$ are equal. Therefore, under the conditions of Theorem 2, we see that f has a local minimum at a if the $(n+1)$ numbers $\Delta_0, \Delta_1, \ldots, \Delta_n$ of the corresponding Jacobian matrix for f are all positive, and a local maximum if these numbers are alternately positive and negative.

We have the following result:

Theorem 3: If $f : E \subset \mathbf{R}^n \to \mathbf{R}$, E open in $\mathbf{R}^n$, has continuous first and second-order partial derivatives at a where a is a critical point of f, and Hf is the Hessian of f at a (refer Unit 5 where we have defined the Hessian of a function) and Δ_k denote the determinant of $k \times k$ matrix obtained by deleting the last $(n-k)$ rows and column of the matrix. Then the following hold:

a) if $\Delta_{2k} < 0$ for some k then a is a saddle point of f,

b) if $\Delta_n \neq 0$ then

(b1) f has a local minimum at a if and only if $\Delta_k > 0$ for all k,
(b2) f has a local maximum at a if and only if $(-1)^k\Delta_k > 0$ for all k,

c) if $\Delta_n = 0$ we call it a degenerate case and the test cannot be applied.

The case $n = 2$ can be handled directly and gives the following criterion.

Theorem 4: Let f be a real-valued function with continuous second-order partial derivatives at a stationary point a in $\mathbf{R}^2$. Let

$$A = D_{1,1}f(a), \quad B = D_{1,2}f(a), \quad C = D_{2,2}f(a).$$

and let

$$\Delta = \det \begin{bmatrix} A & B \\ B & C \end{bmatrix} = AC - B^2.$$

Then we have:

a) If $\Delta > 0$ and $A > 0$, f has a local minimum at a.

b) If $\Delta > 0$ and $A < 0$, f has a local maximum at a.

c) If $\Delta < 0$, f has a saddle point at a.

Note: You may recall that the conditions given in the theorem above resembles that of the one-variable case.

If $\Delta = 0$, then the test fails. Let us consider some examples.

Example 5: Let us check the following function $f : \mathbf{R}^3 \to \mathbf{R}$ for local extrema.

Let $f(x, y, z) = x^2y^2 + z^2 + 2x - 4y + z$.

We first note that

$$\nabla f(x, y, z) = (2xy^2 + 2, 2x^2y - 4, 2z + 1).$$

If a is a critical point of f, then a should satisfy the

$$\begin{aligned} 2xy^2 + 2 &= 0 \\ 2x^2y - 4 &= 0 \\ 2z + 1 &= 0 \end{aligned}$$

We solve these equations. From the third equation we get that $z = -1/2$. From the first two equations we see that x and y are non-zero. Hence $xy^2 = -1$ and $x^2y = 2$ imply $xy^2/x^2y = -1/2 = y/x$ and $x = -2y$. We have $-2y \cdot y^2 = -1$, i.e. $y^3 = 1/2$ and $y = 2^{-1/3}$. From $x = -2y$ we obtain $x = -2^{2/3}$ and conclude that $(-2^{2/3}, 2^{-1/3}, -1/2)$ is the only critical point of f.

A simple calculation shows

$$J_f = \begin{bmatrix} 2y^2 & 4xy & 0 \\ 4xy & 2x^2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

and

$$J_f(-2^{2/3}, 2^{-1/3}, 1/2) = \begin{bmatrix} 2^{1/3} & -4.2^{1/3} & 0 \\ -4.2^{1/3} & 2.2^{4/3} & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Note $\Delta_1 = 2^{1/3} > 0$ and

$$\begin{aligned} \Delta_2 = \begin{vmatrix} 2^{1/3} & -4.2^{1/3} \\ -4.2^{1/3} & 2.2^{4/3} \end{vmatrix} &= 2.2^{5/3} - 16.2^{2/3} \\ &= 4.2^{2/3} - 16.2^{2/3} < 0 \end{aligned}$$

Therefore by Theorem 3, the critical point $(-2^{2/3}, 2^{-1/3}, -1/2)$ is a saddle point of f.

* * *

Your can try some exercises now.

E1) Find the critical points of $f(x, y, z) = x^2y + y^2z + z^2 - 2x$ and check whether they are extreme points.

In the next section we shall discuss the problem of finding maxima or minima subject to certain constraints.

7.3 LAGRANGE MULTIPLIER

Here we start with a practical situation.

Suppose a person consumes n commodities in nonnegative quantities. Then her utility from consuming $x_i \geq 0$ units of commodity i ($i = 1, \ldots, n$) is given by $u(x_1, \ldots, x_n)$, where $u : \mathbf{R}^n_+ \to \mathbf{R}$. Suppose she has an income of $I \geq 0$, and faces the price vector $p = (p_1, \ldots, p_n)$, where $p_i \geq 0$ denotes the unit price of the i-th commodity. Her budget set (i.e., the set of feasible or affordable consumption bundles, given her income I and the prices p) is denoted $B(p, I)$, and is given by

$$B(p, I) = \{x \in \mathbf{R}^n_+ | \sum_{i=1}^{n} p_i x_i \leq I\}$$

Then her objective is to maximize the level of her utility over the set of affordable commodity bundles, i.e., to solve:

Maximize $u(x)$ subject to $x \in B(p, I)$.

There are many situations like this where the values of a given function $f : \mathbf{R}^n \to \mathbf{R}$ are to be maximized or minimized over a given set $E \subseteq \mathbf{R}^n$. Here we shall discuss a method for solving such problems that is developed by the mathematician Lagrange.

Let us consider another problem.

Suppose that $f(x, y, z)$ represents the temperature at the point (x, y, z) in space and we ask for the maximum or minimum value of the temperature on a certain surface. If the equation of the surface is given explicitly in the form $z = h(x, y)$, then in the expression for $f(x, y, z)$ we can replace z by $h(x, y)$ to obtain the temperature on the surface as a function of x and y alone, say $F(x, y) = f[x, y, h(x, y)]$. The problem is then reduced to finding the extreme value of F. However, in practice, certain difficulties arise. The equation of the surface might be given in an implicit form, say $g(x, y, z) = 0$ and it may be impossible, in practice, to solve this equation explicitly for z in terms of x and y, or even for x or y in terms of the remaining variables. The problem might be further complicated by asking for the extreme values of the temperature at those points which lie on a given curve in space. Such a curve can be the intersection of two surfaces, say $g_1(x, y, z) = 0$ and $g_2(x, y, z) = 0$. If we could solve these two equations simultaneously, say for x and y in terms of z, then we could introduce these expressions into f and obtain a new function of z alone, whose extrema we would then seek. In general, however, this procedure cannot be carried out and a more practicable method need to be sought. An elegant and useful method for attacking such problems was developed by Lagrange. The validity of the method is established by the Implicit Function Theorem which we described in Unit 6.

Lagrange's method provides a necessary condition for a point to be an extreme point which we shall explain now.

Let $f : E \subset \mathbf{R}^n \to \mathbf{R}, E \subseteq \mathbf{R}^n$ an open set, be a function whose extreme values are sought when the variables are restricted by a certain number of side conditions, say $g_1(x_1, \ldots, x_n) = 0, \ldots, g_m(x_1, \ldots, x_n) = 0$.

We first form the linear combination

$$L(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) - \lambda_1 g_1(x_1, \ldots, x_n) - \ldots - \lambda_m g_m(x_1, \ldots, x_n), \quad (6)$$

where $\lambda_1, \ldots, \lambda_m$ are m constants. We then differentiate ϕ with respect to each coordinate and consider the following system of n + m equations:

$$\frac{\partial L}{\partial x_i} = 0, \quad i = 1, 2, \ldots, n, \quad (7)$$

$$g_k(x_1, \ldots, x_n) = 0, \quad k = 1, 2, \ldots, m. \quad (8)$$

Lagrange, by his method, proved that if the point $(x_1, x_2, \ldots, x_n)$ is a point of extrema for f, then it will also satisfy this system of $(n + m)$ equations. In practice we solve the "n + m" unknowns $\lambda_1, \lambda_2, \ldots, \lambda_m$ and $x_1, x_2, \ldots, x_n$. The point $(x_1, x_2, \ldots, x_n)$ so obtained is a stationary point. According to the Lagrange's theorem this point can then be tested for maximum or minimum point by the already known methods.

The numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$ are introduced only to help to solve the system for $x_1, x_2, \ldots, x_n$, and they are called **Lagrange's multipliers**. One multiplier is introduced for each side condition. The function L in Equation (6) is called the Lagrangian function. Equations (7) and (8) are called Lagrangian Equations.

Now we state the Lagrange's theorem, the proof of which involves implicit function theorem. We omit the proof here.

Theorem 5: Let $f : E \subset \mathbf{R}^n \to \mathbf{R}$, E an open set in $\mathbf{R}^n$, be such that the partial derivatives of f exists and are continuous on E. Let $g_1, \ldots, g_m$ be m real-valued functions defined on E such that partial derivatives of g_i exists and are continuous on E for $i = 1, \ldots, m$. Let us assume that $m < n$. Let X_0 be that subset of E on which each g_i vanishes for $i = 1, \ldots, m$, that is,

$$X_0 = \{x \in E, g_i(x) = 0 \text{ for } i = 1, \ldots, m\}.$$

Assume that $x_0 \in X_0$ and assume that there exists a ball $B(x_0)$ in $\mathbf{R}^n$ such that $f(x) \leq f(x_0)$ for all x in $X_0 \cap B(x_0)$ or such that $f(x) \geq f(x_0)$ for all x in $X_0 \cap B(x_0)$. Assume also that the m-rowed determinant $\det[D_j g_i(x_0)] \neq 0$. Then there exist m real numbers $\lambda_1, \ldots, \lambda_m$ such that they satisfy following n equations:

$$\frac{\partial f}{\partial x_i}(x_0) - \sum_{k=1}^{m} \lambda_k \frac{\partial g_k}{\partial x_i}(x_0) = 0 \quad (i = 1, 2, \ldots, n). \quad (9)$$

So the solution of an extremum problem by Lagrange's method involves the following step:

Step 1: Form the Lagrangian function given in Equation (6)

Step 2: Form the Lagrangian equations given in Equations (7) and (8). The solution thus obtained is a stationary point.

Step 3: Check the stationary point for extrema by the methods already discussed in Sec. 7.2.

Here we state a sufficient condition for checking extrema when we have a single constraint. In this case the Equation 6 reduces to

$$L(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) - \lambda g(x_1, \ldots, x_n). \tag{10}$$

To check that the stationary point obtained by Lagrange method is local maximum or local minimum, we need to compute the value of $n-1$ principal minors of the following determinant

$$\Delta_{n+1} = \begin{vmatrix} 0 & \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} & \cdots & \frac{\partial g}{\partial x_n} \\ \frac{\partial g}{\partial x_1} & \frac{\partial^2 f}{\partial x_1^2} - \lambda\frac{\partial^2 g}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} - \lambda\frac{\partial^2 g}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} - \lambda\frac{\partial^2 g}{\partial x_1 \partial x_n} \\ \frac{\partial g}{\partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} - \lambda\frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} - \lambda\frac{\partial^2 g}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} - \lambda\frac{\partial^2 g}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\partial g}{\partial x_n} & \frac{\partial^2 f}{\partial x_n \partial x_1} - \lambda\frac{\partial^2 g}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} - \lambda\frac{\partial^g}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} - \lambda\frac{\partial^2 g}{\partial x_n^2} \end{vmatrix}$$

If the signs of minors $\Delta_3, \Delta_4, \Delta_5$ are alternatively positive and negative, then extreme point is a local maximum. But if sign of all minors $\Delta_3, \Delta_4, \Delta_5$ are negative, the extreme point is a local minimum.

Let us see an example.

Example 6: Suppose we want to find the extreme values of the function

$$Z = 2x_1^2 + x_2^2 + 3x_3^2 + 10x_1 + 8x_2 + 6x_3 - 100$$

subject to the constraint

$$x_1 + x_2 + x_3 = 20, \quad x_1, x_2, x_3 \geq 0$$

Solution: Here $n = 3$ and $m = 1$. Let $g(x_1, x_2, x_3) = x_1 + x_2 + x_3 - 20$. Lagrangian function can be formulated as:

$$L(x, \lambda) = 2x_1^2 + x_2^2 + 3x_3^2 + 10x_1 + 8x_2 + 6x_3 - 100 - \lambda(x_1 + x_2 + x_3 - 20)$$

To obtain the stationary points, we solve the following system of equations.

$$\frac{\partial L}{\partial x_1} = 4x_1 + 10 - \lambda = 0; \quad \frac{\partial L}{\partial x_2} = 2x_2 + 8 - \lambda = 0$$

$$\frac{\partial L}{\partial x_3} = 6x_3 + 6 - \lambda = 0; \quad g(x_1, x_2, x_3) = x_1 + x_2 + x_3 - 20 = 0$$

Putting the value of x_1, x_2 and x_3 in the last equation $g(x_1, x_2, x_3) = 0$, and solving for λ, we get $\lambda = 30$. Substituting the value of λ in the other three equations, we get the stationary point: $(x_1, x_2, x_3) = 5, 11, 4$.

To prove the sufficient condition whether the stationary point gives maximum or minimum value of the function we evaluate 2 principal minors as illustrated

in Sec. 7.2.

$$\Delta_3 = \begin{vmatrix} 0 & \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \\ \frac{\partial g}{\partial x_1} & \frac{\partial^2 f}{\partial x_1^2} - \lambda\frac{\partial^2 g}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} - \lambda\frac{\partial^2 g}{\partial x_1 \partial x_2} \\ \frac{\partial g}{\partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} - \lambda\frac{\partial^2 g}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} - \lambda\frac{\partial^2 g}{\partial x_2^2} \end{vmatrix}_{(5,11,4)} = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 4 & 0 \\ 1 & 0 & 2 \end{vmatrix} = -6$$

$$\Delta_4 = \begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 4 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 6 \end{vmatrix} = 48$$

Since sign of Δ_3 and Δ_4 are alternative, the stationary point: $(x_1, x_2, x_3) = (5, 11, 4)$ is a local maximum. At this point the value of the function is, $Z = 281$.

* * *

In the appendix we have given an illustrative example where we have explained how Lagrange Multiplier method can be used for modelling problem in economics.

You can try this exercise now.

E2) Find and clarify the extreme values of the following functions subject to the constraints given along side.

i) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^3$ subject to the constraint $4x_1 + x_2^2 + 2x_3 = 14, x_1, x_2, x_3 \geq 0$.

ii) $f(x_1, x_2) = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2$ subject to the constraint $x_1 + 2x_2 = 2, x_1, x_2 \geq 0$.

With this we come to an end of this unit and to this block.

7.4 SUMMARY

In this unit, we have covered the following points for real-valued functions of vector-variable:

1. We have defined
   i) critical points and stationary points
   Let $f : E \subset \mathbf{R}^n \to \mathbf{R}$, E an open set in $\mathbf{R}^n$, be a function. A point $a \in E$ is called a critical point of f if either
   i) the partial derivatives of f do not exist at a, or
   ii) $\frac{\partial f}{\partial x_i}(a) = 0$ for all i such that $1 \leq i \leq n$.
   The points for which the condition (ii) is satisfied are called **stationary points**.

ii) saddle point:
A point $a \in E$ is called a **saddle point** if every neighbourhood E_a of a contains points $x \in E$ such that $f(x) > a$ and other points $y \in E_a$ such that $y < f(a)$.

iii) local maxima and local minima
Let $f : E \subseteq \mathbf{R}^n \to \mathbf{R}$ be a function where E is an open subset of $\mathbf{R}^n$. A point $a \in E$ is said to be a local maximum for f if there exists a neighbourhood E_a of a such that $f(x) \leq f(a)$ for all $x \in E$.
A local minima is similarly defined.

2. We have established that a necessary condition for a function to have local extrema is $\nabla f = 0$ provide ∇f exists.

3. We have derived a test called second derivative test for finding local extrema.
(Second-derivative test for extrema). Let $f : E \to \mathbf{R}$ be a function define on an open set $E \subset \mathbf{R}^n$. Assume that the second-order partial derivatives $D_{i,j}f$ exist in an n-ball $B(a)$ and are continuous at $a \in \mathbf{R}^n$, where a is a stationary point of f. Let

$$Q(x) = \frac{1}{2}f''(a;x) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} D_{i,j}f(a)x_i x_j$$

where $x = (x_1, \ldots, x_n)$. Then
a) If $Q(x) > 0$ for all $x \neq 0$, f has a relative minimum at a.
b) If $Q(x) < 0$ for all $x \neq 0$, f has a relative maximum at a.
c) If $Q(x)$ takes both positive and negative values, then f has a saddle point at a.

4. We have explained a method for classifying local maxima and local minima using the Hessian.

5. We have explained Lagrange's Multiplier Method.

7.5 HINTS/SOLUTIONS

E1) $\nabla f(x, y, z) = (2xy - 2, x^2 + 2yz, y^2 + 2z)$
and the critical points satisfy the equations

$$2xy - 2 = 0, x^2 + 2yz = 0 \text{ and } y^2 + 2z = 0.$$

Substituting $z = -y^2/2$ into the second equation implies $y^3 = x^2$. Hence, the first equation shows $y^{5/2} = 1$ and we have $y = 1$ and $z = -1/2$. From $xy = -1$ we get $x = 1$ and $(1, 1, -1/2)$ is the only critical point of f. We have

$$H_{f(x,y,z)} = \begin{pmatrix} 2y & 2x & 0 \\ 2x & 2z & 2y \\ 0 & 2y & 2 \end{pmatrix}$$

and

$$H_{f(1,1,-1/2} = \begin{pmatrix} 2 & 2 & 0 \\ 2 & -1 & 2 \\ 0 & 2 & 2 \end{pmatrix}$$

Since $\det(2) > 0$ and

$$\det \begin{pmatrix} 2 & 2 \\ 2 & -1 \end{pmatrix} = -2 - 4 < 0$$

the point $(1, 1, -1/2)$ is a saddle point of f.

E2) i) **Hint:** The extreme point is $\left(\frac{81}{100}, \frac{7}{20}, \frac{7}{25}\right)$. It is a minimum and the minimum value is $\frac{857}{100}$.

ii) **Hint:** The extreme point is $\left(\frac{1}{3}, \frac{5}{6}\right)$. It is a maximum point and the value is 4.166.

Unit 13

Course Structure

- Integration on $\mathbb{R}^n$: Integrals of $f : A \to \mathbb{R}$, where A is a closed rectangle in $\mathbb{R}^n$
- Conditions of integrability

---

11 Introduction

The multiple integral is a definite integral of a function of more than one real variable, for example, $f(x, y)$ or $f(x, y, z)$. Integrals of a function of two variables over a region in $\mathbb{R}^2$ are called double integrals, and integrals of a function of three variables over a region of $\mathbb{R}^3$ are called triple integrals.

Just as the definite integral of a positive function of one variable represents the area of the region between the graph of the function and the x-axis, the double integral of a positive function of two variables represents the volume of the region between the surface defined by the function (on the three-dimensional Cartesian plane where $z = f(x, y)$ and the plane which contains its domain. If there are more variables, a multiple integral will yield hypervolumes of multidimensional functions.

Objectives

After reading this unit, you will be able to

- define the partition of a rectangle in $\mathbb{R}^n$
- define the upper and lower sums of a bounded function defined on a closed rectangle and their relationships with respect to refinements
- define the integral of a bounded function defined on a closed rectangle, if it exists
- learn a necessary and sufficient condition for the existence of the integral of a bounded function over a closed rectangle
- apply the theorems in various problems

11.1 Integral Over a Closed Rectangle

We begin by defining the volume of a rectangle. Let

$$Q = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$$

be a rectangle in $\mathbb{R}^n$. Each of the intervals $[a_i, b_i]$ is called the component interval of Q. The maximum of the numbers $b_1 - a_1, \ldots, b_n - a_n$ is called the width of Q . Their product

$$v(Q) = (b_1 - a_1)(b_2 - a_2) \ldots (b_n - a_n)$$

is called the volume of Q.

In the case $n = 1$, the volume and the width of the (1-dimensional) rectangle $[a, b]$ are the same, namely, the number $b - a$. This number is also called the length of $[a, b]$.

Definition 11.1. Given a closed interval $[a, b]$ of $\mathbb{R}$, a partition of $[a, b]$ is a finite collection P of points of $[a, b]$ that includes the points a and b. We usually index the elements of P in increasing order, for notational convenience, as

$$a = t_0 < t_1 < \cdots < t_k = b;$$

each of the intervals $[t_{i-1}, t_i]$, for $i = 1, \ldots, k$, is called a subinterval determined by P, of the interval $[a, b]$. More generally, given a rectangle

$$Q = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$$

in $\mathbb{R}^n$, a partition P of Q is an n-tuple $(P_1, \ldots, P_n)$ such that P_j is a partition of $[a_j, b_j]$ for each j. If for each j, I_j is one of the subintervals determined by P_j of the interval $[a_j, b_j]$, then the rectangle

$$R = I_1 \times \cdots \times I_n$$

is called a subrectangle determined by P, of the rectangle Q. The maximum width of these subrectangles is called the mesh of P.

Definition 11.2. Let Q be a rectangle in $\mathbb{R}^n$ and let $f : Q \to \mathbb{R}$ be a bounded function. Let P be a partition of Q. For each subrectangle R determined by P, let

$$m_R(f) = \inf\{f(x) : \ x \in R\}, \quad M_R(f) = \sup\{f(x) : \ x \in R\}.$$

We define the lower sum and the upper sum, respectively, of f, determined by P, by the equations

$$\begin{aligned} L(f, P) &= \sum_R m_R(f).v(R), \\ U(f, P) &= \sum_R M_R(f).v(R). \end{aligned}$$

where the summations extend over all subrectangles R determined by P.

Let $P = (P_1, \ldots, P_n)$ be a partition of the rectangle Q. If P'' partition of Q obtained from P by adjoining additional points to some or all of the partitions $P_1, \ldots, P_n$, then P" is called a refinement of P. Given two partitions P and $P' = (P_1^{'}, \ldots, P_n^{'})$ of Q, the partition

$$P'' = (P_1 \cup P_1^{'}, \ldots, P_n \cup P_n^{'})$$

is a refinement of both P and P'; it is called their common refinement.

Passing from P to a refinement of P of course affects lower sums and upper sums; in fact, it tends to increase the lower sums and decrease the upper sums as we have seen in the case of one-dimensional upper and lower sums. That is the substance of the following lemma:

Lemma 11.3. Let P be a partition of the rectangle Q and let $f : Q \to \mathbb{R}$ be a bounded function. If P'' is a refinement of P, then

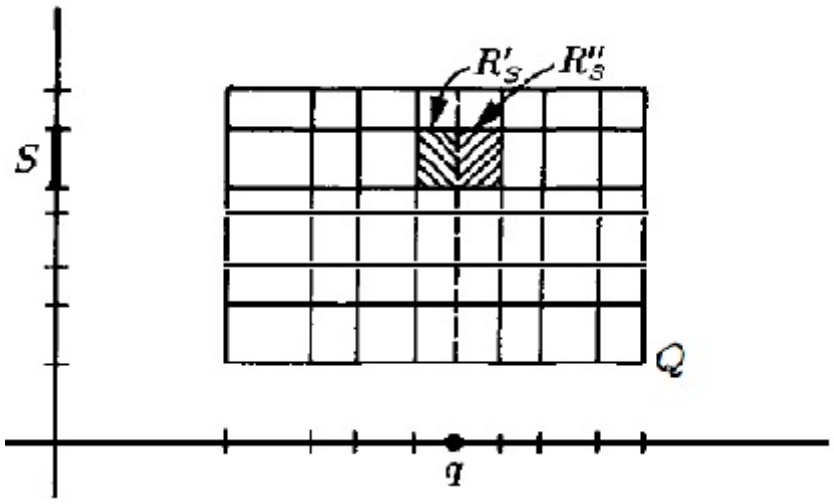
$$L(f, P) \leq L(f, P'') \quad \text{and} \quad U(f, P'') \leq U(f, P).$$

Proof. Let Q be the rectangle

$$Q = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$$

It suffices to prove the lemma when P'' is obtained by adjoining a single additional point to the partition of one of the component intervals of Q. Suppose, to be definite, that P is the partition $(P_1, \ldots, P_n)$ and that P'' is obtained by adjoining the point q to the partition P_1. Further, suppose that P_1 consists of the points

$$a_1 = t_0 < t_1 < \cdots < t_k = b_1$$



and that q lies interior to the subinterval $[t_{i-1}, t_i]$. We first compare the lower sums $L(f, P)$ and $L(f, P'')$. Most of the subrectangles determined by P are also subrectangles determined by P''. An exception occurs for a subrectangle determined by P of the form

$$R_S = [t_{i-1}, t_i] \times S$$

where S is one of the subrectangles of $[a_2, b_2] \times \cdots \times [a_n, b_n]$ determined by $(P_2, \ldots, P_n)$. The term involving the subrectangle R_S disappears from the lower sum and is replaced by the terms involving the two subrectangles

$$R'_S = [t_{i-1}, q] \times S \quad \text{and} \quad R''_S = [q, t_i] \times S,$$

which are determined by P''.

Now since $m_{R_S}(f) \leq f(x)$ for each $x \in R'_S$ and for each $x \in R''_S$, it follows that

$$m_{R_S}(f) \leq m_{R'_S}(f) \quad \text{and} \quad m_{R_S}(f) \leq m_{R''_S}(f).$$

Because $v(R_S) = v(R'_S) + v(R''_S)$ by direct computation , we have

$$m_{R_S}(f)v(R_S) \leq m_{R'_S}(f)v(R'_S) + m_{R''_S}(f)v(R''_S).$$

Since this inequality holds for each subrectangle of the form R_S, it follows that

$$L(f, P) \leq L(f, P'').$$

A similar argument applies to show that $U(f, P'') \leq U(f, P)$. □

Now we explore the relation between upper sums and lower sums. We have the following result:

Lemma 11.4. Let Q be a rectangle and $f : Q \to \mathbb{R}$ be a bounded function. If P and P' are any two partitions of Q, then

$$L(f, P) \leq U(f, P').$$

Proof. In the case where $P = P'$, the result is obvious: For any subrectangle R determined by P, we have $m_R(f) \leq M_R(f)$. Multiplying by $v(R)$ and summing gives the desired inequality.

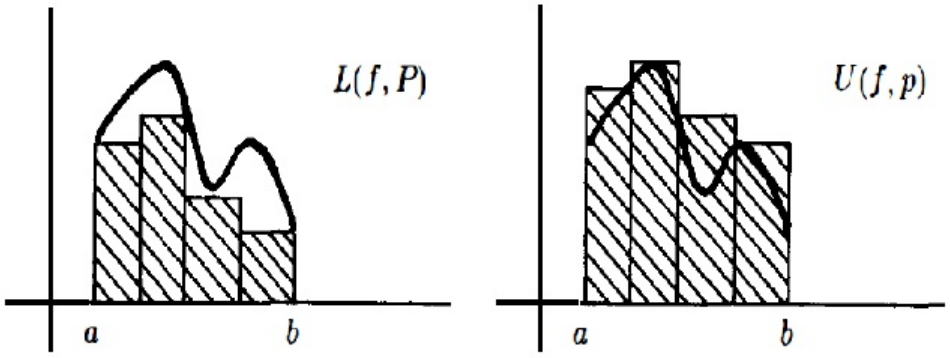


Figure 7: For one-dimensional f in example 11.6

In general, given partitions P and P' of Q , let P'' be their common refinement. Using the preceding lemma, we conclude that

$$L(f, P) \leq L(f, P'') \leq U(f, P'') \leq U(f, P').$$

□

We are now in a position to define the integral.

Definition 11.5. Let Q be a rectangle and $f : Q \to \mathbb{R}$ be a bounded function. As P ranges over all partitions of Q, define

$$\underline{\int_Q} f = \sup_P \{L(f, P)\} \quad \text{and} \quad \overline{\int_Q} f = \inf_P \{U(f, P)\}.$$

These numbers are called the lower integral and upper integral, respectively, of f over Q. They exist because the numbers $L(f, P)$ are bounded above by $U(f, P')$ where P' is any fixed partition of Q; and the numbers $U(f, P)$ are bounded below by $L(f, P')$. If the upper and lower integrals of f over Q are equal, we say that f is integrable over Q , and we define the integral of f over Q as the common value of the upper and lower integrals. We denote the integral of f over Q by either of the symbols

$$\int_Q f \quad \text{or} \quad \int_{x \in Q} f(x).$$

Example 11.6. Let $f : [a, b] \to \mathbb{R}$ be a non-negative bounded function. If P is a partition of $I = [a, b]$, then $L(f, P)$ equals the total area of a bunch of rectangles inscribed in the region between the graph of I and the x-axis, and $U(f, P)$ equals the total area of a bunch of rectangles circumscribed about this region as shown in the figure.

The lower integral represents the so-called ”inner area” of this region, computed by approximating the region by inscribed rectangles, while the upper integral represents the so-called ”outer area,” computed by approximating the region by circumscribed rectangles. If the ”inner” and ”outer” areas are equal, then f is integrable.

Similarly, if Q is a rectangle in $\mathbb{R}^2$ and $f : Q \to \mathbb{R}$ is non-negative and bounded, one can picture $L(f, P)$ as the total volume of a bunch of boxes inscribed in the region between the graph of f and the xy-plane, and $U(f, P)$ as the total volume of a bunch of boxes circumscribed about this region.

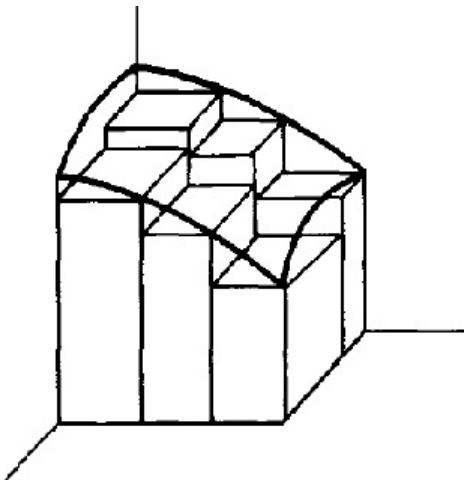


Figure 8: For two-dimensional f in example 11.6

Example 11.7. Let $I = [0, 1]$. Let $f : I \to \mathbb{R}$ be defined by setting

$$\begin{aligned} f(x) &= 0; \text{ if } x \text{ is rational} \\ &= 1; \text{ if } x \text{ is irrational.} \end{aligned}$$

We show that f is not integrable over I.

Let P be a partition of f. If R is any subinterval determined by P, then $m_R(f) = 0$ and $M_R(f) = 1$, since R contains both rational and irrational numbers. Then

$$L(f, P) = \sum_R 0.v(R) = 0, \quad \text{and} \quad U(f, P) = \sum_R 1.v(R) = 1.$$

Since P is arbitrary, it follows that the lower integral of f over I equals 0, and the upper integral equals 1 . Thus f is not integrable over I.

Theorem 11.8. (The Riemann condition) . Let Q be a rectangle and $f : Q \to \mathbb{R}$ is a bounded function. Then

$$\underline{\int_Q} f \leq \overline{\int_Q} f;$$

equality holds if and only if given $\epsilon > 0$, there exists a partition P of Q for which

$$U(f, P) - L(f, P) < \epsilon.$$

Proof. Let P' be a fixed partition of Q. It follows from the fact that $L(f, P) \leq U(f, P)$ for every partition P of Q, that

$$\underline{\int_Q} f \leq U(f, P').$$

Now we use the fact that P' is arbitrary to conclude that

$$\underline{\int_Q} f \leq \overline{\int_Q} f.$$

Suppose now that the upper and lower integrals are equal and let $\epsilon > 0$ be arbitrary. So, there exist a partitions P and P' so that

$$\underline{\int_Q} f - \frac{\epsilon}{2} < L(f, P) \leq \underline{\int_Q} f = \int_Q f.$$

and

$$\int_Q f = \overline{\int_Q} f \leq U(f, P') < \overline{\int_Q} f + \frac{\epsilon}{2}.$$

Let $P'' = P \cup P'$. Then both the above inequalities simultaneously hold for P''. Thus, we get

$$\underline{\int_Q} f - \frac{\epsilon}{2} < L(f, P) \leq L(f, P'') \leq \int_Q f \leq U(f, P'') \leq U(f, P) < \overline{\int_Q} f + \frac{\epsilon}{2},$$

since P'' is the common refinement of P and P'. Thus, we get

$$U(f, P'') - L(f, P'') < \epsilon.$$

Conversely, suppose the upper and lower integrals are not equal. Let

$$\epsilon = \overline{\int_Q} f - \underline{\int_Q} f > 0.$$

Let P be any partition of Q. Then

$$L(f, P) \leq \underline{\int_Q} f < \overline{\int_Q} f \leq U(f, P);$$

which implies that

$$U(f, P) - L(f, P) \leq \overline{\int_Q} f - \underline{\int_Q} f = \epsilon$$

and the Riemann condition does not hold. □

Here is an easy application of this theorem.

Theorem 11.9. Every constant function $f(x) = c$ is integrable. Indeed, if Q is a rectangle and if P is a partition of Q, then

$$\int_Q c = c.v(Q) = c \sum_R v(R),$$

where the summation extends over all subrectangles determined by P.

Proof. If R is a subrectangle determined by P, then $m_R(f) = c = M_R(f)$. It follows that

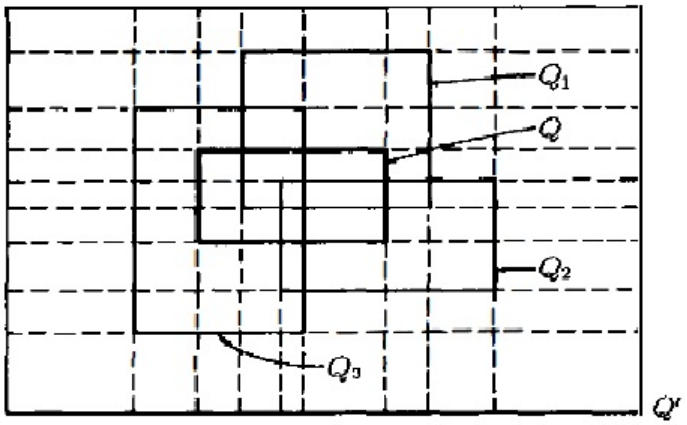
$$L(f, P) = c \sum_R v(R) = U(f, P),$$

so the Riemann condition holds trivially. Thus $\int_Q c$ exists; since it lies between $L(f, P)$ and $U(f, P)$, it must be equal to $c \sum_R v(R)$.

This result holds for any partition P. In particular, if P is the trivial partition whose only subrectangle is Q itself, then

$$\int_Q c = c.v(Q).$$

□



Corollary 11.10. Let Q be a rectangle in $\mathbb{R}^n$. Let $\{Q_1, \ldots, Q_k\}$ be a finite collection of rectangles that covers Q. Then

$$v(Q) \leq \sum_{i=1}^{k} v(Q_i).$$

Proof. Choose a rectangle Q' containing all the rectangles $Q_1, \ldots, Q_k$. Use the end points of the component intervals of the rectangles $Q, Q_1, \ldots, Q_k$ to define a partition P of Q'. Then each of the rectangles $Q, Q_1, \ldots, Q_k$ is a union of sub rectangles determined by P.

From the preceding theorem, we conclude that

$$v(Q) = \sum_{R \subset Q} v(R),$$

where the summation extends over all sub rectangles contained in Q. Because each such subrectangle R is contained in at least one of the rectangles $Q_1, \ldots, Q_k$, we have

$$\sum_{R \subset Q} v(R) \leq \sum_{i=1}^{k} \sum_{R \subset Q_i} v(R).$$

By the preceding theorem, we get

$$\sum_{R \subset Q_i} v(R) = v(Q_i),$$

and the corollary follows. □

In the case of $n = 1$, Q is a closed interval $[a, b]$ in $\mathbb{R}$ and we denote the integral of f over $[a, b]$ by one of the symbols

$$\int_a^b f \quad \text{or} \quad \int_{x=a}^{x=b} f(x)$$

instead of $\int_{[a,b]} f$.

Theorem 11.11. Let Q be a rectangle and $f, g : Q \to \mathbb{R}$ be bounded functions such that $f(x) \leq g(x)$ for $x \in Q$. Then

$$\underline{\int_Q} f \leq \underline{\int_Q} g \quad \text{and} \quad \overline{\int_Q} f \leq \overline{\int_Q} g.$$

Proof. Left as exercise. □

11.2 Few Probable Questions

1. Suppose $f : Q \to \mathbb{R}$ is continuous. Show that f is integrable over Q. Is the converse true? Justify.
2. State and prove the necessary and sufficient condition for integrability of a bounded function f, defined on a closed rectangle Q.
3. Show that any constant function f defined on a closed rectangle Q is always integrable.
4. Show that the function $f : [a, b] \to \mathbb{R}$ is not integrable over $[a, b]$ where

$$\begin{aligned} f(x) &= 0; \text{ if } x \text{ is rational} \\ &= 1; \text{ if } x \text{ is irrational} \end{aligned}$$

5. Let $I = [0, 1]^2 = [0, 1] \times [0, 1]$ and $f : I \to \mathbb{R}$ be defined by

$$\begin{aligned} f(x) &= 0; \text{ if } y \neq x \\ &= 1; \text{ if } y = x. \end{aligned}$$

   Show that f is integrable over I.

6. Let $f : \mathbb{R} \to \mathbb{R}$ be defined as

$$\begin{aligned} f(x) &= \frac{1}{q}; \text{ if } x = \frac{p}{q}, \text{ where } p \ \& \ q \text{ are positive integers having no common factor} \\ &= 0; \text{ otherwise} \end{aligned}$$

   Show that f is integrable over $[0, 1]$.

Unit 14

Course Structure

- Concept of Jordan measurability of a set in $\mathbb{R}^n$
- Some more conditions of integrability of a bounded function on a closed rectangle
- Integrals of type $f : C \to \mathbb{R}$, where $C \subset \mathbb{R}^n$ is not a rectangle

---

12 Introduction

Integration and measure zero sets are related in a very crucial way. We know that, in the one-dimensional case, a function f defined on a closed interval $[a, b]$ is integrable (due to Riemann) if and only if the set of discontinuities of f is of measure zero. We will try to find an analogous theorem for the multivariable case. First, we will define measure zero sets in $\mathbb{R}^n$ and then will move on to derive the necessary and sufficient condition of integrability of a bounded function f defined on a closed rectangle in connection to the measure zero sets.

Also, we so far have dealt with the integration of a bounded function f defined on a closed rectangle. We will see that, with the help of the closed rectangles we can define integrability of a bounded function, on any set, say C in $\mathbb{R}^n$. Let's explore!

Objectives

After reading this unit, you will be able to

- define measure zero sets in $\mathbb{R}^n$
- learn the characteristics of measure zero sets and see certain examples
- learn some more conditions of integrability of a bounded function f, defined on a closed rectangle Q in $\mathbb{R}^n$
- apply them in problems
- define the integration of a bounded function on any set C in $\mathbb{R}^n$, other than a closed rectangle
- learn certain related properties

12.1 Measure zero sets in $\mathbb{R}^n$

Definition 12.1. Let A be a subset of $\mathbb{R}^n$. We say that A has measure zero in $\mathbb{R}^n$ if for every $\epsilon > 0$, there is a cover $Q_1, Q_2, \ldots$ of A by countably many closed rectangles such that

$$\sum_{i=1}^{\infty} v(Q_i) < \epsilon.$$

If this inequality holds, we often say that the total volume of the rectangles $Q_1, Q_2, \ldots$ is less than ϵ.

A set with only finitely many points clearly has measure 0. If A has infinitely many points which can be arranged in a sequence $a_1, a_2, \ldots$, then A also has measure 0, since for $\epsilon > 0$, we can choose Q_i to be a closed rectangle containing a_i with

$$v(Q_i) < \frac{\epsilon}{2^i}.$$

Then,

$$\sum_{i=1}^{\infty} v(Q_i) < \sum_{i=1}^{\infty} \frac{\epsilon}{2^i} = \epsilon.$$

We derive some properties of sets of measure zero.

Theorem 12.2. 1. If $B \subset A$ and A has measure zero in $\mathbb{R}^n$, then so does B.

2. Let A be the union of the countable collection of sets $A_1, A_2, \ldots$ If each A_i has measure zero in $\mathbb{R}^n$, then so does A.

3. A set A has measure zero in $\mathbb{R}^n$ if and only if for every $\epsilon > 0$, there is a countable covering of A by open rectangles $\text{Int}Q_1, \text{Int}Q_2, \ldots$ such that

$$\sum_{i=1}^{\infty} v(Q_i) < \epsilon.$$

4. If Q is a rectangle in $\mathbb{R}^n$, then $\text{Bd}Q$ has measure zero in $\mathbb{R}^n$ but Q does not ($\text{Bd}Q$ is the boundary of Q).

Proof. 1. Let $\epsilon > 0$. Since A is measure zero set, so for the given ϵ, there is a cover $Q_1, Q_2, \ldots$ of A by countably many closed rectangles such that

$$\sum_{i=1}^{\infty} v(Q_i) < \epsilon.$$

Since $B \subset A$, so B satisfies the definition of zero measure in $\mathbb{R}^n$.

2. To prove 2,we cover the set A_j, for each j, by countably many rectangles

$$Q_{1j}, Q_{2j}, \ldots$$

of total volume less than $\epsilon/2^j$. Then the collection of rectangles $\{Q_{ij}\}$ is countable, that covers A, having total volume

$$\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} v(Q_{ij}) < \sum_{j=1}^{\infty} \frac{\epsilon}{2^j} = \epsilon.$$

Hence A is of measure zero.

3. If the open rectangles $\text{Int}Q_1, \text{Int}Q_2, \ldots$ cover A, then so do the rectangles $Q_1, Q_2, \ldots$ Thus the given condition implies that A has measure zero. Conversely, suppose A has measure zero. Cover A by rectangles $Q_1', Q_2', \ldots$ of total volume less than $\epsilon/2$. For each i, choose a rectangle Q_i such that

$$Q_i' \subset \text{Int } Q_i \quad \text{and} \quad v(Q_i) \leq 2v(Q_i').$$

This is possible because $v(Q)$ is a continuous function of the end points of the component intervals of Q. Then the open rectangles $\text{Int}Q_1, \text{Int}Q_2, \ldots$ cover A, and $\sum v(Q_i) < \epsilon$.

4. Let
$$Q = [a_1, b_1] \times \cdots \times [a_n, b_n].$$
The subset of Q consisting of those points x of Q for which $x_i = a_i$ is called one of the ith faces of Q. The other ith face consists of those x for which $x_i = b_i$. Each face of Q has measure zero in $\mathbb{R}^n$; for instance, the face for which $x_i = a_i$ can be covered by the single rectangle
$$[a_1, b_1] \times \cdots \times [a_i, a_i + \delta] \times \cdots \times [a_n, b_n],$$
whose volume may be made as small as desired by taking δ small. Now BdQ is the union of the faces of Q, which are finite in number. Therefore BdQ has measure zero in $\mathbb{R}^n$.

Now we suppose Q has measure zero in $\mathbb{R}^n$, and derive a contradiction . Set $\epsilon = v(Q)$. By 3, we can cover Q by open rectangles IntQ_1, Int$Q_2, \ldots$ with $\sum v(Q_i) < \epsilon$. Because Q is compact, we can cover Q by finitely many of these open sets, say IntQ_1, Int$Q_2, \ldots$ IntQ_k. But
$$\sum_{i=1}^{k} v(Q_i) < \epsilon,$$
which is a contradiction to a previous corollary we read in the previous unit.

□

By the third point of the above theorem, we can easily say that open rectangles may be used instead of closed rectangles in the definition of measure zero sets.

Definition 12.3. Let A be a subset of $\mathbb{R}^n$. We say that A has measure zero in $\mathbb{R}^n$ if for every $\epsilon > 0$, there is a cover $Q_1, Q_2, \ldots Q_n$ of A by finitely many closed rectangles such that
$$\sum_{i=1}^{n} v(Q_i) < \epsilon.$$

If A has content 0, then A clearly has measure 0. Again, open rectangles could be used instead of closed rectangles in the definition.

Theorem 12.4. If $a < b$, then $[a, b] \subset \mathbb{R}$ does not have content 0. In fact, if $Q_1, Q_2, \ldots Q_n$ is a finite cover of $[a, b]$ by closed intervals, then
$$\sum_{i=1}^{n} v(Q_i) \geq b - a.$$

Proof. Clearly we can assume that each $Q_i \subset [a, b]$. Let $a = t_0 < t_1 < t_2 < \cdots < t_k = b$ be all endpoints of all Q_i. Then, each $v(Q_i)$ is the sum of certain $t_j - t_{j-1}$. Moreover, each $[t_{j-1}, t_j]$ lies in at least one Q_i (namely, any one which contains an interior point of $[t_{j-1}, t_j]$), so that
$$\sum_{i=1}^{n} v(Q_i) \geq \sum_{j=1}^{k} (t_j - t_{j-1}) = b - a.$$

□

If $a < b$, it is also true that $[a, b]$ does not have measure 0. This follows from

Theorem 12.5. If A is compact and has measure 0, then A has content 0.

Proof. Let $\epsilon > 0$. Since A has measure 0, there is a cover $\{Q_1, Q_2, \ldots\}$ of A by open rectangles such that

$$\sum_{i=1}^{\infty} v(Q_i) < \epsilon.$$

Since A is compact, a finite subcover $\{Q_1, Q_2, \ldots Q_n\}$ of A for which

$$\sum_{i=1}^{n} v(Q_i) < \epsilon.$$

□

The conclusion of the above theorem is false if A is not compact. For example, let A be the set of rational numbers between 0 and 1; then A has measure 0. Suppose, however, that $\{[a_1, b_1], \ldots, [a_n, b_n]\}$ covers A. Then A is contained in the closed set $[a_1, b_1] \cup \cdots \cup [a_n, b_n]$, and hence $[0, 1] \subset [a_1, b_1] \cup \cdots \cup [a_n, b_n]$. Thus, we get

$$\sum_{i=1}^{n} (b_i - a_i) \geq 1$$

for any such cover, and consequently A does not have content 0.

Recall that $o(f, x)$ denotes the oscillation of f at x.

Lemma 12.6. Let Q be a closed rectangle and let $f : Q \to \mathbb{R}$ be a bounded function such that $o(f, x) < \epsilon$ for all $x \in Q$. Then there is a partition P of Q such that $U(f, P) - L(f, P) < \epsilon.v(Q)$.

Proof. For each $x \in A$, there is a closed rectangle Q_x containing x in its interior, such that $M_{Q_x}(f) - m_{Q_x}(f) < \epsilon$. Since Q is compact, there exists a finite number $Q_{x_1}, \ldots, Q_{x_n}$ of the sets Q_x that cover Q. Let P be a partition for Q such that each subrectangle S of P is contained in some Q_{x_i}. Then $M_S(f) - m_S(f) < \epsilon$ for each subrectangle S of P, so that

$$U(f, P) - L(f, P) = \sum_{S} [M_S(f) - m_S(f)].v(S) < \epsilon.v(A).$$

□

Theorem 12.7. Let Q be a closed rectangle and let $f : Q \to \mathbb{R}$ be a bounded function. Let $B = \{x : \ f$ is not continuous at $x\}$. Then f is integrable if and only if B is a set of measure 0.

Proof. Suppose first that B has measure 0. Let $\epsilon > 0$ and let $B_\epsilon = \{x : \ o(f, x) \geq \epsilon\}$. Then $B_\epsilon \subset B$, so that B_ϵ has measure zero. Since B_ϵ is compact, it has content zero. Thus, there exist a finite collection $Q_1, \ldots, Q_n$ of closed rectangles, whose interiors cover B_ϵ, such that $\sum_{i=1}^{n} v(Q_i) < \epsilon$. Let P be a partition of Q such that every subrectangle S of P is in one of two groups

1. S_1, which consists of subrectangles S, such that $S \subset Q_i$ for some i.
2. S_2, which consists of subrectangles S with $S \cap B_\epsilon = \emptyset$.

Let $|f(x)| < M$ for $x \in Q$. Then $M_S(f) - m_S(f) < 2M$ for every S. Hence

$$\sum_{S \subset S} [M_S(f) - m_S(f)].v(S) < 2M \sum_{i=1}^{n} v(Q_i) < 2M\epsilon.$$

Now, if $S \in S_2$, then $o(f,x) < \epsilon$ for $x \in S$. The previous lemma implies that there is a refinement P' of P such that

$$\sum_{S' \subset S} [M_{S'}(f) - m_{S'}(f)].v(S') < \epsilon.v(S)$$

for $S \in S_2$. Then

$$\begin{aligned} U(f,P') - L(f,P') &= \sum_{S' \subset S \in S_1} [M_{S'}(f) - m_{S'}(f)].v(S') + \sum_{S' \subset S \in S_2} [M_{S'}(f) - m_{S'}(f)].v(S') \\ &< 2M\epsilon + \sum_{S \in S_2} \epsilon.v(S) \\ &\leq 2M\epsilon + \epsilon.v(Q). \end{aligned}$$

Since M and $v(Q)$ are fixed, this shows that we can find a partition P' with $U(f,P') - L(f,P')$ as small as desired. Thus f is integrable.

Suppose, conversely, that f is integrable. Since $B = B_1 \cup B_{1/2} \cup B_{1/3} \cup \cdots$, it suffices to prove that each $B_{1/n}$ has measure 0. In fact we will show that each $B_{1/n}$ has content zero (since $B_{1/n}$ is compact, this is actually equivalent).

Let $\epsilon > 0$, and let P be a partition of Q such that

$$U(f,P) - L(f,P) < \epsilon/n.$$

Let $\mathcal{S}$ be the collection of subrectangles S of P which intersect $B_{1/n}$. Then $\mathcal{S}$ is a cover of $B_{1/n}$. Now, if $S \in \mathcal{S}$, then $M_S(f) - m_S(f) \geq 1/n$. Thus

$$\begin{aligned} \frac{1}{n} \sum_{S \in \mathcal{S}} v(S) &\leq \sum_{S \in \mathcal{S}} [M_S(f) - m_S(f)].v(S) \\ &\leq \sum_{S} [M_S(f) - m_S(f)].v(S) \\ &< \frac{\epsilon}{n}, \end{aligned}$$

and so

$$\sum_{S \in \mathcal{S}} v(S) < \epsilon.$$

□

Exercise 12.8. 1. Show that any finite set in $\mathbb{R}^n$ has measure zero.

12.2 Integrals of functions on sets other than rectangles

We have thus far dealt only with the integrals of functions over rectangles. Integrals over other sets are easily reduced to this type. If $C \in \mathbb{R}^n$, the characteristic function χ_C of C is defined by

$$\begin{aligned} \chi_C(x) &= 0, \ c \notin C, \\ &= 1, \ x \in C. \end{aligned}$$

If $C \subset Q$ for some closed rectangle Q and $f : A \to \mathbb{R}$ bounded, then $\int_C f$ is defined as $\int_A f.\chi_C$ is integrable. This certainly occurs if f and χ_C are integrable.

Theorem 12.9. The function $\chi_C : Q \to \mathbb{R}$ i8 integrable if and only if the boundary of C has measure zero (and hence content zero).

Proof. If x is in the interior of C, then there is an open rectangle U with $x \in U \subset C$. Thus, $\chi_C = 1$ on U and χ_C is clearly continuous at x. Similarly, if x is in the exterior of C, there is an open rectangle U with $x \in U \subset \mathbb{R}^n \setminus C$. Hence $\chi_C = 0$ on U and χ_C is continuous at x. Finally, if x is in the boundary of C, then for every open rectangle U containing x, there is $y_1 \in U \cap C$, so that $\chi_C(y_1) = 1$ and there is $y_2 \in U \cap (\mathbb{R}^n \setminus C)$, so that $\chi_C(y_2) = 0$. Hence χ_C is not continuous at x. Thus, $\{x : \ \chi_C \text{ is not continuous at } x\}$ =boundary of C and the result follows by the previous theorem. □

A bounded set C whose boundary has measure 0 is called Jordan-measurable. The integral $\int_C 1$ is called the n-dimensional content of C, or the n-dimensional volume of C. Naturally one-dimensional volume is often called length, and two-dimensional volume, area.

12.3 Few Probable Questions

1. Define measure zero set in $\mathbb{R}^n$. Show that a countable set in $\mathbb{R}^n$ has measure zero.
2. Deduce a necessary and sufficient condition for a bounded function defined on a closed rectangle to be integrable.
3. Define content zero sets. Show that a content zero set is of measure zero.
4. Deduce a necessary and sufficient condition for a bounded function defined on a bounded set C of $\mathbb{R}^n$ to be integrable.

Units 15, 16

FUBINI'S THEOREM

The problem of calculating integrals is solved, in some sense, by Theorem 3-10, which reduces the computation of integrals over a closed rectangle in $\mathbf{R}^n$, $n > 1$, to the computation of integrals over closed intervals in $\mathbf{R}$. Of sufficient importance to deserve a special designation, this theorem is usually referred to as Fubini's theorem, although it is more or less a

special case of a theorem proved by Fubini long after Theorem 3-10 was known.

The idea behind the theorem is best illustrated (Figure 3-2) for a positive continuous function f: $[a,b] \times [c,d] \to \mathbf{R}$. Let $t_0, \ldots, t_n$ be a partition of $[a,b]$ and divide $[a,b] \times [c,d]$ into n strips by means of the line segments $\{t_i\} \times [c,d]$. If g_x is defined by $g_x(y) = f(x,y)$, then the area of the region under the graph of f and above $\{x\} \times [c,d]$ is

$$\int_c^d g_x = \int_c^d f(x,y)dy.$$

The volume of the region under the graph of f and above $[t_{i-1},t_i] \times [c,d]$ is therefore approximately equal to $(t_i - t_{i-1}) \cdot \int_c^d f(x,y)dy$, for any $x \in [t_{i-1},t_i]$. Thus

$$\int_{[a,b]\times[c,d]} f = \sum_{i=1}^{n} \int_{[t_{i-1},t_i]\times[c,d]} f$$

is approximately $\Sigma_{i=1}^{n}(t_i - t_{i-1}) \cdot \int_c^d f(x_i,y)dy$, with x_i in

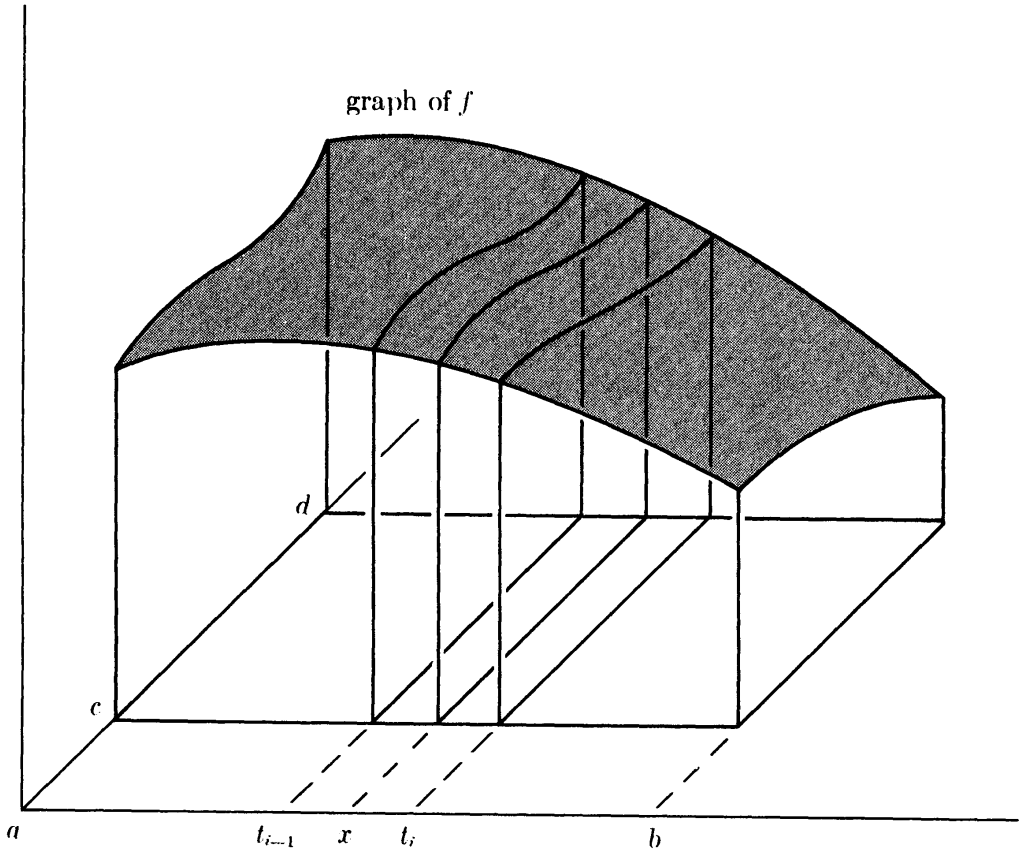


FIGURE 3-2

$[t_{i-1},t_i]$. On the other hand, sums similar to these appear in the definition of $\int_a^b(\int_c^d f(x,y)dy)dx$. Thus, if h is defined by $h(x) = \int_c^d g_x = \int_c^d f(x,y)dy$, it is reasonable to hope that h is integrable on $[a,b]$ and that

$$\int_{[a,b]\times[c,d]} f = \int_a^b h = \int_a^b \left(\int_c^d f(x,y)dy\right) dx.$$

This will indeed turn out to be true when f is continuous, but in the general case difficulties arise. Suppose, for example, that the set of discontinuities of f is $\{x_0\} \times [c,d]$ for some $x_0 \in [a,b]$. Then f is integrable on $[a,b] \times [c,d]$ but $h(x_0) = \int_c^d f(x_0,y)dy$ may not even be defined. The statement of Fubini's theorem therefore looks a little strange, and will be followed by remarks about various special cases where simpler statements are possible.

We will need one bit of terminology. If $f\colon A \to \mathbf{R}$ is a bounded function on a closed rectangle, then, whether or not f is integrable, the least upper bound of all lower sums, and the greatest lower bound of all upper sums, both exist. They are called the **lower** and **upper integrals** of f on A, and denoted

$$\mathbf{L}\int_A f \qquad \text{and} \qquad \mathbf{U}\int_A f,$$

respectively.

3-10 Theorem (Fubini's Theorem). *Let $A \subset \mathbf{R}^n$ and $B \subset \mathbf{R}^m$ be closed rectangles, and let $f\colon A \times B \to \mathbf{R}$ be integrable. For $x \in A$ let $g_x\colon B \to \mathbf{R}$ be defined by $g_x(y) = f(x,y)$ and let*

$$\mathfrak{L}(x) = \mathbf{L}\int_B g_x = \mathbf{L}\int_B f(x,y)dy,$$
$$\mathfrak{U}(x) = \mathbf{U}\int_B g_x = \mathbf{U}\int_B f(x,y)dy.$$

Then $\mathfrak{L}$ and $\mathfrak{U}$ are integrable on A and

$$\int_{A\times B} f = \int_A \mathfrak{L} = \int_A \left(\mathbf{L}\int_B f(x,y)dy\right) dx,$$
$$\int_{A\times B} f = \int_A \mathfrak{U} = \int_A \left(\mathbf{U}\int_B f(x,y)dy\right) dx.$$

(The integrals on the right side are called **iterated integrals** for f.)

Proof. Let P_A be a partition of A and P_B a partition of B. Together they give a partition P of $A \times B$ for which any subrectangle S is of the form $S_A \times S_B$, where S_A is a subrectangle of the partition P_A, and S_B is a subrectangle of the partition P_B. Thus

$$L(f,P) = \sum_S m_S(f) \cdot v(S) = \sum_{S_A,S_B} m_{S_A \times S_B}(f) \cdot v(S_A \times S_B)$$
$$= \sum_{S_A} \Big(\sum_{S_B} m_{S_A \times S_B}(f) \cdot v(S_B) \Big) \cdot v(S_A).$$

Now, if $x \in S_A$, then clearly $m_{S_A \times S_B}(f) \leq m_{S_B}(g_x)$. Consequently, for $x \in S_A$ we have

$$\sum_{S_B} m_{S_A \times S_B}(f) \cdot v(S_B) \leq \sum_{S_B} m_{S_B}(g_x) \cdot v(S_B) \leq \mathbf{L} \int_B g_x = \mathfrak{L}(x).$$

Therefore

$$\sum_{S_A} \Big(\sum_{S_B} m_{S_A \times S_B}(f) \cdot v(S_B) \Big) \cdot v(S_A) \leq L(\mathfrak{L},P_A).$$

We thus obtain

$$L(f,P) \leq L(\mathfrak{L},P_A) \leq U(\mathfrak{L},P_A) \leq U(\mathfrak{U},P_A) \leq U(f,P),$$

where the proof of the last inequality is entirely analogous to the proof of the first. Since f is integrable, $\sup\{L(f,P)\} = \inf\{U(f,P)\} = \int_{A \times B} f$. Hence

$$\sup\{L(\mathfrak{L},P_A)\} = \inf\{U(\mathfrak{L},P_A)\} = \int_{A \times B} f.$$

In other words, $\mathfrak{L}$ is integrable on A and $\int_{A \times B} f = \int_A \mathfrak{L}$. The assertion for $\mathfrak{U}$ follows similarly from the inequalities

$$L(f,P) \leq L(\mathfrak{L},P_A) \leq L(\mathfrak{U},P_A) \leq U(\mathfrak{U},P_A) \leq U(f,P). \quad \blacksquare$$

Remarks. 1. A similar proof shows that

$$\int_{A \times B} f = \int_B \Big(\mathbf{L} \int_A f(x,y)dx \Big) dy = \int_B \Big(\mathbf{U} \int_A f(x,y)dx \Big) dy.$$

These integrals are called *iterated integrals* for f in the reverse order from those of the theorem. As several problems show, the possibility of interchanging the orders of iterated integrals has many consequences.

2. In practice it is often the case that each g_x is integrable, so that $\int_{A\times B} f = \int_A (\int_B f(x,y)dy)dx$. This certainly occurs if f is continuous.

3. The worst irregularity commonly encountered is that g_x is not integrable for a finite number of $x \in A$. In this case $\mathfrak{L}(x) = \int_B f(x,y)dy$ for all but these finitely many x. Since $\int_A \mathfrak{L}$ remains unchanged if $\mathfrak{L}$ is redefined at a finite number of points, we can still write $\int_{A\times B} f = \int_A (\int_B f(x,y)dy)dx$, provided that $\int_B f(x,y)dy$ is defined arbitrarily, say as 0, when it does not exist.

4. There are cases when this will not work and Theorem 3-10 must be used as stated. Let f: $[0,1] \times [0,1] \to \mathbf{R}$ be defined by

$$f(x,y) = \begin{cases} 1 & \text{if } x \text{ is irrational,} \\ 1 & \text{if } x \text{ is rational and } y \text{ is irrational,} \\ 1 - 1/q & \text{if } x = p/q \text{ in lowest terms and } y \text{ is rational.} \end{cases}$$

Then f is integrable and $\int_{[0,1]\times[0,1]} f = 1$. Now $\int_0^1 f(x,y)dy = 1$ if x is irrational, and does not exist if x is rational. Therefore h is not integrable if $h(x) = \int_0^1 f(x,y)dy$ is set equal to 0 when the integral does not exist.

5. If $A = [a_1,b_1] \times \cdots \times [a_n,b_n]$ and f: $A \to \mathbf{R}$ is sufficiently nice, we can apply Fubini's theorem repeatedly to obtain

$$\int_A f = \int_{a_n}^{b_n} \left(\cdots \left(\int_{a_1}^{b_1} f(x^1, \ldots, x^n)dx^1 \right) \cdots \right) dx^n.$$

6. If $C \subset A \times B$, Fubini's theorem can be used to evaluate $\int_C f$, since this is by definition $\int_{A\times B} \chi_C f$. Suppose, for example, that

$$C = [-1,1] \times [-1,1] - \{(x,y): |(x,y)| < 1\}.$$

Then

$$\int_C f = \int_{-1}^{1} \left(\int_{-1}^{1} f(x,y) \cdot \chi_C(x,y)dy \right) dx.$$

Now

$$\chi_C(x,y) = \begin{cases} 1 & \text{if } y > \sqrt{1-x^2} \text{ or } y < -\sqrt{1-x^2}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore

$$\int_{-1}^{1} f(x,y) \cdot \chi_C(x,y)dy = \int_{-1}^{-\sqrt{1-x^2}} f(x,y)dy + \int_{\sqrt{1-x^2}}^{1} f(x,y)dy.$$

In general, if $C \subset A \times B$, the main difficulty in deriving expressions for $\int_C f$ will be determining $C \cap (\{x\} \times B)$ for $x \in A$. If $C \cap (A \times \{y\})$ for $y \in B$ is easier to determine, one should use the iterated integral

$$\int_C f = \int_B \left(\int_A f(x,y) \cdot \chi_C(x,y)dx \right) dy.$$

Problems. **3-23.** Let $C \subset A \times B$ be a set of content 0. Let $A' \subset A$ be the set of all $x \in A$ such that $\{y \in B: (x,y) \in C\}$ is *not* of content 0. Show that A' is a set of measure 0. *Hint:* χ_C is integrable and $\int_{A\times B} \chi_C = \int_A \mathfrak{U} = \int_A \mathfrak{L}$, so $\int_A \mathfrak{U} - \mathfrak{L} = 0$.

3-24. Let $C \subset [0,1] \times [0,1]$ be the union of all $\{p/q\} \times [0, 1/q]$, where p/q is a rational number in $[0,1]$ written in lowest terms. Use C to show that the word "measure" in Problem 3-23 cannot be replaced by "content."

3-25. Use induction on n to show that $[a_1,b_1] \times \cdots \times [a_n,b_n]$ is not a set of measure 0 (or content 0) if $a_i < b_i$ for each i.

3-26. Let $f: [a,b] \to \mathbf{R}$ be integrable and non-negative and let $A_f = \{(x,y): a \leq x \leq b \text{ and } 0 \leq y \leq f(x)\}$. Show that A_f is Jordan-measurable and has area $\int_a^b f$.

3-27. If $f: [a,b] \times [a,b] \to \mathbf{R}$ **is continuous, show that**

$$\int_a^b \int_a^y f(x,y)dx\,dy = \int_a \int_x f(x,y)dy\,dx.$$

Hint: Compute $\int_C f$ in two different ways for a suitable set $C \subset [a,b] \times [a,b]$.

3-28.* Use Fubini's theorem to give an easy proof that $D_{1,2}f = D_{2,1}f$ if these are continuous. *Hint:* If $D_{1,2}f(a) - D_{2,1}f(a) > 0$, there is a rectangle A containing a such that $D_{1,2}f - D_{2,1}f > 0$ on A.

3-29. Use Fubini's theorem to derive an expression for the volume of a set of $\mathbf{R}^3$ obtained by revolving a Jordan-measurable set in the yz-plane about the z-axis.

3-30. Let C be the set in Problem 1-17. Show that

$$\int_{[0,1]} \left(\int_{[0,1]} \chi_C(x,y)dx \right) dy = \int_{[0,1]} \left(\int_{[0,1]} \chi_C(y,x)dy \right) dx = 0$$

but that $\int_{[0,1]\times[0,1]} \chi_C$ does not exist.

3-31. If $A = [a_1,b_1] \times \cdots \times [a_n,b_n]$ and $f\colon A \to \mathbf{R}$ is continuous, define $F\colon A \to \mathbf{R}$ by

$$F(x) = \int_{[a_1,x^1]\times \cdots \times [a_n,x^n]} f.$$

What is $D_iF(x)$, for x in the interior of A?

3-32.* Let $f\colon [a,b] \times [c,d] \to \mathbf{R}$ be continuous and suppose D_2f is continuous. Define $F(y) = \int_a^b f(x,y)dx$. Prove *Leibnitz's rule:* $F'(y) = \int_a^b D_2f(x,y)dx$. *Hint:* $F(y) = \int_a^b f(x,y)dx = \int_a^b (\int_c^y D_2f(x,y)dy + f(x,c))dx$. (The proof will show that continuity of D_2f may be replaced by considerably weaker hypotheses.)

3-33. If $f\colon [a,b] \times [c,d] \to \mathbf{R}$ is continuous and D_2f is continuous, define $F(x,y) = \int_a^x f(t,y)dt$.

(a) Find D_1F and D_2F.

(b) If $G(x) = \int_a^{g(x)} f(t,x)dt$, find $G'(x)$.

3-34.* Let $g_1,g_2\colon \mathbf{R}^2 \to \mathbf{R}$ be continuously differentiable and suppose $D_1g_2 = D_2g_1$. As in Problem 2-21, let

$$f(x,y) = \int_0^x g_1(t,0)dt + \int_0^y g_2(x,t)dt.$$

Show that $D_1f(x,y) = g_1(x,y)$.

3-35.* (*a*) Let $g\colon \mathbf{R}^n \to \mathbf{R}^n$ be a linear transformation of one of the following types:

$$\begin{cases} g(e_i) = e_i & i \neq j \\ g(e_j) = ae_j \end{cases}$$

$$\begin{cases} g(e_i) = e_i & i \neq j \\ g(e_j) = e_j + e_k \end{cases}$$

$$\begin{cases} g(e_k) = e_k & k \neq i, j \\ g(e_i) = e_j \\ g(e_j) = e_i. \end{cases}$$

If U is a rectangle, show that the volume of $g(U)$ is $|\det g| \cdot v(U)$.

(b) Prove that $|\det g| \cdot v(U)$ is the volume of $g(U)$ for any linear transformation $g\colon \mathbf{R}^n \to \mathbf{R}^n$. *Hint:* If $\det g \neq 0$, then g is the composition of linear transformations of the type considered in (a).

3-36. (Cavalieri's principle). Let A and B be Jordan-measurable subsets of $\mathbf{R}^3$. Let $A_c = \{(x,y)\colon (x,y,c) \in A\}$ and define B_c similarly. Suppose each A_c and B_c are Jordan-measurable and have the same area. Show that A and B have the same volume.

PARTITIONS OF UNITY

In this section we introduce a tool of extreme importance in the theory of integration.

3-11 Theorem. *Let $A \subset \mathbf{R}^n$ and let $\mathcal{O}$ be an open cover of A. Then there is a collection Φ of C^∞ functions φ defined in an open set containing A, with the following properties:*

(1) *For each $x \in A$ we have $0 \leq \varphi(x) \leq 1$.*
(2) *For each $x \in A$ there is an open set V containing x such that all but finitely many $\varphi \in \Phi$ are 0 on V.*
(3) *For each $x \in A$ we have $\Sigma_{\varphi \in \Phi} \varphi(x) = 1$* (by (2) for each x this sum is finite in some open set containing x).
(4) *For each $\varphi \in \Phi$ there is an open set U in $\mathcal{O}$ such that $\varphi = 0$ outside of some closed set contained in U.*

(A collection Φ satisfying (1) to (3) is called a C^∞ **partition of unity** for A. If Φ also satisfies (4), it is said to be **subordinate** to the cover $\mathcal{O}$. In this chapter we will only use continuity of the functions φ.)

Proof. *Case 1. A is compact.*

Then a finite number $U_1, \ldots, U_n$ of open sets in $\mathcal{O}$ cover A. It clearly suffices to construct a partition of unity subordinate to the cover $\{U_1, \ldots, U_n\}$. We will first find compact sets $D_i \subset U_i$ whose interiors cover A. The sets D_i are constructed inductively as follows. Suppose that $D_1, \ldots, D_k$ have been chosen so that {interior $D_1, \ldots,$ interior D_k, $U_{k+1}, \ldots, U_n$} covers A. Let

$$C_{k+1} = A - (\text{int } D_1 \cup \cdots \cup \text{int } D_k \cup U_{k+2} \cup \cdots \cup U_n).$$

Then $C_{k+1} \subset U_{k+1}$ is compact. Hence (Problem 1-22) we can find a compact set D_{k+1} such that

$$C_{k+1} \subset \text{interior } D_{k+1} \quad \text{and} \quad D_{k+1} \subset U_{k+1}.$$

Having constructed the sets $D_1, \ldots, D_n$, let ψ_i be a non-negative C^∞ function which is positive on D_i and 0 outside of some closed set contained in U_i (Problem 2-26). Since

$\{D_1, \ldots, D_n\}$ covers A, we have $\psi_1(x) + \cdots + \psi_n(x) > 0$ for all x in some open set U containing A. On U we can define

$$\varphi_i(x) = \frac{\psi_i(x)}{\psi_1(x) + \cdots + \psi_n(x)}.$$

If $f\colon U \to [0,1]$ is a C^∞ function which is 1 on A and 0 outside of some closed set in U, then $\Phi = \{f \cdot \varphi_1, \ldots, f \cdot \varphi_n\}$ is the desired partition of unity.

Case 2. $A = A_1 \cup A_2 \cup A_3 \cup \cdots$, *where each* A_i *is compact and* $A_i \subset$ *interior* A_{i+1}.

For each i let $\mathcal{O}_i$ consist of all $U \cap$ (interior $A_{i+1} - A_{i-2}$) for U in $\mathcal{O}$. Then $\mathcal{O}_i$ is an open cover of the compact set $B_i = A_i -$ interior A_{i-1}. By case 1 there is a partition of unity Φ_i for B_i, subordinate to $\mathcal{O}_i$. For each $x \in A$ the sum

$$\sigma(x) = \sum_{\varphi \in \Phi_i, \text{ all } i} \varphi(x)$$

is a finite sum in some open set containing x, since if $x \in A_i$ we have $\varphi(x) = 0$ for $\varphi \in \Phi_j$ with $j \geq i + 2$. For each φ in each Φ_i, define $\varphi'(x) = \varphi(x)/\sigma(x)$. The collection of all φ' is the desired partition of unity.

Case 3. A *is open.*

Let $A_i =$

$$\{x \in A\colon |x| \leq i \text{ and distance from } x \text{ to boundary } A \geq 1/i\},$$

and apply case 2.

Case 4. A *is arbitrary.*

Let B be the union of all U in $\mathcal{O}$. By case 3 there is a partition of unity for B; this is also a partition of unity for A. ∎

An important consequence of condition (2) of the theorem should be noted. Let $C \subset A$ be compact. For each $x \in C$ there is an open set V_x containing x such that only finitely many $\varphi \in \Phi$ are not 0 on V_x. Since C is compact, finitely many such V_x cover C. Thus only finitely many $\varphi \in \Phi$ are not 0 on C.

One important application of partitions of unity will illustrate their main role—piecing together results obtained locally.

An open cover $\mathcal{O}$ of an open set $A \subset \mathbf{R}^n$ is **admissible** if each $U \in \mathcal{O}$ is contained in A. If Φ is subordinate to $\mathcal{O}$, $f: A \to \mathbf{R}$ is bounded in some open set around each point of A, and $\{x: f \text{ is discontinuous at } x\}$ has measure 0, then each $\int_A \varphi \cdot |f|$ exists. We define f to be **integrable** (in the extended sense) if $\Sigma_{\varphi \in \Phi} \int_A \varphi \cdot |f|$ converges (the proof of Theorem 3-11 shows that the φ's may be arranged in a sequence). This implies convergence of $\Sigma_{\varphi \in \Phi} |\int_A \varphi \cdot f|$, and hence absolute convergence of $\Sigma_{\varphi \in \Phi} \int_A \varphi \cdot f$, which we define to be $\int_A f$. These definitions do not depend on $\mathcal{O}$ or Φ (but see Problem 3-38).

3-12 Theorem.

(1) *If Ψ is another partition of unity, subordinate to an admissible cover $\mathcal{O}'$ of A, then $\Sigma_{\psi \in \Psi} \int_A \psi \cdot |f|$ also converges, and*

$$\sum_{\varphi \in \Phi} \int_A \varphi \cdot f = \sum_{\psi \in \Psi} \int_A \psi \cdot f.$$

(2) *If A and f are bounded, then f is integrable in the extended sense.*

(3) *If A is Jordan-measurable and f is bounded, then this definition of $\int_A f$ agrees with the old one.*

Proof

(1) Since $\varphi \cdot f = 0$ except on some compact set C, and there are only finitely many ψ which are non-zero on C, we can write

$$\sum_{\varphi \in \Phi} \int_A \varphi \cdot f = \sum_{\varphi \in \Phi} \int_A \sum_{\psi \in \Psi} \psi \cdot \varphi \cdot f = \sum_{\varphi \in \Phi} \sum_{\psi \in \Psi} \int_A \psi \cdot \varphi \cdot f.$$

This result, applied to $|f|$, shows the convergence of $\Sigma_{\varphi \in \Phi} \Sigma_{\psi \in \Psi} \int_A \psi \cdot \varphi \cdot |f|$, and hence of $\Sigma_{\varphi \in \Phi} \Sigma_{\psi \in \Psi} |\int_A \psi \cdot \varphi \cdot f|$. This absolute convergence justifies interchanging the order of summation in the above equation; the resulting double sum clearly equals $\Sigma_{\psi \in \Psi} \int_A \psi \cdot f$. Finally, this result applied to $|f|$ proves convergence of $\Sigma_{\psi \in \Psi} \int_A \psi \cdot |f|$.

(2) If A is contained in the closed rectangle B and $|f(x)| \leq M$ for $x \in A$, and $F \subset \Phi$ is finite, then

$$\sum_{\varphi \in F} \int_A \varphi \cdot |f| \leq \sum_{\varphi \in F} M \int_A \varphi = M \int_A \sum_{\varphi \in F} \varphi \leq Mv(B),$$

since $\Sigma_{\varphi \in F}\, \varphi \leq 1$ on A.

(3) If $\varepsilon > 0$ there is (Problem 3-22) a compact Jordan-measurable $C \subset A$ such that $\int_{A-C} 1 < \varepsilon$. There are only finitely many $\varphi \in \Phi$ which are non-zero on C. If $F \subset \Phi$ is any finite collection which includes these, and $\int_A f$ has its old meaning, then

$$\begin{aligned} \left| \int_A f - \sum_{\varphi \in F} \int_A \varphi \cdot f \right| &\leq \int_A \left| f - \sum_{\varphi \in F} \varphi \cdot f \right| \\ &\leq M \int_A \left(1 - \sum_{\varphi \in F} \varphi\right) \\ &= M \int_A \sum_{\varphi \in \Phi - F} \varphi \leq M \int_{A-C} 1 \leq M\varepsilon. \quad \blacksquare \end{aligned}$$

Problems. **3-37.** (a) Suppose that $f\colon (0,1) \to \mathbf{R}$ is a non-negative continuous function. Show that $\int_{(0,1)} f$ exists if and only if $\lim_{\varepsilon \to 0} \int_\varepsilon^{1-\varepsilon} f$ exists.

(b) Let $A_n = [1 - 1/2^n, 1 - 1/2^{n+1}]$. Suppose that $f\colon (0,1) \to \mathbf{R}$ satisfies $\int_{A_n} f = (-1)^n/n$ and $f(x) = 0$ for $x \notin$ any A_n. Show that $\int_{(0,1)} f$ does not exist, but $\lim_{\varepsilon \to 0} \int_{(\varepsilon, 1-\varepsilon)} f = \log 2$.

3-38. Let A_n be a closed set contained in $(n, n+1)$. Suppose that $f\colon \mathbf{R} \to \mathbf{R}$ satisfies $\int_{A_n} f = (-1)^n/n$ and $f = 0$ for $x \notin$ any A_n. Find two partitions of unity Φ and Ψ such that $\Sigma_{\varphi \in \Phi} \int_{\mathbf{R}} \varphi \cdot f$ and $\Sigma_{\psi \in \Psi} \int_{\mathbf{R}} \psi \cdot f$ converge absolutely to different values.

CHANGE OF VARIABLE

If $g\colon [a,b] \to \mathbf{R}$ is continuously differentiable and $f\colon \mathbf{R} \to \mathbf{R}$ is continuous, then, as is well known,

$$\int_{g(a)}^{g(b)} f = \int_a^b (f \circ g) \cdot g'.$$

The proof is very simple: if $F' = f$, then $(F \circ g)' = (f \circ g) \cdot g'$; thus the left side is $F(g(b)) - F(g(a))$, while the right side is $F \circ g(b) - F \circ g(a) = F(g(b)) - F(g(a))$.

We leave it to the reader to show that if g is 1-1, then the above formula can be written

$$\int_{g((a,b))} f = \int_{(a,b)} f \circ g \cdot |g'|.$$

(Consider separately the cases where g is increasing and where g is decreasing.) The generalization of this formula to higher dimensions is by no means so trivial.

3-13 Theorem. *Let $A \subset \mathbf{R}^n$ be an open set and $g: A \to \mathbf{R}^n$ a 1-1, continuously differentiable function such that $\det g'(x) \neq 0$ for all $x \in A$. If $f: g(A) \to \mathbf{R}$ is integrable, then*

$$\int_{g(A)} f = \int_A (f \circ g)|\det g'|.$$

Proof. We begin with some important reductions.

1. Suppose there is an admissible cover $\mathcal{O}$ for A such that for each $U \in \mathcal{O}$ and any integrable f we have

$$\int_{g(U)} f = \int_U (f \circ g)|\det g'|.$$

Then the theorem is true for all of A. (Since g is automatically 1-1 in an open set around each point, it is not surprising that this is the only part of the proof using the fact that g is 1-1 on all of A.)

Proof of (1). The collection of all $g(U)$ is an open cover of $g(A)$. Let Φ be a partition of unity subordinate to this cover. If $\varphi = 0$ outside of $g(U)$, then, since g is 1-1, we have $(\varphi \cdot f) \circ g$

$= 0$ outside of U. Therefore the equation

$$\int_{g(U)} \varphi \cdot f = \int_U [(\varphi \cdot f) \circ g] |\det g'|.$$

can be written

$$\int_{g(A)} \varphi \cdot f = \int_A [(\varphi \cdot f) \circ g] |\det g'|.$$

Hence

$$\begin{aligned} \int_{g(A)} f = \sum_{\varphi \in \Phi} \int_{g(A)} \varphi \cdot f &= \sum_{\varphi \in \Phi} \int_A [(\varphi \cdot f) \circ g] |\det g'| \\ &= \sum_{\varphi \in \Phi} \int_A (\varphi \circ g)(f \circ g) |\det g'| \\ &= \int_A (f \circ g) |\det g'|. \end{aligned}$$

Remark. The theorem also follows from the assumption that

$$\int_V f = \int_{g^{-1}(V)} (f \circ g) |\det g'|$$

for V in some admissible cover of $g(A)$. This follows from (1) applied to g^{-1}.

2. It suffices to prove the theorem for the function $f = 1$.

Proof of (2). If the theorem holds for $f = 1$, it holds for constant functions. Let V be a rectangle in $g(A)$ and P a partition of V. For each subrectangle S of P let f_S be the constant function $m_S(f)$. Then

$$\begin{aligned} L(f,P) &= \sum_S m_S(f) \cdot v(S) = \sum_S \int_{\text{int } S} f_S \\ &= \sum_S \int_{g^{-1}(\text{int } S)} (f_S \circ g) |\det g' \leq \sum_S \int_{g^{-1}(\text{int } S)} (f \circ g) |\det g'| \\ &\leq \int_{g^{-1}(V)} (f \circ g) |\det g'|. \end{aligned}$$

Since $\int_V f$ is the least upper bound of all $L(f,P)$, this proves that $\int_V f \leq \int_{g^{-1}(V)} (f \circ g) |\det g'|$. A similar argument, letting $f_S = M_S(f)$, shows that $\int_V f \geq \int_{g^{-1}(V)} (f \circ g) |\det g'|$. The result now follows from the above Remark.

3. If the theorem is true for $g\colon A \to \mathbf{R}^n$ and for $h\colon B \to \mathbf{R}^n$, where $g(A) \subset B$, then it is true for $h \circ g\colon A \to \mathbf{R}^n$.

Proof of (3).

$$\begin{aligned}\int_{h\circ g(A)} f &= \int_{h(g(A))} f = \int_{g(A)} (f\circ h)\left|\det h'\right| \\ &= \int_A [(f\circ h)\circ g]\cdot[\left|\det h'\right|\circ g]\cdot\left|\det g'\right| \\ &= \int_A f\circ(h\circ g)\left|\det (h\circ g)'\right|.\end{aligned}$$

4. The theorem is true if g is a linear transformation.

Proof of (4). By (1) and (2) it suffices to show for any open rectangle U that

$$\int_{g(U)} 1 = \int_U \left|\det g'\right|.$$

This is Problem 3-35.

Observations (3) and (4) together show that we may assume for any particular $a \in A$ that $g'(a)$ is the identity matrix: in fact, if T is the linear transformation $Dg(a)$, then $(T^{-1}\circ g)'(a) = I$; since the theorem is true for T, if it is true for $T^{-1}\circ g$ it will be true for g.

We are now prepared to give the proof, which preceeds by induction on n. The remarks before the statement of the theorem, together with (1) and (2), prove the case $n = 1$. Assuming the theorem in dimension $n - 1$, we prove it in dimension n. For each $a \in A$ we need only find an open set U with $a \in U \subset A$ for which the theorem is true. Moreover we may assume that $g'(a) = I$.

Define $h\colon A \to \mathbf{R}^n$ by $h(x) = (g^1(x), \ldots, g^{n-1}(x), x^n)$. Then $h'(a) = I$. Hence in some open U' with $a \in U' \subset A$, the function h is 1-1 and $\det h'(x) \neq 0$. We can thus define $k\colon h(U') \to \mathbf{R}^n$ by $k(x) = (x^1, \ldots, x^{n-1}, g^n(h^{-1}(x)))$ and $g = k\circ h$. We have thus expressed g as the composition

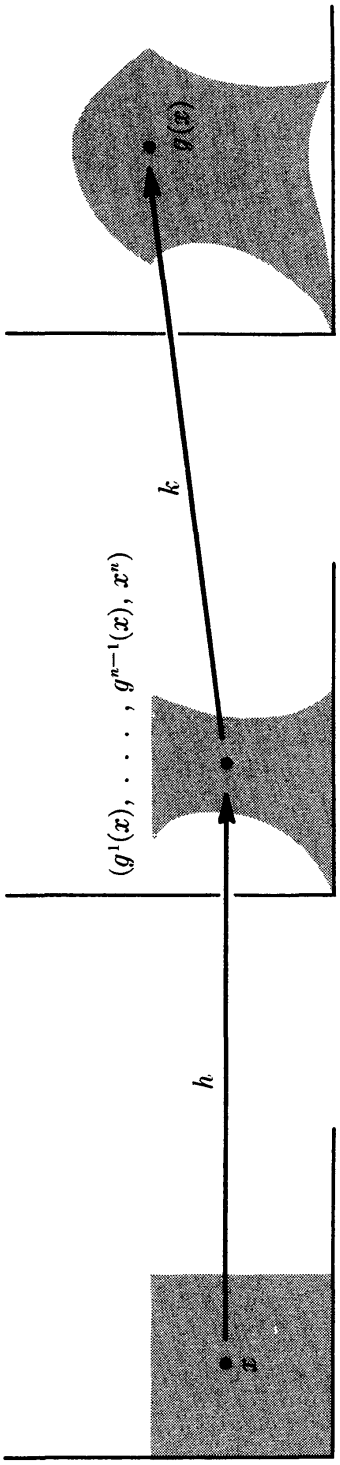


FIGURE 3-3

of two maps, each of which changes fewer than n coordinates (Figure 3-3).

We must attend to a few details to ensure that k is a function of the proper sort. Since

$$(g^n \circ h^{-1})'(h(a)) = (g^n)'(a) \cdot [h'(a)]^{-1} = (g^n)'(a),$$

we have $D_n(g^n \circ h^{-1})(h(a)) = D_n g^n(a) = 1$, so that $k'(h(a)) = I$. Thus in some open set V with $h(a) \in V \subset h(U')$, the function k is 1-1 and $\det k'(x) \neq 0$. Letting $U = k^{-1}(V)$ we now have $g = k \circ h$, where $h\colon U \to \mathbf{R}^n$ and $k\colon V \to \mathbf{R}^n$ and $h(U) \subset V$. By (3) it suffices to prove the theorem for h and k. We give the proof for h; the proof for k is similar and easier.

Let $W \subset U$ be a rectangle of the form $D \times [a_n,b_n]$, where D is a rectangle in $\mathbf{R}^{n-1}$. By Fubini's theorem

$$\int_{h(W)} 1 = \int_{[a_n,b_n]} \left(\int_{h(D \times \{x^n\})} 1 \, dx^1 \cdots dx^{n-1} \right) dx^n.$$

Let $h_{x^n}\colon D \to \mathbf{R}^{n-1}$ be defined by $h_{x^n}(x^1, \ldots, x^{n-1}) = (g^1(x^1, \ldots, x^n), \ldots, g^{n-1}(x^1, \ldots, x^n))$. Then each h_{x^n} is clearly 1-1 and

$$\det (h_{x^n})'(x^1, \ldots, x^{n-1}) = \det h'(x^1, \ldots, x^n) \neq 0.$$

Moreover

$$\int_{h(D \times \{x^n\})} 1 \, dx^1 \cdots dx^{n-1} = \int_{h_{x^n}(D)} 1 \, dx^1 \cdots dx^{n-1}.$$

Applying the theorem in the case $n - 1$ therefore gives

$$\begin{aligned}
\int_{h(W)} 1 &= \int_{[a_n,b_n]} \left(\int_{h_{x^n}(D)} 1 \, dx^1 \cdots dx^{n-1} \right) dx^n \\
&= \int_{[a_n,b_n]} \left(\int_D \left|\det (h_{x^n})'(x^1, \ldots, x^{n-1})\right| dx^1 \cdots dx^{n-1} \right) dx^n \\
&= \int_{[a_n,b_n]} \left(\int_D \left|\det h'(x^1, \ldots, x^n)\right| dx^1 \cdots dx^{n-1} \right) dx^n \\
&= \int_W |\det h'|. \quad \blacksquare
\end{aligned}$$

The condition $\det g'(x) \neq 0$ may be eliminated from the

hypotheses of Theorem 3-13 by using the following theorem, which often plays an unexpected role.

3-14. Theorem (Sard's Theorem). *Let $g: A \to \mathbf{R}^n$ be continuously differentiable, where $A \subset \mathbf{R}^n$ is open, and let $B = \{x \in A: \det g'(x) = 0\}$. Then $g(B)$ has measure 0.*

Proof. Let $U \subset A$ be a closed rectangle such that all sides of U have length l, say. Let $\varepsilon > 0$. If N is sufficiently large and U is divided into N^n rectangles, with sides of length l/N, then for each of these rectangles S, if $x \in S$ we have

$$|Dg(x)(y - x) - g(y) - g(x)| < \varepsilon|x - y| \leq \varepsilon\sqrt{n}\,(l/N)$$

for all $y \in S$. If S intersects B we can choose $x \in S \cap B$; since $\det g'(x) = 0$, the set $\{Dg(x)(y - x): y \in S\}$ lies in an $(n - 1)$-dimensional subspace V of $\mathbf{R}^n$. Therefore the set $\{g(y) - g(x): y \in S\}$ lies within $\varepsilon\sqrt{n}\,(l/N)$ of V, so that $\{g(y): y \in S\}$ lies within $\varepsilon\sqrt{n}\,(l/N)$ of the $(n - 1)$-plane $V + g(x)$. On the other hand, by Lemma 2-10 there is a number M such that

$$|g(x) - g(y)| < M|x - y| \leq M\sqrt{n}\,(l/N).$$

Thus, if S intersects B, the set $\{g(y): y \in S\}$ is contained in a cylinder whose height is $< 2\varepsilon\sqrt{n}\,(l/N)$ and whose base is an $(n - 1)$-dimensional sphere of radius $< M\sqrt{n}\,(l/N)$. This cylinder has volume $< C(l/N)^n\varepsilon$ for some constant C. There are at most N^n such rectangles S, so $g(U \cap B)$ lies in a set of volume $< C(l/N)^n \cdot \varepsilon \cdot N^n = Cl^n \cdot \varepsilon$. Since this is true for all $\varepsilon > 0$, the set $g(U \cap B)$ has measure 0. Since (Problem 3-13) we can cover all of A with a sequence of such rectangles U, the desired result follows from Theorem 3-4. ▮

Theorem 3-14 is actually only the easy part of Sard's Theorem. The statement and proof of the deeper result will be found in [17], page 47.

Problems. **3-39.** Use Theorem 3-14 to prove Theorem 3-13 without the assumption $\det g'(x) \neq 0$.

3-40. If $g: \mathbf{R}^n \to \mathbf{R}^n$ and $\det g'(x) \neq 0$, prove that in some open set containing x we can write $g = T \circ g_n \circ \cdots \circ g_1$, where g_i is of the form $g_i(x) = (x^1, \ldots, f_i(x), \ldots, x^n)$, and T is a linear transformation. Show that we can write $g = g_n \circ \cdots \circ g_1$ if and only if $g'(x)$ is a diagonal matrix.

3-41. Define $f: \{r: r > 0\} \times (0,2\pi) \to \mathbf{R}^2$ by $f(r,\theta) = (r \cos \theta, r \sin \theta)$.

(a) Show that f is 1-1, compute $f'(r,\theta)$, and show that $\det f'(r,\theta) \neq 0$ for all (r,θ). Show that $f(\{r: r > 0\} \times (0,2\pi))$ is the set A of Problem 2-23.

(b) If $P = f^{-1}$, show that $P(x,y) = (r(x,y),\theta(x,y))$, where

$$r(x,y) = \sqrt{x^2 + y^2},$$

$$\theta(x,y) = \begin{cases} \arctan y/x & x > 0,\ y > 0, \\ \pi + \arctan y/x & x < 0, \\ 2\pi + \arctan y/x & x > 0,\ y < 0, \\ \pi/2 & x = 0,\ y > 0, \\ 3\pi/2 & x = 0,\ y < 0. \end{cases}$$

(Here arctan denotes the inverse of the function tan: $(-\pi/2,\pi/2) \to \mathbf{R}$.) Find $P'(x,y)$. The function P is called the **polar coordinate system** on A.

(c) Let $C \subset A$ be the region between the circles of radii r_1 and r_2 and the half-lines through 0 which make angles of θ_1 and θ_2 with the x-axis. If $h: C \to \mathbf{R}$ is integrable and $h(x,y) = g(r(x,y),\theta(x,y))$, show that

$$\int_C h = \int_{r_1}^{r_2} \int_{\theta_1}^{\theta_2} rg(r,\theta)\,d\theta\,dr.$$

If $B_r = \{(x,y): x^2 + y^2 \leq r^2\}$, show that

$$\int_{B_r} h = \int_0^r \int_0^{2\pi} rg(r,\theta)\,d\theta\,dr.$$

(d) If $C_r = [-r,r] \times [-r,r]$, show that

$$\int_{B_r} e^{-(x^2+y^2)}\,dx\,dy = \pi(1 - e^{-r^2})$$

and

$$\int_{C_r} e^{-(x^2+y^2)}\,dx\,dy = \left(\int_{-r}^{r} e^{-x^2}\,dx\right)^2.$$

(e) Prove that

$$\lim_{r\to\infty} \int_{B_r} e^{-(x^2+y^2)}\,dx\,dy = \lim_{r\to\infty} \int_{C_r} e^{-(x^2+y^2)}\,dx\,dy$$

and conclude that

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

"A mathematician is one to whom *that* is as obvious as that twice two makes four is to you. Liouville was a mathematician."

—Lord Kelvin

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER III

SELF LEARNING MATERIAL

PAPER : MATC 3.3
(Pure & Applied Streams)

Block - I : Fuzzy Set Theory
Block - II : Computer Programming in 'C' (Theory)

Directorate of Open and Distance Learning
University of Kalyani
Kalyani, Nadia
West Bengal, India

Course Preparation Team

1. Mr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani	2. Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

November, 2019

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

Director's Massage

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Sankar Kumar Ghosh, Hon'ble Vice-Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as weel as each and every member of PGBOS (DODL), University of Kalyani, Heartfelt thanks is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

CONTENTS

Core Paper

MATC 3.3
Block - I
Marks : 26 (SSE : 20; IA : 06)

Fuzzy Set Theory

Syllabus

• Unit 1 •

Interval Arithmetic: Interval numbers, arithmetic operations on interval numbers, distance between intervals, two level interval numbers

• Unit 2 •

Basic concepts of fuzzy sets: Types of fuzzy sets, -cuts and its properties, representations of fuzzy sets, decomposition theorems, support, convexity, normality, cardinality, standard set-theoretic operations on fuzzy sets, Zadeh's extension principle.

• Unit 3 •

Fuzzy Relations: Crisp versus fuzzy relations, fuzzy matrices and fuzzy graphs, composition of fuzzy relations, relational join, binary fuzzy relations.

• Unit 4 •

Fuzzy Arithmetic: Fuzzy numbers, arithmetic operations on fuzzy numbers (multiplication and division on $\mathbb{R}^+$ only), fuzzy equations.

Unit 1

Course Structure

- Interval Arithmetic: Interval numbers, arithmetic operations on interval numbers,
- Distance between intervals, two level interval numbers

---

1 Introduction

Interval arithmetic is the arithmetic of quantities that lie within specified ranges (i.e., intervals) instead of having definite known values. Interval arithmetic can be especially useful when working with data that is subject to measurement errors or uncertainties. It can be considered a rigorous version of significance arithmetic (a.k.a., automatic precision control).

Interval arithmetic, interval mathematics, interval analysis, or interval computation, is a method developed by mathematicians since the 1950s and 1960s, as an approach to putting bounds on rounding errors and measurement errors in mathematical computation and thus developing numerical methods that yield reliable results. Very simply put, it represents each value as a range of possibilities. For example, instead of estimating the height of someone using standard arithmetic as 2.0 metres, using interval arithmetic we might be certain that that person is somewhere between 1.97 and 2.03 metres.

This concept is suitable for a variety of purposes. The most common use is to keep track of and handle rounding errors directly during the calculation and of uncertainties in the knowledge of the exact values of physical and technical parameters. The latter often arise from measurement errors and tolerances for components or due to limits on computational accuracy. Interval arithmetic also helps find reliable and guaranteed solutions to equations (such as differential equations) and optimization problems.

Mathematically, instead of working with an uncertain real x we work with the two ends of the interval $[a, b]$ that contains x. In interval arithmetic, any variable x lies between a and b, or could be one of them. A function f when applied to x is also uncertain. In interval arithmetic f produces an interval $[c, d]$ that is all the possible values for $f(x)$ for all $x \in [a, b]$.

Objectives

After reading this unit you will be able to

- define interval numbers
- define set operations on intervals numbers and see certain examples related to them
- define arithmetic operations on intervals numbers and see certain examples related to them
- define algebraic properties of interval numbers
- define distance between intervals

1.1 Interval Number System

We are familiar with the closed intervals in the real line, which is denoted by

$$[a, b] = \{x \in \mathbb{R} : \ a \leq x \leq b\}.$$

Here, we will mainly refer to the closed intervals as intervals.

We will denote the endpoints of an interval I as $\underline{I}$ and $\overline{I}$, where these both represent the lower and upper endpoints respectively, that is,

$$I = [\underline{I}, \overline{I}]$$

and two intervals I and J are said to be equal if they are the same sets, that is

$$I = J \ \ \& \ \ \underline{I} = \underline{J}, \ \ \overline{I} = \overline{J}.$$

We say that an interval I is *degenerate* if $\underline{I} = \overline{I}$. Such an interval contains a single real number x. By convention, we agree to identify a degenerate interval $[x, x]$ with the real number x.

1.1.1 Certain Important Definitions

The intersection of two intervals I and J is empty if either $\overline{J} < \underline{I}$ or $\overline{I} < \underline{J}$. In this case, we let $\emptyset$ denote the empty set and write

$$I \cap J = \emptyset,$$

which indicates that I and J have no points in common. We may otherwise define the intersection $I \cap J$ as the interval

$$\begin{aligned} I \cap J &= \{z : \ z \in I \ \ \& \ \ z \in J\} \\ &= [\max\{\underline{I}, \underline{J}\}, \min\{\overline{I}, \overline{J}\}]. \end{aligned}$$

In this latter case, the union of I and J is also an interval

$$\begin{aligned} I \cup J &= \{z : \ z \in I \ \text{ or } \ z \in J\} \\ &= [\min\{\underline{I}, \underline{J}\}, \max\{\overline{I}, \overline{J}\}]. \end{aligned}$$

In general, the union of two intervals is not an interval. However, the interval hull of two intervals, defined by

$$I \underline{\cup} J = [\min\{\underline{I}, \underline{J}\}, \max\{\overline{I}, \overline{J}\}],$$

is always an interval and can be used in interval computations. We have

$$I \cup J \subseteq I \underline{\cup} J,$$

for any two intervals I and J.

Example 1.1. If $I = [-1, 0]$ and $J = [1, 2]$, then $I \underline{\cup} J = [-1, 2]$. $I \cup J$ is a disconnected set and hence is not an interval. But this is not the case if we consider $I \underline{\cup} J$ and $I \cup J$ is still a subset of $I \underline{\cup} J$.

Intersection plays a key role in interval analysis. If we have two intervals containing a result of interest — regardless of how they were obtained — then the intersection, which may be narrower, also contains the result.

Example 1.2. Suppose two people make independent measurements of the same physical quantity q. One finds that $q = 10.3$ with a measurement error less than 0.2. The other finds that $q = 10.4$ with an error less than 0.2. We can represent these measurements as the intervals $I = [10.1, 10.5]$ and $J = [10.2, 10.6]$, respectively. Since q lies in both, it also lies in $I \cup J = [10.2, 10.5]$. An empty intersection would imply that at least one of the measurements is wrong.

Definition 1.3. 1. As the name suggests, the **width** of an interval I is defined as

$$w(I) = \overline{I} - \underline{I}.$$

2. The **absolute value** of I, denoted as $|I|$, is the maximum of the absolute values of its endpoints

$$|I| = \max\{|\underline{I}|, |\overline{I}|\}.$$

Note that, $|x| \leq |I|$ for every $x \in I$.

3. The **midpoint** of I is given by

$$m(I) = \frac{1}{2}(\underline{I} + \overline{I}).$$

Example 1.4. Let $I = [0, 2]$ and $J = [-1, 1]$. Then the intersection and union of I and J are the intervals

$$I \cap J = [0, 1], \quad I \cup J = [-1, 2].$$

We have, $w(I) = w(J) = 2$ and

$$|I| = 2, \quad \& \quad |J| = 1.$$

The midpoint of I and J are 1 and 0 respectively.

The real numbers are ordered by the relation $<$. A corresponding relation can be defined for the intervals as follows

$$I < J \implies \overline{I} < \underline{J}.$$

For example, $[3, 4] < [6, 8]$ and we also have the transitivity relation which says that

$$A < B \ \& \ B < C \implies A < C.$$

We can also define $I > 0$ and $I < 0$. That is, $I > 0$ if $x > 0$ for all $x \in I$ and $I < 0$ if $x < 0$ for all $x \in I$.

We can also define another relation on the set of intervals as the set inclusion relation which says that

$$I \subseteq J \quad \text{iff} \quad \underline{J} \leq \underline{I} \ \& \ \overline{I} \leq \overline{J}.$$

For example, $[1, 2] \subseteq [0, 2]$. This is a partial ordering. This has to be noted that not every pair of intervals is comparable under this relation.

The notion of the degenerate interval permits us to regard the system of closed intervals as an extension of the real number system. Indeed, there is an obvious one-to-one pairing $[x, x] \mapsto x$ between the elements of the two systems. We will next investigate into the arithmetic operations of the intervals.

1.2 Arithmetic Operations on Intervals

We are about to define the basic arithmetic operations between intervals. The key point in these definitions is that computing with intervals is computing with sets. For example, when we add two intervals, the resulting interval is a set containing the sums of all pairs of numbers, one from each of the two initial sets. By definition then, the sum of two intervals I and J is

$$I + J = \{i + j : \ i \in I \ \& \ j \in J\}.$$

We will return to an operational description of addition momentarily (that is, to the task of obtaining a formula by which addition can be easily carried out). But let us define the remaining three arithmetic operations. The difference of two intervals I and J is the set

$$I - J = \{i - j : \ i \in I \ \& \ j \in J\}.$$

The product of I and J is given by

$$I.J = \{ij : \ i \in I \ \& \ j \in J\}.$$

Finally the quotient I/J is defined as

$$I/J = \{i/j : \ i \in I \ \& \ j \in J\}.$$

provided that $0 \notin J$.

We have seen the purpose of introducing the interval number system. So it is redundant to talk about arithmetic operations in terms of the terms in the interval. So, we will find a way to write it in terms of intervals.

1. Addition : Since $i \in I$ and $j \in J$ implies that

   $$\underline{I} \leq i \leq \overline{I} \quad \& \quad \underline{J} \leq j \leq \overline{J},$$

   we see by addition of inequalities that the sum $i + j \in I + J$ must satisfy

   $$\underline{I} + \underline{J} \leq i + j \leq \overline{I} + \overline{J}.$$

   Hence the formula

   $$I + J = [\underline{I} + \underline{J}, \overline{I} + \overline{J}].$$

   Example 1.5. Let $I = [0, 2]$ and $J = [-1, 2]$. Then

   $$I + J = [-1, 3].$$

   This is not the same as $I \cup J = [-1, 2]$

2. Subtraction : Since $i \in I$ and $j \in J$ implies that

   $$\underline{I} \leq i \leq \overline{I} \quad \& \quad -\overline{J} \leq -j \leq -\underline{J},$$

   gives

   $$\underline{I} - \overline{J} \leq i - j \leq \overline{I} - \underline{J}.$$

   It follows that

   $$I - J = [\underline{I} - \overline{J}, \overline{I} - \underline{J}].$$

Note that

$$I - J = I + (-J),$$

where, $-J$ is defined as

$$-J = [-\overline{J}, -\underline{J}] = \{y : \quad -y \in Y\}.$$

Note the reversal of endpoints that occurs when we find the negative of an interval.

Example 1.6. If $I = [-1, 0]$ and $J = [1, 2]$, then

$$-J = [-2, -1], \quad \& \quad I - J = [-3, -1].$$

What happens for $I - I$? Is it necessary that $I - I = 0$ as in the case of any real number? Consider $I = [2, 3]$. Then, as we have seen the definition of interval subtraction,

$$I - I = [2 - 3, 3 - 2] = [-1, 1].$$

In fact, for any interval $I = [\underline{I}, \overline{J}]$, we have

$$I - I = [\underline{I} - \overline{I}, \overline{I} - \underline{I}]$$

which is equal to 0 if and only if I is a degenerate interval.

3. Multiplication : The multiplication of intervals is given in terms of the minimum and maximum of four products of endpoints. Actually, by testing for the signs of the endpoints $\underline{I}$, $\overline{I}$, $\underline{J}$, $\overline{J}$. The formula for the endpoints of the interval product can be broken into nine special cases. In eight of these, only two products need be computed.

Exercise 1.7. 1. Find $I \cap J$ and $I \cup J$ for the following intervals

   (a) $I = [3, 4]$ and $J = [5, 7]$

   (b) $I = [1, 2]$ and $J = [0, 3]$

   (c) $I = [1, 4]$ and $J = [2, 6]$

2. Find $I + J$ and $I \cup J$ if $I = [5, 7]$ and $J = [-2, 6]$.

3. Find $I - J$ if $I = [5, 6]$ and $J = [-2, 4]$.

1.3 Algebraic Properties of Interval Numbers

We will now study certain algebraic properties related to the interval numbers as follows.

1. **Commutative and Associative Properties:** It is easy to show that the interval addition and multiplication are commutative and associative. That is, for any three intervals I, J, K,

$$I + J = J + I, \qquad I + (J + K) = (I + J) + K,$$
$$IJ = JI, \qquad I(JK) = (IJ)K.$$

2. **Additive and Multiplicative elements:** The degenerate intervals 0 and 1 are additive and multiplicative identity elements in the system of intervals

$$0 + I = 0 + I = I, \quad 1.I = I.1 = I, \quad 0.I = I.0 = 0$$

for any interval I.

3. **Nonexistence of Inverse Elements:** We note that $-I$ is not an additive inverse for I. We have

$$I + (-I) = [\underline{I}, \overline{I}] + [-\overline{I}, -\underline{I}] = [\underline{I} - \overline{I}, \overline{I} - \underline{I}],$$

and this is zero only if $\underline{I} = \overline{I}$. If I does not have zero width, then

$$I - I = w(I)[-1, 1].$$

Similarly, $I/I = 1$ only if $w(I) = 0$. In general,

$$\begin{aligned} I/I &= [\underline{I}/\overline{I}, \overline{I}/\underline{I}]; \ 0 < \underline{I}, \\ &= [\overline{I}/\underline{I}, \underline{I}/\overline{I}]; \ \overline{I} < 0. \end{aligned}$$

We don't have additional additive or multiplicative inverses except for degenerate intervals. However, we always have the inclusions $0 \in I - I$ and $1 \in I/I$.

4. **Subdistributivity:** The distributive law

$$x(y + z) = xy + xz$$

of ordinary arithmetic also fails to hold for intervals. An easy counterexample can be obtained by taking $I = [1, 2], \ J = [1, 2], \ K = [-1, 1]$ which gives

$$I(J + K) = [1, 2].([1, 1] - [1, 1]) = [1, 2].[0, 0].$$

Also,

$$IJ + IK = [1, 2].[1, 1] - [1, 2].[1, 1] = [-1, 1].$$

However, the subdistributive law says that

$$I(J + K) \subseteq IJ + IK.$$

We can see this in the example above. Full distributivity does hold in certain special cases. In particular, for any real number x we have

$$x(J + K) = xJ + xK.$$

Interval multiplication can be distributed over a sum of intervals as long as those intervals have the same sign:

$$I(J + K) \subseteq IJ + IK, \ \text{ provided that } \ JK > 0.$$

5. **Cancellation Law:** The cancellation law

$$I + K = J + K \implies I = J$$

holds for interval addition.

We should emphasize that, with the identification of degenerate intervals and real numbers, interval arithmetic is an extension of real arithmetic. It reduces to ordinary real arithmetic for intervals of zero width.

Exercise 1.8. 1. Verify the distributive law for the intervals $I = [1, 2], \ J = [-3, -2], \ K = [-5, -1]$.

2. Prove the Cancellation law. Show that multiplicative cancellation does not hold in interval arithmetic, that is, $IK = JK$ does not imply $I = J$.

1.3.1 Symmetric Intervals

An interval I is said to be symmetric if $\underline{I} = -\overline{I}$. For example, $[-1, 1]$ is symmetric and $[-1, 5]$ is not. Any symmetric interval has midpoint 0. If I is symmetric, then

$$|I| = \frac{1}{2}w(I), \qquad I = |I|[-1, 1].$$

The rules of interval arithmetic are slightly simpler when symmetric intervals are involved. If I, J, K are all symmetric, then

$$\begin{aligned} I + J = I - J &= (|I| + |J|)[-1, 1], \\ IJ &= |I||J|[-1, 1], \\ I(J \pm K) = IJ + JK &= |I|(|J| + |K|)[-1, 1]. \end{aligned}$$

If J is symmetric and I is any interval, then

$$IJ = |I|J.$$

It follows that if J and K are symmetric, then

$$I(J + K) = IJ + IK$$

for any interval I.

1.3.2 Inclusion Isotonicity of Interval Arithmetic

Let $\odot$ stand for interval addition, subtraction, multiplication, or division. If A, B, C and D are intervals such that

$$A \subseteq C \quad \text{and} \quad B \subseteq D,$$

then

$$A \odot B \subseteq C \odot D.$$

These relations follow directly from the definitions given previously. Interval arithmetic is said to be inclusion isotonic. We will now extend the concept of interval expressions to include functions such as $\sin x$ and e^x.

1.4 Interval Functions

Let f be a real-valued function of a single real variable x. Ultimately, we would like to know the precise range of values taken by $f(x)$ as x varies through a given interval I. In other words, we would like to be able to find the image of the set I under the mapping f, which is, $f(I) = \{f(x) : \ x \in I\}$. More generally, given a function $f = f(x_1, \ldots, x_n)$ of several variables, we will wish to find the image set

$$f(I_1, \ldots, I_n) = \{f(x_1, \ldots, x_n) : \ x_1 \in I_1, \ldots, x_n \in I_n\}$$

where $I_1, \ldots, I_n$ are specified intervals.

Definition 1.9. Let $g : M_1 \to M_2$ be a mapping between sets M_1 and M_2, and denote by $S(M_1)$ and $S(M_2)$ the families of subsets of M_1 and M_2, respectively. The united extension of g is the set-valued mapping $\overline{g} : S(M_1) \to S(M_2)$ such that

$$\overline{g}(I) = \{g(x) : \ x \in I, \ I \in S(M_1)\}.$$

The mapping $\overline{g}$ is sometimes of interest as a single-valued mapping on $S(M_1)$ with values in $S(M_2)$. For our purposes, however, it is merely necessary to note that

$$\overline{g}(I) = \cup_{x \in I}\{g(x)\},$$

that is, $\overline{g}(I)$ contains precisely the same elements as the set image $g(I)$. For this reason, and because the usage is common, we shall apply the term united extension to set images such as those described previously.

1.4.1 Elementary Functions of Interval Arguments

For some functions, the image set is easy to compute. For example, consider $f(x) = x^2$, $x \in \mathbb{R}$. If $I = [\underline{I}, \overline{I}]$, it is evident that the set

$$f(I) = \{x^2 : \ x \in I\}$$

can be expressed as

$$\begin{aligned} f(I) &= [\underline{I}^2, \overline{I}^2], \quad 0 \leq \underline{I} \leq \overline{I}, \\ &= [\overline{I}^2, \underline{I}^2], \quad \underline{I} \leq \overline{I} \leq 0, \\ &= [0, \max\{\underline{I}^2, \overline{I}^2\}], \quad \underline{I} < 0 < \overline{I}. \end{aligned}$$

Note that I^2 is not the same as $I.I$. For example

$$[-1, 1]^2 = [0, 1], \qquad [-1, 1].[-1, 1] = [-1, 1].$$

We will use the definition of I^2 for $f(I)$. However, $[-1, 1]$ does contain $[0, 1]$. The overestimation when we compute a bound on the range of I^2 as $I.I$ is due to the phenomenon of interval dependency. Namely, if we assume x is an unknown number known to lie in the interval I, then, when we form the product $x.x$, the x in the second factor, although known only to lie in I must be the same as the x in the first factor, whereas, in the definition of the interval product $I.I$, it is assumed that the values in the first factor and the values in the second factor vary independently.

Interval dependency is a crucial consideration when using interval computations. It is a major reason why simply replacing floating point computations by intervals in an existing algorithm is not likely to lead to satisfactory results.

The reasoning is particularly straightforward with functions f (x) that happen to be monotonic, i.e., either increasing or decreasing with increasing x. Note that, an increasing function f maps an interval $I = [\underline{I}, \overline{I}]$ into the interval $f(I) = [f(\underline{I}), f(\overline{I})]$.

1.4.2 Interval-Valued Extensions of Real Functions

Let us begin with an example. Consider the real-valued function f given by $f(x) = 1 - x$, $x \in \mathbb{R}$. Note carefully that a function is defined by two things: (1) a domain over which it acts, and (2) a rule that specifies how elements of that domain are mapped under the function. Both of these are specified in the definition of f. The elements of Dom$f$ are real numbers x, and the mapping rule is $x \mapsto 1 - x$. Taken in isolation, the entity $f(x) = 1 - x$ is a formula—not a function. Often this distinction is ignored; in many elementary math books, for example, we would interpret the entity as a function whose domain should be taken as the largest possible set over which the formula makes sense (in this case, all of $\mathbb{R}$). However, we will understand that Dom$f$ is just as essential to the definition of f as is the formula $f(x)$.

Now suppose we take the formula that describes the given function f and apply it to interval arguments. The resulting interval-valued function

$$F(I) = 1 - I, \quad I = [\underline{I}, \overline{I}],$$

is an extension of the function f. we have enlarged the domain to include nondegenerate intervals I as well as the degenerate intervals $x = [x, x]$.

Definition 1.10. We say that F is an interval extension of f , if for degenerate interval arguments, F agrees with f, that is, $F([x, x]) = f(x)$.

Let us compare $F(I)$ with the set image $f(I)$. We have according to the laws of interval arithmetic,

$$F(I) = [1,1] - [\underline{I}, \overline{I}] = [1,1] + [-\overline{I}, -\underline{I}] = [1 - \overline{I}, 1 - \underline{I}].$$

On the other hand, as x increases through the interval $[\underline{I}, \overline{I}]$, the the value of $f(x)$ given by $1 - x$ decreases from $1 - \overline{I}$ to $1 - \underline{I}$. So by definition, $f(I) = [1 - \overline{I}, 1 - \underline{I}]$. In this example, we have $F(I) = f(I)$; this particular extension of f obtained by the formula $f(x) = 1 - x$ directly to interval arguments, yields the desired set image $f(I)$. In other words, we have found the united extension of f, which is, $f(I) = 1 - I$. Although the situation is not always so simple, but we will leave it for the time being and move on to the definition of distance between intervals.

1.5 Distance between Intervals

We are very much accustomed with the idea of metric and the basic point set theory, the convergence, completeness, etc. We will now attempt to define metric for the interval numbers.

Definition 1.11. If I and J are two intervals, then the distance between them is defined by

$$d(I, J) = \max\{|\underline{I} - \underline{J}|, |\overline{I} - \overline{J}|\}.$$

We can define the concepts of convergence, continuity with the help of the above definition.

Definition 1.12. Let $\{I_k\}$ be a sequence of intervals. We say that it converges if there exists an interval I^* such that for every $\epsilon > 0$, there is a natural number $N = N(\epsilon)$ such that $d(I_k, I^*) < \epsilon$ whenever $k > N$. As in the case of real sequences, we write

$$I^* = \lim_{k \to \infty} I_k.$$

We know that the interval number system represents an extension of the real number system. In fact, the correspondence $[x, x] \leftrightarrow x$ can be regarded as a function or mapping between the two systems. This mapping preserves distances between corresponding objects. We have

$$d([x, x], [y, y]) = \max\{|x - y|, |x - y|\} = |x - y|$$

for any real x and y. For this reason, it is called an isometry, and we say that the real line is "isometrically embedded" in the metric space of interval numbers.

Exercise 1.13.

1. Show that the definition of distance given between two intervals satisfy the metric axioms.
2. Find the distance between the intervals $I = [1, 2]$ and $J = [3, 5]$.
3. For any intervals I, J, K prove that
   (a) $d(I + K, Y + K) = d(I, J)$;
   (b) $d(I, J) \leq w(J)$ when $I \subseteq J$;
   (c) $d(I, 0) = |I|$.

1.6 Few Probable Questions

1. Define symmetric interval. Show that any interval I can be expressed as the sum of a real number (i.e., degenerate interval) and a symmetric interval:

$$I = m + W, \quad \text{where} \quad m = m(I) \quad \text{and} \quad W = \frac{1}{2}w(I)[-1, 1].$$

2. Show that $I_k \to I$ if and only if $\underline{I}_k \to \underline{I}$ and $\overline{I}_k \to \overline{I}$.

Unit 2

Course Structure

- Types of fuzzy sets, cuts and its properties,
- Representations of fuzzy sets, decomposition theorems, support,
- Convexity, normality, cardinality, standard set-theoretic operations on fuzzy sets,
- Zadeh's extension principle.

---

2 Introduction

In mathematics, fuzzy sets (also known as uncertain sets) are somewhat like sets whose elements have degrees of membership. Fuzzy sets were introduced independently by Lotfi A. Zadeh and Dieter Klaua in 1965 as an extension of the classical notion of set. At the same time, Salii (1965) defined a more general kind of structure called an L-relation, which he studied in an abstract algebraic context. Fuzzy relations, which are used now in different areas, such as linguistics (De Cock, Bodenhofer & Kerre 2000), decision-making (Kuzmin 1982), and clustering (Bezdek 1978), are special cases of L-relations when L is the unit interval $[0, 1]$.

In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition — an element either belongs or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described with the aid of a membership function valued in the real unit interval $[0, 1]$. Fuzzy sets generalize classical sets, since the indicator functions of classical sets are special cases of the membership functions of fuzzy sets, if the latter only take values 0 or 1. In fuzzy set theory, classical bivalent sets are usually called crisp sets. The fuzzy set theory can be used in a wide range of domains in which information is incomplete or imprecise, such as bioinformatics.

Objectives

After reading this unit, you will be able to

- define fuzzy sets and its types
- define a-cuts of fuzzy sets and related properties
- learn various representations of fuzzy sets
- deduce the decomposition theorems of fuzzy sets
- define the set theoretic operations on fuzzy sets and see various related examples
- get an idea of the extension principle

2.1 Fuzzy Sets

A classical (crisp) set is normally defined as a collection of elements or objects $x \in X$ that can be finite, countable, or uncountable. Each single element can either belong to or not belong to a set A, $A \subseteq X$. In the former case, the statement "x belongs to A" is true, whereas in the latter case this statement is false.

Such a classical set can be described in different ways: one can either enumerate (list) the elements that belong to the set; describe the set analytically, for instance, by stating conditions for membership ($A = \{x : \ x \leq 5\}$); or define the member elements by using the characteristic function, in which 1 indicates membership and 0 nonmembership. For a fuzzy set, the characteristic function allows various degrees of membership for the elements of a given set.

Definition 2.1. If X is a collection of objects denoted generically by x, then a fuzzy set $\tilde{A}$ in X is a set of ordered pairs

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) : \ x \in X\}.$$

$\mu_{\tilde{A}}(x)$ is called the membership function or grade of membership (also degree of compatibility or degree of truth) of $x \in \tilde{A}$ that maps X to the membership space M (When M contains only the two points 0 and 1, A is nonfuzzy and $\mu_{\tilde{A}}(x)$ is identical to the characteristic function of a nonfuzzy set). The range of the membership function is a subset of the nonnegative real numbers whose supremum is finite. Elements with a zero degree of membership are normally not listed. The set X is called the universal set and let us denote the set of all fuzzy sets on X by $\mathscr{F}(X)$.

Fuzzy sets are represented in different ways.

1. A fuzzy set is denoted by an ordered set of pairs, the first element of which denotes the element and the second the degree of membership.

   Example 2.2. A realtor wants to classify the house he offers to his clients. One indicator of comfort of these houses is the number of bedrooms in it. Let $X = \{1, 2, \ldots, 10\}$ be the set of available types of houses described by x =number of bedrooms in a house. Then the fuzzy set "comfortable type of house for a four-person family" may be described as

   $$\tilde{A} = \{(1, 0.2), (2, 0.5), (3, 0.8), (4, 1), (5, 0.7), (6, 0.3)\}.$$

   Example 2.3. Let $\tilde{A}$ = real numbers "considerably" larger than 10. Then in this case, the numbers less than or equal to 10 automatically falls out and we must define $\mu_{\tilde{A}}(x)$ in such a way that as x goes farther away from 10, the membership function increases. We define $\mu_{\tilde{A}}(x)$ as

   $$\begin{aligned} \mu_{\tilde{A}}(x) &= 0, \quad x \leq 10 \\ &= \frac{1}{1 + \frac{1}{(x-10)^2}}, \ \ x > 10 \end{aligned}$$

   Example 2.4. Let $\tilde{A}$ = real numbers close to 10. Then

   $$\tilde{A} = \left\{ (x, \mu_{\tilde{A}}(x)) : \ \mu_{\tilde{A}}(x) = \frac{1}{1 + (x - 100)} \right\}.$$

   If we plot the graph of the membership function against the members set elements, then we will get somewhat as given in the figure.

2. A fuzzy set is represented can be sometimes solely by stating its membership function.

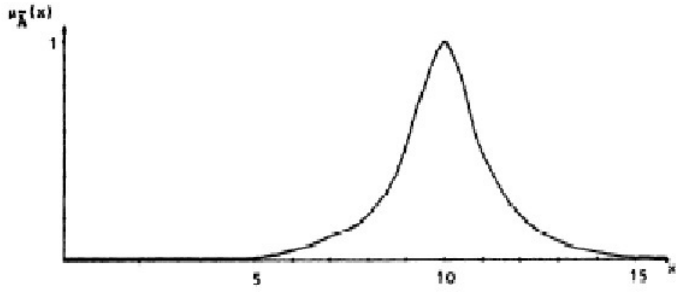


Figure 1: Real numbers close to 10

3.

$$\tilde{A} = \mu_{\tilde{A}}(x_1)/x_1 + \mu_{\tilde{A}}(x_2)/x_2 + \cdots = \sum_{i=1}^{n} \mu_{\tilde{A}}(x_i)/x_i$$

$$\text{or} \quad \int_x \mu_{\tilde{A}}(x)/x.$$

Example 2.5. If $\tilde{A}$ =integers close to 10, then

$$\tilde{A} = 0.1/7 + 0.5/8 + 0.8/9 + 1/10 + 0.8/11 + 0.5/12 + 0.1/13.$$

Also, if $\tilde{A}$ =real numbers close to 10, then

$$\tilde{A} = \int_{\mathbb{R}} \frac{1}{1 + (x-10)^2}/x.$$

It has already been mentioned that the membership function is not limited to values between 0 and 1.

Definition 2.6. A fuzzy set $\tilde{A}$ is called normal if $\sup_x \mu_{\tilde{A}}(x) = 1$.

A non-empty fuzzy set $\tilde{A}$ can always be normalized by dividing $\mu_{\tilde{A}}(x)$ by $\sup_x \mu_{\tilde{A}}(x)$. For convenience, we will consider only normal fuzzy sets. For the representation of fuzzy sets, we will use the notation 1.

A fuzzy set is obviously a generalization of a classical set and the membership function a generalization of the characteristic function. Since we are generally referring to a universal (crisp) set X, some elements of a fuzzy set may have the degree of membership zero. Often it is appropriate to consider those elements of the universe that have a nonzero degree of membership in a fuzzy set.

Definition 2.7. The support of a fuzzy set $\tilde{A}$, $S(\tilde{A})$, is the crisp set of all $x \in X$ such that $\mu_{\tilde{A}}(x) > 0$.

Example 2.8. For example (2.2), the support of $\tilde{A}$ is $S(\tilde{A}) = \{1, 2, 3, 4, 5, 6\}$. The elements $\{7, 8, 9, 10\}$ are not part of the support of $\tilde{A}$.

A more general and even more useful notion is that of an a-level set.

Definition 2.9. The (crisp) set of elements that belong to the fuzzy set $\tilde{A}$ at least to the degree a is called the a-level set or a-cut

$$A_a = \{x \in X : \mu_{\tilde{A}}(x) \geq a\}$$

$A'_a = \{x \in X : \mu_{\tilde{A}}(x) > a\}$ is called strong a-level set or strong a-cut.

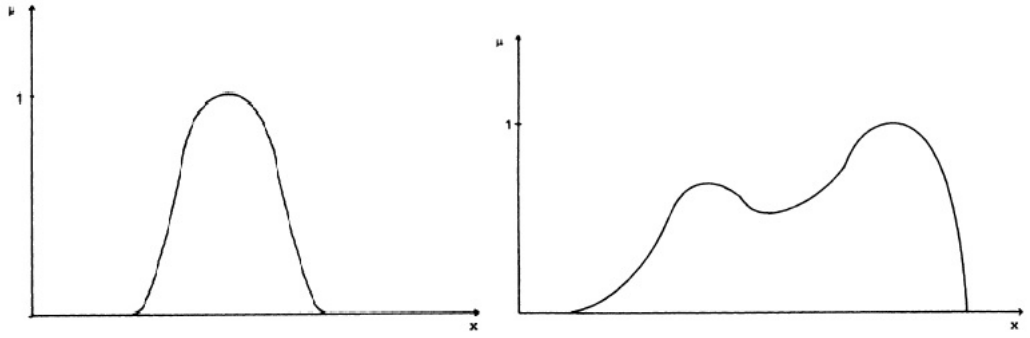


Figure 2: Convex and Non-convex set

Example 2.10. Again we refer to the example (2.2). We list a possible a-level sets.

$$\begin{aligned} A_{0.2} &= \{1,2,3,4,5,6\} \\ A_{0.5} &= \{2,3,4,5\} \\ A_{0.8} &= \{3,4\} \\ A_1 &= \{4\}. \end{aligned}$$

The strong 0.8-level set is $A'_{0.8} = \{4\}$.

Convexity also plays a role in fuzzy set theory. By contrast to classical set theory, however, convexity conditions are defined with reference to the membership function rather than the support of the fuzzy set.

Definition 2.11. A fuzzy set $\tilde{A}$ is convex if

$$\mu_{\tilde{A}}(cx + (1-c)y) \geq \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(y)\}, \;\; x, y \in X, \;\; c \in [0,1].$$

Alternatively, a fuzzy set is convex if all a -level sets are convex. In the figure given above, the set on the right is convex and that on the left is not.

Definition 2.12. For a fuzzy set $\tilde{A}$, the cardinality $|\tilde{A}|$ is defined as

$$|\tilde{A}| = \sum_{x \in X} \mu_{\tilde{A}}(x),$$

and

$$||\tilde{A}|| = \frac{|\tilde{A}|}{|x|}$$

is called the relative cardinality of $\tilde{A}$.

Obviously, the relative cardinality of a fuzzy set depends on the cardinality of the universe . So you have to choose the same universe if you want to compare fuzzy sets by their relative cardinality.

Example 2.13. For the fuzzy set "comfortable type of house for a four-person family" from (2.2), the cardinality is

$$|\tilde{A}| = 0.2 + 0.5 + 0.8 + 1 + 0.7 + 0.3 = 3.5.$$

Its relative cardinality is

$$||\tilde{A}|| = \frac{3.5}{10} = 0.35$$

The relative cardinality can be interpreted as the fraction of elements of X being in $\tilde{A}$, weighted by their degrees of membership in $\tilde{A}$. For infinite X, the cardinality is defined by $|\tilde{A}| = \int_x \mu_{\tilde{A}}(x)dx$. Of course, $|\tilde{A}|$ does not always exist.

2.2 Basic Set-Theoretic Operations for Fuzzy Sets

The membership function is obviously the crucial component of a fuzzy set. It is therefore not surprising that operations with fuzzy sets are defined via their membership function s. We shall first present the concepts suggested by Zadeh in 1965. They constitute a consistent framework for the theory of fuzzy sets. They are, however, not the only possible way to extend classical set theory consistently.

Definition 2.14. The membership function $\mu_{\tilde{C}}(x)$ of the intersection $\tilde{C} = \tilde{A} \cap \tilde{B}$ is pointwise defined by

$$\mu_{\tilde{C}}(x) = \min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \;\; x \in X$$

Definition 2.15. The membership function $\mu_{\tilde{D}}(x)$ of the intersection $\tilde{D} = \tilde{A} \cup \tilde{B}$ is pointwise defined by

$$\mu_{\tilde{D}}(x) = \max\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \;\; x \in X$$

Theorem 2.16. Let $\tilde{A}$ and $\tilde{B}$ be be two fuzzy sets on a universal set X. Then for all $a, b \in [0, 1]$,

1. ${A_a}' \subseteq A_a$;
2. $a \leq b$ implies that $A_b \subseteq A_a$ and $A_b^{'} \subseteq A_a^{'}$;
3. $(A \cap B)_a = A_a \cap B_a$ and $(A \cup B)_a = A_a \cup B_a$;
4. $(A \cap B)_a^{'} = A_a^{'} \cap B_a^{'}$ and $(A \cup B)_a^{'} = A_a^{'} \cup B_a^{'}$.

Proof. 1. By definition, $A_a^{'} = \{x \in X : \; \mu_{\tilde{A}}(x) > a\} \subseteq \{x \in X : \; \mu_{\tilde{A}}(x) \geq a\} = A_a$.

2. Let $a \leq b$. Then, $A_b = \{x \in X : \; \mu_{\tilde{A}}(x) \geq b\} \subseteq \{x \in X : \; \mu_{\tilde{A}}(x) \geq a\} = A_a$. We can similarly show the result for the strong cuts.

3. For $x \in (A \cap B)_a$, we have, $\mu_{\tilde{A} \cap \tilde{B}}(x) \geq a$ and hence $\min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \geq a$. This means that $\mu_{\tilde{A}}(x) \geq a$ and $\mu_{\tilde{B}}(x) \geq a$ and hence $x \in (A_a \cap B_a)$ and hence $(A \cap B)_a \subseteq A_a \cap B_a$. Conversely, for any $x \in A_a \cap B_a$, we have $x \in A_a$ and $x \in B_a$, that is, $\mu_{\tilde{A}}(x) \geq a$ and $\mu_{\tilde{B}}(x) \geq a$. Hence, $\min\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \geq a$ which means that $\mu_{A \tilde{\cap} B}(x) \geq a$. Hence, $x \in (A \cap B)_a$ and consequently, we have $(A \cap B)_a \supseteq A_a \cap B_a$. Thus, we have $(A \cap B)_a = A_a \cap B_a$.

   For the second equality, let $x \in (A \cup B)_a$, we have, $\max\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \geq a$ and hence, $\mu_{\tilde{A}}(x) \geq a$ and $\mu_{\tilde{B}}(x) \geq a$. This implies that $x \in A_a \cup B_a$ and thus $(A \cup B)_a \subseteq (A_a \cup B_a)$. Conversely, for any $x \in A_a \cup B_a$, we have, $x \in A_a$ and $x \in B_a$; that is, $\mu_{\tilde{A}}(x) \geq a$ or $\mu_{\tilde{B}}(x) \geq a$. Hence $\max\{\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\} \geq a$, which means that $\mu_{A \tilde{\cup} B}(x) \geq a$. This means that $x \in (A \cup B)_a$ and hence, $A_a \cup B_a \subseteq (A \cup B)_a$. Hence the result.

4. Left as an exercise.

□

Let us examine the significance of the properties stated in the previous theorem. Property 1 is trivial, expressing that the strong a-cut is always included in the a-cut of any fuzzy set and for any $a \in [0, 1]$; the property follows directly from the definitions of the two types of a-cuts. Property 2 means that the set sequences $\{A_a : \; a \in [0, 1]\}$ and $\{A_a^{'} : \; a \in [0, 1]\}$ of a-cuts and strong a-cuts, respectively are always

monotonic decreasing with respect to a; consequently, they are nested families of sets. Properties 3 and 4 show that the standard fuzzy intersection and fuzzy union are both cutworthy and strong cutworthy when applied to two fuzzy sets or, due to the associativity -of min and max, to any finite number of fuzzy sets.

Theorem 2.17. Let A^i be fuzzy sets over the universal set X for all $i \in I$, where I is an index set. Then,

1. $\bigcup_{i\in I} A^i_a \subseteq \left(\bigcup_{i\in I} A^i\right)_a$ and $\bigcap_{i\in I} A^i_a \subseteq \left(\bigcap_{i\in I} A^i\right)_a$;
2. $\bigcup_{i\in I} A^{i'}_a \subseteq \left(\bigcup_{i\in I} A^i\right)'_a$ and $\bigcap_{i\in I} A^{i'}_a \subseteq \left(\bigcap_{i\in I} A^i\right)'_a$

Proof. 1. Left for the reader.

2. For all $x \in X$,
$$x \in \bigcup_{i\in I} A^{i'}_a$$
if and only if there exists some $i_0 \in I$ such that $x \in A^{i'_0}_a$ (that is, $\mu_{A^{i_0}}(x) > a$). This inequality is satisfied iff
$$\sup_{i\in I} \mu_{A^i}(x) > a,$$
which is equivalent to
$$\mu_{\bigcup_{i\in I} A^i}(x) > a.$$
That is,
$$x \in \left(\bigcup_{i\in I} A^i\right)'_a.$$
Hence the equality in 2 is satisfied.

We now prove the second proposition in 2. For all
$$x \in \left(\bigcap_{i\in I} A^i\right)'_a,$$
we have
$$\mu_{\bigcap_{i\in I} A^i}(x) > a;$$
that is,
$$\inf_{i\in I} \mu_{A^i}(x) > a.$$
Hence, for any $i \in I$, $\mu_{A^i}(x) > a$ which means that $x \in A^{i'}_a$. Hence
$$x \in \bigcap_{i\in I} A^{i'}_a,$$
which concludes the proof.

□

The inequalities in the above theorem can't be replaced by equalities.

Example 2.18. Consider the fuzzy set A^i in the universal set X defined as

$$\mu_{A^i}(x) = 1 - \frac{1}{i}$$

for all $x \in X$ and $i \in \mathbb{N}$. Then for any $x \in X$,

$$\mu_{\bigcup\limits_i A^i}(x) = \sup_i \mu_{A^i}(x) = \sup_i \left(1 - \frac{1}{i}\right) = 1.$$

Let $a = 1$. Then

$$\left(\bigcup_i A^i\right)_1 = X.$$

However, for any $i \in \mathbb{N}$, $A^i_1 = \emptyset$ because, for any $x \in X$,

$$\mu_{A^i}(x) = 1 - \frac{1}{i} < 1.$$

Hence

$$\bigcup_i A^i_1 = \bigcup_i \emptyset = \emptyset \neq X = \left(\bigcup_i A^i\right)_1.$$

This shows that equality is not possible always in case of property 1 of the above theorem. A similar example can be used to show the same for property 2.

Theorem 2.19. Let A and B be two fuzzy sets in the universal set X. Then for all $a \in [0, 1]$,

1. $A \subseteq B$ iff $A_a \subseteq B_a$ and $A \subseteq B$ iff $A'_a \subseteq B'_a$;
2. $A = B$ iff $A_a = B_a$ and $A = B$ iff $A'_a = B'_a$

Proof. 1. To prove the first proposition, we assume that there exists $a_0 \in [0, 1]$ such that $A_{a_0} \not\subseteq B_{a_0}$, that is, there exists $x_0 \in X$ such that $x_0 \in A_{a_0}$ but $x_0 \notin B_{a_0}$. Then, $\mu_A(x_0) \geq a_0$ and $\mu_B(x_0) < a_0$. Hence, $\mu_B(x_0) < \mu_A(x_0)$, which contradicts that $A \subseteq B$. Now assume that $A \not\subseteq B$; that is, there exists $x_0 \in X$ such that $\mu_B(x_0) < \mu_A(x_0)$. Let $a = \mu_A(x_0)$. Then $x_0 \in A_a$ and $x_0 \notin B_a$, which demonstrates that $A_a \subseteq B_a$ is not satisfied for all $a \in [0, 1]$.

Now we prove the second proposition. The first part is similar to the previous proof. For the second part, assume that $A \not\subseteq B$. Then there exists $x_0 \in X$ such that $\mu_A(x_0) > \mu_B(x_0)$. Let a be any number between $\mu_A(x_0)$ and $\mu_B(x_0)$. Then $x_0 \in A'_a$ $x_0 \notin B'_a$. Hence $A'_a \not\subseteq B'_a$, which demonstrates that $A'_a \subseteq B'_a$ is not satisfied for all $a \in [0, 1]$.

2. Left as exercise.

□

The above theorem establishes that the properties of fuzzy set inclusion and equality are both cutworthy and strong cutworthy.

Theorem 2.20. For any fuzzy set A in the universal set X, the following properties hold

1. $A_a = \bigcap\limits_{b<a} A_b = \bigcap\limits_{b<a} A'_b$;
2. $A'_a = \bigcup\limits_{a<b} A_b = \bigcup\limits_{a<b} A'_b$.

Proof. 1. For any $b < a$, we clearly have $A_a \subseteq A_b$. Hence

$$A_a \subseteq \bigcup_{b<a} A_b.$$

Now, for all $x \bigcap_{b<a} A_b$ and for any $\epsilon > 0$, we have $x \in A_{a-\epsilon}$ (since $a - \epsilon < a$), which means that $\mu_A(x) \geq a - \epsilon$. Since ϵ is an arbitrary number, let $\epsilon \to 0$. This results in $\mu_A(x) \geq a$ (that is, $x \in A_a$). Hence,

$$\bigcap_{b<a} A_b \subseteq A_a,$$

which concludes the proof of the first equation. The proof of the second equation is analogous.

2. Left as exercise.

□

We now convert each of the a-cuts into a special fuzzy set ${}_aA$, defined for $x \in X$ as

$$\mu_{{}_aA}(x) = a.\mu_{A_a}(x).$$

Theorem 2.21. (First Decomposition Theorem). For every fuzzy set A in the universal set X,

$$A = \bigcup_{a\in[0,1]} {}_aA,$$

where the symbols have their usual meaning.

Proof. For each particular $x \in X$, let $a = \mu_A(x)$. Then,

$$\begin{aligned} \mu_{\bigcup_{a\in[0,1]} {}_aA}(x) &= \sup_{a\in[0,1]} \mu_{{}_aA}(x) \\ &= \max\{\sup_{a\in[0,\alpha]} \mu_{{}_aA}(x), \sup_{a\in(\alpha,1]} \mu_{{}_aA}(x)\}. \end{aligned}$$

Foe each $a \in (\alpha, 1]$, we have $\mu_A(x) = \alpha < a$ and hence, $\mu_{{}_aA}(x) = 0$. On the other hand, for each $a \in [0, \alpha]$, we have $\mu_A(x) = \alpha \geq a$, therefore, $\mu_{{}_aA}(x) = a$. Hence

$$\mu_{\bigcup_{a\in[0,1]} {}_aA}(x) = \sup_{a\in[0,\alpha]} a = \alpha = \mu_A(x).$$

Since the same argument is valid for each $x \in X$, the validity of the theorem is established. □

Theorem 2.22. (Second Decomposition Theorem). For any fuzzy set A in X, we have

$$A = \bigcup_{a\in[0,1]} {}_aA',$$

where ${}_aA'$ denotes a special fuzzy set defined by

$$\mu_{{}_aA'}(x) = a.\mu_{A'_a}(x)$$

where, $\bigcup$ denotes the standard fuzzy union.

Proof. Since the proof is analogous to the proof of the First Decomposition theorem, we express it in a more concise form. For each particular $x \in X$, let $\alpha = \mu_A(x)$. Then,

$$\begin{aligned} \mu_{\bigcup_{a\in[0,1]} {}_aA'}(x) &= \sup_{a\in[0,1]} \mu_{{}_aA'}(x) \\ &= \max\{\sup_{a\in[0,\alpha)} \mu_{{}_aA'}(x), \sup_{a\in[\alpha,1]} \mu_{{}_aA'}(x)\} \\ &= \sup_{a\in[0,\alpha)} a = \alpha = \mu_A(x). \end{aligned}$$

□

Definition 2.23. The set of all levels $a \in [0, 1]$ that represent distinct a-cuts of a given fuzzy set A is called a level set of A. Formally,

$$\Lambda(A) = \{a : \ \mu_A(x) = a \ \text{ for some } \ x \in X\},$$

where Λ denotes the level set of fuzzy set A defined on X.

Theorem 2.24. (Third Decomposition Theorem). For every fuzzy set A in the universal set X,

$$A = \bigcup_{a\in\Lambda(A)} {}_aA,$$

where $\Lambda(A)$ is the level set of A.

Proof. Analogous to the proofs of the other decomposition theorems. □

Let us see some other definitions related to fuzzy sets. We will then see the Extension Principle for fuzzy sets.

Definition 2.25. The membership function of the complement of a normalized fuzzy set $\tilde{A}$, $\mu_C\tilde{A}(x)$ is defined by

$$\mu_C\tilde{A}(x) = 1 - \mu_{\tilde{A}}(x), \ \ x \in X.$$

Example 2.26. Let $\tilde{A}$ be the fuzzy set in the example (2.2) and $\tilde{B}$ be the fuzzy set "large type of house" defined as

$$\tilde{B} = \{(3, 0.2), (4, 0.4), (5, 0.6), (6, 0.8), (7, 1), (8, 1)\}$$

The intersection $\tilde{C} = \tilde{A} \cap \tilde{B}$ is then

$$\tilde{C} = \{(3, 0.2), (4, 0.4), (5, 0.6), (6, 0.3)\}$$

and the union $\tilde{D} = \tilde{A} \cup \tilde{B}$

$$\tilde{D} = \{(l, 0.2), (2, 0.5), (3, 0.8), (4, 1), (5, 0.7), (6, 0.8), (7, 1), (8, 1)\}$$

The complement $C\tilde{B}$, which might be interpreted as "not large type of house," is

$$C\tilde{B} = \{(1, 1), (2, 1), (3, 0.8), (4, 0.6), (5, 0.4), (6, 0.2), (9, 1), (l0, 1)\}.$$

It has already been mentioned that min and max are not the only operators that could have been chosen to model the intersection or union, respectively, of fuzzy sets. The question arises, why those and not others? Bellman and Giertz addressed this question axiomatically in 1973. They argued from a logical point of view, interpreting the intersection as "logical and," the union as "logical or," and the fuzzy set $\tilde{A}$ as the statement

"The element x belongs to the set $\tilde{A}$" which can be accepted as more or less true. It is very instructive to follow their line of argument, which is an excellent example for an axiomatic justification of specific mathematical models. We shall therefore sketch their reasoning: Consider two statements, S and T, for which the truth values are μ_S and μ_T respectively, where $\mu_S, \mu_T \in [0,1]$. The truth value of the "and" and "or" combination of these statements, $\mu(S \text{ and } T)$ and $\mu(S \text{ or } T)$, both from the interval $[0,1]$, are interpreted as the values of the membership functions of the intersection and union , respectively, of S and T. We are now looking for two real-valued functions f and g such that

$$\begin{aligned} \mu_{S \text{ and } T} &= f(\mu_S, \mu_T) \\ \mu_{S \text{ or } T} &= g(\mu_S, \mu_T). \end{aligned}$$

Bellman and Giertz feel that the following restrictions are reasonably imposed on f and g:

1. f and g are nondecreasing and continuous in μ_S and μ_T.
2. f and g are symmetric, that is,

$$\begin{aligned} f(\mu_S, \mu_T) &= f(\mu_T, \mu_S) \\ g(\mu_S, \mu_T) &= g(\mu_T, \mu_S). \end{aligned}$$

3. $f(\mu_S, \mu_S)$ and $g(\mu_S, \mu_S)$ are strictly increasing in μ_S.
4. $f(\mu_S, \mu_T) \leq \min(\mu_S, \mu_T)$ and $g(\mu_S, \mu_T) \geq \max(\mu_S, \mu_T)$. This implies that accepting the truth of the statement "S and T" requires more, and accepting the truth of the statement "S or T" less than accepting S or T alone as true.
5. $f(1,1) = 1$ and $g(0,0) = 0$.
6. Logically equivalent statements must have equal truth values, and fuzzy sets with the same contents must have the same membership functions, that is,

$$S_1 \text{ and } (S_2 \text{ or } S_3)$$

is equivalent to

$$(S_1 \text{ and } S_2) \text{ or } (S_1 \text{ and } S_3)$$

and therefore must be equally true.

Bellman and Giertz now formalize the above assumptions as follows : Using the symbols $\wedge$ for "and" and $\vee$ for "or", these assumptions amount to the following seven restrictions, to be imposed on the two commutative and associative binary compositions $\wedge$ and $\vee$ on the closed interval $[0,1]$, which distributive with respect to one another.

1. $\mu_S \wedge \mu_T = \mu_T \wedge \mu_S$ and $\mu_S \vee \mu_T = \mu_T \vee \mu_S$.
2. $(\mu_S \wedge \mu_T) \wedge \mu_U = \mu_S \wedge (\mu_T \wedge \mu_U)$ and $(\mu_S \vee \mu_T) \vee \mu_U = \mu_S \vee (\mu_T \vee \mu_U)$.
3. $\mu_S \wedge (\mu_T \vee \mu_U) = (\mu_S \wedge \mu_T) \vee (\mu_S \wedge \mu_U)$ and $\mu_S \vee (\mu_T \wedge \mu_U) = (\mu_S \vee \mu_T) \wedge (\mu_S \vee \mu_U)$.
4. $\mu_S \wedge \mu_T$ and $\mu_S \vee \mu_T$ are continuous and nondecreasing in each component.
5. $\mu_S \wedge \mu_T$ and $\mu_S \vee \mu_T$ are are strictly increasing in μ_S.
6. $\mu_S \wedge \mu_T \leq \min(\mu_S, \mu_T)$ and $\mu_S \vee \mu_T \leq \max(\mu_S, \mu_T)$.

7. $1 \wedge 1 = 1$ and $0 \vee 0 = 0$.

Bellman and Giertz then prove mathematically that $\mu_{S \wedge T} = \min(\mu_S, \mu_T)$ and $\mu_{S \vee T} = \max(\mu_S, \mu_T)$.

For the complement, it would be reasonable to assume that if statement "S" is true, its complement "non S" is false, or if $\mu_S = 1$, then $\mu_{\text{non } S} = 0$ and vice versa.

Exercise 2.27. 1. Model the following expressions as fuzzy sets :

(a) Very small numbers.

(b) Numbers approximately between 10 and 20.

2. Determine all a-level sets and all strong a-level sets for the following fuzzy set

$$\begin{aligned} \tilde{A} = \{(x, \mu_{\tilde{C}}(x) : \ x \in R\} & \\ \text{where } \ \mu_{\tilde{C}}(x) &= 0 \ \text{ for } \ x \leq 10 \\ &= \frac{1}{1 + (x-10)^{-2}}, \ \text{ for } \ x > 10. \end{aligned}$$

3. Let $X = \{1, \ldots, 10\}$. Determine the cardinalities and relative cardinalities of the following fuzzy sets:

(a) $\tilde{B} = \{(2, 0.4), (3, 0.6), (4, 0.8), (5, 1), (6, 0.8), (7, 0.6), (8, 0.4)\}$.

(b) $\tilde{C} = \{(2, 0.4), (4, 0.8), (5, 1), (7, 0.6)\}$.

2.3 Types of Fuzzy Sets

So far we have considered fuzzy sets with crisply defined membership functions or degrees of membership. It is doubtful whether, for instance, human beings have or can have a crisp image of membership functions in their minds. Zadeh therefore suggested the notion of a fuzzy set whose membership function itself is a fuzzy set. If we call fuzzy sets, such as those considered so far, type 1 fuzzy sets, then a type 2 fuzzy set can be defined as follows.

Definition 2.28. A type 2 fuzzy set is a fuzzy set whose membership values are type 1 fuzzy sets on $[0, 1]$.

The operations intersection, union, and complement defined so far are no longer adequate for type 2 fuzzy sets.

Definition 2.29. A type m fuzzy set is a fuzzy set in X whose membership values are type $m-1$, $m > 1$ fuzzy sets on $[0, 1]$.

From a practical point of view, such type m fuzzy sets for large m (even for $m \geq 3$) are hard to deal with, and it will be extremely difficult or even impossible to measure them or to visualize them. We will, therefore, not even try to define the usual operations on them.

There are certain other types of sets. A definition was given by Hirota which is given below.

Definition 2.30. A probabilistic set A on X is defined by a defining function μ_A,

$$\mu_A : X \times \Omega \ \text{ defined as } \ (x, \omega) \mapsto \mu_A(x, \omega) \in \Omega_C$$

where $\mu_A(x, \dot{)}$ is the (B, B_C)-measurable function for each fixed $x \in X$.

For Hirota, a probabilistic set A with the defining function $\mu_A(x,\omega)$ is contained in a probabilistic set B with $\mu_B(x,\omega)$ if for each $x \in X$ there exists an $E \in B$ such that $P(E) = 1$ and $\mu_A(x,\omega) \leq \mu_B(x,\omega)$ for all $\omega \in E$. (Ω, B, P) is called the parameter space.

Further attempts at representing vague and uncertain data with different types of fuzzy sets were made by Atanassov and Stoeva and by Pawlak which are given below.

Definition 2.31. Given an underlying set X of objects, an intuitonistic fuzzy set (IFS) A is a set of ordered triples,

$$A = \{(x, \mu_A(x), \nu_A(x)) : \ x \in X\}$$

where $\mu_A(x)$ and $\nu_A(x)$ are functions mapping from X into $[0,1]$. For each $x \in X$, $\mu_A(x)$ represents the degree of membership of the element x to the subset A of X, and $\nu_A(x)$ gives the degree of nonmembership. For the functions $\mu_A(x)$ and $\nu_A(x)$ mapping into $[0,1]$, the condition $0 \leq \mu_A(x) + \nu_A(x) \leq 1$ holds.

Ordinary fuzzy sets over X may be viewed as special intuitonistic fuzzy sets with the nonmembership function $\nu_A(x) = 1 - \mu_A(x)$.

Definition 2.32. Let U denote a set of objects called universe and let $R \subset U \times U$ be an equivalence relation on U. The pair $A = (U, R)$ is called an approximation space. For $u, v \in U$ and $(u, v) \in R$, u and v belong to the same equivalence class, and we say that they are indistinguishable in A. Hence the relation R is called an indiscernibility relation. Let $[x]_R$ denote an equivalence class (elementary set of A) R containing element x; then the lower and upper approximations for a subset $X \subseteq U$ in A-denoted by $\underline{A}(X)$ and $\overline{A}(X)$ respectively, are defined as follows

$$\underline{A}(X) = \{x \in U : \ [x]_R \subset X\} \quad \text{and} \quad \overline{A}(X) = \{x \in U : \ [x]_R \cap X \neq \theta\}.$$

If an object x belongs to the lower approximation space of X in A, then "x surely belongs to X in A," $x \in \overline{A}(X)$ means that "x possibly belongs to X in A."

For the subset $X \subseteq U$ representing a concept of interest, the approximation space $A = (U, R)$ can be characterized by three distinct regions of X in A: the so-called positive region $\underline{A}(X)$, the boundary region $\overline{A}(X) - \underline{A}(X)$, and the negative region $U - \overline{A}(X)$.

The characterization of objects in X by the indiscernibility relation R is not precise enough if the boundary region $\overline{A}(X) - \underline{A}(X)$ is not empty. For this case it may be impossible to say whether an object belongs to X or not, and so the set X is said to be nondefinable in A, and X is a rough set.

2.4 Extension Principle for Fuzzy Sets

We say that a crisp function

$$f : X \to Y$$

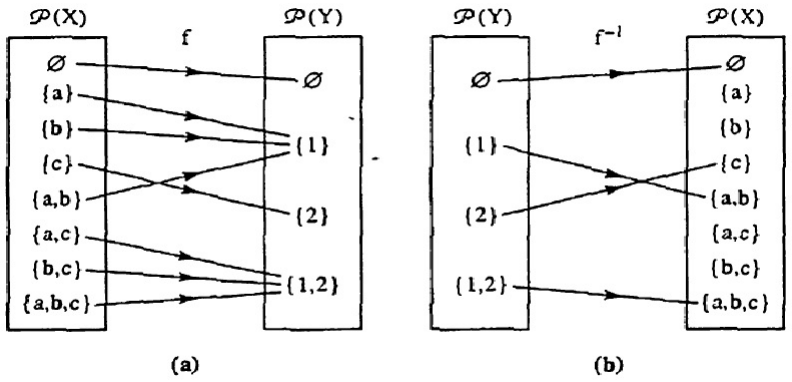
is fuzzified when it is extended to act on fuzzy sets defined on X and Y. That is, the fuzzified function, for which the same symbol f is usually used, has the form

$$f : \mathscr{F}(X) \to \mathscr{F}(Y),$$

and its inverse function, f^{-1} has the form

$$f^{-1} : \mathscr{F}(Y) \to \mathscr{F}(X).$$

A principle for fuzzifying crisp functions (or, possibly, crisp relations) is called an extension principle. Before introducing this principle, let us first discuss its special case, in which the extended functions are restricted to crisp power sets $\mathcal{P}(X)$ and $\mathcal{P}(Y)$. This special case is well established in classical set theory.



Given a crisp function from X to Y, its extended version is a function from $\mathcal{P}(X)$ to $\mathcal{P}(Y)$ that, for any $A \in \mathcal{P}(X)$, is defined by

$$f(A) = \{y : \ y = f(x), \ x \in A\}.$$

Furthermore, the extended version of the inverse of f, denoted by f^{-1}, is a function from $\mathcal{P}(Y)$ to $\mathcal{P}(X)$ that, for any $B \in \mathcal{P}(Y)$, is defined by

$$f^{-1}(B) = \{x : \ \ f(x) \in B\}.$$

Expressing the sets $f(A)$ and $f^{-1}(B)$ by their characteristic functions (viewed here as special cases of membership functions), we obtain

$$\begin{aligned} [f(A)](y) &= \sup_{x|y=f(x)} 1_A(x), \\ \{f^{-1}(B)\}(x) &= 1_B(f(x)). \end{aligned}$$

As a simple example illustrating the meaning of these equations, let $X = \{a, b, c\}$ and $Y = \{1, 2\}$, and let us consider the function

$$\begin{aligned} f : \ a &\rightarrow 1 \\ b &\rightarrow 1 \\ c &\rightarrow 2 \end{aligned}$$

When applying the last two equations to this function, we obtain the extension of f shown in the figure. Allowing now sets A and B in the above equations to be fuzzy sets and replacing the characteristic functions in these equations with membership functions, we arrive at the following extension principle by which any crisp function can be fuzzified.

Extension Principle. Any given function $f : X \rightarrow Y$ induces two functions,

$$\begin{aligned} f : \mathscr{F}(X) &\rightarrow \mathscr{F}(Y), \\ f^{-1} : \mathscr{F}(Y) &\rightarrow \mathscr{F}(X), \end{aligned}$$

which are defined by

$$[f(A)](y) = \sup_{x|y=f(x)} \mu_A(x), \ \ \forall A \in \mathscr{F}(X),$$

and

$$\{f^{-1}(B)\}(x) = \mu_B(f(x)), \ \ \forall B \in \mathscr{F}(Y).$$

2.5 Few Probable Questions

1. Define the a-cut of a fuzzy set. Prove that for all $a \in [0,1]$, $A_a^{'} \subseteq A_a$.
2. Define strong a-cut of a fuzzy set. Show that $(A \cap B)_a^{'} = A_a^{'} \cap B_a^{'}$ for every $a \in [0,1]$.
3. Define the union of two fuzzy sets. Show that $(A \cup B)_a^{'} = A_a^{'} \cup B_a^{'}$ for every $a \in [0,1]$.
4. Show that for any collection of fuzzy sets A^i over a universal set X, where i belongs to the index set I, we have
$$\bigcup_{i \in I} A_a^i \subseteq \left(\bigcup_{i \in I} A^i \right)_a .$$
Can the inequality be replaced by equality? Justify.
5. State and prove the first decomposition theorem.
6. State and prove the second decomposition theorem.
7. Define the level set for a fuzzy set A. State and prove the third decomposition theorem.

Unit 3

In 1965, L. A. Zadeh introduced the concept of fuzzy set theory. Fuzzy set theory is an extension of classical set theory. A logic that is not very precise is called a fuzzy logic. The imprecise way of looking at things and manipulating them is much more powerful than precise way of looking at them and then manipulating them. Fuzzy logic is one of the tools for making computer system capable of solving problems involving imprecision. Fuzzy logic is an attempt to capture imprecision by generalizing the concept of set to fuzzy set.

In every day content most of the problems involve imprecise concept. To handle the imprecise concept, the conventional method of set theory and numbers are insufficient and need to be extended to some other concepts. Fuzzy concept is one of the concepts for this purpose.

A relation is a mathematical description of a situation where certain elements of sets are related to one another in some way. Fuzzy relations are significant concepts in fuzzy theory and have been widely used in many fields such as fuzzy clustering, fuzzy control and uncertainty reasoning. They also play an important role in fuzzy diagnosis and fuzzy modeling. When fuzzy relations are used in practice, how to estimate and compare them is a significant problem. Uncertainty measurements of fuzzy relations have been done by some researchers. Similarity measurement of uncertainty was introduced by Yager who also discussed its application.

1.2 Crisp Relation

To describe the fuzzy relation, first we describe relation by an example of daily life using discrete fuzzy sets. Relationship is described between the colours of a fruit X and the grade of maturity Y. Crisp set X with three linguistic terms is given as

$$X = \{\text{green, yellow, red}\}$$

Similarly the grade of maturity for the other set Y will be

$$Y = \{\text{verdant, half-mature, mature}\}$$

Crisp formulation of a relation $X \rightarrow Y$ between two crisp sets is presented in tabular form

	Verdant	Half-mature	Mature
Green	1	0	0
Yellow	0	1	0
Red	0	0	1

In the above table "0" and "1" describe the grade of membership to this relation. This relation is a new kind of crisp set that is built from the two crisp base set X and Y. This new set is now called R and can be expressed by the rules

1. If the colour of the fruit is green then the fruit is verdant.
2. If the colour of the fruit is yellow then the fruit is half-mature.
3. If the colour of the fruit is red then the fruit is mature.

This crisp relation shows the existence or absence of connection, relations or interconnection between two sets. Now we show the membership grades represented in the fuzzy relation.

	Verdant	Half-mature	Mature
Green	1	0.6	0
Yellow	0.4	1	0.3
Red	0	0.5	1

The table above represents the fuzzy relation.

Crisp relation is defined on the Cartesian product of two universal sets determined as

$$X \times Y = \{(x, y) \mid x \in X, y \in Y\}$$

The crisp relation R is defined by its membership function

$$\mu_R(x, y) = \begin{cases} 1, & (x, y) \in R \\ 0, & (x, y) \notin R \end{cases}$$

Here "1" implies complete truth degree for the pair to be in relation and "0" implies no relation.

When the sets are finite the relation is represented by a matrix R called a relation matrix.

1.2.1 Example

Let $X = \{1, 4, 5\}$ and $Y = \{3, 6, 7\}$

Classical matrix for the crisp relation when $R = x < y$ is

$$R = \begin{array}{c} \\ 1 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccc} 3 & 6 & 7 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \end{array}$$

1.2.2 Example

Let $A = \{2, 4, 6, 8\}$ and $B = \{2, 4, 6, 8\}$

Classical matrix for the crisp relation $R = x = y$

$$R = \begin{array}{c} \\ 2 \\ 4 \\ 6 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccc} 2 & 4 & 6 & 8 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

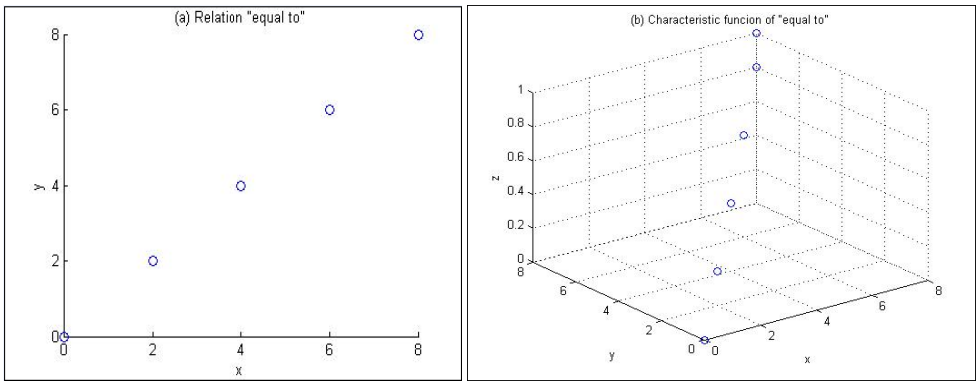


Figure 1.1 Relation “equal to” and its characteristic function

1.3 Fuzzy relation

Let X, $Y \subseteq R$ be universal sets then;

$$R = \left\{ \left((x, y), \mu_R (x, y) \right) | (x, y) \in X \times Y \right\}$$

is called a fuzzy relation in $X \times Y \subseteq \mathrm{R}$

or X and Y are two universal sets, the fuzzy relation $R(x, y)$ is given as

$$R(x, y) = \left\{ \frac{\mu_R (x, y)}{(x, y)} | (x, y) \in X \times Y \right\}$$

Fuzzy relations are often presented in the form of two dimensional tables. A $m \times n$ matrix represents a contented way of entering the fuzzy relation R.

$$R = \begin{array}{c} \\ x_1 \\ \vdots \\ x_m \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & \cdots & y_n \end{array} \\ \begin{bmatrix} \mu_R (x_1, y_1) & \cdots & \mu_R (x_1, y_n) \\ \vdots & \ddots & \vdots \\ \mu_R (x_m, y_1) & \cdots & \mu_R (x_m, y_n) \end{bmatrix} \end{array}$$

1.3.1 Example

Let $X = \{1, 2, 3\}$ and $Y = \{1, 2\}$

If the membership function associated with each order pair (x, y) is given by

$$\mu_R (x, y) = e^{-(x-y)^2}$$

then derive fuzzy relation.

Solution

The fuzzy relation can be defined in two ways using the standard nomenclature we have.

$$R=\left\{\frac{e^{-(1-1)^2}}{(1,1)},\frac{e^{-(1-2)^2}}{(1,2)},\frac{e^{-(2-1)^2}}{(2,1)},\frac{e^{-(2-2)^2}}{(2,2)},\frac{e^{-(3-1)^2}}{(3,1)},\frac{e^{-(3-2)^2}}{(3,2)}\right\}$$

$$R=\left\{\frac{1.0}{(1,1)},\frac{0.37}{(1,2)},\frac{0.37}{(2,1)},\frac{1.0}{(2,2)},\frac{0.02}{(3,1)},\frac{0.37}{(3,2)}\right\}$$

In the second method using the relational matrix, we have

$$R=\begin{bmatrix}1 & 0.37\\ 0.37 & 1\\ 0.02 & 0.37\end{bmatrix}$$

Thus the membership function describes the closeness between set X and Y . From the relational matrix it is obvious that higher values imply stronger relation.

1.4 The maximum-minimum composition of relations

Let X , Y and Z be universal sets and let R be a relation that relates elements from X to Y , i.e.

$$R=\left\{\left((x,y),\mu_R(x,y)\right)\right\}\quad x\in X,y\in Y,R\subset X\times Y$$

and

$$Q=\left\{\left((y,z),\mu_Q(y,z)\right)\right\}\quad y\in Y,z\in Z,Q\subset Y\times Z$$

Then S will be a relation that relates elements in X that R contains to the elements in Z that Q contains, i.e.

$$S=R\circ Q$$

Here "$\circ$" means the composition of membership degrees of R and Q in the max-min sense.

$$S=\left\{\left((x,z),\mu_s(x,z)\right)\right\}\quad x\in X,z\in Z,S\subset X\times Z$$

max-min composition is then defined as

$$\mu_S(x,z) = \max_{y \in Y}\left(min\left(\mu_R(x,y), \mu_Q(y,z)\right)\right)$$

and max product composition is then defined

$$\mu_S(x,z) = \max_{y \in Y}\left(min\left(\mu_R(x,y).\mu_Q(y,z)\right)\right)$$

1.4.1 Example

Let $X = \{x_1, x_2\}$ *and* $Y = \{y_1, y_2\}$ *and* $Z = \{z_1, z_2\}$

$$R = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Then find the max-min composition and max product composition

$$S = R \circ Q$$

$$S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{is the max-min composition.}$$

and $$S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is the max product composition

For crisp relations max-min composition and max product will yield the same result, when X has three elements, Y has four elements and Z has two elements like

$X = \{x_1, x_2, x_3\}$ *and* $Y = \{y_1, y_2, y_{3,} y_4\}$ *and* $Z = \{z_1, z_2\}$ then for relations

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

$$Q = \begin{array}{c} \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \end{array}$$

the max-min composition is

$$S = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} z_1 & z_2 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \end{array}$$

In this example max-min composition and max product have the same result.

1.5 Fuzzy max-min composition operation

Let us consider two fuzzy relations R_1 and R_2 defined on a Cartesian space $X \times Y$ and $Y \times Z$ respectively. The max-min composition of R_1 and R_2 is a fuzzy set defined on a cartesian spaces $X \times Z$ as

$$R_1 \circ R_2 = \left[(x,z), \max\left\{ \min\left\{ \mu_{R_1}(x,y), \mu_{R_2}(y,z) \right\} \right\} \right] | \, x \in X, y \in Y, z \in Z$$

where $R_1 \circ R_2$ is the max-min composition of fuzzy relations R_1 and R_2 and max product composition is defined as

$$\mu_{R_1 \circ R_2} = \max\left[\mu_{R_1}(x,y) . \mu_{R_2}(y,z) \right] | \, x \in X, y \in Y, z \in Z$$

1.5.1 Example

Let $R_1(x,y)$ and $R_2(x,y)$ be defined as the following relational matrices

$$R_1 = \begin{bmatrix} 0.6 & 0.5 \\ 1 & 0.1 \\ 0 & 0.7 \end{bmatrix} \quad \text{and} \quad R_2 = \begin{bmatrix} 0.7 & 0.3 & 0.4 \\ 0.9 & 0.1 & 0.6 \end{bmatrix}$$

We shall first calculate the max-min composition $R_1 \circ R_2$

$$R_1 \circ R_2 = \begin{bmatrix} 0.6 & 0.5 \\ 1 & 0.1 \\ 0 & 0.7 \end{bmatrix} \circ \begin{bmatrix} 0.7 & 0.3 & 0.4 \\ 0.9 & 0.1 & 0.6 \end{bmatrix}$$

Now we calculate

$$\mu_{R_1 \circ R_2}(x_1, z_1) = \max\left(\min(0.6, 0.7), \min(0.5, 0.9)\right) = \max(0.6, 0.5) = 0.6$$

Similarly we can calculate the other entries. The relational matrix for max-min composition in fuzzy relation is thus

$$R_1 o R_2 = \begin{bmatrix} 0.6 & 0.3 & 0.5 \\ 0.7 & 0.3 & 0.4 \\ 0.7 & 0.1 & 0.6 \end{bmatrix}$$

1.5.2 Example

Let $R_1(x, y)$ *and* $R_2(x, y)$ be defined by the following relational matrix

$$R_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccccc} y_1 & y_2 & y_3 & y_4 & y_5 \end{array} \\ \begin{bmatrix} 0.1 & 0.2 & 0 & 1 & 0.7 \\ 0.3 & 0.5 & 0 & 0.2 & 1 \\ 0.8 & 0 & 1 & 0.4 & 0.3 \end{bmatrix} \end{array}$$

$$R_2 = \begin{array}{c} \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & z_3 & z_4 \end{array} \\ \begin{bmatrix} 0.9 & 0 & 0.3 & 0.4 \\ 0.2 & 1 & 0.8 & 0 \\ 0.8 & 0 & 0.7 & 1 \\ 0.4 & 0.2 & 0.3 & 0 \\ 0 & 1 & 0 & 0.8 \end{bmatrix} \end{array}$$

we shall first compute the max-min composition $R_1 \circ R_2(x, z)$

$$\mu_{R_1 \circ R_2}(x_1, z_1) = \max\left(\min(0.1, 0.9), \min(0.2, 0.2), \min(0, 0.8), \min(1, 0.4), \min(0.7, 0)\right)$$

$$= \max(0.1, 0.2, 0, 0.4, 0) = 0.4$$

Similarly we can determine the grades of membership for all pairs

$$(x_i, z_i), i = 1, 2, 3,\ j = 1, \ldots\ldots 4$$

$$R_1 o R_2 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & z_3 & z_4 \end{array} \\ \begin{bmatrix} 0.4 & 0.7 & 0.3 & 0.7 \\ 0.3 & 1 & 0.5 & 0.8 \\ 0.8 & 0.3 & 0.7 & 1 \end{bmatrix} \end{array}$$

for the max product composition, we calculate

$$\mu_{R_1}(x_1, y_1).\mu_{R_2}(y_1, z_1) = 0.1.0.9 = 0.09$$

$$\mu_{R_1}(x_1, y_2).\mu_{R_2}(y_2, z_1) = 0.2.0.2 = 0.04$$

$$\mu_{R_1}(x_1, y_3).\mu_{R_2}(y_3, z_1) = 0.0.8 = 0$$

$$\mu_{R_1}(x_1, y_4).\mu_{R_2}(y_4, z_1) = 1.0.4 = 0.4$$

$$\mu_{R_1}(x_1, y_5).\mu_{R_2}(y_5, z_1) = 0.7.0 = 0$$

hence

$$\mu_{R_1 o R_2}(x_1, z_1) = \max\{0.09, 0.04, 0, 0.4, 0\} = 0.4$$

In the similar way after performing the remaining computation, we obtain

$$R_1 o R_2 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & z_3 & z_4 \end{array} \\ \begin{bmatrix} 0.4 & 0.7 & 0.3 & 0.56 \\ 0.27 & 1 & 0.4 & 0.8 \\ 0.8 & 0.3 & 0.7 & 1 \end{bmatrix} \end{array}$$

1.6 Conclusion

It is clear from the example that max-min composition and max product composition of crisp relations will yield the same result, but in fuzzy max-min composition and max product composition have different result.

Unit 4

2.1 Projection of Fuzzy Relation

Let $R=\{((x,y),\mu_R(x,y))\,|\,(x,y)\in X\times Y\}$ be a fuzzy relation. The projection of $R(x,y)$ on X denoted by R_1 is given by

$$R_1=\left\{\left(x,\max_y \mu_R(x,y)\right)\,|\,(x,y)\in X\times Y\right\}$$

and the projection of $R(x,y)$ on Y denoted by R_2 is given by

$$R_2=\left\{\left(y,\max_x \mu_R(x,y)\right)\,|\,(x,y)\in X\times Y\right\}$$

2.1.1 Example

Let R be a fuzzy relation defined by the following relational matrix

$$R=\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}\begin{array}{c} \begin{array}{cccccc} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \end{array} \\ \begin{bmatrix} 0.1 & 0.2 & 0.4 & 0.8 & 1 & 0.8 \\ 0.2 & 0.4 & 0.8 & 1 & 0.8 & 0.6 \\ 0.4 & 0.8 & 1 & 0.8 & 0.4 & 0.2 \end{bmatrix} \end{array}$$

The projection of $R(x,y)$ on X is calculated as, e.g.

$$\mu_{R_1}(x_1)=\max\{0.1,0.2,0.4,0.8,1,0.8\}=1$$

In the similar way can calculate the grades of membership for all pairs, so the X projection is

$$R_1=\{(x_1,1),(x_2,1),(x_3,1)\}$$

The projection of $R(x,y)$ on Y is calculated as, e.g.

$$\mu_{R_1}(y_1)=\max\{0.1,0.2,0.4\}=0.4$$

In the similar way we can determine the membership grade for all other pairs, so the Y projection

$$R_2=\{(y_1,0.4),(y_2,0.8),(y_3,1),(y_4,1),(y_5,1),(y_6,0.8)\}$$

2.2 Cylindrical extension of fuzzy relation

The cylindrical extension on $X \times Y$ of a fuzzy set A of X is a fuzzy relation cylA whose membership function is equal to

$$cylA(x, y) = A(x), \qquad \forall x \in X, \qquad \forall y \in Y$$

Cylindrical extension from *X*-projection means filling all the columns of the related matrix by the X -projection. Similarly cylindrical extension from Y projection means filling all the rows of the relational matrix by the Y -projection.

2.2.1 Example

The cylindrical extension of R_2 form the previous example is

$$R_2 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccccc} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \end{array} \\ \begin{bmatrix} 0.4 & 0.8 & 1 & 1 & 1 & 0.8 \\ 0.4 & 0.8 & 1 & 1 & 1 & 0.8 \\ 0.4 & 0.8 & 1 & 1 & 1 & 0.8 \end{bmatrix} \end{array}$$

2.3 Reflexive Relation

Let R be a fuzzy relation in $X \times X$ then R is called reflexive, if

$$\mu_R(x, x) = 1 \qquad \forall x \in X$$

2.3.1 Example

Let $X = \{1, 2, 3, 4\}$

$$R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 1 & 0.9 & 0.6 & 0.2 \\ 0.9 & 1 & 0.7 & 0.3 \\ 0.6 & 0.7 & 1 & 0.9 \\ 0.2 & 0.3 & 0.9 & 1 \end{bmatrix} \end{array}$$

is reflexive relation

2.4 Antireflexive relations

Fuzzy relation $R \subset X \times X$ is antireflexive if

$$\mu_R(x, x) = 0, x \in X$$

2.4.1 Example

$$R_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0.6 \\ 0.3 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \end{array}$$ is antireflexive relation

2.5 Symmetric Relation

A fuzzy relation R is called symmetric if,

$$\mu_R(x, y) = \mu_R(y, x) \quad \forall x, y \in X$$

2.5.1 Example

Let $X = \{x_1, x_2, x_3\}$

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 0.8 & 0.1 & 0.7 \\ 0.1 & 1 & 0.6 \\ 0.7 & 0.6 & 0.5 \end{bmatrix} \end{array}$$ is a symmetric relation.

2.6 Antisymmetric Relation

Fuzzy relation $R \subset X \times X$ is antisymmetric iff

$$\text{if } \mu_R(x, y) > 0 \text{ then } \mu_R(y, x) = 0 x, y \in X, x \neq y$$

2.6.1 Example

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0.7 \\ 0.2 & 0 & 0 \\ 0 & 0.2 & 0 \end{bmatrix} \end{array}$$ is antisymmetric relation.

2.7 Transitive Relation

Fuzzy relation $R \subset X \times X$ is transitive in the sense of max-min iff

$$\mu_R(x,z) \geq \max_{y \in X}\left(\min\left(\mu_R(x,y), \mu_R(y,z)\right)\right) x, z \in X$$

since $R^2 = R \circ R$ if

$$\mu_{R^2}(x,z) = \max_{y \in X}\left(\mu_R(x,y), \mu_R(y,z)\right)$$

then R is transitive if $R \circ R = R\left(R \circ R \subseteq R\right)$

and $R^2 \subset R$ means that $\mu_{R^2}(x,y) \leq \mu_R(y,x)$

2.7.1 Example

Let $X = \{x_1, x_2, x_3\}$

is $R = \begin{bmatrix} 0.7 & 0.9 & 0.4 \\ 0.1 & 0.3 & 0.5 \\ 0.2 & 0.1 & 0 \end{bmatrix}$ a transitive relation?

Solution

$$R \circ R = \begin{bmatrix} 0.7 & 0.9 & 0.4 \\ 0.1 & 0.3 & 0.5 \\ 0.2 & 0.1 & 0 \end{bmatrix} \circ \begin{bmatrix} 0.7 & 0.9 & 0.4 \\ 0.1 & 0.3 & 0.5 \\ 0.2 & 0.1 & 0 \end{bmatrix}$$

$$R^2 = \begin{bmatrix} 0.7 & 0.7 & 0.5 \\ 0.2 & 0.3 & 0.3 \\ 0.2 & 0.2 & 0.2 \end{bmatrix}$$

Since $\mu_{R^2}(x_i, x_j)$ is not always less than or equal to $\mu_R(x_i, x_j)$, hence R is not transitive.

2.7.2 Example

Let $X = \{x_1, x_2\}$

$$\text{is } R = \begin{bmatrix} 0.4 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} \text{a transitive relation?}$$

Solution

$$R \circ R = \begin{bmatrix} 0.4 & 0.2 \\ 0.7 & 0.3 \end{bmatrix} \circ \begin{bmatrix} 0.4 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}$$

using max-min composition

$$R^2 = \begin{bmatrix} \max(\min(0.4, 0.4), \min(0.2, 0.7)) & \max(\min(0.4, 0.2), \min(0.2, 0.3)) \\ \max(\min(0.7, 0.4), \min(0.3, 0.7)) & \max(\min(0.7, 0.2), \min(0.3, 0.3)) \end{bmatrix}$$

$$R^2 = \begin{bmatrix} \max(0.4, 0.2) & \max(0.2, 0.2) \\ \max(0.4, 0.3) & \max(0.2, 0.3) \end{bmatrix}$$

$$R^2 = \begin{bmatrix} 0.4 & 0.2 \\ 0.4 & 0.3 \end{bmatrix}$$

$\mu_{R^2}(x_i, x_j)$ is less than or equal to $\mu_R(x_i, x_j)$, so R is transitive.

2.8 Similarity Relations

$R \subset X \times X$ which is reflexive, symmetric and transitive is called the similarity relation.

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{bmatrix} 1 & 0.2 & 1 & 0.6 & 0.2 & 0.6 \\ 0.2 & 1 & 0.2 & 0.2 & 0.8 & 0.2 \\ 1 & 0.2 & 1 & 0.6 & 0.2 & 0.6 \\ 0.6 & 0.2 & 0.6 & 1 & 0.2 & 0.8 \\ 0.2 & 0.8 & 0.2 & 0.2 & 1 & 0.2 \\ 0.6 & 0.2 & 0.6 & 0.8 & 0.2 & 1 \end{bmatrix} \end{array}$$ is a similarity relation.

2.8.1 Theorem

Each equivalence class $R[X]$ is given as

$$R[X] = \bigcup_{\alpha} \alpha / R_{\alpha}[X], \alpha \in [0,1]$$

where $R_{\alpha}[X]$ is the α-cut of $R[\alpha]$.

2.8.2 Definition

$A \subset X$, A is a fuzzy set the α-cut of A is a non fuzzy set denoted by A_{α} and defined by

$$A_{\alpha} = \{x : \mu_A(x) \geq \alpha\}, \quad \alpha \in [0,1]$$

2.8.3 Example

For $R[x_1]$ we have

$$R_{0.2}[x_1] = \{x_1, x_2, x_3, x_4, x_5, x_6\}$$

$$0.2 / R_{0.2}[x_1] = 0.2/x_1 + 0.2/x_2 + 0.2/x_3 + 0.2/x_4 + 0.2/x_5 + 0.2/x_6$$

$$R_{0.6}[x_1] = \{x_1, x_3, x_4, x_6\}$$

$$0.6 / R_{0.6}[x_1] = 0.6/x_1 + 0.6/x_3 + 0.6/x_4 + 0.6/x_6$$

$$R_1[x_1] = \{x_1, x_3\}$$

$$1 / R_1[x_1] = 1/x_1 + 1/x_3$$

Equivalence class for $R[x_1]$

$$R[x_1] = 0.2/x_1 + 0.2/x_2 + 0.2/x_3 + 0.2/x_4 + 0.2/x_5 + 0.2/x_6 + 0.6/x_1 + 0.6/x_3 + 0.6/x_4 + 0.6/x_6 + 1/x_1 + 1/x_3$$

$$R[x_1] = \max(0.2, 0.6, 1)/x_1 + 0.2/x_2 + \max(0.2, 0.6, 1)/x_3 + \max(0.2, 0.6)/x_4 + 0.2/x_5 + \max(0.2, 0.6)/x_6$$

$$R[x_1] = 1/x_1 + 0.2/x_2 + 1/x_3 + 0.6/x_4 + 0.2/x_5 + 0.6/x_6$$

2.8.4 Example

Equivalence class for the similarity relation R is

$$R[x_1] = 1/x_1 + 0.2/x_2 + 1/x_3 + 0.6/x_4 + 0.2/x_5 + 0.6/x_6$$

$$R[x_2] = 0.2/x_1 + 1/x_2 + 0.2/x_3 + 0.2/x_4 + 0.8/x_5 + 0.2/x_6$$

$$R[x_3] = 1/x_1 + 0.2/x_2 + 1/x_3 + 0.6/x_4 + 0.2/x_5 + 0.6/x_6$$

$$R[x_4] = 0.6/x_1 + 0.2/x_2 + 0.6/x_3 + 1/x_4 + 0.2/x_5 + 0.8/x_6$$

$$R[x_5] = 0.2/x_1 + 0.8/x_2 + 0.2/x_3 + 0.2/x_4 + 1/x_5 + 0.2/x_6$$

$$R[x_6] = 0.6/x_1 + 0.2/x_2 + 0.6/x_3 + 0.8/x_4 + 0.2/x_5 + 1/x_6$$

2.9 Antisimilarity Relation

If R is a similarity relation then the complement of R is antisimilarity relation.

$R \subset X \times X$ is a antisimilarity relation if

$$\mu_{R'}(x,y) = 1 - \mu_R(x,y)$$

The antisimilarity relation is antireflexive, symmetric and transitive in the sense of max-min, i.e.

$$\mu_{R'}(x,z) \geq \min_{y \in X}\left(\max\left(\mu_{R'}(x,y), \mu_{R'}(y,z)\right)\right) x, z \in R$$

2.9.1 Example

Prove that $R = \begin{bmatrix} 1 & 0.1 & 0.7 \\ 0.1 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}$ is antisimilarity relation?

Solution

According to definition of antisimilarity relation

$$\mu_{R'}(x,y) = 1 - \mu_R(x,y)$$

$$\mu_{R'}(x,y) = 1 - \begin{bmatrix} 1 & 0.1 & 0.7 \\ 0.1 & 1 & 0.7 \\ 0.7 & 0.7 & 1 \end{bmatrix}$$

$$\mu_{R'}(x,y) = \begin{bmatrix} 0 & 0.9 & 0.3 \\ 0.9 & 0 & 0.3 \\ 0.3 & 0.3 & 0 \end{bmatrix}$$

This is anti-reflexive, symmetric and transitive, so R is antisimilarity relation.

2.10 Weak Similarity

$R \subset X \times X$ which is reflexive and symmetric is called the relation of weak similarity (not transitive).

$$R = \begin{bmatrix} 1 & 0.1 & 0.8 & 0.2 & 0.3 \\ 0.1 & 1 & 0 & 0.3 & 1 \\ 0.8 & 0 & 1 & 0.7 & 0 \\ 0.2 & 0.3 & 0.7 & 1 & 0.6 \\ 0.3 & 1 & 0 & 0.6 & 1 \end{bmatrix}$$ is weak similarity relation

2.11 Order Relation

An order relation $R \subset X \times X$ is transitive relation in the sense of max-min; i.e

$$\mu_R(x,z) \geq \max_{y \in X} \left(\min \left(\mu_R(x,y), \mu_R(y,z) \right) \right), x, z \in X$$

2.12 Pre Order Relations

A pre order relation $R \subset X \times X$ is reflexive and transitive in the max-min sense e.g.

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 1 & 0.7 & 0.8 & 0.5 & 0.5 \\ 0 & 1 & 0.3 & 0 & 0.2 \\ 0 & 0.7 & 1 & 0 & 0.2 \\ 0.6 & 1 & 0.9 & 1 & 0.6 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

2.13 Half Order Relation

A fuzzy half order is a relation $R \subset X \times X$ which is reflexive

$$\mu_R(x, x) = 1 \qquad \forall x \in X$$

and weakly antisymmetric, i.e.

$$\text{if} \quad \mu_R(x, y) > 0 \text{ and } \mu_R(y, x) > 0 \quad \text{then} \quad x = y$$

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \begin{bmatrix} 1 & 0.8 & 0.2 & 0.6 & 0.6 & 0.4 \\ 0 & 1 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 1 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 & 0.6 & 0.4 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$ is half order relation

3.1 Fuzzy Graph

In 1975, Rosenfeld considered fuzzy relations on fuzzy sets. He developed the theory of fuzzy graphs. Bang and Yeh during the same time introduced various connectedness concepts in fuzzy graph. Inexact information is used in expressing or describing human behaviors and mental process. The information depends upon a person subjectively and it is difficult to process objectively.

Fuzzy information can be analyzed by using a fuzzy graph. Fuzzy graph is an expression of fuzzy relation and thus the fuzzy graph is frequently expressed in fuzzy matrix.

Mathematically a graph is defined as $G = (V, E)$ where V denotes the set of vertices and E denotes the set of edges. A graph is called a crisp graph if all the values of arcs are 1 or 0 and a graph is called fuzzy graph if its values is between 0 and 1.Fuzzy graph $G = (\sigma, \mu)$ is a pair of functions $\sigma : S \to [0,1]$ where S is the set of vertices and $\mu : S \times S \to [0,1], \forall x, y \in S$.

Fuzzy graph $H = (\tau, v)$ is called a fuzzy subgraph of G if

$$\tau(x) \le \sigma(x), \forall x \in S \quad \text{and} \quad v(x, y) \le \mu(x, y) \forall x, y \in S$$

3.1.1Example

Fuzzy relation is defined by the following fuzzy matrix the corresponding fuzzy graph is shown in the fig (3.1)

$$\begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{bmatrix} 0.5 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.5 \\ 1.0 & 1.0 & 0.0 \end{bmatrix} \end{array}$$

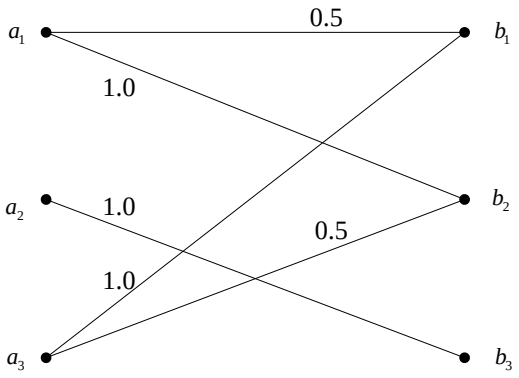


Fig 3.1

Fuzzy graph

3.2 Complement of a Fuzzy Graph

The complement of a fuzzy graph $G:(\sigma,\mu)$ is a fuzzy graph $\overline{G}:(\overline{\sigma},\overline{\mu})$ where $\overline{\sigma}\equiv\sigma$ and

$$\overline{\mu}(u,v)=\sigma(u)\wedge\sigma(v)-\mu(u,v)\quad\forall\ u,v\in V$$

Complement of a fuzzy graph are shown in fig below

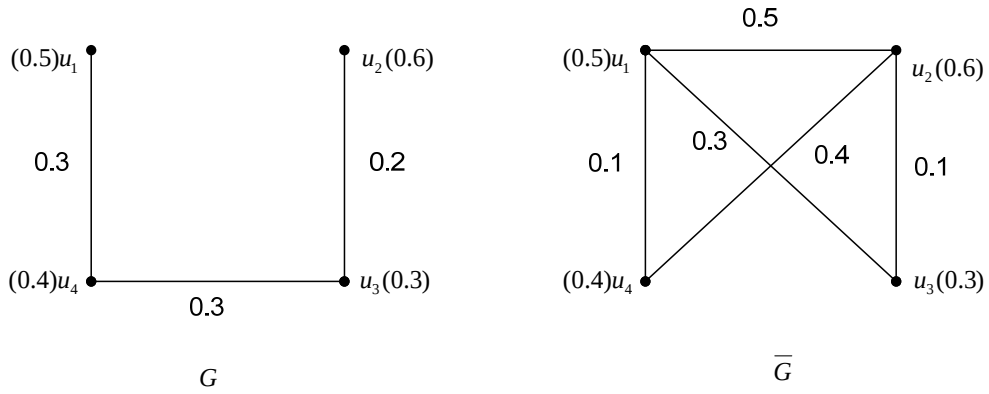


Fig 3.2(a)

Complement of a fuzzy graph

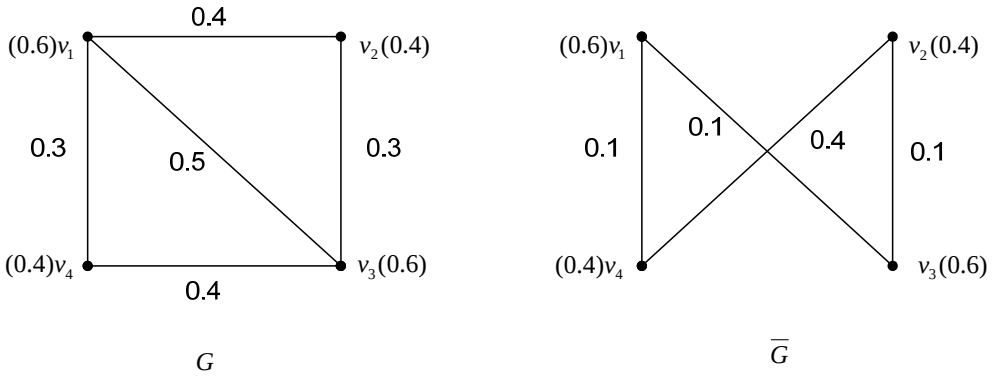


Fig 3.2(b)

Complement of a fuzzy graph

3.3 Model for Predicting Score in Cricket

In this model we can predict score using max-min composition,max product composition and max-av composition.

Speed of bowling = {fast bowling, medium bowling, spin bowling} and

Y= condition on pitches= {good wicket, fair wicket, sporting wicket, green wicket, crumbling wicket, rough wicket}

Let *R* denotes the relationship between speed of bowling and condition on pitch and *Q* denotes the relationship between conditions on pitches and runs on the board.

$$R = \begin{array}{c} \\ fast \\ medium \\ spin \end{array} \begin{array}{c} \begin{array}{cccccc} gd.w & f.w & s.w & gr.w & c.w & r.w \end{array} \\ \begin{bmatrix} 0.6 & 0.5 & 0.4 & 0.1 & 0.9 & 0.5 \\ 0.8 & 0.6 & 0.9 & 0.2 & 0.1 & 0.6 \\ 0.7 & 0.8 & 0.6 & 0.7 & 0.1 & 0.2 \end{bmatrix} \end{array}$$

and

$$Q = \begin{array}{c} \\ gd.w \\ f.w \\ s.w \\ gr.w \\ c.w \\ r.w \end{array} \begin{array}{c} \begin{array}{ccc} low.r & ave.r & hig.r \end{array} \\ \begin{bmatrix} 0.4 & 0.8 & 0.7 \\ 0.3 & 0.8 & 0.8 \\ 0.2 & 0.7 & 0.8 \\ 0.8 & 0.6 & 0.4 \\ 0.7 & 0.5 & 0.4 \\ 0.9 & 0.4 & 0.2 \end{bmatrix} \end{array}$$

$R \circ Q =$ Relationship between speed of the bowling and runs on the board

We calculate $R \circ Q$ by using max-min composition rule

$$\max\{\min(0.6,0.4),\min(0.5,0.3),\min(0.4,0.2),\min(0.1,0.8),\min(0.9,0.7),\min(0.5,0.9)\}$$

$$= \max\{0.4,0.3,0.2,0.1,0.7,0.5\}$$

$$= 0.7$$

Similarly we can calculate the other entries

The relational matrix for max-min composition in fuzzy relational is thus

$$R \circ Q = \begin{array}{c} \\ fast \\ medium \\ spin \end{array} \begin{array}{c} \begin{array}{ccc} low.r & ave.r & hig.r \end{array} \\ \begin{bmatrix} 0.7 & 0.6 & 0.6 \\ 0.6 & 0.8 & 0.8 \\ 0.7 & 0.8 & 0.8 \end{bmatrix} \end{array} \qquad (3.1)$$

Max Product composition

Now by using max product composition we find the relationship between speed of the bowling and runs on the board

$R \circ Q =$ Relationship between speed of the bowling and runs on the board

We calculate $R \circ Q$ by using max product composition rule

$$\max(0.6.0.4, 0.5.0.3, 0.4.0.2, 0.1.0.8, 0.9.0.7, 0.5.0.9)$$

$$= \max(0.24, 0.15, 0.08, 0.08, 0.63, 0.45)$$

$$= 0.63$$

Similarly

$$\max\left(0.48, 0.4, 0.28, 0.06, 0.45, 0.2\right)$$

$$= 0.48$$

and

$$\max\left(0.42, 0.4, 0.32, 0.04, 0.36, 0.1\right)$$

$$= 0.42$$

Similarly we calculate the other entries and the relational matrix for max product composition is

$$R \circ Q = \begin{array}{r} \\ fast \\ medium \\ spin \end{array} \begin{array}{c} \begin{array}{ccc} low.r & ave.r & hig.r \end{array} \\ \begin{bmatrix} 0.63 & 0.48 & 0.4 \\ 0.54 & 0.64 & 0.64 \\ 0.56 & 0.64 & 0.64 \end{bmatrix} \end{array} \tag{3.2}$$

Max-av Composition

Now by using max product composition we find the relationship between speed of the bowling and runs on the board

$R \circ Q =$ Relationship between speed of the bowling and runs on the board

We calculate $R \circ Q$ by using max-av composition rule

$$\frac{1}{2}.\max\left(0.6+0.4, 0.5+0.3, 0.4+0.2, 0.1+0.8, 0.9+0.7, 0.5+0.9\right)$$

$$= \frac{1}{2}.\max\left(1, 0.8, 0.6, 0.9, 0.16, 0.14\right)$$

$$= \frac{1}{2}\left(0.16\right)$$

$$= 0.8$$

for the second entry

$$\frac{1}{2}.\max\left(0.14, 0.13, 0.11, 0.7, 0.14, 0.9\right)$$

$$= \frac{1}{2}\left(0.14\right)$$

$$= 0.7$$

for third entry

$$\frac{1}{2}.\max\left(0.13, 0.13, 0.12, 0.5, 0.13, 0.7\right)$$

$$= \frac{1}{2}\left(0.13\right)$$

$$= 0.65$$

Similarly we calculate the other entries and the relational matrix for max-av composition is

$$R \underset{av}{\circ} Q = \begin{array}{c} \\ fast \\ medium \\ spin \end{array} \begin{array}{c} \begin{array}{ccc} low.r & ave.r & hig.r \end{array} \\ \begin{bmatrix} 0.8 & 0.7 & 0.65 \\ 0.85 & 0.8 & 0.85 \\ 0.75 & 0.8 & 0.8 \end{bmatrix} \end{array} \tag{3.3}$$

By analyzing the results of $(3.1), (3.2)$ and (3.3) we conclude that (3.2) is more reliable.

Core Paper

MATC 3.3

Block - II

Marks : 37 (SSE : 30; IA : 05)

Computer Programming in 'C' (Theory)

Syllabus

• Unit 5 •

Fundamentals of 'C' Language : Basic structure of a 'C' program, Basic Data type, Constants and Variables, Identifier, Keywords, Constants, Basic data type, Variables, Declaration and Initialization, Statements and Symbolic constants. Compilation and Execution of a 'C' program.

• Unit 6 •

Operators and Expressions : Arithmetic, Relational, Logical operators. Increment, Decrement, Control, Assignment, Bitwise, and Special operators. Precedence rules of operators, Type Conversion (casting), Modes of arithmetic expressions, Conditional expressions.

• Unit 7 •

Input / Output Operations : Formatted I/O - Single character I/O (getchar(), putchar()), Data I/O (scanf(), printf()), String I/O (gets(), puts()). Programming problems. Decision Making Statements : Branching – *if* Statement, *if* ..... *else* Statement, Nested *if* .... *else* Statement. *else* ..... *if* and *switch* Statements. Loop Control : *for* Statement, *while* Statement, *do while* Statement. *break, continue* and *exit* Statements. Programming problems.

• Unit 8 •

Functions : Function declaration, Library functions, User defined function, Passing argument to a function, Recursion. Programming problems. Arrays : Array declaration and static memory allocation. One dimensional, two dimensional and multidimensional arrays. Passing arrays to functions. Sparse matrix.

• Unit 9 •

Pointers : Basic concepts of pointer, Functions and Pointers. Pointers and Arrays, Memory allocation, Passing arrays to functions, Pointer type casting. Programming problems. Structures and Unions : Declaring a Structure, Accessing a structure element, Storing methods of structure elements, Array of structures, Nested structure, Self –referential structure, Dynamic memory allocation, Passing arrays to function. Union and rules of Union. Programming problems.

• Unit 10 •

File Operations : File Input / Output operations – Opening and Closing a file, Reading and Writing a file. Character counting, Tab space counting, File-Copy program, Text and Binary files.

Unit 5

Overview of C

Structure

History of C

Importance of C

Sample Programs

Basic Structure Of C Programs

Programming style

Executing A C Program

Unix System

MS- DOS System

Summary

History of C

C is a structured general purpose machine Independent high level programming language developed by Dennis Ritchie at AT & T's Bell Labs of USA in the mid 1970s for the Unix based operating system. Many of the important concepts of C are borrowed from the language BCPL (Basic Combined Programming Language), developed by Martin Richards in 1967. Although originally designed as a systems programming language, C has proved to be a powerful and flexible language that is used for a variety of applications for nearly every available platform. The merit of C lay in the fact that it is easier to read, more flexible and more efficient at using memory. It is particularly popular for personal computer programmers because it requires less memory than other languages. C is the archetype or original model for many modern languages as when we find Language constructs in C, such as "if" statements, "for" and "while" loops, and types of variables, can be found in many later languages. Today, there are very few platforms that do not have a C compiler

In the late, seventies C began to replace the more familiar languages of that time like, ALGOL, PL/I, etc. The drawback of the B language was that it did not know data-types. Both BCPL and B are " type less" system programming languages. By Contrast, C Provides a variety of data types with powerful features. The fundamental data types are integers, characters and floating point numbers of various sizes. In addition there is a hierarchy of derived data types created with arrays, pointers, structures and union.

Since C was developed along with the UNIX operating system, it is has close association with UNIX. Major parts of the popular operating systems like windows, Linux and Unix are coded in C. This is because when it comes to performance nothing beats C. Although C is technically a high-level language, it is one of the "lowest-level" high-level programming languages in the sense; it is much closer to assembly language than are most other high-level languages. This closeness to the underlying machine language allows C programmers to write very efficient code. More over if one is

to extend the operating system to work with new devices one needs to write device driver programs. These programmes are exclusively written in C.

For many years, C was the reference manual, but eventually with the appearance of many C compilers coupled with the wide popularity of UNIX operating system, it gained wide popularity among computer professionals. Today, C is the language of choice while building a variety of hardware and operating system platforms.

The American National Standards Institute (ANSI) constituted a committee in 1983, to provide an updated definition of C. The resulting definition "**ANSI C** "was completed in late 1988, and modern compilers are already supporting most of the features of this standard .The standard is based on the original reference Manual in the first edition, the classic book **"The C Programming Language"** , with little or no changes on the original design of the C language . They ensured that old programs still worked with the new standard, failing that, the compiler would produce warnings of new behavior.

One of the significant contributions of the standard is the definition of a new syntax for the defining and declaration of the function. This extra information makes it easier for compilers to detect errors caused by mismatched arguments. A second significant contribution of the standard is the definition of a library to accompany C. These library functions specifies functions for accessing the operating system, formatted input and output, memory allocation, string manipulation, and the like. A collection of standard headers provides uniform.

3.2. Importance of C

C is an immensely popular language widely used and well understood. Some of the versatile features of C language are**:** reliability, portability, flexibility, interactivity, modularity and finally efficiency and effectiveness. It is a great tool for expressing programming ideas in a way it is easily understood, regardless of the language users are most familiar with. It is in fact the original or archetypal building block for many other currently known languages and it is very close to assembly language. C is a robust language whose rich set of built in functions, and operators can be used to write any complex programs. In C large programs are divided into small programs called functions and data moves freely around the systems from one function to another. Moreover, the C compiler combines the capabilities of an assembly language with the attributes of a high level language and therefore it is useful for writing both system software and business packages without worrying about the hardware platforms where they will be implemented..The great thing about C is that it can be used to write high performance code for both application and system software. Further it can interact with hardware at quite low level. In fact, many of the compilers available in market are written in C. It is the language used for developing system applications that forms major portion of operating systems such as Windows, UNIX and Linux. C is increasingly being used in Database systems, Graphics, Spread sheets, word processors, Compilers /Assemblers, Network drivers and interpreters.

The variety of data types and powerful operators available in C makes C programs very efficient and fast. In C there are only 32 key words and its strength lies in its built-in functions. Some standard functions are available which can be used for developing programs. C Being highly portable, programs written for one computer can be made to run on another system with little or no modification.

C is at once one of the pillars of modern information technology (IT) and computer science (CS). C is a high level language that lets us to write very low level stuff like device drivers that runs as fast as assembly written programs. C's power and fast program execution come from its ability to access low level commands, similar to assembly language, but with high level syntax. It allows low level access to information and commands while still retaining the portability and syntax of a high level language. In this process C imposes few constraints on the programmer. Further it is tailor- made for structured programming, thus requiring the user to think a problem in terms of function modules or blocks. A collection of these modules make a program debugging and testing easier..Thus, C meets the requirements, where speed, space and portability are important.

Another prime feature of C is its ability to extend itself. A program in C is basically a collection of functions that are supported by the C library. We can add our own functions to the C library .With the availability of large number of functions , the programming burden becomes simple. C being simple and easy to understand, most of the operating systems and game software are written in C .

Before discussing some distinct features of C, we shall look at some sample programs in C, and as we proceed, can learn more about the language.

3.3 Sample Programs

Printing A Message: Sample program 1

The only way to learn a new programming language is by writing programs in it. Let us begin by looking at the construction of a very simple program.

The following is the output of the above program code when it is executed:

hello, fine

```
main( )
  {
   /* ……Printing begins…….*/
        Printf(“ hello, fine ”);
    /* ……Printing ends…….*/
  }
```

Fig. 3.1 The first C program to print a single line of text

In the above C program, the code begins executing at the beginning of **main**. **main( )** is a special function used by the C systems to tell the computer where the program begins. This means that every program must have a **main** somewhere. In this example, **main** is defined to be a function that expects no arguments, which is indicated by the empty list ( ). All the statements that belong to **main( )** are enclosed within a pair of braces { } as indicated above. The opening brace **"{"** indicates the beginning of the function **main** and the closing brace **"}"** marks the end of the program. All the statements between these two braces form the function body. The function body contains a set of instructions to perform the given task.

In our example, the function body contains three statements out of which only the **printf** line is an executable statement. A function is called by naming it, followed by parenthesized list of arguments, so this calls the function **printf** with the argument " hello, fine ". **printf** is a library function that prints output , in this case the string of characters (String constant or character string) between quotes.

The two lines

```
/* ……Printing begins…….*/
```

And

```
/* ……Printing ends…….*/
```

Are **comment lines** which in this program tells what the program does. Any characters between /* and .*/ are ignored by the compiler (comments are solely given for the understanding of the programmer or the fellow programmers); they may be used freely to make a program easier to understand . Any number of comments can be written at any place in the program. The normal language rules do not apply to text written with in /* and .*/ . Thus we can type this text in small case, capital, or a combination. Moreover, comment can be split over more than one line, as in,

```
/* printing
   begins.*/
```

Such a comment is often called a multi-line comment. Comments cannot be nested. For example,

```
/* Printing begins /*Printing ends.*/*/
```

Is invalid and therefore results in an error.

Let us come back to the **printf** function, the only executable statement of the program .

```
printf(" hello, fine ");
```

The above quotation can be printed in two lines, by adding another **printf** function, as in,

```
printf("hello,\n");
printf("fine");
```

The information contained between the parentheses is called the **augment** (which are simply strings of character to be printed out) of the function. The argument of the first **printf** contains a combination of two characters \ and **n** at the end of the string. The combination sequence" \n " is called **newline** and it takes the character to the next line. Therefore, you will get the output split over two lines. **\n** is one of the several Escape Sequence (similar in concept to the carriage return key on a type writer, which when printed advances the output to the left margin on the next line) available in C. if you try something like

```
printf("hello, fine
");
```

The C compiler will produce an error message.

No space is allowed between \ and n. **printf** never supplies a new line automatically, so several function calls may be used to build up an output line in stages, as in,

.

```
main( )
  {
  /* ……printing begins…….*/
      printf(" hello,");
     printf(" fine,");
     printf(" \n");
   /* ……printing ends…….*/
  }
```

To produce identical output. Here \n represents only a single character. An escape sequence like \n provides a general and extensible mechanism for representing hard to type or invisible characters. It is also possible to produce multi line output by one printf statement with the use of newline character at appropriate places, as in,

```
printf ("hello\n….fine,\n……I\n……..am ok!");
```

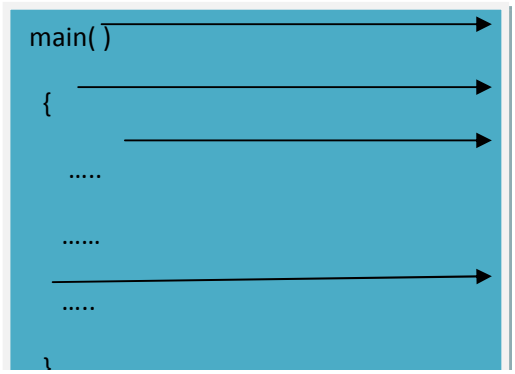
Where the output is

```
hello
…..fine,
…….I
…….am ok !
```

The inclusion of the preprocessor directive **# include < stdio.h >** at the beginning of all programs that use any input/output library functions should not be insisted for functions like, **printf** and **scanf**, **Printf** is a pre defined standard C function (predefined in the sense that it is function that has already been written, compiled, and linked together with the program at the time of linking).

Note that the print line ends with a **semi colon**. Thus the mark **;** acts as a statement terminator. That is, every C statement must end with a **;** mark. In C , everything is written in lowercase letters. However, uppercase letters are used for symbolic names representing constants. we may also use uppercase letters in output strings like "HELLO" and "FINE".

The General format of simple C programs is shown below.



Simple C program Format

The main Function

The main ( ) is a function and is part of every program. There are different forms of main statement in C. viz.,

main ( )

int main ( )

main (void)

void main (void)

int main (void)

The empty pair of parenthesis indicates that the function has no arguments This may be explicitly indicated by using the keyword **void** inside the parenthesis. Just like the way functions in a calculator returns a value, functions in C also return a value to the operating system. That is, It is also possible to specify the keyword **int or void** before the word **main**. Some compilers permit us to return nothing or no information to the operating system from **main ( ).** In such a case we should precede it with the key word **void.** The key word **void** means that the function does not return any value to the operating system and **int** means that the function s returns an integer value to operating system. When **int** is specified, the last statement in the program must be **"return 0".**

Addition of Two numbers: Sample program 2

Consider another program, which performs addition on two numbers. This program explains the need for the use of declaration of variables, and use of operators.

/Program to add two numbers:/

```
/* addition of  two numbers */
main ( )
 {
      int  num;
      float amount;
      num = 10;
      amount = 20.25+29.85;
      printf ( " % d\n",num);
      printf ("%5.2f",amount);
 }
```

On execution of this program we will get the following output:

10

50.10

The first line of the program is a **comment line**. Comment line in the beginning give information such as name of the program, author, date etc. To indicate line numbers comment characters can also be used. in other lines. The words **num** and **amount** are variable names used to store numeric data. The numeric data may be either in **real or integer** form. In C, all variables must be declared before they are used, usually at the beginning of the function before any executable statement. The type declaration statement is written at the beginning of main ( ) function. In lines 4 and 5, the declarations

int num;

float amount;

tells the compiler that num is an integer (**int)** and amount is a floating (**float**) point (numbers with fractional part) numbers. All declaration statements ends with a **semicolon.** The words such as **int** and **float** are called keywords and cannot be used as variable names .The range of both **int** and **float** depends on the machine you are using; 16- bit i**nts**, which lie between -32768 and +32768 , are common, as are 32-bit **ints**. A float number is typically 32-bit quantity, with at least six significant digits and magnitude generally between about 10^{-38} and 10^{+38}. While declaring the type of variable one can also initialize it as shown in line 7 and 9.That is , the statements

num = 10;

amount = 20.25+29.85;

are called the assignment statement. **Every assignment statement must have a semicolon at the end.**

The order in which we define the variables is sometimes important sometimes and sometimes not. For example,

int i =10, j =25;

is same as

int j= 25, i=10;

However,

float a= 1.5, b = a + 3.2;

Is alright. But

float b= a+3.2, a = 1.5 ;

Is not, because we are trying to use **a** even before defining it.

Moreover, the following statements would work

int a,b,c,d

a = b = c = d = 10;

However the following statement would not work

Int a= b= c= d =10;

The next statement of the program is an output statement that prints the value of **number.** The print statement

printf (" % d\n", num);

contains two arguments..The first argument **"%d'** tells the compiler that the value of the second argument **num** should be printed as a *decimal integer*. These arguments are separated by **comma. T**he newline character "\n " causes the next output to appear on a new line.

The last statement of the program

printf ("%5.2f", amount);

print out the value of **amoun**t in floating point format. The format specification **"%5.2f "** tells the compiler that the output must be floating type , with five places in all and two places to the right of the decimal point.

Calculation of Interest: Sample Program 3

C supports the basic four arithmetic operators (-, +, * . /) along with various others. The use of such operators along with other variable declarations, the while loop construct and # define preprocessor directive are illustrated in the program below. The program calculates the value of money at the end of each year of investment, assuming the interest rate at 11 percent with an initial investment of 50 000 for 10 years .In this program, the variable **value** represents the value of money at the end of the year and the **amount** represents the value of the money at the start of the year. The statement

amount = value ;

makes the value at the end of the current year as the value at the beginning of the *next* year .

The preprocessor compiler directive **#define,** defines a symbolic constant. Whenever a symbolic name is encountered, the compiler automatically substitutes the value associated with the name. If you want to change the value you have to simply change the definition. **#define** line should not end with a semicolon and are usually written in upper case letters(so that they can be readily distinguished from the lower case variable names), usually placed at the beginning before the **main ( )** function. They are not declared in the declaration section. The declaration section of the program declares **year** as integer and **amount ,value and rate** as floating point numbers. When two or more variables are declared in one statement, they are separated by commas. It is also possible to declare the floating point variables as multiple statements as in,

float amount;

float value;

float rate;

```
/*  ..... INVESTMENT PROBLEM …….. */
# define PERIOD       10
#define   PRINCIPAL  50000.00
/* ..... MAIN PROGRAM BEGINS …….. */
main ( )
 {  /*  .....DECLARATION STATEMENTS …….. */
        int year;
        float amount, value, rate;
 /*  ..... ASSIGNMENT STATEMENTS …….. */
       amount = PRINCIPAL ;
       rate = 0.11;
       year = 0;
/*  ..... ..... COMPUTATION STATEMENTS… …….. */

 /*  .....  COMPUTATION  USING while LOOP …….. */
       While (year < = PERIOD )
          {
            printf ( " % 2d     % 8.2 f \n" , year, amount );
            value = amount + rate * amount;
            year = year +1;
             amount = value;
          }
  /*  .....  .....  while LOOP ENDS… …….. */
 }
 /*  ..... ..... PROGRAM  ENDS              … …….. */
```

Fig.3.5 The Investment Program

In the **while** loop all computation and printing are accomplished. The body of a **while** loop can be one or more statements enclosed in braces . The parenthesis after the **while** contain a condition that is tested. So long as this condition remains true all , all statements within body of the while loop keep getting executed repeatedly. When the condition becomes false , the control passes to the first statement that follows the body of the **while** loop..In this case as long as the value of the **year** is less than or equal to the **PERIOD**, the four statements grouped by braces that follows the **while** are executed. The loop ends when year becomes greater than **PERIOD.**

Sample Program 4: Use of Sub routines:

A very simple program that explains the use of **mul ( )** function is shown below. It uses a user defined

```
/*  ..... PROGRAM USING FUNCTION …….. */
int mul  ( int a, int b);          /*  DECLARATION….. */
/*  ..... MAIN PROGRAM STARTS…….. */
      main ( )
        {
            int a, b,c;
            a =7;
            b =10;
            c = mul (a,b);
            printf ( “multiplication of %d and % d is  % d”, a,b,c);
         }
/*  ..... MAIN PROGRAM ENDS
                    MUL FUNCTION STARTS…….. */
     int mul (int x, int y)
     int p;
     {
          p = x * y;
          return ( p);
      }
/*                  MUL  (  )  FUNCTION ENDS                    . */
```

function equivalent to subroutine in **FORTRAN** or Sub program in **BASIC**. The Execution of the program will print the output

Multiplication of 7 and 10 is 70

The **mul ( )** function multiplies the value of variables x and y and the result is returned to the **main ( )** function when it is called in the statement

c = **mul (a,b);**

The **mul ( )** function has two arguments x and y (declared as integers) and when called the values of a and b are passed onto x and y respectively. This example also shows a bit more of how **printf** works.

Sample Program 5: Use of Math Functions:

There are many occasions where we often use standard mathematical functions like cos, sin, exp, etc.

```
/*  ... PROGRAM USING COSINE FUNCTION …….. */
# include < math.h >
# define PI  3.1416
# define MAX 180
main ( )
  {
     int angle;
    float x,y;
    angle = 0;
    Printf ( “Angle      Cos(angle) \n\n “);
    While (angle < = MAX)
     {
          x = ( PI/MAX) * angle;
          y = cos (x);
          printf ( “% 15 d % 13.4 f\ n “, angle, y);
          angle = angle +10;
      }
  }
```

The standard mathematical functions are defined and kept as a part of **C math library** for use in programs. The use of any of these mathematical functions in the program can be accomplished by means of **# include** instruction in the program. The **#include** directive tells the preprocessor to treat the contents of a specified file as if those contents had appeared in the source program at the point where the directive appears Like **# define**, it is also a compiler directive and tells the compiler to link the specified mathematical functions from the library. The instruction is of the form

 # include < math.h >

math.h is the file name containing the required information. Program code,(Figure 3.1) explains the use of cosine function. Another # include instruction that is often used is

include <stdio.h>

<stdio.h> refers to the standard I/O header file containing standard Input output functions. That is, it adds the contents of the file named **stdio.h** to the source program and the ankle brackets cause the preprocessor to search the directories specified by the Include environment variable for stdio.h, after searching directories specified by the / I compiler option. For example, to use the function printf( ) in a program, the line

#include <stdio.h>

Should be at the beginning of the source file, because the definition for printf() is found in the file stdio.h.

As explained earlier, C programs are divided into modules or functions. To use any of the standard functions, the appropriate header file should be included...Header files contain definitions of functions and variables which can be incorporated into any C program by using the pre-processor *#include* statement. This is done at the beginning of the C source file . To access the functions stored in the C library, it is necessary to tell the compiler about the files to be accessed. This is achieved by the use of pre processor directive

#include <filename>

Placed at the beginning of the program. Note here that **filename** is the name of the library file that contains the required function definition.

3.4 Basic Structure Of C Programs

The programs in C so far discussed illustrates that it can be viewed as a group of building blocks called functions. A function is a segment that groups a number of program statements to perform specific task. To write a c program , we must first create functions and then put them together.

The different sections of a C program as shown in figure 3.2..The documentation section consists of a set of comment lines giving the name of a program, author, date and other details, which the programmer would like to use later .The link section provides instructions to the compiler to link functions from the system library. All symbolic constants are defined in the definition section. Global

variables (variables that are used in more than one function) and all the user defined functions are declared in the global declaration section that is out side of all the functions.

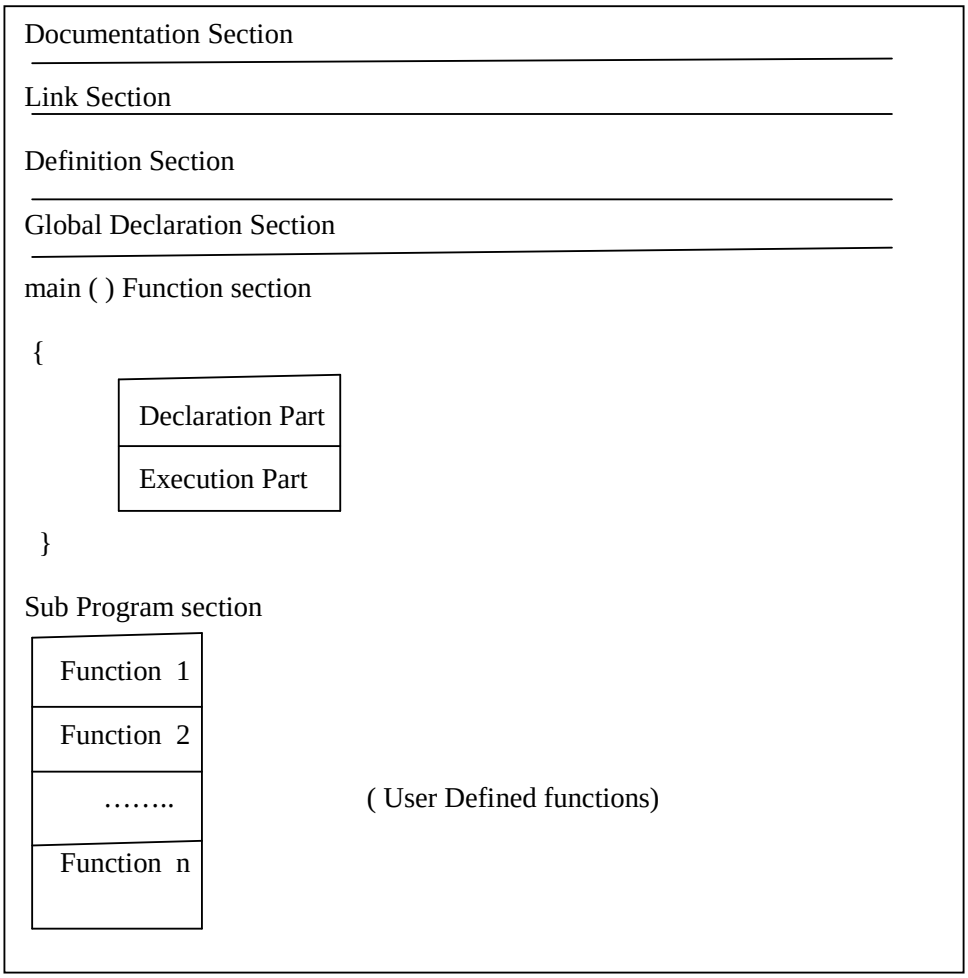


Fig.3.2 An over view of C program

Every C program must have one main ( ) function section that contains two parts, the declaration and executable part, appearing between the opening and closing braces. In the declaration part all those variables used in the executable part are declared..There is at least one statement in the executable part. The program execution begins at the opening brace and ends at the closing brace which marks the logical end of the program. Every statements in the declaration and executable parts end with a semi colon **(;)**.\

The sub program section contains all the user defined functions that are called in the **main** function. User defined functions are generally placed immediately after the **main** function, although they may appear in any order. All sections , except the main function may be absent when they are not required.

3.5 Programming Style

Programming style is a set of rules or guidelines used when writing the source code for a computer program. It is often claimed that following a particular programming style will help programmers to read and understand source code conforming to the style, and help to avoid introducing errors.

C has no specific rules for the position at which a statement is to be written. That's why it is often called a free –form language. First of all, all statements are entered in small case letters. Upper case letters are used only for symbolic constants. The statements in the program must appear in the same order in which we wish to be executed.; unless of course the logic of the problem demands a deliberate "jump", which is out of sequence. These statements are terminated with a semi-colon **(;)**, and are collected in sections known as functions. By convention, a statement should be kept on its own line. Blank spaces may be inserted between two words to improve the readability of the statement. However , no blank spaces are allowed with in a variable, constant or key word.

Since C is a free-form language, we can group statements together on one line. The statements

```
a = b;
x = y-1;
z = a-1;
```

can be written on one line as

```
a = b;  x = y-1; z = a-1;
```

The program

```
main ( )
{
    Print f ("hello");
}
```

May be written in one line like

```
main ( )  { Print f ("hello")};
```

However, this style makes the program more difficult to understand. Rather than putting everything on one line, it is much more readable to break up long lines so that each statement and declaration goes on its own line.

Comments in code can be useful and they provide the easiest way to set off specific parts of code (and their purpose); as well as providing a visual "split" between various parts of your code. Having good comments throughout your code will make it much easier to remember what specific parts of your code do. Care should be taken not to over emphasize generous use of comments inside the code. For debugging as well as testing of the code Judiciously inserted comments is very helpful and it improves the code readability as well as the understandability of the code logic.

3.6 Executing A C Program

C program Execution involves the following steps

1. **Creating the program**
2. **Compiling the program**
3. **Linking the program with functions that are needed from the C library**
4. **Executing the program.**

Although these steps remain the same irrespective of the operating system, system commands for implementing the steps and conventions for naming the files may differ on different systems. An operating system is a program that controls the entire operations of a computer system. All I/O operations are channeled through the operating system. It is an interface between the hard ware and the user. The most popular ones today are UNIX and MS-DOS .Figure 3.10 illustrates the steps involved in the execution of C program.

3.7Unix System：Creating the program

Once you have written the program you need to type it and instruct the machine to execute it. Once we

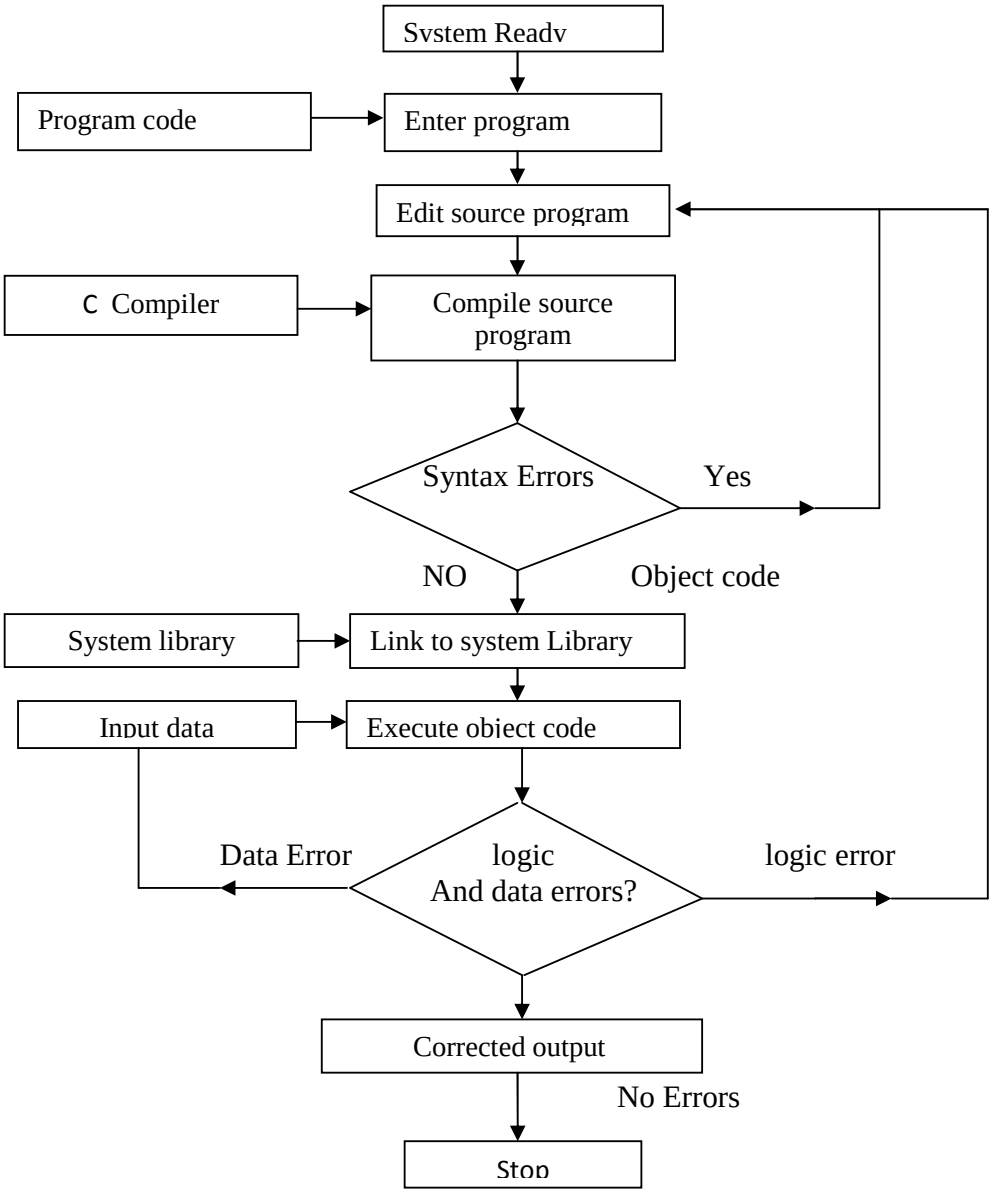


Fig.3.10 Process of compiling and running in C

load the UNIX OS in to the memory , the computer is ready to receive the program. The program must be entered into a file. The file name can consists of ,letters, digits and special characters followed by a dot and a letter c.

For eg,

hello.c

The file is created with the help of another program called text editor., either ed or vi. The command for calling the editor and creating the file is

ed filename

if the file existed before , it is loaded up. If not the file has to be created so that it is ready to receive the new program. Any corrections to the program are done under the editor. when the editing is over it is saved on the disk .It can the be referenced at any time later by its file name. The program that is entered into the file is known as ***source program*** **.**A source program is a program coded in a languages other than machine language, ad it is translated into machine language before being executed.

Compiling and Linking

Once you have written the program you need to type it and instruct the machine to execute it. To type the C program you need another program called **Editor.** Once the program has been typed it needs to be converted to machine language (0s and 1s) before the machine can execute it. To carry out this conversion we need another program called **compiler.** Assume that the source program has been created in a file named kmv.c The compilation command to achieve this task under UNIX is

cc kmv.c

The source program instructions are now translated into a form that is suitable for execution by the compiler. The translation is done after examining each instruction for its correctness. If everything is alright, the compilation proceeds silently and the translated program is stored in another file with the name ***kmv.o.*** This program is called the **object code.**

Linking is the process of putting together other programs files and functions that are required by the program. Under **UNIX**, the linking is automatically done when the **cc** command is used. Errors, if any should be should be corrected in the source program with the help of **editor** and the compilation is done again..The compiled and link program is called the **executable object code** and is stored automatically in another file named **a.out.**

Executing The Program

On compiling the program its machine language equivalent is stored as an EXE file which is an executable file. The command **a.out** would load the executable object code into the computer memory and execute the instructions .During execution, the program may request for some data to be entered *through the keyboard.*

Here are the steps that you need to follow to compile and execute your C program using Turbo C or C++.

1. start the compiler at **C >** prompt. The compiler (TC.EXE is usually present in C:\TC\BIN directory).
2. Select **New** from **File menu**
3. Type the program.
4. Save the program using **F2** under a proper name(say prog.c)
5. Use **Ctrl +F9** to compile and execute the program
6. Use **Alt +F5** to view the output.

Creating your own Executable File

Note while linking, the linker always assign the same name a.out. while Compiling a new program , this file will be over written by the executable object code of the new program .To prevent this from happening , we should rename the file immediately using the command

mv a.out *name*

Or

use the cc command option

cc-o *name* source-file

This **cc command** option will store the executable object code in the file name and prevent the old file **a.out** from being destroyed..

To compile and link multiple source program files, we must append all the filenames to the cc command.

Cc filename-1.c…… filename-n.c

These files will be separately compiled into object files called

filename-i.o

and then linked to produce an executable program file **a.out .** Also it is possible to compile each file separately and link them later .The commands,

c c - c mod1.c

c c - c mod2.c

will compile the source files mod1.c and mod2.c into object files **mod1.o and mod2.o**. They can be linked together by the command

c c mod1.o mod2.o

Further, the source and object files can be combined as

C c mod1.c mod2.o

Here only **mod1.c** is compiled and then linked with the object file **mod2.o.** This approach helps in situations when one of the source files need to be changed and recompiled or an existing object file is to be used along with the program to be compiled.

3.8 MS- DOS System

In **MS-DOS** system, the program is created by any word processing software in non document mode and should end with the characters **" .c.** ". For example, program.c ,pay.c , etc. Then the command

MSC pay.c

Would load the program stored in the file **pay.c** and generate the object code. This code is stored in another file under the name **pay.obj**. The linking is done by the command

LINK pay.obj

Which generates the executable code. with the file name **pay.exe**. Now the command would execute the program and give the results.

3.9 Summary

1. Every C program needs a main() function.
2. The execution of a function begins at the opening brace of the function and ends at the corresponding closing brace.
3. C programs are written in lowercase letters. Upper case letters are used for symbolic and output strings.
4. Every program statement must end with a semicolon.
5. All variables must be declared for their types before they are used in the program.
6. Include header files using # include directive for reference to special names and functions that it does not define. They should not end with a semicolon. The # sign must appear in the first column of the line.
7. When braces are used to group statements, the opening brace must have a corresponding closing brace,
8. A comment can be inserted anywhere to increase readability and understandability of the program. Comments help the users in testing and debugging. Care must be taken to match the symbols /* and */

Unit 6

Constants Variables and Data Types

Structure

4.1 Introduction

To communicate with a computer we have to speak a language which the computer understands since a computer speaks in bits bytes. . This means that, English or for that matter any other natural language by them cannot be used to perform the task of communication with computer. For this we have to have a language that is close to human language and far removed from machine language. A programming language is a methodical/systematic language designed to communicate instruction to a machine, especially to a computer and it can be used to create programs that control the behavior of a machine . However, learning C as a programming language is very much like learning English language. Learning English language begins with learning first of all the alphabets, then learning how to combine these alphabets to form words, combining words to form sentences, and finally learning to combine sentences to form paragraphs. On the same analogy , Learning C is not different. Instead of straight away learning how to write programs, we must incrementally learn (1) what alphabets, numbers, and special symbols are used in C, (2) how using these alphabets, numbers and special symbols, constants, variables and keywords are constructed, and (3) finally how these are combined to form an instruction and how groups of instructions are combined in accordance with " rules for sentence building" or syntax to form a program. The steps in learning C language is depicted below in the Figure 4.1 ,

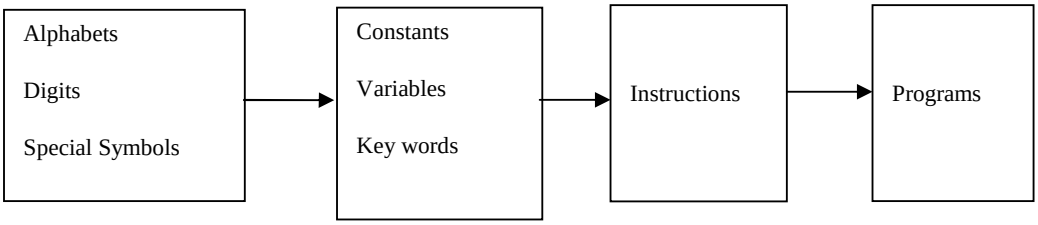


Figure 4.1: Steps in Learning C Language

As in any language, C language has its own grammar (*or syntax rules*) and each program instruction must conform precisely to the syntax of the language. In this chapter we will discuss the concepts of constants, variables and their types.

4.2 The C Character Set

A C character set denotes any valid alphabet, digit or special symbol, to represent an information. The set of characters that can be used to write a source program is called source character set and the set of characters available during program execution is called execution character set. Very often, in most implementations of C, both character sets are taken as identical. Generally, a character data type holds a single character(or one byte), enclosed with in single quotes, to represent a character constant. For e.g., the expressions 'a' , 'b',and '0' represent character constants. Remember that "a" is used to represent a string of characters(or sequence of characters enclosed with in double quotes) and is different from 'a'. Further, '\n' is used to represent a new line character, that is used to move the cursor to a new line on the screen. Figure 4.2 shows the entire character set (i.e., the valid alphabets, numbers, special characters and white spaces) allowed in C. The compiler ignores white spaces unless they are part of a string constant. White spaces may be used to separate words, and are prohibited between characters of key words and identifiers.

Trigraph Characters

Some characters from the C character set are not available in all environments, because keyboard may not have keys to cover the entire characters set of the language. A Trigraph, is a three character replacement for a special character in the C character set. ANSI C introduces the concept of "**Trigraph**" Sequences to provide a way to enter certain characters that are not available on some keyboards. Actually, each T**rigraph** sequence contains three characters (i.e., two question marks followed by another character) as in Figure 4.3. i.e., Each trigraph sequence is introduced by two question marks followed by a third character that indicates the character to be represented. For eg., , if a key board does not support square brackets , we can still use them in a program using the **Trigraphs ?**? (and ??).

Alphabets	Upper case letters A,B,……., Z Lower case letters a,b,…… .., z
Digits	All decimal digits 0,1,2,…….9
Special Characters	; semicolon , comma & ampersand . period * asterisk + plus sign ‘ apostrophe ? question mark < opening bracket or less than sign > closing bracket or greater than sign ^ caret ~ tilde ! exclamation mark | vertical bar (left parenthesis) right parenthesis \ backlash [left bracket] right bracket $ dollar sign } right brace _ under score { left brace = equal sign % percent sign # number sign / slash @ commercial at - hyphen or minus sign “ quotation mark
	White Spaces Blank spaces Horizontal Tab Carriage Return New Line

Figure 4.2 : The C Character Set

4.3 C Tokens

A **token** is a source program text that the compiler does not break down into atomic units. They are the basic building blocks/elements of the C language, constructed together to make a C program. That is, each and every smallest individual units in a C program are called **Tokens.** The **Tokens** in C language include:

1. Key words (eg: float, double etc.,)
2. Constants (eg: 100, -10.0 etc.,)
3. Strings (eg: “ABC”, “month” etc.,)
4. Operators (eg: +, - etc.,)
5. Identifiers (eg: main, total etc.,)
6. Special Symbols (eg: [ ],( ) etc.,)

C Programs are written using these **tokens** and the syntax of the language.

Trigraph Sequence	Translation
??=	# number sign
??(	[left bracket
??)	] right bracket
??<	{ left brace
?? >	} right brace
??!	| vertical bar
??/	\ back slash
??|	^ caret
??~	~ tilde

Fig. 4.3 ANSI C Trigraph Sequences

4.4 Key Words and Ident

Every C word fall under two categories, viz,. either a **key word** or an **Identifier**. C **Key words** (also called Reserved words) are the words that convey a special meaning to the C Compiler. They are the system defined **identifiers** that do have a fixed meaning (i.e., it does not change) and cannot be used as variable names. They are the basic building blocks for program statements and are written in lowercase letters. C language supports **32 (Thirty Two) keywords** and are listed in Figure 4.4.below.

auto	float	double	long
short	signed	unsigned	const
goto	else	switch	break
if	do	while	for
typedef	extern	static	struct
default	enum	return	sizeof
register	union	int	case
void	char	continue	volatile

Figure 4.4 Key words in C

An Identifier reters to the names of variables (i.e., the one which changes during program execution), names of functions, arrays, and structures. They are user defined names consisting of a combinations of alphabets, digits with a letter as the first character and underscore. The under score symbol is treated as a letter in the C character set and it helps in the readability of long variable names. That is, they are the names given to C entities such as , variables, types, functions, structures and labels in the program. However, the lengths of identifiers in C, vary from one implementation to another. In general, Identifier are created to give a unique name to C entities so as to identify it during the execution of the program. For example: **int** apple; Here apple is an identifier which denote a variable of ***integer*** type. In fact, **Keywords** (either C or Microsoft) are not used as **identifiers.**(i.e., they are reserved for special use). **Identifiers** are in general, used to name constants, functions, files and the like, apart from variables.

Rules for Identifiers

1. The first character must be an alphabet(uppercase or lowercase) or an under score.
2. All succeeding characters must be letters or digits.
3. Key words should not be used as identifiers.
4. Name of identifier is case sensitive i.e. num and Num are two different variables.
5. Identifier name cannot be exactly same as constant name which have been declared in the header file of C and you have included that header file.
6. Name of identifier cannot be exactly same as of name of function with in the scope of the function.
7. Name of function cannot be global identifier.
8. No two successive underscores are allowed.
9. Only first 31 characters are significant.
10. No special characters or punctuation symbols are used except the under score.

4.5 Constants

A **constant in C** refers to a piece of data that does not change throughout the execution of the program**. That is, C**onstants in C are expressions with a fixed value that are not changed during the execution of the program and are declared with the *define* keyword .In general, C constants can be divided into two major categories

1. Primary constants
2. Secondary constants.

These constants are further categorized as shown in Figure 4.5.

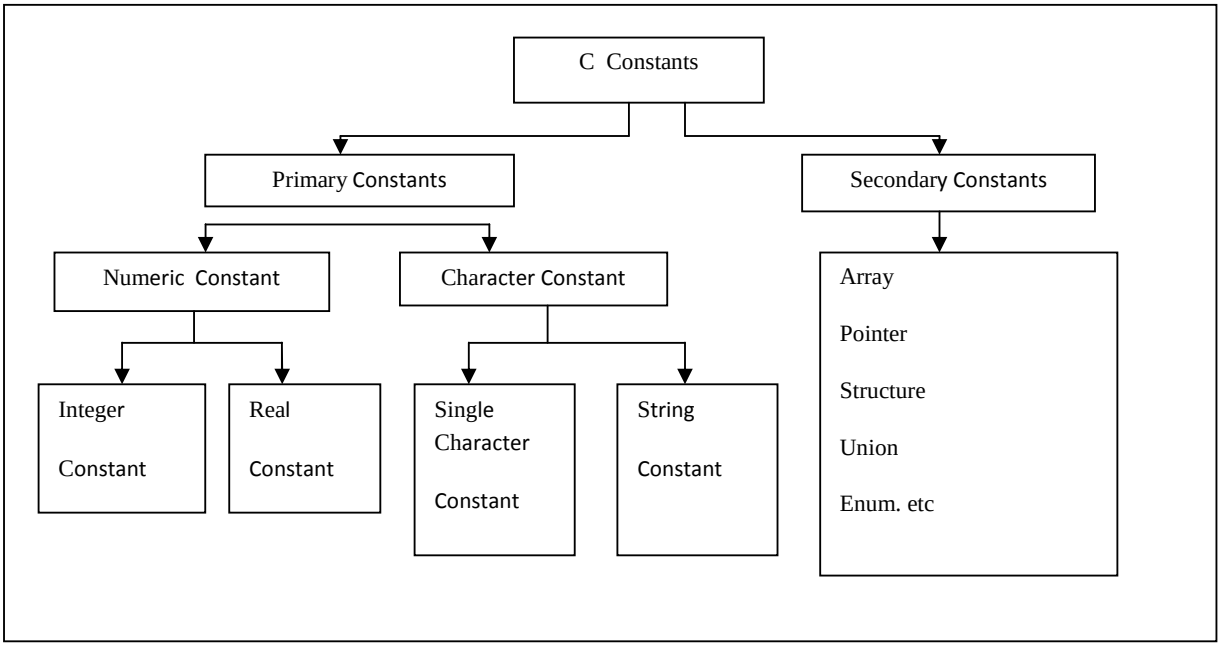


Fig. 4.5 Types of C Constants

At this stage, we would restrict our discussion to only primary constants(or basic constants) namely, Integer, Real and Character constants. Let us discuss the features of these constants..

Integer Constants

Integer constants are the numeric constants (Constants associated with number) without any fractional or exponential part. Integer constants take one of the following forms:

1. A **decimal integer.** , *e.g., 1 , 134, 10005 (* Decimal integers are a set of digits, 0 through 9, preceded by an optional – or + sign). Embedded spaces, commas, and non digit characters are not allowed between digits.
2. An **Octal integer constant** (base 8), e.g., *0 1 , 134, 0303242* . An octal constant is introduced

by a leading 0 and digits, the digits are 0 through 7 .

3. A **Hexa decimal** (base 16) Number. *e.g., 1 , 0x1, 0X186A2.* A hex constant is preceded by a leading 0X or 0x and the digits are 0 through 9 followed by A through F (Note that upper and lower case Letters are allowed) .

4. A character Constant.

Integer constants can also be suffixed with an identifier U (or u) or L (or l), which is used to indicate that the constant is unsigned or long, respectively. For e.g., 567U or 567u These suffixes may be combined *as in .e.g.,* 989712343UL or 989712343ul . The largest integer value that can be stored is machine dependent. It is 32767 on 16-bit and 2147483647 on 32-bit machines. For constructing the integer constants, certain rules have been laid down. These rules are as under:

Rules for constructing Integer constants

1.An integer constant must have at least one digit

2.It must not have a decimal point.

3.It can be either + ve or - ve.(If no sign precedes, it is assumed to be + ve.).

4. No Commas or Blanks are allowed within an integer constant.

5. The allowable range is between -32768 to 32767(For 16 bit compiler).

Real Constant

Certain quantities that vary continuously, such as prices, distances, temps, and so on, are represented by numbers containing fractional parts like 10.246. Such numbers are called **Real or Floating** point constants. That is, a real constant is one of :

- A fractional constant followed by an optional exponent
- A digit sequence followed by an exponent.

In either case followed by an optional of f, l (for single precision) , F. L(For double Precision), where:

- An optional digit sequence followed by a decimal point followed by a digit sequence.
- A digit sequence followed by a decimal point.

Further, an exponent is one of :

- E or e followed by an optional + or – followed by a **digit sequence** (A digit sequence is an arbitrary combination of one or more digits).

Floating point constants are normally represented as double precision quantities. Following rules must be observed while constructing real constants in fractional form:

1. A real constant must have at least one digit
2. It must have a decimal point
3. It could be either positive or negative
4. If no sign precedes an integer constant, it is assumed to be positive.
5. No commas or blanks are allowed within the real constant.

The exponential form of representation of real constants is usually used if the value of the constant is either too small or too large . In this form of representation, the real constant is represented in two parts. The part appearing before 'e' is called mantissa, whereas the part following 'e' is called exponent. Thus 0.000213 is represented in exponential form as 2.13e-4 . The General form is

mantissa e exponent

Following rules must be observed while constructing real constants expressed in exponential form:

1. The mantissa and exponential part should be separated by a letter e or E.

2. The mantissa part may have + ve or –ve sign.(default sign is positive).

3. The exponent must have at least one digit , which must be a +ve or _ve integer. Default sign is +ve.

4 .Range of real constants expressed in exponential form is -3.4e38 to 3.4e38.

Character Constant

Character constants are the constant which use single quotation around characters. example, `b`, `k`, `l` etc. In general, A character constant is a single alphabet, a single digit, or a single special symbol enclosed with in single quotes(or inverted commas). For both the inverted commas(single quotes) should point to the left. For example, 'C' is a valid character constant while ' C' is not. In C, characters are small integers, so you can use a character constant anywhere you can use an integer constant and *vice versa.* More over, the maximum length of a character constant can be 1 character.

String Constants

It is a collection of characters enclosed in double quotes. It may contain letters, digits, special characters and blank space. Examples are:

"Hello!" "How Are You " " ? " "X "

Note that a character constant (e.g., 'X') is not equal to the single character string constant(e.g., "X") . Further, a single character string constant does not have an equivalent integer value while a character constant has an integer value. More over, character strings are often used in programs to build meaningful programs. Moreover, the entity having two consecutive double quotes without any characters in between them, i.e., " ", is called a null string. Here, the quotes acts as delimiters and are not part of the string.

Backlash character constants

Sometimes, it is necessary to use newline(enter), tab, quotation mark etc. in the program which either cannot be typed or has special meaning in C programming. Such characters with special meaning should be preceded by a backlash symbol to make use of special function of them.. The backlash (\ ) causes "escape" from the normal way the characters are interpreted by the compiler. Each backlash character constant represents one character, although they consist of two characters. These character combinations are called escape sequences. Given below (Table 4.1)is the list of special characters and their purpose .

4.6 Variables

Every language should support the basic data objects namely, variables and constants. **Variables** are memory location in computers memory to store data. To indicate the memory location, each variable should be given a unique name called **identifier**. Variable names are just the symbolic representation of a memory location. These memory locations can contain integer, real or character constants. Unlike constants that remain unchanged during the execution of program , a variable may take different values at different times during execution. Examples of variable names are : sum, count, bike, interest etc. A variable name can be chosen by the programmer in a meaningful manner so as to reflect its function. Variables are to be declared before using it in the program.

Rules for writing Variable names in C

1. Variable names can be composed of letters(upper & lower case) , digits, and underscore. There is no rule for the length of a variable. A variable name is any combination of 1 to 31 alphabets.
2. The first letter of a variable should be either a letter or an under score. Note that upper and lower case are significant
3. No commas or blanks are allowed with in a variable name.
4. No special symbol other than underscore can be used in the variable name.
5. It should not be a key word.
6. White spaces are not allowed.

Constant	*Meaning*
‘\a’	audible alarm
‘\b’	back space
‘\f’	form feed
‘\n’	new line
‘\r’	carriage return
‘\t’	horizontal Tab
‘\v’	vertical tab
‘\”’	double quote
‘\’’	single quote
‘\?’	question mark
‘\\’	backlash
‘\0’	null

Table 4.1

4.7 Data Types.

Like other computer languages, **C** supports data types namely, of **integer, character and** of **float** type. In C, all variables must be declared before they are used, usually at the beginning of the function before an executable statements. A declaration announces the properties of variables; it consists of a type name and a list of variables such as

int Celsius;

int count;

The type **int** means that the variables listed are integers. ANSI C supports three classes of data types:

1. Primary data types
2. Derived data Types
3. User defined data Types.

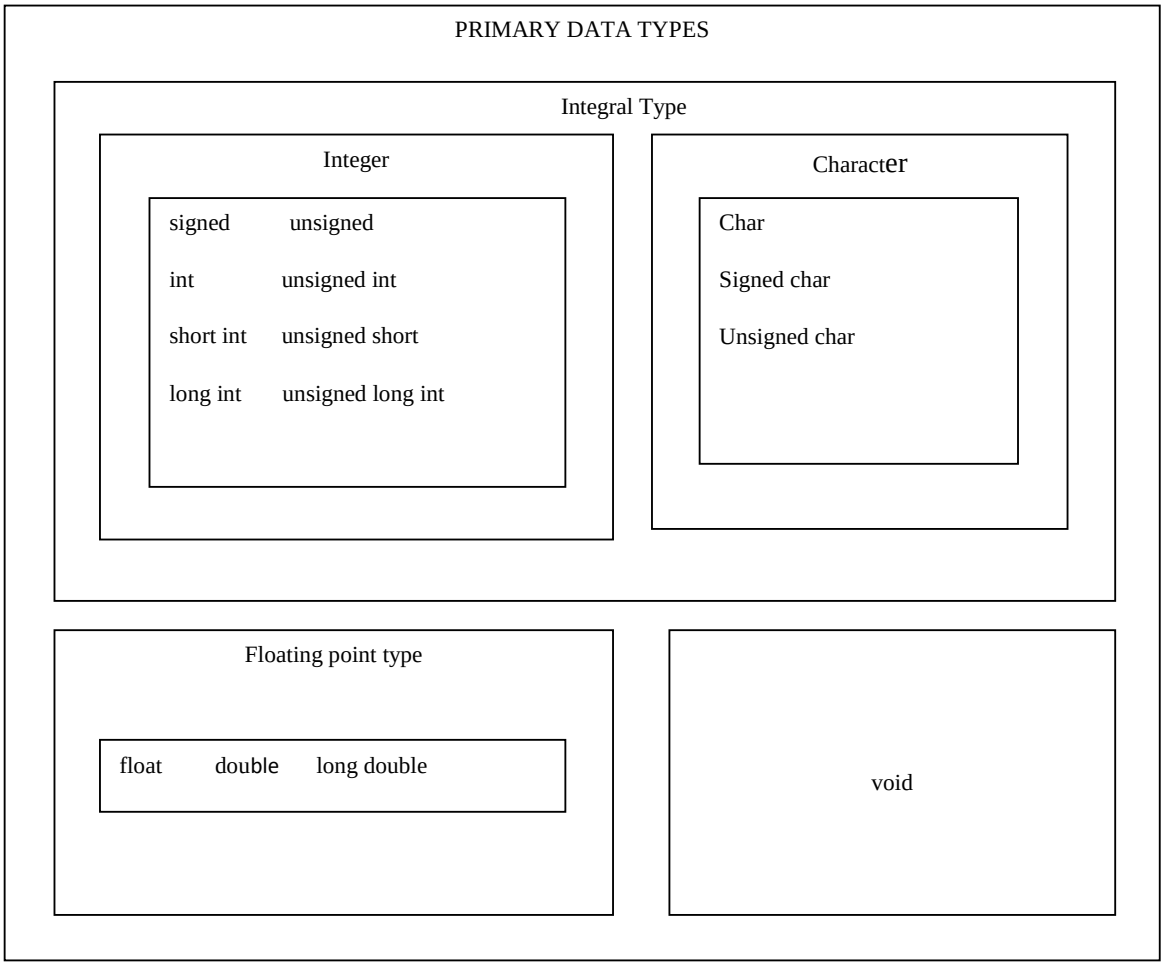


Fig. 4.6 Primary data types in C

All **C** Compilers support five fundamental data types, namely integer(**int**) , character(**char)**, Floating point(**float**), double precision floating point(**double)** and **void**. Extended data types like **long int ,long double** are also in use in C. Figure 4.6 gives an overview of primary data types in C.

Integer Types

This data type allows a variable to store numeric values. **int** keyword is used to refer integer data type. The. integers are whole numbers with a range of values supported by a particular machine (that is, the storage size of **int** data type is 2 or 4 or 8 byte. It varies with the processor in the CPU that we use). Generally, the C integer types were intended to allow code to be portable among machines with different inherent data sizes (word sizes), so each type may have different ranges on different machines. The problem with this is that a program often needs to be written for a particular

range of integers, and sometimes must be written for a particular size of storage, regardless of what machine the program runs on. In fact, integers occupy one word of storage, and since the word size of machines vary, the size of integer that can be stored depends on the computer. For a 16 bit word length, the size of the integer value is limited to the range -2^{15} to $2^{15}-1$. A signed integer uses one bit for sign and 15 bits for the magnitude of the number.

In order to provide control over the range of numbers and storage space, the C language defines several integer data types: **integer, short integer, long integer, and character, all both in signed and unsigned varieties.** For eg., **Short int** represents fairly small integer values and requires half the amount of storage space as a regular **int** number uses. Unlike **signed integers**, unsigned integers use all the bits for the magnitude of the number and are always positive. To increase the range of values we declare long and unsigned integers

Floating point types

C uses the key word **float** to define floating point numbers . Floating point numbers are stored in 32-bit, with six digits precision. Key word **double i**s used to define **big** floating point numbers. It reserves twice the storage for the number. A **doubl**e data type number uses 64 bits giving a precision of 14 digits. On PC's this is likely to be 8 bytes. The **double** type represents the same data type that **float** represents, but with a greater precision. To extend the precision further, the key word **long double** with 80 bits are used.

Void types

Void is an empty data type normally used as a return type in C to declare that no value will be returned by the function. It can also play the role of generic type, meaning that it can represent any of the other standard types.

Character type

A single character of the character set of C, can be defined as a **character (or char)** type data . Key word **char** is used for declaring the variable of character type. Usually, a character enclosed between a pair of single quotes denotes a character constant. The size of **char** is 1 byte(or 8 bits of internal storage)..The qualifier s**igne**d or **unsigned** may explicitly applied to **char.**

4.8 Declaration of Variables

In order to use a variable in C, we must first declare it before they are used in the program. Declaration does two things:

1. It tells the compiler what the variable name (type name) is
2. It specifies what type of data (or properties) the variable will hold

The type declaration statement is written at the beginning of **main ( )** function.

Primary type instruction

A variable can be used to hold a value of any data type in a memory location. After assigning variable names, we have to declare them. The syntax for declaring a variable is:

data-type v1,v2,….vn;

Here **v1,v2,….vn** are the variable names and are separated by commas A declaration statement must end with a semicolon. For example,

i**nt** num, sum**;**

int code**;**

double ratio;

are valid declarations. Here, Keywords **int** and **double** are used to represent integer and real type data respectively. When qualifier is applied to the data type then it changes its size (The size qualifiers are :**shor**t and **long**) or its sign (sign qualifiers are: **signed** and **unsigned).** While using qualifiers like, **short, long, unsigned** without specifying the basic data type , the **C** compiler will treat the data type as **int** . Moreover, if we want to declare a character variable as **unsigned**, then we must do so by using both the terms like **unsigned char**

User Defined Declaration

In C language, a user can define an identifier that represents an existing data type. The user defined data type identifier can later be used to declare variables. The General syntax is:

typedef **type identifier;**

Here *type* represents existing data type and "identifier" refers to the row name given to the data type.

Example:

typedef int amount;

typedef float sum;

Here amount symbolizes **int** and sum symbolizes **float**. They can be later used to declare variables as follows:

amount dep1,dept2**;**

sum section1[20],section2[20];

Therefore dept1 and dept2 are indirectly declared as integer data type and section1 and section 2 are indirectly float data type.

Another user defined data type is enumerated data type provided by ANSI C standard which is defined as follows:.

enum identifier { value1,value2,…..valuen};

The "identifier " here , is a user- defined enumerated data type which can be used to declare variables that can have one of the values enclosed with in the braces . After the definition we can declare variables to be of this 'new' type as below.

enum identifier v1,v2,…..vn;

The enumerated variables v1,v2,…vn can have only one of the values value1, value2 ….. value n.

Th assignments of the following type:

v1 = value3;

v5 = value1;

are valid.

For example:

enum day { Monday, Tuesday,…….,Sunday};

enum day week_ st,week_end;

week_ st = Monday;

week_end = Friday;

If (week_st = = Tuesday)

week_end = Saturday;

The C compiler automatically assign integer digits beginning with 0 to all the enumeration constants. That is, the enumeration constant value 1 is assigned 0, value 2 is assigned 1, and so on. The automatic assignment can be overridden if we assign enumeration constant values explicitly as;

enum day { Monday = 1 , Tuesday,…….,Sunday};

Here Monday is assigned the value 1.The remaining constants are assigned values that increase successively by 1.

The definition and declaration of enumerated variables can be combined in one statement as in :

enum day { Monday, Tuesday,…….,Sunday} week _st, week_end;

4.9 Declaration of Storage Class

C has a concept of "storage class" that defines the scope and life time of variables and/ or functions within a program. Storage class specifier helps to specify the type of storage used for data objects, C program uses the following storage classes specifiers:

- auto
- register
- static
- extern

In a declaration only one storage class specifiers is permitted, as there is only one way of storing things and if the storage class specifiers in a declaration is omitted then a default is chosen, depending on whether the declaration is made outside or inside the function. For external declarations the default storage class specifiers will be *extern* and for internal declaration it will be *auto.* It is the default storage class for all local variables. The variables with local life time are allocated new storage each time execution control passes to the block in which they are defined. When execution returns, the variables no longer have meaningful values,.

r**egister** is used to define local variable (or used for variables that need quick access-such as counters) that should be stored in a register instead of RAM. The variable declared as *registe*r is stored in the CPU register, the default value of that variable is the garbage value.. That is, the variable has a maximum size equal to the register size (usually one word) and cannot have unary '&' operator applied to it (as it does not have a memory location).The scope of the variable is local to the block in which it is defined (or it contains) and the variable is alive till the control remains with in the block in which the variable is defined..

static is the default storage class for global variables. The variable that is declared as static is stored in the memory, default value of which is zero. Life of variable persist between different function calls. The *static* storage class provides a life time over the entire duration of program and are not available to the linker. Therefore, another compilation unit can contain an identical declaration that refers to different object. A *static o*bject can be declared anywhere (or it does not have to be at the beginning of the block). s*tatic* variables may be initialized in their declarations; the initializes must be constant expressions, and it is done only once at compile time when memory is allocated for the static variable. Further, the scope of the *static automatic variables* is identical to that of automatic variables; however the storage allocated becomes permanent for the duration of the program.

The *extern storage c*lass is used to give reference of a global variable or function in another file, that is visible to all program files. It is the default class for objects with file scope .The variable

declared as extern is stored in the memory, the default value of that variable is being zero. Variable is alive as long as the program's execution does not come to an end . External variable can be declared outside all the functions or inside functions using 'e*xtern*' keywords. External variables may be declared outside any function block in a source code file the same way another variable is declared, by specifying the type and name(extern keyword may be omitted).Typically, when declared at the beginning of the source file, the *extern* key word is omitted. When you use *extern* the variable cannot be initialized as all it does is point the variable name at a storage location that has been previously defined. If the program is in several source files and the variable is defined in several files, collect *extern* declarations of variables and functions in separate header file then included by using # include when you have multiple files and you define a global variable function which will be used in another files also then *extern* will be used in another file to reference of defined variable or function.

The extern class specifies the same storage duration as static objects, but the object of function is not hidden from the linker .Using the *extern* key word in a declaration, results in external linkage and results in static duration of the object Memory for such variables is allocated when the program begins execution, and remains allocated until the program terminates. The storage class is another qualifier(like long and unsigned) that can be used in the variable declaration as given below:

```
auto int count;
register char ch;
static int y;
extern long sum;
```

The ***extern*** and ***static*** class variables are automatically initialized to zero. ***Auto*** variables , on the other hand contain undefined (or garbage)values unless they are initialized explicitly.

4.10 Assigning Values To Variables

Variables are used in program statements. Any variable used in the program must be declared before using it in any statement. In fact, the type declaration statement is written at the beginning of main( ) function. While all the variables are declared for their type, the variables that are used in expressions (on the right side of equal sign) must be assigned values before they are encountered in the program. First we will discuss the subtle variations of the type declarations as:

(a) While declaring the type variable we can also initialize it as:

```
int i = 5, j = 15;
float a = 1.2, b = 1.99;
```

(b) The order in which we define variable is sometimes important and sometimes not.

For e.g**., int** i = 10, j = 12; . is same as

i**nt** j= 12, i = 10.

However, **float** a= 1.5, b= a +3.2; is alright

But **floa**t b = a+ 3.2 ,a = 1.5 is not,

Because, here we are trying to use **a** even before defining it.

(c) The following statements work better

int a,b,c,d;

a = b = c = d = 10;

However the following statement would not work

int a = b = c = d = 10, an instance of using **b** (to assign to a) before defining it.

The Assignment statement

We can assign values to the variables using the assignment operator = as follows:

variable_name = constant;

Multiple assignments in one line are permitted in C. For eg.,

initial _value= 0; final value = 10; is a valid statement.

It is also possible to assign a value to variable at the time the variable is declared. This takes the following form:

data-type variable_name = constant;

More than one variable can be initialized in a single statement as:

a= b = c = 2;

x = y = z = MIN;

Note here that, **MIN** is a symbolic constant defined at the beginning.

Reading Data from Key board

There is a function in C, called the **scanf** function, which allows the programmer to accept input from the key board(or pass data to our C program). That is, Once executed our program will wait for the user inputs , once it came across any **scanf** function during program execution. It is a general input function available in **C** and is very similar in concept to the **printf** function. That is, **printf** and

scanf are two standard C programming language functions for console input and output. **scanf** works much like an INPUT statement in BASIC language. The syntax of **scanf** function is:

scanf("format string", &argument list);

The format string must be a text enclosed in **double quotes** and it contains the format of data being received for connecting it into internal representation in memory. e.g., integer (%d), float (%f), character (%c), or string (%s). The argument list contains a list of variables each preceded by the **address list** and separated by comma. The number of argument is not fixed. However corresponding to each argument there should be a format specifier. Inside the format string the number of argument should tally with the number of format specifier. For eg., if i is an integer and j a floating point number , to input these two numbers we may use s**canf("%d %f", &i, &j);.** The **& symbol** before each variable name is an **operator** that specify the variable name's address. We must always use this address. Let us look at an eg'.,

scanf("%d", &number);

when this statement is encountered by the computer, the execution stops and waits for the value of the **variable number** to be typed in. Since the control string **"%d**" specifies that it is an integer to be read from the terminal , we have to type in the value in the integer form. Once the number is typed in and the return key is pressed, the computer then proceeds to the next statement. The required header for the **scanf** function is **# include < stdio.h >.**

4.11 .Defining Symbolic Constants

A symbolic constant is a name that substitute for a sequence of characters (characters may be a numeric constant, a character constant, or a string corresponding to a character sequence) that cannot be changed..When the program is compiled, each occurrence of a symbolic constant is replaced by its corresponding character sequence compiled. They are usually defined at the beginning of the program. The symbolic constants may then appear later in the program in place of the numeric constants, character constants, etc, that the symbolic constants represent. The syntax of the Symbolic constant is:

#define symbolic- name value of constant

For example, consider a C program with the following symbolic constant definitions:

```
#define PI  3.141593
#define TRUE 1
#define FALSE 0
```

define PI 3.141593 defines a symbolic constant PI whose value is 3.141593.When the program is preprocessed, all the occurrences of the symbolic constant pi are replaced with the replacement text 3.141593. Here the preprocessor statements begin with # symbol. and are not end with a semi colon. By convention preprocessor constants are written in UPPER CASE. Further during run time, the value of a symbolic constant does not change. Symbolic names are sometimes called constant identifiers. Since symbolic names are constants, they do not appear in declarations.

Rules for Symbolic Constants

1. Symbolic names have the same form as variable names written in UPPER CASE.
2. No blank space between **'#'** and the word **define.**
3. **'#'** must be the first character in the line.
4. A blank space is required between **#define** and **symbolic name** and between **symbolic name** and the **constant.**
5. **#define** (**#define** is a preprocessor compiler directive) statements do not end with a semi colon.
6. After definition, the *symbolic name* should not be assigned any other value within the program by using an assignment statement.
7. *symbolic names* are not declared for data types. Its data type depends on the type of constant.
8. **#define** statements may appear anywhere in the program but before it is referenced in the program.

4.12 Declaring a Variable as Constant

In environments that support C, we may like the value of certain variables to remain constant during Program execution. We can achieve this by declaring the variable with the qualifier *const* at initialization as in e.g.,

const int *class_size* = 20;

The **const** is a new data type qualifier defined by ANSI C. This tells the compiler that the value of the **int** variable *class_size* must not be modified by the program. However, it can be used on the RHS of an assignment statement like any other variable.

4.13 Declaring a Variable as Volatile

Although we have phrased the discussion in terms of declaring a variable as constant , by far the most frequent use of another qualifier **volatile,** that could be used to tell explicitly the compiler that a variables value may be changed at any time by any external source is imminent. For example:

volatile int date;

This means that the value of **date** may be altered by some external factors even if it does not appear on the LHS of an assignment statement. When we declare a variable as **volatile,** the compiler will examine the value of the variable each time it is encountered to see whether any external alteration has changed the value.

If we wish that the value of a variable must not be modified by the program while it may be altered by some other process, then we may declare it as both **const** and **volatile** as :

volatile const int date = 12 ;

4.14 Summary

1. The three primary constants and variable types in C are **int, float** and **character**.
2. A variable name can be of maximum 31 character.
3. Do not use a key word as a variable name.
4. Each variable used must be declared for its type at the beginning of the program or function.
5. Each variable must be initialized before they are used in the program.
6. Integer constants, by default, assume int types. To make the numbers **long** or **unsigned** , append **L** or **U** to them.
7. Floating point default to **doubl**e To make them to denote **float** or **long double** , append letters F or L to the numbers.
8. Do not use l for long.
9. Use single quote for character constants and double quotes for string constants.
10. Do not combine declarations with executable statements.
11. A variable can be made constant either by using **#define** at the beginning of program or by declaring it with the qualifier **const** at the time of initialization.
12. **'#'** must be the first character in the line
13. No blank space between **'#'** and the word **define** is allowed.
14. A **variable** defined before the main function is available to all the functions in the functions in the program.
15. A **variable** defined inside a function is local to that function and not available to other functions.
16. Input/output in C can be achieved using **scanf** ( ) and **printf(**) functions.
17. No blank space are allowed within a variable, constant or keyword.

Unit 7

Operators And Expressions

Structure

Introduction:
Arithmetic Operators
Relational Operators
Logical Operators.
Assignment Operators.
Increment and Decrement operators.
Conditional Operator
Bitwise Operators
Special Operators
Arithmetic Expressions
Evaluation of Expression
Precedence of Arithmetic Operators
Some computational problems
Type conversion in expressions
Operator Precedence and associativity
Mathematical Functions
Summary

1.1 Introduction:

C language has a wide range of built –in operators to perform various operations. The symbols which are used to perform logical and mathematical operations in a C program are called operators. These C operators are used to join individual constants and variables to frame expressions. Moreover, operators, functions, constants and variables are combined to shape expressions. That is, operators are used with operands to build expressions. For example , the following is an expression containing two operands and one operator ' +' (an operator to perform addition).

8 + 5

whose value is 13. The value can be any type other than void. C offers the following operator Groups.

- Arithmetic
- Assignment
- Logical/relational
- Incremental and decrement operators
- Conditional
- Special Operators
- Bit wise operators.

1..2 Arithmetic Operators

The C arithmetic operators are the +, -, /, * and the modulo operator % . These C arithmetic operators are used to carry out mathematical calculations like addition, multiplication, division and modulus in c programs. Unlike /, which returns quotient, the % returns the reminder, the integer division truncates any fractional part. That is, the expression

x % y

produces the remainder when x is divided by y, and thus is zero when y divides x exactly. Note that the **operator ' % '** cannot be applied on floating point or double type data. Further, C does not have an operator for exponentiation. The operators in C with their meaning are listed in **Table 5.1** below.

Integer Arithmetic

When both the operands in a single arithmetic expression are integers, the expression is called an integer expression, and the operation is called integer arithmetic. Integer arithmetic always yields an integer value. For example, for integer operands such as **a** and **b** with assigned values respectively, 15 and 5, we have:

a + b = 20

a - b = 10

a * b = 75

a / b = 3

a % b = 0

During integer division , if both operands are of the same sign, the result is truncated to zero. If one of them is negative, the direction of truncation is machine dependant. .That is , 6/7 = 0 and -6/-7 = 0 but -6/7 may be zero or -1(that is , machine dependent).

Similarly, **during modulo operation, the sign of the result is sign of the first operand**., as in:

-16 % 3 = - 1

-16 % -3 = - 1

16 % - 3 = 1

Operator	**Meaning**
+	Addition(unary plus)
-	Subtraction(Unary minus)
*	Multiplication
/	Division
%	Modulo division (reminder after division)

Table 5.1 Arithmetic Operators

The Precedence to the operations associated with the operators are listed as:

Operator type	Precedence	priority
Unary Minus	1	Highest
*, / , %	2	Second
+, -	3	Third

That is, when an expression is given for evaluation, they are evaluated from Left to Right, based on the precedence associated with the operators. On the other hand, if the precedence's associated with the operators are to be overridden, it is necessary to use parenthesis in the expression. However, the expression within the parenthesis is evaluated on the basis of the precedence rule , with parentheses again evaluated from left to right. For expressions with nested parentheses, we evaluate the innermost one first, then the one immediately outside and so on.

Real Arithmetic

The C language contains the basic real arithmetic operators. An arithmetic operation involving only real operands is called real arithmetic. A real operand may accept values either in decimal or exponential form. An arithmetic operation between an integer and integer gives an integer result, while , the result of applying the real operators to real is another real. For floating point values, it is rounded to the number of significant digits permissible, and the final value is an approximation of the corrected result. For example, if operands x, y ,z are floats, then we will have,

x = 6.0/ 7.0 = 0.857143

y = 1.0/ 3.0 = 0.333333

z = -2.0 /3.0 = -0. 666667.

The operator % cannot be used with real operands

Mixed Mode Arithmetic

If operands in an expression contains both integer and real constants or variables then it is a mixed mode arithmetic expression. That is, When one of the operands is real, an operation between an integer and real always gives a **real** result. In this operation, the integer is first promoted to a real one and then operation is performed. The expression thus obtained is called a Mixed mode arithmetic expression. For e.g., 25/ 10.0 = 2.5 while, 25/10= 2.

1.3 Relational Operators

Relational operators are used to check relationship between two operands. If the relation is true, it returns value 1 and if the relation is false, it returns value zero. The relational operators are

>, > = , < , < =

They all have the same precedence. C offers six relational operators in all. These operators and their meanings are listed in Table 5.2.

Operator	**Meaning**
<	is less than
<=	is less than or equal to
>	is greater than
>=	is greater than or equal to
= =	is equal to
!=	is not equal to

Table 5.2 Relational Operators.

A simple relational expression contains only one relational operator . When arithmetic operations are used on either side of a relational operator, arithmetic expressions will be evaluated first and then the results are compared. Relational operators have lower precedence than arithmetic operators and are used in decision making and loops(i.e., in statements like If and while) in C programming..The Syntax Is:

ae-1 relational operator ae-2

with **ae-1 and ae-2** representing arithmetic expressions**.**

For e.g., 4.6 < = 10 TRUE

4.6 < - 10 FALSE

x+y = y+z TRUE only if sum of values of x and y are equal to the sum of values of y and z

Relational operator complements

Among the six relational operators, each one is complement of another operator. They are as:

- > is complement of < =
- < is complement of > =
- = = is complement of ! =

We can simplify an expression involving the not and less than operators using the complements as :

! (x < y) simplified to x > = y

! (x > y) simplified to x < = y

! (x ! = y) simplified to x= = y

! (x < = y) simplified to x > y

! (x > = y) simplified to x < y

! (x = = y) simplified to x !> = y

1.4 Logical Operators.

Logical operators are used to combine expressions containing relational operators. These operators perform logical operations on the given expressions .In C there are 3 logical operators (Table 5.3) and are:

Operator	Meaning of operator
&&	logical AND
||	logical OR
!	logical NOT

. **Table 5.3**

Logical operators perform logical-AND (&&) and logical –OR (| |) operations. Its Syntax is:

logical-AND-expression:

inclusive-OR- expression

logical –AND- expression & & inclusive- OR- expression

logical-OR-expression:

logical –AND- expression

logical -OR- expression | | logical - AND- expression

some example of usage of logical expression is:

1. If (age > 60 & & salary < 300 000)

2.If (number < 0 | | number > 1000) .

Logical operators & & and | | are used when we want to test more than one condition and to make decisions. They do not perform the usual arithmetic conversions. Instead, they evaluate each operand in terms of its equivalence to 0.The result of logical operation is either 0 or 1 and is of **int** type**.** The operands of logical-AND and logical-OR are evaluated from left to right. If the value of the first operand is sufficient to determine the result of the operation, the second operand is not evaluated . The C logical operators are described in Table 5.4 belo

Operator	**Description**
&&	If both operand are non zero logical AND produces the value 1.If either operand is equal to zero, the result is zero and if the first operand is equal to zero, the second operand is not evaluated.
| |	The logical-OR performs an inclusive - OR operation on its operands. The result is 0 if both operands have 0 values. If either operands has a non zero value, the second operand is not evaluated.

Table 5.4

While using compound expressions, care should be taken in using the precedence of relational and logical operators. The relative precedence are listed as:

! Highest

> > = < < =

= = ! =

& &

| | Lowest.

1. 5 Assignment Operators.

The assignment operators perform an arithmetic operation on the 1value and assign the result to the 1value.The usual assignment operator is the '=' ,. In addition, C has a set of less frequent *shorthand* assignment operators of the form (+ +, - =, * =, / =, % =). The syntax s;

v op = exp;

where ***v*** is a variable, ***exp*** is an expression and ***op*** is a C binary arithmetic operator.(or *short hand* binary operator). For e.g., consider the statement x + = y +1 ; this is same as x= x+(y+1). Here the operator + = means add ' y + 1 to x ' (or increment x by y + 1) . Some of the commonly used *short hand* assignment operators with their description is shown in Table 5.6. In all expressions involving these operators, the type of an assignment expression is the type of its left operand, and the value is the value after the assignment.

Statement with simple assignment operator	**Statement with assignment operator**
a = a + 1	a + = 1
a = a - 1	a - = 1
a = a* (n+1)	a* = n+1
a = a/(n+1)	a / = n+1
a = a % b	a % = b

Table 5.6. Short hand assignment operators

1.6 Increment and Decrement operators.

C provides two operators ++ and - - called increment and decrement operators and these operators are useful in controlling the loops through an index variable. The + + operator adds 1 to its operand while the decrement operator - - subtracts 1. Both of these operators are unary operators. (That is, used on single operand. ++ adds 1 to operand and - - subtracts 1 to operand respectively). For example:

Let a = 3 and b = 7

a ++ ; becomes 4 and a - - becomes 6

The unusual aspect is that ++ and - - may be used either as prefix (before the variable as in ++a) or post fix (after the variable as in a ++) . In both case effect is to increment a. But the expression ++a increments a before its value is used, while a ++ increments a after its value has been used. This means that in a context where the value is being used, not just the effect, + + a and a++ are different. For e.g., in the assignment statement x = i ++, if i =5, then x = i++ sets x= 5 , but x = ++ i sets x to 6. In both case i becomes 6. The increment and decrement operators can only be applied to variables, an expression like (i +j) ++ is illegal. In general, a prefix operator first adds 1 to the operand and then the result is assigned to the variable on the left. On the other hand, a post fix operator first assigns the value to the variable on left and then increments the operand.

Similar is the case, when we use ++ or - - in subscripted variable. That is, the statement

```
a[ i++ ] = 5;
```

Is equivalent to

```
a[i] =5;
i = i+1;
```

Rules for increment (++) and decrement (- -) operators.

1.They are unary operators and require variable as their operands.

2.A postfix ++ or - - operator used with a variable in an expression, the expression is evaluated first using the original value of the variable and then the variable is incremented(or decremented by one).

3 When prefix ++ or - - is used in an expression, the variable is incremented (or decremented) first and then the expression is evaluated using the new value of the variable.

4.The precedence and associativity of ++ and - - operators are the same as those of unary + and unary -

1.7 Conditional Operator

Conditional operator (? :) is a ternary operator (that demands three operands) consisting of symbols " **?"** and "**:** " and are used for decision making in C. The operator works by evaluating test expression, returning a value if that expression is TRUE and different one if the expression is evaluated as FALSE. The general syntax is:

identifier = (test expression) ? expression1 : expression2;

This is an expression, not a statement, so it represents a value. If the condition (or test expression) is true , it evaluates and returns expression1, otherwise it evaluates and returns expression2 .Conditional operator can be used as a short hand for some **if-else** statements. For example, consider the statements,

```
a = 10;
b = 20;
x = ( a > b) ? a : b;
```

Here in this example, x will be assigned the value of b. This can be achieved using the **if…..else** statement as follows:

```
If ( a > b)
    x = a ;
else
   x = b;
```

.1.8 Bitwise Operators

Bit wise operations in C are carried out by using operations on bits(or lowest form of data that can be accessed in digital hardware) at individual level. That means , Bit wise operators are used to perform bit operations on given two variables. Four commonly used bit wise operators in C are ~ , & ,| , and ^. Generally, Bitwise operators manipulate the value of individual bits(i.e., 1 or 0). Further, to understand "<< "and ">>" , there are two shift operators which are used to shift the position of a bit (or a set of bits) to another location, in a multi-bit value. Moreover, these operators work only on a limited number of types: **int** and **char.** That means**,** they may not be applied to data types : **float** and **double**. Bit wise operators supported by C are listed in the following **Table 5.7.**

Operator	Description of the operator
&	Binary AND operator copies a bit to the result if it exists in both operands(or Bitwise AND)
|	Binary OR operator copies a bit if it exists in either operand(or Bitwise Inclusive OR).
^	Binary XOR operator copies the bit if it is set in one operand but not both (or Bitwise Exclusive OR).
~	Binary Ones complement operator is unary and it has the effect of flipping bits(or Bitwise ones complement).
<<	Binary left shift operator(or bitwise left shift). The left operands value is moved left by the number of bits specified by the right operand.
>>	Binary right shift operator (or bitwise right shift). The left operands value is moved right by the number of bits specified by the right operand

Table 5.7 Bit wise operators

1.9 Special Operators

C language provides a number of special operators which have no counter parts in other languages. These operators include **comma** operator, **sizeof** operator, **pointer** operators(& and *) and member selection operator (. and -- >) . Pointer operators will be discussed while introducing pointers and member selection operators will be discussed with structures and union. The c**omma** and **sizeof** operators are discussed in this section.

The Comma Operator

This operator is used to link the related expressions together. A coma -linked list of expressions are evaluated left to right and the value of right most expression is the value of the combined expression. For example, the statement

```
int x, y,z;
z =  ( x =10, y = 20, x * y);
```

Here the 1st statement will create three integer type variables : x, y,z . In the 2nd statement, R.H.S will be evaluated first. As a result, 10 will be stored in variable x, then 20 will be stored in variable y and then values in x and y will be multiplied, result of which will be stored in the variable z as 200 at the end of the execution. Since comma operator has the lowest precedence of all operators, the use of parentheses are necessary.

The size of Operator

The **sizeof** operator works on variables, constants and even on data types. It returns the number of bytes, the operand occupies in the memory. It is a compile time operator and when used with an operand, it returns the number of bytes occupied by its operand on that particular machine.

Examples include:

```
m = sizeof (sum);
n =  sizeof( long int);
o =  sizeof ( 235L) ;
```

The **sizeof** operator is normally used to determine the lengths of arrays and structures when their sizes are not known to the programmer and is also used during program execution, for dynamic memory space allocation of variables.

1.10 Arithmetic Expressions

Arithmetic expressions have numbers and variables combined with the regular numeric operators (+ , - , *, /) , as per syntax of the language and simplify to a single number .Some of the examples of C expressions are (table 5.8) given below:

Algebraic Expression	**C Expression**
a×b-c	a*b-c
ab/c	a*b/c
ax^2+bx+c	a*x*x+b*x+c

Table 5.8 C Expressions

1.11 Evaluation of Expression

Every expression is formed out of operands and operators. Expressions in C, are evaluated using an assignment statement of the form:

variable = expression;

Usually when a statement is encountered, the expression (on the RHS) is first evaluated and the result obtained thus, is used to replaces the previous value of the variable on the LHS. All variables used in the expression must be assigned values before evaluation is attempted. An example of a valid evaluation expression is;

x = a* b-c;

Remember that blank space around an operator is optional and adds only to improve the readability..

1.12. Precedence of Arithmetic Operators

The two distinct priority levels of arithmetic operators in C are:

* / % High priority

+ - Low priority

An arithmetic operation without parentheses will be evaluated from *left to right*, using the rules of operator precedence. The basic evaluation procedure involves *two left to right pass* through the expression..During the 1st pass, high priority operators (if any) are applied. and during the 2nd pass low priority operators, if any , are applied as they are encountered. For example, consider the statement,

x = a-b/3 + c*2-1

when a= 9, b=12, and c =3 , the statement becomes

x = 9- 12/3 + 3*2 -1

1st pass

Step 1: x = 9- 4 + 3*2 -1

Step 2: x = 9-4+6-1

Second pass

Step 3: 5+6-1

Step 4: 11-1

Step 5: 10

However, one can change the order if evaluation, by introducing parentheses into the expression. The same above expression in parentheses reads as:

x = 9- 12/(3 + 3)*(2 -1)

Whenever parentheses are used, the expression contained in the left most set is evaluated first and the expression on the right most the last. The steps are as follows:

First pass:

Step 1: 9-12/6*(2-1)

Step 2: 9-12/6*1

Second Pass

Step 3: 9-2*1

Step 4: 9-2

Third pass

Step 5: 7

Though the procedure here, involves three left to right passes, number of evaluation steps is equal to the number of arithmetic operators. That is, the number of evaluation steps is same (equal to 5) for evaluation without and with parentheses

It may happen that parentheses may be nested, in which case evaluation will proceed outward from the inner most set of parentheses as in eg;, x = 9- (12/(3 + 3)*2) -1 = 4 .

Rules for evaluation of Expression

1. The arithmetic expressions are evaluated from left to right using the rules of precedence.
2. When parentheses are used , the expression with in the parentheses assume highest priority
3. First parenthesized sub expressions from left to right are evaluated.
4. The precedence rule is applied in determining the order of application of operators in evaluating sub expressions.
5. The associativity rule is applied when two or more operators of the same precedence level appear in a sub expression.
6. If parentheses are nested, the evaluation begins at the inner most sub expression

1.13 Some computational problems

On most computers, any attempt to divide a number by zero will result in an abnormal termination of the program. In such instances, care should be taken to test the denominator that is likely to assume zero value so that the division by zero error may be avoided. Further, one must specify the correct type of operands and it should be of the correct range, so that any error due to over flow / under flow may be eliminated.

1.14 Type conversion in expressions

C lets mixing of constants and variables of different types in an expression. It automatically, converts any intermediate values to the proper type so that expressions can be evaluated without loosing any significance. This automatic conversion is called *implicit type conversion*. If the operands are of different types, the lower type is automatically converted to the higher type before the operation proceeds. The result is of higher type. The sequence of rules to be followed while evaluating an expression are given below.

Rules for evaluating expressions

All s**hort** and **char** are automatically converted to **int:** then

1. If one of the operand is **long double**, the other will be converted to **long double** and the result will be **long double**.
2. else, if one of the operands is **double**, the other will be converted to **double** and the result will be **double**.
3. else, if the operand is **float**, the other will be converted to **float** and the result will be **float**;
4. else if one of the operand is **unsigned long int**, the other will be converted to u**nsigned long int** and the result will **be unsigned long int**.
5. else, if one of the operands is **long int** and the other is **unsigned int**, then

(a) If **unsigned int** can be converted to **long int**, the **unsigned int** operand will be converted as such and the result will be **long int;**

(b) else, both operands will be converted to **unsigned long int** and the result will be **unsigned long int;**

6. else, one of the operands is **long int**, the other will be converted **to long int** and the result will be **long int;**
7. else, if one of the operands is **unsigned int**, the other will be converted to **unsigned int** and the result will be **unsigned int**.

Explicit conversion

Explicit conversion is used to tell the compiler to treat a variable as of a different type in a specific context. The compiler will automatically change one type of data in to another (or locally convert) to make it sense. For instance, if you assign an integer value to a floating point variable, the compiler will insert code to convert the **int** to a **float**. The **general syntax** is:

(type-name)expression

Where type-name is one of the standard C data types. The expression may be a constant, variable or an expression. Casting allows you to make this type conversion explicit, or to force it when it would not normally happen. To perform casting, put the desired type including modifiers like unsigned inside parentheses to the left of the variable or constant you want to cast. For Example

```
float a = 5.25;
int b = (int)a;     /*Explicit casting from float to int */
```

The value of b here is 5.

1.15 Operator Precedence and associativity

Two operator characteristics (or precedence and associatively of operators) determines how operators group with operators. Precedence is the priority for grouping different types of operators with their operands. Associativity is the left to right or right to left order for grouping operand to operators that have the same precedence. An operator's precedence is meaningful only if other operators with higher to lower precedence are present. Expressions with higher-precedence operators are evaluated first. The grouping of operands can be forced by using parentheses Operators that have the same rank have the same precedence.

For example, in the following statements, the value of `1` is assigned to both `a` and `b` because of the right-to-left associativity of the `=` operator. The value of `c` is assigned to `b` first, and then the value of `b` is assigned to `a`.

```
b = 2;
c = 1;
a = b = c;
```

Because the order of sub expression evaluation is not specified, you can explicitly force the grouping of operands with operators by using parentheses.

In the expression

```
a + b * c / d
```

the `*` and `/` operations are performed before `+` because of precedence. `b` is multiplied by `c` before it is divided by `d` because of associativity. Table 5.8 gives a complete list of C operators, their precedence levels , and their rules of association.

Operator	Description	Associativity
( ) [ ]	Function call Array element reference	Left to right Right to Left
+	Unary plus	Right to left
-	Unary minus	
++	increment	
- -	decrement	
!	Logical negation	
~	Ones complement	
*	Pointer reference	
&	address	
Sizeof (type)	Size of an object Type cast	
*	multiplication	Left to right
/	division	
%	Modulo	

+	addition	Left to right
-	subtraction	
<<	Left shift	Left to right
>>	Right shift	
<	Less than	Left to right
< =	Less than or equal	
>	Greater than	
> =	Greater than or equal to	
= =	equality	Left to right
! =	In equality	
&	Bitwise AND	Left to right
^	Bitwise XOr	Left to right
|	Bitwise OR	Left to right
&&	Logical AND	Left to right
||	Logical Or	Left to right
?:	Conditional expression	Right to left
=	Assignment operators	Right to left
* = /= % =		
+ = - = & =		
^ = | =		
< < = > > =		
,	Comma operator	Left to right

Table 5.8 Precedence and Associativity of operators

1.16 Mathematical Functions

Mathematical functions such as cos, sqrt,log etc are frequently used in the analysis of real life problems. Most C compilers support these basic type functions. To use any of these functions in a program, we should include the line

include stdio.h.

In the beginning of the program. Table 5.9 shows some standard mathematical functions

1.17 Summary:

1. An operator in C is used with operands to build functions.
2. Each expression in C should end with a semicolon.
3.Associativity is applied when more than one operator of the same precedence are used in an expression.
4. All mathematical functions implement **double** type parameters and return double type values.
5. On either side of binary operator, always use spaces to increase readability.
6.Care should be taken to increment/decrement operators to floating point variables.
7.Assignment =. Operator should not be confused with equality operator = = .

Function	Meaning of function
Trigonometric	
acos(x)	arc cosine of x
asin(x)	arc sine of x
atan(x)	arc tangent of x
atan2(x,y)	arctangent of x/y.
cos(x)	cosine of x
sin(x)	sine of x.
tan(x)	tangent of x.
Hyperbolic	
cosh(x)	hyperbolic cosine of x.
sinh(x)	hyperbolic sine of x.
tanh(x)	hyperbolic tangent of x.
Other functions	
exp(x)	e to the power of x.
fabs(x)	absolute value of x.
floor(x)	x rounded down to the nearest integer.
fmod(x,y)	remainder of x/y.
log(x)	natural log of x, x>0.
pow(x,y)	x to the power y.
sqrt(x)	square root of x, x > = 0.

Fig 5.9 Mathematical Functions

Unit 8

Managing Input and output Operations

Structure

Introduction

Reading a Character

Writing a Character

Formatted Input

Formatted output

Summary

2.1 Introduction

In order to learn a program effectively in C language, one should know, how to manage input and output of data to and from the screen and the key board. Most programs take some data as input and display the processed data, often as results, on a suitable medium. The two methods so far used, for providing data to program variables, rely on : (1) Assigning values to variables through assignment statements and (2) using the input function **scanf (**to read data from a key board). For getting the output results, usually the **printf** function that sends results out to a terminal, is used.

The Input and output operations are convenient for program that interact with the user, takes input from the user and print the message. Unlike, other higher level languages, C does not provide any input-output (I/O) statements as part of its syntax. Instead , a set of library functions provided by the operating system for input and output operations are borrowed and used by C. The standard library for I/O operations used in C is **stdlib**. That is , Standard input (or **stdin**) is a data stream used to receive input from user / collects characters typed at the keyboard and **stdout, is** the data stream for sending output to a device such as monitor etc., . In otherwords, to include input and output functionality in C programs, the **stdio** header is needed. Each program that uses a standard I/O function must contain the statement

 # include < stdio.h >

at the beginning. This instruction tells the compiler, ‘to search for a file named **stdio.h** and place its contents at the appropriate place in the program . Indeed, the contents of the header file become part of the **source code** when it is compiled. In fact, this statement can be avoided in situations, where the functions **printf** and **scanf** have been defined as part of the C language. Here, in this chapter, a brief introduction of some common I/O function that can be used in many machines without much change is discussed.

2.2 Reading a Character

The simplest of all I/O operations is reading a character from the standard input unit(or key board) and writing it to the standard output unit(or the screen). The most basic way of reading input is by calling the function **getchar**. The C library function **getchar** gets a character from **stdin**, regardless of what it is, and

returns it to the program. That is, it is used to get a character from console, and echoes to the screen. It is the most basic input function in C, included in the **stdio.h** header file. The **getchar** takes the following form:

variable_name = getchar( );

Variable name is a valid C name that has been declared as of **char** type. When this statement is encountered, the computer waits until a key is pressed and then assigns this character as a value to g**etchar** function. Since **getchar** is used on the RHS of an assignment statement, the character value of **getcha**r is in turn assigned to the variable name on the left. For example,

char = name;

name **= getchar ( );**

Will assign the character "**a"** to the variable name when we press the key **a** on the keyboard. Since **getchar** is a function, it requires a set of parentheses as shown. The use of **getchar** function is illustrated in the program (Table 6.1) below..

Program	**Output**

#include <stdio.h>
#include<conio.h>
int main( )
{
char a;
clrscr( );
printf("Enter a character\n");
a=getchar( );
printf("The character entered is %c \n",a);
getchar( );
return 0;
}
	Enter a character
	b
	The character entered is b

Table 6.1: *use of* **getchar** *function*

The **getchar** function may be called successively to read the characters contained in a line of text..The following program me segment , for example, reads characters from key board one after another until the 'return key' is pressed

```
------------

------------

call  character;

character = '  ';

while ( character ! = '\n' )

   {

         character = getchar ( );

    }

-------------

-------------
```

The **getchar** returns the character it reads, or, if there are no more characters accessible, it will return the special value EOF ("end of file") .That is, The **getchar** function accepts any character keyed in, This includes TAB and RETURN . In other words, when we enter single character input, the newline character is waiting in the input queue after **getchar( )** returns. A dummy **getchar or fflush function (**to flush out unwanted functio**n)** may be used to get away the unwanted new line character , when we use **getchar** in a loop interactively. However, **getc** is used to accept a character from standard input.

2.3 Writing a Character

Often there do occur circumstances, where we want to solve computational problems and to display the characters therein on the console. The two special functions in C, that is designed to handle the output of character to monitor is ***putch*** and ***putchar* . That is,** Like **getchar**, there is an analogous companion C library function **putchar** that writes a single character to the standard output stream, (or console), specified by the argument **char** to s**tdout(i.e., it** is same as calling **putc(c,stdout).** The **putchar** function displays a single character on the screen. The syntax is:

putchar (variable_name);

where variable_name is a type **char** variable containing a character. **For e.g.,** the statement

```
answer = 'N'
putchar (answer);
```

will display the character N on the screen. The statement

putchar ('\n');

would cause the cursor on the screen to move to the beginning of the next line. The following example (Fig.6.1) explains the use of **putchar( )** function. ***Putch( )*** function, on the other hand is useful in writing the output, character by character, on the display.

The puts Function

The ***puts*** function stands for put string (or a bit of text) to the screen and this function works inside the main function. That means, the function puts( ) writes ***str*** to ***stdou***t, then writes a new line character. The general form of the function is:

int puts (char A [ ]);

A ***puts()*** function automatically appends a new line character at the end of any text it display and it uses a character array as parameter which is displayed on the screen. The ***puts()*** function performs a function that is similar to printf( ) with a %s conversion specifier (or formatted text display). However, **putc** is used for sending a single character to standard output.

```
# include <stdio.h>

int main ( )

 {

   char ch ;

   for (ch = 'A'; ch < = 'Z' ; ch++)  {

        putchar (ch);

   }

  return (0);

}
```

Output

ABCDEFGHIJKLMNOPQRSTUVWXYZ

Fig.6.1 Program to read and write all the letters in English alphabet

2.4 Formatted input

The standard formatted input function in C is **scanf (**that supply input in a fixed format**)** and is the input analog of **printf,** providing many of the conversion facilities in the opposite direction**..** The **scanf** contains two important things –the **format string** and the **address list** and it reads characters from the input file and converts them to internal form.. That is, **scanf** reads characters from the standard input, interprets them according to the specifications in format, and stores the results through the remaining arguments. Very often, This is the function used to read an input from the command line. The general format of an input statement is:

scanf(“ format string”, arg1,arg2,……, arg n);

Here the format string gives information to the computer on the type of data stored in the list of arguments arg1, arg2,….arg n and in how many columns (or address of locations) they are found. That is, format string specifies, how each input is read(.i.e., as a decimal integer, a decimal float, a character, a string and so on in matching arguments). The argument must be a pointer to a data type that is being read. In fact, format string and arguments are separated by commas.

scanf stops when it exhausts its format string, or when some input fails to match the control specification. It returns as its value the number of successfully matched and assigned input items. This can be used to decide how many items were found. On end of file, EOF is returned; note that this is different ' from 0, which means that the next input character does not match the first specification in the format string. The next call to **scanf** resumes searching immediately after the last character already converted. The format string usually contains conversion specifications, which are used to control conversion of input. The format string may contain:

- Blanks or tabs, which are ignored.
- Ordinary characters (not %), which are expected to match the next non-white space
- character of the input stream.
- Conversion specifications, consisting of the character %, an optional assignment suppression
- character *, an optional number specifying a maximum field width, an optional h, 1, or L indicating the width of the target, and a conversion character

A conversion specification directs the conversion of the next input field. Normally the result is placed in the variable pointed to by the corresponding argument. If assignment suppression is indicated by the * character, however, the input field is skipped; no assignment is made. An input field is defined as a string of non-white space characters; it extends either to the next white space character or until the field width, if specified, is exhausted. This implies that **scanf** will read across line boundaries to find its input, since newlines are white space

Inputting Integer l numbers

The field specification for reading an integer number is

% w sd

The percentage sign (%) indicates that a conversion specification follows.. **w** is an integer number specifying the field width of the number to be read and **d** the data type. For example, in the statement

scanf(“%3d %5d”, &num1,&num2);

the two variables in which numbers are to be stored are num1 and num2 and are of integer type. The input data items must be separated by spaces, tabs or new lines. A sample data line may thus be;

500 31246

The value 500 is assigned to num1 and 31246 to num2**.** Observe that the symbol & (ampersand) should precede each variable name, that is used to indicate the address of the variable name.

The **scanf** statement causes data to be read from one or more lines till numbers are stored in all the specified variable names. Also no blanks are permitted between characters in the format-string. The data type character d may be preceded by l to read long integers and h to read short integers.

Inputting real numbers

The **scanf** reads real numbers using the specification %f for both decimal and exponential notation. The input field specification may be separated by any arbitrary blank spaces. If the number to be read is of double type, then

Program	**Output**
main( )	values for x and y is : 12.3456 17.5e-2
{	x=12.345600
float x,y;	y=0.175000
double p,q;	
printf(“values for x and y is :\n”);	values of p and q is :4.142857142857
scanf(“%f %e” , &x ,&y);	18.5678901234567890
printf(“\n”);	p= 4.142857142857
printf(“x= %f\n y= %f\n\n”, x, y);	q= 1.8567890123456e+001
printf(“values of p and q is: ”);	
scanf(“%lf %lf ”, &p, &q);	
printf(“\n\np = % .12lf \np = %.12e”, p, q);	
}	

Table 6.2 : Reading of real numbers.

the specification should be %**lf.** Consider the statement

scanf(“%f %f %f”, &p,&q, ,&r) ;

with the data line

462.85 41.23E-1 543

It will assign the value 462.85 to p, 41.23E-1 to q and 543.0 to r. T he program (Table 6.2) below shows how to read real numbers in both decimal and exponential notation

Inputting character strings

A **scanf** function can input strings containing more than one character. The syntax is:

%ws or %wc

The corresponding arguments should be a pointer to character array. When the argument is a pointer to a ***char*** variable, then %c may be used to read a single character**.** Some **scanf** versions support the following string conversion specification.:

% [characters]

% [^ characters]

The specification **% [characters**] imply that only the characters within brackets are permissible in the input string. Any encounter of other string characters, will terminate the string. The specification % **[^characters]** does exactly the reverse. That is , characters after the ^ are not permitted in the input string, The reading of the string will be terminated at the encounter of one of these characters.

Reading Mixed data types

scanf can be used to input data containing mixed mode type. When one attempts to read an item that does not match the type , the **scanf** function does not read any further and immediately returns the value read. For e.g.,

scanf("%d %c %f", c %s " , &count, &code, &ratio, &name) ;

will read the data line

15 p 1.453 coffee

Correctly and assign values in the order in which they appear.

Rules for scanf

- Each variable to be read need a filed specification and a variable address of proper type.
- For any non -white space character used in the format string there must be a matching character in the user input.
- Ending the format string with white space will result in error.
- The **scanf** reads until:

1. A whitespace character is found in the numeric specification or
2. Maximum number of characters have been read
3. An error is detected.
4 .The EOF is reached

2.5 Formatted output

Formatted output refers to an output data that has been arranged in a particular format, using certain features, that are effectively exploited to control the alignment and spacing of print-outs on the terminals.. The main output routine is **printf** , which writes a formatted string to the **stdout** stream. The **printf( )** function is used to print the character, string, float, integer, octal and hexa decimal values on to the output screen and it returns the number of characters that was written if an error occurs, it will return a negative value. The required header for the **printf function** is:

#include <stdio.h>

The general form of **printf** statement is :

printf (" control string"' arg1,arg2,…., arg n);

Control string consists of three types:

1.character that will be printed on the screen as they appear.

2.format specification

3.escape sequence characters like, \n,\t, and \n.

The control string specifies the number of arguments (or variables whose values are formatted and printed according to the specification of control string) that follow with their types. The arguments should match in number, order and type with the format specification. A simple format specification is as:

% w. p type-specifier

Where w , is an integer specifying the total number of columns for output value and p is another integer that specifies the total number of digits to the right of the decimal point or the number of characters to be printed from a string.

Printf formatting is controlled by 'format identifiers' which in the simplest form are listed below:

%d % i	decimal signed integer.
% o	octal integer
%x % X	Hex integer
%u	unsigned integer
% c	character
%s	string
%f	double
%e %E	double
%p	pointer
%n	number of characters written by this printf, no argument expected
%%	% .No argument expected.

Output of Integer Numbers

The format specification for printing an integer number is:

% w d

Where **w** specifies the minimum field width for the output and d , the value to be printed as an integer. However, if a number (right justified in the given field width with leading blanks) is greater than the specified field width, it will be printed in full, over riding the minimum specification. It is possible to force the printing to be left- justified by placing a minus sign directly after the % character. More over, it is possible to pad with zeros the leading blanks by placing a zero before the field width specifier. Here, The minus (-) and zero (0) are named as flags. For printing short integers we may specify **hd** . And for printing long integers the specifier **ld** is used in place of **d** in the format specifier. Some examples of different format are:

Format	output
Printf(%d"', 1076)	1076
Printf(%6d"', 1076)	1076
Printf(%-6d"', 1076)	1076
Printf(%06d"', 1076)	001076

Output of Real Numbers:

Using the following form specification, the output of a real number may be displayed in decimal form:

% w.p f

The integer **w** indicates the number of positions that are to be used for the display of the value and the integer p represents the number of digits to be displayed after the decimal point. That is, the values when displayed, is rounded to p decimal places with right justification in the field of w columns, with leading trails and blanks. The default precision is actually 6 decimal places. The negative numbers will be printed with the minus sign and of the form [-] mmm-nnn.

A real number can be displayed in exponential form using the specification:

% w. p e

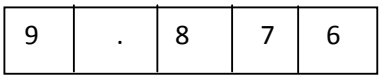
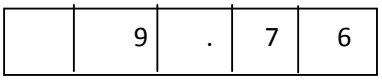
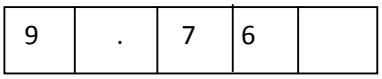
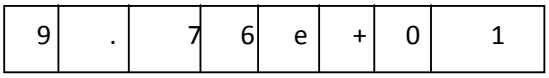
The display is of the form

[-] m.nnnne[±]xx

Where the length of the string n 's is specified by the precision p with the default precision being 6..Moreover, the field width w should satisfy the condition

$w \geq p + 7$

and will be rounded off and printed right justified in the field of w columns. Further, padding the leading blanks with zeros and printing with left justification using flags 0 or – before the field specifier is also possible. Following are some examples:

Format	output
Printf("%5.3f",x)	
Printf(%5.2f"',x)	
Printf(%-5.2f"',x)	
Printf(% -8.2e"',x)	

For dynamic format specification during run time (i.e., with field width and precision given as arguments for w and p) we have the special field specification:

printf("%*.*f" , width, precision, number);

For e.g.,

printf('%*.*f", 7,2, number);

Is equivalent to

printf('%7.2f", number);

Printing of a single character

A single character can be displayed in the keyboard at the desired position , right justified in the field of w column (with default value for w being 1) using the format

% wc

Printing of strings

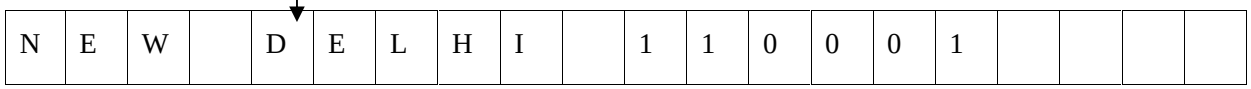
The format specification for outputting strings is similar to that of real numbers.. The format being:

% w. ps

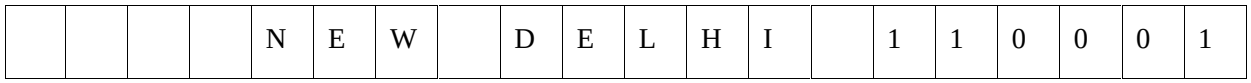
With w the field width for display and p indicates that only first p characters of the string are to be displayed with right justification..Some examples are:

Table showing specification and out put

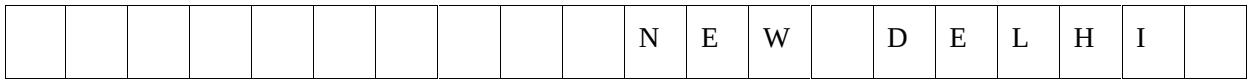
%s (specification) output



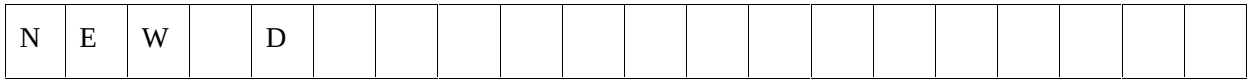
%20s(specification) output



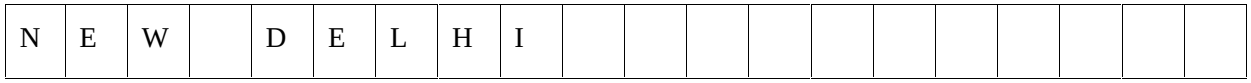
% 20.10s(specification) output



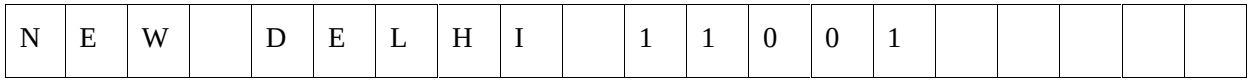
%.5s(specification) output



%-20.10s(specification) output



%5s(specification) output



Mixed data output

Mixed data types in one printf statement is permitted in C. For e.g.,

p**rintf("%d % f % s %c ,a,b,c,d);** is a valid one.

code Meaning

%c	Print a single character	
%d	Print a decimal number	
%e	Print a floating point number in exponent form	
%f	Print a floating point number	Without
%g		exponent form
%i	Print a floating point number	Either e-
% o	type	
%s		or f-
%u	type	
%x	Print a signed decimal integer	
	Print an octal integer without leading zero.	
	Print a string	
	Print an unsigned decimal integer	
	Print a hexagonal integer, without leading 0.s	

Table 6.1 printf format codes

Remember that, the format specification should match the variables in number, order and type. Table 6.1 below shows commonly used **printf** format codes

The letters used as prefix for certain conversion characters are:

h short integer

l long or double

L for long double .

2.6 Summary

1. While using **getchar**, clear all unwanted characters on the console.
2. While using I/O functions always use the header < stdio.h >.
3. For functions that use character handling use the header< ctype.h>
4. For any variable to be read or printed, the proper field specification is to be done.
5. Always enclose format control strings in double quotes.
6. While using scanf the address specifier & ampersand is to be used.
7 Single character constants are to be enclosed in single quotes.
8. Avoid white space at the end of format string and use comma after he format string in **scanf** statements.
9. Do not use commas in the format string of a **scanf** statement.

Unit 9

This unit is designed as an introduction to control structures: branching and looping. Programming languages by default execute in sequence, line by line. This is very useful since in this mode of execution, it is done in an orderly manner . But if we need to make decisions and evaluate some input and decide which path to take depending on that input then we use Control Structures. Control Structures allow programmers, to change that default sequential execution. In most Programming Languages such as C, PHP, C++, C#, Java, JavaScript, and others, we have Control Structures. The first Control Structure we are going to talk about is the "if" and "if … else". What this structure does is to evaluate the condition of the "if" statement and determine if it's true or false; then if it's true executes the statements inside the "if" body, otherwise executes statements in the else body or continues executing the rest of the program. Actually, the flow of control in a computer program may be altered in two ways. One involves alternate paths provided by if…else or switch statements; the other is through the repetitive execution of a set of instructions. The first mechanism is called branching, the second called looping. Branching is deciding what actions to take and looping is deciding how many times to take a certain action. In the first unit of the module, you are guided through the structure of the various branching constructs like, if…else, else…if, switch etc., with sample programs. The next unit is a tour through the control structure through looping: viz, while, do…while, for(,,) loop ,and continue statement.

Decision Making And Branching

Structure

1.1 Introduction.

Decision making is one of the most important concepts in C programming. That is, the programs should be able to make logical decisions based on the conditions they are in. **C** language has three major decision making instructions- the **if** statement, the **if else** statement, and the **switch** statement. These statements 'control' the flow of program execution (or they specify the order in which computations are performed), and are known as **control statements**. Here we will learn each of these, and discuss their features, capabilities and applications in more detail.

1.2 Decision Making with if statement

The key word, **if** statement, is a conditional branching statement. **It,** instructs the compiler that, what follows is a decision control instruction. That is, it allows the program to select an action (i.e., a condition is evaluated, and if it is true the statement is executed, and, the program skips past it if it is found false) based upon the user's input. The condition following the keyword if is always enclosed within a pair of parenthesis. It takes the form:

If (test expression)

A decision control instruction can be implemented in C using (1) The simple if statement, (2) The if – else statement (3) nested if-else statement and (4) else if ladder.

1.3 The Simple If Statement

The general form of **if** statement looks as:

```
if (test expression)
{
    statement block;
 }
 statement –x;
```

Here the expression can be any valid expression including a relational expression. We can even use arithmetic expressions in the **if** statement. In fact a compound statement composed of several statements enclosed with in braces (braces are used to group declarations and statements together into a compound statement or block), can replace the single statement. Remember, there is no semicolon after the right brace that ends a block. If the test expression evaluates to true, then the compound statement is executed. Otherwise the control jumps to the statement following the right brace ignoring the compound statement.. ***Please do remember that in C, a non zero value is considered to be true, where as a zero is considered to be false.*** Here is a simple program (Figure 7.1) using simple ***if statement***:

```
/* Demonstration of  if  statement*/
# include < stdio.h >
# include < conio.h>
int main ( )
 {
   int number;
   clrscr ( );
   printf ( " enter a number\n");
   scanf(" %d", &number);
   If (number > 0)
        printf(" The given number is positive\n");
   getch( );
   return 0;
}
```

output

```
enter a number
5
The given number is positive
```

Fig.7.1 program for illustration of simple if statement

On execution of this program, if you type a number greater than zero, you will get a message on the

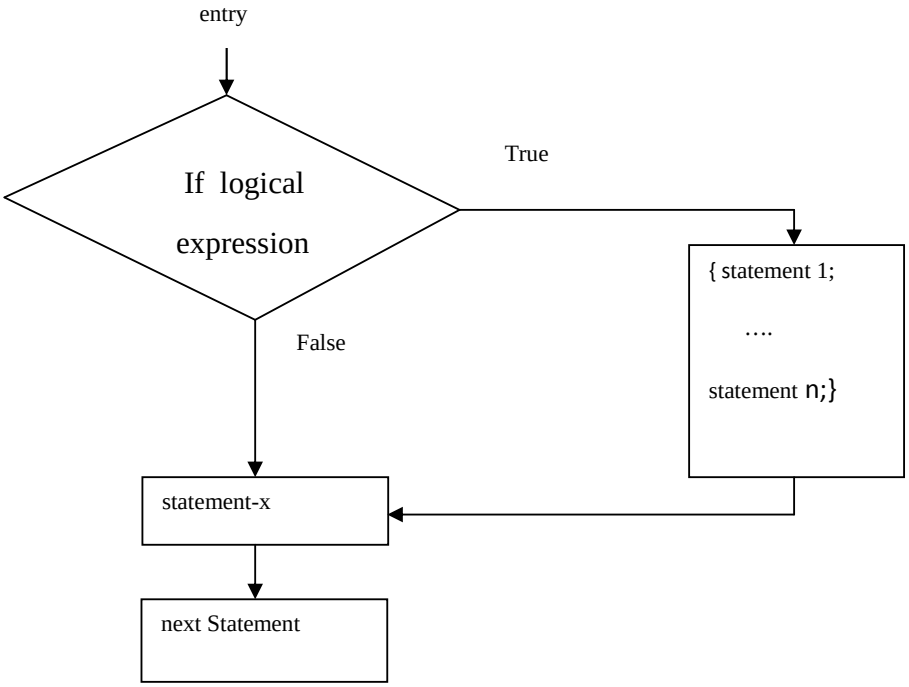


Fig.7.2 Flow chart I illustrating simple **If conditional** statement

screen through **printf( ).** If you type some other number(i.e., a number less than 0, the program does not do anything. The Flow chart given in Fig. 7.2 help you understand the flow of control in **simple if** statement.

1.4 The IF…..ELSE Statement.

The **if** statement by itself will execute a group of statements or a single statement, when the expression following it evaluates to **true** and it does nothing when it evaluates to f**als**e .In fact, the **if –else** statement is an extension of the simple **if** statement and is used to express decisions. It permits the programmer to write a single comparison, and then execute one of the two statements depending on whether the test expression (in parentheses) is true or false. That is, the **if…else** statement is used, the intention of the programmer is- to execute the group of statements denoted as true (.i.e., the true block of statements immediately following the **if** statements), or else the test expression statements denoted as false are executed..In either case, either a true or a false block of codes/statements, are executed not both .In both cases, control is transferred to the subsequent statement-x. This is interpreted in the flow chart of Fig.7.3.

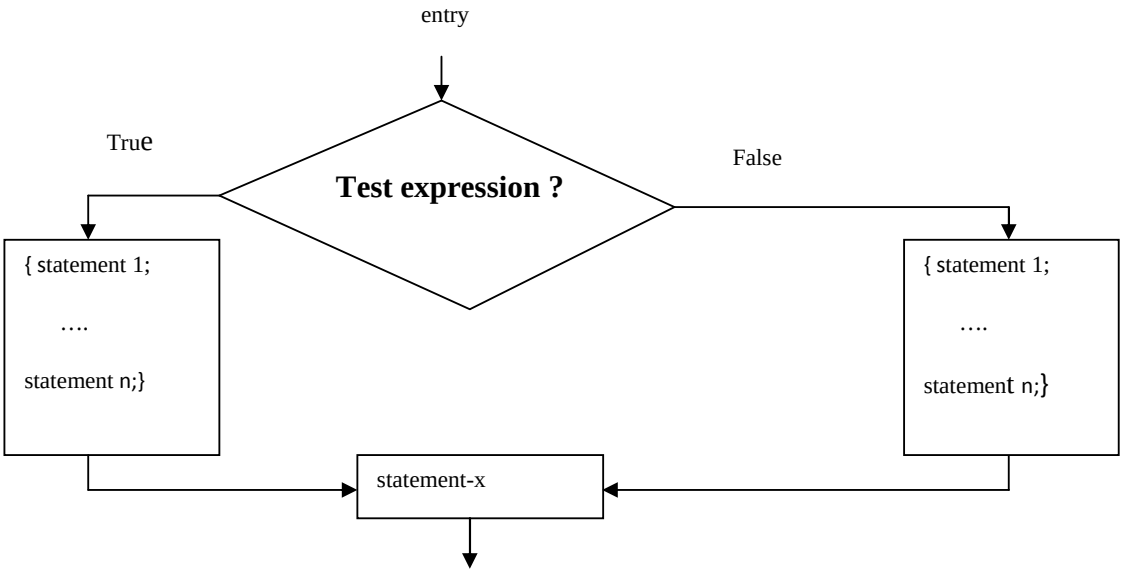


Fig.7.3 Flow chart I illustrating simple If **–else conditional** statement

Example 7.1: A program to check whether the number is odd or even?

```
# include < stdio.h >

int main ( ) {

      int number;

      printf(“ Enter a number.\n”);

      scanf(“%d”, &number);

      if ((number  % 2) = = 0)

            printf(“%d is even,” , number);

     else

            printf(“%d is odd..” , number);

           return 0;

}
```

Output

```
Enter  a number
22
22 is even.
```

Fig.7.4 A program to illustrate the If ….else statement

There are a few points that deserve worth mentioning:

1.The group of statements after the if up to and not including the else is the ‘ if block’. Similarly, the statements after the else form the ‘else block’.

2.The statements in the if and those in the else block have been indented to the right.

3. As with the if statement, the default scope of else is also the statement immediately after the else. In order to override this default scope, a pair of braces must be used.

1.5 Nested *If-else* statements.

The if….else statement can be used in nested form when a serious decision are involved. In nested if ..else construct, we write an entire if-else construct with in either the body of the if statement or the body of an else statement. The logic of execution is shown in Fig.7.5.The syntax is:

```
if  (test condition-1)
     {
       if  (test condition-2);
       {
         statement-1;
       }
       else
        {
         statement-2;
        }
   }
 else
  {
    statement-3;
  }
  statement-x;
```

Here, if the test expression -1 is false, the statement -3 will be executed; otherwise control of the program jumps to perform the second test condition. If the condition- 2 is true, the statement-1 will be evaluated, otherwise the statement-2 will be evaluated and then the control is transferred to the statement-x.

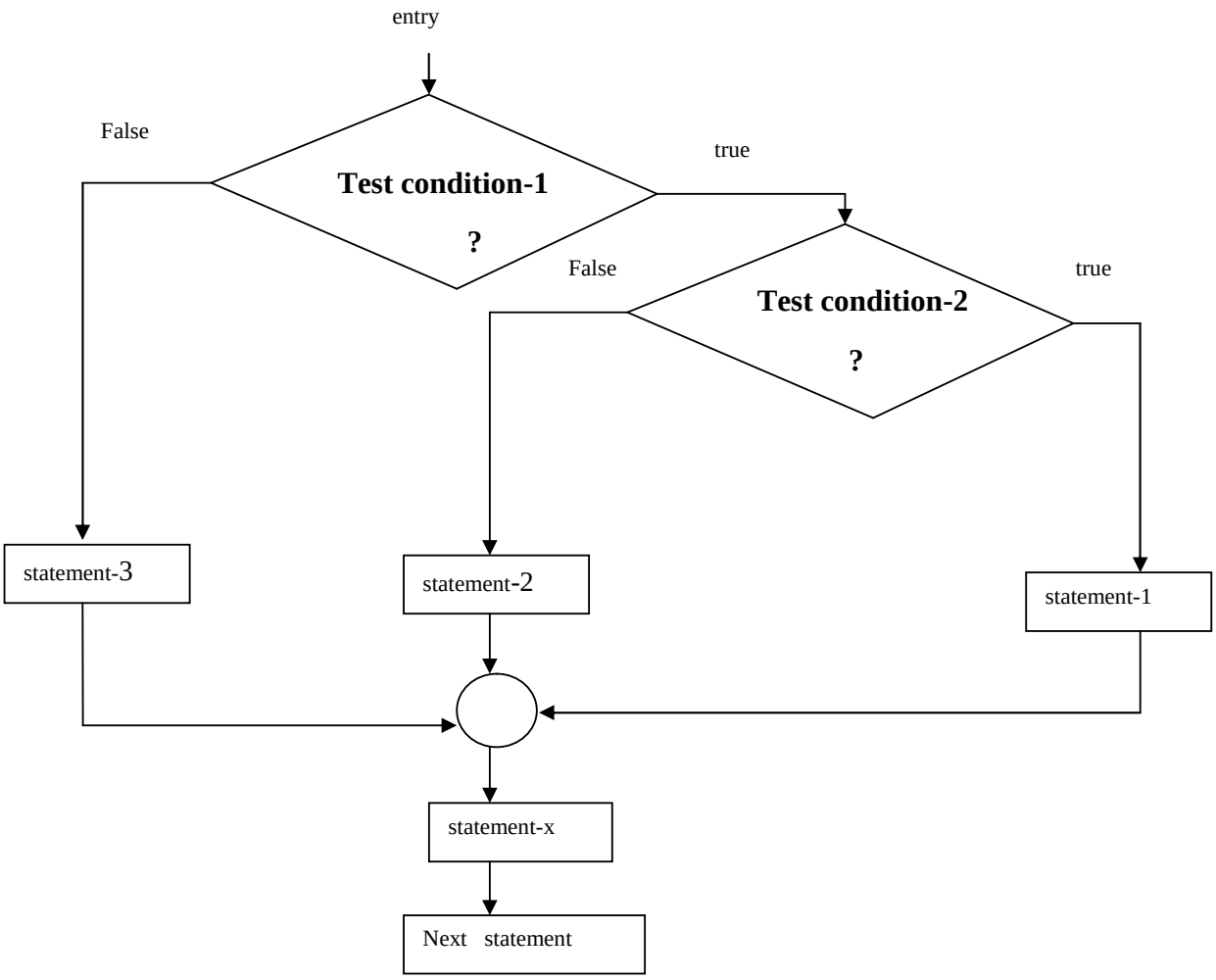


Fig.7.3 Flow chart I illustrating nested **If –else** statement

Example 7.2: A program to check whether the two numbers is <, than or > than or equal.

```
# include < stdio.h >

int main ( ) {

      int num1, num2;

      printf(“ Enter two integers.”,\n);

      scanf(“%d %d”; & num1, &num2);

      if (num1= = num2)

           printf( result: %d=%d”, num1,num2);

     else

          if(num1> num2)

               printf(“result:%d > %d”, num1,num2);

           else

               print(“result: %d >%d “,num2,num1);

 return 0;

}
```

Output

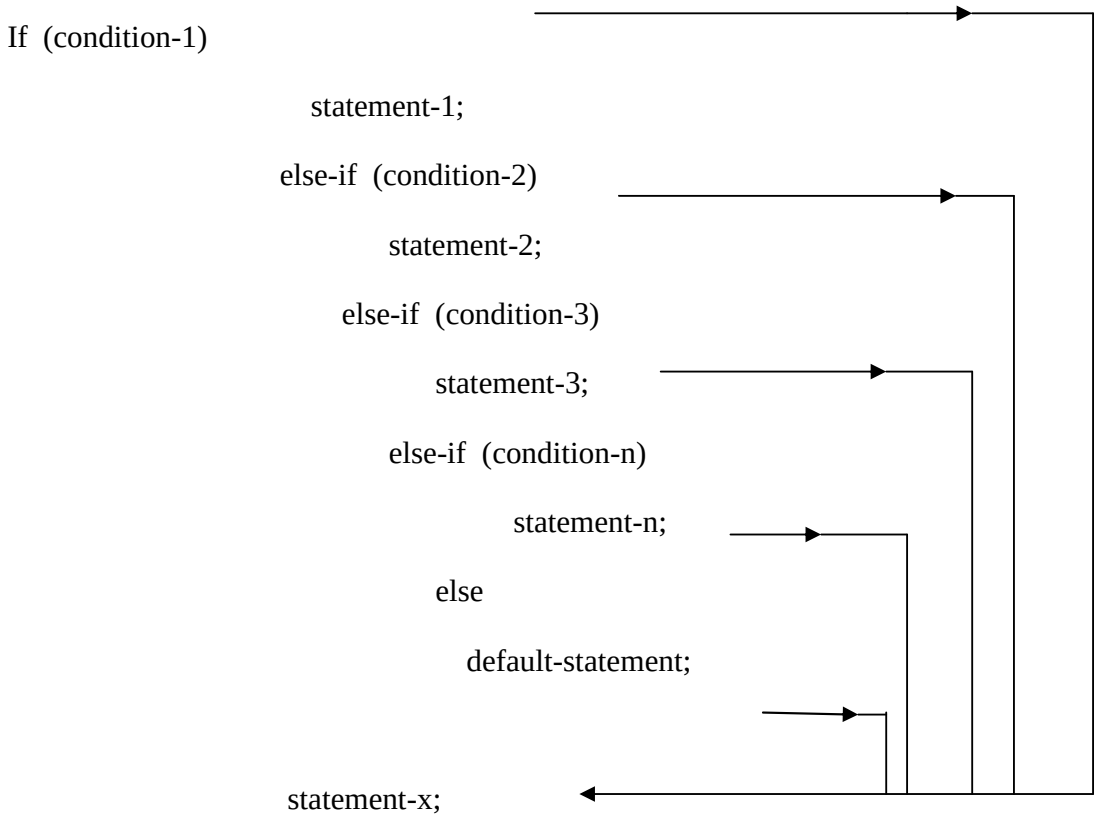
```
Enter two integers

4

2

Result:4>2
```

Fig.7.7: program illustrating **nested if -else**

1.6 The else -If Ladder

Another way of describing the nested **if-else** is the **else-if** ladder, where, every **else** is associated with an **if** statement. That is, **else-if**, is a combination of **if** and **else**. Like **else**, it extends an **if** statement to execute a different statement in case the original **if** expression is evaluated as False.

The syntax is:



This construct is called the **else-if ladder** and is useful where two or more alternatives are available for selection. In **else-if** ladder various conditions are evaluated one by one starting from top to bottom, on reaching a condition evaluating to TRUE the statement group associated with it are executed and skip other statements. If none of the expressions is evaluated to true, then the statement or group of statements associated with the final **else** is executed. In this construct nesting is allowed only in the **else** part . In fact, In **else……if** ladder, we do not have to pair **if** statements with **else** statements. That is, there is no need to remember the number of braces opened as in nested **if….else.** Moreover, **else….if** ladder produces the same effect as **nested if-else** with the benefit that it is easy to code. The flow chart corresponding to **else-if** ladder is shown in fig.7.8

In this construct, the conditions are checked, starting from the top of the **else-if** ladder, moving downwards. That is, firstly, condition-1 is checked, and if it is true, statement-1 is executed and control is transferred to statement-x. On the other hand, **If** condition-1 is false, condition-2 is checked and if true, statement -2 is executed and control is transferred to statement-x skipping the rest of the ladder .When all the n conditions are false, then the final default-statement is executed followed by the execution of statement-x. The following program(Fig.7.9) explains the **else-if** construct.

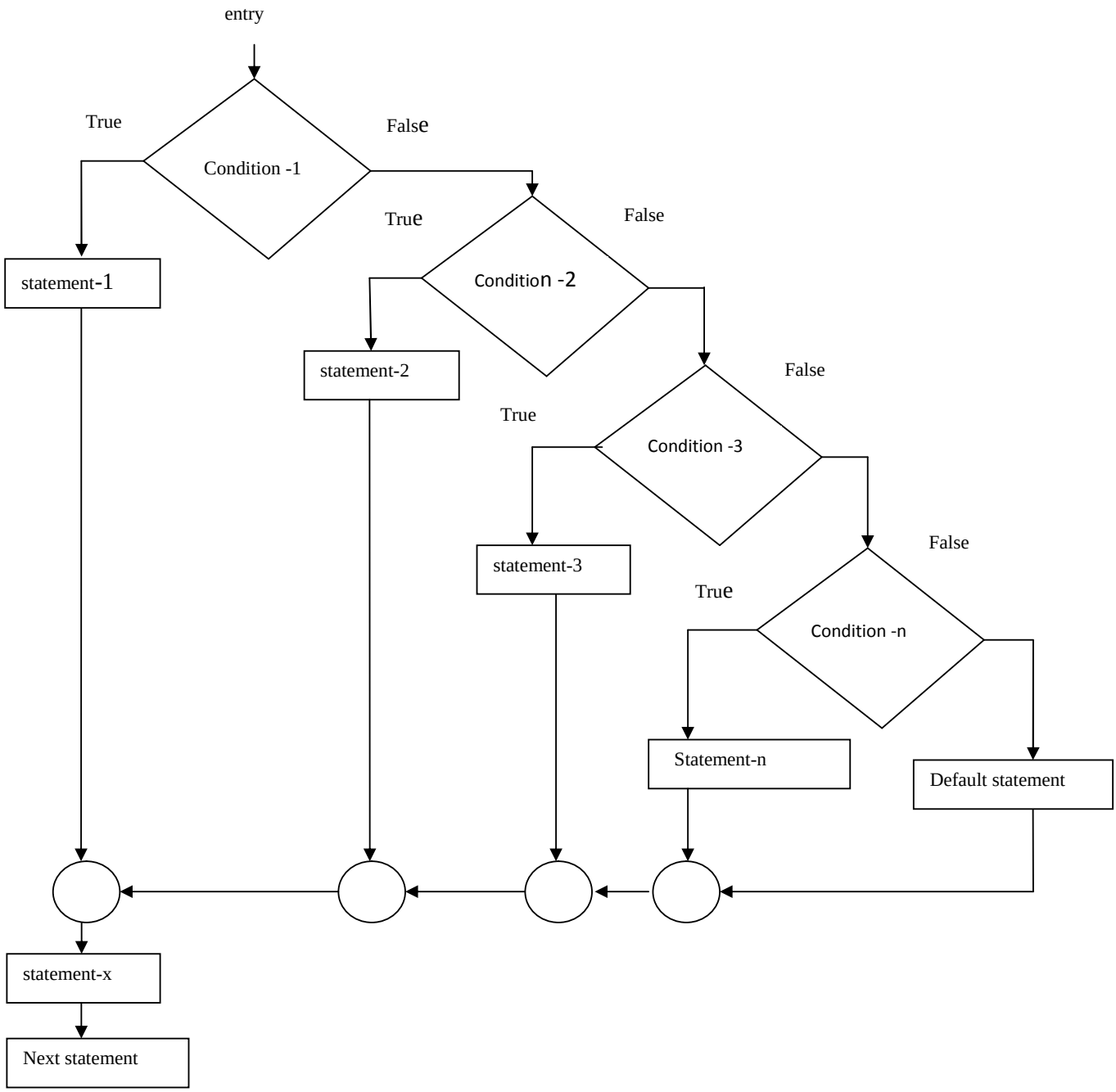


Fig.7.8 Flow chart I illustrating else- **If ladder**

```
#include < stdio.h>
#include <conio.h >
void main ( )
 {
   int num;
   clrscr( );
   printf("enter a number.\n");
   scanf("%d", &num);
   If( num = =0)
     Printf("Given number is Zero.\n");
   else if (number > 0)
     printf("Given number is positive.\n");
   else
     printf("Given number is negative.\n");
   getch ( );
  }
```

Output

```
Enter a number.
5
Given number is positive.
```

Fig.7.9. program for **else if** ladder demonstration.

Rules for Indentati

The sections of this page cover the guidelines of acceptable code indentation. Indentation is important for clarity and sticking to standard. The guidelines that are to be followed while using indentation , for control statements are listed below:

1. Indent statements that are dependent on the previous statements; provide at least three spaces of indentation.
2.Align vertically else clause with their matching **if** clause.
3.Use braces on separate lines to identify a block of elements.
4.Indent the statements in the block by at least three spaces to the right of the braces.
5.Align the opening and closing braces.
6. Indent the nested statements as per the above rules.
7. Code only one statement/clause on each line.

1.7 The Switch Statement

The switch statement is much like a nested **if** statement and it allows us to make a decision from a number of choices. In fact, it is a powerful decision making statement that allows a variable to be tested for equality against a list of values. The condition of a **switch** statement is a value. The **case** says that if it has the value of whatever is after that **case** then do whatever follows the colon. That is,.each value is called a **case,** and the variable being switched on is checked for each **switch case**. More correctly, a **switch-case default** (since these keywords go together to make up the control statement) accepts single input from the user and based on that input executes a particular block of statements. The break is used to **break** out of the case statements, and is usually surrounded by braces, which it is in. The syntax is:

```
switch   (integer expression)
 {
    case value-1;
                block-1
                break;
   case value-2;
                block-2
                break;
   ……
   …….
   default:
                default-block
                break;
    }
statement-x;
```

The integer expression following the key word **switch i**s any C expression that yields an integer value. It could be an integer constant or an expression that evaluates to an integer. The keyword **case** is followed by an integer or a character constant. Each constant in each **case** must be different from all the others. When the **switch** is executed, the value of the expression is compared against the values value-1,value-2,…When a match is found, the program executes the statements following that case, and all subsequent case and default statements as well .If no match is found, with any of the

case statements, only the statements following the default are executed. Moreover, the **switch** statement transfers control to a statement within its body. Control passes to the statement whose **case** constant-expression matches the value of **switch (expression).** Further, execution of the statement body begins at the selected statement and proceeds until the end of the body or until a break statement transfers control out of the body. A default is optional. When present, it will be executed if the value of the expression does not match any of these **case** values .if not present, no action takes place if all matches fail and the control goes to the statement-x.

The selection process of **switch** statement is explained by the following flow diagram (Fig.7.10).

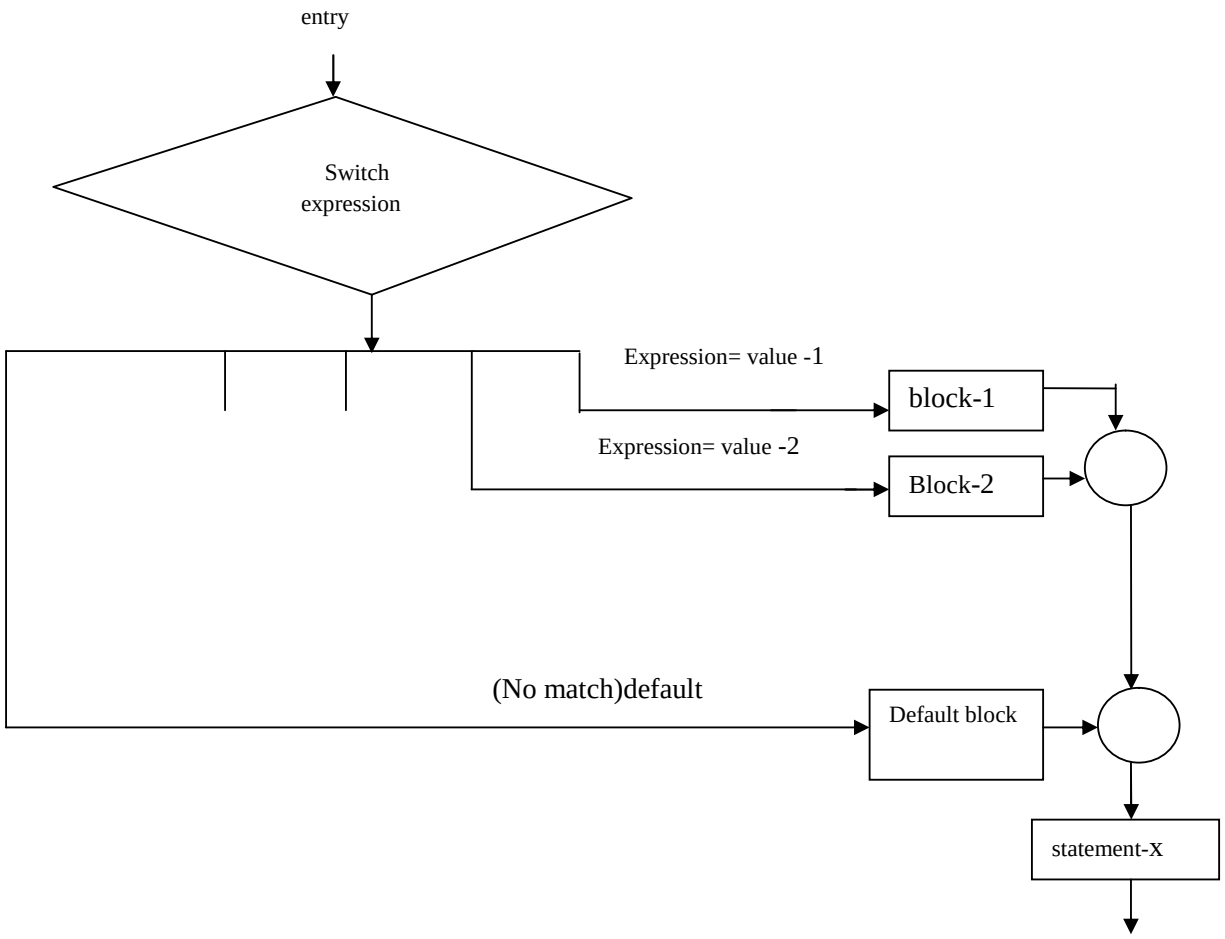


Fig.7.10 Flow chart I illustrating switch statement

The following program explain how this control structure works. Here is a program (Fig.7.11)using switch statement:

```
#include <stdio.h>
int main ( )
 {
    char grade = 'B';
    switch (grade)
     {
     case 'A' :
            Printf( "very good!\n" );
            Break;
      case 'B':
      case'C' :
            Printf("good\n");
            Break;
     case 'D':
            Printf("passed\n");
            Break;
    case 'F':
           Printf("pl try again\n");
           Break;
    default :
           Printf("grade invalid\n");
    }
     Printf("grade is %c\n", grade);
     Return 0;
}
```

Fig. 7.11 : An example showing switch statement

This program on execution gives the following output:

Output

Good

Your grade is B.

Rules for using switch case :

1.The expression used in a **switch** statement must be an integral or enumerated type.

2.With in a **switch** statement one can have any number of **case** statements, with each **case** followed by the

v**alu**e to be compared to and a colon.

3.**case** label must be unique , and must be constants or constant expressions. case labels must end with

semicolon

4.**case** label must of integral type and should not be of floating point type.

5.When the variable being switched on is equal to a **case**, the statements following that **cas**e will execute

until a **break** statement is reached.

6.**switch** case should have at most one **default** label and can be placed anywhere in the **switch,** usually

placed at the end . d**efault** label is optional. No **break** is needed in the **default** case.

7.**break** statements takes control out of the **switc**h (or **switch** terminates and the flow of control jumps to

the next line following **switch** statement) and it is possible to share two or more case statement to

have one **break** statement.

8.Nesting(switch within switch) is permitted for **switch** statement.

9.It is not necessary that every case needs a break statement. If no break appears, the flow of control will

fall through to subsequent cases until a break is reached.

10 relational operators are not allowed in switch case statement .

1.8 The ?: Operator

The operator ?: is just like an **if..else** statement except that because it is an operator one can use it within expressions. This is a ternary operator in that it takes three values. The general form of use of this operator is:

conditional expression ? expression 1 : expression 2

Here, the conditional expression is evaluated first and the result if it is non zero, then expression 1 is evaluated and its value is returned as the value of the conditional expression. Otherwise, expression 2 is evaluated and its value is returned. For example the code segment,

```
If (x < 0)
flag = 0;
else
flag = 1;
```

can be written as

```
flag = (x< 0) ? 0 : 1;
```

consider evaluation of yet another function

y = 1.5x+3 for x≤ 2

2x +4 for x >2.

This can be done using the conditional operator ? : as:

```
y = ( x >2)  ? (2*x+4) : (1.5 *x+3);
```

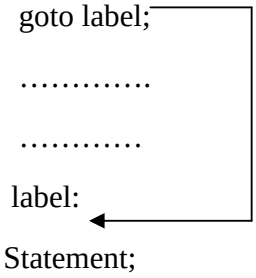
```
# include < stdio.h >

# include < conio.h >

Void main ( )

{

int a,b.c, maxm;

printf(“ program to find maxm value of three numbers:\n”);

printf(“enter the first number:\n”);

scanf(“%d”, &a);

printf(“enter the second number:\n”);

scanf(“%d”, &b);

printf(“enter the third number:\n”);

scanf(“%d”, &c);

max= a>b? (a>c?a: (b > c?b:c )) : (b>c? b:c);

printf(“the maximum  number is %d:”, maxm\n”);

}
```

Fig 7.12: illustration of the conditional operator.

.On execution of the program, the maximum variable gives the maximum value of the three numbers .

1.9 The GOTO statement

In C, GO TO statement is used for altering the normal sequence of program execution by transferring control to some other part of the program. That is ,A **goto** statement provides an unconditional jump from the **go** to a labeled statement in the function. The general form of a go to statement is:



In this syntax label; is an identifier, to identify the place where the branch is to be made.? That is, when the control of program reaches to go to statement, it will jump to the label:, and execute the codes after it. Control may be transferred to anywhere within the current function. The **label** is placed immediately before the statement where the control is to be transferred. A **label:** is any valid variable name, followed by a colon and can be any where in the program either before or after the go to label; statement. During program execution when a statement like

go to begin;

Is met, the control flow will jump to the statement immediately following the label begin; This happens unconditionally.

Note that though, using **goto** statement give power to jump to any part of program, using **goto** makes the logic of the program complex and tangled .It breaks the normal sequential execution of the program. If the label: is used before the statement **goto** label; a loop will be formed and some statements will be executed repeatedly. Such a jump is called as a forward jump. On the other hand, if the label: is placed after the **goto** label; some statements will be skipped and the jump is called a forward jump.

A **goto i**s often used at the end of a program to direct the control to go to the input statement, to read further data, in fact, such goto statements puts one to enter in a permanent loop called infinite loop, until one take some special steps to terminate the program. Such infinite loops are to be avoided. Another use of goto is to transfer control out of a loop 9or nested loop) when certain peculiar conditions are encountered. Use of **goto** statement is highly discouraged in any programming language because it makes difficult to trace the control flow of a program, making the program hard to understand and hard to modify. An example to explain the control flow of goto statement is shown in fig 7, 12.Here in this program,

we want to display the numbers from 0 to 9. For this, we have defined the label statement **loop** above the **goto** statement. The given program declares a variable n initialized to 0. The **n++** increments the value of n till the loop reaches 10. Then on declaring the **goto** statement, it will jumps to the label statement and prints the value of n.

1.10 Summary:

1. There are three ways of taking decisions in a C program. - The **if** statement, the **if else** statement, and the **switch** statement. The default scope of the if statement is only the next statement.

2 An ***if*** block need not always be associated with an else block. However, an ***else*** block is always

Associated with an **if** statement.\

```
#include< stdio.h>
#include< conio.h>
int main( )
 {
  int n =0;
  loop: ;
  printf(" \n%d",  n);
  n++;
 if(n <10)
  {
 goto loop;
  }
getch( );
return 0
}
```

Fig.7.12 Use of *go to* statement

3. If the outcome of an ***if else*** ladder is only one of two answers then the ladder should be replaced either with an ***else-if*** or by logical operators.

4. When we need to choose one among number of alternatives, a ***switch*** statement is used.

5. The ***switch*** key word is followed by an integer or an expression that evaluates to an integer. the case key word is followed by an integer or a character constant. the control jumps through all the cases unless the break statement is given.

6. The usage of ***goto*** is to be avoided as it obstructs the normal flow of execution.

Decision making and looping

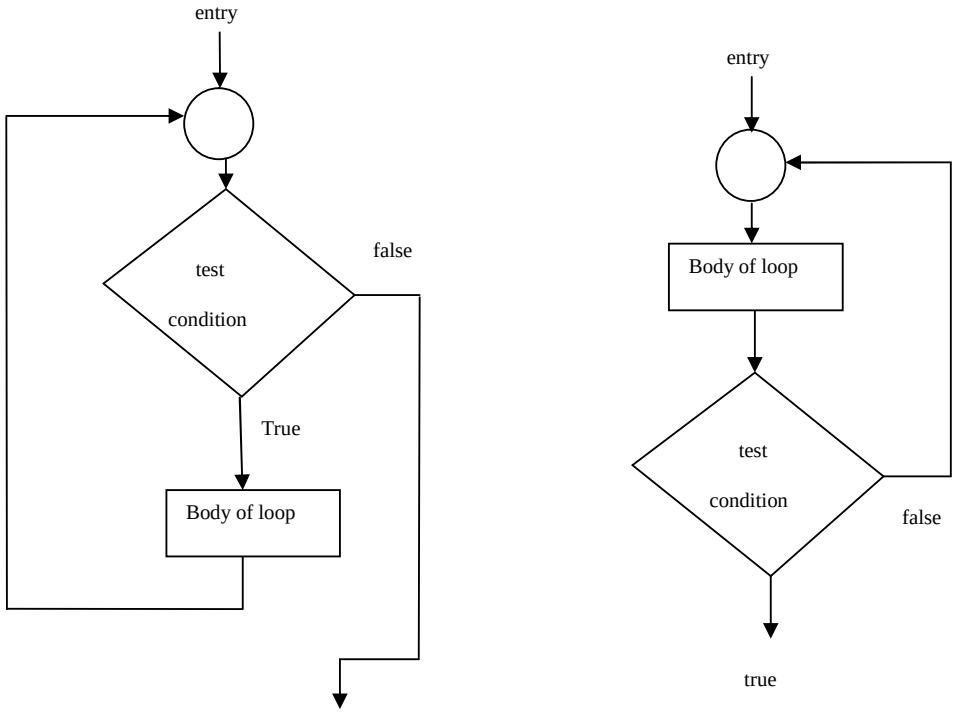
Structure

Introduction
The *While* statement
The *Do* Statement
The For Statement
Jumps in loops
The continue statement
Summary:

2.1 Introduction

The multifunctional ability of the computer lies in its adaptability to perform a set of instructions repeatedly. This involves repeating some portion of the program either a specified number of times or until a particular condition is being satisfied. This repetitive operation is done through a loop control instruction. During looping, a set of statements are executed until some conditions for termination of the loop is encountered .A program loop consists of two segments, one is the **body of the loop** and the other known as the **control statement**. The control is tested always for execution of body of the loop.

Depending on the position of the control statement in the loop, a control may be classified as the **entry controlled loop** or as the **exit controlled one** (Fig.8.1). In the **entry** controlled loop, the control condition is tested first and if satisfied then only body of the loop is executed. In the **exit controlled** loop, the test is made at the end of the body, so the body is executed unconditionally first time.



(a) Entry controlled loop

(b) Exit controlled loop

Fig.8.1 loop control s

A looping process, in general, would include the following four steps:

1. Setting and initialization of a counter.
2. Execution of the statement in the loop
3. Test for a specified condition for execution of the loop.
4. Incrementing the counter.

The three loop constructs in C language for performing loop operations are:

1. The ***while*** statement
2. The ***do-while*** statement
3. The ***for*** statement.

Sentinel loops

Based on the nature of control variable, and the type of value assigned to it, for testing the control expression, there are two types of loops:

1. counter controlled

2. sentinel controlled loops (repetition).

Counter controlled repetitions are the loops which the number of repetitions needed for the loop is known before the loop begins; these loops have control variables to count repetitions. Counter controlled repetitions need initialized control variable (loop counter), an increment (or decrement) statement and a condition used to terminate the loop (continuation condition).

Sentinel controlled repetitions are loops with an indefinite repetitions; this type of loop use a special value, called **sentinel** value, to change the loop control expression from true to false(i.e., to indicate end of iteration) .

2.2 The *While* statement.

While statement is a s**entinel c**ontrolled repetition which can be iterated infinite number of times. Number of iterations is controlled using the **sentinel** variable (test expression). It is one of the simplest looping structures. The basic format of the **while** statement is:

```
While (test condition)
  {
      body of the loop
  }
```

The ***while*** is an **entry-controlled** loop statement. The test condition is evaluated and only if the condition is true the body is executed. After execution of the body, the test-condition is once again evaluated and if it is true, the body is executed once again. This process of repeated execution of the body continues until the test-condition finally becomes false and the control is transferred out of the loop. On exit, the program continues with the statement immediately after the body of the loop. If the body contains only one statement it is not necessary to put the braces, but placing them is a good programming practice. Let us look at a simple example, which uses a ***while*** loop.

```
# include< stdio.h>
int main( )
 {
    int p,n,count;
    float r,si;
    count =1;
    while(count <= 4)
     {
            printf (”enter values for p,n,r\n”);
            scanf ( “%d %d %f “, &p,&n,&r”);
            si = p*n*r/100;
            printf(“Simple interest is: Rs. %\n f”, si);
            count = count +1;
       }
       return 0;
  }
```

Fig 8.2: program to illustrate ***whil***e loop

Here, the program executes all the statements after ***while*** 4 times. The logic for calculating the simple interest is written within a pair of braces (i.e., the statements form body of while loop) immediately after the keyword while. The parentheses after the while contain a condition. So long as this condition remains true, all statements within the body of the ***while*** loop keeps getting executed repeatedly. .Also, to start with, the variable **count** is initialized to 1 and every time the logic of simple interest is executed, the value of count is incremented by one .The index variable **coun**t here, is called the loop counter .

The following points about ***while*** are worth noting.

1. The statements within *w**hile*** loop would keep on getting executed till the condition being tested remains true. When the condition becomes false, the control passes to the first statement that follows the body of the ***while*** loop.
2. In the place of condition there can be any other valid expression. So long as the expression evaluates to a non zero value, the statements within the loop would get executed.
3. The condition being tested may be relational or logical operators as in the example below.

   while (i < = 4)

   while (i > = 4 && j < = 5)

   while (i >. = 4 && (j < 5 || c< 10))

4. The statements within the loop may be a single line(i.e., here braces optional) or a block of Statements as in example shown below.

```
while( i < =5)
  i = i+1;
```

is same as,

```
while( i < =5)
{
   i = i+1;
}
```

5. Almost always, the ***while*** must test a condition that will eventually become false, otherwise the loop Will be executed for ever.

6. Instead of incrementing a loop counter (not necessarily integer it can be a float), one can Decrement it and can still manage the body of the loop to be executed repeatedly.

2.3 The *Do* Statement

The ***do while*** loop is also a kind of loop, which is similar to the ***while*** loop, in contrast to while loop, the ***do while*** loop tests at the bottom of the loop after executing the body of the loop. Since the body of the loop is executed first and then the loop condition is checked we can be assured that the body of the loop is executed at

least once. The ***while*** on the other hand, will not execute its statements if the condition fails for the first time. That is, the *w**hile*** tests the condition before executing any of the statements within the ***while*** loop. As against this, the ***do-while*** tests the condition after having executed the statements within the loop. Since the test condition is evaluated at the bottom of the loop, *the* ***do-while*** statement is

```
do
{
        body of the loop
}
while (test condition);
```

an **exit controlled** loop statement. The ***do-while*** loop looks like this: Here the statement is executed first, and next the expression is evaluated. If the condition in the expression is true then the body is executed again and this process continues till the conditional expression becomes false. When the expression becomes false the loop terminates. This difference is brought about more clearly by the following program.

```
#include<stdio.h>
int main ( )
 {
    while ( 4<1)
          printf("hello\n");
    return 0;
 }
```

Here the, since the condition fails the first time itself, the printf ( ) will not get executed at all. The same program using **the *do-while*** construct is

```
#include<stdio.h>
int main ( )
{
   do
    {
         printf("hello\n");
      } while ( 4<1);
      return 0
  }
```

In this program, the printf ( ) would be executed once, since first the body of the loop is executed and then the condition is tested. ***Break*** and ***continue*** are used with ***do while*** just as they would be in a ***while***. A ***break*** takes one out of the ***do-while*** by passing the conditional test. A ***continue*** sends you straight to the test at the end of the loop.

2.4 The For Statement

The for loop is another entry-controlled loop that provides a more concise loop control structure. It is a counter controlled repetition. Therefore the number of iterations **must** be known before the loop starts (or predetermined). The body of a **for** statement is executed zero or more times until an optional condition becomes false. Also one can use optional expressions with in the **for** statement to initialize and change values during the for statements execution. **The** general form of the **for** loop is:

```
for (initialization; test condition; increment;)
{
      body of the loop
}
```

That is, in the control block of the **for** loop statement there are three expressions separated by semicolon (;).The execution of the for loop is as :

1. **The initialization**: Initialization of the control variables is done first using assignment statements .It is typically used to initialize a loop counter variable.
2. The value of the control variable is tested using the **test condition**. The test condition is a relational expression, such as i <5 that determines when the loop will exit. That is, the loop condition expression is evaluated at the beginning of each iteration. The execution of the loop continues until the loop condition evaluates to false.
3. **Increment**: The increment expression is evaluated at the end of each iteration. It is used to increase or decrease the loop counter variable.

Let us write down the simple interest program(which we have written earlier using **while** statement) using **for (Fig.8.3).** If this program is compared with the one written using **while** construct, we can see that , the three steps of f**o**r loop construct have now been incorporated in the **for** statement. Here in this program (fig 8,3), when the **for** statement is executed for the first time, the value of c**ount** is set to an initial value 1. Next the condition count <=3 is tested. Since the count was set to 1, the condition is satisfied and the body of the loop is executed for the first time. Up On reaching the closing brace of for**, control** is sent back to the **for** statement, where the value of count

is incremented by 1. Again the test is performed to check whether the new value of count exceeds 3. If the value of count is less than or equal to 3, the statements within braces of for are executed again,. The body of the for loop continues to get executed till count does not exceed the final value 3.The control exits from the loop , when count reaches the value 4.and the control is transferred to the statement(if any) immediately after the body of **for.**

```
#include<stdio.h>
int main( )
 {
      int p,n,si;
      float,si;
      for(count =1; count <=3; count= count+1)
       {
            printf(enter the values for p.n,r\n");
            scanf("%d %d %f",&p,&n,&r);
            si = p*n*r/100;
            printf(" simple interest + rs. %f\n", si);
            return 0
       }
 }
```

Fig 8.3: Program using for loop

Additional Features of **for** loop

1. More than one variable can be initialized at a time in the **for** statement as in :

for (p =1, n =6; n <11; $^{++}$n)

Statement. That is, initialization section has two parts p = 1 and n = 6 , separated by comma..Like initialization section, increment section too can have more than one part. The multiple arguments in the increment section too are separated by commas.

2. The test condition may have any compound relation and the testing need not be limited only to the Loop control variable. For eg:

```
sum = 0;
for ( i =1 ;i<10 && sum< 19; ++i )
{
  S = s+1;
 printf("%d %d \n",i,sum):
}
```

Here the loop uses a compound test condition with the counter variable i and variable sum .The loop is executed as long as both the conditions i<10 && sum < 19 are true. The sum is evaluated inside the loop.

3. It is also permissible to use expressions in the assignment statements of initialization and increment Sections. For eg. A statement of the type

```
for( x= (m + n)/2; x > 0; x = x/2)
```

is valid.

4. One or more sections can be omitted if necessary as in eg.,

```
----------
m=5;
for (; m ! = 100 ;)
 {
        printf( " %d\n", m);
        m = m+3;
 }
```

Here, both initialization and increment sections are omitted in the **for** statement. The initialization has been done before the **for** statement and the control variable is incremented inside the loop. Though the sections remains blank, the semicolons separating the sections must remain. If the test condition is not present, the **for** statement sets up an infinite loop. Such loops can be broken using **break or goto** statements in the loop..

5.Time delay loops in **for** loop can be set up using the null statement as:

```
for ( i = 100; i > 0; i = i-1)
;
```

Here this loop is executed 100 times without any output. The body of the loop contains only a semicolon.

Known as null statement.

Nesting of For Loops

The way IF statements can be nested, similarly whiles and **for**s can also be nested; two loops can be nested as follows:

```
…….
……..
for ( i =1; i <10; ++ i)
{
      ……
    for ( j= 1; j!  = 5: ++ j)
    {
      ……..
     }
      ……..
}
      ……..
```

Inner loop

Outer loop

The nesting may continue up to any desired level. To understand how nested loops work, we look at the program below.

```
# include< stdio.h>
int main ( )
 {
    int r,c,sum;
    for ( r =1; r < =3; r ++)
     {
          for( c=1; c<=2; c++)
           {
               sum = r+c;
               printf("r= %d sum = %d \n", r,c,sum);
            }
     }
       return 0;
 }
```

output

```
r =1 c=1 sum=2
 r =1 c=2 sum=3
r =2 c=1 sum=3
r =2 c=2 sum=4
r =3 c=1 sum=4
r =3 c=2 sum=5
```

Fig 8.4. Program to explain *neste*d **for**

Here for each value of r, the inner loop cycles through twice, with variable c taking values 1and 2.The inner loop terminates when c exceeds 2 and the outer loop terminates when r exceeds 3.

2.5 Jumps in loops

We often come across situations, where we want to jump out of a loop instantly, without waiting to get back to the conditional test. The keyword **break** allows to do this. When **break** is encountered in a loop , control automatically passes to the first statement after the loop. A **beak** is usually associated with an **if.** The key word **break, b**reaks the control only from the **while** in which it is placed. As an example we have :

```
# include < stdio.h>
int main( )
 {
    int num, i;
    printf(“ enter a number”);
    scanf(“%d”, & num);
    i =2;
    while( i < = num-1)
     {
       if (num% ! = = 0)
        {
            printf( “not a prime number\n”);
            break;
        }
        i++;
     }
     if ( i = = num)
          printf(“prime number\n”);
  }
```

Fig 8.5 use of **brea**k statement

2.6 The continue statement

The keyword **continu**e, allows us to take the control to the beginning of the loop, by passing the statements inside the loop, which have not yet been executed. That is , when the key word **continue** is encountered inside any loop, control automatically passes to the beginning of the loop .A **continue** is usually associated with an **if.** The **syntax** is:

Continue;

```
#include < stdio.h >

main()
{
    int i;
    int j = 10;

    for( i = 0; i <= j; i ++ )

    {

       if( i == 5 Goods 1

)
       {
          continue; Goods 1

       }
       printf("goods %d\n", i );
    }
}

Output

Goods 1

Goods 2

Goods 3

Goods 4

Goods 5

Goods 6

Goods 7

Goods 8

Goods 9

Goods 10
```

Fig .8. 6 .Use of **continue** statement

As an example consider the program of Fig.8.6. The use of **continue s**tatement in loops is illustrated in fig 8.7.In **while** and **do while** loops, **continue**, causes the control to go directly to the test condition and then to continue the iteration process. In the case of **for** loop, , the increment section of the loop is executed before the test condition is evaluated.

```
While (test condition)        do                          for(initialization; test condition; increment)
{                             {                           {
……..                          ……                          ……..
If (……)                       if(……)                      if(……..)
  Continue;                     continue;                   continue;
……                            ……..                        ……..
……                            ……                          ……..
}                             } (while test condition );  }
```

Fig.8.7 *continue* **command in while, do while and for loop statements**

Jumping out o

We have seen that we can jump out of a loop using either the **break** or **goto** statement. In the same way we can jump out of a program by using the library function exit( ).. The use of exit( ) function is shown in fig. 8.8 below:

```
……..
…….
If (test condition) exit (0);
……
……
```

Fig.8.8. use of **exit** ( ) function.

2.7 Summary:

1.The three types of loops available in C are for, while, and do while.
2. A Break statement takes the execution control out of the loop.
3.a continue skips the execution of the statements after it and takes he control to the beginning of the loop.
4. A do while loop is used to ensure that the statements with in the loop are executed at least once.
5 when we need to choose one among number of alternatives, a switch statement is used.
6.The switch key word is followed by an integer or an expression that evaluates to an integer.
7. the case keyword is followed by an integer or a character constant.
8. the usage of goto keyword should be avoided as it usually violates the normal flow of execution.

Unit 10

This module is designed as an introduction to Data structures. It is about structuring and organizing data as a fundamental aspect of developing computer application. The standard data structures which are often used and which forms the basis for complex data structures is the array. An array is a homogenous data structure in which all elements are of the same type. In the first unit of the module, we describe different types of arrays in general. The next unit is devoted to a useful introduction to User defined functions.

Arrays

Structure

Introduction

One dimensional Arrays.

Declaration of one dimensional Arrays

Initialization of one dimensional Array.

Two dimensional Arrays.

Initializing 2-D arrays

Multi dimensional Arrays

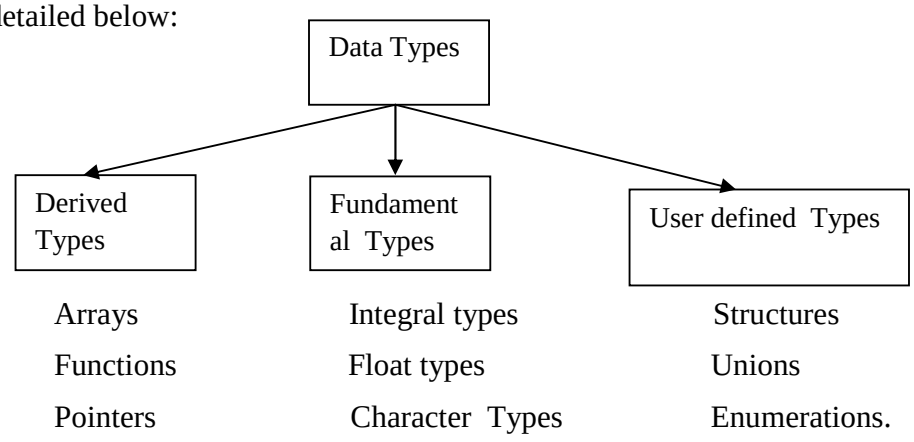
Dynamic Arrays

Summary:

1.1 Introduction

An array is a collection of similar elements. These similar elements could be all integers, or all floats, , or all characters, etc. Usually, an array of characters is called a ‘ string’, where as an array of integers or floats is simply called an **array**. All elements of any given array must be of the same type. That is, we cannot have an array of 10 numbers, of which five are of integers and five of float type.

C supports a rich set of derived and user defined data types, in addition to a variety of fundamental data types.as detailed below:



Arrays and structures are referred to as **structured data types** because they can be used to represent data values that have a structure of some sort. Structured data types provide an organizational scheme that shows the relationship among the individual elements and facilitate efficient data manipulation. In programming language such data types are known as **data structures**.

1.2 One dimensional Arrays.

As already discussed**,** an array is a collective name given to a group of similar variables .The values in an array is called as **elements** of array, and are accessed by numbers called **subscripts**. The array which is used to represent and store data in a linear form (or accessing its elements involve only a single subscript) is called as single or **one dimensional array**. As an example consider the C declaration:

```
int number [5];
```

Here in this declaration, the array variable **number** contain 5 elements of any value available to the i**nt** type .and the computer reserves 5 storage locations. The values to the array elements can be assigned as:

```
number [0]= 12;
number [1]=13;
number [2]=15;
number [3]=20;
number [4]=25;
```

This would cause the array number to store the values as shown below:

number [0]	12
number [1]	13
number [2]	15
number [3]	20
number [4]	25

These elements may be used in programs just like any C variable

1.3 Declaration of one dimensional Arrays

To begin with, like other variables an array needs to be declared before they are used so that the compiler will know what kind of an array and how large an array we want. The general form of array declaration is:

type variable-name [size];

The **type** specifies the type of element that will be contained in the array, such as int, float or char and the **siz**e indicates the maximum number of elements that can be stored inside the array .For example,

int marks [10];

Declares the marks as an array to contain a maximum of 10 integer constants. This number is often called the **dimension** of the array .The bracket ([ ]) tells the compiler that we are dealing with an array.

The C treats character strings simply as array of characters. The size in a character string represents the maximum number of characters that the string can hold. For instance,

char name[13];

Declares the name as a character array(string) variable that can hold a maximum of 13 characters. Suppose we read the following string constant in to the string variable name

"GOOD MORNING"

In this, each character of the string is treated as an element of the array name and is stored in the memory as:

'G'
'O'
'O'
'D'
' '
'M'
'O'
'R'
'N'
'I'
'N'
'G'
'\o'

When the compiler sees a character string , it terminates with an additional null character \o. Thus the element name[13] holds the null character '\o'. Remember that, while declaring character arrays, we must allow one extra space for the null terminator.

1.4 Initialization of one dimensional Array.

After an array is declared, its elements must be initialized. If they are not given any specific value, they are supposed to contain garbage values. An array can be initialized at either of the following stages:

- at compile time
- at run time

Compile time initialization

Whenever we declare an array we can initialize it directly at compile time. In this type of initialization, we assign certain set of values to array elements before executing program The general form of initialization of arrays is:

type array-name[size] = [list of values];

the values in the list are separated by commas. The type size can be specified directly as :

int num [5] = { 2.3,4,5,6};

Here the size of the array is specified directly as 5 in the initialization statement. The compiler will assign values to the particular elements of the array. i.e., At the time of compilation all, the elements are at specified positions as shown below.

num [0] = 2

num [1] = 3

num [2] = 4

num [3] = 5

num [4] = 6

Also the type size can be specified indirectly as in:

int num [ ] = { 2.3,4,5,6};

The compiler counts the number of elements written with in the braces and determines the size of the array.

Character arrays may be initialized in the same manner. Thus the statement

char name [ ] = { 'j', 'o', 'h', 'n', '\o'};

Declares the name to be an array of five characters, initialized with the string 'john' ending with the null character. Alternatively, we can assign the string literal directly as :

char name [ ] = 'john';

Run time initialization

An array can also be explicitly initialized at run time usually; .this approach is applied for initialization of large arrays. For example, consider the following program segment;

```
for (i  = 0;  i < 5; i++)
{
 scanf ( "% d "'  & x [ i ] );
}
```

The above segment will initialize the array elements with the values entered through the keyword .In this type of initialization (run time initialization) of the arrays. looping elements are almost compulsory. Looping statements are used to initialize the values of the arrays one by one by using assignment operator or through the keyboard by the user. we can also use read function such as **scanf** to initialize an array as in example below.

```
int x [2] ;
```

```
# include < stdio.h >
void  main  ( )
 {
    int array [3], i;
    printf( " enter 3 numbers to store them in an array\n" );
    for ( i =0; i < 3; i ++)
     {
        scanf ( " % d ", & array [ i] ) ;
     }
     printf ( " elements  in the array are: \ n");
     for  i =0; i < 3; i ++)
      {
        printf (" elements stored at a [ %d] = %d\n",i, array [ i]);
      }
      getch ( );
}
```

output

```
enter  3 elements in the array : 2 3 4
elememts in the array are :
element stored at a[ 0] = 2
element stored at a[ 1] = 3
element stored at a[ 2] = 4
```

Fig 9.1: program to illustrate an array

scanf (" %d % d", & x[0], & x[1]);

will initialize the array elements with the values entered through the key word. Here is a sample program (Fig.9.1) to store the elements in the array and to print them from this array.

Searching and sorting are two operations performed on arrays. Searching is the process of arranging elements in the list according to their values, in ascending or descending order. An ordered list is a sorted one. The three simple and important sorting methods are:

Bubble sort

Selection sort

Insertion sort.

Other sorting methods include, Merge sort, quick sort and Shell sort.

Searching is the process of finding the location of the specified element in a list. The specified element is often called the **search key**. If the process of searching finds a match of the search key with a list element value, then the search is sad to be successful. Otherwise it is unsuccessful. Two most commonly used searching methods are ;

Sequential search

Binary Search.

1.5 Two dimensional Arrays.

So far, we have explored arrays with only one dimension. It is also possible to have two or more dimensions. The 2-D array is also called a matrix. The 2-D arrays are declared as :

type array-name [size of row] [column size];

2-D arrays are stored in memory as shown below. In memory, whether, it is single or two dimensional array, the array elements are stored in one continuous chain .Each dimension of the array is indexed from zero to its maximum size minus one: the first index selects the row and the second index selects the column within that row,

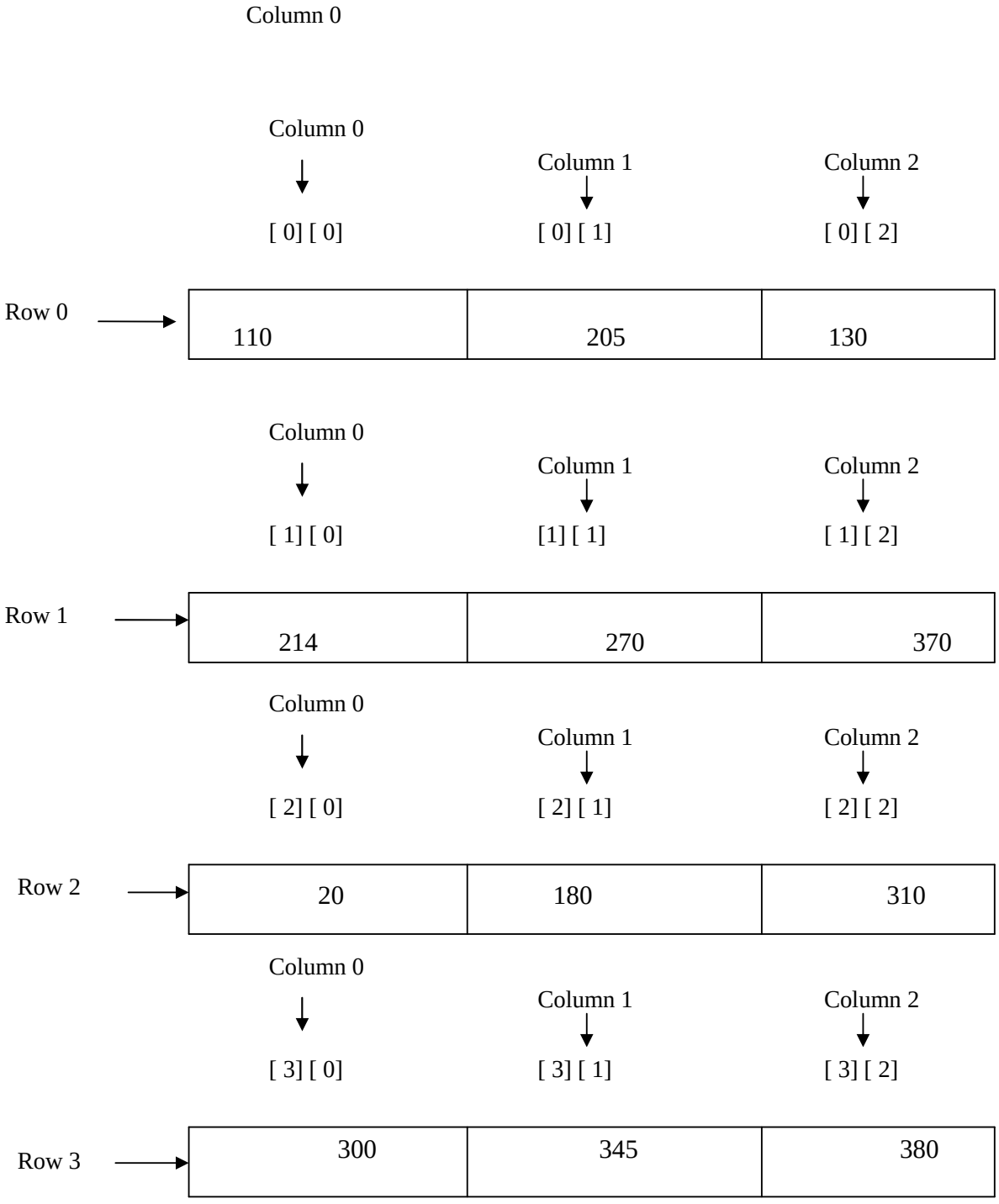


Fig 9,2 : Representation of 2-D array in memory

Here is a sample program:

```
# include< stdio.h>
int main( )
{
    int students [4] [2];
    int i,j;
    for (i = 0; i < = 3; i ++)
     {
          printf ( “enter the roll no of student and marks\n”);
          scanf(“ %d %d”,  &student [i] [0], &student[i][1]);
       }
    for (i =0; i< = 3; i ++)
           printf( “ %d %d “, student [i] [0],student [i] [1]);
    return 0;
  }
```

9.3. program to illustrate 2-D array

This program stores the roll number and marks obtained by a student side by side in a matrix. In the first part of the program, i.e., in the first f**or** loop, we read in the values of roll number and marks, where as in the second **for** loop, we print out these values. Also, in the first **scanf** , the first subscript of the variable student is row number which changes for every student. The second subscript tells which of the two columns are we talking about- the zeroth column which contains the roll number or the first column which contains the mark. The counting of rows and columns begins with zero. Remember that two dimensional array is a collection of a number of one dimensional arrays placed one below the other .In this program, the array elements have been stored row wise and accessed row wise. Although it is possible to access the elements column wise, row-wise strategy is accepted widely.

1.6 Initializing 2-D arrays

Like 1-D arrays, 2-D arrays could be initialized by following their declaration with a list of initial values enclosed in braces as in ,

```
int table [2][3] = { 0,0,0,1,1,1};
```

which initializes the first row to zero and second row to one. Equivalently one can write the above statement as:

```
int table [2][3] = {{ 0,0,0} ,{ 1,1,1}};
```

We can also initialize a 2-D array in matrix form as:

```
int table [2][3] = {
                    {0,0,0},
                    {1,1,1}
                   };
```

More over, the declaration

```
int table [ ][3] = {
                    { 0,0,0},
                    {1,1,1}
                   };
```

Is perfectly valid.

If the values are missing in the initatializer, they are automatically set to zero. For instance, the statement

```
int table [2][3] = {
                    {1,1}
                    {2}
                   };
```

will initialize the first two elements of the first row to one, the first element of the second row to 2 and all other elements to zero.

In situations where we have to initialize all the elements to zero, a short cut method as in,

```
int m [3] [5] = { { 0}, { 0},{0} };
```

may be used. Here the first element of each row is explicitly initialized to zero, while all other elements are automatically initialized to zero. the following statement would also work.

int m [3] [5] =- { 0,0};

1.7 Multi dimensional Arrays

The general form of a multidimensional Array is:

Type array-name [s1] [s2] [s3] …….[sm] ;

Where si is the size of the ith dimension. A 3-D array can be thought of as an array of arrays of array. The outer array has three elements, each of which is 2-d array of four 1-D arrays., each of which contains two integers. That is, a 1-D array of two elements is constructed first, followed by placing four 1-D arrays placed one below the other. So that a 2-d array containing four rows is obtained. Thereafter, three 2-D arrays are placed one behind the other to yield a 3-D array containing three 2-D arrays.

1.8 Dynamic Arrays

In C it is possible to allocate memory to arrays at run time. The arrays created at run time are called dynamic arrays .Dynamic arrays are created using memory management functions like malloc, calloc, realloc, that are included in the header file< stdlib.h > The concept of dynamic arrays is used in creating and manipulating data structures like lists, stack and queues.

1.9 Summary:

1,An array is similar to an ordinary variable except that it can store multiple elements of similar type.

2.The array variable acts as a pointer to the zeroth element of the array. In 1-D array, zeroth element is a

single valued one, whereas in a 2-D array this element is a 1-D array. During multidimensional

initialization, omission of array size other than the first dimension may result an error.

3. While initializing character array, enough space is to be provided for the terminating null character.

4. The subscript variables in a array need to be initialized before they are used.

Unit 11

User Defined Functions

Structure

2.1 Introduction

The **C** language is similar to most modern programming languages in that, it allows the use of **functions** (i.e., a self contained block or module of program code), to get its tasks done. In general, **C** functions contain a set of instructions enclosed by braces'{ }' , that can perform a coherent task of same kind. They are easy to define and are reusable. That means, it can be executed from as many different points ***in a C program as required. Broadly speaking, the two categories of functions in C are (1) Library*** functions and (2) ***user defined*** functions. **Library functions** are in built functions that are grouped and placed together in a common place called 'library', and are capable of performing specific operations. The main difference between a **library** and **user defined function** is that **library** functions are not required to be written by the user where as a **user defined function** has to be developed by us at the time of writing a program. In fact, a **user defined function** later becomes a part of the C program library. **main** is a specially recognized function in C and **is a**n example of **user defined function** while the functions **printf a**nd **scanf** belong to the category of **library** function.

2.2 Need for User defined Functions

A **function in C,** is a module of a program code (or a block of code that takes information in, does some computation, and returns a new piece of information based on the parameter information) which deals with a particular task. In fact, every program can be thought of as a collection of these functions. That is, **functions** groups a number of program statements into a unit and this unit can be invoked from other parts of a program. This division approach clearly results in a number of advantages:

1.It results in high level modular programming, (Fig.10.1) wherein the high level logic of the overall problem is solved first while the details of each lower level functions are addressed later.

2. By using functions at the appropriate places, the length of the source program can be reduced.

3. A function may be used by many other programs.

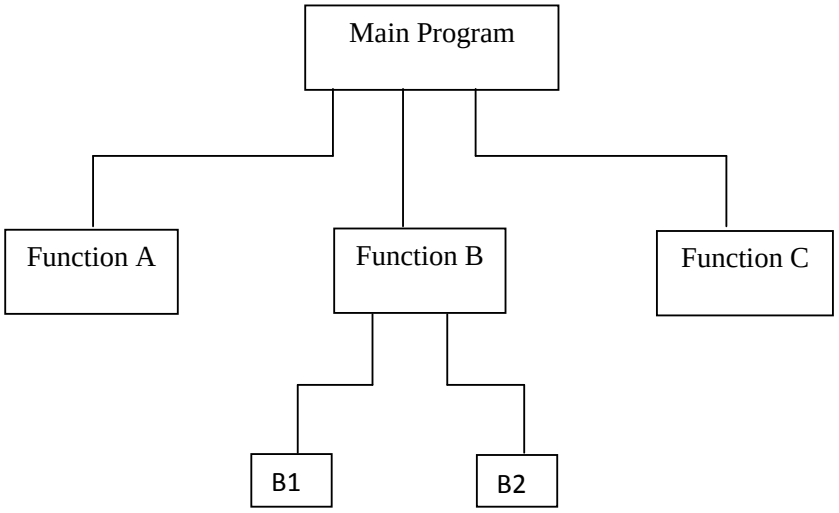


Fig.10.1 Top down Modular Programming using **functions**

2.3 A Multi function Program

As was pointed earlier, a **function** is a self contained block of instructions that perform a coherent task of some kind. Moreover, a **function** can be accessed from any location with in the C program. Making **functions** is a way of isolating one block of code from other Independent blocks of code. A **function** can take a number of parameters, do required processing and then return a value. When a function is

```
void message( );
int main ( );
{
  message( );
  printf ( “this explains  the use\n”);
  return 0;
 }
void message( )
 {
  printf ( “this is function definition \n”);
}
```

Fig 10.2

Defined at any place in the program then it is called function definition. That means, once a function is defined and packed, then it takes some data from the main program and returns a value. Actually, we will be looking at two things - a function that calls the function and the function itself. Let us consider the above chunk of program(fig.10.2).

And here is the output…..

this is function definition

this explains the use

Here we have defined two user defined functions- **main ( )** and **message ( ).** In fact, we have used the word message at three places in the program. During the execution of the main, the first statement encountered is

message( );

which indicates that the function **message** is to be executed. At this point , the program transfers its control to the function **message**. After executing the **message** function (here no value is returned as was indicated by the key word **void**)., the control is transferred back to the **main**. Now, the execution continues at the point where the function call (by definition) was executed. After executing the **printf** statement, the control is again transferred to the function **message ( )** if being called by **main ( ).** That means the activity of **main ( )** is temporarily suspended while the **message ( )** function

wakes up and goes to work. When the message ( ) function runs out of statements to execute, the control returns to **main ( ),** which comes to be active again by executing its code at the exact point where it left off. Thus, **main ( )** becomes the *calling* and message ( ) becomes the *called* function.

Any function can call any other function, In fact, it can call itself. Further, a called function can call another function. Also, a function can be called more than once in any program. Moreover, there are no predetermined relationships, rules of precedence or hierarchies (except at the starting point), among the functions that make up the complete program. The functions can be placed in any order and the called function can be placed either before or after the calling function. The best practice is to put all the called functions at the end. Figure 10.3 illustrates the flow of control in a multifunction program

2.4 Elements of User defined Functions

So far we have discussed and used a variety of data types and variables in our programs .Nevertheless, declaration and use of these variables were primarily done inside the main function. We can therefore define functions and use them like any other variables in C program. Both functions names and variables are considered as identifiers and therefore they must follow the rules for identifiers..Further, Like variables, functions have type associated with them and the function names and types must be declared and defined before they are used in program. Every user defined functions has three elements.

- Function definition
- Function Call
- Function Declaration.

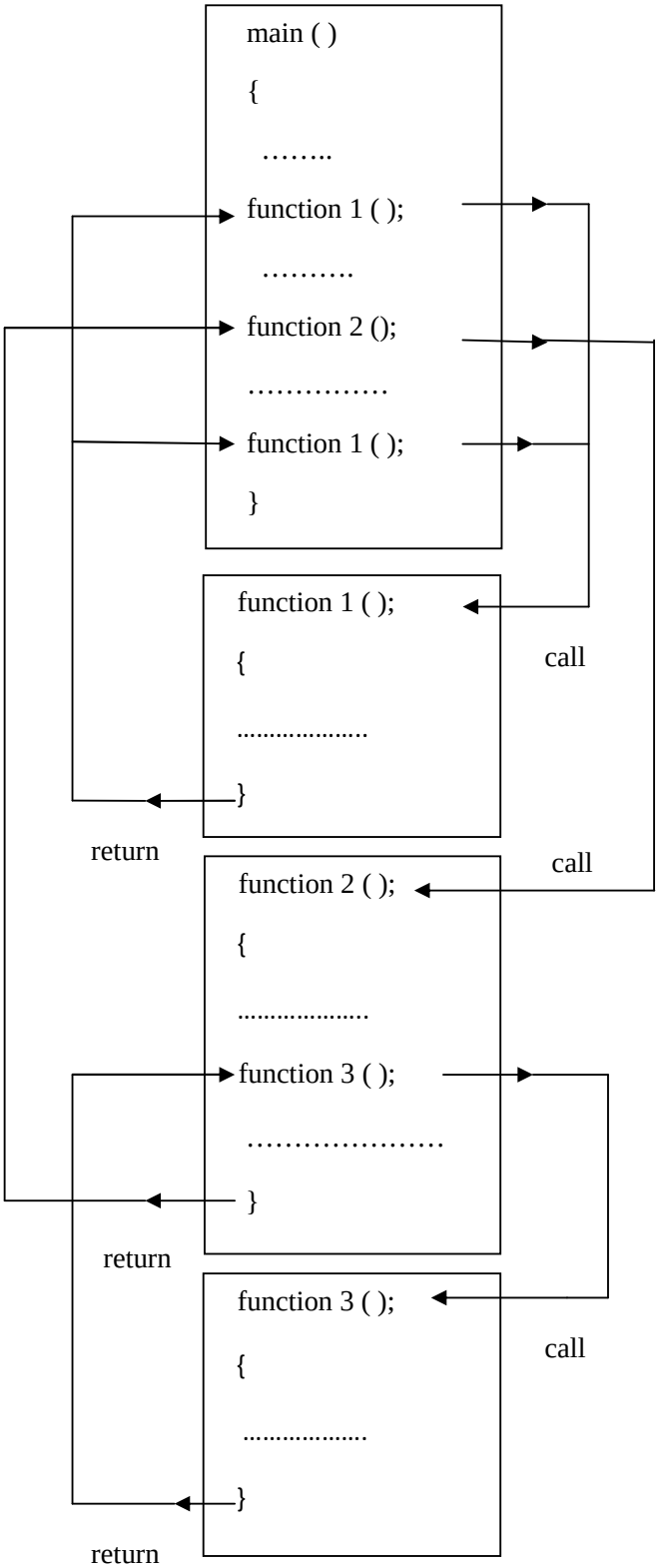


Fig 10.3 Flow of control in a multifunction program

The ***function definition*** is independent program modules that is specially written or apply the requirements of the function. To use this block or function, we need to call down it at the required place in the program, known as the functions. A function is defined when function name is followed by a pair of braces in which one or more statements may be present. The program that calls the function is referred to as the calling program or calling functions. The calling program should declare any function that is used later in the program. This is termed the function declaration or ***function prototype.***

2..5 Definition of Functions

The function ***definition*** which is the heart of function, is an independent program module that is specially written to suit to the requirements of the function. A function definition shall include the following elements

- Function name
- Function type
- List of parameters.
- Local variable declarations
- Function statements
- A return statement

All the six elements are grouped in two parts namely,

1. Function header (first three elements)

2. Function body (Second three elements)

A general format of function definition to implement these two parts(Fig.10.4) is:

```
function_ type  function_name (parameter list)
{
  local variable declaration;
  executable statement1;
  executable statement2;
  …….
  …….
  return statement;
 }
```

The first line **function_ type function_name (parameter list)** is known as the function header and the statements within the opening and closing braces constitute the function body.

Function header

The function header consists of three parts: function type, function name and the function parameter list. Semicolon is not used at the end of the function header.

Function name and type

Function type may specify the data type that one may use (like ***float ,int or double*** whatever according to ones needs) .If data type is not specified then C will assume it as ***int*** type and if the function does not return any value then ***void*** is used.

Function name may consist of any variable that is suitable for users understanding. That means, it is any valid C identifier that must follow the same rules of formation as other variable names in C. A function gets called when the function name is followed by a semicolon.

Parameter List

It declares the variables that are to be used in the function and that are going to be called in the program. Actually, they serve as input data to the function to carry out the specified task and are also be used to send values to calling programs. They are often termed as ***formal*** parameters(or arguments). The parameter list contains declaration of variables separated by commas and enclosed in parentheses with no semicolon after the closing parentheses. Note that combined declaration of parameter variables is invalid. That is, *int sum(int a,b)* is not a valid declaration of parameter list. To indicate an empty parameter list, usually we use the key word ***void*** between the parentheses as in

```
void printline (void)
{
  ……..
 }
```

Many compilers do accept an empty set of parentheses, without specifying anything as in

```
void printline ( )
```

Again, its nice to have *void* to indicate a nill parameter list.

Function Body

The function body contains the declarations and statements necessary for performing the required task. The bodies enclosed in braces contain three parts:

- Local declaration that specify the variables needed by the function
- Function statements that perform the task of the function
- A ***return*** statement that returns the value evaluated by the function.

If the called function is not going to return any meaningful value to the calling function, the use of ***return*** statement can be omitted. Nevertheless, its return type should be specified as ***void***. But it is better to have a return statement even for ***void*** functions.

2.6 Return Values and their types

As pointed out earlier, a ***return*** statement is a statement that returns the value evaluated by the function to the calling program. If a function does not return any value, one can omit the ***return*** statement. When a ***return*** is encountered, the control is immediately passed back to the calling function. A function can ***return*** only one value at a time per call and the ***return*** statement can take one of the following forms:

return;

or

return (expression) ;

Here, the first 'plain' return does not return any value (or it acts as the closing brace of function).The second form of return returns the value of the expression. For example, the function

```
int mul ( int x, int y)
{
  int z;
  z = x* y;
  return (z);
 }
```

Returns the value of z . It is possible to have more than one ***return*** statement for a function as in:

```
if ( x < = 0 )
    return ( 0 );
else
    return (1 );
```

All functions by default return ***int*** type data. We can force a function to return a particular type of data by specifying the ***type specifier*** in the function header. For functions that use ***doubles***, yet returns ***ints***, the returned value will be truncated to an integer as in:.

```
int product (void)
 {
 return (2.5* 3.0);
}
```

Will return the value 7, only the integer part of the computation.

2.7 Function Calls .

In order to use functions user need to call on it at a required place in the program. This is known as the function call. A function can be called by simply using the function name followed by a list of actual parameters, if any, enclosed in parentheses. For example,

```
main ( )
 {
  int y;
  y = mul (10, 5);              /* function call * /
  printf ( " %d\n",y);
 }
```

Here in the **main( )** program the mul(10, 5) function has been called. The C compiler, when it encounters a function call, the control is transferred to the function mul ( ). This function is then executed line by line and a value is returned (which is assigned to y) , when a return statement is encountered.

A Function that returns value can be used in expressions like any other variable.

e.g; y = mul (p,q)/(p+q);

Of course, a function cannot be used on the RHS side of an assignment statement. Thus, the statement

mul (a,b) = 15;

Is wrong. Moreover a function , that does not return any value may not be used in expressions; but

can be used to perform certain tasks specified in the function. Such functions may be called in by simply stating their names as independent statements. For example,

```
main ( )
{
   printline ( );
}
```

2.8 Function Declaration

The program or a function that called a function is referred to as the calling function or calling program. The calling program should declare any function that is to be used later in the program. This is known as the function declaration (also known as function prototype). Like variables, all the C functions must be declared, before they are called on. A function declaration involves four parts. viz,

- Function type
- Function name
- Parameter list
- Terminating semicolon.

The general format is:

Function- type function –name (parameter list);

The format is similar to the function header line except the terminating semicolon. Further, when a function does not take any parameters and does not return any value, its proto type , written as:

```
void  display (void);
```

A proto type declaration may be placed in two places in a program:

1. Above all functions including main (also called Global prototype);
2. Inside a function definition.(also called local prototype).

Global declarations are available for all the functions in the program where as local prototype type declarations are used by the functions containing them. The place of declaration of a function defines a region (also called *scope* of the function) in a program in which the function may be used by other functions. It is nice to declare prototypes in the global declaration section before ***main*** so that the user gets a quick reference to the functions used in the program thereby enhancing the documentation.

2.9 Category of Functions .

A function depending on whether arguments are present or not and whether a value is returned or not, may be categorized into:

- Functions with no arguments and no return values
- Functions with arguments and no return values
- Functions with arguments and one return values
- Functions with no arguments but return a value
- Functions that return multiple values

Let us have a look category of functions one by one.

2.10 Functions with no arguments and no return values

When a function has no arguments, the called function does not receive any data from the calling function and it does not return any data back to the calling function. Hence there is no data transfer between the called and calling function. This is pictorially represented in Fig. 10.5.Let us understand this with the help of a program (Fig 10.6)

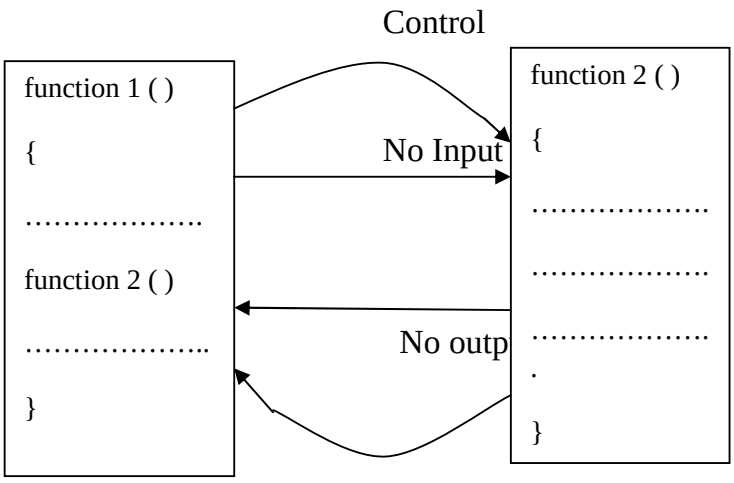


Fig 10.5

Control

```
void main ( )
{
 read_value ( );
}
read_value ( );
{
char name [10];
printf(“enter your name\n”);
scanf(“%s”, name);
printf(“your name is % s, name”);
}
```

output

```
enter your name
salu
your name is salu
```

Fig.10.6

2.11 Function with Arguments but no return value:

Here the called function receives the data from the calling function but the called function does not

```
# include < stdio.h>
# include < conio.h>
 void main ( )
{
 int a,b;
printf(“enter the value for a and b\n”);
scanf(“%d %d”, &a, &b );
largest (a,b);
largest (c,d);
int c,d;
{
if ( c > d)
{
printf(“ largest = % d\n”);
}
else
{
printf(“ largest = % d\n”);
}
return ( );
}
```

output

```
enter  the value for a and b
5
3
largest = 5
```

Fig 10.7 Program to find largest of two numbers

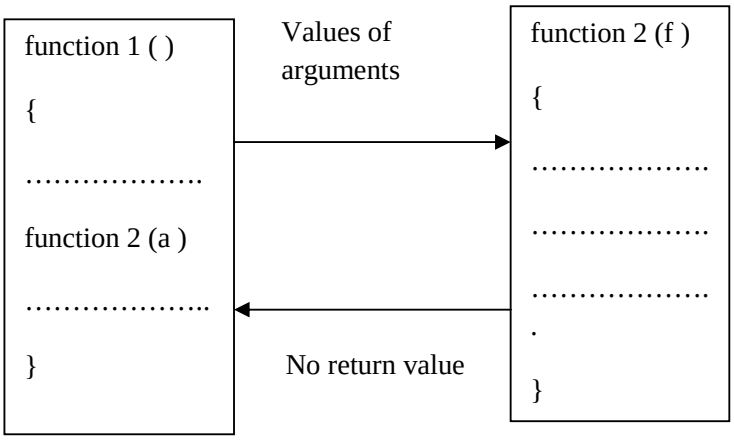


Fig 10.8

return any value back to the calling function. This is depicted in Fig 10.8.The dotted lines in Fig 10.8 Indicates that there is only transfer of control but not data.. A sample program to illustrate this is shown in Fig 10.7

2.12. Arguments with return values

In this type of functions, functions accepts arguments and returns value back to the calling program That means, a self contained and independent function receives a predetermined form of input and outputs a desired value. Thus it is a two way communication between a calling function and a called function (fig.10.9)

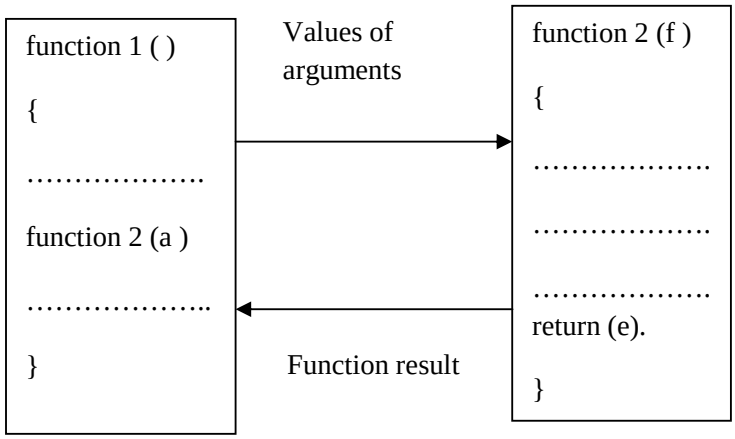


Fig 10.9

For example, the program (Fig.10.10) illustrates the use of two way data communication between calling and called functions.

```
#include < stdio.h >
float  calculate_ area ( int);
int main ( )
{
   int radius;
   int area;
   printf (“ enter the radius:\n”);
   scanf( “%d”, & radius);
   area = calculate_area(radius);
   printf(“ area of circle :”, area);
   return (0);
 }
float  calculate_ area ( int radius);
{
  float area of circle;
  area of circle = 3.14 * radius * radius;
  return(area of circle);
}
```

output

```
enter the radius: 1
area of circle = 3.14
```

Fig 10.10: program to show functions with argument and return value

2.13 Functions with no arguments but returns a value

In this type, the called function does not receive any data from the calling function. It is also a one way data communication between the calling function and the called function. To understand this following program (fig 10.11) will help.

```
# include < stdio.h >

# include < conio.h >

void main ( )

{

float sum;

float total ( );

clrscr ( );

sum = total ( );

printf ( " sum = % f \n", sum);

}

float total ( )

{

float a,b;

a = 2;

b= 8;

return (a+b);

}
```

output

```
sum = 10.000000
```

Fig 10.11: function with no arguments but returns a value

2.14 Functions that return multiple values

Using a return statement, a function in C can return only one value. If we want the function to return more than one value of same data types, we could return the pointer to array of that data types. We can also make the function return multiple values by using the arguments of the function. That is , by providing the pointers as arguments. In fact, when a function needs to return several values, we use one pointer in return instead of several pointers as arguments. Here, the mechanism of sending back information through arguments is achieved by using what are known as address operator (&) and indirection operator (*).For e.g., consider the program code:

```
void mathoperation ( int x, int y, int *s, int *d);
main ( )
{
   int x =10, y = 8, s,d;
    mathoperation (x, y, &s, &d);
    printf ( " s = % d \n d = % d\n", s,d);
}
void mathoperation 9 int a, in b, int * sum, int * diff )
 {
   *sum = a+b;
   *diff = a-b;
}
```

In this code, in the function call, when we pass the actual values of x and y to the function, we pass the address of locations where the values of s and d are stored in the memory. When a function call I s passed, the following assignments takes place.

Value of x to a

Value of y to b

Address of s to sum

Address of d to diff

The indirection operator *** (The name indirection means that it gives indirect reference to variable through its address)** in the declarations **sum** and **diff** in the header indicates these variables are to store addresses and not the actual values of variables. That means, the variables **sum** and **diff** point to the memory location of s and d respectively. In the body of the function, the statements

*sum = a + b;

* diff = a - b;

Imply that the value stored in the location pointed to by **sum** is the value of s and the value of a-b is stored in the location pointed to by **diff** is the value of d. The variables * sum and * diff are pointers and **sum** and **diff** are *pointer variables*..Since they are declared as i**nt** , they can point to locations of **int** type data. The use of pointer variables for communicating the data between functions is termed **call by reference (or call by address/ pass by pointers).**

2.15 Recursion

In C programming, it is possible for the functions to call themselves or the process of defining.

```
#include  < stdio.h>
int sum (int n);
int main ( )
{
    int num, add;
    printf( " enter a positive integer:\n");
    scanf(" %d", & num);
    add = sum (num);
    printf(" sum = %d ", add );
}
int sum(int n)
 {
   if (n = = 0 )
        return n ;
   else
       return n+ sum (n-1);
  }
output
enter a positive number
3
6
```

Fig.10.12 program code for the sum on n natural numbers

Something in terms of itself is known as recursion. A very simple example to find the sum of n natural numbers using recursion(or call a function inside the same function) is shown in Fig 10.12.In this example, the function sum ( ) is invoked from the same function. If n is not zero then the function calls itself by passing argument 1 less when the previous argument it was called with. When n becomes equal to zero, the value of n is returned .In this example, a better visualization of recursion for n = 3, assumes the form:

sum (3)

= 3+ sum (2)

= 3+2+sum(1)

= 3+2+1+sum(0)

= 3+2+1+0

= 3+2+1

= 3+3

= 6

That is, every recursive function must be accommodated with a way to end the recursion. when n is zero, there is no recursive function call and the recursion ends here.

2.16 passing arrays to function

In C programming it is possible to pass a single array or an entire array to a function. Also, both one and multidimensional array can be passed to function as argument. To pass a 1-d array to a called function, listing the name of the array without any subscripts, and size of the array as argument is sufficient. That means, while passing arrays to the argument, the name of the array is passed as an argument. Also, Single element of an array can be passed in the same way as passing variables to a function. For example, the following code

```
#include < stdio.h>
void display( int a)
   {
    printf("%d",a);
   }
 int main( ) {
   int c [ ] = {2,3,4};
   display ( c[2]);  /* passing array element c[2] */
   return 0;
}
```

Explains the passing single element of an array (**that is c[2])** to a function. The output of this program is 4.In C, the name of the array represents the address of its first element. By passing the array name in fact deals with passing the address of the array to the called function. The array in the called function refers to the same array stored in the memory. That is, any changes in the array in the called function will be reflected in the original array. Remember that one cannot pass a whole array by value, as we do in the case of ordinary variables. Also, when we deal with array arguments, care should be taken to incorporate the changes made to the original array that passed to the function, if the function changes the values of the elements of an array.

Two dimensional arrays

Like simple arrays, to pass two dimensional array to a function as an argument, the starting address of memory area reserved is passed .An example, to pass 2-D arrays to function is shown below.

```
# include < stdio.h>
void function(int c[2][2]);
int main(){
int c[2][2],i,j;
    printf("enter 4 numbers:\n");
    for(i=0;i<2;++i)
        for(j=0;j<2;++j){

scanf("%d",&c[i][j]);
        }
    function(c);
    return 0;
}
void function(int c[2][2]){
    int i,j;
    printf("displaying:\n");
    for(i=0;i<2;++i)
        for(j=0;j<2;++j)
```

The output of this program is:

```
Enter 4 numbers
1
2
3
4
Displaying
1
2
3
4
```

The function defined in the program can be used in the main function to display 4 numbers in the array .

2.17. Passing strings to functions

The strings are treated as character arrays in C and therefore the rules for passing strings to functions are same as those for passing arrays to functions .The rules are as follows:

1. The strings to be passed must be declared as a formal argument of the function when it is defined.
2. The function prototype must show that the argument is a string. eg., **void display (char str [ ]);**
3. A call to the function must have a string array name without subscripts as its actual argument.

   eg. **display (names);**

2.18 Summary

1. Function declaration specifies the return type of the function and the types of parameters it accepts. A function can return only one value at a time.
2. There is no restriction on the number of return statements that may be present in a function. Also return statements need not always be present at the end of the called function.
3. A return statement is needed if the return type is anything other than **void.** If a function does not return any value, return type must be declared as **void**

4. Any number of arguments can be passed to a function being called. However, the type, order, and the number of actual and formal arguments must be same .If the value of the formal argument is changed in the called function, the corresponding change does not take place in the calling function.
5. Where more functions are used, they may be placed in any order.
6. If a function has no parameters, the parameter list must be declared as **void** .Functions return integer value by default.
7. Functions cannot be defined as assignment.
8. A function with void return type cannot be used in the RHS of an assignment statement.
9. Function definition defines the body of the function .it may be placed either after or before the main function.
10. Variables declared in a function are not available to other functions in a program.
11. A function can be called either by value or by reference.
12. Recursion offers a better solution than loops.
13. If a function is to be made to return more than one value at a time, then return these values indirectly by using a call by reference.

10 Use parameter passing by values as far as possible.

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER III

SELF LEARNING MATERIAL

PAPER : MATA 3.4
(Applied Stream)

- Block - I : Mathematical Biology
- Block - II : Dynamical System

Directorate of Open and Distance Learning
University of Kalyani
Kalyani, Nadia
West Bengal, India

Course Preparation Team

1. Mr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani	2. Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

November, 2019

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

Director's Massage

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Sankar Kumar Ghosh, Hon'ble Vice-Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as weel as each and every member of PGBOS (DODL), University of Kalyani, Heartfelt thanks is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Elective Paper

MATA 3.4

Block - I

Marks : 50 (SSE : 40; IA : 10)

Mathematical Biology (Applied Stream)

SINGLE SPECIES POPULATION MODELS

Structure

8.1 INTRODUCTION

Ecology has attracted attention of scientists and philosophers from the early ages of human civilisation. Some of the writings of great Greek philosophers like Hippocrates, Aristotle, etc. dealt with ecological materials although the term "ecology" was not know to them. The word "ecology" was first coined by the german biologist Ernst Haeckel in 1869 to define "the science of the interrelations between living organisms and their environment". The word "ecology" owes its origin to the Greek word "oikos" meaning "house" or "place to live". Because of growing environmental awareness now-a-days, ecology has become a branch of science that is most relevant to everyday life.

The fact that ecology is essentially a mathematical subject is becoming more widely accepted. Population biology or mathematical ecology deals with the increase and fluctuations of populations (e.g. plant population, animal population, or other organic population). The mathematical study of the problems in ecology is not of recent origin. In fact, Lotka (1924) and Volterra (1926) were early pioneers developing foundation work in this field. They established their works on the expression of predator-prey and competing species relations in terms of differential/integral equations.

In this unit we shall first define some fundamental concepts used in ecological studies and then develop mathematical models of some basic principles in ecology dealing with the growth of single species biological populations. We shall talk about two species biological population in Unit 9.

Objectives

After reading this unit, you should be able to

- express population growth processes in a mathematical framework.
- apply your knowledge of differential calculus, integral calculus and differential equations in building mathematical models of population dynamics.

- to solve mathematical models or population aynamics.
- analyse inathematical relations obtained to understand how the change of a population can be predicted

8.2 FUNDAMENTAL CONCEPTS

In nature, an individual living organism of any species does not live in isolation – the organisms live in groups which are called populations. The term population means a group of iridividuals of any one kind of living organism. Ecological studies start at the population level.

The basic characteristic of a population is indicated by its density. The density of a population is its size.in relation to some unit of space; it is generally expressed as the number of individuals or biomass per unit area or volume. For example, 300 trees per acre of land or 2 quintals of fish per acre of water surface or 20,000 bacteria per cubic metre of a test tube, etc..

Since a population changes over time, we are interested in knowing how it is changing or more precisely, what is its time-rate of change which we call the growth-rate. The growth-rate of a population is tlie rate of change of its density or size per unit time; it is determined by the birth-rate and the death-rate. The birth-rate of a population is the maximum production of new individuals per unit time under ideal conditions (i.e. without any ecological limiting factor: reproduction being limited by physiological factors only). Death-rate may be expressed as the number of individuals dying per unit time.

With these few definitions we now proceed to develop an exponential growth population model.

8.3 EXPONENTIAL GROWTH MODEL

The principle of exponential growth for human populations was first propounded. by Thomas R. Malthus (1766-1834) an English clergyman and political economist in the first edition of his famous book entitled An Essay on the Principle of Population published in 1798. Malthus achieved notoriety through this work for publishing that human population grows at a (geometrical) rate that is faster than tlie (arithmetical) rate of growth of the supply of commodities necessary for life. He predicted famine and wars as a consequence.

Let us now discuss the exponential growth model propounded by Malthus.

8.3.1 Formulation of the Model

How does one predict the growth of a population? If we are interested in a single population, we can thinlc of species as being contained in a compartment (a Petrie dish, an island, a country etc.) and study the growth process as one-compartment system. While the population say $x(t)$ is always an integer, it is usually assumed to be large enough so that very little error is introduced in assuming that $x(t)$ is a continuous function. In fact, to avoid this problem, $x(t)$ is often taken to be the population density or to be biomass rather than the number of individuals.

Malthus, while formulating the population growth model made the following three assumptions

i) The population is sufficiently large

ii) Population is homogeneous that is, it is evenly spread over the living space.

iii) There are no limitations to growth, i.e., no limitations of food, space and so on. Population changes only by the occurrence of hirths and deaths. Let us now discuss the model formulated under these conditions.

Let $x(t)(>0)$ be the size of the population at time t and $x(0) = x_0$. Suppose that the population changes only by the occurrence of births and deaths-there is no immigration or emigration. Let $B(t)$ and $D(t)$ denote, respectively, the numbers of births and deaths that have occurred by time t. Then the per capita birth rate b and the death rate m are given by

$$b = \frac{1}{x(t)}\frac{dB}{dt}, \tag{1}$$

$$m = \frac{1}{x(t)}\frac{dD}{dt} \tag{2}$$

and the per capita growth rate of the population at any time t is given by

$$\frac{1}{x(t)}\frac{d}{dt}(B-D) = b-m \tag{3}$$

$$\text{or, } \frac{1}{x(t)}\frac{dx(t)}{dt} = b-m = r \text{ (constant) say}$$

$$\text{or, } \frac{dx(t)}{dt} = rx(t), \text{ where } x(0) = x_0 \tag{4}$$

where constant r **represents the net growth rate.**

Eqn.(4) is the model equation for the population growth as given by Malthus. Let us now solve this equation and see what it represents.

8.3.2 Solution And Interpretation

The simple ordinary differential Eqn.(4) can be solved by the method of separation of variables,

We have

$$\frac{dx(t)}{dt} = rx(t), x(0) = x_0 \tag{5}$$

Integrating this equation, we obtain

$$\ln x(t) = C + rt, C \text{ being } \textit{a} \text{ constant}$$

To obtain C we use the initial condition $x(0) = x_0$, and get

$$\ln x_0 = C \tag{6}$$

Solution to Eqn.(5) then reduces to

$$\ln\frac{x(t)}{x_0} = rt = \ln e^{rt}$$

$$\text{or, } x(t) = x_0 e^{rt}. \tag{7}$$

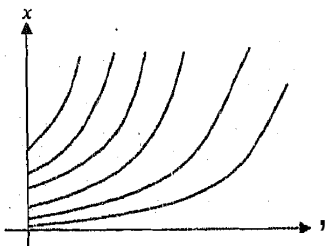


Fig. 1

Eqn. **(7).**gives the population size at any time t. If the net growth rate $r > 0, x(t)$ grows exponentially without any bound as shown in Fig. **1 For** $r < 0, x(t) \to 0$ as $t \to \infty$, implying that the population is ultimately driven to extinction. Both these outcomes are extreme and are not found to occur in the nature. In this sense the Malthus model **has** severe limitations which we shall discuss now, but before that, let us solve this example.

Example 1: In a population of birds, the proportionate birth rate and the proportionate death rate are both constant, being 0.45 per year and 0.65 per year respectively. Formulate a model of the population and discuss its long-term behaviour.

Solution: Let $x(t)$ denotes the size of the population at any time $t > 0$. The per capita birth-rate $b = 0.45$ and per capita death-rate $m = 0.65$. Hence

$$\begin{aligned} \frac{dx}{dt} &= (0.45 - 0.65)x \\ &= -0.2x \end{aligned} \tag{8}$$

Integrating Eqn.(8) we get

$$x(t) = x_0 e^{-0.2t}$$

where x_0 is the initial size of the bird population.

Eqn.(8) gives the size of the bird population at any time t. Since $e^{-0.2t} \longrightarrow 0$ as $t \longrightarrow \infty, x(t) \longrightarrow 0$ whatever (finite) value is assigned to x_0 . This shows that the bird population goes to extinction in the long run.

8.3.3 Limitations

Under ideal conditions when the availability of space, food and other resources do not inhibit growth, many biological populations are observed to grow initially at an approximately exponential rate After some time, when the population size becomes considerably large, there is lack of food, space and other resources; also there is pollution due to overcrowding. All these consequences are collectively called "**crowding** effects". The crowding effects force the growth rate to decline. These considerations make it clear that the growth rate r cannot. be constant, but must depend on the size or density of the population. This is where the limitations of the Malthus model precisely lie.

The above discussion suggests that Eqn.(4) should be modified as

$$\frac{dx}{dt} = r(x)x(t), x(0) = x_0 \tag{9}$$

where $r(x)$ after certain stage decreases as x increases.

When $r(x)$ is a decreasing function of x, the model is said to describe a process of "feed back" or "compensation". The natural biological population9 usually exhibit compensatory growth processes.

It is thus clear that Eqn.(4) does not provide a very accurate model for the population growth when the population itself is very large. Therefore, there is a need to improve this model. In the next section we shall develop a model called Logistic model which takes care of the large population.

And now some exercises for you.

E1) In a population of birds, the proportionate birth rate and death rate are both constant, being 0.48 per year and 0.65 per year respectively. Immigration occurs at a constant rate of 2000 birds and emigration at a constant rate of 1000 birds per year, Use these assumptions to formulate a model of the population. Solve the model and describe the long-term behaviour of the population in the two cases when the initial population is 3000 or 8000.

E2) The population of fish in a reservoir is affected by both fishing and restocking. The proportionate birth rate is constant at 0.6 per year and the proportionate death rate is constant at 0.65 per year. The reservoir

is restocked at a constant rate of 4000 fish per year and fishermen are allowed to catch 3500 fish per year.

Use these assumptions to derive a model for the population. Solve the model and describe the long-term behaviour of the fish population in the two cases when the initial population is 5000 or 15000.

We shall now discuss the **Logistic** Model based on a density dependent, compensatory growth process.

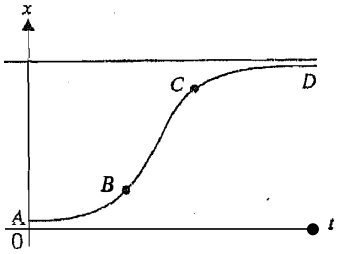


Fig.2

8.4 LOGISTIC MODEL

When a population is growing in a limited space, the density gradually rises until eventually the presence of other organisms reduce the fertility and longevity of the population. This reduce the rate of increase of the population until eventually the population ceases to grow. The growth curve defined by such a population follows sigmoid, or S-shaped pattern when density is plotted against time (see Fig. 2). This curve was first suggested to describe the growth of human populatioiis by PF. Verhulst in 1838. The sigmoid curve arises due to greater and greater action of detrimental factors (environmental resistance) as the density of population increases. The simplest, case that can be conceived is the one in which the detrimental factors are linearly proportional to the density. Such simple or ideal growth form is called **"logistic"** and the corresponding growth equation is called the **"logistic** equation".

If you, look at the shape of the curve in Fig.2 you will notice that the curve consists of three different patterns AB, BC and CD. From A to B the curve gradually rises, from B to C it is almost; an exponential increase and from C to D it gets flattened. This curve is found to represent adequately the population growth which has steady growth initially until the growth rate is reduced due to various factors like crowding effects, epidemics etc. and ultimately tending almost to zero. In other words, we can say that ultimately the population gets stabilized/reaches an equilibrium value without any appreciable increase or decrease. We now take up mathematical formulation of the logistic model.

8.4.1 Formulation

Assuming $r(x)$ to be positive and putting $r(x) = r_1\left(1 - \frac{x}{K}\right)$ in Eqn.(9) where, constants $r_1 > 0$ and $K > 0$ we get the Verhulst's famous "logistic equation".

$$\frac{dx}{dt} = r_1 x\left(1 - \frac{x}{K}\right), x(0) = x_0 \tag{10}$$

Since $r'(x) = -\frac{r_1}{K} < 0$ for all $x > 0$, the per capita growth rate $r(x)$ declines as the density x increases. This decrease in $r(x)$ is brought about by environmental resistance term $\frac{x}{K}$ which is linearly proportional to the density. [Since $r(x) \approx r_1$ for small x, r is called the "intrinsic **growth rate**" i.e., growth rate free from environmental constraints.]

Note that Eqn.(10) is non-linear, first order equation. It is easy to solve it by the method of separation of variables. Before we do that let us discuss the qualitative behaviour of the solution by using geometric reasoning.

The graph of $\frac{dx}{dt}$ against x, where $\frac{dx}{dt}$ is given by Eqn.(10) gives the graph of the logistic growth function as shown in. Fig.3. The graph is a parabola with

intercepts at $(0,0)$ and (K,0) and with vertex at $\left(\frac{K}{2}, r\frac{K}{4}\right)$

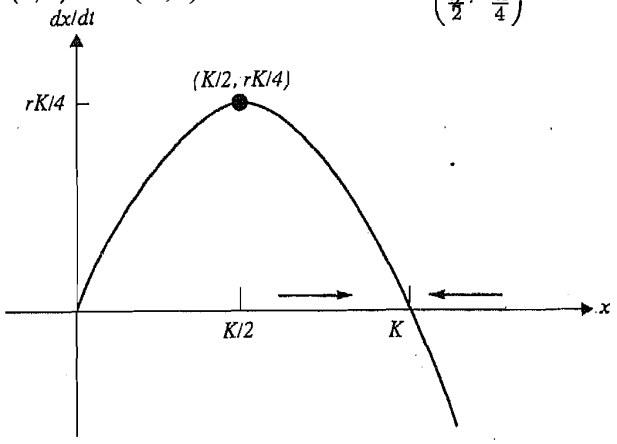


Fig.3: The logistic growth function

When $0 < x < K$, $\frac{dx}{dt} > 0$ so that x increases towards K. When $x > K$, $\frac{dx}{dt} < 0$ so that x decreases towards K.

This shows that the population level $x(t)$ always approaches K.

We can infact, express this by writing

$$\lim_{t\to\infty} x(t) = K \text{ provided } x_0 > 0. \tag{11}$$

A constant solution of a differential equation is called an **equilibrium solution.**

If $x = 0$ or $x = K$, then $\frac{dx}{dt} = 0$ and $x(t)$ does not change. The constant solutions $x = 0$ and $x = K$ are called equilibrium solutions. Corresponding to them the points $x = 0$ and $x = K$ are called equilibrium points or critical points. You would notice in Unit 9 when we discuss the stability of the critical points in detail that the constant K defined by Eqn.(11) is an asymptotically stable equilibrium. It is a "saturation level" or "upper limit" of the population. It is called the 'carrying capacity' of the population- the maximum number of individuals that can persist under the conditions specified.

In many situations it is sufficient to have the qualitative information about the solution $x(t)$ of Eqn.(10). However, if we wish to have a more detailed description of logistic growth - for example, if we wish to know the population at some particular time, then we must solve Eqn.(10). Let us now do that.

8.4.2 Solution and Interpretation

Consider Eqn.(10), viz.,

$$\frac{dx}{dt} = r_1 x\left(1 - \frac{x}{K}\right), x(0) = x_0$$

It can be easily solved by the method of separation of variables, by writing it in the form

$$\frac{Kdx}{x(K-x)} = r_1 dt, \tag{12}$$

We can write Eqn.(12) in the form

$$\left[\frac{1}{x} + \frac{1}{K-x}\right] dx = r_1 dt. \tag{13}$$

If we assume that $x < K$, then Eqn.(13) on integration yields

$$\ln x - \ln(K-x) = r_1 t + C, \text{ C being a constant} \tag{14}$$

$$\text{or, } \ln\frac{x}{K-x} = r_1 t + C. \tag{15}$$

Using the initial condition $x(0) = x_0$, we obtain $C = \ln\frac{x_0}{K-x_0}$.
Therefore,

$$\ln\frac{x}{K-x} = \ln e^{r_1 t} + \ln\frac{x_0}{K-x_0}$$

$$= \ln\frac{x_0 e^{r_1 t}}{K-x_0} \quad (16)$$

$$\text{o} \quad \frac{x}{K-x} = \frac{x_0 e^{r_1 t}}{K-x_0} \quad (17)$$

$$\text{or,} x\left[(K-x_0) + x_0 e^{r_1 t}\right] = K x_0 e^{r_1 t}$$

$$\text{or,} x = \frac{K}{1 + \left(\frac{K-x_0}{x_0}\right) e^{-r_1 t}} \quad (18)$$

$$\text{Therefore,} x(t) = \frac{K}{1 + C_1 e^{-rt}} \quad (19)$$

$$\text{where } C_1 = \frac{K-x_0}{x_0} \text{ is a constant.} \quad (20)$$

You may observe here that we made the assumption that $x < K$ in order to derive Eqn.(19). But this restriction is unnecessary, because you can easily verify that Eqn.(19) gives the solution of the logistic equation whether $x < K$ or $x \geq K$. We are leaving it for you to verify

---

E3) Verify that solution of Eqn.(10) for $x \geq K$ is given by Eqn.(19).

Thus the solution of Eqn.(10) as given by **Eqn.(19)represents** the size of the population at any time t. It is evident from Eqn.(19) that $x(t) \to K$ as $t \to \infty$. Thus a population that satisfies the logistic equation is not like a naturally growing population; it does not grow without bound: but instead approaches the finite limiting population K as $t \to \infty$. But because $\frac{dx}{dt} > 0$ in this case, we see that population is steadily increasing.

Moreover, differentiating Eqn.(10) with respect to t, we have

$$\frac{d^2x}{dt^2} = r\left[\frac{dx}{dt} - \frac{2x}{K}\frac{dx}{dt}\right] \quad (21)$$

$$= \frac{r}{K}(K-2x)\frac{dx}{dt}$$

$$= \frac{r^2}{K^2} x (K-x)(K-2x). \text{ (using Eqn.(10)).} \quad (22)$$

If $(K-2x) > 0$, we have $K - x > x > 0$. Then $\frac{d}{dt}\left(\frac{dx}{dt}\right) > 0$. So that the rate of increase $\frac{dx}{dt}$ increases with time. This shows that there is an accelerated growth of the population in the range $0 < x < \frac{K}{2}$.

On the other hand, if $\frac{K}{2} < x < K$, then $K - 2x < 0$ and $K - x > 0$, so that $\frac{dx}{dt}$ is a decreasing function of time. Thus there is a retarded growth of the population in range $\frac{K}{2} < x < K$.

We have shown the two typical solution curves $x(t)$ of the logistic Eqn.(10) in Fig.4.

The graphs of solution of Eqn.(10) must have the general shape shown in Fig.4, regardless of the values of r and K. The horizontal lines are the equilibrium solutions $x(t) = 0$ and $x(t) = K$. If the initial population level $x_0 > K$, $x(t)$ monotonically decreases towards K; the upper curve depicts this situation. The lower curve, with its characteristic "sigmoid" or "ogive" shape, is usually referred to as the "logistic growth curve".

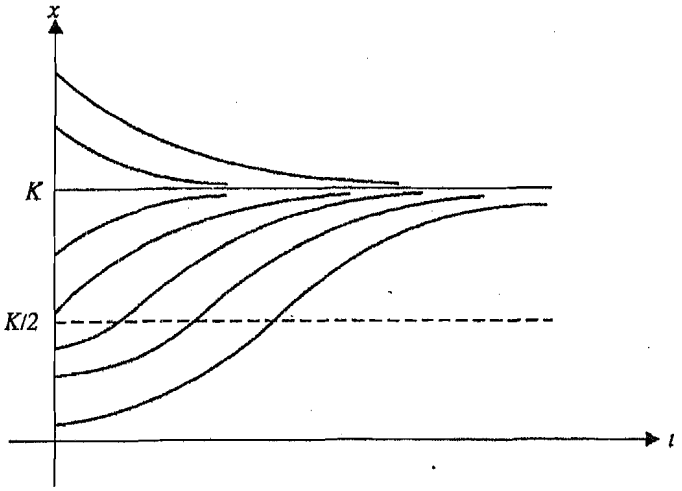


Fig.4: The solution curves of the logistic growth equation

The logistic curve rises at an increasing rate to start with, like an exponential curve, and then gradually slows down and finally flattens out to approach the horizontal line x = K as t becomes very large, The time-period before the population reaches half its equilibrium value (K/2) is a period of "accelerated growth". Thereafter, the rate of growth diminishes and gradually becomes zero.

If you compare Fig.1 and Fig.4, you would notice that solutions of the non-linear Eqn.(10) are strikingly different from those of the linear Eqn.(5), at least for large values of t. Regardless of the value as K, that is, no matter how small the non-linear term in Eqn.(10), solution of the equation approach a finite value of $t \to \infty$, whereas solution of Eqn.(5) grow (exponentially) without bound as $t \to \infty$. Thus, even a tiny non-linear term in the differential equation has a decisive effect on the solution for large t.

Let us now consider the following examples,

Example 2: the logistic model has been applied to the natural growth of the halibut population in certain areas of the Pacific Ocean. Let x(t), measured in kilograms he the total mass, or biomass, of the halibut population at time t. The parameters in the logistic equation are estimated to have the values $r_1 = 0.71$ / year and $K = 80.5 \times 10^6$ Kg. If the initial biomass is $x_0 = 0.25K$, find the biomass two years latter. Also find the time τ for which $x(\tau) = 0.75K$.

Solution: We can rewrite Eqn.(19) in the form

$$\frac{x(t)}{K} = \frac{\frac{x_0}{K}}{\left(\frac{x_0}{K}\right) + \left[1 - \left(\frac{x_0}{K}\right)\right] e^{-r_1 t}} \tag{23}$$

Using the data given in the problem we find that

$$\frac{x(2)}{K} = \frac{0.25}{0.25 + 0.75e^{-1.42}} \cong 0.5797$$

Hence $x(2) \cong 46.7 \times 10^6$Kg.

To find T, we solve Eqn.(23) for t and obtain

$$e^{-r_1 t} = \frac{(x_0/K)\left[1 - \left(\frac{x}{K}\right)\right]}{(x/K)\left[1 - \left(\frac{x_0}{K}\right)\right]}$$

Hence

$$t = -\frac{1}{r_1} \ln \frac{(x_0/K)\left[1 - \left(\frac{x}{K}\right)\right]}{(x/K)\left[1 - \left(\frac{x_0}{K}\right)\right]}$$

Using $r_1 = 0.71$, $\frac{x_0}{K} = 0.25$ and $\frac{x}{K} = 0.75$, we find

$$\tau = -\frac{1}{0.71}\ln\frac{(0.25)(0.25)}{(0.75)(0.75)} = \frac{1}{0.71}\ln 9 \cong 3.095 \text{ years.}$$

Example **3:** The population of fish in a large lake has been stable for some time. Prior to this situation the population was decreasing from an initially relatively high level. When the population was 4000, the proportionate birth rate was 10% and the proportionate death rate was 70% . When the population **was** 3000, the proportionate birth rate was 30% and the proportionate death rate was 60%.

A model of the population is based on the following assumptions:

(i) there is no exploitation and no restocking;

(ii) the proportionate birth rate is a decreasing linear function of the population;

(iii) the proportionate death rate is an increasing linear function of the population.

Show that the rnodel based on these assumptions and the above data predicts that population falls according to the logistic model; find the equilibrium population size.

Restocking of the lake now takes place at a rate of 20% of the population per year. Find the equilibrium population in this case.

Solution: Let $x(t)$ denotes the size of the fish population at any time $t > 0$. By the given conditions, the proportionate birth rate is

$$b(x) = \lambda_1 - \mu_1 x \tag{24}$$

and the proportionate death rate is

$$m(x) = \lambda_2 + \mu_2 x \tag{25}$$

where $\lambda_i, \mu_i (i = 1, 2)$ are all positive constants.

Then the net proportionate growth rate is

$$b(x) - m(x) = \lambda_1 - \lambda_2 - (\mu_1 + \mu_2)x \tag{26}$$

$$= A - px \tag{27}$$

where $\lambda = \lambda_1 - \lambda_2$ and $\mu = \mu_1 + \mu_2$ are constants. Here λ may have any sign but μ is always positive.
Using the given conditions,

$$\lambda - 4000\mu = \frac{1}{10} - \frac{7}{10} = -\frac{6}{10} \tag{28}$$

$$\text{and } \lambda - 3000\mu = \frac{3}{10} - \frac{6}{10} = -\frac{3}{10} \tag{29}$$

Subtracting, $-1000\mu = -\frac{3}{10}$
Therefore,

$$\mu = 3 \times 10^{-4}$$

$$\text{and } \lambda = 3000 \times \mu = \frac{3}{10} = 0.6.$$

Hence, $b(x) - m(x) = \lambda - ux = 0.6 - 3 \times 10^{-4}x$
This implies that

$$\frac{1}{x}\frac{dx}{dt} = b(x) - m(x) = \lambda - ux \tag{30}$$

$$\text{or, } \frac{dx}{dt} = x(\lambda - \mu x) \tag{31}$$

This is the logistic growth equation with carrying capacity $\frac{\lambda}{\mu}$ for the population. As we have seen in the 'logistic growth model', this carrying capacity is the stable equilibrium population.

Hence the equilibrium population level is $\frac{\lambda}{\mu} = \frac{0.6}{3\times10^{-4}} = 2000$.

When restocking of the population is allowed, the governing Eqn.(31) is modified to the form

$$\begin{aligned}\frac{dx}{dt} &= \left(0.6 - 3\times10^{-4}x\right)x + 0.2x \\ &= \left(0.8 - 3\times10^{-4}x\right)x \end{aligned} \tag{32}$$

This also represents the logistic law of growth with the new carrying capacity $= \frac{0.8}{3\times10^{-4}} = 2227$ approximately.

Hence the new equilibrium level of the fish population after restocking is, 2667.

And now a few exercises for you.

E4) A colony of birds has a stable population. Prior to this situation the population iucreasod from an initially low level. When the population was 10,000 the proportionate birth rate was 50% per year and the proportionate death rate was 10% per year. When the population was 20,000 the proportionate birth rate was 30% and the proportionate death rate was 20%.

A model of the population is based on the following assumptions:

(i) there is no migration and no exploitation (such as shooting);

(ii) the proportionate birth rate is a decreasing linear function of the population;

(iii) the proportionate death rate is an increasing linear function of population.

Show that a model based on these assumptions and above data predicts that the population grows according to the logistic model and find the stable population size.

Shooting of the birds is now allowed at a rate of 20% of the population per year. Find the new equilibrium population.

E5) For the model

$$\frac{dx}{dt} = r_1x\left(1 - \frac{x}{K}\right) - Ex, x(0) = K, \tag{33}$$

where r_1, E and K are constant, determine $x(t)$ explicitly. Verify from the form of the solution that for $x > k(1 - \frac{E}{s})$ if $E \leq r_1$, then $x(t) \to K(1 - E/r_1)$ as $t \to \infty$ whereas if $E > r_1$, then $x(t) \to 0$ exponentially as $t \to \infty$.

We shall now discuss several limitations of the logistic model.

8.4.3 Limitations

(i) The logistic model is not suitable for a population of small size. The reason is obvious; for small x, $r_1x(1 - \frac{x}{K}) \approx r_1x$ neglecting the second-order small quantity x^2. The logistic equation reduces to that of Malthus for small x.

(ii) It has been observed in both laboratory and natural population that the growth of many populations (of microorganisms, plants and animals) exhibit a sigmoid pattern, although such populations do **not** increase according to the logistic equation. Almost any equation in which the negative factors increase in some manner with density will yield sigmoid curves. The S-shaped logistic curve is an adequate description for the laboratory growth of paramecium, yeast and other organisms with simple life cycles. Population growth in organisms with more complex life cycles seldom follows the logistic very closely.

(iii) The basic assumption in the logistic model that,"the environmental resistance increases linearly with density" is violated in many growing populations when tested through direct experiments. This holds for populations with simple life histories, as for example, yeast growing in a limited space.

(iv) Some populations, fluctuate periodically between,two values. These fluctuations occur when certain populations reach a sufficiently high density, they become susceptible to epidemics. The epidemic brings the population down to a lower value where it again begins to increase, until when it is large enough, the epidemic strikes again. But any kind of fluctuation is ruled out in a logistic curve.

In addition to the above, the following limitations pertain to both the population models considered in this unit.

(1) The models of population growth operate in a closed system, without input or output. Only self-crowding or other internal factors are modeled. The real world consists of open systems in which the input and output environments play major roles in the behaviour of the component considered. This short coming is especially apparent, when it comes to modelling the growth form of human populations. Clearly, the technological developmenis, pollution consideration and sociological trends have significant influence on the coefficients r and K.

(2) We have considered the population as made up of one homogeneous group of individuals. We should subdivide it into different age groups, into males and females since tlie reproduction rate in a population usually depends more on the number of females than on the number of males.

8.5 EXTENSION OF THE LOGISTIC MODEL

In the logistic model just discussed the function r(x) is positive and linear. We shall now consider a simple extension of this model with an assumption of r(x) being a non-linear function of x. Three types of density dependent r(x) are depicted in Fig.5.

Fig.5 (a) shows that logistic growth decreases linearly with density i.e. $r(x) = r_1 (1 - \frac{r}{k})$ which corresponds to the model discussed in Sec.8.4.1 (ref. Eqn.(10))

Fig.5(b) corresponds to the function which has a maximum at an intermediate point. The function r(x) corresponding to this case is of the form

$$r(x) = a_1 + a_2x + a_3x^2 \qquad (34)$$

with $a_2 > 0$ and $a_3 < 0$. This represents a situation in which a population has a maximal intrinsic growth rate at intermediate density. This is known

as the Allee effect.

The general character of this density dependent function $r(x)$ can be summarised by the inequalities $r(x) > 0.(x < \eta)$ and $r(x) < 0(x > \eta)$ (where η is the density for optimal reproduction) $0 < \eta < k$, where k is the carrying capacity.

Fig.5 (c) corresponds to $r(x) = -\ln x$. You may observe that this is a non-linear curve which becomes negative for $x > 1$ and not defined at $x = 0$. This represents a situation in which there is a negative logarithmic dependence of growth rate on population size. This relation is not valid for very small populations since function is not defined at $x = 0$. This model is known as the Gompertz law, which is used mainly for depicting the growth of solid tumors. Let us consider a simple example of an Allee effect.

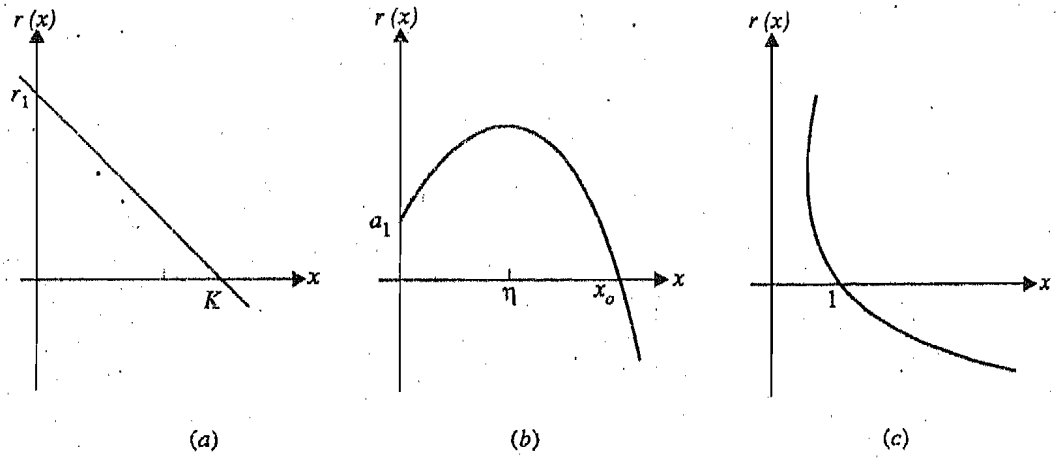


Fig.5

Example 4 Discuss the Allee effect given that

$$\frac{dx}{dt} = x\left[r_0 - \alpha(x-\eta)^2\right]\left(\eta < \sqrt{\frac{r_0}{\alpha}}\right)$$

where r_0 and η are positive constants. Can you relate $r(x)$ corresponding to this situation with Fig.5(b).

Solution: Here $r(x) = r_0 - \alpha(x-\eta)^2$. Comparing this with Eqn.(34) we get $a_1 = r_0 - \alpha\eta^2, a_2 = 2\alpha\eta > 0, a_3 = -a < 0$. As in Fig.5 (b) this would be an inverted parabola which intersects the axis at $r(x) = a_1 = r_0 - aq^2$ it has a maximum of r_0 when $x = \eta$ and drops below zero when $x > x_0 = \eta \pm \sqrt{\frac{r_0}{\alpha}}$. Thus for densities above x_0, the population begins to decline. $x = x_0$ is a stable equilibrium for the population.

You may now try this exercise.

E6) A solid turner usually grows at a declining rate because its interior has no access to oxygen and other necessary substances that the circulation supplies. This has been modeled empirically by the Gompetz growth law

$$\frac{dN}{dt} = \gamma N \text{ where } \frac{d\gamma}{dt} = -\alpha\gamma$$

γ is the effective tumor growth rate, which will decrease exponentially

by this assumption. Show that equivalent ways of writing this are

$$\frac{dN}{dt} \doteq \gamma_0 e^{-\alpha t} \qquad N = (-\alpha \ln N)N$$

We now end this unit by giving a summary of what we have covered in it.

8.6 SUMMARY

In this unit, we have covered the following:

(1) Mathematical study of the problems in ecology deals with the increase and fluctuations of populations. (e.g. plant: population, animal population or other organic population).

(2) Malthus model and Logistic model deal with the growth of a single species biological populations.

(3) For a population of size $x(t)(> 0)$ at any time t, the Malthus model is given by the equation

$$\frac{dx(t)}{dt} = rx(t), x(0) = x_0$$

where $r > 0$ is a constant and is the growth rate of the population.

(4) Malthus model works well only for small populations. For large populations the growth rate r cannot be constant, but'depends on the size or density of the population.

(5) For large populations the logistic model given by the equation

$$\frac{dx}{dt} = r_1 x\left(1 - \frac{x}{K}\right), x(0) = x_0$$

(the constant $K > 0$ being the saturation level of the population) gives a type of growth which follows an S-shaped or sigmoid pattern when density is plotted against time.

(6) In nature, growth of many populations of plants and animals exhibit a sigmoid pattern though they do not increase according to the logistic equation.

8.7 SOLUTIONS/ANSWERS

E1) The differential equation describing the growth of the population is

$$\frac{dx}{dt} = -0.2x + 2000 - 1000 = -0.2x + 1000 \qquad (35)$$
$$x(0) = x_0$$

This can be written as

$$\frac{-0.2dx}{1000 - 0.2x} = -0.2dt, x(0) = x_0 \qquad (36)$$

Integrating, $\ln|(1000 - 0.2x)| = -0.2t + \ln|C_1|$
Therefore, $1000 - 0.2x = C_1 e^{-0.2t}$

Using the initial condition and finding the value of C, we obtain
$x(t) = 5000 - x_0 e^{-0.2t}$

As $t \to \infty, e^{-0.2t} \to 0$ and hence $x(t) \to 5000 = \bar{x}$ (say). Thus the stable population level is 5000 whatever (finite) value the initial population level x_0 may have.

Hence, if the initial population be 3000, it rises upto 5000; if the initial population be 8000, it ultimately drops to 5000.

E2) Proceeding exactly as above obtain the stable population level as 10,000.

E3) Eqn.(13) can be written as

$$\left[\frac{1}{x} - \frac{1}{x-K}\right] dx = r_1 dt$$

For $x \geq K$ we get

$$\ln\frac{x}{x-K} = r_1 t + C$$

Using initial condition $x(0) = x_0$, we obtain

$$\ln\frac{x}{x-K} = \ln e^{r_1 t} + \ln\frac{x_0}{x_0 - K}$$

$$\text{or, } \frac{x}{x-K} = \frac{x_0 e^{r_1 t}}{x_0 - K}$$

$$\text{or, } x = \frac{K}{1 + \left(\frac{K-x_0}{x_0}\right) e^{-r_1 t}}$$

E4) Using the same notations as in Example-2, we have

$$\lambda - 10,000\mu = \frac{1}{2} - \frac{1}{10} = \frac{4}{10}$$

$$\text{and } \lambda - 20,000\mu = \frac{3}{10} - \frac{2}{10} = \frac{1}{10}$$

$$\text{Subtracting, } 10,000\mu = \frac{4}{10} - \frac{1}{10} = \frac{1}{10}$$

$$\text{Therefore, } \mu = 3 \times 10^{-5}$$

$$\text{Therefore, } \lambda = \frac{4}{10} + 10,000$$

$$\mu = \frac{4}{10} + \left(\frac{3}{10}\right) = 0.7$$

$$\text{Therefore } \lambda - \mu x = \left(0.7 - 3 \times 10^{-5} x\right)$$

Hence the equation governing,the growth of the population is

$$\frac{dx}{dt} = \left(0.7 - 3 \times 10^{-5} x\right) x \qquad (37)$$

This represents logistic law of growth with carrying capacity

$$= \frac{0.7}{3 \times 10^{-5}} = 2.3 \times 10^4 = 23333 \qquad (38)$$

Hence the equilibrium population level is equal to 23333. When shooting of the birds is allowed, the Eqn.(26) is modified into the form

$$\frac{dx}{dt} = \left(0.7 - 3 \times 10^{-5} x\right) x - 0.2x \qquad (39)$$

$$= \left(0.5 - 3 \times 10^{-5} x\right) x \qquad (40)$$

This also represents a logistic law of growth with carrying capacity $= \frac{0.5}{3 \times 10^{-5}} = 16667$. Hence the new equilibrium population level after shooting is allowed is 16667.

E5)

$$\frac{dx}{dt} = r_1\left(1-\frac{x}{k}\right) - E(x), x(0) = k \quad (41)$$

$$= \frac{r_1 x}{k}\left[\frac{k(r_1-E}{r_1} - x\right] \quad (42)$$

$$\left[\frac{1}{x} + \frac{1}{\frac{k(r_1-E}{r_1} - x}\right] dx = (r_1 - E)dt \quad (43)$$

(44)

Integrating and using the boundary condition $x(0) = k$, we get for $x > k\left(1-\frac{E}{r}\right)$

$$x(t) = \frac{k\left(1-\frac{E}{r_1}\right)}{1-\frac{E}{r_1}e^{-(r_1-E)t}}$$

Now for $E \leq r_1, x(t) \to k\left(1-\frac{E}{r_1}\right)$ as $t \to \infty$ and for $eE > r_1, x(t) \to 0$ as $t \to \infty$.

E6) Hint: Use the fact, $\frac{1}{N}\frac{dN}{dt} = \frac{d}{dt}(\ln N)$

SINGLE SPECIES POPULATION MODELS

Structure

INTRODUCTION

Ecological studies have assumed greater significance in view of human concern for environmental degradation witnessed during the last four decades or so. Ecological modelling has been revived with great interest in the recent years due to the following:

i) Measurements in ecosystems are becoming more accurate and precise due to instrumental facilities.

ii) Satellite data has the potential to provide estimates of vegetation in the water body.

iii) Present capability of models to deal with a large system of non-linear equation due to improving computing facilities.

Mathematical ecology can be described as the study of inter dependence of several species in a variable eco-system. The approach based on modelling pre-supposes certain degree of idealization so that mathematical techniques can be brought into play. Although the so-called idealized models might not capture the full diversity of dynamic environmental landscape, they promise to provide an insight into growth pattern of various populations. This knowledge has been exploited fruitfully in management of renewable resources, ecological control of pests, evolution of pesticide resistant strains, etc. The prospective application of studies on bacteria and viruses to various bio-medical sciences has attracted many researchers to this area.

In this unit, we shall discuss some discrete and continuous models of single species population growth. In Sec. 4.2, we give some basic concepts of modelling population growth. Many species have no overlap between successive generations and so population growth for such species is in discrete steps. Some discrete population models are discussed in Sec. 4.3. For continuous population growth the delay model incorporating the effect of time delay in growth and reproduction of individuals in the population is discussed in Sec.4.4. For understanding the discussion in this unit the knowledge of the equilibrium points of a difference equation and their stability analysis is

essential. We have given these details in an appendix at the end of the unit. You must go through the appendix while reading this unit.

Objectives

After studying this unit, you should be able to:

- apply the knowledge of calculus, analysis, differential equations and difference equations in building mathematical models of population dynamics both for discrete as well as continuous population growth;
- analyse the models both quantitatively and qualitatively;
- carry out the stability analysis of models of population growth.

BASIC CONCEPTS OF MATHEMATICAL MODELLING IN POPULATION DYNAMICS

A mathematical model in population dynamics represents the pattern of growth of a given population in the presence of various environmental factors. Real life eco-system consists of many species, which interact with each other in many different ways. In addition, various environmental factors like migration, territorial behaviour and climatic fluctuations also play important role. But all these factors cannot be accommodated in one model since otherwise the model would become too complex for solution by known mathematical techniques. The crucial decision, while modelling a given eco-system, lies in the choice of most relevant variables. The simplest approach is to include one factor at a time and subsequently modify the model by adding another factor.

Thus, a good mathematical model in population dynamics relies on thorough understanding and appreciation of biological problems, mathematical description of the relevant biological phenomena and subsequent derivation and interpretation of mathematical results for the use of biologists. It acts like a bridge in the description of theoretical details of a population and their mathematical abstractization.

We shall now discuss in brief some of the concepts which we shall be using for the study in this unit.

Classification

The mathematical models in population dynamics can be classified based on various factors like nature of growth rate, its fluctuation, environmental factors, size of the population and so on.

On the basis of growth rate, the population models can be further classified as follows:

1) **The deterministic models** which presuppose a fairly large population size and ignore random fluctuation in the environmental factors with time.
2) **The stochastic models** which are more appropriate when populations are small or when there is significant random fluctuation in the parameters.

The deterministic models are further subdivided into two subclasses:

a) **Continuous deterministic models** which are employed when population is very large and its growth rate is also fast so that variables can be modelled by continuous functions and the growth rate by their derivatives. These models lead to differential equations. (Their typical examples are those dealing with insect population, grass and zooplanktons).

b) **Discrete deterministic models** which are applicable to the cases when moderately large populations such as humans, large animals like lions, elephant are modelled. In such models the variations in population are studied by means of difference equations.

Analysis of Models

The models in population dynamics can be either quantitative or qualitative. Quantitative analysis of models involves consideration of actual parameters and prediction in terms of numbers. On the other hand, qualitative analysis concentrates on study of the nature and pattern of growth and their dependency on parameters relevant to the model. Thus while a quantitative study may require the knowledge of numerical analysis, the qualitative study of a model needs comprehensive knowledge of stability analysis techniques of difference/differential equations.

Stability Analysis

Consider a system involving a number of variables $x_1, x_2, x_3, \ldots, x_n$ denoting the densities of species composing the system. The state of the system can then be represented by a point in n-dimensional phase space. To each point in this space we can attach a vector or an arrow indicating how the system moves. These vectors can be joined to form trajectories which show the long term behaviour of the system.

A **stationary point** or an **equilibrium point** of a system is one which has associated with it a vector of zero length. That is, if a system is at a stationary point at an instant it will remain at that point at the next instant.

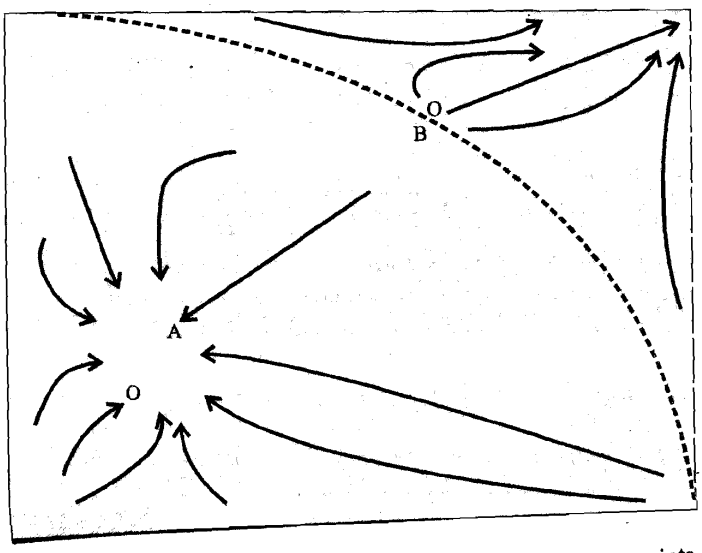


Fig. 1: A System with A as stable and B as unstable equilibrium points.

The equilibrium point may be stable or unstable. If a system slightly disturbed from equilibrium point returns to that point, then, the equilibrium point is said to be **stable**. More precisely, it is said to have neighbourhood stability. On the other hand when a system disturbed slightly from an equilibrium point continues to move further away from that point, then the equilibrium point is said to be **unstable**. Fig. 1 on previous page shows the trajectories and equilibrium points A and B of a system.

If a system moves to a particular equilibrium point irrespective of the point from where it started that equilibrium point is said to have **global stability**. You may note that in Fig. 1 the equilibrium point A has neighborhood stability but not global stability.

The stability of equilibrium point can well be understood by the analogy of a landscape in which troughs represent stable points and peaks represent unstable ones and the behaviour of the system is represented by a rolling ball. Mathematically speaking the neighborhood stability analysis is the easiest one. By considering only small displacements from an equilibrium point, it is possible to linearize the equations and discuss their behaviour. We shall be using this technique for the discussion of the models considered in this unit.

In the next section we shall discuss various discrete population models of single species.

DISCRETE POPULATION MODELS

In many species, the populations do not overlap. There is a fixed interval of time between successive generations and so population grows in discrete steps. The size of steps can vary from species to species. In general the population of a given generation depends on that of the preceding generations. Thus if we scale the time step to be 1 and denote the initial population by x_0 and the population of nth generation at time n by x_n then population of next generation x_{n+1} at time $(n+1)$ can be written in the form of following difference equation:

$$x_{n+1} = x_n \; F(x_n) = f(x_n) \qquad (1)$$

where $f(x_n)$ is, in general, a non-linear function of x_n. Thus, study of discrete population models is equivalent to the study of difference equations of the form given by Eqn. (1).

The efficiency of a model representing a specific population growth lies in determining the appropriate form of $F(x_n)$ to reflect the factual situation of the species in question. The function $F(x_n)$ is called **recruitment function.** Taking various forms for $F(x_n)$, we can construct different models of population growth. We shall now discuss them one-by-one.

Exponential/Constant Growth Model

In the simplest of the cases, we consider the species for which the birth and death rates are constant. For example, many insects and annual desert plants reproduce once and then they fade away. The surviving offspring forms the

basis for the next generation. Thus, the population increases or decreases by the same amount each year. In such a situation we can take the recruitment function as constant and arrive at linear or exponential model of discrete population. This model was first propounded by Thomas R. Malthus (1766-1834), an English clergyman and political economist and hence is also called **Malthusian growth model**.

Formulation

Let us assume that the population is closed i.e., it changes only by births and deaths and there is no migration into or out of the region. We further suppose that the birth rate b and death rate d are constants. This means that the birth and death of individuals in a given population are proportional to the population size.
Then

$$x_{n+1} - x_n = (b - d)\, x_n$$

or

$$\begin{aligned} x_{n+1} &= x_n + (b - d)\, x_n \\ &= (1 + b - d)\, x_n \end{aligned} \qquad (2)$$

Substituting $r = (1 + b - d)$ in Eqn. (2) we get the following linear homogeneous difference equation

$$x_{n+1} = r\, x_n \qquad (3)$$

where constant r represents the **net growth rate**, also called **net reproductive rate**. Eqn. (3) together with the prescribed initial population size x_0 determines the population size in each generation.

Solution and Interpretation

By a solution of the difference equation with initial value x_0, we mean a sequence $\{x_n\}$ such that $x_{n+1} = rx_n$ for $n = 1, 2, \ldots$ with x_0 as prescribed. The difference Eqn. (3) can be solved iteratively by substituting

$$x_n = r\, x_{n-1},\ x_{n-1} = r\, x_{n-2},\ \ldots,\ x_1 = r\, x_0$$

Thus unique solution is given by

$$x_n = r^n\, x_0,\ n = 1, 2, \ldots$$

In general, we can say that if $|r| < 1$ then $x_n \to 0$ as $n \to \infty$ implying that the population is ultimately driven to extinction and for $|r| > 1$, x_n grows unbounded as $n \to \infty$. In the present situation the case $r < 0$ is ruled out. Therefore, if $0 \le r < 1$ then x_n decreases monotonically to zero and if $r > 1$, then x_n increases to $+\infty$.

If $r > 1$, the growth of population over each time interval of length n occurs by the same rate but not by the same amount as shown in Fig. 2.

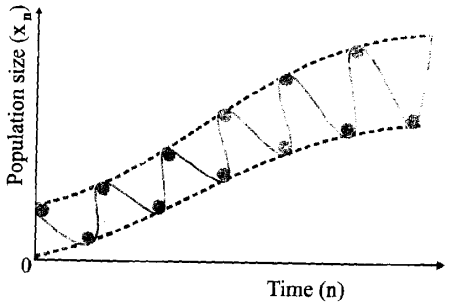


Fig. 2: Discrete Population Model

In case of species which reproduce annually and death occurs throughout the year, the population growth curve resembles a jagged saw blade with a sharp increase, resulting from births followed by gradual decrease from death during the rest of the year. The overall curve will rise exponentially, because the growth rate is positive. The size of each tooth in the growth curve will increase year after year because same proportional increase will add more individuals to a large population. If the time interval between reproductive periods reduces, the occurrence of teeth on the graph gets closer. Finally if the time interval is infinitesimally small the curve is no longer jagged but smooth and resembles the corresponding curve for exponential model of continuous growth.

This simple model of population growth is not very realistic for most populations nor for long times but, even so, it has been used successfully with some justification for the early stages of growth of certain bacteria. Another variation of the exponential growth model can be considered by incorporating a constant migration rate m per generation which we shall consider.

Formulation

Let the positive value of m represents immigration and negative value denotes emigration. Then the difference Eqn. (3) becomes

$$x_{n+1} = r\, x_n + m \qquad (4)$$

Solution and Interpretation

Eqn. (4) can be solved iteratively by considering

$$x_1 = r\, x_0 + m$$

$$x_2 = r\, x_1 + m$$

$$= r(r\, x_0 + m) + m = r^2\, x_0 + r\, m + m$$

$$x_3 = r\, x_2 + m$$

$$= r(r^2\, x_0 + r\, m + m) + m = r^3\, x_0 + r^2\, m + r\, m + m$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\vdots \qquad \vdots \qquad \text{and so on;}$$

Then by induction method we can prove that,

$$x_n = r^n\, x_0 + m(r^{n-1} + r^{n-2} + \cdots + r + 1)$$

$$= r^n\, x_0 + m\left(\frac{1-r^n}{1-r}\right)$$

$$= \left(x_0 - \frac{m}{1-r}\right) r^n + \frac{m}{1-r}$$

Now, if $r > 1$, then x_n grows unbounded for $m > (1-r)\, x_0$ but x_n reaches zero if $m < (1-r)\, x_0$; thus sufficiently large emigration will wipe out a population that would otherwise grow unbounded. If $0 < r < 1$, then as $n \to \infty$, x_n tends to the limit $m/(1-r) > 0$ for $m > 0$, while x_n tends to zero for $m < 0$. Thus, immigration may help survival of a population that would otherwise become extinct. Even this model is not very realistic. We now list the limitations of this model.

Limitations

- The model has constant per capita birth and death rates and generates limitless growth. This is highly unrealistic.
- The model ignores lags. The growth rate does not depend upon the past. Moreover the population responds instantaneously to change in the current population size.
- There is no consideration of temporal variability.

In the next sub-subsection we shall discuss a model which is an improvement over this model but let us first consider some applications of this model.

Example 1 (Fibonacci Sequence): Fibonacci, in the 18^{th} century, set a modelling exercise involving an hypothetical growing rabbit population. Start with a pair (male and female) of immature rabbits which after one reproductive season produce two pairs of male and female immature rabbits after which the parents stop reproducing. Their offspring pairs then do exactly the same and so on. The question is to determine the number of pairs of rabbits at each reproductive period. If we denote the number of pairs of rabbits by x_n at the n^{th} reproductive period then we have the model equation as

$$x_{n+2} = x_{n+1} + x_n,\ n = 1,\ 2, \ldots$$

with $x_0 = 0,\ x_1 = 1$

what is known as the Fibonacci sequence, namely

1, 1, 2, 3, 5, 8, 13, ...

To find the number of pairs of rabbits at the n^{th} reproductive period consider the equation

$$x_{n+2} = x_{n+1} + x_n$$

The equation can be written as

$$(E^2 - E - 1)\, x_n = 0 \text{ where, } E^p x_k = x_{k+p},$$

whose roots are given by

$$m^2 - m - 1 = 0.$$

Solving we obtain

$$m = \left(\frac{1 \pm \sqrt{5}}{2} \right).$$

The solution of the above model can be written as

$$x_n = c_1 \left(\frac{1+\sqrt{5}}{2} \right)^n + c_2 \left(\frac{1-\sqrt{5}}{2} \right)^n.$$

After applying the conditions $x_0 = 0$, $x_1 = 1$ we get

$$x_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n$$

which gives the number of pairs of rabbits at the n^{th} reproductive period.

***

Example 2: Suppose that a population of yeast satisfying exponential growth model increases by 10% in an hour. If the initial population of yeast is 100,000 then find the population of yeast after four hours. How much time is required by the population to grow to double of its initial size?

Solution: The population of yeast satisfies the equation

$$P_{n+1} = (1 + 0.1) P_n \text{ with } P_0 = 100{,}000.$$

The population after one hour is $P_1 = 1.1\ P_0 = 110,000$. After two hours, $P_2 = 1.1\ P_1 = (1.1)^2 P_0 = 121,000$. Thus, after 4 hours, $P_4 = (1.1)^4 P_0 = 146,410$.

For the population to double, it must reach $2P_0 = 200,000$. Thus, we must solve $2P_0 = (1.1)^n P_0$ or $2 = (1.1)^n$.

By taking the logarithms of both sides we have $\ln(2) = \ln(1.1)^n = n \ln(1.1)$ or $n = \ln(2)/\ln(1.1) = 7.27$ hours as the required time.

***

You may **notice** that the discrete Malthusian growth model is closely related to **compound interest problems**. If interest is compounded annually, then the amount of principal in any year n satisfies the discrete Malthusian growth model. The general formula for determining the amount of principal when interest rate is r (annual), which is compounded k times a year for n years, given an initial amount of P_0 satisfies:

$$P_n = (1 + r/k)^{kn} P_0,$$

where P_n is the amount of principal after n years.

In population studies, one can use this concept to examine growth rates for a population growing according to the Malthusian growth model for differing periods of time.

You may now try the following exercise.

E1) Suppose that a business is started with constructing a cow shed for 85 cows and a decision is taken that there will be an addition of 38 cows every year. Now, if mortality rate of cows is 5 percent per year then what is the population size of the cow shed after 15 years.

Before proceeding further, we give you a **graphical method** called **cobwebbing** of solving the models of discrete population growth.

Cobwebbing

Consider the difference Eqn. (1) viz.,

$$x_{n+1} = x_n F(x_n) = f(x_n).$$

Generally, in a population when the population size becomes very large, there is lack of food, space and other resources and pollution due to overcrowding. We expect $f(x_n)$ to have some maximum at x_m say, with f as a function of x_n decreasing for $x_n > x_m$. A typical growth form of f is as shown in Fig. 3.

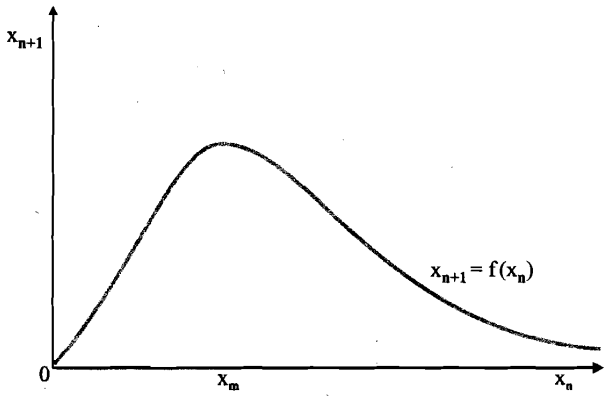


Fig. 3: Typical growth form in the model $x_{n+1} = f(x_n)$.

The steady states of Eqn. (1) are solutions x^* of

$$x^* = f(x^*) = x^* F(x^*)$$

so that $x = x^*$ is a constant solution of Eqn. (1).
Thus, the steady states are given by $x^* = 0$ or $F(x^*) = 1$.

Graphically the steady states are points of intersections of the curve $x_{n+1} = f(x_n)$ and the straight-line $x_{n+1} = x_n$ as shown in Fig. 4.

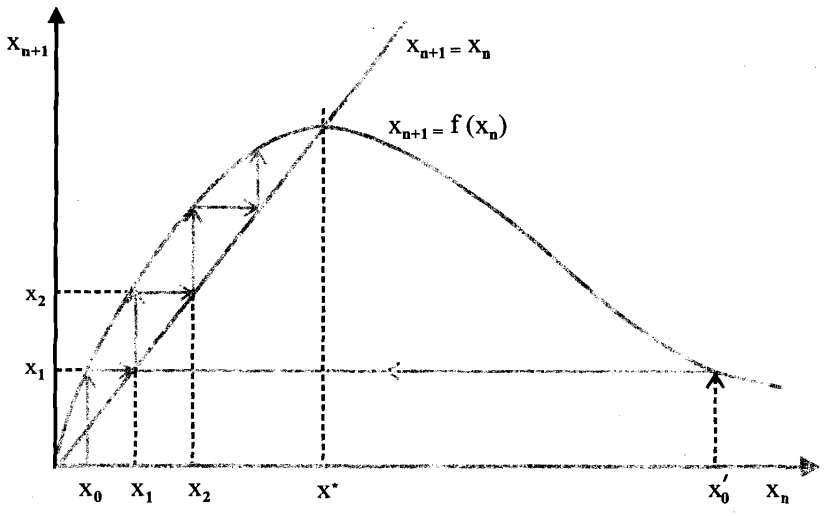


Fig. 4: Graphical determination of the steady-states of $x_{n+1} = f(x_n)$

We begin by drawing the point x_0 as shown in Fig. (4). Then x_1 is given by simply moving vertically to the curve $x_{n+1} = f(x_n)$. Then we go horizontally to the line $x_{n+1} = x_n$. We can then employ x_1 to get x_2 in similar fashion and proceed to arrive at points $x_3, x_4, \ldots$ and so on. The arrows show the path sequence. The path is simply a series of reflections in the line $x_{n+1} = x_n$. We observe that $x_n \to x^*$ monotonically as $n \to \infty$ as illustrated in Fig. (5). This behaviour has already been obtained analytically for the case $f(x_n) = rx_n$.

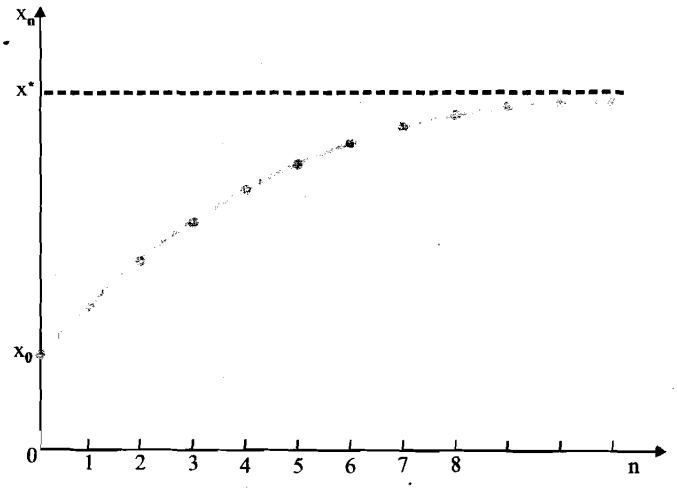


Fig. 5: Continuous curve showing populations at different time-steps for more clarity.

The Cobwebbing method can be applied to any difference equation of the form $x_{n+1} = F(x_n)$. It gives information about the behaviour of the solution and is particularly useful for difference equations whose analytic solutions are complicated.

You may now try the following exercise.

E2) A discrete model for a population N_t consists of

$$N_{t+1} = \frac{r\, N_t}{1 + b\, N_t^2} = f(N_t)$$

where t is the discrete time and r and b are positive parameters. What do r and b represent in this model? Show, with the help of a cobweb, that after a long time the population N_t is bounded by

$$N_{min.} = \frac{2r^2}{(4+r^2)\sqrt{b}} \leq N_t \leq \frac{r}{2\sqrt{b}}.$$

Prove that, for any r, the population will become extinct if $b > 4$.

In the long run there must be some adjustment to such exponential growth. P. F. Verhulst in 1838 proposed that a self-limiting process should operate when a population becomes too large and suggested a model called logistic growth model which we shall discuss now.

Logistic Growth Model

As we have already mentioned the exponential growth model is applicable for early phases of population growth. As the population grows it faces crowding effects due to factors like toxic buildup, self-regulation or space and resource limitations, etc. In order to incorporate such factors, a negative quadratic term is added to the recruitment function reflecting consequent decrease in the growth of the population. This leads to the Logistic Growth Model.

Formulation

Assuming that the population size is large and the growth rate decreases with the increase in population on account of crowding effects, a difference equation model can be formulated as

$$x_{n+1} = r\, x_n \left(1 - \frac{x_n}{K}\right) \tag{5}$$

where $r > 0$ represents **intrinsic growth** rate i.e., growth rate free from environmental constraints and K denotes the **carrying capacity** of the population which physically means the maximum sustainable population size that a particular environment can support over a long period of time. Eqn. (5) is called the **logistic difference equation**.

Let us analyze the model qualitatively. For the background needed for the qualitative analysis of the given model refer to the **appendix** given at the end of this unit.

Stability Analysis

The logistic growth model given by Eqn. (5) can be rescaled by substituting $u_n = \frac{x_n}{K}$, so that the carrying capacity is 1. The Eqn. (5) thus takes the form

$$u_{n+1} = r\,u_n(1-u_n), \quad r>0$$

where we assume that $0<u_0<1$ and we are interested in solutions $u_n \geq 0$. The steady states of Eqn. (6) are given by

$$u^* = f(u^*) = u^* F(u^*) \Rightarrow u^* = 0 \text{ or } F(u^*) = 1$$

Here $F(u^*) = r(1-u^*)$ so $F(u^*) = 1 \Rightarrow u^* = \dfrac{r-1}{r}$

thus

$u_1^* = 0$ and $u_2^* = \dfrac{(r-1)}{r}$ are the two steady states of Eqn. (6).

Further, $u_1^* = 0$ gives $\lambda_1 = f'(0) = r - 2ru^*\big|_{u^*=0} = r$

and $u_2^* = \dfrac{r-1}{r}$ gives $\lambda_2 = f'\left(\dfrac{r-1}{r}\right) = 2-r$.

Thus, the eigen values corresponding to u_1^* and u_2^* are given by $\lambda_1 = r$ and $\lambda_2 = 2-r$ respectively. If $0<r<1$, the steady state u_1^* is stable since $0<\lambda_1<1$. As r increases and crosses the value 1, the steady state u_1^* becomes unstable while the positive steady state u_2^* becomes stable as long as $-1<\lambda_2<1$. Hence the first bifurcation occurs at $r=1$. The second bifurcation occurs at $r=3$, where the positive steady state u_2^* undergoes a qualitative change. If r lies between 2 and 3 the equilibrium u_2^* is stable and as soon as r exceeds 3, it becomes unstable. Fig. 6 shows a schematic diagram of stable solutions of Eqn. (6) as r passes through bifurcation values. At each bifurcation the preceding steady state becomes unstable and has been represented by the dashed lines.

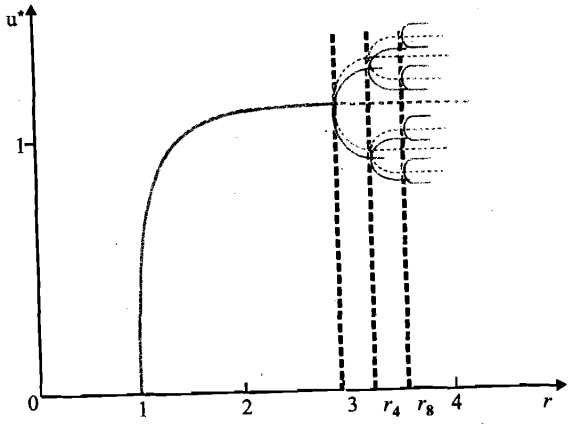


Fig. 6: Schematic diagram for the stability of Eqn. (6).

To see what is happening when r passes through the bifurcation value $r=3$, we introduce the iterative procedure as follows:

$$u_1 = f(u_0)$$
$$u_2 = f(f(u_0)) = f^2(u_0)$$
.....
$$u_n = f^n(u_0)$$

Thus for Eqn. (6) the first iterative is simply Eqn. (6) while the second iterative is

$$u_{n+2} = f^2(u_n, r) = f(f(u_n, r))$$
$$= r[r\,u_n(1-u_n)]\,[1 - r\,u_n(1-u_n)]$$

It can be easily shown that there are two more equilibria u_3^* and u_4^* for the second iterative $u_{n+2} = f^2(u_n, r)$ when r exceeds 3 and that both these steady states are stable. This means that there is a stable equilibrium of the second iterative which results into a stable periodic solution of period 2 of Eqn. (6).

As r continues to increase, the eigen values at u_3^* and u_4^* again undergo a qualitative change and so these 2-period solutions become unstable. At this stage we consider the steady states of fourth iterative of Eqn. (6) and a 4-cycle periodic solution is obtained. Fig. 7 shows a 4-cycle periodic solution schematically.

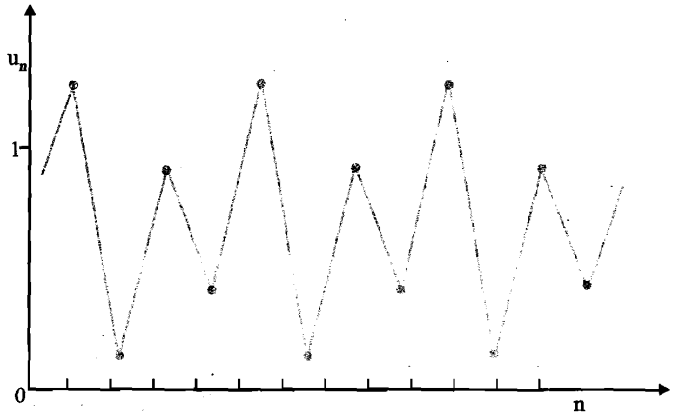


Fig. 7: A 4-cycle periodic solution

Thus as r increases the solution passes through a series of bifurcation, doubling the period of periodic solution each time. There is a limiting value r_c at which instability sets in for all periodic solutions and a chaotic or aperiodic solution occurs. This critical value in the present model as given by $r_c \approx 3.828$ [ref. 'Mathematical Biology' by J. D. Murray]. The solutions in this case oscillate randomly. The fluctuations in a chaotic solution do not arise by chance factor or randomness. Once the parameters are specified, the same erratic population track will be obtained. The main property of chaotic solutions is that a very small change in initial conditions can lead to very different population behaviour. There is widespread interest among scientists regarding chaotic behaviour. The research in this direction has resulted in many different applications of chaos.

The discrete logistic model is very significant from biological as well as mathematical point of view, because of the fact that such a simple model is capable of predicting an apparently unpredictable behaviour. However, there are limitations of this model too which we list below.

Limitations

Sigmoids are tilted S-shaped curves that resemble trends in the life-cycle of many living organisms and phenomenon.

- The logistic model is not suitable for a population of small size.
- Although many populations exhibit the sigmoid pattern of growth still they do not increase according to the logistic equation eventually.
- The carrying capacity of the population has been taken as constant whereas in view of the changing environmental factors, it keeps varying.
- The population has been considered as a homogeneous group of individuals where death and birth take place simultaneously. But in real populations, there may be some delay on account of development time, breeding seasons and other environmental factors. Further, the intensity of these processes is different for different age groups.

Let us now consider some applications of this model.

Example 3: Consider the model given by Verhulst

$$x_{n+1} = \frac{rx_n}{x_n + A},\ r > 0,\ A > 0.$$

In this case we have

$$f(x_n) = \frac{rx_n}{x_n + A} \text{ and } F(x_n) = \frac{r}{x_n + A}.$$

Hence the equilibrium points are given by $x_n = 0$ and $x_n = r - A$. Thus, if $r < A$ the only equilibrium corresponding to a non negative population size is $x_n = 0$. Since $f'(0) = r/A < 1$, at this point system is **asymptotically stable** and every solution tends to zero (see Appendix). If $r > A$ there are two steady states at $x_n = 0$ and $x_n = r - A$. Since, $f'(0) = r/A > 1$, the equilibrium at $x_n = 0$ is **unstable** and since $f'(r - A) = \frac{A}{r} < 1$ the equilibrium $x_n = r - A$ is **asymptotically stable.**

***

Example 4: Consider an example of the Ricker model given by the equation

$$N_{t+1} = N_t \exp\left[r\left(1 - \frac{N_t}{K}\right)\right]. \tag{7}$$

The dynamics of this model is similar to the continuous time logistic model if population growth rate is small $(0 < r < 0.5)$. However, if the population growth rate is high, then the model may exhibit more complex dynamics including damping oscillations, cycles or chaos. It can be shown that Ricker's model is stable if $0 < r < 2$. When r assumes the value 2 the first bifurcation occurs leading to a 2-cycle periodic solution. If r further increases these solution become unstable and at $r = 2.6$, a 4-cycle periodic solution appears. Thus with increasing values of r the solution passes through a series of bifurcations, each time doubling the period. When r reaches near the critical

value 3 then an aperiodic or chaotic solution exists. There are simulations of population dynamics using Ricker's method with different values of r, as shown in Fig. 8.

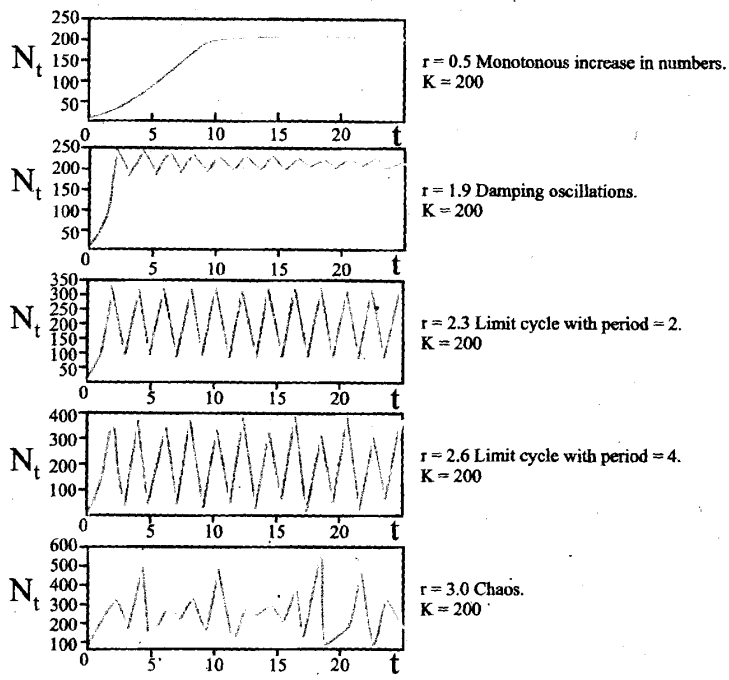


Fig. 8: Solutions N_t of the Model (7) for various r.

In the upper two figures the model has a stable equilibrium, only the patterns of approaching the equilibrium are different. In the lower three figures there is no stable equilibrium. Non-equilibrium dynamics may be of 2 types: a limit cycle when the trajectory repeats itself, and chaotic when the trajectory does not repeat itself.

Chaos dynamics looks like stochastic noise, however the model is absolutely deterministic. Chaotic models are widely used for random number generation in computers.

***

You may now try the following exercises.

E3) Find the two equilibria u_3^* and u_4^* of the second iterative given by Eqn. (8) and verify that they are stable.

E4) Consider a population of peacock modeled by the equation $P_{i+1} = P_i + r^*h^*P_i^*(1 - P_i/K) - h^*H$, i = 0, 1, 2, 3, ... where H is the harvesting rate.

Let $P_0 = 100$, $t_0 = 0$, $r = 0.0987$, $h = 1.0$ year and $K = 194.6$; all populations are expressed in thousands.

The removal of members of a population at a specified rate is termed as **harvesting**.

a) Determine and graph the populations of peacocks for a period of 50 years using the harvesting rates $H = 0, 2, 4,$ and 6. For each value of H describe the trend of the peacock population. Does it appear to approach a stable state? If so, what is that value?

b) What is the first value of H (to the nearest tenth) that does not produce a steady state population?

In general the harvesting rate depends on time and is represented by the function H(t). If H(t) is constant then harvesting is called **constant yield harvesting**. If H(t) is a linear function of population size then it is called **proportional** or **constant effort harvesting**.

E5) Determine the non negative steady state and discuss their linear stability of the following discrete time population models:

a) $N_{t=1} = N_t\left[1 + r\left(1 - \frac{N_t}{K}\right)\right]$.

b) $N_{t=1} = \frac{r\,N_t}{(1 + a\,N_t)^b}$.

In earlier models of discrete population growth, a common assumption was that all the members of a generation contribute to the growth. This is not true in general because for most of species the individuals reproduce only after attaining a certain degree of maturity in age. We shall now discuss a model that incorporates the effect of this delay.

Delay Model

For most of the animals and other species there is a substantial maturation time to sexual maturity. Such a delay can also be caused by delayed response to environmental changes. If the delay is observed to be m time steps then the basic difference equation of the model takes the form

$$x_{n+1} = f(x_n, x_{n-m}) \tag{8}$$

Now we study such a model.

Formulation

We assume that the growth rate of a population reduces with the increase in population due to crowding effects. Further, unlike the logistic model, the population x_{n+1} at $(n+1)$th generation remains positive. Thus the model can be formulated as

$$x_{n+1} = x_n \exp r\left(1 - \frac{x_{n-m}}{K}\right) \tag{9}$$

where $r > 0$ is the intrinsic growth rate and $K > 0$ is the carrying capacity. Rescaling the model by substituting $u_n = x_n / K$, it can be written as

$$u_{n+1} = u_n \exp\,[r(1 - u_{n-m})],\ r > 0 \tag{10}$$

As a simple case we consider the delay to be of one time step. Then Eqn. (10) can be written as

$$u_{n+1} = u_n \exp\,[r\,(1 - u_{n-1})] \tag{11}$$

The steady states of the difference Eqn. (11) are given by

$u_1^* = 0$ and $u_2^* = 1$.

We now discuss the stability analysis of the steady states.

Stability Analysis

Let us discuss the stability of the equilibrium states $u_1^* = 0$ and $u_2^* = 1$.

Since $|f'(u_1^*)| = e^r > 0$.

The steady state $u_1^* = 0$ turns out to be **unstable**.

Next we linearize the equation about the steady state $u_2^* = 1$ by substituting $u_n = 1 + \varepsilon_n$, where $|\varepsilon_n| << 1$.

We obtain

$$1 + \varepsilon_{n+1} = (1 + \varepsilon_n) \exp(-r\varepsilon_{n-1})$$

Omitting higher powers of ε_n because $|\varepsilon_n| << 1$.

We have

$$1 + \varepsilon_{n+1} = (1 + \varepsilon_n)(1 - r\varepsilon_n)$$

or, $1 + \varepsilon_{n+1} = 1 + \varepsilon_n - r\,\varepsilon_{n-1}$

Hence the difference equation to determine the steady state is given by

$$\varepsilon_{n+1} - \varepsilon_n + r\,\varepsilon_{n-1} = 0 \tag{12}$$

Substituting $\varepsilon_n = z^n$ Eqn. (12) takes the form

$$z^2 - z + r = 0 \tag{13}$$

If $r < 1/4$ then the roots of Eqn. (13) are real and are given by

$$z_1 = 1/2\,[1 + (1 - 4r)^{1/2}]$$

and $z_2 = 1/2\,[1 - (1 - 4r)^{1/2}]$

if $r > 1/4$ then the roots of Eqn. (13) are imaginary and are given by

$$z_1 = 1/2\,[1 + i\,(4r - 1)^{1/2}]$$

$$z_2 = 1/2\,[1 - i\,(4r - 1)^{1/2}]$$

The complete solution of Eqn. (13) is given by

$$\varepsilon_n = c_1 z_1^n + c_2 z_2^n \tag{14}$$

where c_1 and c_2 are arbitrary constants.

If $0 < r < 1/4$ then both z_1 and z_2 lies between 0 and 1 and hence $\varepsilon_n \to 0$ as $n \to \infty$. Therefore the state $u^* = 1$ is linearly **stable equilibrium** state.

If $r > \frac{1}{4}$ then the two complex roots of quadratic Eqn. (13) can be written as

$$z_1 = \rho e^{i\theta} \quad z_2 = \rho e^{-i\theta} \text{ with } z_2 = \bar{z}_1$$

where $\rho = r^{1/2}$ and $\theta = \tan^{-1}(4r - 1)^{1/2}$

and $z_1 z_2 = |z_1|^2 = \rho^2 = r$.

Thus for $\frac{1}{4} < r < 1$, $|z_1|\,|z_2| < 1$.

In this case solution (14) is

$$\varepsilon_n = c_1 z_1^n + c_2 \bar{z}_1^n \qquad (15)$$

and since it is real we must have $c_2 = \bar{c}_1$.

If we now let $c_1 = \alpha e^{i\gamma}$ then $c_2 = \alpha e^{-i\gamma}$ then for complex roots Eqn. (15) can be written as

$$\varepsilon_n = \alpha\,[\rho^n e^{i(n\theta+\gamma)} + \rho^n e^{-i(n\theta+\gamma)}]$$

or, $\varepsilon_n = 2\alpha\rho^n \cos(n\theta+\gamma)$ where $\gamma = \tan^{-1}\dfrac{c_1}{c_2}$ and $\theta = \tan^{-1}(4r-1)^{1/2}$.

Hence $\varepsilon_n \to 0$ as $n \to \infty$ and the state $u^* = 1$ is **stable.**

As $r \to 1$, $\theta \to \tan^{-1}\sqrt{3} = \dfrac{\pi}{3}$.

As r passes through the critical $r_c = 1$ i.e. $r > 1$ then $|z_1| > 1$ and ε_n given by Eqn. (15) grows unboundedly with $n \to \infty$ and $u^* = 1$ is then unstable. Since $\theta \approx \pi/3$ for $r \approx 1$ and $\varepsilon_n \approx 2\alpha\cos(n\pi/3+\gamma)$, which has a period of 6 we expect the solution of Eqn. (11), at least for r just greater than $r_c(=1)$ to exhibit 6 cycle periodic solution.

Limitations

- In the model discussed above, the discrete delay due to development time of the species has been incorporated. But this model does not accurately model populations with continuous growth and time lags.
- Sometimes the delay can be caused by the factors limiting the species. This is generally the case when the limiting factor itself is a species subject to delay due to prolonged development.

Let us consider the following example.

Example 5: An animal may feed on the number of plants which is annual, so that the total food available depends on the number of the plants in the previous year and thus in turn is a function of the number of animals present in the previous years. The logistic equation can then be written as a time delay equation of the form.

$$\frac{n(T_{m+1}) - n(T_m)}{T} = n(T_m)\left(\frac{1 - n(T_{m-1})}{N}\right)$$

with a delay of one period, equal to T units of time.

We can easily find out that the equilibrium value of $n(T_m)$ in terms of N. To find out how $n(T_m)$ is going to behave near the value N we will use the expression given below in the place of above equation for small y,

$$y(T_{m+1}) - y(T_m) \approx y(T_{m-1})\,T/\tau \qquad (16)$$

where we have neglected the quadratic term in y.

The general solution of an equation of the form (16), which is called a recurrence relation, is

$$y(T_m) = Az_1^m + Bz_2^m \qquad (17)$$

where A and B are constants which can be determined by using the initial conditions and z_1 and z_2 are the roots of the quadratic equation

$$z^2 - z + \frac{T}{\tau} = 0. \qquad (18)$$

Solving Eqn. (18) we obtain

$$z_1 = \frac{1}{2}\left(1 + \sqrt{(1 - 4T/\tau)}\right) \text{ and } z_2 = \frac{1}{2}\left(1 - \sqrt{1 - 4T/\tau}\right)$$

z_1, z_2 are both real if $\frac{T}{\tau} \leq \frac{1}{4}$ and then the usual logistic behaviour occurs.

When $\frac{T}{\tau} > \frac{1}{4}$, the two roots z_1, z_2 are complex conjugate and can be written in the form $re^{\pm i\theta}$ where, $r = \sqrt{\frac{T}{\tau}}$, $\cos\theta = \sqrt{\frac{\tau}{4T}}$. $\left(0 \leq \theta \leq \frac{\pi}{2}\right)$.

The solution of Eqn. (18) can then be written as

$$y(T_m) = C\, r^m \sin(\theta m + D)$$

where C and D are constants determined by the initial conditions in the same way as A and B. In these conditions $n(T_m)$ oscillate about its equilibrium value with a period $2\pi T/\theta$. When $T/\tau > 1$ the oscillation get larger with m, so that Eqn. (17) is not valid in this case. When $T/\tau < 1$ the oscillation gets smaller, so that $n(T_m)$ converges to its equilibrium value N. This shows that a time delay makes the population less stable.

***

You may now try the following exercise.

E6) The population n(m) of a certain species after the m^{th} breeding season is related to $n(m-1)$ by

$$n(m) = N\,\alpha\, n(m-1)/(N + n(m-1))$$

where α is a number greater than one, and N is a constant. Find the equilibrium value of n and show that n approaches this value monotonically.

In the next section we shall consider various models of continuous population growth.

CONTINUOUS POPULATION MODELS

As mentioned earlier, if the population size is very large and births take place continuously we may assume the population growth to be continuous.

The continuous models of exponential growth or Malthusian model and logistic model which are analogues of discrete models given by Eqn. (3) and Eqn. (5) respectively, have been dealt exhaustively in Block 3, MTE-14 of mathematical modelling course of our Bachelors Degree programme. We shall

assume that you are familiar with these models. In case you are not, then you can go through Unit 8, Block 3 of MTE-14.

You may **observe** that both the logistic and exponential models of continuous growth, share the assumption of absence of no migration, genetic variation or age structure in a given population. In addition logistic model assumes constant carrying capacity, limited resources and density dependent growth. This means that when an individual is added to the population the per capita growth rate decreases immediately. But this is not the case in general. In many population the density dependent response assumes some time lags. Individuals do not immediately adjust into their growth and reproduction. Seasonal availability of resources and age structure can also cause time lag in population growth. These delays can affect population dynamics of a given species significantly. We shall now study the effect of such delays.

Delay Model

Consider the logistic growth model which is given by the following equation

$$\frac{du}{dt} = ru(t)\left(1 - \frac{u(t)}{K}\right) \tag{19}$$

where $u(t)$ is the population at time t and the positive constants r and K denote the intrinsic growth rate and the carrying capacity of the population respectively. In this model we now incorporate the effect of time delay in growth and reproduction of individuals in the population and formulate the delay model.

Formulation

Let us assume that there is a time lag of length $T > 0$ between the change in population size and its effect on population growth rate. Consequently, the growth rate $\frac{du}{dt}$ of the population at time t, depends on the size of population at time $(t - T)$. Thus, in the logistic growth model incorporating the effect of delay, Eqn. (19) takes the form

$$\frac{du}{dt} = ru(t)\left(1 - \frac{u(t-T)}{K}\right) \tag{20}$$

where r, K and T are positive constants. Eqn. (20) is the **differential delay equation**. The behaviour of Eqn. (20) depends on the length of time lag T as well as the response time which is inversely proportional to the growth rate r [Gotelli, 1995].

The Analytical solution of Eqn. (20) cannot be obtained in general so we have to analyze it qualitatively.

Stability Analysis: Periodic Solutions

The steady states or the equilibrium points of Eqn. (20) are given by

$$u^* = 0 \text{ and } u^* = K$$

For computational convenience we rescale the model by substituting

$N(t) = \frac{u(t)}{K}$, $t^* = rt$, $T^* = rt$. Thus the model (20) becomes

$$\frac{dN}{dt^*} = N(t^*)\ [1 - N(t^* - T^*)]$$

For the sake of simplicity we drop the asterisks from the above model and write the delay model in the form

$$\frac{dN}{dt} = N(t)\ [1 - N(t - T)] \tag{21}$$

The equilibrium points of Eqn. (21) are $N_1 = 0$ and $N_2 = 1$.

Let us first consider the linearization about equilibrium point $N_1 = 0$.

We assume

$$N = N_1 + n(t), \text{ where } |n(t)| << 1. \tag{22}$$

Using Eqn. (22), Eqn. (21) reduces to

$$\frac{dn(t)}{dt} = (N_1 + n(t))\ [1 - (N_1 + n(t - T))]$$

or, $$\frac{dn(t)}{dt} = n(t)\,[1 - n(t - T)] \qquad [\because N_1 = 0]$$

Omitting higher powers of $n(t)$ in the above equation we get

$$\frac{dn(t)}{dt} = n(t) \tag{23}$$

which on integration yields

$$n(t) = A\, e^t$$

where A is a constant. This implies that $n(t) \to \infty$ as $t \to \infty$.

Hence, the steady state $N_1 = 0$ is **unstable** with **exponential growth**.

Next we consider, perturbation about the steady state $N_2 = 1$.

Using Eqn. (22), Eqn. (21) in this case reduces to

$$\frac{dn(t)}{dt} = (1 + n(t))\ [1 - (1 + n(t - T))] \qquad [N_2 = 1]$$

or, $$\frac{dn(t)}{dt} = (1 + n(t))\,[-n(t - T)]$$

Omitting the higher powers of $n(t)$, we obtain

$$\frac{dn(t)}{dt} = -n\ (t - T) \tag{24}$$

We now look for solutions of Eqn. (24) in the form

$$n(t) = c\, e^{\lambda t} \tag{25}$$

where c is a constant.

Substituting the value of $n(t)$ from Eqn. (25) into Eqn. (24) we obtain

$$c\lambda\, e^{\lambda t} = -c\, e^{\lambda(t-T)}$$

or, $\lambda = -e^{-\lambda T}$ (26)

You can thus see that the eigen values λ of n(t) are the solution of Eqn. (26) which is a transcendental equation and it is not easy to solve it analytically.

However, from a stability point of view we are interested to know whether there are any solutions with $\mathrm{Re}\,\lambda > 0$ for which Eqn. (25) implies instability since n(t) grows exponentially with time. That is, if we set $\lambda = \mu + i\omega$ then does there exists a real number μ_0 such that all solution λ of Eqn. (26) satisfy $\mathrm{Re}\,\lambda < \mu_0$. To see this, we have from Eqn. (26)

$$|\lambda| = |-e^{-\lambda T}| = |e^{-(\mu+i\omega)T}| = |e^{-\mu T}||e^{-i\omega T}| = e^{-\mu T}$$

Thus, if $|\lambda| \to \infty$ then $e^{-\mu T} \to \infty$ provided $\mu \to -\infty$.
Thus there must exist a real number μ_0 so that $\mathrm{Re}\,\lambda$ i.e., μ is bounded above by μ_0.

If we now introduce $z = 1/\lambda$ and $\omega(z) = 1 + ze^{-T/z}$ then $z = 0$ is an essential singularity of $\omega(z)$. By Picard's theorem $\omega(z) = 0$ will then have infinitely many complex roots in the neighborhood of $z = 0$. This means there are infinitely many roots λ (ref. Sec. 65, Page-232, of Brown and Churchill).

Let us now consider the real and imaginary parts of Eqn. (26), namely

$$\mu = -e^{-\mu T}\cos\omega T \quad (27)$$

$$\omega = e^{-\mu T}\sin\omega T \quad (28)$$

We want to determine the range of T such that $\mu < 0$. That is, we want to find the conditions such that the upper limit μ_0 on μ is negative. Here we consider two cases viz., $\omega = 0$ and $\omega \neq 0$.

Case I: $\omega = 0$

When $\omega = 0$, λ becomes real. You can see that for $\omega = 0$, Eqn. (28) is satisfied and Eqn. (27) becomes

$$\mu = -e^{-\mu T} \quad (29)$$

which has no positive roots $\mu > 0$. Since $e^{-\mu T} > 0$ for all μT.

Case II: $\omega \neq 0$.

If ω is a solution of Eqns. (27) and (28) then $-\omega$ also satisfies these equations. Thus, without any loss of generality, we can consider $\omega > 0$. From Eqn. (27) you may observe that

$$\mu < 0 \text{ requires } \omega T < \pi/2 \text{ since } -e^{-\mu T} < 0 \;\forall \mu T.$$

Since Eqns. (27) and (26) defines $\mu(T)$ and $\omega(T)$, we are interested in finding that value of T when $\mu(T)$ first crosses from $\mu < 0$ to $\mu > 0$.

As T increases from zero then μ first becomes zero only when $\omega T = \pi/2$. Now we see, that for $\mu = 0$ Eqn. (28) has the only relevant solution $\omega = 1$ occurring at $T = \pi/2$. Since this is the first zero of μ as T increases this gives the bifurcating value $T = T_c = \pi/2$.

Alternately, we can use the argument that

$e^{-\mu T} \sin \omega T = \omega$

$\Rightarrow$ $T\, e^{-\mu T} \sin \omega T = \omega T < \pi/2$

$\Rightarrow$ $T\, e^{-\mu T} \sin \omega T < \pi/2$

$\Rightarrow$ $0 < T < \pi/2$ (Since $\sin \omega T < 1$ and $e^{-\mu T} > 1$ if $\mu < 0$).

which gives the condition on T for the stability of $N_2 = 1$.

Returning to our Eqn. (20) we can thus say that the steady state

$u^*(t) = K$ is **stable** if $0 < rT < \pi/2$ and **unstable** if $rT > \pi/2$.

In the latter case we expect the solution to exhibit stable limit cycle behaviour. The critical value $rT = \pi/2$ is the bifurcation value, that is the value of the parameter, rT here, where the character of the solution of Eqn. (20) changes abruptly or bifurcates from stable steady state to a time varying solution. The effect of delay in models is usually to increase potential for instability. As T is increased beyond the bifurcation value $T_c = \pi/2r$, the steady state becomes unstable.

Let us now obtain the first estimate of the period of the bifurcating oscillatory solution. Consider Eqn. (21) and let

$$T = T_c + \varepsilon = \frac{\pi}{2} + \varepsilon, \quad 0 < \varepsilon << 1 \tag{30}$$

For $T = \frac{\pi}{2}$, $\mathrm{Re}\,\lambda$ is largest and the solution $\lambda = \mu + i\omega$ of Eqns. (27) and (28) is $\mu = 0$, $\omega = 1$. We would then expect that for ε small, μ and ω also differ from $\mu = 0$ and $\omega = 1$ by small quantities. Accordingly, we let

$$\mu = \alpha, \ \omega = 1 + \beta, \quad 0 < \alpha << 1, \ |\beta| << 1 \tag{31}$$

where α and β are to be determined. Substituting these values of μ and ω in Eqn. (28) and expanding for small α, β and ω, we get

$$1 + \beta = \exp\left[-\alpha\left(\frac{\pi}{2} + \varepsilon\right)\right] \sin\left[(1+\beta)\left(\frac{\pi}{2} + \varepsilon\right)\right] \Rightarrow \beta \approx \frac{-\pi\alpha}{2} \tag{32}$$

to the first order of α, β and ε. Similarly Eqn. (27) gives

$$\alpha = -\exp\left[-\alpha\left(\frac{\pi}{2} + \varepsilon\right)\right] \cos\left[(1+\beta)\left(\frac{\pi}{2} + \varepsilon\right)\right] \Rightarrow \alpha \approx \varepsilon + \frac{\pi\beta}{2}. \tag{33}$$

Thus on solving Eqns. (32) and (33) we obtain

$$\alpha \approx \frac{\varepsilon}{1 + \frac{\pi^2}{4}}, \quad \beta \approx \frac{-\varepsilon\pi}{2\left(1 + \frac{\pi^2}{4}\right)} \tag{34}$$

and hence, near the bifurcation, using Eqn. (25), $N(t) = 1 + n(t)$ reduces to

$$N(t) = 1 + \mathrm{Re}\{c \exp[\alpha t + i(1+\beta)t]\}$$

$$\approx 1 + \mathrm{Re}\left\{c \exp\left[\frac{\varepsilon t}{1+\pi^2/4}\right] \exp\left[it\left\{1 - \frac{\varepsilon\pi}{2(1+\pi^2/4)}\right\}\right]\right\}$$

This shows that the instability is by growing oscillations with period

$$\frac{2\pi}{1 - \dfrac{\varepsilon\pi}{2(1+\pi^2/4)}} \approx 2\pi$$

for small ε. Now since $rT = \pi/2$, the period of oscillation is then $4T$.

The model discussed above is also not perfect as it has its limitations which we are stating below.

Limitations

- This model incorporates the delay due to maturation period but ignores the age structure. Further the sex ratio also plays an important role in determining the birth rate which is not considered. The sex ratio may be different in different species.
- In addition to the term corresponding to the crowding effect the absolute death rate should be considered. The age structure also affects the death rate.

There are models with age distribution which incorporates the effect of age structure of the population for both discrete and continuous growth but we shall not be discussing these models here.

Let us consider the following application of the delay model.

Example 6: Consider a laboratory population of the blowfly Lucilia Cupriva kept in a cage and given a limited supply of food. Along with the adult blowflies the cage contains larvae also, which are supplied with unlimited food. Let $x(t)$ be the number of adult blowflies at time t, c be the constant mortality rate per unit time and τ be the time taken by the egg to develop into an adult. Assuming that the laying of eggs be proportional to the initial population. The growth equation of blowflies can be written as

$$\frac{dx}{dt} = -cx(t) + \frac{1}{2}mksx(t-\tau) \qquad (35)$$

where k is the constant of proportionality and s is the probability that an egg will grow into an adult and m is the mass of the blowflies.

Putting $-c = a$ and $\frac{msk}{2} = b$ Eqn. (35) can be reduced to the form

$$\frac{dx}{dt} = ax(t) = bx(t-\tau) \qquad (36)$$

where, $a < 0$ and $b > 0$ are real constants.

Model (36) is known as the Nicolson's model and its solution is to be obtained under the initial conditions $x(t) = 0$ for $t < 0$ and $x(0) = \tau$.

Taking the Laplace transform of both sides of Eqn. (36) and using the given initial conditions we obtain

$$sX(s) - \tau = aX(s) + b\, e^{-s}\, X(s).$$

Hence,

$$x(t) = \frac{\tau}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{e^{st}}{s - a - be^{-s}}\, ds.$$

The only singularity of the integrand are at the zeros of $s - a - be^{-s}$ and the solution is the sum of the residues of the integrand at these poles. To calculate the zeros, we put $s = \alpha + i\beta$ where α and β are real. This gives

$$\alpha + i\beta = a + be^{-(\alpha+i\beta)} \tag{37}$$

Equating real and imaginary parts of Eqn. (37), we obtain the simultaneous equations

$$\alpha = a + be^{-\alpha}\cos\beta \tag{38}$$

$$\beta = -be^{-\alpha}\sin\beta \tag{39}$$

Clearly, one solution of Eqn. (39) is $\beta = 0$, and then Eqn. (38) becomes

$$\alpha = a + be^{-\alpha}. \tag{40}$$

When $b \geq 0$, Eqn. (40) has just one real root which is positive if $a + b > 0$ and negative otherwise. If α_0 is a real root of Eqn. (40) then the corresponding contribution to $x(t)$ from the residue has a form $\exp(\alpha^0 t)$. Hence the condition for exponential divergence are $a + b > 0$ or $a > \tau$ and $b > -e^{a-1}$.

If $\beta \neq 0$ then the stability boundaries for the oscillatory solutions are given parametrically by

$$a = \beta\cot\beta,\ b = -\frac{\beta}{\sin\beta},\ \beta > 0.$$

***

You may now try the following exercises.

E7) Consider the budworm population dynamics to be governed by the equation

$$\frac{dN}{dt} = r_B N\left(1 - \frac{N}{K_B}\right) - p(N) \tag{41}$$

where r_B is the linear birth rate of the budworm and K_B is the carrying capacity. The $p(N)$ term is predation generally by birds. Find out the steady states and do the stability analysis of Eqn. (41).

E8) An animal may feed on the number of plant which is annual, so that the total food available depends on the number of the plants in the previous year and thus in term is a function of the number of animals present in the previous years. Thus the logistic equation can be written as a time delay equation of the form

$$\frac{dn(t)}{dt} = \frac{n(t)}{\tau}\left(1 - \frac{n(t-T)}{N}\right).$$

Discuss the steady state and carry out the perturbation analysis.

E9) A continuous time model for the baleen whale is the delay equation

$$\frac{dN}{dt} = -\mu N(t) + \mu N(t-T)\,[1 + q\{1 - [N(t-T)/K]^z\}]:$$

Here $\mu > 0$ is a measure of mortality and $q > 0$ is the maximum increase in fecundity the population is capable of, K is the unharvested carrying capacity, T is the time of sexual maturity and $z > 0$ is the measure of density of the population. Discuss the steady states and carry out the perturbation analysis.

We now end this unit by giving a summary of what we have covered in it.

SUMMARY

In this unit, we have covered the following:

1. Mathematical ecology can be described as the study of inter-dependence of several species in a variable eco-system employing mathematical modeling.
2. A mathematical model in population dynamics represents quantitatively the pattern of growth of a given population in the presence of various environmental factors.
3. The mathematical models in population dynamics can be classified on account of various factors like nature of growth rate, its fluctuation and environmental factors, size of the population and so on.
4. On the basis of growth rate, the population models can be classified as
   i) Deterministic models
   ii) Stochastic models
5. The models can be analyzed in two ways i.e., quantitatively and qualitatively. The quantitative analysis of models involves consideration of actual parameter and prediction in terms of numbers. On the other hand qualitative analysis concentrates on study of the nature and pattern of growth and their dependency on parameters relevant to the model.
6. The discrete deterministic models are applicable to the cases when moderately large populations such as humans, large animals like lions, elephant are modelled. In such models the variations in population are studied by means of difference equations.
7. Continuous deterministic models are employed when populations are very large and their growth rate is also fast so that variables can be modelled by continuous functions and their growth rate by their derivatives. These models lead to differential equations.

SOLUTIONS/ANSWERS

E1) **Hint:** Proceed as in Example 2. Apply formula for compound interest.

E2) For the model

$$N_{t+1} = \frac{r\,N_t}{1+bN_t^2} = f(N_t).$$

Let us take $N_{t+1} = N_t = N^*$ then for stability we have

$$N^* = \frac{r\,N^*}{1+bN^{*2}} = f(N^*)$$

The steady states are

$$N_1^* = 0,\ N_2^* = \sqrt{\frac{r-1}{b}}$$

$$f'(N^*) = \frac{(1+bN^{*2})\,r - 2\,rbN^{*2}}{(1+bN^{*2})}.$$

$$f'(N_1^*) = r \text{ and } f'(N_2^*) = \frac{2-r}{r} = \frac{2}{r} - 1.$$

For $r < 1$, N_1^* is stable.

For $r > 1$, N_1^* becomes unstable and N_2^* is stable.

E3) With little manipulation Eqn. (7) can be written as

$$u^*\,[ru^* - (r-1)]\ [r^2u^{*2} - r(r+1)u^* + (r+1)] = 0$$

which has solutions

$$u^* = 0 \text{ or } u^* = \frac{r-1}{r} > 0 \text{ if } r > 1$$

$$u^* = \frac{(r+1) \pm [(r+1)\,(r-3)]^{1/2}}{2r} > 0 \text{ if } r > 3$$

$$\therefore\ u_3^* = \frac{(r+1) + [(r+1)\,(r-3)]^{1/2}}{2r} \text{ and } u_4^* = \frac{(r+1) - [(r+1)\,(r-3)]^{1/2}}{2r}.$$

$$[f^2(u, r)]' = r^2(1-2u)\,(1-2ru+2ru^2)$$

$$\therefore\ [f^2(u, r)]'\big|_{u=u_3^*} = -(r^2+2r+2) < 1 \text{ for } r > 3$$

$\Rightarrow$ u_3^* is a stable steady state.

Similarly show that u_4^* is stable steady state.

E4) Apply the same method as in Example 3 and draw the results as in Rickers model in Example 4.

E5) a) The steady states of the given model are
$N_1 = 0$, $N_2 = K$ and the corresponding eigenvalues are
$\lambda_1 = 1+r$, $\lambda_2 = 1-r$.
For $0 < r < 1$, the only equilibrium corresponding to non-negative population sizes is $N_1 = 0$, which is unstable.
For $r = 1$, $\lambda_1 = \lambda_2$.

For $r > 1$, N_1 is unstable and N_2 is stable.

b) Proceed as in a) above.

E6) Proceed as in Example 5 and get the solution.

E7) In the given model put right hand side equal to zero to find out equilibrium state i.e.

$$r_B N\left(1-\frac{N}{K_B}\right)-p(N)=0.$$

Then to do the stability analysis use perturbation i.e.

$N = N^* + n_1$ and $\frac{dN}{dt} = \frac{dn_1}{dt}$ then

$$\frac{dn_1}{dt} = r_B\,(N^*+n_1)\left(1-\frac{N^*+n_1}{K_B}\right)-p(N^*+n_1)$$

expanding the above expression and applying Taylor's series we get

$$= r_B\ N^* - \frac{r_B N^{*2}}{K_B} - \frac{r_B N^* n_1}{K_B} + r_B n_1 - \left\{p(N^*)+\frac{n_1}{1!}p'(N^*)+\frac{n_1^2}{2!}p''(N^*)+\cdots\right\}$$

Neglecting the higher order terms and rearranging we get

$$\frac{dn_1}{dt} = \left[1\frac{r_B N^*}{K_B} + r_B - p(N^*)\right] n_1.$$

After integrating we get

$$n_1 = e^{-At+c} \text{ where } A = -\frac{r_B N^*}{K_B} + r_B - p(N^*)$$

if $t \to \infty$ then $n_1 \to$ zero.

So we can draw the conclusion that when time tends to infinity at that time species n_1 tends to zero i.e., it will go to extinct.

E8) This problem can be converted into delay logistic equation by substituting $N^* = \frac{n}{N}$, $t^* = \frac{t}{\tau}$, $T^* = \frac{t}{\tau}$ and then solved.

E9) First of all we will find out steady states for that we will do the perturbation $n(t)$ about the positive equilibrium and we will get

$$\frac{dn(t)}{dt} \approx -\mu n(t) - \mu(qz-1)\ n(t-T)$$

and hence that the stability of the equilibriums is determined by $\mathrm{Re}\lambda$.

$$\lambda = -\mu - \mu(qz-1)\ e^{-\lambda T}$$

After this we can perform the stability analysis and find out its stability condition and oscillating conditions.

—x—

APPENDIX

The stability of a population relates to its persistence for a large number of generations. For this, we wish to know if our model possesses any stable steady state or equilibrium. You already know that steady-state or an equilibrium of a difference equation of the form

$$x_{n+1} = f(x_n) \tag{1}$$

is a value x^* such that $x^* = f(x^*)$ so that $x = x^*$ is a constant solution of the difference equation. In other words, $x = x^*$ is a fixed point of the mapping $y = f(x^*)$. Knowing the equilibrium points of Eqn. (1) the next step is to investigate the nature of these equilibrium points.

Stability Analysis

The behaviour of solution near an equilibrium can be studied by using the process of linearisation analogous to that used for differential equations.

To investigate the linear stability of x^*, we write

$$x_n = x^* + \varepsilon_n, \quad |\varepsilon_n| << 1$$

Substituting this in Eqn. (1) and expanding for small ε_n, using Taylor expansion, we get

$$\begin{aligned} x^* + \varepsilon_{n+1} &= f(x^* + \varepsilon_n) \\ &= f(x^*) + \varepsilon_n \, f'(x^*) + O(\varepsilon_n^2) \end{aligned}$$

Since $x^* = f(x^*)$, the equation for determining the linear stability of x^* is given by

$$\varepsilon_{n+1} = \varepsilon_n \, f'(x^*) + \cdots, \quad n = 0, 1, 2, \ldots \tag{2}$$

Let $\lambda = f'(x^*)$ be the **eigenvalue** of the first iterate at the steady-state point x^*. Then Eqn. (2) can be written as

Eigenvalues are the scalar values for which the nontrivial solution of the system exists.

$$\varepsilon_{n+1} = \lambda \, \varepsilon_n$$

and its solution is given by

$$\varepsilon_n = \lambda^n \varepsilon_0$$

which gives $\varepsilon_n \to 0$ as $n \to \infty$ if $|\lambda| < 1$ and $\varepsilon_n \to \pm\infty$ as $n \to \infty$ if $|\lambda| > 1$.

Thus x^* is **stable if** $-1 < f'(x^*) < 1$ and **unstable if** $|f'(x^*)| > 1$.

The steady state x^* is stable if any small perturbation from this equilibrium state decays to zero monotonically when $0 < f'(x^*) < 1$ and with decreasing oscillations when $-1 < f'(x^*) < 0$. On the other hand x^* is unstable if any perturbation grows monotonically when $f'(x^*) > 1$ and by growing oscillations when $f'(x^*) < -1$. This argument can be derived by graphical method also.

If $|f'(x^*)| < 1$ then the equilibrium has the property that every solution of Eqn. (1) with initial value x_0 close enough to x^*, remains close to it and as n tends to infinity, the solution approaches to x^*. This property is called **asymptotic stability** of the equilibrium. In biological applications the asymptotic stability is more important than stability, because an asymptotically stable equilibrium is not significantly affected by perturbations.

Bifurcation and Chaos

The population models of single species involve at least one parameter say r. As this parameter varies the solution of the general model

$$x_{n+1} = f(x_n, r)$$

usually undergoes some qualitative changes at specific values of r resulting in so called **bifurcation**. Such bifurcation can lead to periodic solutions. The frequency of periods increases with the increasing value of the parameter r and when the value of r exceeds some critical value say r_c, chaotic solutions are obtained; the term reflecting their random oscillations. From the graphical analysis by Cobwebbing method, it can be concluded that bifurcation occurs when the eigenvalues of the system pass through $\lambda = 1$, 0 and $\lambda = -1$.

TWO-SPECIES POPULATION MODELS

Structure

9.1 INTRODUCTION

In Unit 8, we discussed two mathematical models on the growth of a single species biological population. In reality, any ecosystem consists of several species which are interrelated amongst themselves. 11 is, therefore, necessary to study multi-species population models to understand the nature and diversity of natural ecosystem. For the sake of simplicity, we shall confine our discussions in this unit to two species only. We shall develop simple mathematical models for the growth of two populations having interrelations in the form of pre-predator or competition. But, to start with in Sec 9.2, we have discussed different types of interactions between two diffcrcnt species living in the same ecosystem. The prey-predator model developed by Vito-Volterra is discussed in Sec 9.3. The population growth model for two competing species is discussed in Sec 9.4. For understanding the discussion in this unit the knowledge of the critical points of a system of differential equations and their stability is essential. For those who are not familiar with the stability of the system of equations we are giving the details in the appendix. You must go through the appendix carefully while reading this unit this will provide you with the necessary background.

Objectives

After reading this unit, you should be able to

- identify different types of interactions between the populations of two species.
- get acquainted with the fundamental mathematical model of a pre-predator system developed by Lotka-Volterra.
- identify some of the major limitations of prey-predator model,
- learn the different features of the basic model for two competing species.

o pinpoint some of the major drawbaclcs of the competition model.

9.2 TYPES OF INTERACTIONS.BETWEEN TWO SPECIES

There may exist various types of interactions between two different species living in the same habitat. Theoretically, the interaction between populations of two species may be described by combinations of neutral organism (0), positive organism(+) and a negative organism (−) in the following patterns:

$$00, --, ++, +0, -0, \text{and } +-.$$

Note that the number of combinations of the three symbols 0, +, −, taken two at a time is ${}^3C_2 = 6$.

Three of these combinations (+ +, −−, and + −) are Further subdivided to get nine types of interactions. These interactions are as follows:

(1) When neither population is affected by association with the other, the type of interaction is called **Neutralism**(00).

(2) When both the species actively inhibit the growth of each other, the type of interaction is Mutual Inhibition Competition(− −).

(3) When both the species compete for a common source of food, they adversely affect each other if the supply of that food is very limited. This type of interaction is Resource **Use** Competition (−−).

(4) When the growth of one species is inhibited by association with a second species and that of the second species is not at all influenced by the first one, the type of interaction is called **Amensalism**(− 0).

(5) When one population adversely affects the other by direct attack and is dependent on the other, the type of interaction is predation or Parasitism (+−).

(6) When one species is benefited by association with a second species and the second species remains unaffected, the type of interaction is called Commensalism (+ 0).

(7) Wlien both populations benefit by the association but the relations between them are not obligatory, the type of interaction is called **Protocooperation**(+ +).

(9) When growth and survival of both the species are strengthened and neither of them can survive under natural conditions without the other, the type of interactions is called **Mutualism**(+ +).

We sum up the above discussion and give the analysis of the interaction between two-species population in the form*of Table 1-2 below.

Table **1**
Analysis of two-species interaction.

0	indicates no significant interaction.
+	indicates growth, survival or other population attributes benefited(positive term added to growth equation)
−	Indicates population growth or other attributes inhibited (negative term added to growth equation).

Table 2
Analysis of two-species population interaction.

NO.	Types of interaction	species 1.	species 2	General nature of interaction
1	Neutralism	0	0	Neither population affects the other
2	Competition: Direct interference type	−	−	Direct inhibition of each species by the otlier
3	Competition: Resource use type	−	−	Indirect inhibition when common resource is in short supply
4	Amensalism	−	0	Population 1 inhibited, 2 not affected
5	Parasitism	+	−	Population 1, the parasite, generally smaller than 2, the host
6	Predation	+	−	Population 1, the predator, generally larger than 2, the
7	Commensalism	+	0	Population 1, the commensal benefits while 2, the host, is not affected
8	Protocooperation	+	+	Interaction favourable to both but not obligatory
9	Mutualism	+	+	Interaction favourable to both and obligatory

And now **a word** of **caution.** Care should be taken in using the various terms. It is seen that the term **Symbiosis** which literally means living together is sometimes used in the same sense as **mutualism;** it is also used at times to cover **commensalism** and parasitism,

Parasite is a small organism that lives on or in another organism, irrespective of its effects being positive(+), negative(-) or neutral (0).

From the above discussion it is clear that interaction between,species may have positive or negative results, For example, mutualism and commensalism are positive interactions arid competition and predation are negative interactions. Ecologists have studied negative interactions involving competition and predation much more than positive interaction. The impact of positive interactions on population growth has rarely been demonstrated, and until more quantitative analyses are done, we cannot evaluate the impact of positive interaction on population abundance.

In this unit we shall discuss the two negative interaction viz., those involving predation and compgtition. The best known models of these phenomena are the Lotka-Volterra equations, which were derived independently by Lotka in 1925 in the United States and by Volterra in 1926 in Italy. Lotkn (1880-1949), an American biophysicist was born in what is now the Ukraine, and was educated mainly in Europe. He is remembered mainly for his formulation of the Lotka-Volterra equations. Vito-Volterra (1860-1940) was born in Ancona, Italy. He gave his theory of interacting species when he was motivated by data collected by a friend, the Italian biologist, Umberto D'Ancona, who was unable to explain the causes of increase of both the selachians (predator shark species) and food fish (prey) in the Mediterranean at the time of first World War when the level of fishing was greatly reduced.

Vito-Volterra (1860-1940)

We start with the mathematical model for the prey-predator relationship

between two species.

9.3 PREY-PREDATOR MODEL

In the predator-prey situation involving two species, one species–the predator–feeds on the other species – the prey – which in turn feeds on some third food item readily available in the environment. For example, population of foxes and rabbits in a woodland; the foxes (predators) eat rabbits (the prey), while the rabbits eat certain vegetation in the woodland. Other examples are sharks (predator) and food fish(prey), bass(predator) and sunfish(prey), ladybugs(predator) and aphids(prey), beetles(predators) and scale insects(prey) etc.

We now give the mathematical formulation of the prey-predator model.

9.3.1 Formulation

To construct a mathematical model, let the first species, the number of prey (or host) at any time t be taken as $x(t)$ and the second species, the size of predator (or parasite) be taken as $y(t)$. Let us assume further the following simplifying assumptions:

1) In the absence of predators, the prey population would grow at a natural rate, with $\frac{dx}{dt} = ax, a > 0$.

2) In the absence of prey, the predator population would decline at a natural rate, with $\frac{dy}{dt} = -cy, c > 0$.

3) When both predator and prey are present, there occurs, in combination with these natural rates of growth and decline, a decline in the prey population and a growth in the predator population, each at a rate proportional to the frequency of encounters between individuals of two species. We assume further that the frequency of such encounters is proportional to the product xy, reasoning that doubling either population alone should double the frequency of encounters, while doubling both populations ought to quadruple the frequency of encounters. Consequently the effect of predators eating prey is an interaction rate of decline $-bxy$ in the prey population $x(t)$, and an interaction rate of growth mxy of the predator population $y(t)$, with b and m being positive constants.
When we add the natural and interaction rates described above, we obtain the predator-prey equations

$$\begin{aligned} \frac{dx}{dt} &= x(a - by) \\ \frac{dy}{dt} &= y(mx - n) \end{aligned} \tag{1}$$

where a, b, m, n are positive constants; a and n are the growth rate of the prey and death rate of the predator respectively, and b and m are measures of the effect of the interaction between the two species.

Eqn.(1) along with the initial conditions $x(0) = x_0$ and $y(0) = y_0$ are known as Lotka-Volterra equations.

Let us now find the solution of the system of differential Eqns. (1).

9.3.2 Solution And Interpretation

Volterra argued that if the size of the prey population x (food fish) be sufficiently large, the predator population y (selachians) has an abundant supply of food and hence y increases. As y goes on increasing, more and more of the prey x is consumed as food and this leads to a rapid decrease of x. As the prey x becomes scarce, y stops increasing due to lack of food, thus allowing the rernaining x to increase again. This cycle of phenomena is repeated over and over again.

When $y(t) = 0$ and $x(t) > 0$, the first equation of system (1) becomes

$$\frac{dx}{dt} = ax$$

This corresponds to Eqn.(5) of Unit 8, i.e., the exponential growth model.

The solution, as you have seen in Eqn.(7) of Unit 8, is $x(t) = x_0e^{at}$, where $x(0) = x_0$.

This shows that, **in the absence of the predators, the prey grows exponentially** according to the Malthusian law of population growth.

On the other hand, when $x(t) = 0$ and $y(t) > 0$, second equation of system(1) becomes

$$\frac{dy}{dt} = -ny$$

This is again similar to Eqn.(5) of Unit 8, the only difference being the coefficient of y is negative. This as you will see now will change the nature of the solution, i.e. instead of an exponentially increasing solution, we get an exponentially decaying solution of the form

$$y(t) = y_0e^{-nt} \text{ where } y(0) = y_0$$

This relation implies that, **in the,absence of the prey, the predator population dies out exponentially** (due to lack of food).

Again if you consider the system of Eqns. (1), you would see that

$$\frac{dx}{dt} = 0 \qquad \Rightarrow x = 0 \qquad \text{or,} \qquad y = \frac{a}{b}$$

and

$$\frac{dy}{dt} = 0 \qquad \Rightarrow x = \frac{n}{m} \qquad \text{or,} \qquad y = 0.$$

Hence the critical (or equilibrium) points of the system given by $\frac{dx}{dt} = 0 = \frac{dy}{dt}$ are $O(0,0)$ and $P(n/m, a/b)$. Here O is a trivial steady state and P is a non-trivial one. The critical point P is of interest, it specifies a constant population $\frac{n}{m}$ of prey and $\frac{a}{b}$ of predator that can coexist with one another in the environment.

A critical point of the system of equations $\frac{dx}{dt} = F(x,y)$, $\frac{dy}{dt} = G(x,y)$ is a point (x^*,y^*) s.t. $F(x^*,y^*) = G(x^*,y^*) = 0$. **Also** then the constant valued functions $x(t) = x^*, y(t) = y^*$ **satisfying the system** is **called an equilibrium solution.**

Before actually solving the system (4 let us analyse what the system represents geometrically.

Geometrical Interpretation

If $y = 0$ and $x > 0$ at some instant, we find $\frac{dy}{dt} = 0$ and $\frac{dx}{dt} = ax > 0$. This means that the predator population continues to remain at the zero level

while the prey population goes on increasing. Geometrically this means that 'the positive x-axis ($y = 0$) is an orbit of the system.

On the other hand, if $x = 0$ and $y > 0$ at-any time, we have $\frac{dx}{dt} = 0$ while $\frac{dy}{dt} = -ny < 0$.

This implies that the preypopulation continues to remain at the zero level while the predator population goes on decreasing. Geometrically this means that the negative y-axis ($x = 0$) is an orbit of the system.

This analysis reconfirms our previous observations that

(i) the prey grows exponentially in the absence of the predator and

(ii) the predator dies out exponentially in the absence of the prey.

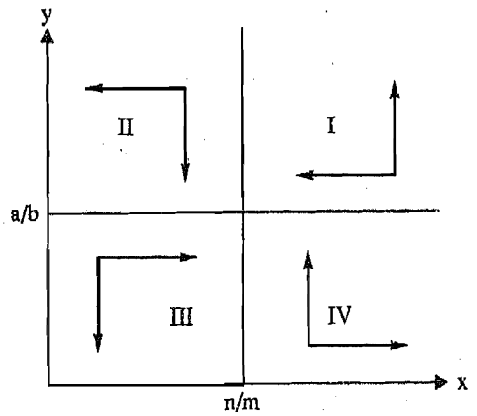
Thus the equilibrium solution $x(t) = 0$, $y(t) = 0$ corresponding to the critical point (0, 0) describes simultaneous extinction of both species.

Since $x(t) \geq 0$ and $y(t) \geq 0$ for all times, all other orbits of the system lie entirely in the first quadrant of the x-y plane.

To get an idea of the other orbits, we first note that $y = a/b$ and $x = n/m$ divide the first quadrant into four regularly shaped regions. We see that $\frac{dx}{dt} > 0$ if $y < \frac{a}{b}$ while $\frac{dx}{dt} < 0$ if $y > \frac{a}{b}$.

This means that x increases in the regions III and IV, and decreases in the regions I and II in Fig, 1.

Moreover, $\frac{dy}{dt} > 0$ if $x > \frac{n}{m}$ and $\frac{dy}{dt} < 0$ if $x < \frac{n}{m}$. This implies that y increases in the regions I and IV while decreases in II and III (Fig.1)



Region	$\frac{dx}{dt}$	$\frac{dy}{dt}$
I	< 0	> 0
II	< 0	< 0
III	> 0	< 0
IV	> 0	> 0

Fig. 1

It is clear from Fig.1 that the orbit will follow a counter clockwise direction about the critical point $(\frac{n}{m}, \frac{a}{b})$ whatever be the initial sizes of the populations. For example, if there are small numbers of prey and predator initially, i.e. if the orbit begins in the region III, then the prey increases and the predator decreases. This is what is expected in reality. For, a small number of foxes poses little threat to the rabbits so that the rabbits go on increasing in number. On the other hand, scarcity of the rabbits forces the fox population to decline. When the size of the rabbit population exceeds the critical value $\frac{n}{m}$, the orbit is in region IV and then the fox population also begins to increase due to availability of sufficient food (rabbits).

When the fox population exceeds the critical value $\frac{a}{b}$, the orbits enters the region I. Now foxes being plenty in number to endanger the rabbits, the

rabbit population begins to decrease. Ultimately when the rabbit population declines below the critical level $\frac{n}{m}$, the orbit enters the region II. As a result of declining rabbit population, now the fox population also begins to decline due to shortage in food supply. When the fox population declines below the critical value $\frac{a}{b}$, there is a small number of foxes to endanger the lives of the rabbits existing at that point of time. As a result, the rabbits start growing and we are again in region IV. This cycle of phenomena continues to repeat again and again. **Thus the fluctuation of the populations follows some kind of cyclical pattern about the critical point** $(\frac{n}{m}, \frac{a}{b})$. Let us denote this critical point as (x^*, y^*).

The above discussion gives you a qualitative description of the growth of the two species. In order to get the quantitative or accurate description, we need to solve the system and get an algebraic relation between x and y,

Analytic Solution

To find the solution of system of Eqns. (1) for $x(t) > 0, y(t) > 0$ with initial conditions $x(0) = x_0, y(0) = y_0$ we write

$$\frac{dy}{dx} = \frac{y(mx - n)}{x(a - by)}$$

$$\text{or, } \frac{a - by}{y}dy = \frac{mx - n}{x}dx$$

$$\text{or, } a\frac{dy}{y} - bdy = mdx - n\frac{dx}{x}$$

which on integration gives,

$$a \ln y - by - mx + n \ln x = \ln K_1,$$

where K_1 is the constant of integration to be determined using the initial conditions.

We have,

$$\ln y^a - \ln e^{by} - \ln e^{mx} + \ln x^n = \ln K_1$$

$$\text{or, } \ln \frac{y^a}{e^{by}} \frac{x^n}{e^{mx}} = \ln K_1$$

Hence,

$$\frac{y^a}{e^{by}} \frac{x^n}{e^{mx}} = K_1 \tag{2}$$

Thus Eqn.(2) which represents a family of closed curves gives the solution of system of Eqns. (1).

We may write Eqn.(2) in the form

$$\left(\frac{y}{e^{\frac{b}{a}y}}\right)^a \left(\frac{x}{e^{\frac{m}{n}x}}\right)^n = K_1$$

$$\text{or, } \left(\frac{y}{e^{y/y^*}}\right)^a \left(\frac{x}{e^{x/x^*}}\right)^n = K_1,$$

where $x^* = n/m$, and $y^* = a/b$.

Using the transformation $X = x/x^*, Y = y/y^*$, we have

$$\left(\frac{Yy^*}{e^Y}\right)^a \left(\frac{Xx^*}{e^X}\right)^n = K_1$$

$$\text{or, } \left(\frac{e^X}{X}\right)^n \left(\frac{e^Y}{Y}\right)^a = \frac{1}{K_1} y^{*a} x^{*n} = C(\text{say}) \tag{3}$$

where the constant K and hence C has to be determined using the initial conditions $x(0) = x_0$ and $y(0) = y_0$. Using these conditions we obtain

$$C = \frac{x^{*n} y^{*a}}{K_1} = \left(\frac{e^{x_0/x^*}}{x_0/x^*}\right)^n \left(\frac{e^{y_0/y^*}}{y_0/y^*}\right)^a \tag{4}$$

Thus for a given value of (x_0, y_0) value of C is known. So' the final solution

$$\left(\frac{e^x}{X}\right)^n \left(\frac{e^y}{Y}\right)^a = C \tag{5}$$

can be obtained. We have already had a qualitative picture of this solution through geometrical considerations (see Fig. 1) and hence we know what to expect. in order to verify the findings of Fig. 1 we have to plot the curve given by Eqn.(5). You may observe that Eqn.(5) does not represent a curve with which you are already familiar say, ellipse or parabola etc. Without going into the details we give below the plot of the curve (5) in Fig. 2. For each fixed value of C the graph is a closed curve enclosing the point $\frac{n}{m}, \frac{a}{b}$.

It may be observed that as C increases, x and y show oscillations of increasing amplitude (Fig.2). At the minimal value, these curves shrink into a point with coordinates (x*,y*).

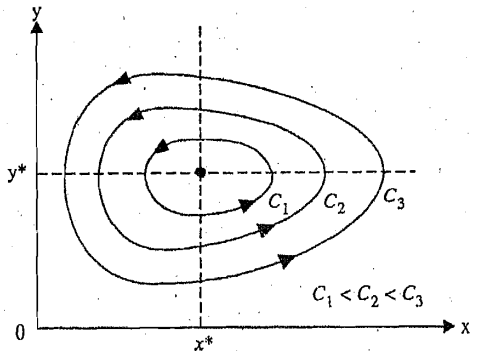


Fig..2

Note the direction of the arrows in Fig. 2.' These are drawn based on our earlier geometric considerations.

For **a** linear system, critical point P is **stable** if for initial population (x_0, y_0) close to P, the population $(x(t), y(t))$ remain near it for all $t > O$

Stability

To check the stability of the critical point $P\left(\frac{n}{m}, \frac{a}{b}\right)$, and get an idea of the' pattern of the orbits near the critical point, i.e. whether the orbits are moving towards the critical point or moving away from it or exhibiting some other type of behaviaur, we use the perturbation technique. The basic idea of this technique is to perturb or disturb the equilibrium slightly and then to see whether the system remains in the neighbourhood of the equilibrium or deviates far away from it. Mathematically, we change the equilibrium values of x and y slightly by adding to them very small quantities.

$$\text{Let } x = \frac{n}{m}(1+u), y = \frac{a}{b}(1+v) \tag{6}$$

where u, v are very small quantities. This transformation indicates small departure from the equilibrium point $\left(\frac{n}{m}, \frac{a}{b}\right)$.

We have from Eqns.(1) and (6),

$$\begin{aligned} \frac{du}{dt} &= -av - auv \\ \frac{dv}{dt} &= nu + nuv \end{aligned} \tag{7}$$

Clearly the system of Eqns. (7) is almost linear system and has (0, 0) as the critical point corresponding to the critical point $\left(\frac{n}{m}, \frac{a}{b}\right)$ of the system of Eqns. (1).

In order to check the nature and stability of the critical point of system (7) we consider the related linear system

$$\begin{aligned} \frac{du}{dt} &= -av \\ \frac{dv}{dt} &= nu \end{aligned} \tag{8}$$

The eigenvalues of system (8) are given by the equation

$$|A - \lambda I| = 0 \tag{9}$$

where $A = \begin{pmatrix} 0 & -a \\ n & 0 \end{pmatrix}$

We obtain from Eqn.(9)

$$\begin{vmatrix} -\lambda & -a \\ n & -\lambda \end{vmatrix} = 0$$

$$\Rightarrow \lambda^2 + an = 0, \text{ or } \lambda = \pm i\sqrt{an} \tag{10}$$

Thus the eigenvalues of the system (8) are pure imaginary. We thus conclude that critical point (0, 0) of the system (8) is a **center.** Further differentiating the two equations of the system (8) with respect to t, we obtain

$$\frac{d^2u}{dt^2} = -a\frac{dv}{dt} = -anu, \tag{11}$$

$$\frac{d^2v}{dt^2} = n\frac{du}{dt} = -anv, \tag{12}$$

Could you recognise Eqns.(11) and (12)?

Yes! both these equations represent a simple harmonic motion of periodic time $T = \frac{2\pi}{(an)^{1/2}}$ (Ref. Unit 4, Sec. 4.4).

Thus the trajectories of the system (8) are closed curves exhibiting periodic oscillations of period $\frac{2\pi}{(an)^{1/2}}$ in the neighbourhood of point (0, 0). We can further show that these closed curves are ellipses in this case.

We multiply the first equation of system (8) by nu, and the second by av and then adding together, we have

$$nudu + avdv = 0.$$

This gives on integration

$$\begin{aligned} nu^2 + av^2 &= A, \\ \text{or, } \frac{u^2}{\lambda/n} + \frac{v^2}{\lambda/a} &= 1. \end{aligned} \tag{13}$$

where λ is an arbitrary nonnegative constant of integration.

Thus the trajectories of the system (8) are ellipses around the critical point $(0, 0)$. Some of these ellipses are shown in Fig. 3.

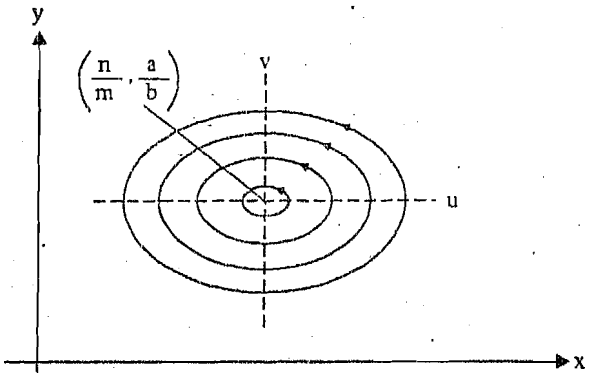


Fig. 3

We have shown that the critical point $(0, 0)$ is a stable center of the linear system (8). Wc now need to assess its character for the almost linear system ('7). Here as we know, our theory for almost linear systems fails (ref. Table 1 of appendix). The effect of the nonlinear terms may be to change the center into a stable spiral point, or into an unstable spiral point, or it rnay remain as a stable center. Fortunately, in this case we have actually solved the nonlinear Eqns.(1) and seen,what happens. We have shown in Fig. (2) that the graph of this equation for a fixed value of C in Eqn.(5) is a closed curve (not an ellipse bpt deformed ellipse) enclosing the critical point $\left(\frac{n}{m}, \frac{a}{b}\right)$. Thus the predator and prey have a cyclic variation about the critical point $\left(\frac{n}{m}, \frac{a}{b}\right)$ and the critical paint is also a center of the system (1).

Let us now consider the following examples.

Example-1: For the system of equations

$$\begin{aligned} \frac{dx}{dt} &= x - y + xy \\ \frac{dy}{dt} &= 3x - 2y - xy \end{aligned} \tag{14}$$

verify that $(0, 0)$ is a critical point. Show that the system is almost linear and discuss the type and stability of the critical point $(0, 0)$.

Solution: Clearly $(0, 0)$ is a critical point of the system (14). System(14) can be written in the form

$$\begin{aligned} \frac{dx}{dt} &= x - y + f(x, y) \\ \frac{dy}{dt} &= 3x - 2y + g(x, y) \end{aligned}$$

where $f(x, y) = xy$ and $g(x, y) = -xy$

For checking the condition for almost linear system it is convenient to use polar coordinates by letting $x = r\cos\mathbf{0}, y = r\sin\theta$.

Now $\frac{f(x, y)}{r} = \frac{r^2 \cos 0 \sin\theta}{r} = r\cos\theta\sin\theta \to 0$ as $r \to 0$

dso $\frac{g(x, y)}{r} = \frac{-r^2\cos\theta\sin\theta}{r} = -r\sin\theta\cos\theta \to 0$ as $r \to 0$

Thus system (14) is almost linear. The related linear system in the neighbourhood of $(0, 0)$ is

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 3 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \tag{15}$$

Eigenvalues of **(15)** are the roots of the equation $\begin{vmatrix} 1-\lambda & -1 \\ 3 & -2-\lambda \end{vmatrix} = 0$

$\Rightarrow \lambda^2 + \lambda + 1 = 0$ or $\lambda = \dfrac{-1 \pm i\sqrt{3}}{2}$

$\therefore \lambda_1 = \dfrac{-1+i\sqrt{3}}{2}$ and $\lambda_2 = \dfrac{-1-i\sqrt{3}}{2}$. Since the eigenvalues are conjugate complex of the form $\lambda \pm i\mu$, λ, μ real. Critical point (0, 0) of the system **(15)** is a spiral. Also since $\lambda < 0$, it is asymptotically stable point. Since the system **(14)** is almost linear, critical point (0, 0) of the system is also **asymptotically stable** spiral point.

Example-2: Consider the systern of equations

$$\begin{aligned} \frac{dx}{dt} &= x \\ \frac{dy}{dt} &= -x + 2y \end{aligned} \tag{16}$$

Find the critical point of the system. Discuss the type and stability of the critical point. Write down the general solution of the system **(16)** and sketch the graph of its trajectories.

Solution: Clearly (0, 0) is the critical point of the system **(16).**

Eigenvalues of **(16)** are the roots of the equation

$$\begin{vmatrix} 1-\lambda & -0 \\ -1 & -2-\lambda \end{vmatrix} = 0$$

$\Rightarrow \lambda^2 - 3\lambda + 2 = 0$

$\Rightarrow \lambda_1 = 1, \lambda_2 = 2.$

Eigenvalues are real, distinct and of the same sign so the critical point is a mode. Also since $\lambda_1 > 0, \lambda_2 > 0$ it is unstable.

To find the general solution of system (16), we find the eigonvectors corresponding to the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 2$.

Eigenvector corresponding to the eigenvalue $\lambda_1 = 1$ is the solution of the equation

$$\begin{pmatrix} 0 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We see that $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is one possible eigenvector. Similarly $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ is one possible eigenvector corresponding to the eigenvalue $\lambda_2 = 2$.

Then the general solution of system **(16)** can be written as

$$\begin{pmatrix} x \\ y \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^t + c_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{2t}$$

$$\begin{aligned} \Rightarrow x &= c_1 e^t \\ y &= c_1 e^t + c_2 e^{2t} \end{aligned} \tag{17}$$

where c_1, c_2 are arbitrary constants.

For $c_1 = 0, x = 0$ and $y = c_2 e^{2t}$. In this case the trajectory is positive y axis when $c_2 > 0$ and it is negative y axis when $c_2 < 0$ and also since $y \to \infty$ as $t \to \infty$, each path approaches ∞ as $t \to \infty$.

For $c_2 = 0, x = e_1 e^t, y = c_1 e^t$. This trajectory is a half line $y = x, x > 0$ when $c_1 > 0$ and the half line $y = x, x < 0$ when $c_1 < 0$ and again both paths $\to \infty$ as $t \to \infty$.

When both c_1 and c_2 are $\neq 0$, the trajectories are parabolas $y = x + (c_2/c_1^2)x^2$ which passes through the origin with slope 1. Each of these trajectories also approach ∞ as $t \to$ do. The sketch of the trajectories is shown in Fig.(4).

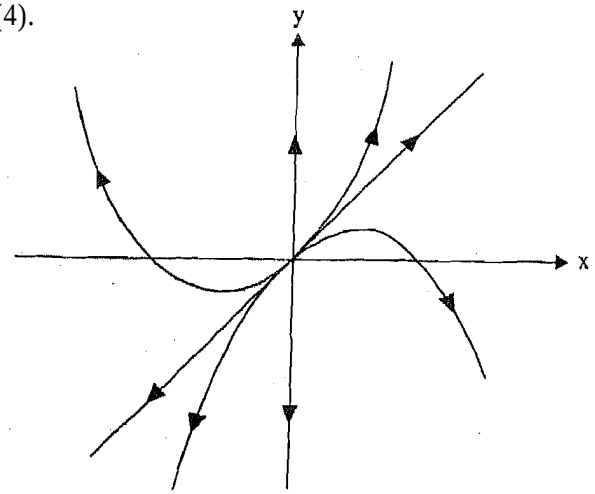


Fig. 4

Example 3 : Determine the type and stability of the critical point (0, 0) of the almost linear system

$$\begin{aligned} \frac{dx}{dt} &= 4x + 2y + 2x^2 - 3y^2, \\ \frac{dy}{dt} &= 4x - 3y + 7xy. \end{aligned} \tag{18}$$

Find the general solution of the corresponding linear system and sketch its trajectories.

Solution : The auxiliary equation of the associated linear system

$$\begin{aligned} \frac{dx}{dt} &= 4x + 2y \\ \frac{dy}{dt} &= 4x - 3y \end{aligned} \tag{19}$$

is $(4 - \lambda)(-3 - \lambda) - 8 = (\lambda - 5)(\lambda + 4) = 0$

The roots $\lambda_1 = -4$ and $\lambda_2 = 5$ are real unequal and have opposite sign. So the critical point (0, 0) is an unstable saddle point of the system (19) and hence of the system (18).

Eigenvector corresponding to $\lambda_1 = -4$ is $\begin{pmatrix} 1 \\ -4 \end{pmatrix}$ and that corresponding to $\lambda_2 = 5$ is $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$. So the general solution of (19) can be written as

$$\begin{aligned} x &= c_1 e^{-4t} + 2c_2 e^{5t} \\ y &= -4c_1 e^{-4t} + c_2 e^{5t} \end{aligned} \tag{20}$$

For $c_1 = 0, x = 2c_2 e^{5t}, y = c_2 e^{5t}$. This trajectory is the half line $y = \frac{x}{2}, x > 0$ when $c_2 > 0$ and half line $y = \frac{x}{2}, x < 0$ when $c_2 < 0$. Also $x \to \infty, y \to \infty$ as $t \to \infty$.

For $c_2 = 0, x = c_1 e^{-4t}, y = -4c_1 e^{-4t}$. This trajectory is the half line $y = -4x, x > 0$ when $c_1 > 0$ and the half line $y = -4x, x < 0$ when $c_1 < 0$. Both the trajectories approach and enter the origin as $t \to \infty$.

If $c_1 \# 0 c_2 \neq 0$, solution (20) represents curved trajectories none of which approaches (0, 0) as $t \to \infty$. Fig. (5) gives a qualitative picture of this behaviour.

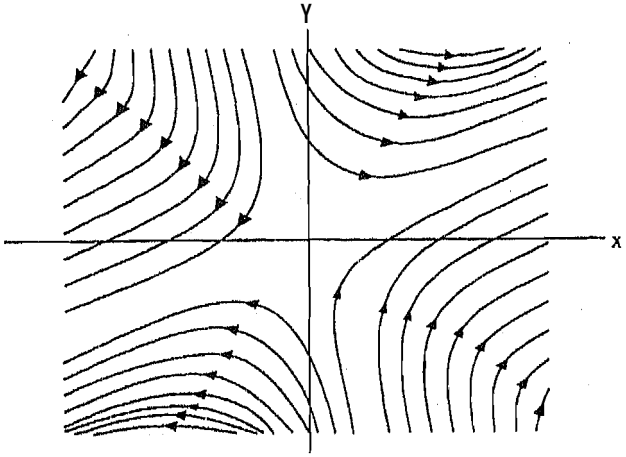


Fig. 5

And now a few exercises for you.

E1) In each of the following problems, verify that (0, 0) is a critical point, show that the system is almost linear, arid discuss the type and stability of the critical point (0, 0).

a) $\frac{dx}{dt} = y + x(1 - x^2 - y^2)$

$\frac{dt}{dx} = -x + y(1 - x^2 - y^2)$

b) $\frac{dx}{dt} = 2x + y + xy^3$

$\frac{dy}{dt} = x - 2y - xy$

E2) Determine all real critical points of each of the following system of equations and discuss their type and stability.

a) $\frac{dx}{dt} = x + y^2$

$\frac{dy}{dt} = x + y$

b) $\frac{dx}{dt} = 1 - xy$

$\frac{dy}{dt} = x - y^3$

E3) One improvement in the predator-prey model is to modify the equation for the prey so that it has the forrn of a logistic equation in the absence of the predator. What will be the form of system of equations modelling this situation. Determine all critical points of the system and discuss their nature and stability.

E4) Consider the linear system

$\frac{dx}{dt} = -x + hy$

$\frac{dy}{dt} x - y$

Discuss the nature; and stability of the critical point (0, 0) if

a) $h = 0$ b) $h < 0$ c) $0 < h < 1$

What do you conclude from the above three cases?

E5) Find the critical point of the system

$$\frac{dx}{dt} = -y$$

$$\frac{dy}{dt} = x$$

and discuss its nature and stability. Find the general solution of the system and sketch its trajectories.

Before taking up the model for the competing species let us discuss the limitations of this model.

9.3.3 Limitations

From the above discussion it is clear that Volterra system possesses no mechanism to maintain its non-trivial steady state. It is seen from Fig. 2 that the prey-predator system switches from one orbit to another for arbitrary small changes in the phase coordinates (x^*, y^*). In the mathematical sense, we describe this behaviour of the system by saying that Volterra orbits lack "roughness"

We also observed that

(i) in the absence of predators, the prey population grow, unbounded exponentially and

(ii) in the absence of prey, the predator population goes to extinction due to lack of food.

These phenomena arc not found to occur in reality. In the absence of predators, the prey population is expected to increase rapidly to start with; after considerable increase in its size, its growth must be retarded due to crowding effects and ultimately, it cannot increase beyond a limiting level. On the other hand, when prey (food) is not available, the predator population is expected to decrease rapidly in the beginning; after some time, the predators are likely to adjust themselves with the situation by finding alternative sources of food.

So far we discussed mathematical model for two species in which one species preyed upon the other. In contrast to this, we shall now consider two species which compete with each other for the food available in their common environment.

9.4 COMPETING SPECIES

In the broadest sense, the term "competition" between two living organisms refers to the interaction between them when they strive for the same thing. As we have already seen in Table 2, this competition may be of two types – indirect or resource competition and direct or interference competition, Resource competition occurs when a number of organisms (of the same or of different species) utilise common resources that are in short supply. Interference occurs when the organisms seeking a resource harm one another, in the process, even if the resource is not in short supply. These competitions or interactions between the populations of two or more species adversely affects their growth and survival. The tendency for competition to bring about an ecological separation of closely related, or otherwise similar species is known as the competitive exclusion principle.

Note that the competition may be interspecific (between two or more

different species) or intraspecific (between members of the same species). Here we restrict ourselves to the study of populations of two species whir.": are competing for a common resource (food, space, light, etc.).

Competition occurs over resources, and a variety of resources may become the center of competitive interactions. For plants, light, nutrients, and water may be important resources, but plants may compete for pollinators or for attachment sites. Water, food and males arc possible sources of competition 'for animals. Competition for space also occurs in some animals and may involve many types of specific requirements, such as nesting sites, wintering sites, or sites that are safe from predators. Thus resources are diverse and complex.

Let us now formulate the mathematical model for the interspecific competition for a common resources.

9.4.1 Formulation

Suppose that there are two species living in the same environment and having.a common source of food. At any time t, let x(t) and y(t)) denote the number of individuals in the populations of the two species and $x(0) = x_0$ and $y(0) = y_0$ be their initial populations. If $r_1 (> 0)$ and $r_2 (> 0)$ be their growth coefficients, then the differential equations for this competing system may be written as

$$\begin{aligned} \frac{dx}{dt} &= r_1 x - \alpha_1 xy \\ \frac{dy}{dt} &= r_2 y - \alpha_2 xy \end{aligned} \qquad (21)$$

where α_1 and α_2 are two positive constants, These equations are also given by **Lotka-Volterra**.

Since the competition between two species has the effect of a rate of decline in each population proportional to their product xy, the terms $\alpha_1 xy$ and $\alpha_2 xy$ in system (21) indicate interaction between the x and y species. If you compare the two systems (1)and (21), modelling the prey-predator and competing species respectively what difference do you find between them? You would notice that competition between two species for a common resource has a declining effect (xy term negative) on the rate of growth of both the species, whereas the ,interaction between the two species has a declining effect on the growth rate of prey and increasing effect (xy term positive) on the growth rate of predator.

The coefficients α_1 and α_2 in Eqns.(21) are called the **coefficients** of **interspecific competition** for the two species.

111 the absence of the second species (i.e. when y = 0), the first equation of system (23.) becomes $\frac{dx}{dt} = r_1 x$ of which the solution (as we have already seen in Unit 8) is $x = x_0 e^{r_1 t}$, x_0 being the initial density of the first species.

This shows that the first species grows exponentially in the absence of the second species, being the sole user of the food resource. We arrive at a similar conclusion for the second species (i.e. when x = I)). Thus **each species grows unbounded in the absence of the other.**

When both the species are present, their growth rates are bound to decrease due to sharing of food. To quantify this amount of decrease in the respective growth rate, we may argue, as follows:

The decrease in the growth rate of x-species is proportional to x when y is constant, and is proportional to y when x is constant. Then the decrease in the growth rate of x-species is proportional to the product xy when both x and y vary. Similar arguments holds for the decrease in the growth of y-species. Based on these characteristics of the two species, the model bas been formulated here.

We now try to find the solution of the system of differential Eqns. (21) and discuss it.

9.4.2 Solution and Interpretation

The equilibrium solution of the system of Eqns.(21) is given by

$$\frac{dx}{dt} = 0 \qquad \Rightarrow x = 0 \qquad \text{or,} \qquad y = r_1/\alpha_1$$

and

$$\frac{dy}{dt} = 0 \qquad \Rightarrow y = 0 \qquad \text{or,} \qquad x = r_2/\alpha_2.$$

Thus the non-trivial steady-state or equilibrium point of the system is (x^*, y^*) where $x^* = r_2/\alpha_2$ and $y^* = r_1/\alpha_1$.

It is interesting to note that the equilibrium density of one species depends on the proportional growth rate and the coefficient of inter-specific coefficient of the other species.

Just as we did in the case of prey-predator model, we analyse the system of Eqns.(21) geometrically.

Geometrical Interpretation

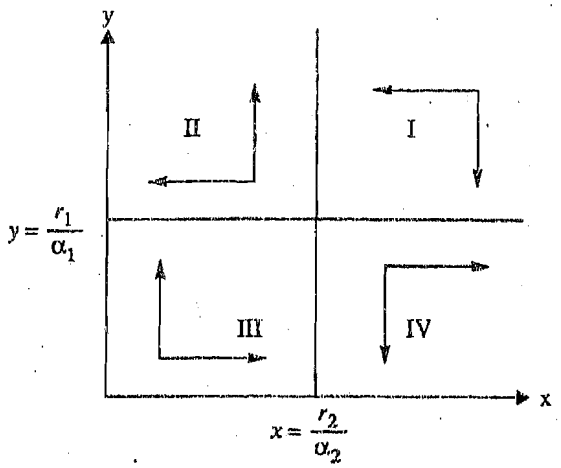
If $y = 0$ and $x > 0$ at some instant, we find $\frac{dy}{dt} = 0$ and $\frac{dx}{dt} = r_1 x > 0$. This means the population of second species continues to remain at the zero level while the first goes on increasing. The positive x-axis ($y = 0$) is an orbit of the system in this case and we say that x-species outcornpetes the y-species.

Similarly, if $x = 0$ and $y > 0$ at any time, we have $\frac{dx}{dt} = 0$ and $\frac{dy}{dt} = r_2 y > 0$. This implies that the first species continues to remain at the zero level while the second goes on increasing. In this case, the positive y axis ($x = 0$) in an orbit of the system and y-species outcornpetes the x-species. This sort of phenomena in population biology is known as the principle of competitive exclusion.

This analysis reconfirms our previous observation that each species grows unbounded in the absence of the other.

When $x(t) \geq 0$ and $y(t) \geq 0$, all other orbits of the system lie entirely in the first quadrant of the x-y plane. Now $x = \frac{r_2}{\alpha_2}$ and $y = \frac{r_1}{\alpha_1}$ divide the first quadrant into four regions. We see that $\frac{dx}{dt} > 0$ if $y < \frac{r_1}{\alpha_1}$ while $\frac{dx}{dt} < 0$ if $y > \frac{r_1}{\alpha_1}$.

This means that x increases in the III and IV regions and decreases in I and II Regions. Similarly, $\frac{dy}{dt} > 0$ if $x < \frac{r_2}{\alpha_2}$ and $\frac{dy}{dt} < 0$ if $x > \frac{r_2}{Q2}$ implying that y increases in regions II and III and decreases in I and IV (see Fig. 6).



Region	$\frac{dx}{dt}$	$\frac{dy}{dt}$
I	< 0	< 0
II	< 0	> 0
III	> 0	> 0
IV	> 0	< 0

Fig. 6

It is clear from Fig. 6 that orbits do not follow any kind of cyclic pattern. Thus the fluctuations of the populations do not follow any kind. of cyclic pattern.

Stability

We now examine the stability of the steady *state* (x^*, y^*) by using the perturbation technique.

Let $x = x^*(1 + u)$ and $y = y^*(1 + v)$ (22)

where u and v are very small and indicate small deviations from the equilibrium.

Using Eqns.(22) in (21), we have

$$\frac{du}{dt} = -r_1 v - r_1 uv$$
$$\text{and } \frac{dv}{dt} = -r_2 u - r_2 uv \tag{23}$$

System (23) is almost linear and has (0, 0) as the critical point corresponding to the critical point $\left(\frac{r_2}{\alpha_2}, \frac{r_1}{\alpha_1}\right)$ of the system (21).

To examine the stability of the critical point (0, 0) of the system (23), we consider the related linear system

$$\frac{du}{dt} = -r_1 v$$
$$\frac{dv}{dt} = -r_2 u \tag{24}$$

Eigenvalues of the system (24) are $\lambda = \pm\sqrt{r_1 r_2}$, i.e. eigenvalues are real, distinct and of opposite signs. Thus the critical point is a **saddle point.** Differentiating tho equations of the system (24) once again w.r.t. t and eliminating the first derivative terms we obtain

$$\frac{d^2u}{dt^2} = r_1 r_2 u \tag{25}$$
$$\text{and } \frac{d^2v}{dt^2} = r_1 r_2 v \tag{26}$$

The general solution of Eqn.(25) is of the form

$$u = A_1 e^{\sqrt{(r_1 r_2)}t} + A_2 e^{-\sqrt{(r_1 r_2)}t}$$

where A_1 and A_2 are arbitrary constants.

We thus find that $u \longrightarrow \infty$ as $t \longrightarrow \infty$.

Similarly, on solving Eqn.(26), we find that $v \longrightarrow \infty$ as $t \longrightarrow \infty$.

Thus critical point (0, 0) is unstable saddle point of system (24) and hence of the system **(23)** (ref Table-1 of the appendix).

It is, therefore, clear that the steady state (x^*, y^*) of the system (21) is unstable. The point (x^*, y^*) moves on to either x-axis or y-axis in the **(x,** y)−plane, depending on the initial conditions.

You may be wondering in this case why we did not we solve the system of Eqns.(21) analytically like we did for the prey-predator model. Yes! you can solve the system (21) and find its analytical solution in this case also. We are leaving it for you to do it yourself.

E6) Solve the system of Eqns.(21) analytically.

9.4.3 Limitations

The major limitation of this model lies in the extreme outcome that one species may be such a strong conipetitor that it, may force the other species to go extinct. In the natural environment, however, populations are distributed over space, and space is strongly inhomogeneous. A species that is completely out-competed by another species, may find various refuges where it can continue to survive, at least in small numbers.

It is also found in natural environment that two species competing for a common resource for their survival coexist. This model fails to exhibit such coexistence of two competing species.

Another limitation of the model lies in the observation that each species grows unbounded in the absence of the other. This can never happen in reality -.there must be carrying capacity for the growing species.

We now end this unit by giving a summary of what we have covered in it.

9.5 SUMMARY

In this unit, we have covered the following:

(I) Any ecosystem consists of several species which are interrelated amongst themselves. It is therefore necessary to study multi-species population models to understand the nature and diversity of ecosystem,

(2) The prey-predator model due to Lotka-Volterra involves two species in which, one species – the predator feeds on the other species – the prey.

(3) For a prey and predator population of sizes $x(t)$ and $y(t)$ respectively, at any time t the Lotka-Volterra equations are

$$\frac{dx}{dt} = x(a - by), x(0) = x_0,$$

$$\text{and } \frac{dy}{dt} = y(mx - n), y(0) = y_0$$

where a, b, m, n are positive constants.

(4) The above system of equations has two critical points $O(0, 0)$ and $P\left(\frac{n}{m}, \frac{a}{b}\right)$. The critical point O is a trivial steady state but point P is of interest. It specifies a population of prey and predator that can coexist with one another in the environment.

(5) Geometrical analysis of the system in **(3)** shows that prey-predator population follows an orbital path. Elliptic orbits are obtained around the critical point $\left(\frac{n}{m}, \frac{a}{b}\right)$.

(6) The prey-predator model has some limitations as it possess no mechanism to maintain its non-trivial steady state. The prey predator system switches from one orbit to another for small changes in the coordinates $\left(\frac{n}{m}, \frac{a}{b}\right)$. Also, in the absence of predators, the prey grow unbounded while in the absence of prey, the predator population goes to extinction. But such phenomenon are not found to occur in reality.

(7) Besides prey-predator relationship there is competition between two living organisms in the nature. There is interaction between them when they strive for the same thing. Model for two competing species having population $x(t)$ and $y(t)$ at any time t, is given by

$$\frac{dx}{dt} = r_1 x - \alpha_1 xy, x(0) = x_0,$$

$$\text{and } \frac{dy}{dt} = r_2 y - \alpha_2 xy, y(0) = y_0$$

where $r_1 > 0$ and $r_2 > 0$ are their growth coefficients and α_1, α_2 are positive constants.

(8) The non-trivial steady-state or the critical point $\left(\frac{r_2}{\alpha_2}, \frac{r_1}{\alpha_1}\right)$ of the above system is unstable. An outcome of this model is that one species may be such a strong competitor that it may force the other species to go extinct,, which is the major limitations of the model. Also, it gives that one species grows unbounded in the absence of the other. But these things can never happen in reality in nature.

9.6 SOLUTIONS/ANSWERS

E1) a) Spiral point, unstable

b) Saddle point, unstable

E2) a) Critical points are $(0, 0)$ and $(-1, 1)$. Point $(0, 0)$ is a node or spiral point, it is unstable.
To check the stability of $(-1, 1)$
Let $x = 4 - 1$ and $y = v + 1$
So the given system reduces to

$$\frac{du}{dt} = u + 2v + v^2$$

$$\frac{dv}{dt} = u + v.$$

which has $(0, 0)$ as the critical point corresponding to the critical point $(-1, 1)$ of the given system. Now verify that critical point is unstable saddle point.

b) $(1,1)$ node or spiral point, asymptotically stable
$(-1, -1)$ saddle point unstable.

E3) Equations modelling this situation are of the form

$$\frac{dx}{dt} = x(a - by - cx)$$

$$\frac{dy}{dt} = y(mx - n)$$

where a, b, c, m, n are all positive constant. The critical points of this system are $(0,0)\left(\frac{a}{c},0\right)\left(\frac{n}{m},\frac{am-cn}{mb}\right)$ Where (0, 0) is a saddle point, $\left(\frac{a}{c},0\right)$ is a saddle point and $\left(\frac{n}{m},\frac{am-cn}{mb}\right)$ is an asymptotically stable node if $\left(\frac{nc}{m}\right)^2 - 4nb\left(\frac{a}{c}-\frac{n}{m}\right) \geq 0$ and an asymptotically stable spiral if $\left(\frac{nc}{m}\right)^2 - 4nb\left(\frac{a}{c}-\frac{n}{m}\right) < 0$.

E4) a) For $h = 0, (0,0)$ is asymptotically stable node

b) For $h < 0, (0,0)$ is asymptotically spiral point

c) For $0 < h < 1, (0,0)$ is asymptotically stable node. Small perturbation of the system $x' = -x, y' =$ y, can change the type of the critical, point (0, 0) without affecting its stability.

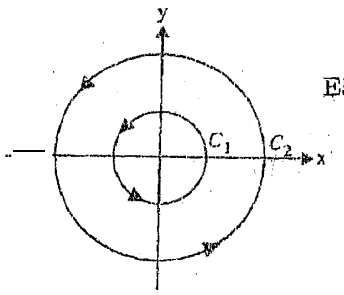


Fig.7

E5) Critical point (0, 0) is a center. General solution of the system is

$x = -c_1 \sin t + c_2 \cos t$

$y = c_1 \cos t + c_2 \sin t$

So x(t) and y(t) are periodic and each trajectory is a closed curve surrounding the origin.

Also we have from given system $\frac{dy}{dx} = -\frac{x}{y}$ whose general solution is $x^2 + y^2 = c^2$ this yields all the curves which are circles (see Fig.7).

Also from the given differential equations, for the region $x > 0, y > 0$, we see that $\frac{dx}{dt} < 0$, mean x decreases with t, $\frac{dy}{dt} > 0$, means y increases with t. Thus the trajectories are anticlockwise round the circle.,

E6) Prom Eqns. (21) we have

$$\frac{dy}{dx} = \frac{y(r_2 - \alpha_2 x)}{x(r_1 - \alpha_1 y)}$$

$$\text{or, } \left(\frac{r_1 - \alpha_1 y}{y}\right) dy = \left(\frac{r_2 - \alpha_2 x}{x}\right) dx$$

$$\text{or, } r_1 \frac{dy}{y} - \alpha_1 dy = \frac{r_2}{x} dx - r_2 dx$$

Integrating

$$r_1 \ln y - \alpha_1 y - r_2 \ln x + \alpha_2 x = \ln K$$

$$\text{or, } \ln \frac{y^{r_1} e^{\alpha_2 x}}{x^{r_2} e^{\alpha_1 y}} = \ln K$$

Hence, the general solution of Eqns.(21) is

$$\left(\frac{y^{r_1/\alpha_1}}{e^y}\right)^{\alpha_1} \left(\frac{e^x}{x^{r_2/\alpha_2}}\right)^{\alpha_2} = K$$

where the constant K has to be determined using the initial conditions $x(0) = x_0$ and $y(0) = y_0$. Using these conditions we obtain

$$K = \left(\frac{y^{x^*}}{e^{y_0}}\right)\left(\frac{e^{x_0}}{x^{y^*}}\right)^{\alpha_2}$$

APPENDIX

A system of two first order equations of the form

$$\frac{dx}{dt} = F(x,y) \qquad (1)$$
$$\frac{dy}{dt} = G(x,y)$$

is said to be **autonomous** when the independent variable t does not appear explicitly; We assume that the functions F and G are continuously differentiable in some region R in the xy-plane, which is called the **phase plane** for the system (1). Then according to the existence and uniqueness theorem, ***(ref:*** Unit 1 , **Block** 1 of MTE-08) given t_0 and any point (x_0, y_0) of R, there exists a unique solution $x = x(t), y = y(t)$ of (I) that is defined on some open internal $a < t_0 < b$ and satisfies the initial conditions

$$x(t_0) = x_0, y(t_0) = y_0 \qquad (2)$$

The equation $x = x(t), y = y(t)$ then describes a solution *in* the phase plane. Any such solution curve **is** called a **trajectory** or the orbit of the system (1) and precisely one trajectory passes through each point of R.

A **critical point** of the system (1j is a point (x^*, y^*) obtained by setting $\frac{dx}{dt} = 0, \frac{dy}{dt} = 0$ and such that

$$F(x^*, y^*) = G(x^*, y^*) = 0 \qquad (3)$$

Conversely, if (x^*,y^*) is a **critical** point of the system, then the constant valued functions

$$x(t) = x^*, y(t) = y^* \qquad (4)$$

satisfy system (1) and are called **equilibrium** solutions of the system.

Note that the trajectory of the equilibrium solution in Equ.(3) 'consists of the single point (x^*, y^*).

In practical situations, the equilibrium solutions and trajectories are of most interest. For example, suppose that the system $x' = F(x, y), y' = G(x, y)$ models two populations x(t) and y(t) of animals that cohabit the same environment, and compete for the same food or prey on one mother. x(t) might denote the number of foxes and y(t) the number of rabbits present at time t. Then a critical point (x^*,y^*) of the system specifies a constant population x^* of foxes and a constant population y^* of rabbits that can coexist with one another in the environment. If (x^*,y^*) is not; a critical **point** of the system, it is not possible for constant populations of x^* foxes **and** y^* rabbits to coexist; one or both must change with time.

As you know, it is **not always** possible to obtain the analytical solution of the system of the form (1). We thus make the qualitative study of system (1) to learn as much as we can about the system. Let us assume that, the nonlinear system (1) is of the form

$$\frac{dx}{dt} = a_1x + b_1y + f(x,y)$$
$$\frac{dy}{dt} = a_2x + b_2y + g(x,y)$$

or in matrix form as

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} f(x,y) \\ g(x,y) \end{pmatrix} \qquad (5)$$

Elective Paper

MATA 3.4

Block - II

Marks : 50 (SSE : 40; IA : 10)

Dynamical System (Applied Stream)

1 One-Dimensional ODE Dynamics

The first main objective in this section will be to outline the basic effects in nonlinear ordinary differential equations (ODEs). The basic notation for ODEs we are going to use is

$$x' := \frac{\mathrm{d}x}{\mathrm{d}t} = f(x), \qquad x = x(t) \in \mathbb{R}^d, \tag{1.1}$$

where $t \in [0, T] = \mathcal{I}$ for some $T \in (0, +\infty]$, $f : \mathbb{R}^d \to \mathbb{R}^d$ is a sufficiently differentiable map, and the initial condition $x_0 = x(0)$ is given. In this context, $\mathbb{R}^d$ is called the **phase space**, $x : \mathcal{I} \to \mathbb{R}^d$ is a **trajectory** (or an**orbit**), and we shall also refer to f as the **vector field**; see Figure TODO. For this section we shall take $d = 1$.

Example 1.1. Consider the **linear** population-growth model

$$x' = x = f(x), \tag{1.2}$$

which is linear since the vector field $f : \mathbb{R} \to \mathbb{R}$ is a linear map, and where we could interpret $x = x(t)$ as the size of a population at time t. One can easily solve (1.2) by **separation of variables**, i.e., by (formally) re-writing the original equation as

$$\frac{1}{x}\,\mathrm{d}x = \mathrm{d}t,$$

re-labelling the variables and then integrating both sides, which yields

$$\int_{x_0}^{x(t)} \frac{1}{y}\,\mathrm{d}y = \int_0^t 1\,\mathrm{d}s.$$

This just implies $\ln x(t) - \ln x_0 = t$ so that taking exponentials gives

$$x(t) = x_0 \mathrm{e}^t. \tag{1.3}$$

From this explicit formula we can easily plot a one-dimensional **phase portrait** as shown in Figure TODO. Hence, any starting solution with $x_0 \neq 0$ grows exponentially. The point $x^* = 0$ is an **equilibrium point**, sometimes also called **steady state** or **stationary point**. ♦

The concept of equilibrium points is evidently more general as the next definition shows.

Definition 1.2. Consider (1.1), then a point x^* is called an **equilibrium point** if $f(x_*) = 0$.

For general *nonlinear* ODEs, i.e., if f is not a linear map, there are no general explicit solution techniques available. Mathematically, this may look unfortunate. But it is actually very fortunate from the viewpoint of real-life considerations since a purely linear world would be really boring.

Example 1.3. Indeed, the population growth model (1.2) is not very realistic as it assumes growth is just proportional to the current population size. It is more natural to assume resource-limitations kick in for very large populations. So we consider our first **nonlinear** ODE

$$x' = rx\left(1 - \frac{x}{p}\right), \tag{1.4}$$

where $r, p > 0$ are parameters interpreted as *growth rate* and *carrying capacity* respectively. The equation (1.4) is also known as the **logistic equation** or as the **Verhulst model**. We can actually eliminate one parameter in (1.4) by a **time re-scaling**

$$t = \tilde{t}/r \quad \Rightarrow \quad \frac{\mathrm{d}x}{\mathrm{d}\tilde{t}} = \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}\tilde{t}} = \frac{1}{r}x',$$

so that we obtain that

$$\frac{\mathrm{d}x}{\mathrm{d}\tilde{t}} = x\left(1 - \frac{x}{p}\right) = f(x), \tag{1.5}$$

where we can just drop the tilde to obtain an ODE in the usual notation. The process we just went through was an example of the more general technique known as **non-dimensionalization**, which means scaling variables such as t and/or x, to get remove parameters; it is actually possible to use a scaling of x to also remove p (exercise!). Our next goal is to think *geometrically* and analyze (1.5) *without solving it*. Equilibria are found by the zeros of f, which gives

$$x^* = 0 \qquad \text{or} \qquad x^* = p.$$

We can plot f over the one-dimensional phase space $\mathbb{R}$ as shown in Figure TODO to determine the *direction* of the motion/flow. Hence, we see that $x^* = 0$ is **unstable** and $x^* = p$ is **stable**. ♦

More formally, we can define stability as follows:

Definition 1.4. An equilibrium point x^* of (1.1) is **(locally asymptotically) stable** if there exists a neighbourhood $\mathcal{U}$ of x_* such that all trajectories starting in $\mathcal{U}$ converge to x_* as $t \to +\infty$. If all trajectories leave $\mathcal{U}$, then x^* will be called **unstable**.

For nonlinear systems, it is usually not easy to check **global stability** of an equilibrium, which just means we could extend the neighbourhood $\mathcal{U}$ in Definition 1.4 to the entire phase space.

Example 1.5. (Example 1.3 continued) Clearly $x^* = 0$ is not globally stable, as it is unstable. Our phase portrait in Figure TODO shows that all initial conditions with $x_0 < 0$ escape to $-\infty$, while all initial conditions $x_0 > 0$ satisfy

$$\lim_{t\to\infty} x(t) = p \qquad \text{if } x(0) > 0.$$

So if we view (1.5) purely mathematically on a phase space $\mathcal{X} := \mathbb{R}$, then x^* is not globally stable, while considering a phase space definition $\mathcal{X} := [0, \infty)$, then x^* is globally stable. This illustrates the key interplay between modelling and analysis. ♦

Our geometric argument can also be formalized using an important observation for *one-dimensional* ODEs.

Proposition 1.6. *Consider the logistic equation defined on*

$$x' = x\left(1 - \frac{x}{p}\right), \qquad x_0 > 0. \tag{1.6}$$

Then $x(t) \to p$ as $t \to +\infty$.

Proof. Let us first translate the equilibrium point $x^* = p$ to the origin by $y = x - p$, which yields

$$y' = (y + p)\left(1 - \frac{y + p}{p}\right) = -y\left(\frac{y}{p} + 1\right), \qquad y(0) > -p, \tag{1.7}$$

and moves the equilibrium point we are interested in to $y^* = 0$. Now observe that any *one-dimensional* ODE can be written as a **gradient system**

$$y' = -\nabla V(y) = -\dot{V}(y), \qquad \dot{} = \frac{\mathrm{d}}{\mathrm{d}y},$$

which holds by just finding the anti-derivative of the vector field and changing the sign. Applying this to (1.6) gives

$$V(y) = \int \frac{y^2}{p} + y \, \mathrm{d}y = \frac{1}{2}y^2 + \frac{1}{3p}y^3.$$

V vanishes at the equilibrium $y^* = 0$ since $V(0) = 0$, and $V(y) > 0$ for $y > -p$ follows easily as well. Furthermore, we have using the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}t}V(y) = \dot{V}(y)y' = -(\dot{V}(y))^2 < 0$$

for $y > -p$ and $y \neq 0$, so V is strictly decreasing along all trajectories we are interested in. This allows us to conclude that $y^* = 0$ is globally stable for the transformed system (1.7) so $x^* = p$ is globally stable for (1.6). $\square$

The function V in the last proof is also known as a **potential** in analogy to a physical energy. Furthermore, V is a special case of a (strict) **Lyapunov function** L for an equilibrium $y^* = 0$ for $y' = f(y)$, which has to satisfy the three conditions

(L1) $L(0) = 0$,

(L2) $L(y) > 0$ for $y \neq 0$,

(L3) $\frac{\mathrm{d}L}{\mathrm{d}t}(y) < 0$ for $y \neq 0$.

Arguments based upon energy/physical considerations and Lyapunov functions can sometimes be useful but for general nonlinear systems such arguments fail. Hence, we essentially have to give up the dream to always understand global stability. Let us be more modest and try *local* analysis for an ODE (1.1) near a given steady state x^* and $d = 1$. Let us consider a perturbation

$$x(t) = x^* + \varepsilon X(t), \qquad X(t) \in \mathbb{R},\ \varepsilon > 0,$$

where ε is assumed to be small. Plugging this into (1.1) and using Taylor expansion gives

$$\begin{aligned} x' &= (x^* + \varepsilon X)' = \varepsilon X' \\ &= f(x^* + \varepsilon X) = f(x^*) + \varepsilon f'(x^*)X + \frac{1}{2}\varepsilon^2 f''(x^*)X^2 + \cdots \\ &= \varepsilon f'(x^*)X + \frac{1}{2}\varepsilon^2 f''(x^*)X^2 + \cdots, \end{aligned}$$

where we used that $f(x^*) = 0$ in the last step. Dividing through by ε and then dropping all the terms, which are still multiplied by ε gives

$$X' = f'(x^*)X. \tag{1.8}$$

It is intuitive that the **linearized system** (1.8) is only valid as an approximation to the dynamics in a neighbourhood $\mathcal{U}$ of x^*, and we were only allowed to discard the higher-order Taylor terms if $f'(x^*) \neq 0$.

Example 1.7. (Example 1.5 continued) Let us look at the two equilibrium points again. We just calculate

$$f(x) = x(1 - x/p) \qquad \Rightarrow \quad f'(x) = 1 - 2x/p.$$

So for $x^* = 0$, we find the linearized system

$$X' = f'(0)X = X$$

and we know from the explicit solution in Example 1.1 that solutions are going to diverge away from $X = 0$ so $x^* = 0$ is unstable; see Figure TODO. For $x^* = p$ we get

$$X' = f'(p)X = -X$$

so small perturbations X modelled by $X(0) \neq 0$ are going to decay back towards $X = 0$ so $x^* = p$ is locally asymptotically stable for the logistic equation. ♦

The Taylor series argument above can be extended quite easily using a d-dimensional Taylor series, so we shall give the following definition in quite some generality.

Definition 1.8. Consider an ODE $x' = f(x)$, $f : \mathbb{R}^d \to \mathbb{R}^d$, $d \in \mathbb{N}$ with equilibrium x^*. Then the **linearized system** or **linearization** near x^* is given by

$$X' = [\mathrm{D}f(x^*)]X, \qquad X \in \mathbb{R}^d, \tag{1.9}$$

where $\mathrm{D}f(x^*)$ is the **Jacobian matrix** (or just **Jacobian**) consisting of all partial derivatives evaluated at the equilibrium, i.e., we have

$$\mathrm{D}f(x^*) = \left(\frac{\partial f_i}{\partial x_j}\right)_{i,j} =: \left(\partial_{x_j} f_i\right)_{i,j} \quad i, j \in \{1, 2, \ldots, d\}.$$

2 Bifurcations of One-Dimensional ODEs

We have already seen in the logistic equation from Section 1 that it naturally comes with *parameters*. In fact, essentially all mathematical models are parametrized. Therefore, let us look at one-parameter families of one-dimensional ODEs

$$x' = f(x,p), \qquad x \in \mathbb{R},\ p \in \mathbb{R}, \tag{2.1}$$

where f has again sufficiently many derivatives. Suppose x^* is an equilibrium point at a fixed parameter value p^*. The local dynamics near $(x,p) = (x^*,p^*)$ is approximated using Taylor expansion

$$\begin{aligned} x' = f(x,p) \approx\ & f(x^*,p^*) + \nabla f|_{(x^*,p^*)} \cdot \begin{pmatrix} x-x^* \\ p-p^* \end{pmatrix} \\ & + \frac{1}{2} \begin{pmatrix} x-x^* \\ p-p^* \end{pmatrix}^\top \begin{pmatrix} \partial_{xx} f & \partial_{xp} f \\ \partial_{px} f & \partial_{pp} f \end{pmatrix}\Big|_{(x^*,p^*)} \begin{pmatrix} x-x^* \\ p-p^* \end{pmatrix} + \mathcal{O}(3), \end{aligned}$$

where $\mathcal{O}(3)$ denotes terms at least cubic in $x-x^*$ and $p-p^*$. Writing out the different terms and using $f(x^*,p^*) = 0$ gives

$$x' = \partial_x f(x^*,p^*)(x-x^*) + \partial_p f(x^*,p^*)(p-p^*) + \frac{1}{2}\partial_{xx} f(x^*,p^*)(x-x^*)^2 + \cdots,$$

where we have dropped cubic terms in $(x-x^*)$, mixed quadratic terms and quadratic terms in $(p-p^*)$; to prove that this is indeed rigorously possible under *generic* assumptions on f is a more advanced result. From the calculation we see that it is quite cumbersome to always write $p-p^*$ and $x-x^*$ so it is more natural to set

$$\tilde{x} = x - x^*, \qquad \tilde{p} = p - p^*$$

to translate the point we are interested in to the origin. Dropping the tildes and using the simplified notation $(0,0) =: 0$ gives

$$x' = \partial_x f(0)x + \partial_p f(0)p + \frac{1}{2}\partial_{xx} f(0)x^2 + \cdots. \tag{2.2}$$

If $\partial_x f(0) \neq 0$, then we expect that the dynamics near the equilibrium is governed by the linearized system. This gives a natural definition.

Definition 2.1. An equilibrium point x^* of an ODE $x' = f(x)$, $x \in \mathbb{R}^1$, is called **hyperbolic** if $\partial_x f(x^*) \neq 0$.

How to deal with the hyperbolic case locally, we have already learned in Example 1.1. However, we have now a free parameter p so it seems natural that *generically* we can expect that there exist distinguished isolated parameter values, say wlog $p^* = 0$, such that $\partial_x f(x^*, 0) = 0$. Then we have from (2.2) that

$$x' = \partial_p f(0)p + \frac{1}{2}\partial_{xx} f(0)x^2 + \cdots . \tag{2.3}$$

Example 2.2. The ODE (2.3) motivates us to look at

$$y' = q + y^2. \tag{2.4}$$

The vector field (2.4) can be analyzed graphically as in Section 1. Figure TODO shows the situation for different parameter values near $q = 0$. For $q < 0$, we have two equilibria $y^{\pm} = \pm\sqrt{-q}$. One observes from the signs of the vector field or checks from the linearization that y^+ is unstable and y^- is unstable. At $q = 0$, the two equilibria coalesce into a non-hyperbolic equilibrium, and for $p > 0$ there are no equilibria. ♦

Although we have not quite proven the next result fully, the main ideas are clear from our previous discussion.

Theorem 2.3 (**fold bifurcation**). *Consider an ODE* (2.1) *and let* $0 = (0, 0)$. *Assume that the following conditions hold*

(A1) $f(0) = 0$, $\partial_x f(0) = 0$;

(A2) $\partial_{xx} f(0) \neq 0$, $\partial_p f(0) \neq 0$.

Then there exists ***(generic) fold bifurcation****, i.e., there is a transition from two to zero equilibria locally upon varying* p *near* 0*; see also Figure TODO.*

Remark: To be precise, the conclusion of the last theorem just means that there exists a homeomorphism $h : (-p_0, p_0) \to (-q_0, q_0)$ on parameter space such that the general vector field (2.1) and (2.4) at $h(p) = q$ are **topologically equivalent**, which just means that the phase portraits are homeomorphic (preserving the direction of time); see Figure TODO.

Definition 2.4. The ODE $y' = q \pm y^2$ is called the **normal form** of the fold bifurcation.

We shall encounter many different application models throughout this course, which exhibit fold bifurcations. Although the fold bifurcation is the one we expect most frequently, let us use an application-inspired view to gain insight which other types of bifurcation we could expect for one-dimensional ODEs.

Example 2.5. (Example 1.7 continued) Recall our logistic equation $x' = x(1 - x/p)$. No matter, how we select the carrying capacity p of the population density x, there is always the equilibrium $x^* = 0$. This makes perfect sense from macroscopic population biology. In many other physical situations, similar modelling assumptions have to be made, e.g., chemical reactions are completely stationary if there are no reactants. ♦

So let us assume that we are now *restricting* to vector fields, which always have the **trivial branch** of equilibria $x^* = 0$ for any parameter value p

$$x' = f(x, p) = xg(x, p), \qquad x \in \mathbb{R},\ p \in \mathbb{R}, \tag{2.5}$$

where g is a sufficiently smooth function. We shall now derive another interesting example in the form

Example 2.6. Suppose we want to model an infectious disease, e.g. think of the flu. Consider a population with two types: **susceptibles** with density y and **infected** with density x. Suppose when infected and susceptibles meet, the disease is transmitted at rate $\alpha > 0$ and infected recover at rate $r > 0$, then we get the **susceptible-infected-susceptible (SIS) model**

$$\begin{aligned} y' &= -\alpha xy + rx, \\ x' &= \alpha xy - rx. \end{aligned}$$

If the population is constant and normalized to say $x+y = 1$, then it suffices to just look at one of the equations, say the infected density x, which yields

$$x' = \alpha x(1 - x) - rx = x(\alpha - r - \alpha x),$$

which almost has the form we considered in (2.5). Using a non-dimensionalization we can just scale time by $1/\alpha$ and set $p := r/\alpha$ to get

$$x' = x(1 - p - x). \tag{2.6}$$

The model (2.6) can now be analyzed. There are two equilibria, $x^* = 0$ and $x^* = 1 - p$; note that $x^* = 1 - p$ is not biologically relevant for $p > 1$; see Figure TODO. We check that the linearized problem is

$$X' = f'(x^*)X = (1 - p - 2x^*)X.$$

Therefore, $x^* = 0$ is locally stable for $p > 1$ but unstable for $p < 1$, which makes sense as the latter case corresponds to high infection rate relative to the recovery rate; in epidemiology the threshold $p = 1$ is also referred to as R_0, which is the parameter value at which there are more secondary

infections than recoveries. The equilibrium $x^* = 1 - p$ has the associated linearized system

$$X' = f'(x^*)X = (1 - p - 2(1 - p))X = (p - 1)X,$$

so x^* is locally stable for $0 < p < 1$ and unstable for $p > 1$. This analysis gives us the **bifurcation diagram** shown in Figure TODO. ⧫

In general, bifurcation diagrams just depict parameter values together with a representative phase portrait for each parameter value. The diagram in Figure TODO is an example of a **transcritical bifurcation**, which has normal form

$$y' = y(q - y). \tag{2.7}$$

Theorem 2.7 (**transcritical bifurcation**)**.** *Consider an ODE* (2.5) *and let* $0 = (0, 0)$*. Assume that the following conditions hold*

(A1) $g(0) = 0$*;*

(A2) $\partial_x g(0) \neq 0$*,* $\partial_p g(0) \neq 0$*.*

Then there exists ***(generic) transcritical bifurcation****, i.e., there is an* ***exchange-of-stability*** *between two equilibria locally upon varying* p *near* 0*; see also Figure TODO.*

As before, the last theorem means that the dynamics is equivalent to the normal form (2.7) using a homeomorphism of parameter space and then one for phase space at each parameter. The next example shows that in addition to trivial branches, additional modelling considerations may play a role.

Example 2.8. Consider a beam with a load as sketched in Figure TODO. If we increase the load by varying a parameter p, then it is intuitive from mechanics that eventually the beam is going to bend suddenly; see Figure TODO. This is known as **Euler buckling**. Note that there is **symmetry** in the problem, i.e., if we look at Figure TODO, there is no preference for buckling left or right. A simple toy model for this mechanical system is

$$x' = px + x^3 - x^5 = f(x, p), \qquad x \in \mathbb{R},\ p \in \mathbb{R}. \tag{2.8}$$

Note that the vector field f respects the **reflection symmetry** $x \mapsto -x$ as it remains invariant under this transformation since

$$f(\gamma x, p) = \gamma f(x, p) \qquad \text{for } \gamma = -1.$$

The last equation can be interpreted more abstractly by saying that there is a group $\mathbb{Z}_2$ formed by the elements 1 and -1 under multiplication under which the vector field is **equivariant**

$$f(\gamma x, p) = \gamma f(x, p) \qquad \forall \gamma \in \mathbb{Z}_2.$$

Of course, the more classical presentation of $\mathbb{Z}_2$ via addition modulo 2 and elements $0, 1$ is isomorphic to the group used here. The ODE (2.8) can be analyzed as before and we obtain the bifurcation diagram in Figure TODO. ♦

If we were only interested of the bifurcation in Example 2.8, we can drop the fifth-order term and obtain the normal form

$$y' = y(q + y^2) \tag{2.9}$$

of a **subcritical pitchfork bifurcation**. More generally, we have the following theorem:

Theorem 2.9 (**pitchfork bifurcation**)**.** *Consider an ODE* (2.1) *and let* $0 = (0, 0)$*. Assume that the following conditions hold in a neighbourhood of* 0*, i.e., locally*

(A1) $f(-x, p) = -f(x, p)$*,* $\partial_x f(0) = 0$*;*

(A2) $\partial_{xx} f(0) = 0$*,* $\partial_{xxx} f(0) = \beta \neq 0$*,* $\partial_p f(0) = 0$*,* $\partial_{xp} f(0) \neq 0$*.*

Then there exists ***(generic) pitchfork bifurcation****; see Figure TODO. The pitchfork is* ***subcritical*** *if for the* ***normal form coefficient*** *we have* $\beta > 0$ *and it is* ***supercritical*** *if* $\beta < 0$*.*

It is a good exercise to try to re-write the conditions (A1)-(A2) of Theorem 2.9 for $x' = f(x, p)$ in the format used in Theorem 2.7, we we use the formulation $x' = xg(x, p)$, and also do the exercise vice versa.

3 Stationary Two-Dimensional ODE Dynamics

We start with a motivating example and a little summary of linear ODEs.

Example 3.1. Consider a spring with spring constant $k > 0$ and with a mass $m > 0$ suspended at one end; see Figure TODO. Then Newton's Law gives for the position y of the mass

$$m\frac{\mathrm{d}^2 y}{\mathrm{d}t^2} = my'' = -ky, \qquad \text{with } y(0), y'(0) \text{ given}, \tag{3.1}$$

which is the classical **second-order ODE** known as the **harmonic oscillator**. Although (3.1) looks like a one-dimensional ODE at first, it is really two-dimensional as we can reduce it via the trick

$$x_1 := y, \qquad x_2 := \frac{\mathrm{d}x_1}{\mathrm{d}t}$$

to a coupled system of two *linear* **first-order ODEs**

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= -\tfrac{k}{m}x_1, \end{aligned} \tag{3.2}$$

with initial conditions implicitly understood from (3.1). Using a nondimensionalization $x = \tilde{x}_1\sqrt{m}$, $x_2 = \tilde{x}_2\sqrt{k}$, $t = \tilde{t}\sqrt{m/k}$, dropping the tildes after the scaling, and re-writing the system (3.2) in matrix form leads to

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}}_{=:A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad \text{or} \qquad x' = Ax. \tag{3.3}$$

There are several views on finding the exact solution of the harmonic oscillator. For example, one could make the ansatz $y(t) = \mathrm{e}^{\lambda t}$ for (3.1), or one could use a matrix exponential technique $x(t) = \mathrm{e}^{tA}x(0)$, or one could just guess that

$$x_1(t) = x_2(0)\sin t + x_1(0)\cos t, \quad x_2(t) = -x_1(0)\sin t + x_2(0)\cos t \tag{3.4}$$

solves (3.3) and then verify the solution. The phase portrait is shown in Figure TODO. The equilibrium point $x^* = 0$ is not locally asymptotically stable, neither is it unstable. It is called a **center**. ♦

Centers are one reason for another definition of local stability. This definition just means that starting near x_*, we stay near it.

Definition 3.2. An equilibrium point x^* of (1.1) is **Lyapunov stable** if given any $\varepsilon > 0$ there exists $\delta > 0$ such that if $|x_* - x(0)| < \delta$ then $|x(t) - x_*| < \varepsilon$ for all $t > 0$.

In the last definition we used the simpler notation $|\cdot| = \|\cdot\|_2$ for the **Euclidean norm**, which is a convention employed from now on. Next, observe that a general linear system

$$x' = Ax, \qquad x(0) = x_0,\ A \in \mathbb{R}^{d\times d}, \tag{3.5}$$

can be solved using the **matrix exponential** so that we get

$$x(t) = \mathrm{e}^{tA}x_0 = \left(\sum_{j=0}^{\infty} \frac{(tA)^j}{j!}\right) x_0. \tag{3.6}$$

In practice, this means transforming A to a simpler form,, e.g., **Jordan canonical form** B, via an invertible linear transformation $M \in \mathbb{R}^{d\times d}$ so that $A = M^{-1}BM$. This means we have to compute

$$x(t) = \left(\sum_{j=0}^{\infty} \frac{(tM^{-1}BM)^j}{j!}\right) x_0,$$

where we can use the key observation $(M^{-1}BM)^j = M^{-1}B^jM$ as well as the fact that it is easy to compute the matrix exponential of a matrix in Jordan canonical form (exercise!).

Example 3.3. (Example 3.1 continued) Although the center dynamics in Figure TODO is mathematically correct, is obviously not what we see in many experiments; see also Figure TODO. In the spring experiment, there will be *friction*, which should induce *damping* of the oscillations. This is modeled by the ODE

$$my'' = -ky - cy', \qquad \text{with } y(0), y'(0) \text{ given}, \tag{3.7}$$

for a viscous damping coefficient $c > 0$. Following a similar procedure as in Example 3.1 (exercise!), we get the linear ODE system

$$x' = \begin{pmatrix} 0 & 1 \\ -1 & -p \end{pmatrix} x, \tag{3.8}$$

where $p > 0$ indicates (positive) damping and $p < 0$ negative damping. Indeed, the eigenvalues of A should give us information about decay or growth of solutions via formula (3.6) so we compute

$$\det(A - \lambda\mathrm{Id}) \stackrel{!}{=} 0 \qquad \Rightarrow \quad \lambda = \lambda_\pm = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - 1}.$$

From this we easily conclude from (3.6) that for $p > 0$ we have

$$|x(t)| = \sqrt{x_1(t)^2 + x_2(t)^2} \to 0 \qquad \text{for } t \to +\infty$$

as sketched in Figure TODO. For $p \in (0, 2)$, the situation is called a **spiral sink**. For $p \in (-2, 0)$, we get a **spiral source** as sketched in Figure TODO. ♦

We have seen in Example 3.1 that solutions of linear systems with purely complex eigenvalues neither decay nor grow, while Example 3.3 indicated that if the real parts of eigenvalues are non-zero, then we get growing and decaying directions for linear systems. This motivates the following definition:

Definition 3.4. Consider an ODE $x' = f(x)$ with equilibrium point x^* and linearized system $X' = AX$, where $A = \mathrm{D}f(x^*)$ is the Jacobian. If all eigenvalues λ of A satisfy $\mathrm{Re}(\lambda) \neq 0$, then we say x^* is a **hyperbolic equilibrium point**.

For linear planar systems, there exists a very convenient general classification of equilibrium points.

Theorem 3.5 (**trace-determinant plane**)**.** *Consider a planar linear system*

$$X' = AX, \qquad \textit{for } A \in \mathbb{R}^{2\times 2}. \tag{3.9}$$

The stability of the equilibrium point $X = 0$ can be classified just using the trace $\mathrm{tr}(A)$ *and determinant* $\det(A)$ *as shown in Figure TODO.*

Proof. Let λ denote an eigenvalue of A. We simplify the notation by writing

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \Rightarrow \quad 0 \overset{!}{=} \det\left(A - \lambda \mathrm{Id}\right) = (a-\lambda)(d-\lambda) - bc.$$

Simplifying the characteristic equation for the eigenvalues gives

$$0 = \lambda^2 - \underbrace{(a+d)}_{=\mathrm{tr}(A)=:\tau}\lambda + \underbrace{ad-bc}_{=\det(A)=:\Delta} = \lambda^2 - \tau\lambda + \Delta.$$

Therefore, the eigenvalues of A are

$$\lambda_\pm = \frac{\tau \pm \sqrt{\tau^2 - 4\Delta}}{2}.$$

If $\lambda_+ \neq \lambda_-$, and the initial condition is written as $x(0) = c_+ v_+ + c_- v_-$ for eigenvectors $v_\pm$, i.e., $Av_\pm = \lambda_\pm v_\pm$ then it is easy to check from (3.6) that the general solution of the linear system (3.9) can be written as

$$x(t) = c_+ \mathrm{e}^{\lambda_+ t} v_+ + c_- \mathrm{e}^{\lambda_- t} v_-.$$

Using this formula, we can verify most of Figure TODO easily, e.g., if $\lambda_\pm$ are real then we have a **saddle** for $\det(A) < 0$. If $\det(A) > 0$, we get a **stable node** for $\tau^2 - 4\Delta > 0$ and $\tau < 0$, while for $\tau^2 - 4\Delta > 0$ and $\tau > 0$ we get an unstable **unstable node**. Similarly, we can analyze spirals for $\tau^2 - 4\Delta < 0$. With a bit more work using the matrix exponential, we can analyze several special transition regions, e.g., $\tau = 0$ and $\Delta > 0$ corresponds to a **center** as in Example 3.1. □

We can also write the definition of the types of equilibria in a different form for a general planar system:

Definition 3.6. Consider an ODE $x' = f(x)$, $x \in \mathbb{R}^2$, $f \in C^1$, and suppose $f(x_*) = 0$. Let $\lambda_{1,2}$ be eigenvalues of the Jacobian $\mathrm{D}_x f(x_*)$. Then x_* is

(D1) a **saddle** if $\lambda_{1,2} \in \mathbb{R}$ and $\lambda_1 < 0 < \lambda_2$,

(D2) a **stable** (respectively **unstable**) **node** if $\lambda_{1,2} \in \mathbb{R}$ and $\lambda_{1,2} < 0$ (respectively $\lambda_{1,2} > 0$),

(D3) a **stable** (respectively **unstable**) **spiral** if $\lambda_{1,2} \in \mathbb{C}\backslash\mathbb{R}$ and $\mathrm{Re}(\lambda_{1,2}) < 0$ (respectively $\mathrm{Re}(\lambda_{1,2}) > 0$),

(D4) a **center** if $\lambda_{1,2} \in \mathbb{C} \setminus \mathbb{R}$ and $\mathrm{Re}(\lambda_{1,2}) = 0$.

There is also a nice geometry associated to equilibrium points. Let us just illustrate this concept for a simple saddle point in a linear system

$$X' = AX, \qquad X \in \mathbb{R}^2,\ A \in \mathbb{R}^{2\times 2} \tag{3.10}$$

where A has eigenvalues $\lambda_- < 0 < \lambda_+$ and associated eigenvectors v_- and v_+. Then we have two natural linear spaces

$$\begin{aligned} E^{\mathrm{s}}(0) &:= \mathrm{span}(v_-), \\ E^{\mathrm{u}}(0) &:= \mathrm{span}(v_+), \end{aligned}$$

called the **stable** and **unstable eigenspaces**; see Figure TODO. In these spaces the solution of the linear system (3.10) converges in forward time respectively backward time to the equilibrium zero. There is a natural generalization for nonlinear systems

$$x' = f(x), \qquad x \in \mathbb{R}^2. \tag{3.11}$$

as shown in Figure TODO for a saddle in a nonlinear system.

Definition 3.7. Let x_* be an equilibrium point of (3.11), let $x(t)$ be a solution of (3.11) starting at $x(0) = x_0$, and define

$$W^{\mathrm{s}}(x^*) := \{x_0 \in \mathbb{R}^2 : x(0) = x_0,\ x(t) \to x^* \text{ as } t \to +\infty\},$$
$$W^{\mathrm{u}}(x^*) := \{x_0 \in \mathbb{R}^2 : x(0) = x_0,\ x(t) \to x^* \text{ as } t \to -\infty\},$$

which are called the **stable** and **unstable manifolds**.

Here we shall not deal with the existence of these objects or general definition of manifolds; see [4]. Instead, it suffices to think of stable and unstable manifolds as points, curves, or surfaces as sketched in Figure TODO. Having classified equilibria locally, we would like to use our results to a concrete application.

Example 3.8. Consider an autocatalytic reaction given by four chemicals

$$P \to Y, \qquad Y \to X, \qquad Y + 2X \to 3X, \qquad X \to Z,$$

where X, Y are the main reactants, P is a 'pool'-chemical and Z is a product. Standard mass-action kinetics and non-dimensionalization leads to the system of ODEs

$$\begin{aligned} x' &= yx^2 + y - x =: f_1(x, y), \\ y' &= \varepsilon(p - yx^2 - y) =: f_2(x, y), \end{aligned} \tag{3.12}$$

where $f := (f_1, f_2)^\top$, $p > 0$ is a parameter, and $\varepsilon > 0$ is small parameter arising from widely differing reaction rates. Now we can simply use the previous methods we developed to analyze the equilibrium points (or chemically/physically: **steady states**) of (3.12). For example, suppose for simplicity $\varepsilon = 0$, then y is fixed and can be viewed as a parameter so we are left with a one-dimensional ODE

$$x' = yx^2 + y - x.$$

Equilibrium points of this ODE satisfy

$$0 = yx^2 + y - x \qquad \Rightarrow \quad y = \frac{x}{1 + x^2}.$$

Therefore, plotting the equilibria and/or checking the conditions (A1)-(A2) from Theorem 2.3 yields a fold bifurcation. Going back to the full two-dimensional model (3.12), equilibria (x_*, y_*) satisfy

$$y_* = \frac{x_*}{1 + x_*^2}, \qquad 0 = p(1 + x_*^2) - x_*^3 - x_*.$$

Having solved the last system and computing the Jacobian $A = \mathrm{D}f(x_*, y_*)$ one can then use Theorem 3.5 to classify the hyperbolic equilibrium points. However, when hyperbolicity of equilibrium points is *lost*, we have to look carefully for bifurcations. ♦

The last example shows that for a general nonlinear system the local stability computations for equilibria might be difficult using pen-and-paper, yet it is necessary to practice them for simple systems. Furthermore, it is good to think (exercise!), how you would implement these operations on a computer.

4 Periodic Two-Dimensional ODE Dynamics

In one-dimensional ODEs defined by sufficiently smooth vector fields, non-trivial periodic trajectories cannot exist, which follows from standard uniqueness theory of ODEs; see Figure TODO.

Example 4.1. (Example 3.1 continued) We have seen that the solutions (3.4) of the harmonic oscillator are linear combinations of $\sin t$ and $\cos t$. In particular, they are periodic with period 2π. So non-trivial periodic trajectories can exist already for flows in $\mathbb{R}^2$. ♦

Definition 4.2. A trajectory $\{x(t)\}_{t\in\mathcal{I}}$ defined for some time interval $\mathcal{I}$ is called a **periodic orbit** if there exists a time $T > 0$ such that

$$x(t) = x(t+T) \qquad \forall t \in \mathcal{I}.$$

The minimal such T is also called the **period**.

Remark: The last definition excludes "trivial" periodic orbits with period zero, e.g., we do not regard an equilibrium point as a periodic orbit.

Periodic orbits are already quite difficult to deal with since - in contrast to equilibrium points - it is often highly non-trivial to prove that periodic orbits exist, or to exclude their existence. For some special systems, the latter task is possible.

Theorem 4.3 (no periodic orbits in gradient systems)**.** *Consider a* ***gradient system***

$$x' = -\nabla V(x) \qquad V : \mathbb{R}^d \to \mathbb{R} \tag{4.1}$$

for $V \in C^1(\mathbb{R}^d, \mathbb{R})$. Then (4.1) *has no periodic orbits.*

Proof. Note that if ∇V is constant, the result is trivial. So we assume that $|\nabla V(x(t))|$ for some t along the periodic orbit. We argue by contradiction and suppose $x(t)$ is a periodic solution with period $T > 0$. Clearly, we have $V(x(0)) = V(x(T))$ by periodicity. So V is constant along x. Furthermore, a direct calculation yields

$$\begin{aligned} V(x(T)) - V(x(0)) &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}t} V(x(t))\ \mathrm{d}t = \int_0^T (\nabla V(x(t)))^\top x'(t)\ \mathrm{d}t \\ &= -\int_0^T |\nabla V(x(t))|^2\ \mathrm{d}t < 0, \end{aligned}$$

which yields the required contradiction. □

The next theorem gives another non-existence criterion, which is nicely based upon one of the most important calculus results.

Theorem 4.4 (**Dulac's criterion**). *Consider* $x' = f(x)$ *with* $f \in C^(\Omega, \mathbb{R}^2)$, *where* Ω *is simply connected (i.e., simply connected means all closed loops are continuously deformable to a point in* Ω*). If there exists* $g \in C^1(\Omega, \mathbb{R})$ *such that*

$$\nabla \cdot (gx') \text{ has one sign on } \Omega, \tag{4.2}$$

then Ω *contains no periodic orbit.*

Proof. Again we argue by contradiction and suppose $\gamma = \gamma(s)$ be periodic say parametrized by **arclength** s; see Figure TODO. We recognize the condition (4.2) as being related to contraction or expansion as $\nabla\cdot$ = div is the classical **divergence**. Denote the area enclosed by γ by Γ and the outer unit normal to γ by $\vec{n}$. Then **Green's Theorem** (or more generally a special case of **Stokes' Theorem**) yields that

$$\int_\Gamma \nabla \cdot (gx') \, \mathrm{d}A = \int_\gamma g \underbrace{x' \cdot \vec{n}}_{=0} \, \mathrm{d}s = 0.$$

However, the left-hand side is either positive or negative by assumption and we have obtained a contradiction. □

Example 4.5. Glycolysis describes the conversion of sugars into energy in eukaryotes. It is a fundamental process in systems biology. One very simple model for the interaction between important molecules in the process is given by

$$\begin{aligned} x' &= -x + py + x^2 y, \\ y' &= \tfrac{1}{2} - py - x^2 y, \end{aligned} \tag{4.3}$$

where $p > 0$ is a parameter. We restrict phase space to $(x, y) \in [0, \infty) \times [0, \infty)$ One easily checks that there is a unique equilibrium point

$$(x^*, y^*) = \left(\frac{1}{2}, \frac{2}{1 + 4p} \right). \tag{4.4}$$

Furthermore, one calculates by linearization that this equilibrium is unstable for p sufficiently small $p > 0$. So how does the dynamics look like as $t \to +\infty$? The phase portrait in Figure TODO suggests that we might be able to find a periodic orbit in the positive quadrant in this case. ♦

Theorem 4.6 (**Poincaré-Bendixson Theorem**). *Consider a* planar *ODE* $x' = f(x)$ *with* $f \in C^1(\mathbb{R}^2, \mathbb{R}^2)$ *and let* Ω *be a compact set containing no equilibrium points. If* Ω *contains a trajectory* $\gamma = \gamma(t)$ *for all* $t \geq t_0$, *then* Ω *contains a periodic orbit.*

Proof. Let $\gamma(0) = \gamma_0$ be the starting point of the trajectory *trapped* inside Ω. Since $\{\gamma(t) : t \geq 0\}$ lies inside a compact set, the following set must be non-empty

$$\omega(\gamma_0) := \{x_\infty \in \mathbb{R}^2 : \lim_{n\to\infty} x(t_n) = x_\infty, x(t_0) = \gamma_0, t_n > t_{n-1}, t_n \to \infty\} \subset \Omega,$$

which is also called the **ω-limit set** of the point γ_0. We aim to show that any point $Y \in \omega(\gamma_0)$ must lie on a periodic orbit. One may check that $\omega(Y) \subseteq \omega(\gamma_0)$ and that $\omega(Y)$ is non-empty. Wlog we can assume up to a translation of coordinates that $(0,0) \in \omega(Y)$. Let $\mathcal{V}$ be a small neighbourhood around 0. Since $(0,0)$ cannot be an equilibrium point by assumption, we may apply the **flow box theorem** (or **rectification theorem**) to put the flow of the ODE inside $\mathcal{V}$ into the form

$$\begin{aligned} x_1' &= 1, \\ x_2' &= 0. \end{aligned} \tag{4.5}$$

Let us consider the local **section** transverse to the flow given by part of x_2-axis

$$\Sigma := \{x \in \mathbb{R}^2 : x_1 = 0, x_2 \in \mathcal{J}\}$$

for a sufficiently small open interval $\mathcal{J}$ containing zero. Since $(0,0)$ is in $\omega(Y)$, there exists a trajectory $x(t)$ starting from Y such that $x(t_n) \to (0,0)^\top$ as $t_n \to +\infty$; see Figure TODO. Since there are infinitely many $x(t_n)$ inside $\mathcal{V}$ by this argument, the flow-box structure (4.5) implies that we can find times $t_\mathrm{a} > t_\mathrm{b}$ such that $x(t_\mathrm{a}), x(t_\mathrm{b}) \in \Sigma$. The next claim is key: we also aim to show that the solution starting from Y can intersect Σ only *once*, i.e.,

$$\{y(t) : t \geq 0, y(0) = Y, y \text{ a solution}\} \cap \Sigma = \{\text{one point}\}. \tag{4.6}$$

If (4.6) holds, then we can conclude $x(t_\mathrm{a}) = x(t_\mathrm{b})$ so

$$x(t_\mathrm{a} - t_\mathrm{b}) = x(0), \qquad t_\mathrm{a} > t_\mathrm{b}$$

and since there are no equilibrium points, the result follows. The claim (4.6) crucially uses the *planar geometry*. Indeed, suppose there are two intersection points on the local section, say $Y_- = (0, y_-)$ and $Y_+ = (0, y_+)$ in Σ, for a trajectory starting at Y; see Figure TODO. We can assume wlog $y_- < 0$ and $y_+ > 0$. Since $Y_\pm \in \omega(\gamma_0)$ there are infinitely many returns to two sections

$$\Sigma_+ := \Sigma \cap \{x_2 > 0\} \quad \text{and} \quad \Sigma_- := \Sigma \cap \{x_2 < 0\}$$

of the orbit starting at Y. The intersection points $Y_1, Y_2, Y_3, \ldots$ are ordered monotonically along the trajectory. However, the intersections are

not monotonically ordered along Σ as Σ_+ and Σ_- are disjoint; see Figure TODO. This turns out to be a contradiction; the argument is sketched in Figure TODO. Let's look at two points Y_1, Y_2 sketched in Figure TODO. Let $\mathcal{L} \subset \Sigma$ be the line segment between them, denote by Γ the part of the trajectory connecting Y_1 to Y_2, and denote by $\mathcal{D}$ the set enclosed by Γ and $\mathcal{L}$ as shown in Figure TODO. Now we have two sets

$$\mathcal{D} \qquad \text{and} \qquad \mathcal{D}^{\mathrm{c}} := \mathbb{R}^2 \setminus \mathcal{D}.$$

Since we are in the plane, the **Jordan Curve Theorem**, says the two sets partition the plane, i.e., any continuous curve starting in $\mathcal{D}$ and ending in $\mathcal{D}^{\mathrm{c}}$ must cross the boundary $\partial\mathcal{D}$. Therefore, $\mathcal{D}^{\mathrm{c}} = \mathbb{R}^2 \setminus \mathcal{D}$ is **positively invariant** in time, i.e., solutions starting in this set cannot leave it and re-enter $\mathcal{D}$. Indeed, it is impossible to enter via $\mathcal{L}$ considering the direction of the vector field (4.5) while entering through Γ is not possible due to local uniqueness of solutions. Hence, there cannot exist a point $Y_3 \in \mathcal{L}$ between Y_1 and Y_2, which occurs later along the trajectory. Hence, the intersection points $Y_1, Y_2, Y_3, \ldots$ must also be ordered along the section Σ, which is the required contradiction. □

The main problem with applying the Poincaré-Bendixson Theorem is to find one, or more, trajectories, which are trapped.

Example 4.7. (Example 4.5 continued) From the numerical simulations shown in Figure TODO, we expect that it is possible to define a region in the positive quadrant within which all trajectories are trapped for $t > 0$. We consider the points

$$P_0 = (0, 0), \quad P_1 = \left(0, \frac{1}{2p}\right), \quad P_2 = \left(\frac{1}{2}, \frac{1}{2p}\right).$$

Let L_1 be a line through P_2 with slope -1 and let

$$N_x := \{(x, y) \in \mathbb{R}^2 : x' = 0\} = \left\{y = \frac{x}{p + x^2}\right\}$$

be the **nullcline** of the x-equation, i.e., on this curve, there is only movement in the y-direction. Then define the points

$$P_3 := N_x \cap L_1, \qquad P_4 := L_2 \cap \{y = 0\},$$

where L_2 is vertical line through P_3 as shown in Figure TODO. Then let $\mathcal{U}$ be the **convex hull** of P_0 to P_4, i.e., the smallest convex set containing all five points as shown in Figure TODO. One can actually check that the

vector field is pointing inside $\mathcal{U}$ on $\partial\mathcal{U}$ so that trajectories are trapped inside $\mathcal{U}$ if they start there. For example, we have on the axes

$$x'|_{\{x=0,y>0\}} = -0 + py + 0^2 y > 0,$$

so x is pointing right between P_0 and P_1, while

$$y'|_{\{y=0,x>0\}} = \frac{1}{2} - p \cdot 0 - x^2 \cdot 0 = \frac{1}{2} > 0,$$

so y is pointing upwards between P_0 and P_4. This also shows that the quadrant $\{x \geq 0, y \geq 0\}$ is positively invariant as required from the modelling. We leave the remaining arrows in Figure TODO as an exercise. Now we can almost apply the Poincaré-Bendixson Theorem as we have to exclude the equilibrium point $z^* := (x^*, y*)$ given by (4.4) from our region via

$$\Omega := \mathcal{U} \setminus \mathcal{B}(z^*; \varepsilon), \quad \mathcal{B}(z^*; \varepsilon) = \{w \in \mathbb{R}^2 : |w| < \varepsilon\},$$

by excluding a small ball $\mathcal{B}(z^*; \varepsilon)$ for $\varepsilon > 0$ sufficiently small. If z^* is locally *unstable*, which holds for the parameter p being sufficiently small as discussed in Example 4.5. ♦

5 Bifurcations of Two-Dimensional ODEs

In the last section, we have seen, how difficult it is to deal with *global* trajectories, so we start with equilibrium points again.

Example 5.1. Consider the toy model

$$\begin{aligned} x' &= p - x^2, \\ y' &= -y. \end{aligned} \tag{5.1}$$

For $p > 0$, we have two equilibrium points $Z_\pm = (\pm\sqrt{p}, 0)$, one being a saddle and one being a node; see also Figure TODO. The two points collide at $p = 0$ and there are no equilibrium points in $\mathbb{R}^2$ for $p < 0$. Hence, we have found the bifurcation diagram shown in Figure TODO of a fold bifurcation; this example also explains, why the fold bifurcations is sometimes called **saddle-node bifurcation**. Note that since the equations are *de-coupled* and the x-axis $\{y = 0\}$ is **invariant** for $t \in \mathbb{R}$, we may directly apply Theorem 2.3. ♦

The last example illustrates that the topological changes of the phase portrait we know from the fold, transcritial and pitchfork bifurcation basically carry over as phenomena to higher dimensions. However, directly applying one-dimensional theorems can be more complicated, which is a topic discussed in [4].

Example 5.2. In classical nonlinear electrical circuit theory, which then has key applications to neuroscience, one finds often planar systems in **Liénard form**

$$\begin{aligned} x' &= y - g(x, p), \\ y' &= -x, \end{aligned} \tag{5.2}$$

where we can roughly think of x, y as variables describing voltages or currents in different parts of a circuit. One common example to study is a *cubic nonlinearity*

$$g(x, p) = x^3 - px, \qquad \text{where } p \in \mathbb{R} \text{ is a parameter.}$$

In this case, the only equilibrium point is $(x^*, y^*) = (0, 0) =: 0$. Calculating the linearization at the origin gives

$$X' = \underbrace{\begin{pmatrix} p & 1 \\ -1 & 0 \end{pmatrix}}_{=:A} X.$$

The eigenvalues of A are

$$\lambda_\pm = \frac{1}{2}\left(p \pm \sqrt{p^2 - 4}\right).$$

Note carefully that for $-2 < p < 0$, there is a *complex conjugate pair* of eigenvalues associated to a spiral sink. For $p = 0$, the eigenvalues are on the imaginary axis in the complex plane $\mathbb{C}$ and the equilibrium at the origin is *not hyperbolic*. For $0 < p < 2$, we get a spiral source. A phase potrait plotted for small positive p actually shows a *periodic orbit*, so we have actually observed an example of **Hopf bifurcation**. ♦

What should a normal form of the Hopf bifurcation look like? Guessing the simplest polynomial vector field seems difficult at first, until we remember *polar coordinates*. Indeed, let us consider for $r \in [0, \infty)$ and $\varphi \in \mathbb{S}^1 = \mathbb{R}/\mathbb{Z}$ the vector field

$$r' = r(q + l_1 r^2), \tag{5.3}$$
$$\varphi' = 1, \tag{5.4}$$

where q is our main **bifurcation parameter** and $l_1 \neq 0$ is an important auxiliary coefficient, to be described below. Note that $r = 0$ can be viewed as an equilibrium point as it corresponds to $(y_1, y_2) = (0, 0)$ for all times as

$$y_1 = r\cos\varphi \qquad \text{and} \qquad y_2 = r\sin\varphi. \tag{5.5}$$

Linearization of (5.3) around zero easily gives that $r = 0$ is locally asymptotically stable for $q < 0$, it is non-hyperbolic for $q = 0$, and it is unstable for $q > 0$. The other equilibrium of (5.3) is

$$r^* = \sqrt{-\frac{q}{l_1}} \tag{5.6}$$

and correspond to *periodic orbits*. In particular, we get the two bifurcation diagrams in Figure TODO.

Remark: The formula (5.6) implies that the amplitude of the periodic orbits locally grows like the square-root in the distance to the bifurcation in parameter space.

Theorem 5.3 (**Hopf bifurcation**). *Any smooth* generic *two-dimensional, one-parameter ODE*

$$x' = f(x, p), \qquad x \in \mathbb{R}^2,\ p \in \mathbb{R}, \tag{5.7}$$

satisfying $f(0, 0) = 0$ *and such that* $\mathrm{D}_x f(0, p)$ *has a complex conjugate pair of eigenvalues*

$$\lambda_{1,2}(p) = \mu(p) \pm \mathrm{i}\omega(p)$$

with $\mu(0) = 0$, $\mu'(0) \neq 0$, $\omega(0) > 0$, *has the normal form*

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} q & -1 \\ 1 & q \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + l_1(y_1^2 + y_2^2) \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \tag{5.8}$$

The bifurcation is ***supercritical*** *if* $l_1 < 0$ *and* ***subcritical*** *if* $l_1 > 0$ *as in Figure TODO.*

We cannot give a general proof here, which will be outlined in [4]. Let us just motivate that if we had managed to get a system into the form (5.8), then we can at least recognize the coefficient l_1. Using polar coordinates (5.5) in (5.8) gives for the first component

$$\begin{aligned} y_1' &= r'\cos\varphi - r\varphi'\sin\varphi, \\ y_1' &= qr\cos\varphi - r\sin\varphi + l_1 r^3\cos\varphi, \end{aligned}$$

and for the second component

$$\begin{aligned} y_2' &= r'\sin\varphi + r\varphi'\cos\varphi, \\ y_2' &= r\cos\varphi + qr\sin\varphi + l_1 r^3\sin\varphi. \end{aligned}$$

The last four equations give us a closed system, which can actually be solved (exercise!) for r', φ'. After a few trigonometric manipulations we obtain precisely (5.3)-(5.4), which can thus be also viewed as a normal form.

Definition 5.4. l_1 is called the **first Lyapunov coefficient**.

From the viewpoint of applications, it is not too difficult to validate the eigenvalue conditions for the Hopf bifurcation. However, we may wonder if there is a simple way to calculate the sign of l_1? Clearly, it does make a difference, whether we transition to a small-amplitude periodic orbit (supercritical), or there is only an unstable equilibrium left (subcritical) upon increasing our parameter.

Theorem 5.5 (Lyapunov coefficient formula; [2])**.** *For a planar system* (5.7) *under Hopf bifurcation conditions at* $(x,p) = (0,0)$ *given already in simplified form at this value by*

$$\begin{pmatrix} Y' \\ Z' \end{pmatrix} = \begin{pmatrix} 0 & -\omega \\ \omega & 0 \end{pmatrix} \begin{pmatrix} Y \\ Z \end{pmatrix} + \begin{pmatrix} F(Y,Z) \\ G(Y,Z) \end{pmatrix}, \tag{5.9}$$

with some $\omega > 0$*, the sign of the first Lyapunov coefficient can be calculated from*

$$\begin{aligned} \tilde{l_1} = {} & \partial_{YYY}F + \partial_{YZZ}F + \partial_{YYZ}G + \partial_{ZZZ}G + \partial_{YZ}F(\partial_{YY}F + \partial_{ZZ}F) \\ & -\partial_{YZ}G(\partial_{YY}G + \partial_{ZZ}G) - \partial_{YY}F\partial_{YY}G + \partial_{ZZ}F\partial_{ZZ}G \end{aligned}$$

and all partial derivatives are evaluated at $(0,0)$*, i.e., we have* $\operatorname{sign}(l_1) = \operatorname{sign}(\tilde{l_1})$.

Remark: The form (5.9) can be obtained at a Hopf bifurcation via a linear transformation and a time-rescaling to get the matrix in (5.9). So one just has to evaluate partial derivatives in the end without having to look at periodic orbits explicitly.

So far, all the bifurcations we have studied have been *local* happening near an equilibrium point. However, already in the planar situation, a lot more complicated *global* bifurcations can occur.

Example 5.6. Consider the following second-order ODE

$$x'' - x + x^3 = 0,$$

which can be viewed as a *nonlinear oscillator*. Re-writing as a first-order system. we get

$$\begin{aligned} x' &= y, \\ y' &= x - x^3. \end{aligned} \tag{5.10}$$

The equilibrium point at the origin is easily checked to be a saddle point. Furthermore, the structure of (5.10) can be written in the form

$$\begin{aligned} x' &= \partial_y H(x,y), \\ y' &= -\partial_x H(x,y), \end{aligned} \qquad H(x,y) = \frac{y^2}{2} + \frac{x^4}{4} - \frac{x^2}{2}, \tag{5.11}$$

where $H : \mathbb{R}^2 \to \mathbb{R}$ is called the **Hamiltonian** and systems in the form (5.11) are called **Hamiltonian systems**. ♦

Theorem 5.7 (Hamiltonian level sets)**.** *Consider a Hamiltonian system. Any solution is constant along a given* ***level set*** $\{H(x,y) = \text{constant}\}$.

Proof. One can simply calculate

$$\frac{\mathrm{d}}{\mathrm{d}t} H(x,y) = x'\partial_x H + y'\partial_y H = x'y' - x'y' = 0,$$

so the result follows. □

Example 5.8. (Example 5.6 continued) Since the system is Hamiltonian and planar, we know now that levels are solution curves. So we look at the level sets and find a figure-eight curve containing zero as shown in Figure TODO. In particular, there are two trajectories $\gamma_j(t)$, $j \in \{1,2\}$, asymptotic in forward and backward time to the saddle point at the origin both satisfying the conditions

$$\lim_{t\to-\infty} \gamma_j(t) = (0,0)^\top = \lim_{t\to+\infty} \gamma_j(t). \tag{5.12}$$

Such a trajectory/orbit bi-asymptotic to a saddle is called a **homoclinc orbit**. Now imagine we add small perturbation terms in our original equations

$$\begin{aligned} x' &= y + \varepsilon f_1(x, y), \\ y' &= x - x^3 + \varepsilon f_2(x, y). \end{aligned}$$

Then we would expect that the Hamiltonian structure breaks for $\varepsilon \neq 0$ for most functions $f_{1,2}$. We expect a similar fate for our homoclinic orbit, so the phase portrait changes. Getting two non-equivalent phase portraits under parameter variation is a bifurcation, yet in this case the bifurcation is *global* as not only an equilibrium is involved; see Figure TODO. ♦

6 Multiple Scales and Perturbation Theory

We have already seen in Example 5.8 that the relatively simple nonlinear systems with a cubic nonlinearity can generate very complicated dynamics. Now we shall be more specific and focus on the analysis of one model problem, which has motivated large parts of nonlinear dynamics.

Example 6.1. We consider the **van der Pol equation**

$$x'' + p(x^2 - 1)x' + x = 0, \tag{6.1}$$

which is a prototypical nonlinear oscillator, whose features appear in almost all branches of nonlinear modelling in some form. There are two interesting limits for (6.1) given by

$$p = \delta \to 0, \qquad p = \frac{1}{\sqrt{\varepsilon}} \to \infty,$$

which correspond to very small ($0 < \varepsilon \ll 1$) and very large ($0 < \delta \ll 1$) *nonlinear damping*. It actually turns out that the standard trick $y' = x$ is not the best solution to re-write (6.1) as a first-order system. Instead, we notice

$$0 = x'' + p(x^2 - 1)x' + x = \frac{\mathrm{d}}{\mathrm{d}t}\left(x' + p\left(\frac{x^3}{3} - x\right)\right) + x.$$

Furthermore, we look back at Example 5.2, which leads us to define

$$y := x' + p\left(\frac{x^3}{3} - x\right) \qquad \Rightarrow \quad y' = -x.$$

Considering the case $p = 1/\sqrt{\varepsilon}$, yields

$$\begin{aligned} x' &= y - \tfrac{1}{\sqrt{\varepsilon}}\left(\tfrac{x^3}{3} - x\right), \\ y' &= -x. \end{aligned} \tag{6.2}$$

This system can be simplified even further by scaling using $y = \tilde{y}/\sqrt{\varepsilon}$, $t = \tilde{t}\sqrt{\varepsilon}$, and then employing this scaling and dropping the tildes gives

$$\begin{aligned} x' &= y - \tfrac{x^3}{3} + x =: f(x, y), \\ y' &= -\varepsilon x =: g(x, y). \end{aligned} \tag{6.3}$$

For small $\varepsilon > 0$, it is relatively easy to draw a phase plane. Indeed, away from the curve

$$\mathcal{C}_0 := \big\{(x, y) \in \mathbb{R}^2 : f(x, y) = 0\big\}$$

the vector field (6.3) is almost horizontal as $x' \gg y'$ so x is fast in comparison to y. Assuming formally the limit $\varepsilon = 0$ gives a one-dimensional ODE

$$x' = y - \frac{x^3}{3} + x,$$

where y is viewed as a parameter. The bifurcation diagram is shown in Figure TODO with two fold bifurcations corresponding to the two local extrema of the curve $\mathcal{C}_0$. However, if we are exactly on $\mathcal{C}_0$ and $\varepsilon > 0$, then $x' = 0$ so the slow motion defined by

$$y' = -\varepsilon x$$

governs the dynamics as shown in Figure TODO. In summary, our analysis seems to predict that there is a periodic orbit consisting of two fast and two slow segements as shown in Figure TODO. These types of periodic orbits occur in numerous applications and are called **relaxation oscillations**. ♦

Some observations about the van der Pol equation can be generalized to more general planar **fast-slow systems**

$$\begin{aligned} x' &= f(x,y), \\ y' &= \varepsilon g(x,y). \end{aligned} \tag{6.4}$$

The set $\mathcal{C}_0$ is also called the **critical set**. One also refers to $\mathcal{C}_0$ more commonly as the **critical manifold** but we shall only cover more general invariant manifold theory in [4]. $\mathcal{C}_0$ consists of equilibrium points for the **fast subsystem**

$$\begin{aligned} x' &= f(x,y), \\ y' &= 0. \end{aligned} \tag{6.5}$$

Re-scaling time as $s := \varepsilon t$ in gives

$$\begin{aligned} \varepsilon \tfrac{\mathrm{d}x}{\mathrm{d}s} &= \varepsilon \dot{x} = f(x,y), \\ \tfrac{\mathrm{d}y}{\mathrm{d}s} &= \dot{y} = g(x,y). \end{aligned} \tag{6.6}$$

Taking the limit $\varepsilon \to 0$ gives the **slow subsystem**

$$\begin{aligned} 0 &= f(x,y), \\ \dot{y} &= g(x,y), \end{aligned} \tag{6.7}$$

which is a **differential-algebraic equation (DAE)**. Since the fast and slow subsystems are different types of differential equations, one refers to $\varepsilon \to 0$ as a **singular limit** and to the van der Pol equation as a **singular perturbation** problem. One may actually prove (very difficult!) that under the hypothesis $\partial_x f(z) \neq 0$ for $z \in \mathcal{C}_0$, the fast-slow decomposition we used above to understand phase space just using the fast and slow subsystems. However, the next example shows that we are bound to encounter problems near the two fold bifurcations of the fast subsystem.

Example 6.2. (Example 6.1 continued) So far, we have only analyzed with the van der Pol (vdP) equation in the limit $\varepsilon = 0$ and observed that $\mathcal{C}_0$ is a key object consisting basically of slow trajectory segments. How would this object look for $0 < \varepsilon \ll 1$? Let us consider a simpler model problem

$$\begin{aligned} \varepsilon \dot{x} &= y - x^2, \\ \dot{y} &= -1, \end{aligned} \tag{6.8}$$

which captures one part of the region for the vdP equation near the local minimum of its critical manifold; see also Figure TODO. We have $\mathcal{C}_0 = \{x = \pm\sqrt{y}\}$ and notice that (6.8) can be re-written

$$\varepsilon \frac{\mathrm{d}x}{\mathrm{d}y} = \varepsilon \frac{\mathrm{d}x}{\mathrm{d}s}\frac{\mathrm{d}s}{\mathrm{d}y} = -y + x^2, \tag{6.9}$$

which is a **non-autonomous** ODE as the "time" y appears also in the vector field. We want to find a trajectory close to part of the parabola $\mathcal{C}_0$ for $\varepsilon > 0$ using (6.9) and an **asymptotic expansion**

$$x(y) \sim h_0(y) + \varepsilon h_1(y) + \varepsilon^2 h_2(y) + \cdots \qquad \text{as } \varepsilon \to 0. \tag{6.10}$$

The ansatz (6.10) just means each term in the series is supposed to be smaller than the previous one as $\varepsilon \to 0$. We can just plug (6.10), say up to quadratic terms in ε, into (6.9) to obtain

$$\varepsilon \frac{\mathrm{d}h_0}{\mathrm{d}y} + \varepsilon^2 \frac{\mathrm{d}h_1}{\mathrm{d}y} + \mathcal{O}(\varepsilon^3) = -y + h_0^2 + 2\varepsilon h_0 h_1 + \varepsilon^2(h_1^2 + 2h_0 h_2) + \mathcal{O}(\varepsilon^3).$$

Before we proceed, let us briefly clarify the order $\mathcal{O}(\cdot)$-notation. ♦

Definition 6.3. Consider two functions $k_1(\varepsilon)$ and $k_2(\varepsilon)$. Let $\varepsilon \to 0$ be the relevant asymptotic limit. Then we write

(D1) $k_1 = \mathcal{O}(k_2)$ if $\lim_{\varepsilon\to 0} \frac{k_1(\varepsilon)}{k_2(\varepsilon)} < +\infty$;

(D2) $k_1 \sim k_2$ if $k_1 = \mathcal{O}(k_2)$ and $k_2 = \mathcal{O}(k_1)$;

(D3) $k_1 \ll k_2$ if $\lim_{\varepsilon\to 0} \frac{k_1(\varepsilon)}{k_2(\varepsilon)} = 0$.

Example 6.4. (Example (6.2) continued) Continuing with our calculation above, we can just start to *collect the terms of different orders*. Starting at order $\mathcal{O}(1)$, we get

$$0 = y - h_0(y)^2, \qquad \Rightarrow \quad h_0(y) = \pm\sqrt{y}.$$

Therefore, the leading-order solution $x(y) \sim \pm\sqrt{y} + \mathcal{O}(\varepsilon)$ really does correspond to the critical manifold $\mathcal{C}_0 = \{y = x^2\}$. Let us just consider a **perturbation** for the right-branch

$$\mathcal{C}_0^{\mathrm{a}} = \{(x, y) \in \mathcal{C}_0 : x > 0\}$$

and fix $h_0(y) = \sqrt{y}$; the calculation for the other branch is very similar (exercise!). Now for the next two orders we get

$$\begin{aligned} \mathcal{O}(\varepsilon): &\quad \tfrac{\mathrm{d}h_0}{\mathrm{d}y} = 2h_0h_1. \\ \mathcal{O}(\varepsilon^2): &\quad \tfrac{\mathrm{d}h_1}{\mathrm{d}y} = h_1^2 + 2h_0h_2. \end{aligned}$$

It is very important to note that we now need h_0 to solve for h_1, i.e., the asymptotic series solution has to proceed order-by-order. We get purely algebraic equation

$$\frac{\mathrm{d}h_0}{\mathrm{d}y} = \frac{1}{2\sqrt{y}} = 2\sqrt{y}h_1(y) = 2h_0h_1,$$

which gives us $h_1(y) = 1/4y$. A slightly longer, yet easy (exercise!), calculation gives $h_2(y) = \frac{5}{32y^{5/2}}$. Therefore, an asymptotic solution perturbing near $\mathcal{C}_0$ gives up to second-order terms

$$x(y) \sim \sqrt{y} + \frac{1}{4y}\varepsilon - \frac{5}{32y^{5/2}}\varepsilon^2 + \mathcal{O}(\varepsilon^3). \tag{6.11}$$

Figure TODO shows that this solution essentially coincides with a numerical solution starting near $\mathcal{C}_0^{\mathrm{a}}$ for some $y(0) > 0$. However, as the numerical solution approaches the non-hyperbolic fold bifurcation of the fast subsystem $(x, y) = (0, 0)$, our asymptotic expansion starts to deteriorate and deviates substantially. Indeed, this is not unexpected since the formula (6.11) is no longer a *well-ordered* asymptotic series as

$$y \sim \varepsilon^{2/3} \quad \Rightarrow \quad \varepsilon^{1/3} \sim \frac{1}{4\varepsilon^{2/3}}\varepsilon \sim \frac{5}{32\varepsilon^{5/3}}\varepsilon^2.$$

This is another clear indicator that we work with a singular perturbation problem as in a neighbourhood of $(0, 0)$ with size $(x, y) \sim (\varepsilon^{1/3}, \varepsilon^{2/3})$. In fact, the ansatz (6.10) we made is a **regular** (or **naive**) **perturbation** ansatz. ♦

The last example is *not a special case* or *just bad luck*. There are many situations in multiscale dynamics, where direct perturbation methods work in one regime but have to replaced by more clever ideas in another regime. We are not even safe in the linear case as the next example shows.

Example 6.5. Following up on the van der Pol equation (6.1), we can structurally write the case $p = \delta$ as a special case of

$$x'' + x + \delta H(x, x') = 0, \tag{6.12}$$

which are **weakly nonlinear oscillators**. A special case of (6.12) are oscillators with weak *linear* damping, e.g.,

$$x'' + 2\delta x' + x = 0, \qquad x(0) = 0, \quad x'(0) = 1. \tag{6.13}$$

Let us try a regular perturbation approach for this problem

$$x(t) \sim x_0(t) + \delta x_1(t) + \mathcal{O}(\delta^2).$$

Inserting the ansatz into (6.13) gives the two ODEs

$$\begin{aligned} \mathcal{O}(1): &\quad 0 = x_0'' + x_0. \\ \mathcal{O}(\delta): &\quad 0 = x_1'' + 2x_0' + x_1, \end{aligned} \tag{6.14}$$

as well as the initial conditions

$$x_0(0) = 0, \quad x_1(0) = 0, \quad x_0'(0) = 1, \quad x_1'(0) = 0. \tag{6.15}$$

Solving the first equation in (6.14) and using (6.15) gives

$$x_0(t) = \sin t.$$

So plugging this into (6.14) means we have to solve

$$x_1'' + x_1 = -2\cos t, \qquad x_1(0) = 0, \quad x_1'(0) = 0. \tag{6.16}$$

The ODE (6.16) is a **(periodically) forced oscillator**; in this case, the forcing is also called **resonant** as it appears as a linear factor in the unforced harmonic oscillator. One verifies easily that the solution to (6.16) is

$$x_1(t) = -t \sin t.$$

Therefore, our asymptotic solution is

$$x(t) \sim \sin t - \delta t \sin t + \mathcal{O}(\delta^2).$$

So is this solution correct? Since the problem is linear, we can find after some calculations an exact solution

$$x(t) = \frac{\mathrm{e}^{-\delta t}}{\sqrt{1-\delta^2}} \sin\left(\sqrt{1-\delta^2}t\right). \tag{6.17}$$

The true solution does not grow unbounded as $t \to +\infty$ but the asymptotic solution contains a **secular term** $-\delta t \sin t$, which starts to grow rapidly when $t = \mathcal{O}(1/\delta)$. One may check that the Taylor series in δ of the exact solution does actually produce the same for the first two terms as our asymptotic solution (exercise!) but since our goal is to find a simple approximation with just a few terms also for *nonlinear* problems without explicit solutions, we have to be aware of secular terms. ♦

7 Two-Timing and Matched Asymptotics

In this section, we outline two of the most classical **multiscale methods**. We have seen in Example 6.5 that secular terms can appear in oscillators of the form (6.12). The first method we are going to look at aims to remove these secular terms.

Example 7.1. A famous example for the class (6.12) is the **Duffing equation**

$$x'' + x + \delta x^3 = 0, \qquad x(0) = 1,\ x'(0) = 0. \tag{7.1}$$

One checks with the same type of calculation as in Example 6.5 that secular terms appear in a naive perturbation ansatz. Example (6.5) shows that we expect *two time scales* to play a role t and $s := \delta t$. So we make the **two-timing** ansatz

$$x(t) = X_0(t, s) + \delta X_1(t, s) + \mathcal{O}(\delta^2). \tag{7.2}$$

One also refers to this ansatz as the **method of multiple scales**; the name is quite common and also very unfortunate since there are *many methods* for multiscale problems. Starting from (7.2), the chain rule yields

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \left(\frac{\partial X_0}{\partial t} + \frac{\partial X_0}{\partial s}\frac{\mathrm{d}s}{\mathrm{d}t}\right) + \delta\left(\frac{\partial X_1}{\partial t} + \frac{\partial X_1}{\partial s}\frac{\mathrm{d}s}{\mathrm{d}t}\right) + \cdots$$

and since $\mathrm{d}s/\mathrm{d}t = \delta$ we find

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\partial X_0}{\partial t} + \delta\left(\frac{\partial X_0}{\partial s} + \frac{\partial X_1}{\partial t}\right) + \mathcal{O}(\delta^2).$$

Differentiating one more time leads to

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = \frac{\partial^2 X_0}{\partial t^2} + \delta\left(2\frac{\partial^2 X_0}{\partial s \partial t} + \frac{\partial^2 X_1}{\partial t^2}\right) + \mathcal{O}(\delta^2). \tag{7.3}$$

Substituting (7.3) and (7.2) into the weakly nonlinear Duffing oscillator (7.1) gives for the first two orders

$$\mathcal{O}(1): \quad \frac{\partial^2 X_0}{\partial t^2} + X_0 = 0, \tag{7.4}$$

$$\mathcal{O}(\delta): \quad \frac{\partial^2 X_1}{\partial t^2} + X_1 = -X_0^3 - 2\frac{\partial^2 X_0}{\partial s \partial t}. \tag{7.5}$$

Observe carefully that (7.5) is a partial differential equation (PDE). The general solution of (7.4) is just

$$X_0(t, s) = A(s)\mathrm{e}^{\mathrm{i}t} + \overline{A(s)}\mathrm{e}^{-\mathrm{i}t}$$

where $\overline{A(s)}$ denotes complex conjugate of the **amplitude** $A(s)$. Calculating the nonlinearity for the next order means looking at

$$\begin{aligned} X_0^3 &= A(s)^3 \mathrm{e}^{3\mathrm{i}t} + \overline{A(s)}^3 \mathrm{e}^{-3\mathrm{i}t} + 3|A(s)|^2 A(s)\mathrm{e}^{\mathrm{i}t} + 3|A(s)|^2 \overline{A(s)}\mathrm{e}^{-\mathrm{i}t}, \\ 2\tfrac{\partial^2 X_0}{\partial s \partial t} &= 2\left(\mathrm{i}\mathrm{e}^{\mathrm{i}t}\tfrac{\mathrm{d}A}{\mathrm{d}s} - \mathrm{i}\mathrm{e}^{-\mathrm{i}t}\tfrac{\mathrm{d}\overline{A}}{\mathrm{d}s}\right). \end{aligned}$$

Therefore, the right-hand side of (7.5) is

$$\mathrm{e}^{\mathrm{i}t}\left(-3|A|^2 A - 2\mathrm{i}\frac{\mathrm{d}A}{\mathrm{d}s}\right) + \mathrm{e}^{-\mathrm{i}t}\left(-3|A|^2\overline{A} + 2\mathrm{i}\frac{\mathrm{d}\overline{A}}{\mathrm{d}s}\right) - \mathrm{e}^{3\mathrm{i}t}A^3 - \mathrm{e}^{-3\mathrm{i}t}\overline{A}^3.$$

Since $\mathrm{e}^{\mathrm{i}t}$ and $\mathrm{e}^{-\mathrm{i}t}$ solve the homogeneous problem on the left-hand side of (7.5) we must assure that their coefficients vanish on the right-hand to *avoid secular terms* so we require

$$0 = -3|A|^2 A - 2\mathrm{i}\frac{\mathrm{d}A}{\mathrm{d}s}\ , \tag{7.6}$$

$$0 = -3|A|^2\overline{A} + 2\mathrm{i}\frac{\mathrm{d}\overline{A}}{\mathrm{d}s}\ . \tag{7.7}$$

The **amplitude equations** (7.6) and (7.7) are complex conjugates and hence redundant. One may just solve, say (7.6), to find $A(s)$. Using polar coordinates for the complex plane and thus writing $A(s) = R(s)\mathrm{e}^{\mathrm{i}\theta(s)}$ yields

$$\begin{aligned} \tfrac{\mathrm{d}R}{\mathrm{d}s} &= 0, \\ \tfrac{\mathrm{d}\theta}{\mathrm{d}s} &= \tfrac{3}{2}R^2, \end{aligned}$$

which implies $A(s) = R(0)\exp(\mathrm{i}\theta(0) + \frac{3}{2}\mathrm{i}R^2(0)s)$. Therefore, the zeroth-order solution is

$$X_0(t,s) = 2R(0)\cos\left(\theta(0) + \frac{3}{2}R^2(0)s + t\right).$$

The initial conditions $x(0) = 1$, $x'(0) = 0$ determine $R(0)$ and $\theta(0)$. Since $x(0) = 1$ we must have

$$X_0(0,0) = 1, \qquad X_1(0,0) = 0, \qquad \ldots$$

and $x'(0) = 0$ translates into (using $s = \delta t$)

$$\frac{\partial X}{\partial t}(0,0) = 0, \qquad \frac{\partial X_1}{\partial t}(0,0) = -\frac{\partial X_0}{\partial s}(0,0), \qquad \ldots.$$

To satisfy the initial conditions we must require $R(0) = \frac{1}{2}$ and $\theta(0) = 0$. It follows that the zeroth-order solution on the original time scale t is given by

$$x_0(t) = \cos\left[t\left(1 + \frac{3}{8}\delta\right)\right] + \mathcal{O}(\delta) \tag{7.8}$$

as $\delta \to 0$, where we have avoided secular terms. ♦

Returning to the van der Pol equation from Example 6.4, we still have to address that certain expansions can fail locally in phase space. To solve this problem in generality here is too advanced but the basic idea is extremely important for many different classes of multiscale problems. The next example illustrates the key principle of **matched asymptotics**.

Example 7.2. As a model problem, we are going to study the non-autonomous ODE

$$x'' + (1 + \varepsilon t)x' + \varepsilon x = 0. \tag{7.9}$$

We turn it into a **boundary value problem (BVP)** for $t \in [0, 1]$ assuming the boundary conditions

$$x(0) = 1 \qquad \text{and} \qquad x(1) = 1. \tag{7.10}$$

Example 6.4 indicates that if we turn the vdP equation into a BVP, then we expect for "most" conditions that the initial movement of trajectories is fast, i.e., it as a **fast/inner layer** near $t = 0$ and then a slow variation, also called **outer layer** over the time remaining interval; see Figure TODO. This gives us the idea to try to use *two* expansions. The **inner expansion**

$$x(t) = g_0(t) + g_1(t)\varepsilon + \mathcal{O}(\varepsilon^2). \tag{7.11}$$

Substituting (7.11) into (7.9) and collecting terms we obtain

$$g_0'' + g_0' = 0 \quad \text{and} \quad g_1'' + g_1' + tg_0' + g_0 = 0.$$

Taking into account the boundary conditions $g_0(0) = 1$ and $g_1(0) = 0$ yields

$$g_0(t) = 1 + A_0(\mathrm{e}^{-t} - 1) \quad \text{and} \quad g_1(t) = -t + A_0\left(-\frac{1}{2}t^2\mathrm{e}^{-t} + t\right) + A_1(\mathrm{e}^{-t} - 1),$$

where we have *no information yet*, how to determine the constants of integration $A_{0,1}$. Since the problem is multiscale, let us also look at the second natural scale $s = t\varepsilon$ in (7.9), which gives

$$\varepsilon\frac{\mathrm{d}^2x}{\mathrm{d}s^2} + (1 + s)\frac{\mathrm{d}x}{\mathrm{d}s} + x = \varepsilon\ddot{x} + (1 + s)\dot{x} + x = 0. \tag{7.12}$$

Let us use our standard ansatz

$$x(s) = h_0(s) + h_1(s)\varepsilon + \mathcal{O}(\varepsilon^2). \tag{7.13}$$

Substituting (7.13) into (7.12) and collecting terms of different orders in ε gives the equations

$$(1 + s)\dot{h}_0 + h_0 = 0 \quad \text{and} \quad (1 + s)\dot{h}_1 + h_1 + \ddot{h}_0 = 0,$$

where we must satisfy the boundary conditions $h_0(1) = 1$ and $h_1(1) = 0$. The equations are easily solved

$$h_0(s) = 2(1+s)^{-1} \quad \text{and} \quad h_1(s) = 2(1+s)^{-3} - \frac{1}{2}(1+s)^{-1}.$$

The outer (7.13) is not valid at $s = 0$ as it does not even satisfy the boundary condition at $s = 0$. However, one can just expand the solution in a series under the assumption $s \to 0^+$, i.e. this is an expansion for the slow/outer solution in the fast/inner limit and requires $s \ll 1$. This idea yields

$$x(s) = 2 + \mathcal{O}(\varepsilon, s) \qquad \text{as } s \to 0^+,\ \varepsilon \to 0.$$

Similarly, one can expand the fast/inner solution in the outer/slow limit $t \to \infty$. As long as $1 \ll s/\varepsilon = t$ and $s \ll 1$ hold we also have $\varepsilon t \to 0^+$. Therefore, expanding the fast solution in terms of $\varepsilon t \approx 0$ leads to

$$x(t) = 1 - A_0 + \mathcal{O}(\varepsilon t) = 1 - A_0 + \mathcal{O}(s).$$

The expressions $1 - A_0 + \mathcal{O}(s)$ and $2 + \mathcal{O}(\varepsilon, s)$ should agree to get a **composite expansion** so that we must have $A_0 = -1$ to get order $\mathcal{O}(1)$. It can be shown that for the expansion up to order $\mathcal{O}(\varepsilon)$ one obtains $A_1 = -\frac{3}{2}$. In particular, we have shown that the two expansions can agree on an (in-time) **overlap domain** to get a uniformly valid asymptotic expansion. ♦

8 Discrete-Time Dynamics

Instead of differential equations, there are several motivations to study discrete-time problems. Here we just mention a few, beyond the directly obvious requirements that specific *models* may enforce. Consider the ODE

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x' = f(x), \qquad x = x(t) \in \mathbb{R}. \tag{8.1}$$

Numerical schemes discretize (8.1) on a **time grid** $0 = t_0 < t_1 = t_0 + h < t_2 + 2h \cdots$ with regular spacing $h > 0$. For example, the **forward Euler method** gives

$$x' \approx \frac{x(t_{j+1}) - x(t_j)}{h} = f(x(t_j)).$$

Defining $x(t_j) =: y_j$ and $y_j + f(y_j) =: g(y_j)$, the Euler method is an iteration

$$y_{j+1} = g(y_j) \qquad g : \mathbb{R} \to \mathbb{R},\ y_0 \text{ given}. \tag{8.2}$$

The rule (8.2) defines an **iterated map** or **difference equation**.

Example 8.1. Let us re-consider the van der Pol oscillator (6.3)

$$\begin{aligned} x' &= y - \tfrac{1}{3}x^3 + x, \\ y' &= -\varepsilon x, \end{aligned} \tag{8.3}$$

which displays periodic relaxation oscillations. How can we study the (local) stability of these oscillations as sketched in Figure TODO? Although the problem is *global* around the orbit, there is an important way to localize the dynamics. Define a **section** (cf. proof of Theorem 4.6)

$$\Sigma := \{(x, y)^\top \in \mathbb{R}^2 : y = 0, x > 0\}.$$

Solving (8.3) induces a **Poincaré map** (or **return map**). More precisely, let $z_0 \in \Sigma$ and let T_{z_0} be the first return time of the trajectory for (8.3) starting at $z_0 = z(0) = (x(0), y(0)) = (x_0, y_0)$ then we define

$$P : \Sigma \to \Sigma, \quad P(x_0, y_0) = (x(T_{z_0}), y(T_{z_0})) =: (x_1, y_1).$$

The map P is effectively *one-dimensional* as we can just look at the first component and define $x_1 := g(x_0)$. This can be iterated so we obtain a map

$$x_{j+1} = g(x_j), \qquad x_j \in (0, \infty). \tag{8.4}$$

A point (x_0^*, y_0^*), which lies on a periodic orbit for (8.3), becomes invariant under g, so that $g(x_0^*) = x_0^*$. ♦

The concept of a Poincaré map can be generalized to higher-dimensional ODEs [4]. The following concept is key and we state it for general maps.

Definition 8.2. Consider an iterated map $g : \mathcal{X} \to \mathcal{X}$ for $\mathcal{X} \subseteq \mathbb{R}^d$. A point $x^* \in \mathcal{X}$ is called a **fixed point** if $g(x^*) = x^*$.

To simplify our notation we write instead of composition of maps $g(g(\dots))$ just $g^2(y) := g(g(x))$, $g^3(x) := g(g(g(x)))$, and emphasize that one has to be careful not to confuse the notation with taking powers.

Definition 8.3. A fixed point x^* of an iterated map is **locally asymptotically stable** if there exists a neighbourhood $\mathcal{U}$ of x^* such that

$$\lim_{k\to\infty} g^k(x) = x^*$$

for all $x \in \mathcal{U}$; cf. Definition 1.4 for equilibrium points of vector fields.

Hence, we have to understand, how to calculate local stability of maps.

Example 8.4. Based upon Example 1.1, it seems wise to start from a general linear one-dimensional toy problem

$$x_{j+1} = g(x_j) := \mu x_j, \tag{8.5}$$

where $\mu \in \mathbb{R}$ is a parameter. The map (8.5) can be iterated/solved explicitly

$$x_j = \mu x_{j+1} = \mu^2 x_{j-2} = \mu^3 x_{j-3} = \cdots = \mu^j x_0.$$

So the *absolute value of* μ is the important stability indicator as for $|\mu| < 1$ we get $x_j \to 0$ as $j \to +\infty$ so the fixed point $x^* = 0$ is (even globally) **stable**. For $|\mu| > 1$, the fixed point is **unstable**. ♦

Definition 8.5. Let $g : \mathcal{X} \to \mathcal{X}$ with $\mathcal{X} \subseteq \mathbb{R}$ be a one-dimensional map with fixed point x^*. Then x^* is called **hyperbolic** if $|g'(x^*)| \neq 1$. The value $g'(x^*)$ is called a **multiplier**.

The next result is an immediate consequence of the discussion in Example (8.4).

Proposition 8.6. *For $\mu > 1$ the fixed point x^* is **unstable**, while for $\mu < 1$, the fixed point x^* is locally asymptotically stable.*

Hence, the stability question of the periodic orbit we observed in the van der Pol equation in Example 8.1 is equivalent to the analysis of a one-dimensional map. One-dimensional maps also appear as Poincaré maps in many other contexts so we shall investigate them in some detail. There is a natural generalization of the fold bifurcation.

Theorem 8.7 (**fold bifurcation of maps**). *Consider a one-dimensional map*

$$x_{j+1} = g(x_j, p), \qquad x_j \in \mathbb{R},\ p \in \mathbb{R}. \tag{8.6}$$

*and fixed point $x^* = 0$ for $p = 0$. A **fold bifurcation** occurs if $\partial_x g(0,0) = 1$. Suppose the conditions $\partial_{xx} g(0,0) \neq 0$ and $\partial_p g(0,0) \neq 0$ hold, then the dynamics near the bifurcation is equivalent to the normal form*

$$y_{j+1} = p + y_j \pm y_j^2.$$

We shall not prove the normal form result here but see [4]. Note that the normal form needs genericity conditions similar to Theorem 2.3 to be valid as a proxy for the general dynamics. The next example already indicates that one-dimensional maps have a lot richer dynamics than 1- and 2-dimensional vector fields.

Example 8.8. Since we already learned a lot from the one-dimensional logistic equation starting from Example 1.3, it is very natural to generalize this model to a discrete-time version

$$x_{j+1} = p x_j (1 - x_j) =: g(x_j, p), \qquad x_j \in \mathbb{R}, \tag{8.7}$$

and $p > 0$ is a parameter. Biologically one can justify (8.7), e.g., since some species only reproduce during fixed phases during the year. A natural domain for (8.7) seems to be $\mathcal{X} = [0,1]$; see Figure TODO. We want to *iterate* the map $g : \mathcal{X} \to \mathcal{X}$, which is only guaranteed if

$$g(x,p) = px(1-x) \leq 1 \qquad \forall x \in [0,1].$$

The maximum of f can simply be found by looking at

$$\partial_x g(x,p) = p - 2px \stackrel{!}{=} 0 \quad \Rightarrow \quad x = \frac{1}{2}.$$

The value at the maximum is $g(1/2, p) = p/4$ so a natural definition of the **logistic map** is

$$x_{j+1} = g(x_j, p) = p x_j (1 - x_j), \qquad g : [0,1] \to [0,1], \quad p \in (0,4]. \tag{8.8}$$

Since $g(0,p) = 0$, it follows that 0 is always a fixed point with the associated locally linearized map

$$X_j = (\partial_x g(0,p)) X = (p - 2p \cdot 0) X = pX.$$

Therefore, 0 is locally asymptotically stable for $p \in (0,1)$. In fact, the fixed point is even globally stable in this parameter range for any $x_0 \in \mathcal{X}$, which

can be illustrated using the **cobweb** construction in Figure TODO. More generally, fixed points of (8.8) can be found solving

$$x = g(x,p) \quad \Rightarrow \quad x = 0 \text{ or } 1 = p(1-x) \iff x = x^* := 1 - 1/p.$$

Note that the fixed point $x^* = 1 - 1/p$ does not lie in $\mathcal{X}$ for $p \in (0,1)$. One easily checks computing the linearization that a bifurcation occurs at $p = 1$, where 0 and x^* exchange stability at a **transcritical bifurcation (of maps)**; see Figure 1(a). Linearizing at x^* yields

$$X' = (p - 2px^*)X = (p - 2p(1-1/p))X = (-p+2)X.$$

Therefore, x^* is locally stable for any $p \in (1,3)$. However, at $p = 3$ we get a new situation since a multiplier $\mu = -1$ occurs. Since there are no additional fixed points generated, we may wonder, what happens at $p = 3$. The cobweb diagram shown in Figure TODO indicates that a periodic orbit may emerge. To find periodic orbits of *period two* we look at the **second-iterate map**

$$g(g(x,p),p) = p^2x(1-x)(1-px(1-x)) = (p^2x - p^2x^2)(1 - px + px^2).$$

Fixed points of g can be found by solving

$$x = (p^2x - p^2x^2)(1 - px + px^2). \tag{8.9}$$

Solving this fourth-order polynomial equation looks difficult at first. Then we realize that the two fixed points $x = 0$ and $x^* = 1 - 1/p$ must still solve (8.9) as they are trivial fixed points of the second iterate map. Removing these two solutions by long division and solving the remaining quadratic equation yields (exercise!) two solutions

$$x^{\pm} = \frac{1 + p \pm \sqrt{(p-3)(p+1)}}{2p}.$$

So a two-cycle does indeed exist for $p \in (3,4]$; see Figure TODO. ♦

The bifurcation we just observed is a special case of the following:

Theorem 8.9 (flip bifurcation)**.** *Consider a one-dimensional map* (8.6)*. The normal form for a generic* ***flip bifurcation*** *occurring when* $g(0,0) = 0$ *and* $\mu = -1$*, is given by*

$$y_{j+1} = -(1+p)y_j \pm y_j^3, \qquad y_j \in \mathbb{R},\ p \in \mathbb{R}. \tag{8.10}$$

We shall not discuss the genericity conditions for a flip bifurcation here as they are a bit more intricate to derive than for the other bifurcations of maps we have considered here but see [4]. The flip bifurcation is also often referred to as **period-doubling bifurcation**.

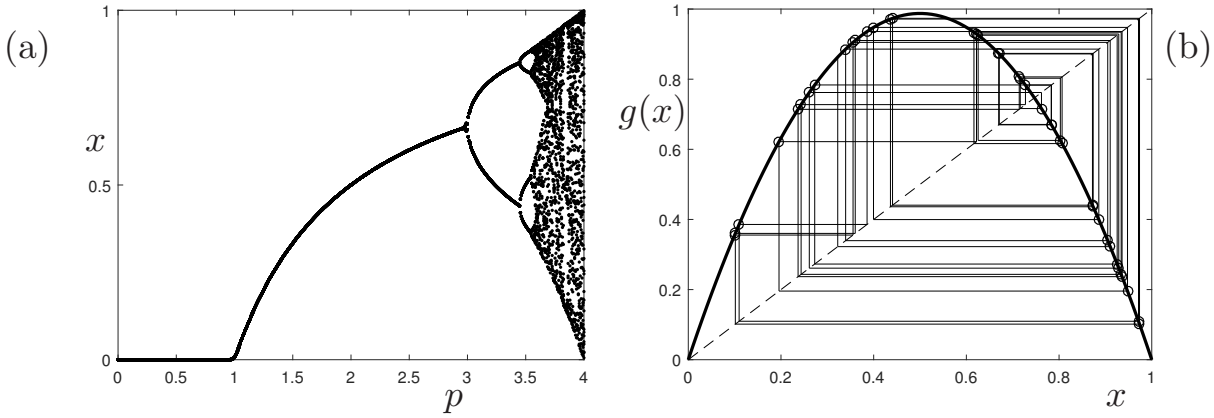


Figure 1: (a) Bifurcation diagram of the logistic map (8.8) obtained by direct simulation for a 500 point equally spaced mesh of the parameter space $p \in [0, 4]$. Transients have been removed (here: first 100 iterations), then 40 iterates are plotted for each value of p. (b) Illustration of the cobweb construction for $p = 3.95$ showing clearly the extremely complicated dynamics of the logistic map in this regime.

Example 8.10. (Example 8.8 continued) Having understood the flip bifurcation at $p = 3$, one may wonder, what happens to the **two-cycle** for $p \in (3, 4]$. Analytically, this analysis is beyond our scope here but Figure 1(a) shows a bifurcation diagram for $p \in (0, 4]$. The results are quite staggering showing a **period-doubling cascade** showing bifurcations to various k-cycle orbits. In fact, there are parts of parameter space, where we do not seem to see anything but "chaotic" behaviour of our iteration. ♦

9 Chaos in Iterated Maps

The bifurcation sequence in the logistic map has already provided a good indication that one-dimensional maps produce complicated dynamics. The next example is a simplified version of the logistic map.

Example 9.1. Consider the unit interval $\mathcal{X} = [0, 1]$ and define the **tent map**

$$x_{j+1} = g(x_j; p) = \begin{cases} px_j & \text{if } x_j \in [0, 1/2) \\ p(1 - x_j) & \text{if } x_j \in [1/2, 1] \end{cases} \tag{9.1}$$

as shown in Figure TODO for the standard case $p = 2$, which we shall consider from now on. The tent map inherites as a key feature from the logistic map that it is **unimodal** having a unique maximum in its domain of definition. ⧫

In more generality, we shall consider continuous interval maps

$$g : [0, 1] \to \mathbb{R}, \qquad g \in C^0.$$

Instead of tracking every point precisely, it will be easier to just capture to look at certain subintervals. So let $0 = y_0 < y_1 < \cdots y_{n-1} < y_n = 1$ be a partition of $[0, 1]$ with associated intervals $\mathcal{I}_j = [y_{j-1}, y_j]$. The interval $\mathcal{I}_j$ is said to **g-cover** $\mathcal{I}_k$ for m-times if there exist m open disjoint subintervals $\mathcal{K}_1, \ldots, \mathcal{K}_m$ of $\mathcal{I}_j$ such that

$$g\left(\overline{\mathcal{K}_r}\right) = \mathcal{I}_k, \qquad \text{for } r \in \{1, 2, \ldots, m\}.$$

An example of the covering property is shown for the tent map discussed in Example 9.3.

Definition 9.2. A (generalized) **transition graph** of g is a direct generalized graph with vertices $\mathcal{I}_j$. There exist m edges from $\mathcal{I}_j$ to $\mathcal{I}_k$ if $\mathcal{I}_j$ does g-cover $\mathcal{I}_k$ for m-times.

Example 9.3. For the standard tent map (9.1) with $p = 2$, we see that $\mathcal{I}_1 = [0, 1/2]$ and $\mathcal{I}_2 = [1/2, 1]$ each cover the other interval and itself once under iteration. The resulting graph is shown in Figure TODO. ⧫

Instead of tracking the iteration of every point, we now just look at sequences of intervals, which is actually a first glimpse at the concept of **symbolic dynamics**. An **allowed path** is a sequence of intervals $\mathcal{I}_{a_1}\mathcal{I}_{a_2} \ldots \mathcal{I}_{a_{n+1}}$, where $a_j \in \mathbb{N}$, and there is an edge from $\mathcal{I}_{a_j}$ to $\mathcal{I}_{a_{j+1}}$ for each $j \in \{1, 2, \ldots, n\}$.

Lemma 9.4. *Consider an allowed path with $a_1 = a_{n+1}$. Then there exists a point $x \in \mathcal{I}_{a_1}$ such that $g^n(x) = x$ and $g^j(x) \in \mathcal{I}_{a_j}$ for $j \in \{2, \ldots, n\}$. In particular, x is periodic with period n.*

Proof. From the definition of an allowed path, there exist subintervals $\mathcal{K}_j \subseteq \mathcal{I}_{a_j}$ such that

$$g(\mathcal{K}_j) = \mathcal{K}_{j+1} \quad \text{for } j = 1, \ldots, n \qquad \text{and} \qquad \mathcal{K}_{n+1} = \mathcal{I}_{a_{n+1}} = \mathcal{I}_{a_1}.$$

Therefore, $g^n(\mathcal{K}_1) = \mathcal{I}_{a_1}$ and $\mathcal{K}_1 \subseteq \mathcal{I}_{a_1}$ so applying the **Intermediate Value Theorem** [6] gives a fixed point $g^n(x) = x$ for $x \in \mathcal{K}_1$. We have by construction that $g(\mathcal{K}_j) = \mathcal{K}_{j+1} \subseteq \mathcal{I}_{a_{j+1}}$ so the last part of the lemma also follows. □

To restrict our attention to periodic points with *minimal period*, we say that a path is **irreducible** if it is not the periodic repetition of a shorter path.

Example 9.5. Consider the tent map from Example 9.3. Then $\mathcal{I}_1\mathcal{I}_2\mathcal{I}_1\mathcal{I}_2\mathcal{I}_1$ is not irreducible while $\mathcal{I}_1\mathcal{I}_2\mathcal{I}_1$ is. ♦

The next result is one of the classical landmark results in the field. It was discovered first by Sharkovskii in more general form as stated in Theorem 9.8 below but the following version is easier to prove and to remember.

Theorem 9.6 (Li-Yorke Theorem; "Period-Three implies Chaos")**.** *Suppose $g : [a, b] \to \mathbb{R}$ with $a < b$, has a periodic orbit with minimal period three. Then g has periodic orbits of all periods.*

Proof. Wlog we may translate and scale coordinates to restrict to $[a, b] = [0, 1]$. Let $p_1, p_2, p_3 \in [0, 1]$ be points on the periodic orbit with $p_1 < p_2 < p_3$ and

$$g(p_1) = p_2, \qquad g(p_2) = p_3, \qquad g(p_3) = p_1.$$

The last assumption is wlog upon reversing the interval direction. Define $\mathcal{I}_1 = [p_1, p_2]$ and $\mathcal{I}_2 = [p_2, p_3]$. By the Intermediate Value Theorem, it follows that $\mathcal{I}_2 \subseteq g(\mathcal{I}_1)$ and $\mathcal{I}_1 \cup \mathcal{I}_2 \subseteq g(\mathcal{I}_2)$; see Figure TODO. Therefore, the transition graph of g is given as in Figure TODO. There exists a fixed point by Lemma (9.4) as $\mathcal{I}_2\mathcal{I}_2$ is an allowed path. Similarly, we get a period two orbit as $\mathcal{I}_2\mathcal{I}_1\mathcal{I}_2$ is an allowed path. Period three exists by assumption and minimal periods with $n > 3$ are constructed by the irreducible and allowed path $\mathcal{I}_2\mathcal{I}_1(\mathcal{I}_2)^{n-2}\mathcal{I}_2$. □

In fact, one may get a lot finer information on the appearance of different families of periodic orbits.

Definition 9.7. The **Sharkovskii ordering** $\triangleleft$ on $\mathbb{N}$ is defined by

$$1 \triangleleft 2 \triangleleft 4 \triangleleft 2^3 \triangleleft \cdots \triangleleft 2^n \triangleleft 2^{n+1} \cdots$$
$$\cdots \triangleleft 9 \cdot 2^{n+1} \triangleleft 7 \cdot 2^{n+1} \triangleleft 5 \cdot 2^{n+1} \triangleleft 3 \cdot 2^{n+1} \triangleleft \cdots$$
$$\cdots \triangleleft 9 \cdot 2^n \triangleleft 7 \cdot 2^n \triangleleft 5 \cdot 2^n \triangleleft 3 \cdot 2^n \triangleleft \cdots \triangleleft 9 \triangleleft 7 \triangleleft 5 \triangleleft 3.$$

This ordering seems odd at first but may not be entirely unexpected since we have already seen in Example 8.10 that we first got period 2, then period 4, and so on upon parameter variation.

Theorem 9.8 (Sharkovskii Theorem)**.** *Suppose $g : [a, b] \to \mathbb{R}$ has a periodic orbit with minimal period n. Then g has periodic orbits of all periods $k \triangleleft n$.*

The proof idea of Theorem 9.6 analyzing transition graphs carefully also applies to the (lengthy!) proof of Sharkovskii's Theorem. One can make precise that the occurence of many different types of periodic orbits is one of the key elements of **chaos**. To perform a precise analysis is beyond this course [4] but we shall encounter another chaotic dynamical system in the next section.

10 Attractors and Time Series

In the previous sections, we have seen that already one-dimensional maps can be chaotic. Here we shall illustrate, how these results link directly into applications. In fact, applications and models first generated chaos, and then the mathematics of one-dimensional maps was studied.

Example 10.1. Consider a fluid confined between two plates; see Figure TODO. The standard model describing fluid motion at the **Navier-Stokes equations**, which are nonlinear partial differential equations (PDEs). Although directly studying Navier-Stokes is beyond this course, it turns out that one can try to study a simpler approximation based upon a Fourier series approximation of the solution of Navier-Stokes. This approximation can be written as a three-dimensional system of ODEs

$$\begin{aligned} x_1' &= \sigma(x_2 - x_1), \\ x_2' &= \rho x_1 - x_2 - x_1 x_3, \\ x_3' &= x_1 x_2 - \beta x_3, \end{aligned} \tag{10.1}$$

with parameters $\sigma, \rho, \beta > 0$. The ODEs are also known as the **Lorenz system**. A numerical simulation of (10.1) is shown in Figure 2.

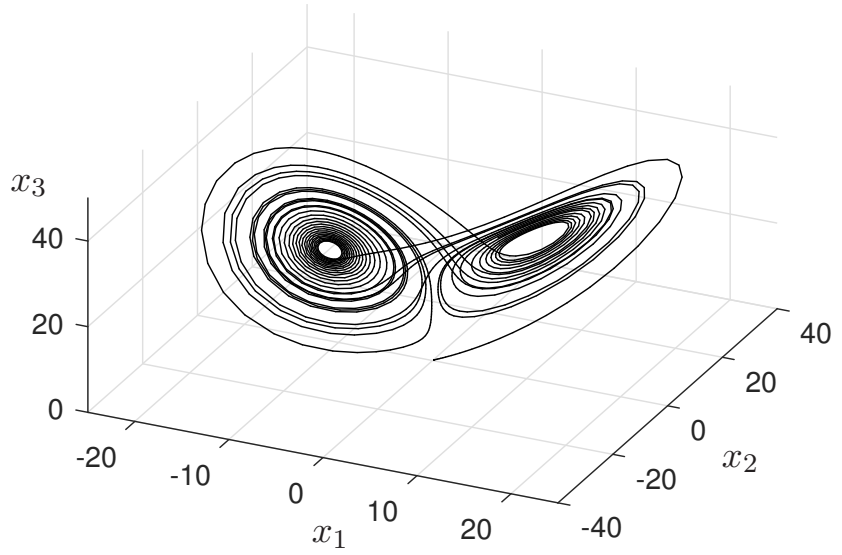


Figure 2: Forward integration of the Lorenz system (10.1) for parameter values $\sigma = 10$, $\beta = 8/3$, $\rho = 28$, and initial condition $x(0) = (0.1, 0.1, 0.1)^\top$.

For the chosen parameter values, one observes that trajectories all seem to tend to a complicated structure, also called the **Lorenz attractor**. ♦

Although we cannot provide the advanced details for the Lorenz attractor here, let us make more precise that trajectories stay bounded.

Definition 10.2. A compact set $\mathcal{U}$ is called a **trapping region** for a vector field provided that all orbits starting in $\mathcal{U}$ are contained in the interior of $\mathcal{U}$ after some time $t > 0$.

Proposition 10.3. *There exists a trapping region for the Lorenz system* (10.1)*.*

Proof. To show that a trapping region exists for the flow generated by (10.1) consider the function

$$L(x) := \frac{x_1^2 + x_2^2 + (x_3 - \rho - \sigma)^2}{2}.$$

Essentially, L can be thought of as a Lyapunov-like function for an ellipsoid since one calculates (exercise!)

$$\frac{\mathrm{d}}{\mathrm{d}t} L(x) = -\sigma x_1^2 - x_2^2 - \beta \left(x_3 - \frac{\rho + \sigma}{2} \right)^2 + \frac{\beta(\rho + \sigma)^2}{4}.$$

We have $L' < 0$ if

$$\sigma x_1^2 + x_2^2 + \beta \left(x_3 - \frac{\rho + \sigma}{2} \right)^2 > \frac{\beta(\rho + \sigma)^2}{4} \tag{10.2}$$

which occurs outside of an ellipsoid $\mathcal{E}$ with boundary $\partial\mathcal{E}$ defined by considering equality in (10.2). $\mathcal{E}$ attracts all trajectories starting outside it and $\mathcal{E}$ is invariant in forward time. □

Although the Lorenz system is a nice model, and can be analyzed a lot further, let us take a step back and think, how we could actually link our models more directly to *applications*. The typical situation is that experiments are performed or field data is gathered. Usually the only common ground is that we can get a time series

$$x(t_0), x(t_1), \ldots, x(t_M)$$

for a certain fixed number of points $M > 0$. An easy example to think of is a chemical reaction, where we measure the concentration of a certain chemical at fixed time intervals. However, in more generality, we expect just to be able to measure a function of phase space.

Definition 10.4. Let $\mathcal{X}$ be the phase space. A function $v : \mathcal{X} \to \mathbb{R}$ is also called an **observable**.

Usually, one requires some additional properties (e.g., measurability, differentiability, smoothness, etc.) but we shall state them as needed.

Example 10.5. (Example 10.1 continued) Let us suppose we do not know the Lorenz equations but just observe some parts of its output, say for simplicity $x_1(t_j)$ for certain times t_j to be specified so the observable is

$$v(x_1, x_2, x_3) = x_1.$$

Can we reconstruct the dynamics shown in Figure TODO just from the time series $x_1(t_j)$? This seems like an under-determined question as we just know one variable. A key trick is to consider a **delay embedding** for a fixed delay $\tau > 0$, i.e., we consider

$$\alpha(t) = (\alpha_1(t), \alpha_2(t))^\top := (x_1(t), x_1(t-\tau))^\top. \tag{10.3}$$

which can be computed from a time series if the spacing of samples is chosen commensurate with τ. Of course, this can be generalized and we could use three components

$$\alpha(t) = (\alpha_1(t), \alpha_2(t), \alpha_3(t))^\top := (x_1(t), x_1(t-\tau), x_1(t-2\tau))^\top. \tag{10.4}$$

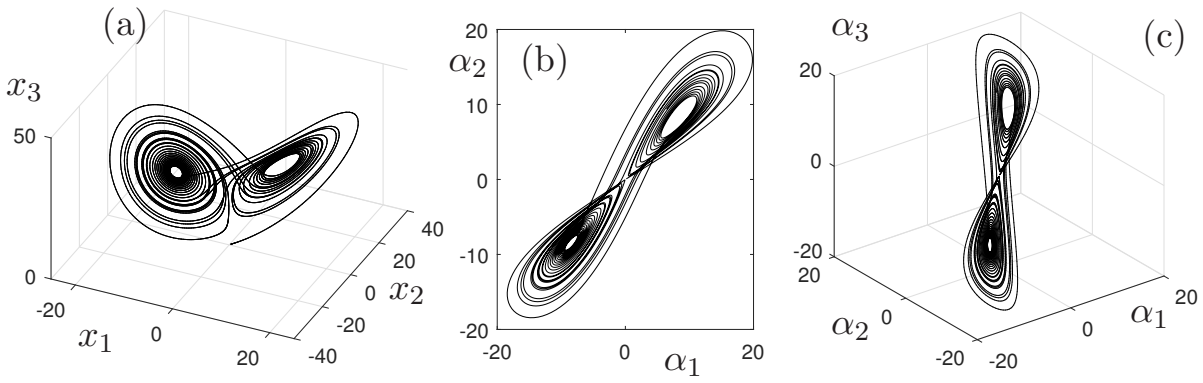


Figure 3: (a) Plot of the Lorenz attractor as in Figure 2. (b) Delay coordinate embedding (10.3) for delay $\tau = 0.055$. (c) Delay coordinate embedding (10.4) for delay $\tau = 0.055$.

Figure 3 shows the original Lorenz attractor, as well as a two- and three-dimensional coordinates α from the delay embedding. We do get a very accurate view of the structure of the attractor in the delay coordinate *phase space* although we have only used a *single* observation function. We also observe by comparing Figure 3(b) and Figure 3(c) that two dimensions almost seem to suffice to obtain a good representation of the attractor. In fact, one can computationally check that the Lorenz attractor has a non-integer dimension slightly larger than two; see [4] for more discussion regarding non-integer dimension. ♦

The process we have carried out in Example 10.5 is known as **attractor reconstruction**, or more generally **phase space reconstruction**. In fact, there is a theorem providing very general conditions that a delay embedding works. Although we cannot discuss it here in full technical detail, it is interesting to just take a first look.

Theorem 10.6 (**Takens' Theorem**). *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact m-dimensional manifold. Consider an ODE*

$$x' = f(x), \qquad x \in \mathbb{R}^d,\ f \in C^2(\mathbb{R}^d, \mathbb{R}^d),\ x(0) = x_0,$$

and an observable $v : \mathcal{M} \to \mathbb{R}$. Then the mapping $\Phi : \mathcal{M} \to \mathbb{R}^{2m+1}$

$$\Phi(x_0) := (v(x(0)), v(x(\tau)), \ldots, v(x(2m\tau))), \qquad \textit{with } \tau > 0,$$

is, under generic conditions, a diffeomorphism onto its image.

Here *generic conditions* means that most C^2 vector fields f are going to satisfy these conditions. Usually one takes as $\mathcal{M}$ just a closed bounded set without boundary containing the attractor; C^2 manifold just means we can locally parametrize $\mathcal{M}$ via a C^2-mapping from some open set in $\mathbb{R}^m$. The conclusion that we get a diffeomorphism onto its image means that $\Phi(x_0)$ can be used to define a trajectory in our reconstructed phase space $\mathbb{R}^{2m+1}$; see Figure TODO.

POST GRADUATE DEGREE PROGRAMME (CBCS) IN

MATHEMATICS

SEMESTER III

SELF LEARNING MATERIAL

PAPER : MATP 3.4
(Pure Stream)

- Block - I : Topological Groups
- Block - II : Measure Theory

Directorate of Open and Distance Learning
University of Kalyani
Kalyani, Nadia
West Bengal, India

Course Preparation Team

1. Mr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani	2. Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

November, 2019

Directorate of Open and Distance Learning, University of Kalyani

Published by the Directorate of Open and Distance Learning

University of Kalyani, 741235, West Bengal

Director's Massage

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Sankar Kumar Ghosh, Hon'ble Vice-Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as weel as each and every member of PGBOS (DODL), University of Kalyani, Heartfelt thanks is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Elective Paper

MATP 3.4

Block - I

Marks : 50 (SSE : 40; IA : 10)

Topological Groups (Pure Stream)

Unit 1

2 Background on topological spaces and abstract groups

2.1 Background on abelian groups

Generally a group G will be written multiplicatively and the neutral element will be denoted by e_G or simply e or 1 when there is no danger of confusion. For a subset $A, A_1, A_2, \ldots, A_n$ of a group G we write

$$A^{-1} = \{a^{-1} : a \in A\}, \quad \text{and} \quad A_1A_2 \ldots A_n = \{a_1 \ldots a_n : a_i \in A_i, i = 1, 2, \ldots, n\} \qquad (*)$$

and we write A^n for $A_1A_2 \ldots A_n$ if all $A_i = A$. Moreover, for $A \subseteq G$ we denote by $c_G(A)$ the *centralizer* of A, i.e., the subgroup $\{x \in G : xa = ax \text{ for every } a \in A\}$.

We use additive notation for abelian groups, consequently 0 will denote the neutral element in such a case. Clearly, the counterpart of (*) will be $-A$, $A_1 + A_2 + \ldots + A_n$ and nA.

A standard reference for abelian groups is the monograph [46]. We give here only those facts or definitions that appear very frequently in the sequel.

For $m \in \mathbb{N}_+$, we use $\mathbb{Z}_m$ or $\mathbb{Z}(m)$ for the finite cyclic group of order m. Let G be an abelian group. The subgroup of torsion elements of G is $t(G)$ and for $m \in \mathbb{N}_+$

$$G[m] = \{x \in G : mx = 0\} \quad \text{and} \quad mG = \{mx : x \in G\}.$$

For a family $\{G_i : i \in I\}$ of groups we denote by $\prod_{i\in I} G_i$ the direct product G of the groups G_i. The underlying set of G is the Cartesian product $\prod_{i\in I} G_i$ and the operation is defined coordinatewise. The direct sum $\bigoplus_{i\in I} G_i$ is the subgroup of $\prod_{i\in I} G_i$ consisting of all elements of finite support. If all G_i are isomorphic to the same group G and $|I| = \alpha$, we write $\bigoplus_\alpha G$ (or $G^{(\alpha)}$, or $\bigoplus_I G$) for the direct sum $\bigoplus_{i\in I} G_i$

A subset X of an abelian group G is independent, if $\sum_{i=1}^n k_i x_i = 0$ with $k_i \in \mathbb{Z}$ and distinct elements x_i of X, $i = 1, 2, \ldots, n$, imply $k_1 = k_2 = \ldots = k_n = 0$. The maximum size of an independent subset of G is called *free-rank* of G and denoted by $r_0(G)$. An abelian group G is *free* , if G has an independent set of generators X. In such a case $G \cong \bigoplus_{|X|} \mathbb{Z}$.

For an abelian group G and a prime number p the subgroup $G[p]$ is a vector space over the finite field $\mathbb{Z}/p\mathbb{Z}$. We denote by $r_p(G)$ its dimension over $\mathbb{Z}/p\mathbb{Z}$ and call it *p-rank* of G.

Let us start with the structure theorem for finitely generated abelian groups.

Theorem 2.1. *If G is a finitely generated abelian group, then G is a finite direct product of cyclic groups. Moreover, if G has m generators, then every subgroup of G is finitely generated as well and has at most m generators.*

Definition 2.2. An abelian group G is

(a) *torsion* if $t(G) = G$;

(b) *torsion-free* if $t(G) = 0$;

(c) *bounded* if $mG = 0$ for some $m > 0$;

(d) *divisible* if $G = mG$ for every $m > 0$;

(e) *reduced* if the only divisible subgroup of G is the trivial one.

Example 2.3. (a) The groups $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are torsion-free. The class of torsion-free groups is stable under taking direct products and subgroups.

(b) The groups $\mathbb{Z}_m$ $\mathbb{Q}/\mathbb{Z}$ are torsion. The class of torsion groups is stable under taking direct sums, subgroups and quotients.

(c) Let $m_1, m_2, \ldots, m_k > 1$ be naturals and let $\alpha_1, \alpha_2, \ldots, \alpha_k$ be cardinal numbers. Then the group $\bigoplus_{i=1}^{k} \mathbb{Z}_{m_i}^{(\alpha_i)}$ is bounded. According to a theorem of Prüfer every bounded abelian group has this form [46]. This generalizes the Frobenius-Stickelberger theorem about the structure of the finite abelian groups (see Theorem 2.1).

Example 2.4. (a) The groups $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{T}$ are divisible.

(b) For $p \in \mathbb{P}$ we denote be $\mathbb{Z}(p^\infty)$ the *Prüfer group*, namely the p-primary component of the torsion group $\mathbb{Q}/\mathbb{Z}$ (so that $\mathbb{Z}(p^\infty)$ has generators $c_n = 1/p^n + \mathbb{Z}$, $n \in \mathbb{N}$). The group $\mathbb{Z}(p^\infty)$ is divisible.

(c) The class of divisible groups is stable under taking direct products, direct sums and quotients. In particular, every abelian group has a maximal divisible subgroup $d(G)$.

(d) [46] Every divisible group G has the form $(\bigoplus_{r_0(G)} \mathbb{Q}) \oplus (\bigoplus_{p\in\mathbb{P}} \mathbb{Z}(p^\infty)^{(r_p(G))})$.

If X is a set, a set Y of functions of X to a set Z *separates the points* of X if for every $x, y \in X$ with $x \neq y$, there exists $f \in Y$ such that $f(x) \neq f(y)$. Now we see that the characters separate the points of a discrete abelian groups.

Theorem 2.5. *Let G be an abelian group, H a subgroup of G and D a divisible abelian group. Then for every homomorphism $f : H \to D$ there exists a homomorphism $\overline{f} : G \to D$ such that $\overline{f} \restriction_H = f$.*

If $a \in G \setminus H$ and D contains elements of arbitrary finite order, then $\overline{f}$ can be chosen such that $\overline{f}(a) \neq 0$.

Proof. Let H' be a subgroup of G such that $H' \supseteq H$ and suppose that $g : H' \to D$ is such that $g \restriction_H = f$. We prove that for every $x \in G$, defining $N = H' + \langle x \rangle$, there exists $\overline{g} : N \to D$ such that $\overline{g} \restriction_{H'} = g$. There are two cases.

If $\langle x \rangle \cap H' = \{0\}$, then take any $y \in D$ and define $\overline{g}(h + kx) = g(h) + ky$ for every $h \in H'$ and $k \in \mathbb{Z}$. Then $\overline{g}$ is a homomorphism. This definition is correct because every element of N can be represented in a unique way as $h + kx$, where $h \in H'$ and $k \in \mathbb{Z}$.

If $C = \langle x \rangle \cap H' \neq \{0\}$, then C is cyclic, being a subgroup of a cyclic group. So $C = \langle lx \rangle$ for some $l \in \mathbb{Z}$. In particular, $lx \in H'$ and we can consider the element $a = g(lx) \in D$. Since D is divisible, there exists $y \in D$ such that $ly = a$. Now define $\overline{g} : N \to D$ putting $\overline{g}(h + ky) = g(h) + ky$ for every $h + kx \in N$, where $h \in H'$ and $k \in \mathbb{Z}$. To see that this definition is correct, suppose that $h + kx = h' + k'x$ for $h, h' \in H'$ and $k, k' \in \mathbb{Z}$. Then $h - h' = k'x - kx = (k' - k)x \in C$. So $k - k' = sl$ for some $s \in \mathbb{Z}$. Since $g : H' \to D$ is a homomorphism and $lx \in H'$, we have

$$g(h) - g(h') = g(h - h') = g(s(lx)) = sg(lx) = sa = sly = (k' - k)y = k'y - ky.$$

Thus, from $g(h) - g(h') = k'y - ky$ we conclude that $g(h) + ky = g(h') + k'y$. Therefore $\overline{g}$ is correctly defined. Moreover $\overline{g}$ is a homomorphism and extends g.

Let $\mathcal{M}$ be the family of all subgroups H_i of G such that $H \leq H_i$ and of all homomorphisms $f_i : H_i \to D$ that extend $f : H \to D$. For $(H_i, f_i), (H_j, f_j) \in \mathcal{M}$ put $(H_i, f_i) \leq (H_j, f_j)$ if $H_i \leq H_j$ and f_j extends f_i. In this way $(\mathcal{M}, \leq)$ is partially ordered. Let $\{(H_i, f_i)\}_{i\in I}$ a totally ordered subset of $(\mathcal{M}, \leq)$. Then $H_0 = \bigcup_{i\in I} H_i$ is a subgroup of G and $f_0 : H_0 \to D$ defined by $f_0(x) = f_i(x)$ whenever $x \in H_i$, is a homomorphism that extends f_i for every $i \in I$. This proves that $(\mathcal{M}, \leq)$ is inductive and so we can apply Zorn's lemma to find a maximal element $(H_{\max}, f_{\max})$ of $(\mathcal{M}, \leq)$. It is easy to see that $H_{\max} = G$.

Suppose now that D contains elements of arbitrary finite order. If $a \in G \setminus H$, we can extend f to $H + \langle a \rangle$ defining it as in the first part of the proof. If $\langle a \rangle \cap H = \{0\}$ then $\overline{f}(h + ka) = f(h) + ky$ for every $k \in \mathbb{Z}$, where $y \in D \setminus \{0\}$. If $\langle a \rangle \cap H \neq \{0\}$, since D contains elements of arbitrary order, we can choose $y \in D$ such that $\overline{f}(h + ka) = f(h) + ky$ with $y \neq 0$. In both cases $\overline{f}(a) = y \neq 0$. □

Corollary 2.6. *Let G be an abelian group and H a subgroup of G. If $\chi \in \mathrm{Hom}(H, \mathbb{T})$ and $a \in G \setminus H$, then χ can be extended to $\overline{\chi} \in \mathrm{Hom}(G, \mathbb{T})$, with $\overline{\chi}(a) \neq 0$.*

Corollary 2.7. *If G is an abelian group, then* $\mathrm{Hom}(G, \mathbb{T})$ *separates the points of G.*

Corollary 2.8. *If G is an abelian group and D a divisible subgroup of G, then there exists a subgroup B of G such that $G = D \times B$.*

Proof. Consider the homomorphism $f : D \to G$ defined by $f(x) = x$ for every $x \in D$. By Theorem 2.5 we can extend f to $\overline{f} : G \to G$. Then put $B = \ker \overline{f}$ and observe that $G = D + B$ and $D \cap B = \{0\}$; consequently $G \cong D \times B$. □

Corollary 2.9. *Every abelian group G can be written as $G = d(G) \times R$, where RT is a reduced subgroup of G.*

Proof. By Corollary 2.8 there exists a subgroup R of G such that $G = d(G) \times R$. To conclude that R is reduced it suffices to apply the definition of $d(G)$. □

The ring of endomorphisms of the group $\mathbb{Z}(p^\infty)$ will be denoted by $\mathbb{J}_p$, it is isomorphic the inverse limit $\varprojlim \mathbb{Z}/p^n\mathbb{Z}$, known also as the ring of *p-adic integers*. The field of quotients of $\mathbb{J}_p$ (i.e., the field of *p-adic numbers*) will be denoted by $\mathbb{Q}_p$. Sometimes we shall consider only the underlying groups of these rings (and speak of "the group p-adic integers", or "the group p-adic numbers).

2.2 Background on topological spaces

We assume the reader is familiar with the basic definitions and notions related to topological spaces. For the sake of completeness we recall here some frequently used properties related to compactness.

Definition 2.10. *A topological space X is*

- compact *if for every open cover of X there exists a finite subcover;*
- Lindelöff *if for every open cover of X there exists a countable subcover;*
- locally compact *if every point of X has compact neighborhood in X;*
- σ-compact *if X is the union of countably many compact subsets;*
- of first category*, if $X = \bigcup_{n=1}^{\infty} A_n$ and every A_n is a closed subset of X with empty interior;*
- of second category*, if X is not of first category;*
- connected *if for every proper open subset of X with open complement is empty.*

Here we recall properties of maps:

Definition 2.11. For a map $f : (X, \tau) \to (Y, \tau')$ between topological spaces and a point $x \in X$ we say:

- f is *continuous* at x if for every neighborhood U of $f(x)$ in Y there exists a neighborhood V of x in X such that $f(V) \subseteq U$,
- f is *open* in $x \in X$ if for every neighborhood V of x in X there exists a neighborhood U of $f(x)$ in Y such that $f(V) \supseteq U$,
- f is continuous (resp., open) if f is continuous (resp., open) at every point $x \in X$.
- f is *closed* if the subset $f(A)$ of Y is closed for every closed subset $A \subseteq X$.

Some basic properties relating spaces to continuous maps are collected in the next lemma:

Lemma 2.12.
- *If $f : X \to Y$ is a continuous surjective map, then Y is compact (resp., Lindelöff, σ-compact, connected) whenever X has the same property.*
- *If X is a closed subspace of a space Y, then X is compact (resp., Lindelöff, σ-compact, locally compact) whenever Y has the same property.*
- *If $X = \prod_{i \in I} X_i$, then X is compact (resp., connected) iff every space X_i has the same property. If I is finite, the same holds for local compactness and σ-compactness.*

A partially ordered set $(A, \leq)$ is *directed* if for every $\alpha, \beta \in A$ there exists $\gamma \in A$ such that $\gamma \geq \alpha$ and $\gamma \geq \beta$. A subset B of A is *cofinal*, if for every $\alpha \in A$ there exists $\beta \in B$ with $\beta \geq \alpha$.

A *net* in a topological space X is a map from a directed set A to X. We write x_α for the image of $\alpha \in A$ so that the net can be written in the form $N = \{x_\alpha\}_{\alpha \in A}$. A *subnet* of a net N is $S = \{x_\beta\}_{\beta \in B}$ such that B is a cofinal subset of A.

A net $\{x_\alpha\}_{\alpha \in A}$ in X *converges* to $x \in X$ if for every neighborhood U of x in X there exists $\beta \in A$ such that $\alpha \in A$ and $\alpha \geq \beta$ implies $\alpha \in U$.

Lemma 2.13. *Let X be a topological space.*

(a) If Z is a subset of X, then $x \in \overline{Z}$ if and only if there exists a net in Z converging to x.

(b) X is compact if and only if every net in X has a convergent subnet.

(c) A function $f : X \to Y$ (where Y is a topological space) is continuous if and only if $f(x_\alpha) \to f(x)$ in Y for every net $\{x_\alpha\}_{\alpha \in A}$ in X with $x_\alpha \to x$.

(d) The space X is Hausdorff if and only if every net in X converges to at most one point in X.

Let us recall that the *connected component* of a point x in a topological space X is the largest connected subset of X containing x. It is always a closed subset of X. The space X is called *totally disconnected* if all connected components are singletons.

In a topological space X the *quasi-component* of a point $x \in X$ is the intersection of all clopen sets of X containing x.

Lemma 2.14. (Shura-Bura) *In a compact space X the quasi-components and the connected components coincide.*

A topological space X *zero-dimensional* if X has a base of clopen sets. Zero-dimensional T_2 spaces are totally disconnected (as every point is an intersection of clopen sets).

Theorem 2.15. (Vedenissov) *Every totally disconnected locally compact space is zero-dimensional.*

By βX we denote the *Čech-Stone compactification* of a topological Tychonov space X, that is the compact space βX together with the dense immersion $i : X \to \beta X$, such that for every function $f : X \to [0, 1]$ there exists $f^\beta : \beta X \to [0, 1]$ which extends f (this is equivalent to ask that every function of X to a compact space Y can be extended to βX). Here βX will be used only for a discrete space X.

Theorem 2.16 (Baire category theorem). *A Hausdorff locally compact space X is of second category.*

Proof. Suppose that $X = \bigcup_{n=1}^{\infty} A_n$ and assume that every A_n is closed with empty interior. Then the sets $D_n = G \setminus A_n$ are open and dense in X. To get a contradiction, we show that $\bigcap_{n=1}^{\infty} D_n$ is dense, in particular non-empty (so $G \neq \bigcup_{n=1}^{\infty} A_n$, a contradiction).

We use the fact that a Hausdorff locally compact space is regular. Pick an arbitrary open set $V \neq \emptyset$. Then there exists an open set $U_0 \neq \emptyset$ with $\overline{U}_0$ compact and $\overline{U}_0 \subseteq V$. Since D_1 is dense, $U_0 \cap D_1 \neq \emptyset$. Pick $x_1 \in U_0 \cap D_1$ and an open set $U_1 \ni x_1$ in X with $\overline{U}_1$ compact and $\overline{U}_1 \subseteq U_0 \cap D_1$. Proceeding in this way, for every $n \in \mathbb{N}_+$ we can find an open set $U_n \neq \emptyset$ in G with $\overline{U}_n$ compact and $\overline{U}_n \subseteq U_{n-1} \cap D_n$. By the compactness of every $\overline{U}_n$ there exists a point $x \in \bigcap_{n=1}^{\infty} \overline{U}_n$. Obviously, $x \in V \cap \bigcap_{n=1}^{\infty} D_n$. □

Lemma 2.17. *If G is a locally compact σ-compact space, then G is a Lindelöff space.*

Proof. Let $G = \bigcup_{\alpha \in I} U_\alpha$. Since G is σ-compact, $G = \bigcup_{n=1}^{\infty} K_n$ where each K_n is a compact subset of G. Thus for every $n \in \mathbb{N}_+$ there exists a finite subset F_n of I such that $K_n \subseteq \bigcup_{n \in F_n} U_n$. Now $I_0 = \bigcup_{n=1}^{\infty} F_n$ is a countable subset of I and $K_n \subseteq \bigcup_{\alpha \in I_0} U_\alpha$ for every $N \in \mathbb{N}_+$ yields $G = \bigcup_{\alpha \in I_0} U_\alpha$. □

Let X be a topological space. Let $\mathcal{C}(X, \mathbb{C})$ be the $\mathbb{C}$-algebra of all continuous complex valued functions on X. If $f \in \mathcal{C}(X, \mathbb{C})$ let

$$\|f\|_\infty = \sup\{|f(x)| : x \in X\}.$$

Theorem 2.18 (Stone-Weierstraß theorem). *Let X be a compact topological space. A $\mathbb{C}$-subalgebra $\mathcal{A}$ of $\mathcal{C}(X, \mathbb{C})$ containing all constants and closed under conjugation is dense in $\mathcal{C}(X, \mathbb{C})$ for the norm $\| \ \|_\infty$ if and only if $\mathcal{A}$ separates the points of X.*

We shall need in the sequel the following local form of Stone-Weierstraß theorem.

Unit 2

3 General properties of topological groups

3.1 Definition of a topological group

Let us start with the following fundamental concept:

Definition 3.1. Let G be a group.

- A topology τ on G is said to be a *group topology* if the map $f : G \times G \to G$ defined by $f(x, y) = xy^{-1}$ is continuous.
- A *topological group* is a pair (G, τ) of a group G and a group topology τ on G.

If τ is Hausdorff (resp., compact, locally compact, connected, etc.), then the topological group (G, τ) is called Hausdorff (resp., compact, locally compact, connected, etc.). Analogously, if G is cyclic (resp., abelian, nilpotent, etc.) the topological group (G, τ) is called cyclic (resp. abelian, nilpotent, etc.). Obviously, a topology τ on a group G is a group topology iff the maps

$$\mu : G \times G \to G \text{ and } \iota : G \to G$$

defined by $\mu(x, y) = xy$ and $\iota(x) = x^{-1}$ are continuous when $G \times G$ carries the product topology.

Here are some examples, starting with two trivial ones: for every group G the discrete topology and the indiscrete topology on G are group topologies. Non-trivial examples of a topological group are provided by the additive group $\mathbb{R}$ of the reals and by the multiplicative group $\mathbb{S}$ of the complex numbers z with $|z| = 1$, equipped both with their usual topology. This extends to all powers $\mathbb{R}^n$ and $\mathbb{S}^n$. These are abelian topological groups. For every n the linear group $GL_n(\mathbb{R})$ equipped with the topology induced by $\mathbb{R}^{n^2}$ is a non-abelian topological group. The groups $\mathbb{R}^n$ and $GL_n(\mathbb{R})$ are locally compact, while $\mathbb{S}$ is compact.

Example 3.2. For every prime p the group $\mathbb{J}_p$ of p-adic integers carries the topology induced by $\prod_{n=1}^{\infty} \mathbb{Z}(p^n)$, when we consider it as the inverse limit $\varprojlim \mathbb{Z}/p^n\mathbb{Z}$. The same topology can be obtained also when we consider $\mathbb{J}_p$ as the ring of all endomorphims of the group $\mathbb{Z}(p^\infty)$. Now $\mathbb{J}_p$ embeds into the product $\mathbb{Z}(p^\infty)^{\mathbb{Z}(p^\infty)}$ carrying the product topology, while $\mathbb{Z}(p^\infty)$ is discrete. We leave to the reader the verification that this is a compact group topology on $\mathbb{J}_p$. Basic open neighborhoods of 0 in this topology are the subgroups $p^n\mathbb{J}_p$ of $(\mathbb{J}_p, +)$ (actually, these are ideals of the ring $\mathbb{J}_p$) for $n \in \mathbb{N}$. The field $\mathbb{Q}_p$ becomes a locally compact group by declaring $\mathbb{J}_p$ open in $\mathbb{Q}_p$ (i.e., an element $x \in \mathbb{Q}_p$ has as typical neighborhoods the cosets $x + p^n\mathbb{J}_p$, $n \in \mathbb{N}$.

Other examples of group topologies will be given in §3.2.

If G is a topological group written multiplicatively and $a \in G$, then the *translations* $x \mapsto ax$ and $x \mapsto xa$ as well as the *internal automorphism* $x \mapsto axa^{-1}$ are homeomorphisms. Consequently, the group G is discrete iff the point 1 is isolated, i.e., the singleton $\{1\}$ is open. In the sequel aM will denote the image of a subset $M \subseteq G$ under the (left) translation $x \mapsto ax$, i.e., $aM := \{am : m \in M\}$. This notation will be extended also to families of subsets of G, in particular, for every filter $\mathcal{F}$ we denote by $a\mathcal{F}$ the filter $\{aF : F \in \mathcal{F}\}$.

Making use of the homeomorphisms $x \mapsto ax$ one can prove:

Exercise 3.3. *Let $f : G \to H$ be a homomorphism between topological groups. Prove that f is continuous (resp., open) iff f is continuous (resp., open) at $1 \in G$.*

For a topological group G and $g \in G$ we denote by $\mathcal{V}_{G,\tau}(g)$ the filter of all neighborhoods of the element g of G. When no confusion is possible, we shall write briefly also $\mathcal{V}_G(g)$, $\mathcal{V}_\tau(g)$ or even $\mathcal{V}(g)$. Among these filters the filter $\mathcal{V}_{G,\tau}(1)$, obtained for the neutral element $g = 1$, plays a central role. It is useful to note that for every $a \in G$ the filter $\mathcal{V}_G(a)$ coincides with $a\mathcal{V}_G(1) = \mathcal{V}_G(1)a$. More precisely, we have the following:

Theorem 3.4. *Let G be a group and let $\mathcal{V}(1)$ be the filter of all neighborhoods of 1 in some group topology τ on G. Then:*

(a) *for every $U \in \mathcal{V}(1)$ there exists $V \in \mathcal{V}(1)$ with $V \cdot V \subseteq U$;*

(b) *for every $U \in \mathcal{V}(1)$ there exists $V \in \mathcal{V}(1)$ with $V^{-1} \subseteq U$;*

(c) *for every $U \in \mathcal{V}(1)$ and for every $a \in G$ there exists $V \in \mathcal{V}(1)$ with $aVa^{-1} \subseteq U$.*

Conversely, if $\mathcal{V}$ is a filter on G satisfying (a), (b) and (c), then there exists a unique group topology τ on G such that $\mathcal{V}$ coincides with the filter of all τ-neighborhoods of 1 in G.

Proof. To prove (a) it suffices to apply the definition of the continuity of the multiplication $\mu : G \times G \to G$ at $(1, 1) \in G \times G$. Analogously, for (b) use the continuity of the map $\iota : G \to G$ at $1 \in G$. For item (c) use the continuity of the internal automorphism $x \mapsto axa^{-1}$ at $1 \in G$.

Let $\mathcal{V}$ be a filter on G satisfying all conditions (a), (b) and (c). Let us see first that every $U \in \mathcal{V}$ contains 1. In fact, take $W \in \mathcal{V}$ with $W \cdot W \subseteq U$ and choose $V \in \mathcal{V}(1)$ with $V \subseteq W$ and $V^{-1} \subseteq W$. Then $1 \in V \cdot V^{-1} \subseteq U$.

Now define a topology τ on G whose open sets O are defined by the following property:

$$\tau := \{O \subseteq G : (\forall a \in O)(\exists U \in \mathcal{V}) \text{ such that } aU \subseteq O\}.$$

It is easy to see that τ is a topology on G. Let us see now that for every $g \in G$ the filter $g\mathcal{V}$ coincides with the filter $\mathcal{V}_{(G,\tau)}(g)$ of all τ-neighborhoods of g in (G, τ). The inclusion $g\mathcal{V} \supseteq \mathcal{V}_{(G,\tau)}(g)$ is obvious. Assume $U \in \mathcal{V}$. To see that $gU \in \mathcal{V}_{(G,\tau)}(g)$ we have to find a τ-open $O \subseteq gU$ that contains g. Let $O := \{h \in gU : (\exists W \in \mathcal{V})\ hW \subseteq gU\}$. Obviously $g \in O$. To see that $O \in \tau$ pick $x \in O$. Then there exists $W \in \mathcal{V}$ with $xW \subseteq gU$. Let $V \in \mathcal{V}$ with $V \cdot V \subseteq W$, then $xV \subseteq O$ since $xvV \subseteq gU$ for every $v \in V$.

We have seen that τ is a topology on G such that the τ-neighborhoods of any $x \in G$ are given by the filter $x\mathcal{V}$. It remains to see that τ is a group topology. To this end we have to prove that the map $(x, y) \mapsto xy^{-1}$ is continuous. Fix x, y and pick a $U \in \mathcal{V}$. By (c) there exists a $W \in \mathcal{V}$ with $Wy^{-1} \subseteq y^{-1}U$. Now choose $V \in \mathcal{V}$ with $V \cdot V^{-1} \subseteq W$. Then $O = xV \times yV$ is a neighborhood of (x, y) in $G \times G$ and $f(O) \subseteq xV \cdot V^{-1}y^{-1} \subseteq xWy^{-1} \subseteq xy^{-1}U$. □

In the above theorem one can take instead of a filter $\mathcal{V}$ also a *filter base*, i.e., a family $\mathcal{V}$ with the property

$$(\forall U \in \mathcal{V})(\forall V \in \mathcal{V})(\exists W \in \mathcal{V})W \subseteq U \cap V$$

beyond the proprieties (a)–(c).

A neighborhood $U \in \mathcal{V}(1)$ is *symmetric*, if $U = U^{-1}$. Obviously, for every $U \in \mathcal{V}(1)$ the intersection $U \cap U^{-1} \in \mathcal{V}(1)$ is a symmetric neighborhood, hence every neighborhood of 1 contains a symmetric one.

Let $\{\tau_i : i \in I\}$ be a family of group topologies on a group G. Then their supremum $\tau = \sup_{i\in I} \tau_i$ is a group topology on G with a base of neighborhoods of 1 formed by the family of all finite intersection $U_1 \cap U_2 \cap \ldots \cap U_n$, where $U_k \in \mathcal{V}_{\tau_{i_k}}(1)$ for $k = 1, 2, \ldots, n$ and the n-tuple $i_1, i_2, \ldots, i_n$ runs over all finite subsets if I.

Exercise 3.5. *If (a_n) is a sequence in G such that $a_n \to 1$ for every member τ_i of a family $\{\tau_i : i \in I\}$ of group topologies on a group G, then $a_n \to 1$ also for the supremum $\sup_{i\in I} \tau_i$.*

3.2 Examples of group topologies

Now we give several series of examples of group topologies, introducing them by means of the filter $\mathcal{V}(1)$ of neighborhoods of 1 as explained above. However, in all cases we avoid the treat the whole filter $\mathcal{V}(1)$ and we prefer to deal with an essential part of it, namely a base. Let us recall the precise definition of a base of neighborhoods.

Definition 3.6. Let G be a topological group. A family $\mathcal{B} \subseteq \mathcal{V}(1)$ is said to be a *base of neighborhoods of* 1 (or briefly, a *base at 1*) if for every $U \in \mathcal{V}(1)$ there exists a $V \in \mathcal{B}$ contained in U (such a family will necessarily be a filterbase).

3.2.1 Linear topologies

Let $\mathcal{V} = \{N_i : i \in I\}$ be a filter base consisting of normal subgroups of a group G. Then $\mathcal{V}$ satisfies (a)–(c), hence generates a group topology on G having as basic neighborhoods of a point $g \in G$ the family of cosets $\{gN_i : i \in I\}$. Group topologies of this type will be called *linear topologies*. Let us see now various examples of linear topologies.

Example 3.7. Let G be a group and let p be a prime:

- the *pro-finite* topology, with $\{N_i : i \in I\}$ all normal subgroups of finite index of G;
- the *pro-p-finite* topology, with $\{N_i : i \in I\}$ all normal subgroups of G of finite index that is a power of p;
- the *p-adic* topology, with $I = \mathbb{N}$ and for $n \in \mathbb{N}$, N_n is the subgroup (necessarily normal) of G generated by all powers $\{g^{p^n} : g \in G\}$.
- the *natural* topology (or $\mathbb{Z}$-*topology*), with $I = \mathbb{N}$ and for $n \in \mathbb{N}$, N_n is the subgroup (necessarily normal) of G generated by all powers $\{g^n : g \in G\}$.
- the *pro-countable* topology, with $\{N_i : i \in I\}$ all normal subgroups of at most countable index $[G : N_i]$.

The next simple construction belongs to Taimanov. Now neighborhoods of 1 are subgroups, that are not necessarily normal.

Exercise 3.8. *Let G be a group with trivial center. Then G can be considered as a subgroup of $Aut\,(G)$ making use of the internal automorphisms. Identify $Aut\,(G)$ with a subgroup of the power G^G and equip $Aut\,(G)$ with the group topology τ induced by the product topology of G^G, where G carries the discrete topology. Prove that:*

- *the filter of all τ-neighborhoods of 1 has as base the family of centralizers $\{c_G(F)\}$, where F runs over all finite subsets of G;*
- *τ is Hausdorff;*
- *τ is discrete iff there exists a finite subset of G with trivial centralizer.*

3.2.2 Topologies generated by characters

Let G be an abelian group. A *character* of G is a homomorphism $\chi : G \to \mathbb{S}$. For characters χ_i, $i = 1, \dots, n$, of G and $\delta > 0$ let

$$U_G(\chi_1, \dots, \chi_n; \delta) := \{x \in G : |\mathrm{Arg}\ (\chi_i(x))| < \delta, i = 1, \dots, n\}, \tag{1}$$

where the argument Arg (z) of a complex number z is taken in $(-\pi, \pi]$.

Exercise 3.9. *Let G be an abelian group and let H be a family of characters of G. Then the family*

$$\{U_G(\chi_1, \dots, \chi_n; \delta) : \delta > 0, \chi_i \in H, i = 1, \dots, n\}$$

is a filter base satisfying the conditions (a)–(c) of Theorem 3.4, hence it gives rise to a group topology $\mathcal{T}_H$ on G (this is the initial topology of the family H, i.e., the coarsest topology that makes continuous all the characters of H).

We refer to the group topology $\mathcal{T}_H$ as *topology generated by the characters* of H. The topology $\mathcal{T}_{G^*}$, generated by all characters of G, is called *Bohr topology* of G.

For an abelian group G some of the linear topologies on G are also generated by appropriate families of characters.

Exercise 3.10. *Let G be an abelian group.*

1. *Prove that the profinite topology of G is contained in the Bohr topology of G. Give an example of a group G where these two topologies differ.*
2. *Let H be the family of all characters χ of G such that the subgroup $\chi(G)$ is finite. Prove that the topology $\mathcal{T}_H$ coincides with the pro-finite topology on G.*
3. *Let H be the family of all characters χ of G such that the subgroup $\chi(G)$ is finite and contained in the subgroup $\mathbb{Z}(p^\infty)$ of $\mathbb{T}$. Prove that the topology $\mathcal{T}_H$ coincides with the pro-p-finite topology on G.*

This exercise suggests to call a character $\chi : G \to \mathbb{T}$ *torsion* is there exists $n > 0$ such that χ vanishes on the subgroup $nG := \{nx : x \in G\}$. (Equivalently, the character $n \cdot \chi$ coincides with the trivial character, where the character $n \cdot \chi : G \to \mathbb{T}$ is defined by $(n \cdot \chi)(x) := n\chi(x)$.)

Exercise 3.11. *Let G be an abelian group. Prove that:*

1. *if H is a family of characters of G, then the topology $\mathcal{T}_H$ is contained in the pro-finite topology of G iff every character of H is torsion.*
2. *if G is bounded, then the Bohr topology of G coincides with the profinite topology of G.*
3. *if the Bohr topology of G coincides with the profinite topology of G, then G is bounded.*

3.2.3 Pseudonorms and pseudometrics in a group

According to Markov a *pseudonorm* in an abelian group G is a map $\nu : G \to \mathbb{R}_+$ such that for every $x, y \in G$:

(1) $\nu(1) = 0$;

(2) $\nu(x^{-1}) = \nu(x)$;

(3) $\nu(xy) \leq \nu(x) + \nu(y)$.

The norms defined in a real vector space are obviously pseudonorms (with the additional property, in additive notation, $\nu(0) = 0$ iff $x = 0$).

Every pseudonorm ν generates a pseudometric d_ν on G defined by $d_\nu(x, y) := \nu(x^{-1}y)$. This pseudometric is *left invariant* in the sense that $d_\nu(ax, ay) = d_\nu(x, y)$ for every $a, x, y \in G$. Denote by τ_ν the topology induced on G by this pseudometric. A base of $\mathcal{V}_{\tau_\nu}(1)$ is given by the open balls $\{B_{1/n}(1) : n \in \mathbb{N}_+\}$.

In order to build metrics inducing the topology of a given topological group (G, τ) we need the following lemma (for a proof see [67, 8.2], [79]). We say that a pseudometric d on G is *continuous* if the map $d : G \times G \to \mathbb{R}_+$ is continuous. This is equivalent to have the topology induced by the metric d coarser than the topology τ (i.e., every open set with respect to the metric d is τ-open).

Lemma 3.12. *Let G be a topological group and let*

$$U_0 \supseteq U_1 \supseteq \ldots \supseteq U_n \supseteq \ldots \tag{2}$$

be symmetric neighborhoods of 1 with $U_n^3 \subseteq U_{n-1}$ for every $n \in \mathbb{N}$. Then there exists a continuous left invariant pseudometric d on G such that $U_n \subseteq B_{1/n}(1) \subseteq U_{n-1}$ for every n.

Exercise 3.13. *Prove that in the previous lemma $H = \bigcap_{n=1}^{\infty} U_n$ is a closed subgroup of G with the property $H = \{x \in G : d(x, 1) = 0\}$. In particular, d is a metric iff $H = \{1\}$.*

If the chain (2) has also the property $xU_nx^{-1} \subseteq U_{n-1}$ for every $x \in G$ and for every n, the subgroup H is normal and d defines a metric on the quotient group letting $\tilde{d}(xH, yH) := d(x, y)$. The metric $\tilde{d}$ induces the quotient topology on G/H.

3.2.4 Permutation groups

Let X be an infinite set and let G briefly denotes the group $S(X)$ of all permutations of X. A very natural topology on G is defined by taking as filter of neighborhoods of $1 = id_X$ the family of all subgroups of G of the form

$$S_F = \{f \in G : (\forall x \in F)\ f(x) = x\},$$

where F is a finite subset of X.

This topology can be described also as the topology induced by the natural embedding of G into the Cartesian power X^X equipped with the product topology, where X has the discrete topology.

This topology is also the point-wise convergence topology on G. Namely, if $(f_i)_{i \in I}$ is a net in G, then f_i converges to $f \in G$ precisely when for every $x \in X$ there exists an $i_0 \in I$ such that for all $i \geq i_0$ in I one has $f_i(x) = f(x)$.

Exercise 3.14. *If $S_\omega(X)$ denotes the subset of all permutations of finite support in $S(X)$ prove that S_ω is a dense normal subgroup of G.*

Exercise 3.15. *Prove that $S(X)$ has no proper closed normal subgroups.*

3.3 Subgroups and quotients of topological groups

Let G be a topological group and let H be a subgroup of G. Then H becomes a topological group when endowed with the topology induced by G. Sometimes we refer to this situation by saying H is a topological subgroup of G.

Let G and H be topological groups and let $f : G \to H$ be a continuous homomorphism. If f is simultaneously an isomorphism and a homeomorphism, then f is called a *topological isomorphism*. If $f : G \to f(G) \subseteq H$ is a topological isomorphism, where $f(G)$ carries the topology induced by H, then f is called *topological group embedding*, or shortly *embedding*.

Proposition 3.16. *Let G be a topological group and let H be a subgroup of G. Then:*

(a) *H is open in G iff H has a non-empty interior;*

(b) *if H is open, then H is also closed;*

(c) *if H is discrete and G is T_1, then H is closed.*

Proof. (a) Let $\emptyset \neq V \subseteq H$ be an open set and let $h_0 \in V$. Then $1 \in h_0^{-1}V \subseteq H = h_0^{-1}H$. Now $U = h_0^{-1}V$ is open, contains 1 and $h \in hU \subseteq H$ for every $h \in H$. Therefore H is open.

(b) If H is open then every coset gH is open and consequently the complement $G \setminus H$ is open. So H is closed.

(c) Since H is discrete there exists $U \in \mathcal{V}(1)$ with $U \cap H = \{1\}$. Choose $V \in \mathcal{V}(1)$ with $V^{-1} \cdot V \subseteq U$. Then $|xV \cap H| \leq 1$ for every $x \in G$, as $h_1 = xv_1 \in xV \cap H$ and $h_2 = xv_2 \in xV \cap H$ give $h_1^{-1}h_2 \in V^{-1} \cdot V \cap H = \{1\}$, hence $h_1 = h_2$. Therefore, if $x \notin H$ one can find a neighborhood $W \subseteq xV$ of x with $W \cap H = \emptyset$, i.e., $x \notin \overline{H}$. Indeed, if $xV \cap H = \emptyset$, just take $W = xV$. In case $xV \cap H = \{h\}$ for some $h \in H$, one has $h \neq x$ as $x \notin H$. Then $W = xV \setminus \{x\}$ is the desired neighborhood of x. □

Exercise 3.17. *Let H be a discrete non-trivial group and let $G = H \times N$, where N is an indiscrete non-trivial group. Prove that $H \times \{1\}$ is a discrete non-closed subgroup of G.*

Let us see now how the *closure* $\overline{H}$ of a subset H of a topological group G can be computed.

Lemma 3.18. *Let H be a subset of G. Then with $\mathcal{V} = \mathcal{V}(1)$ one has*

(a) $\overline{H} = \bigcap_{U \in \mathcal{V}} UH = \bigcap_{U \in \mathcal{V}} HU = \bigcap_{U,V \in \mathcal{V}} UHV$;

(b) *if H is a subgroup of G, then $\overline{H}$ is a subgroup of G; if H a normal subgroup, then also $\overline{H}$ is normal subgroup;*

(c) $N = \overline{\{1\}}$ *is a closed normal subgroup.*

Proof. (a) For $x \in G$ one has $x \notin \overline{H}$ iff there exists $U \in \mathcal{V}$ such that $xU \cap H = \emptyset = Ux \cap H$. Pick a symmetric U, i.e., $U = U^{-1}$. Then the latter property is equivalent to $x \notin UH \cup HU$. This proves $\overline{H} = \bigcap_{U \in \mathcal{V}} UH = \bigcap_{U \in \mathcal{V}} HU$. To prove the last equality in (a) note that the already established equalities yield

$$\bigcap_{U,V \in \mathcal{V}} UHV = \bigcap_{U \in \mathcal{V}} (\bigcap_{V \in \mathcal{V}} UHV) = \bigcap_{U \in \mathcal{V}} \overline{UH} \subseteq \bigcap_{U \in \mathcal{V}} U^2H = \bigcap_{W \in \mathcal{V}} WH = \overline{H}.$$

(b) Let $x, y \in \overline{H}$. According to (a), to verify $xy \in \overline{H}$ it suffices to see that $xy \in UHU$ for every $U \in \mathcal{V}$. This follows from $x \in UH$ and $y \in HU$ for every $U \in \mathcal{V}$. If H is normal, then for every $a \in G$ and for $U \in \mathcal{V}$ there exists a symmetric $V \in \mathcal{V}$ with $aV \subseteq Ua$ and $Va^{-1} \subseteq a^{-1}U$. Now for every $x \in \overline{H}$ one has $x \in VHV^{-1}$, hence $axa^{-1} \in aVHV^{-1}a^{-1} \subseteq UaHa^{-1}U \subseteq UHU$. This proves $axa^{-1} \in \overline{H}$ according to (a).

(c) follows from (b) with $H = \{1\}$. □

Exercise 3.19. *Prove that:*

- *the subgroup $H \times \{1\}$ from Exercise 3.17 of G is dense.*
- *for every infinite set X and every group topology on the permutation group $S(X)$ the subgroups $S_x = \{f \in S(X) : f(x) = x\}$, $x \in X$, are either closed or dense. (Hint. Prove that S_x is a maximal subgroup of $S(X)$, see Fact 3.56.)*

Exercise 3.20. *Prove that every proper closed subgroup of $\mathbb{R}$ is cyclic.*

(Hint. If H is a proper closed non-trivial subgroup of $\mathbb{R}$ prove that the set $\{h \in H : h > 0\}$ has a greatest lower bound h_0 and conclude that $H = \langle h_0 \rangle$.)

Let G be a topological group and H a normal subgroup of G. Consider the quotient G/H with the quotient topology, namely the finest topology on G/H that makes the canonical projection $q : G \to G/H$ continuous. Since we have a group topology on G, the quotient topology consists of all sets $q(U)$, where U runs over the family of all open sets of G (as $q^{-1}(q(U))$ is open in G in such a case). In particular, the canonical projection q is open.

The next theorem is due to Frobenius.

Theorem 3.21. *If G and H are topological groups, $f : G \to H$ is a continuous surjective homomorphism and $q : G \to G/\ker f$ is the canonical homomorphism, then the unique homomorphism $f_1 : G/\ker f \to H$, such that $f = f_1 \circ q$, is a continuous isomorphism. Moreover, f_1 is a topological isomorphism iff f is open.*

Proof. Follows immediately from the definitions of quotient topology and open map. □

Independently on its simplicity, this theorem is very important since it produces topological isomorphisms. Openness of the map f is its main ingredient, so from now on we shall be interested in providing conditions that ensure openness (see also §4.1).

Lemma 3.22. *Let X, Y be topological spaces and let $\varphi : X \to Y$ be a continuous open map. Then for every subspace P of Y with $P \cap \varphi(X) \neq \emptyset$ the restriction $\psi : H_1 \to P$ of the map φ to the subspace $H_1 = \varphi^{-1}(P)$ is open.*

Proof. To see that ψ is open choose a point $x \in H_1$ and a neighborhood U of x in H_1. Then there exists a neighborhood W of x in X such that $U = H_1 \cap W$. To see that $\psi(U)$ is a neighborhood of $\psi(x)$ in P it suffices to note that if $\varphi(w) \in P$ for $w \in W$, then $w \in H_1$, hence $w \in H_1 \cap W = U$. Therefore $\varphi(W) \cap P \subseteq \varphi(U) = \psi(U)$. □

We shall apply this lemma when $X = G$ and $Y = H$ are topological group and $\varphi = q : G \to H$ is a continuous open homomorphism. Then the restriction $q^{-1}(P) \to P$ of q is open for every subgroup P of H. Nevertheless, even in the particular case when q is surjective, the restriction $H_1 \to \varphi(H_1)$ of q to an arbitrary closed subgroup H_1 of G need not be open.

In the next theorem we see some isomorphisms related the quotient groups.

Teoema 3.23. *Let G be a topological group, let N be a normal closed subgroup of G and let $p : G \to G/N$ be the canonical homomorphism.*

(a) *If H is a subgroup of G, then the homomorphism $i : HN/N \to p(H)$, defined by $i(xN) = p(x)$, is a topological isomorphism.*

(b) *If H is a closed normal subgroup of G with $N \subseteq H$, then $p(H) = H/N$ is a closed normal subgroup of G/N and the map $j : G/H \to (G/N)/(H/N)$, defined by $j(xH) = (xN).(H/N)$, is a topological isomorphism.*

(Both in (a) and (b) the quotient groups are equipped with the quotient topology.)

Proof. (a) As $HN = p^{-1}(p(H))$ we can apply Lemma 3.22 and conclude that p' is an open map. Now Theorem 3.21 applies to the restriction $p' : HN \to p(H)$ of p.

(b) Since $H = HN$, item (a) implies that the induced topology of $p(H)$ coincides with the quotient topology of H/N. Hence we can identify H/N with the topological subgroup $p(H)$ of G/N. Since $H = HN$, the set $(G/N) \setminus p(HN) = p(G \setminus HN)$ is open, hence $p(H)$ is closed. Finally note that the composition $f : G \to (G/N)/(H/N)$ of p with the canonical homomorphism $G/N \to (G/N)/(H/N)$ is open, being the latter open. Applying to the open homomorphism f with $\ker f = H$ Theorem 3.21 we can conclude that j is a topological isomorphism. □

Exercise 3.24. *Let G be an abelian group equipped with its Bohr topology and let H be a subgroup of G. Prove that:*

- *H is closed in G;*
- *the topological subgroup topology of H coincides with its Bohr topology;*
- *the quotient topology of G/H coincides with the Bohr topology of G/H.*
- ** G has no convergent sequences [36, §3.4].*

Exercise 3.25. *Let H be a discrete subgroup of a topological group G. Prove that:*

- $H \cap \overline{\{1\}} = \{1\}$;
- $\overline{H}$ *is isomorphic to the semi-direct product of* H *and* $\overline{\{1\}}$, *carrying the product topology, where* H *is discrete and* $\overline{\{1\}}$ *is indiscrete.*

3.4 Separation axioms

Lemma 3.18 easily implies that every topological group is regular, hence:

Proposition 3.26. *For a topological group G the following are equivalent:*

(a) G *is Hausdorff;*

(b) G *is* T_0.

(c) G *is* T_3 *(where* T_3 *stands for "regular and* T_1*").*

(d) $\overline{\{1\}} = \{1\}$.

A topological group G is *monothetic* if there exists $x \in G$ with $\langle x \rangle$ dense in G.

Exercise 3.27. *Prove that:*

- *a Hausdorff monothetic group is necessarily abelian.*
- $\mathbb{T}$ *is monothetic.*

Is $\mathbb{T}^2$ *monothetic? What about* $\mathbb{T}^{\mathbb{N}}$*?*

Now we relate proprieties of the quotient G/H to those of the subgroup H of G.

Lemma 3.28. *Let G be a topological group and let H be a normal subgroup of G. Then:*

(1) the quotient G/H is discrete if and only if H is open;

(2) the quotient G/H is Hausdorff if and only if H is closed.

Let us see now that every T_0 topological group is also a Tychonov space.

Theorem 3.29. *Every Hausdorff topological group is a Tychonov space.*

Proof. Let F be a closed set with $1 \notin F$. Then we can find a chain (2) of open neighborhoods of 1 as in Lemma 3.12 such that $F \cap U_0 = \emptyset$. Let d be the pseudometric defined in Lemma 3.12 and let $f_F(x) = d(x, F)$ be the distance function from F. This function is continuous in the topology induced by the pseudometric. By the continuity of d it will be continuous also with respect to the topology of G. It suffices to note now that $f_F(F) = 0$, while $f_F(1) = 1$. This proves that the space G is Tychonov, as the pseudometric is left invariant, so the same argument provides separation of a generic point $a \in G$ from a closed set F that does not contain a. □

Let G be an abelian group and let H be a family of characters of G. Then the characters of H separate the points of G iff for every $x \in G$, $x \neq 0$, there exists a character $\chi \in H$ with $\chi(x) \neq 1$.

Exercise 3.30. *Let G be an abelian group and let H be a family of characters of G. Prove that the topology $\mathcal{T}_H$ is Hausdorff iff the characters of H separate the points of G.*

Proposition 3.31. *Let G be an infinite abelian group and let $H = Hom(G, \mathbb{S})$. Then the following holds true:*

(a) *the characters of H separate the points of G,*

(b) *the Bohr topology $\mathcal{T}_H$ is Hausdorff and non-discrete.*

Proof. (a) This is Corollary 2.7.

(b) According to Exercise 3.30 item (a) implies that the topology $\mathcal{T}_H$ is Hausdorff. Suppose, for a contradiction, that $\mathcal{T}_H$ is discrete. Then there exist $\chi_i \in H, i = 1, \ldots, n$ and $\delta > 0$ such that $U(\chi_1, \ldots, \chi_n; \delta) = \{0\}$. In particular, $H = \bigcap_{i=1}^{n} \ker \chi_i = \{0\}$. Hence the diagonal homomorphism $f = \chi_1 \times \ldots \times \chi_n : G \to \mathbb{S}^n$ is injective and $f(G) \cong G$ is an infinite discrete subgroup of $\mathbb{S}^n$. According to Proposition 3.16 $f(G)$ is closed in $\mathbb{S}^n$ and consequently, compact. The compact discrete spaces are finite, a contradiction. □

Most often the topological groups in the sequel will be assumed to be Hausdorff.

Example 3.32. Contrary to what we proved in Theorem 3.29 Hausdorff topological groups need not be normal as topological spaces (see Exercise 3.37). A nice "uniform" counter-example to this was given by Trigos: for every uncountable group G the topological group $G^{\#}$ is not normal as a topological space (countable groups are ruled out since every every countable Hausdorff topological group is normal, being a regular Lindelöff space).

Theorem 3.33. (Birkhoff-Kakutani) *A topological group is metrizable iff it has a countable base of neighborhoods of* 1.

Proof. The necessity is obvious as every point x in a metric space has a countable base of neighborhoods. Suppose now that G has countable base of neighborhoods of 1. Then one can build a chain (2) of neighborhoods of 1 as in Lemma 3.12 that form a base of $\mathcal{V}(1)$, in particular, $\bigcap_{n=1}^{\infty} U_n = \{1\}$. Then the pseudometric produced by the lemma is a metric that induces the topology of the group G because of the inclusions $U_n \subseteq B_{1/n} \subseteq U_{n-1}$. □

Exercise 3.34. *Prove that subgroups and quotients of metrizable topological groups are metrizable.*

Exercise 3.35. *Prove that every topological abelian group admits a continuous isomorphism into a product of metrizable abelian groups.*

[*Hint.* For $x \in G$, $x \neq 0$ choose an open neighborhood U of 0 with $x \in U$. Build a sequence $\{U_n\}$ of symmetric open neighborhoods of 0 with $U_0 \subseteq U$ and $U_n + U_n \subseteq U_{n-1}$. Then $H_U = \bigcap_{n=1}^{\infty} U_n$ is a closed subgroup of G .Let τ_U be the group topology on the quotient G/H_U having as a local base at 0 the family $\{f_U(U_n)\}$, where $f_U : F \to G/H_U$ is the canonical homomorphism. Show that $(G/H, \tau_U)$ is metrizable. Now take the product of all groups $(G/H, \tau_U)$. To conclude observe that the diagonal map of the family f_U into the product of all groups $(G/H, \tau_U)$ is continuous and injective.]

Exercise 3.36. *Let G be a Hausdorff topological group. Prove that the centralizer of an element $g \in G$ is a closed subgroup. In particular, the center $Z(G)$ is a closed subgroup of G.*

Exercise 3.37. * *The group $\mathbb{Z}^{\aleph_1}$ equipped with the Tychonov topology (where $\mathbb{Z}$ is discrete) is not a normal space [67].*

Furstenberg used the natural topology ν of $\mathbb{Z}$ (see Example 3.7) to find a new proof of the infinitude of prime numbers.

Exercise 3.38. *Prove that there are infinitely many primes in $\mathbb{Z}$ using the natural topology ν of $\mathbb{Z}$.*

(Hint. If $p_1, p_2, \ldots, p_n$ were the only primes, then consider the union of the open subgroups $p_1\mathbb{Z}, \ldots, p_n\mathbb{Z}$ and use the fact that every integer $n \neq 0, \pm 1$ has a prime divisor, so belongs to $\bigcup_{i=1}^{n} p_i\mathbb{Z}$.)

3.5 Connectedness in topological groups

For a topological group G we denote by $c(G)$ the connected component of 1 and we call it briefly *connected component of G*.

Before proving some basic facts about the connected component, we need an elementary property of the connected sets in a topological groups.

Lemma 3.39. *Let G be a topological group.*

(a) If $C_1, C_2, \ldots, C_n$ are connected sets in G, then also $C_1C_2 \ldots C_n$ is connected.

(b) If C is a connected set in G, then the set C^{-1} as well as the subgroup generated by C are connected.

Proof. (a) Let us conisder the case $n = 2$, the general case easily follows from this one by induction. The subset $C_1 \times C_2$ of $G \times G$ is connected. Now the map $\mu : G \times G \to G$ defined $\mu(x, y) = xy$ is continuous and $\mu(C_1 \times C_2) = C_1C_2$.

(a) For the first part it suffices to note that C^{-1} is a continuous image of C under the continuous map $x \mapsto x^{-1}$.

To prove the second assertion consider the set $C_1 = CC^{-1}$. It is connected by the previous lemma and obviously $1 \in C_1$. Moreover, $C_1^2 \supseteq C \cup C^{-1}$. It remains to note now that the subgroup generated by C_1 coincides with the subgroup generated by C. Since the former is the union of all sets C_1^n, $n \in \mathbb{N}$ and each set C_1^n is connected by item (a), we are done. □

Proposition 3.40. *The connected component $c(G)$ a topological group G is a closed normal subgroup of G. The connected component of an element $x \in G$ is simply the coset $xc(G) = c(G)x$.*

Proof. To prove that $c(G)$ is stable under multiplication it suffices to note that $c(G)c(G)$ is still connected (applying item (a) of the above lemma) and contains 1, so must be contained in the connected component $c(G)$. Similarly, an application of item (b) implies that $c(G)$ is stable also w.r.t. the operation $x \mapsto x^{-1}$, so $c(G)$ is a subgroup of G. Moreover, for every $a \in G$ the image $ac(G)a^{-1}$ under the conjugation is connected and contains 1, so must be contains in the connected component $c(G)$. So $c(G)$ is stable also under conjugation. Therefore $c(G)$ is a normal subgroup. The fact that $c(G)$ is closed is well known.

To prove the last assertion it suffices to recall that the maps $y \mapsto xy$ and $y \mapsto yx$ are homeomorphisms. □

Our next aim is to see that the quotient $G/c(G)$ is totally disconnected. We need first to see that connectedness and total connectedness are properties stable under extension:

Proposition 3.41. *Let G be a topological group and let N be a closed normal subgroup of G.*

(a) If both N and G/N are connected, then also G is connected.

(b) If both N and G/N are totally disconnected, then also G is totally disconnected.

Proof. Let $q : G \to G/N$ be the canonical homomorphism.

(a) Let $A \neq \emptyset$ be a clopen set of G. As every coset aN is connected, one has either $aN \subseteq A$ or $aN \cap A = \emptyset$. Hence, $A = q^{-1}(q(A))$. This implies that $q(A)$ is a non-empty clopen set of the connected group G/N. Thus $q(A) = G/N$. Consequently $A = G$.

(b) Assume C is a connected set in G. Then $q(C)$ is a connected set of G/N, so by our hypothesis, $q(C)$ is a singleton. This means that C is contained in some coset xN. Since xN is totally disconnected as well, we conclude that C is a singleton. This proves that G is totally disconnected. □

Lemma 3.42. *If G is a topological group, then the group $G/c(G)$ is totally disconnected.*

Proof. Let $q : G \to G/c(G)$ be the canonical homomorphism and let H be the inverse image of $c(G/c(G))$ under q. Now apply Proposition 3.41 to the group H and the quotient group $H/c(G) \cong c(G/c(G))$ to conclude that H is connected. Since it contains $c(G)$, we have $H = c(G)$. Hence $G/c(G)$ is totally disconnected. □

For a topological group G denote by $Q(G)$ the quasi-component of the neutral element 1 of G (i.e., the intersection of all clopen sets of G containing 1) and call it *quasi-component* of G.

Proposition 3.43. *For a topological group G the quasi-component $Q(G)$ is a closed normal subgroup of G. The quasi-component of $x \in G$ coincides with the coset $xQ(G) = Q(G)x$.*

Proof. Let $x, y \in Q(G)$. To prove that $xy \in Q(G)$ we need to verify that $xy \in O$ for every clopen set O containing 1. Let O be such a set, then $x, y \in O$. Obviously Oy^{-1} is a clopen set containing 1, hence $x \in Oy^{-1}$. This implies $xy \in O$. Hence $Q(G)$ is stable under multiplication. For every clopen set O containing 1 the set O^{-1} has the same propriety, hence $Q(G)$ is stable also w.r.t. the operation $a \mapsto a^{-1}$. This implies that $Q(G)$ is a subgroup. Moreover, for every $a \in G$ and for every clopen set O containing 1 also its image aOa^{-1} under the conjugation is a clopen set containing 1. So $Q(G)$ is stable also under conjugation. Therefore $Q(G)$ is a normal subgroup. Finally, as an intersection of closen sets, $Q(G)$ is closed. □

Remark 3.44. It follows from Lemma 2.14 that $c(G) = Q(G)$ for every compact topological group G. Actually, this remains true also in the case of locally compact groups G (cf. 4.22).

In the next remark we discuss zero-dimensionality.

Remark 3.45. (a) It follows immediately from Proposition 3.16 that every linear group topology is zero-dimensional; in particular, totally disconnected.

(b) Every countable Hausdorff topological group is zero-dimensional (this is true for topological spaces as well).

We shall see in the sequel that for locally compact abelian groups or compact groups the implication from item (a) can be inverted (see Theorem 4.18). On the other hand, the next example shows that local connectedness is essential.

Example 3.46. The group $\mathbb{Q}/\mathbb{Z}$ is zero-dimensional but has no proper open subgroups.

3.6 Group topologies determined by sequences

Let G be an abelian group and let (a_n) be a sequence in G. The question of the existence of a Hausdorff group topology that makes the sequence (a_n) converge to 0 is not only a mere curiosity. Indeed, assume that some Hausdorff group topology τ makes the sequence (p_n) of all primes converge to zero. Then $p_n \to 0$ would yield $p_n - p_{n+1} \to 0$ in τ, so this sequence cannot contain infinitely many entries equal to 2. This would provide a very easy negative solution to the celebrated problem of the infinitude of twin primes (actually this argument would show that the shortest distance between two consecutive primes converges to ∞).

Definition 3.47. [89] A sequence $A = \{a_n\}_n$ in an abelian group G is called a *T-sequence* is there exists a Hausdorff group topology on G such that $a_n \to 0$. [2]

Let (a_n) be a T-sequence in an abelian group G. Hence the family $\{\tau_i : i \in I\}$ of Hausdorff group topologies on the group G such that $a_n \to 0$ in τ_i is non-empty. Let $\tau = \sup_{i\in I} \tau_i$, then by Exercise 3.5 $a_n \to 0$ in τ as well. Clearly, this is the finest group topology in which a_n converges to 0. This is why we denote it by τ_A or $\tau_{(a_n)}$.[3]

Before discussing the topology $\tau_{(a_n)}$ and how T-sequences can be described in general we consider a couple of examples:

Example 3.48. (a) Let us see that the sequences (n^2) and (n^3) are not a T-sequence in $\mathbb{Z}$. Indeed, suppose for a contradiction that some Hausdorff group topology τ on $\mathbb{Z}$ makes n^2 converge to 0. Then $(n+1)^2$ converges to 0 as well. Taking the difference we conclude that $2n+1$ converges to 0 as well. Since obviously also $2n+3$ converges to 0, we conclude, after substraction, that the constant sequence 2 converges to 0. This is a contradiction, since τ is Hausdroff. We leave the case (n^3) as an exercise to the reader.

(b) A similar argument proves that the sequence $P_d(n)$, where $P_d(x) \in \mathbb{Z}[x]$ is a fixed polynomial with $\deg P_d = d > 0$, is not a T-sequence in $\mathbb{Z}$.

Protasov and Zelenyuk [88] established a number of nice properties of the finest group topology $\tau_{(a_n)}$ on G that makes (a_n) converge to 0.

For an abelian group G and subsets $A_1, \dots, A_n \dots$ of G we denote by $\pm A_1 \pm \dots \pm A_n$ the set of all sums $g = g_1 + \dots + g_n$, where $g_i \in \{0\} \cup A_i \cup -A_i$ for every $i = 1, \dots, n$. Let

$$\pm A_1 \pm \dots \pm A_n \pm \dots = \bigcup_{n=1}^{\infty} \pm A_1 \pm \dots \pm A_n.$$

If $A = \{a_n\}_n$ is a sequence in G, for $m \in \mathbb{N}$ denote by A_m the "tail" $\{a_m, a_{m+1}, \dots\}$. For $k \in \mathbb{N}$ let $A(k,m) = \pm A_m \pm \dots \pm A_m$ (k times).

Remark 3.49. The existence of a finest group topology τ_A on an abelian group G that makes an arbitrary given sequence $A = \{a_n\}_n$ in G converge to 0 is easy to prove as far as we are not interested on imposing the Hausdorff axiom. Indeed, as a_n converges to 0 in the indiscrete topology, τ_A is simply the supremum of all group topologies τ on G such that a_n converges to 0 in τ. This gives no idea on how this topology looks like. One can easily describe it as follows.

Let $m_1, \dots, m_n, \dots$ be a sequence of natural numbers. Denote by $A(m_1, \dots, m_n, \dots)$ the set

$$\pm A_{m_1} \pm \dots \pm A_{m_n} \pm \dots$$

and by $\mathcal{B}_A$ the family of all sets $A(m_1, \dots, m_n, \dots)$ when $m_1, \dots, m_n, \dots$ vary in $\mathbb{N}^{\mathbb{N}}$. Then $\mathcal{B}_A$ is a filter base, satisfying the axioms of group topology. The group topology τ defined in this way satisfies the required conditions. Indeed, obviously $a_n \to 0$ in (G, τ) and τ contains any other group topology with this property. Consequently, $\tau = \tau_A$.

Note that

$$A(k,m) \subseteq A(m_1, \dots, m_n, \dots), \tag{1}$$

for every $k \in \mathbb{N}$, where $m = \max\{m_1, \dots, m_k\}$. The sets $A(k,m)$, for $k, m \in \mathbb{N}$, form a filter base, but the filter they generate need not be the filter of neighborhoods of 0 in a group topology. The utility of this family becomes clear now.

[2] We shall see below that the sequence (p_n) of all primes is not a T-sequence in the group $\mathbb{Z}$ (see Exercise 4.32). So the above mentioned possibility to resolve the problem of the infinitude of twin primes does not work.

[3] To simplify things we consider only sequences without repetition, hence the convergence to zero $a_n \to 0$ depends only on the set $A = \{a_n\}_n$, it does not depend on the enumeration of the sequence.

Theorem 3.50. *A sequence $A = \{a_n\}_n$ in an abelian group G is a T-sequence iff*

$$\bigcap_{m=1}^{\infty} A(k,m) = 0 \text{ for every } k \in \mathbb{N}. \tag{2}$$

Proof. Obviously the sequence $A = \{a_n\}_n$ is a T-sequence iff the topology τ_A is Hausdorff. Clearly, τ_A is Hausdorff iff $\bigcap_{m_1,\dots,m_n,\dots}^{\infty} A(m_1,\dots,m_n,\dots) = 0$. If τ_A is Hausdorff, then (2) holds by (1). It remains to see that (2) implies $\bigcap_{m_1,\dots,m_n,\dots}^{\infty} A(m_1,\dots,m_n,\dots) = 0$. First of all note that $A(m_1,\dots,m_n,\dots) \supseteq A(m_1^*,\dots,m_n^*,\dots)$, where $m_n^* = \max\{m_1,\dots,m_n\}$. Moreover, the sequence (m_n^*) is increasing. Hence

$$\bigcap_{m_1,\dots,m_n,\dots}^{\infty} A(m_1,\dots,m_n,\dots) = \bigcap_{m_1^*,\dots,m_n^*,\dots}^{\infty} A(m_1^*,\dots,m_n^*,\dots),$$

where the second intersection is taken only over the increasing sequences (m_n^*). Obviously, for every increasing sequence (m_n^*) one has

$$A(m_1^*,\dots,m_n^*,\dots) \subseteq \bigcup_{k=1}^{\infty} A(k,m_1^*).$$

This yields

$$\bigcap_{m_1^*,\dots,m_n^*,\dots}^{\infty} A(m_1^*,\dots,m_n^*,\dots) \subseteq \bigcap_{m_1^*=1}^{\infty} \bigcup_{k=1}^{\infty} A(k,m_1^*) = \bigcup_{k=1}^{\infty} \bigcap_{m_1^*=1}^{\infty} A(k,m_1^*) = 0.$$

□

Since every infinite abelian group G admits a non-discrete metrizable group topology, there exist non-trivial (i.e., having all members non-zero) T-sequences.

A notion similar to T-sequence, but defined with respect to only topologies induced by characters, will be given in §6.2. From many points of view it turns out to be easier to deal with than T-sequence. In particular, we shall see easy sufficient condition for a sequence of integers to be a T-sequence.

We give without proof the following technical lemma that will be useful in §6.2.

Lemma 3.51. [89] *For every T-sequence $A = \{a_n\}$ in $\mathbb{Z}$ there exists a sequence $\{b_n\}$ in $\mathbb{Z}$ such that for every choice of the sequence (e_n), where $e_n \in \{0,1\}$, the sequence q_n defined by $q_{2n} = b_n + e_n$ and $q_{2n-1} = a_n$, is a T-sequence.*

Exercise 3.52. *(a)* Prove that there exists a T-sequence (a_n) in $\mathbb{Z}$ with $\lim_n \frac{a_{n+1}}{a_n} = 1$ [89] (see also Example 6.12).*

(b) Every sequence (a_n) in $\mathbb{Z}$ with $\lim_n \frac{a_{n+1}}{a_n} = +\infty$ is a T-sequence [89, 7] (see Theorem 6.11).*

(c) Every sequence (a_n) in $\mathbb{Z}$ such that $\lim_n \frac{a_{n+1}}{a_n} \in \mathbb{R}$ is transcendental is a T-sequence [89].*

3.7 Markov's problems

3.7.1 The Zariski topology and the Markov topology

Let G be a Hausdorff topological group, $a \in G$ and $n \in \mathbb{N}$. Then the set $\{x \in G : x^n = a\}$ is obviously closed in G. This simple fact motivated the following notions due to to Markov [76].

A subset S of a group G is called:

(a) *elementary algebraic* if there exist an integer $n > 0$, $a_1,\dots,a_n \in G$ and $\varepsilon_1,\dots,\varepsilon_n \in \{-1,1\}$ such that

$$S = \{x \in G : x^{\varepsilon_1} a_1 x^{\varepsilon_2} a_2 \dots a_{n-1} x^{\varepsilon_n} = a_n\},$$

(b) *algebraic* if S is an intersection of finite unions of elementary algebraic subsets,

(c) *unconditionally closed* if S is closed in *every* Hausdorff group topology of G.

Since the family of all finite unions of elementary algebraic subsets is closed under finite unions and contains all finite sets, it is a base of closed sets of some T_1 topology $\mathfrak{Z}_G$ on G, called the *Zariski topology*[4]. Clearly, the $\mathfrak{Z}_G$-clsoed sets are precisely the algebraic sets in G.

Analogously, the family of all unconditionally closed subsets of G coincides with the family of closed subsets of a T_1 topology $\mathfrak{M}_G$ on G, namely the infimum (taken in the lattice of all topologies on G) of all Hausdorff group topologies on G. We call $\mathfrak{M}_G$ the *Markov topology* of G. Note that $(G, \mathfrak{Z}_G)$ and $(G, \mathfrak{M}_G)$ are quasi-topological groups, i.e., the inversion and translations are continuous. Nevertheless, when G is abelian $(G, \mathfrak{Z}_G)$ and $(G, \mathfrak{M}_G)$ are not group topologies unless they are discrete.

Since an elementary algebraic set of G must be closed in every Hausdorff group topology on G, one always has $\mathfrak{Z}_G \subseteq \mathfrak{M}_G$. In 1944 Markov [76] asked if the equality $\mathfrak{Z}_G = \mathfrak{M}_G$ holds for every group G. He himself showed that the answer is positive in case G is countable [76]. Moreover, in the same manuscript Markov attributes to Perel'man the fact that $\mathfrak{Z}_G = \mathfrak{M}_G$ for every Abelian group G (a proof has never appeared in print until [37]). An example of a group G with $\mathfrak{Z}_G \neq \mathfrak{M}_G$ was given by Gerchard Hesse [66].

Exercise 3.53. *Show that if $(G, \cdot)$ is an abelian group, then every elementary algebraic set of G has the form $\{x \in G : x^n = a\}$, $a \in G$.*

3.7.2 The Markov topology of the symmetric group

Let X be an infinite set. In the sequel we denote by τ_X the pointwise convergence topology of the infinite symmetric group $S(X)$ defined in §3.2.4. It turns out that the Markov topology of $S(X)$ coincides with τ_X:

Theorem 3.54. *Then Markov topology on $S(X)$ coincides with the topology τ_X of pointwise convergence of $S(X)$.*

This theorem follows immediately from the following old result due to Gaughan.

Theorem 3.55. *([36]) Every Hausdorff group topology of the infinite permutation group $S(X)$ contains the topology τ_X.*

The proof of this theorem follows more or less the line of the proof exposed in [36, §7.1] with several simplifications. The final stage of the proof is preceded by a number of claims (and their corollaries) and two facts about *purely algebraic* properties of the group $S(X)$ (3.56 and 3.59). The claims and their corollaries are given with complete proofs. To give an idea about the proofs of the two algebraic facts, we prove the first one; the proof of the second one can be found in [36, Lemmas 7.1.4, 7.1.8] (actually, only a fragment of the proof of [36, Lemmas 7.1.8] is needed for the proof of item (b) of Fact 3.59).

We say for a subset A of $S(X)$ that A is m-transitive for some positive integer m if for every $Y \subseteq X$ of size at most m and every injection $f : Y \to X$ there exists $a \in A$ that extends f. [5] The leading idea is that a transitive subset A of $S(X)$ is placed "generically" in $S(X)$, whereas a non-tranisitve one is a subset of some subgroup of $S(X)$ that is a direct product $S(Y) \times S(X \setminus Y)$. (Here and in the sequel, for a subset Y of X we tacitly identify the group $S(Y)$ with the subgroup of $S(X)$ consisting of all permutations of $S(X)$ that are identical on $X \setminus Y$.)

The first fact concerns the stabilizers $S_x = S_{\{x\}} = \{f \in S(X) : f(x) = x\}$ of points $x \in X$. They consitute a prebase of the filter of neighborhoods of id_X in τ_X.

Fact 3.56. *For every $x \in X$ the subgroup S_x of $S(X)$ is maximal.*

Proof. Assume H is a subgroup of $S(X)$ properly containing S_x. To show that $H = S(X)$ take any $f \in S(X)$. If $y = f(x)$ coincides with x, then $f \in S_x \subseteq H$ and we are done. Assume $y \neq x$. Get $h \in H \setminus S_x$. Then $z = h(x) \neq x$, so $x \notin \{z, y\}$. There exists $g \in S(X)$ such that $g(x) = x, g(y) = z$ and $g(z) = y$. Then $g \in S_x \subseteq H$ and $f(x) = g(h(x)) = y$, so $h^{-1}g^{-1}f(x) = x$ and $h^{-1}g^{-1}f \in S_x \cap G \subseteq H$. So $f \in ghH = H$. □

Claim 3.57. *Let T be a Hausdorff group topology on $S(X)$. If a subgroups of $S(X)$ of the form S_x is T-closed, then it is also T-open.*

Proof. As S_x is T-closed, for every fixed $y \neq x$ the set $V_y = \{f \in S(X) : f(x) \neq y\}$ is T-open and contains 1. So there exists a symmetric neighborhood W of 1 in T such that $W.W \subseteq V_y$. By the definition of V_y this gives $Wx \cap Wy = \emptyset$. Then either $|X \setminus Wx| = |X|$ or $|X \setminus Wy| = |X|$. Suppose this occurs with x, i.e., $|X \setminus Wx| = |X|$.

[4] Some authors call it also the *verbal topology* [20], we prefer here *Zariski topology* coined by most authors [10].

[5] Note that a countable subset H of $S(X)$ cannot be transitive unless X itself is countable.

Then one can find a permutation $f \in S(X)$ that sends $Wx \setminus \{x\}$ to the complement of Wx and $f(x) = x$. Such an f satisfies:

$$fWf^{-1} \cap W \subseteq S_x$$

as $fWf^{-1}(x)$ meets Wx precisely in the singleton $\{x\}$ by the choice of f. This proves that S_x is T-open.

Analogous argument works for S_y when $|X \setminus Wy| = |X|$. □

Corollary 3.58. *If T be a Hausdorff group topology on $S(X)$ that does not contain τ_X, then all subgroups of $S(X)$ of the form S_x are T-dense.*

Proof. Since the subgroups S_x of $S(X)$ form a prebase of the filter of neighborhoods of id_X in $S(X)$, out hypothesis implies that some subgroup S_x is not T-open. By Claim 3.57 S_x is not T-closed either. By Fact 3.56 S_x is T-dense. Since all subgroups of the form S_y are conjugated, this implies that stabilizers S_y are T-dense. □

This was the first step in the proof. The next step will be establishing that $S_{x,y}$ are never dense in any Hausdorff group topology on $S(X)$ (Corollary 3.62).

In the sequel we need the subgroup $\widetilde{S}_{x,y} := S_{x,y} \times S(\{x,y\})$ of $S(X)$ that contains $S_{x,y}$ as a subgroup of index 2. Note that $\widetilde{S}_{x,y}$ is precisely the subgroup of all permutations in $S(X)$ that leave the doubleton $\{x,y\}$ set-wise invariant.

Fact 3.59. *For any doubleton x, y in X the following holds true:*

(a) the subgroup $\widetilde{S}_{x,y}$ of $S(X)$ is maximal;

(b) every proper subgroup of $S(X)$ properly containing $S_{x,y}$ coincides with one of the subgroups S_x, S_y or $\widetilde{S}_{x,y}$.

Claim 3.60. *Let T be a Hausdroff group topology on $S(X)$, then there exists a T-nbd of 1 that is not 2-transitive.*

Proof. Assume for a contradiction that all T-neighborhoods of id_X that are 2-transitive. Fix distinct $u, v, w \in X$. We show now that the 3-cycle $(u, v, w) \in V$ for every arbitrarily fixed T-neighborhood of id_X. Indeed, choose a symmetric T-neighborhood W of id_X such that $W^2 \subseteq V$. Let f be the transposition (uv). Then $U = fWf \cap W \in T$ is a neighborhood of 1 and $fUf = U$. Since U is 2-transitive there exists $g \in U$ such that $g(u) = u$ and $g(v) = w$. Then $(u, v, w) = gfg^{-1}f \in W \cdot (fUf) \subseteq W^2 \subseteq V$. □

Claim 3.61. *Let T be a group topology on $S(X)$. Then*

(a) every T-nbd V of id_X in $S(X)$ is transitive iff every stabilizer S_x is T-dense;

(b) every T-nbd V of id_X in $S(X)$ is m-transitive iff every stabilizer S_F with $|F| \leq m$ is T-dense.

Proof. Assume that some (hence all) S_z is T-dense in $S(X)$. To prove that V is transitive consider a pair $x, y \in X$. Let $t = (xy)$. By the T-density of S_x the T-nbd $t^{-1}V$ of t^{-1} meets S_x, i.e., for some $v \in V$ one has $t^{-1}v \in S_x$. Then $v \in tS_x$ obviously satisfies $vx = y$.

A similar argument proves that transitivity of each T-nbd of 1 entails that every stabilizer S_x is T-dense.

(b) The proof in the case $m > 1$ is similar. □

What we really need further on (in particular, in the next corollary) is that the density of the stabilizers $S_{x,y}$ imply that every T-nbd V of id_X in $S(X)$ is 2-transitive.

Corollary 3.62. *Let T be a Hausdroff group topology on $S(X)$. Then $S_{x,y}$ is T-dense for no pair x, y in X.*

Proof. Follows from claims 3.60 and 3.61 □

Proof of Theorem 3.55. Assume for a contradiction that T is a Hausdroff group topology on $S(X)$ that does not contain τ_X. Then by corollaries 3.58 and 3.62 all subgroups of the form S_x are T-dense and no subgroup of the form $S_{x,y}$ is T-dense. Now fix a pair $x, y \in X$ and let $G_{x,y}$ denote the T-closure of $S_{x,y}$. Then $G_{x,y}$ is a proper subgroup of $S(X)$ containing $S_{x,y}$. Since S_x is dense, $G_{x,y}$ cannot contain S_x, so $S_x \cap G_{x,y}$ is a proper subgroup of S_x containing $S_{x,y}$. By Claim 3.56 applied to $S_x = S(X \setminus \{x\})$ and its subgroup $S_{x,y}$ (the stabilizer of y in S_x), we conclude that $S_{x,y}$ is a maximal subgroup of S_x. Therefore, $S_x \cap G_{x,y} = S_{x,y}$. This shows that $S_{x,y}$ is a T-closed subgroup of S_x. By Claim 3.57 applied to $S_x = S(X \setminus \{x\})$ and its subgroup $S_{x,y}$, we conclude that $S_{x,y}$ is a T-open subgroup of S_x. Since S_x is dense in $S(X)$, we can claim that $G_{x,y}$ is a T-open subgroup of $S(X)$. Since S_x is a proper dense subgroup of $S(X)$, it is clear that S_x cannot contain $G_{x,y}$. Analogously, S_y

cannot contain $G_{x,y}$ either. So $G_{x,y} \neq S_{x,y}$ is a proper subgroup of $S(X)$ containing $S_{x,y}$ that does not coincide with S_x or S_y. Therefore $G_{x,y} = \widetilde{S}_{x,y}$ by Fact 3.59. This proves that $\widetilde{S}_{x,y}$ is T-open. Since all subgroups of the form $\widetilde{S}_{x,y}$ are pairwise conjugated, we can claim that all subgroups $\widetilde{S}_{x,y}$ is T-open.

Now we can see that the stabilizers S_F with $|F| > 2$ are T-open, as

$$S_F = \bigcap \{\widetilde{S}_{x,y} : x, y \in F, x \neq y\}.$$

This proves that all basic neighborhoods S_F of 1 in τ_X are T-open. In particular, also the subgroups S_x are T-open, contrary to our hypothesis.

3.7.3 Existence of Hausdorff group topologies

According to Proposition 3.31 every infinite abelian group admits a non-discrete Hausdorff group topology, for example the Bohr topology. This gives immediately the following

Corollary 3.63. *Every group with infinite center admits a non-discrete Hausdorff group topology.*

Proof. The center $Z(G)$ of the group G has a non-discrete Hausdorff group topology τ by the above remark. Now consider the family $\mathcal{B}$ of all sets of the form aU, where $a \in G$ and U is a non-empty τ-subset of $Z(G)$. It is easy to see that it is a base of a non-discrete Hausdorff group topology on G. □

In 1946 Markov set the problem of the existence of a (countably) infinite group G that admits no Hausdorff group topology beyond the discrete one. Let us call such a group a *Markov group*. Obviously, G is a Markov group precisely when $\mathfrak{M}_G$ is discrete. A Markov group must have finite center by Corollary 3.63.

According to Proposition 3.26, the closure of the neutral element of every topological group is always a normal subgroup of G. Therefore, a simple topological group is either Hausdorff, or indiscrete. So a simple Markov group G admits only two group topologies, the discrete and the indiscrete ones.

The equality $\mathfrak{Z}_G = \mathfrak{M}_G$ established by Markov in the countable case was intended to help in finding a countably infinite Markov group G. Indeed, a countable group G is Markov precisely when $\mathfrak{Z}_G$ is discrete. Nevertheless, Markov failed in building a countable group G with discrete Zariski topology; this was done much later, in 1980, by Ol′shanskii [78] who made use of the so called *Adian groups* $A = A(m, n)$ (constructed by Adian to negatively resolve the famous 1902 Burnside problem on finitely generated groups of finite exponent). Let us sketch here Ol′shanskii's elegant short proof.

Example 3.64. [78] Let m and n be odd integers ≥ 665, and let $A = A(m, n)$ be Adian's group having the following properties

(a) A is generated by n-elements;

(b) A is torsion-free;

(c) the center C of A is infinite cyclic.

(d) the quotient A/C is infinite, of exponent m, i.e., $y^m \in C$ for every $y \in A$.[6]

By (a) the group A is countable. Denote by C^m the subgroup $\{c^m : c \in C\}$ of A. Let us see that (b), (c) and (d) jointly imply that the Zariski topology of the infinite quotient $G = A/C^m$ is discrete (so G is a countably infinite Markov group). Let d be a generator of C. Then for every $x \in A \backslash C$ one has $x^m \in C \backslash C^m$. Indeed, if $x^m = d^{ms}$, then $(xd^{-s})^m = 1$ for some $s \in \mathbb{Z}$, so $xd^{-s} = 1$ and $x \in C$ by (b). Hence

$$\text{for every } u \in G\backslash\{1\} \text{ there exists } a \in C\backslash C^m, \text{ such that either } u = a \text{ or } u^m = a. \tag{3}$$

As $|C/C^m| = m$, every $u \in G\backslash\{1\}$ is a solution of some of the $2(m-1)$ equations in (3). Thus, $G\backslash\{e\}$ is closed in the Zariski topology $\mathfrak{Z}_G$ of G. Therefore, $\mathfrak{Z}_G$ is discrete.

Now we recall an example, due to Shelah [92], of an uncountable group which is non-topologizable. It appeared about a year or two earlier than the ZFC-example of Ol′shanskii exposed above.

Example 3.65. [92] Under the assumption of CH there exists a group G of size ω_1 satisfying the following conditions (a) (with $m = 10000$) and (b) (with $n = 2$):

(a) there exists $m \in \mathbb{N}$ such that $A^m = G$ for every subset A of G with $|A| = |G|$;

[6] i.e., the finitely generated infinite quotient A/C negatively resolves Burnside's problem.

(b) for every subgroup H of G with $|H| < |G|$ there exist $n \in \mathbb{N}$ and $x_1, \dots, x_n \in G$ such that the intersection $\bigcap_{i=1}^n x_i^{-1} H x_i$ is finite.

Let us see that G is a Markov group (i.e., $\mathfrak{M}_G$ is discrete)[7]. Assume $\mathcal{T}$ be a Hausdorff group topology on G. There exists a $\mathcal{T}$-neighbourhood V of e_G with $V \neq G$. Choose a $\mathcal{T}$-neighbourhood W of e_G with $W^m \subseteq V$. Now $V \neq G$ and (a) yield $|W| < |G|$. Let $H = \langle W \rangle$. Then $|H| = |W| \cdot \omega < |G|$. By (b) the intersection $O = \bigcap_{i=1}^n x_i^{-1} H x_i$ is finite for some $n \in \mathbb{N}$ and elements $x_1, \dots, x_n \in G$. Since each $x_i^{-1} H x_i$ is a $\mathcal{T}$-neighbourhood of e_G, this proves that $e_G \in O \in \mathcal{T}$. Since $\mathcal{T}$ is Hausdorff, it follows that $\{e_G\}$ is $\mathcal{T}$-open, and therefore $\mathcal{T}$ is discrete.

One can see that even the weaker form of (a) (with m depending on $A \in [G]^{|G|}$), yields that every proper subgroup of G has size $< |G|$. In the case $|G| = \omega_1$, the groups with this property are known as *Kurosh groups* (in particular, this is a *Jonsson semigroup* of size ω_1, i.e., an uncountable semigroup whose proper subsemigroups are countable).

Finally, this remarkable construction from [92] furnished also the first consistent example to a third open problem. Namely, a closer look at the above argument shows that the group G is simple. As G has no maximal subgroups, it shows also that taking Frattini subgroup[8] "does not commute" with taking finite direct products (indeed, $Fratt(G) = G$, while $Fratt(G \times G) = \Delta_G$ the "diagonal" subgroup of $G \times G$).

3.7.4 Extension of group topologies

The problem of the existence of (Hausdorff non-discrete) group topologies can be considered also as a problem of extension of (Hausdorff non-discrete) group topologies.

The theory of extension of topological spaces is well understood. If a subset Y of a set X carries a topology τ, then it is easy to extend τ to a topology τ^* on X such that (Y, τ) is a subspace of (X, τ^*). The easiest way to do it is to consider $X = Y \cup (X \setminus Y)$ as a partition of the new space (X, τ^*) into clopen sets and define the topology of $X \setminus Y$ arbitrarily. Usually, one prefers to define the extension topology τ^* on X in such a way to have Y dense in X. In such a case the extensions of a given space (Y, τ) can be described by means of appropriate families of open filters of Y (i.e., filters on Y having a base of τ-open sets).

The counterpart of this problem for groups and group topologies is much more complicated because of the presence of group structure. Indeed, let H be a subgroup of a group G and assume that τ is a group topology of H. Now one has to build a group topology τ^* on G such that (H, τ) is a topological subgroup of (G, τ^*). The first idea to extend τ is to imitate the first case of extension considered above by declaring the subgroup H a τ^*-open topological subgroup of the new topological group (G, τ^*). Let us note that this would immediately determine the topology τ^* in a unique way. Indeed, every coset gH of H must carry the topology transported from H to gH by the translation $x \mapsto gx$, i.e., the τ^*-open subsets of gH must have the form gU, where U is an open subset of (H, τ). In other words, the family $\{gU : \emptyset \neq U \in \tau\}$ is a base of τ^*. This idea has worked in the proof of Corollary 3.63 where H was the center of G. Indeed, this idea works in the following more general case.

Lemma 3.66. *Let H be a subgroup of a group G such that $G = Hc_G(G)$. Then for every group topology τ on H the above described topology τ^* is a group topology of G such that (H, τ) is a topological subgroup of (G, τ^*).*

Proof. The first two axioms on the neighborhood base are easy to check. For the third one pick a basic τ^*-neighborhood U of 1 in G. Since H is τ^*-open, we can assume wlog that $U \subseteq H$, so U is a τ-neighborhood of 1. Let $x \in G$. We have to produce a τ^*-neighborhood V of 1 in G such that $x^{-1}Vx \subseteq U$. By our hypothesis there exist $h \in H, z \in c_G(G)$, such that $x = hz$. Since τ is a group topology on H there exist $V \in \mathcal{V}_{H,\tau}(1)$ such that $h^{-1}Vh \subseteq U$. Then

$$x^{-1}Vx = z^{-1}h^{-1}Vhz \subseteq z^{-1}Uz = U$$

as $z \in c_G(G)$. This proves that τ^* is a group topology of G . □

Clearly, the condition $G = Hc_G(G)$ is satisfied when H is a central subgroup of G. It is satisfied also when H is a direct summand of G. On the other hand, subgroups H satisfying $G = Hc_G(G)$ are normal.

Two questions are in order here:

- is the condition $G = Hc_G(G)$ really necessary for the extension problems;
- is it possible to definite the extension τ^* in a different way in order to have *always* the possibility to extend a group topology?

[7] Hesse [66] showed that the use of CH in Shelah's construction of a Markov group of size ω_1 can be avoided.
[8] the Frattini subgroup of a group G is the intersection of all maximal subgroups of G.

Our next theorem shows that the difficulty of the extension problem are not hidden in the special features of the extension τ^*.

Theorem 3.67. *Let H be a normal subgroup of the group G and let τ be a group topology on H. Then the following are equivalent:*

(a) the extension τ^* is a group topology on G;

(b) τ can be extended to a group topology of G;

(c) for every $x \in G$ the automorphism of H induced by the conjugation by x is τ-continuous.

Proof. The implication (a) $\to$ (b) is obvious, while the implication (b) $\to$ (c) follows from the fact that the conjugations are continuous in any topological group. To prove the implication (c) $\to$ (a) assume now that all automorphisms of N induced by the conjugation by elements of G are τ-continuous. Take the filter of all neighborhoods of 1 in (H, τ^*) as a base of neighborhoods of 1 in the group topology τ^* of G. This works since the only axiom to check is to find for every $x \in G$ and every τ^*-nbd U of 1 a τ^*-neighborhood V of 1 such that $V^x := x^{-1}Vx \subseteq U$. Since we can choose U, V contained in H, this immediately follows from our assumption of τ-continuity of the restrictions to H of the conjugations in G. □

Now we give an example showing that the extension problem cannot be resolved for certain triples G, H, τ of a group G, its subgroup H and a group topology τ on H.

Example 3.68. In order to produce an example when the extension is not possible we need to produce a triple G, H, τ such that at least some conjugation by an element of G is not τ-continuous when considered as an automorphism of H. The best tool to face this issue is the use of semi-direct products.

Let us recall that for groups K, H and a group homomorphism $\theta : K \to Aut(H)$ one defines the semi-direct product $G = H \rtimes_\theta K$, where we shall identify H with the subgroup $H \times \{1\}$ of G. In such a case, the conjugation in G by an element k of K restricted to H is precisely the automorphism $\theta(k)$ of H. Now consider a group topology τ on H. According to Theorem 3.67 τ can be extended to a group topology of G iff for every $k \in K$ the automorphism $\theta(k)$ of H is τ-continuous. (Indeed, every element $x \in G$ has the form $x = hk$, where $h \in H$ and $k \in K$; hence it remains to note that the conjugation by x is composition of the (continuous) conjugation by h and the conjugation by k.)

In order to produce the required example of a triple G, H, τ such that τ cannot be extended to G it suffices to find a group K and a group homomorphism $\theta : K \to Aut(H)$ such that at least one of the automorphisms $\theta(k)$ of H is τ-discontinuous. Of course, one can simplify the construction by taking the cyclic group $K_1 = \langle k \rangle$ instead of the whole group K, where $k \in K$ is chosen such that the automorphisms $\theta(k)$ of H is τ-discontinuous. A further simplification can be arranged by taking k in such a way that the automorphism $f = \theta(k)$ of H is also an involition, i.e., $f^2 = id_H$. Then H will be an index two subgroup of G.

Here is an example of a topological abelian group (H, τ) admitting a τ-discontuous involition f. Then the triple G, H, τ such that τ cannot be extended to G is obtained by simply taking $G = H \rtimes \langle f \rangle$, where the involition f acts on H. Take as (H, τ) the torus group $\mathbb{T}$ with the usual topology. Then $\mathbb{T}$ is algebraically isomorphic to $(\mathbb{Q}/\mathbb{Z}) \oplus_{\mathfrak{c}} \bigoplus \mathbb{Q}$, so $\mathbb{T}$ has $2^{\mathfrak{c}}$ many involutions. Of these only the involutions $\pm id_{\mathbb{T}}$ of $\mathbb{T}$ are continuous.

Let us conclude now with a series of examples when the extension problem has always a positive solution.

Example 3.69. Let p be a prime number. If the group of p-adic integers $N = \mathbb{Z}_p$ is a normal subgroup of some group G, then the p-adic topology of N can be extended to a group topology on G. Indeed, it suffices to note that if $\xi : N \to N$ is an automorphism of N, then $\xi(p^n N) = p^n N$. Since the subgroups $p^n N$ define the topology of N, this proves that every automorphism of N is continuous. Now Theorem 3.67 applies.

Clearly, the p-adic integers can be replaced by any topological group N such that every automorphism of N is continuous (e.g., products of the form $\prod_p \mathbb{Z}_p^{k_p} \times F_p$, where $k_p < \omega$ and F_p is a finite abelian p-group).

3.8 Cardinal invariants of topological groups

Here we shall be interested in measuring the minimum size of a base (of neighborhoods of 1) in a topological group H, as well as other cardinal functions related to H.

It is important to relate the bases (of neighborhoods of 1) in H to those of a subgroup G of H.

Exercise 3.70. *If G is a subgroup of a topological group H and if $\mathcal{B}$ is a base (of neighborhoods of 1) in H then a base (of neighborhoods of 1) in G is given by $\{U \cap G : U \in \mathcal{B}\}$.*

Now we consider the case when G is a *dense* subgroup of H.

Lemma 3.71. *If G is a dense subgroup of a topological group H and $\mathcal{B}$ is a base of neighborhoods of 1 in G, then $\{\overline{U}^H : U \in \mathcal{B}\}$ is a base of neighborhoods of 1 in H.*

Proof. Since the topological group H is regular, the closed neighborhoods form a base at 1 in H. Hence for a neighborhood $V \ni 1$ in H one can find another neighborhood $V_0 \ni 1$ such that $\overline{V_0} \subseteq V$. Since $G \cap V_0$ is a neighborhood of 1 in G, there exists $U \in \mathcal{B}$ such that $U \subseteq G \cap V_0$. There exists also an open neighborhood W of 1 in H such that $U = W \cap G$. Obviously, one can choose $W \subseteq V_0$. Hence $\overline{U}^H = \overline{W}$ as G is dense in H and W is open in H. Thus $\overline{U}^H = \overline{W} \subseteq \overline{V}_0 \subseteq V$ is a neighborhood of 1 in H. □

Lemma 3.72. *Let G be a dense subgroup of a topological group H and let $\mathcal{B}$ be a base of neighborhoods of 1 in H. Then $\{gU : U \in \mathcal{B}, g \in G\}$ is a base of the topology of H.*

Proof. Let $x \in H$ and let $x \in O$ be an open set. Then there exists a $U \in \mathcal{U}$ symmetric with $xU^2 \subseteq O$. Pick a $g \in G \cap xU$. Then $x \in gU \subseteq O$. □

For a topological group G set $d(G) = \min\{|X| : X \text{ is dense in } G\}$,

$$w(G) = \min\{|\mathcal{B}| : \mathcal{B} \text{ is a base of } G\} \text{ and } \chi(G) = \min\{|\mathcal{B}| : \mathcal{B} \text{ a base of neighborhoods of 1 in } G\}.$$

Lemma 3.73. *Let H be a subgroup of a topological group G. Then:*

(a) $w(H) \leq w(G)$ and $\chi(H) \leq \chi(G)$;

(b) if H is dense in G, then $w(G) = w(H)$ and $\chi(G) = \chi(H)$.

Lemma 3.74. *$w(G) = \chi(G) \cdot d(G)$ for every topological group G.*

Proof. The inequality $w(G) \geq \chi(G)$ is obvious. To see that $w(G) \geq d(G)$ choose a base $\mathcal{B}$ of size $w(G)$ and for every $U \in \mathcal{B}$ pick a point $d_U \in U$. Then the set $D = \{d_U : U \in \mathcal{B}\}$ is dense in G and $|D| \leq w(G)$. This proves the inequality $w(G) \geq \chi(G) \cdot d(G)$.

The inequality $w(G) \leq \chi(G) \cdot d(G)$ follows from the previous lemma. □

The *cardinal invariants* of the topological groups are cardinal numbers, say $\rho(G)$, associated to every topological group G such that if G is topologically isomorphic to the topological group H, then $\rho(G) = \rho(H)$. For example, the size $|G|$ is the simplest cardinal invariant of a topological group, it does not depend on the topology of G. Other cardinal invariants are the *weight* $w(G)$, the *character* $\chi(G)$ and the *density character* $d(G)$ defined above. Beyond the equality $w(G) = \chi(G) \cdot d(G)$ proved in Lemma 3.74, one has also the following inequalities:

Lemma 3.75. *Let G be a topological group. Then:*

(a) $d(G) \leq w(G) \leq 2^{d(G)}$;

(b) $|G| \leq 2^{w(G)}$ if G is Hausdorff.

Proof. (a) $d(G) \leq w(G)$ has already been proved in Lemma 3.74 (a). To prove $w(G) \leq 2^{d(G)}$ note that G is regular, hence every open base $\mathcal{B}$ on G contains a base $\mathcal{B}_r$ of the same size consisting of regular open sets[9]. Let $\mathcal{B}$ be a base of G of regular open sets and let D be a dense subgroup of G of size $d(G)$. If $U, V \in \mathcal{B}$, with $U \cap D = V \cap D$, then $\overline{U} = \overline{U \cap D} = \overline{V \cap D} = \overline{V}$. Being U and V regular open, the equality $\overline{U} = \overline{V}$ implies $U = V$. Hence the map $U \mapsto U \cap D$ from $\mathcal{B}$ to the power set $P(D)$ is injective. Therefore $w(G) \leq 2^{d(G)}$.

(b) To every point $x \in G$ assign the set $O_x = \{U \in \mathcal{B} : x \in U\}$. Then the axiom T_2 guarantees that map $x \mapsto O_x$ from G to the power set $P(\mathcal{B})$ is injective. Therefore, $|G| \leq 2^{w(G)}$. □

Remark 3.76. Two observations related to item (b) of the above lemma are in order here.

- The equality in item (b) can be attained (see Theorem 4.46).
- One cannot remove Hausdorffness in item (b) (any large indiscrete group provides a counter-example). This dependence on separation axioms is due to that the presence of the size of the group in (b). We see in the next exercise that the Hausdorff axiom is not relevant as far as the other cardinal invariants are involved.

[9] an open set is said to be *regular open* if it coincides with the interior of its closure.

3.9 Completeness and completion

A net $\{g_\alpha\}_{\alpha\in A}$ in a topological group G is a *Cauchy net* if for every neighborhood U of 1 in G there exists $\alpha_0 \in A$ such that $g_\alpha^{-1} g_\beta \in U$ and $g_\beta g_\alpha^{-1} \in U$ for every $\alpha, \beta > \alpha_0$.

Exercise 3.78. *Let G be a dense subgroup of a topological group H. If (g_α) is a net in G that converges to some element $h \in H$, then (g_α) is a Cauchy net.*

By the previous exercise, the convergent nets are Cauchy nets. A topological group G is *complete (in the sense of Raĭkov)* if every Cauchy net in G converges in G. We omit the tedious proof of the next theorem.

Theorem 3.79. *For every topological Hausdorff group G there exists a complete topological group $\widetilde{G}$ and a topological embedding $i : G \to \widetilde{G}$ such that $i(G)$ is dense in $\widetilde{G}$. Moreover, if $f : G \to H$ is a continuous homomorphism and H is a complete topological group, then there is a unique continuous homomorphism $\widetilde{f} : \widetilde{G} \to H$ with $f = \widetilde{f} \circ i$.*

Therefore every Hausdorff topological abelian group has a unique, up to topological isomorphisms, *(Raĭkov-)completion* $(\widetilde{G}, i)$ and we can assume that G is a dense subgroup of $\widetilde{G}$.

Definition 3.80. *A net $\{g_\alpha\}_{\alpha\in A}$ in G is a* left [resp., right] Cauchy net *if for every neighborhood U of 1 in G there exists $\alpha_0 \in A$ such that $g_\alpha^{-1} g_\beta \in U$ [resp., $g_\beta g_\alpha^{-1} \in U$] for every $\alpha, \beta > \alpha_0$.*

Lemma 3.81. *Let G be a Hausdorff topological group. Every left (resp., right) Cauchy net in G with a convergent subnet is convergent.*

Proof. Let $\{g_\alpha\}_{\alpha\in A}$ be a left Cauchy net in G and let $\{g_\beta\}_{\beta\in B}$ be a subnet convergent to $x \in G$, where B is a cofinal subset of A. Let U be a neighborhood of 1 in G and V a symmetric neighborhood of 1 in G such that $VV \subseteq U$. Since $g_\beta \to x$, there exists $\beta_0 \in B$ such that $g_\beta \in xV$ for every $\beta > \beta_0$. On the other hand, there exists $\alpha_0 \in A$ such that $\alpha_0 \geq \beta_0$ and $g_\alpha^{-1} g_\gamma \in V$ for every $\alpha, \gamma > \alpha_0$. With $\gamma = \beta_0$ we have $g_\alpha \in xVV \subseteq xU$ for every $\alpha > \alpha_0$, that is $g_\alpha \to x$. □

A topological group G is *complete in the sense of Weil* if every left Cauchy net converges in G.

Every Weil-complete group is also complete, but the converse does not hold in general. It is possible to define the Weil-completion of a Hausdorff topological group in analogy with the Raĭkov-completion.

Exercise 3.82. *Prove that if a Hausdorff topological group G admits a Weil-completion, then in G the left Cauchy and the right Cauchy nets coincide.*

Exercise 3.83. *Let X be an infinite set and let $G = S(X)$ equipped with the topology described in §3.2.4. Prove that:*

(a) a net $\{f_\alpha\}_{\alpha\in A}$ in G is left Cauchy iff there exists $f \in X^X$ so that $f_\alpha \to f$ in X^X, prove that such an f must necessarily be injective;

(b) a net $\{f_\alpha\}_{\alpha\in A}$ in G is right Cauchy iff there exists $g \in X^X$ so that $f_\alpha^{-1} \to g$ in X^X;

(c) the group $S(X)$ admits no Weil-completion. (Hint. Build a left Cauchy net in $S(X)$ that is not right Cauchy and use items (a) and (b), as well as the previous exercise.)

(d) $S(X)$ is Raĭkov-complete. (Hint. Use items (a) and (b).)

Exercise 3.84. *(a) Let G be a linearly topologized group and let $\{N_i : i \in I\}$ be its system of neighborhoods of 1 consisting of open normal subgroups. Then the completion of G is isomorphic to the inverse limit $\varprojlim G/N_i$ of the discrete quotients G/N_i.*

(b) Show that the completion in (a) is compact iff all N_i have finite index in G.

Units 3, 4

4 Compactness and local compactness in topological groups

Clearly, a topological group G is locally compact if there exists a compact neighborhood of e_G in G (compare with Definition 2.10). We shall assume without explicitly mentioning it, that all locally compact groups are Hausdorff.

As an immediate consequence of Tychonov's theorem of compactness of products we obtain the following the first example of a compact abelian group (it will become clear with the duality theorem that this is the most general one).

Remark 4.1. Let us see that for every abelian group G the group $G^* = \mathrm{Hom}\,(G, \mathbb{S})$ is closed in the product $\mathbb{S}^G$, hence G^* is compact. Consider the projections $\pi_x : \mathbb{S}^G \to \mathbb{T}$ for every $x \in G$ and the following equalities

$$G^* = \bigcap_{h,g\in G} \{f \in \mathbb{S}^G : f(h+g) = f(h)f(g)\} = \bigcap_{h,g\in G} \{f \in \mathbb{S}^G : \pi_{h+g}(f) = \pi_h(f)\pi_g(f)\}$$
$$= \bigcap_{h,g\in G} \{f \in \mathbb{S}^G : (\pi^{-1}_{h+g}\pi_h\pi_g)(f) = 1\} = \bigcap_{h,g\in G} \ker(\pi^{-1}_{h+g}\pi_h\pi_g).$$

Since π_x is continuous for every $x \in G$ and $\{1\}$ is closed in $\mathbb{S}$, then all $\ker(\pi^{-1}_{h+g}\pi_h\pi_g)$ are closed; so $\mathrm{Hom}\,(G, \mathbb{S})$ is closed too.

The next lemma contains a well known useful fact – the existence of a "diagonal subnet".

Lemma 4.2. *Let G be an abelian group and let $N = \{\chi_\alpha\}_\alpha$ be a net in G^*. Then there exist $\chi \in G^*$ and a subnet $S = \{\chi_{\alpha_\beta}\}_\beta$ of N such that $\chi_{\alpha_\beta}(x) \to \chi(x)$ for every $x \in G$.*

Proof. By Tychonov's theorem, the group $\mathbb{S}^G$ endowed with the product topology is compact. Then N has a convergent (to χ) subnet S. Therefore $\chi_{\alpha_\beta}(x) \to \chi(x)$ for every $x \in G$ and $\chi \in G^*$, because G^* is closed in $\mathbb{S}^G$ by 4.1. □

4.1 Specific properties of (local) compactness

Here we shall see the impact of local compactness in various directions (the open mapping theorem, properties related to connectedness, etc.).

Lemma 4.3. *Let G be a topological group and let C and K be closed subsets of G:*

(a) if K is compact, then both CK and KC are closed;

(b) if both C and K are compact, then CK and KC are compact;

(c) if K is contained in an open subset U of G, then there exists an open neighborhood V of 1 such that $KV \subseteq U$.

Proof. (a) Let $\{x_\alpha\}_{\alpha\in A}$ be a net in CK such that $x_\alpha \to x_0 \in G$. It is sufficient to show that $x_0 \in CK$. For every $\alpha \in A$ we have $x_\alpha = y_\alpha z_\alpha$, where $y_\alpha \in C$ and $z_\alpha \in K$. Since K is compact, then there exist $z_0 \in K$ and a subnet $\{z_{\alpha_\beta}\}_{\beta\in B}$ such that $z_{\alpha_\beta} \to z_0$. Thus $(x_{\alpha_\beta}, z_{\alpha_\beta})_{\beta\in B}$ is a net in $G \times G$ which converges to (x_0, z_0). Therefore $y_{\alpha_\beta} = x_{\alpha_\beta} z_{\alpha_\beta}^{-1}$ converges to $x_0 z_0^{-1}$ because the function $(x, y) \mapsto xy^{-1}$ is continuous. Since $y_{\alpha_\beta} \in C$ for every $\beta \in B$ and C is closed, $x_0 z_0^{-1} \in C$. Now $x_0 = (x_0 z_0^{-1}) z_0 \in CK$. Analogously it is possible to prove that KC is closed.

(b) The product $C \times K$ is compact by the Tychonov theorem and the function $(x, y) \mapsto xy$ is continuous and maps $C \times K$ onto CK. Thus CK is compact.

(c) Let $C = G \setminus U$. Then C is a closed subset of G disjoint with K. Therefore, for the compact subset K^{-1} of G one has $1 \notin K^{-1}C$. By (a) $K^{-1}C$ is closed, so there exists a symmetric neighborhood V of 1 that misses $K^{-1}C$. Then KV misses C and consequently KV is contained in U. □

Compactness of K cannot be omitted in item (a). Indeed, $K = \mathbb{Z}$ and $C = \langle\sqrt{2}\rangle$ are closed subgroups of $G = \mathbb{R}$ but the subgroup $K + C$ of $\mathbb{R}$ is dense (see Exercie 3.20 or Proposition 4.45).

Lemma 4.4. *Let G be a topological group and K a compact subgroup of G. Then the canonical projection $\pi : G \to G/K$ is closed.*

Proof. Let C be a closed subset of G. Then CK is closed by Lemma 4.3 and so $U = G \setminus CK$ is open. For every $x \notin CK$, that is $\pi(x) \notin \pi(C)$, $\pi(U)$ is an open neighborhood of $\pi(x)$ such that $\pi(U) \cap \pi(C)$ is empty. So $\pi(C)$ is closed. □

Lemma 4.5. *Let G be a topological group and let H be a closed subgroup of G.*

(1) If G is compact, then G/H is compact.

(2) If H and G/H are compact, then G is compact.

Proof. (1) is obvious.

(2) Let $\mathcal{F} = \{F_\alpha : \alpha \in A\}$ be a family of closed sets of G with the finite intersection property. If $\pi : G \to G/H$ is the canonical projection, $\pi(\mathcal{F})$ is a family of closed subsets with the finite intersection property in G/H by Lemma 4.4. By the compactness of G/H there exists $\pi(x) \in \pi(F_\alpha)$ for every $\alpha \in A$. So $x \in \bigcap_{\alpha\in A} F_\alpha H$. Let $x = f_\alpha h_\alpha$ with $h_\alpha \in H$ and $f_\alpha \in F_\alpha$. It is not restrictive to assume that $\mathcal{F}$ is closed for finite intersections. Define a partial order on A by $\alpha \leq \alpha'$ if $F_\alpha \supseteq F_{\alpha'}$. Then $(A, \leq)$ is a right-filtered partially ordered set and so $\{f_\alpha\}_{\alpha\in A}$ is a net in G. By the compactness of H we can assume wlog that h_α converges to $h \in H$ (otherwise pass to a convergent subnet). But then $f_\alpha = x h_\alpha^{-1} \to x h^{-1}$. Since f_α is contained definitively in F_α, also the limit $xh^{-1} \in F_\alpha$. So the intersection of all F_α is not empty. □

Lemma 4.6. *Let G be a locally compact group, H be a closed subgroup of G and $\pi : G \to G/H$ be the canonical projection. Then:*

(a) G/H is locally compact too;

(b) If C is a compact subset of G/H, then there exists a compact subset K of G such that $\pi(K) = C$.

Proof. Let U be an open neighborhood of 1 in G with compact closure. Consider the open neighborhood $\pi(U)$ of 1 in G/H. Then $\pi(\overline{U}) \subseteq \overline{\pi(U)}$ by the continuity of π. Now $\pi(\overline{U})$ is compact in G/H, which is Hausdorff, and so $\pi(\overline{U})$ is closed. Since $\pi(U)$ is dense in $\pi(\overline{U})$, we have $\overline{\pi(U)} = \overline{\pi(\overline{U})} = \pi(\overline{U})$. So G/H is locally compact.

(b) Let U be an open neighborhood of 1 in G with compact closure. Then $\{\pi(sU) : s \in G\}$ is an open covering of G/H. Since C is compact, a finite subfamily $\{\pi(s_i U) : i = 1, \ldots, m\}$ covers C. Then we can take $K = (s_1 U \cup \cdots \cup s_m U) \cap \pi^{-1}(C)$. □

Lemma 4.7. *A locally compact group is Weil-complete.*

Proof. Let U be a neighborhood of 1 in G with compact closure and let $\{g_\alpha\}_{\alpha\in A}$ be a left Cauchy net in G. Then there exists $\alpha_0 \in A$ such that $g_\alpha^{-1} g_\beta \in U$ for every $\alpha, \beta \geq \alpha_0$. In particular, $g_\beta \in g_{\alpha_0} U$ for every $\beta > \alpha_0$. By the compactness of $g_{\alpha_0}\overline{U}$, we can conclude that there exists a convergent subnet $\{g_\beta\}_{\beta\in B}$ (for some cofinal $B \subseteq A$) such that $g_\beta \to g \in G$. Then also g_α converges to g by Lemma 3.81. □

Lemma 4.8. *A locally compact countable group is discrete.*

Proof. By the Baire category theorem 2.16 G is of second category. Since $G = \{g_1, \dots, g_n, \dots\} = \bigcup_{n=1}^{\infty}\{g_n\}$, there exists $n \in \mathbb{N}_+$ such that $\operatorname{Int}\{g_n\}$ is not empty and so $\{g_n\}$ is open. □

Now we prove the open mapping theorem for topological groups.

Theorem 4.9 (Open mapping theorem). *Let G and H be locally compact topological groups and let f be a continuous homomorphism of G onto H. If G is σ-compact, then f is open.*

Proof. Let U be an open neighborhood of 1 in G. There exists an open symmetric neighborhood V of 1 in G such that $\overline{VV} \subseteq U$ and $\overline{V}$ is compact. Since $G = \bigcup_{x\in G} xV$ and G is Lindelöff by Lemma 2.17, we have $G = \bigcup_{n=1}^{\infty} x_n V$. Therefore $H = \bigcup_{n=1}^{\infty} h(x_n\overline{V})$, because h is surjective. Put $y_n = h(x_n)$, hence $H = \bigcup_{n=1}^{\infty} y_n h(\overline{V})$ where each $h(\overline{V})$ is compact and so closed in H. Since H is locally compact, Theorem 2.16 yields that there exists $n \in \mathbb{N}_+$ such that $\operatorname{Int} h(\overline{V})$ is not empty. So there exists a non-empty open subset W of H such that $W \subseteq h(\overline{V})$. If $w \in W$, then $w \in h(\overline{V})$ and so $w = h(v)$ for some $v \in \overline{V} = \overline{V}^{-1}$. Hence

$$1 \in w^{-1}W \subseteq w^{-1}h(\overline{V}) = h(v^{-1})h(\overline{V}) \subseteq h(\overline{VV}) \subseteq h(U)$$

and this implies that $h(U)$ is an open neighborhood of 1 in H. □

The following immediate corollary is frequently used:

Corollary 4.10. *If $f : G \to H$ is a continuous surjective homomorphism of Hausdorff topological groups and G is compact, then f is open.*

Now we introduce a special class of σ-compact groups that will play an essential role in determining the structure of the locally compact abelian groups.

Definition 4.11. *A group G is* compactly generated *if there exists a compact subset K of G which generates G, that is $G = \langle K\rangle = \bigcup_{n=1}^{\infty}(K \cup K^{-1})^n$.*

Lemma 4.12. *If G is a compactly generated group then G is σ-compact.*

Proof. By the definition $G = \bigcup_{n=1}^{\infty}(K \cup K^{-1})^n$, where every $(K \cup K^{-1})^n$ is compact, since K is compact. □

It should be emphasized that while σ-compactness is a purely topological property, being compactly generated involves essentially the algebraic structure of the group.

Exercise 4.13. *(a) Give examples of σ-compact groups that are not compactly generated.*

(b) Show that every connected locally compact group is compactly generated.

Lemma 4.14. *Let G be a locally compact group.*

(a) If K a compact subset of G and U an open subset of G such that $K \subseteq U$, then there exists an open neighborhood V of 1 in G such that $(KV) \cup (VK) \subseteq U$ and $\overline{(KV) \cup (VK)}$ is compact.

(b) If G is compactly generated, then there exists an open neighborhood U of 1 in G such that $\overline{U}$ is compact and U generates G.

Proof. (a) By Lemma 4.3 (c) there exists an open neighborhood V of 1 in G such that $(KV) \cup (VK) \subseteq U$. Since G is locally compact, we can choose V with compact closure. Thus $K\overline{V}$ is compact by Lemma 4.3. Since $KV \subseteq K\overline{V}$, then $\overline{KV} \subseteq K\overline{V}$ and so $\overline{KV}$ is compact. Analogously $\overline{VK}$ is compact, so $\overline{(KV) \cup (VK)} = \overline{KV} \cup \overline{VK}$ is compact.

(b) Let K be a compact subset of G such that K generates G. So $K \cup \{1\}$ is compact and by (a) there exists an open neighborhood U of 1 in G such that $U \supseteq K \cup \{1\}$ and $\overline{U}$ is compact. □

In the case of first countable topological groups Fujita and Shakmatov [49] have described the precise relationship between σ-compactness and the property of being compactly generated.

Theorem 4.15. *A metrizable topological group G is compactly generated if and only if G is σ-compact and, for every open subgroup H of G, there exists a finite set $F \subseteq G$ such that $F \cup H$ algebraically generates G [49].*

This gives the following:

Corollary 4.16. *A σ-compact metrizable group G is compactly generated in each of the following cases (for the definition of total boundedness see Definition 4.25):*

- *G has no open subgroups*
- *the completion $\widetilde{G}$ is connected;*
- *G is totally bounded.*

Moreover,

Theorem 4.17. *A countable metrizable group is compactly generated iff it is algebraically generated by a sequence (possibly eventually constant) converging to its neutral element.*

Examples showing that the various conditions above cannot be omitted can be found in [49].

The question when will a topological group contain a compactly generated dense subgroup is considered in [50].

Now we see that linearity and total disconnectedness of group topologies coincide for compact groups and for locally compact abelian groups.

Theorem 4.18. *Every locally compact totally disconnected group has a base of neighborhoods of 1 consisting of open subgroups. In particular, a locally compact totally disconnected group that is either abelian or compact has linear topology.*

This can be derived from the followis more precise result:

Theorem 4.19. *Let G be a locally compact topological group and let $C = c(G)$. Then :*

(a) *C coincides with the intersection of all open subgroups of G;*

(b) *if G is totally disconnected, then every neighborhoodof 1 contains an open subgroup of G.*

If G is compact, then the open subgroups in items (a) and (b) can be chosen normal.

Proof. (a) follows from (b) as G/C is totally disconnected hence the neutral element of G/C is intersection of open (resp. open normal) subgroups of G/C. Now the intersection of the inverse images, w.r.t. the canonical homomorphism $G \to G/C$, of these subgroups coincides with C.

(b) Let G be a locally compact totally disconnted group. By Vedenissov's Theorem G has a base $\mathcal{O}$ of clopen symmetric compact neighborhoods of 1. Let $U \in \mathcal{O}$. The $U = \overline{U} = \bigcap_{V \in \mathcal{O}} UV$. Then every set $U \cdot V$ is compact by Lemma 4.3, hence closed. According to Lemma 2.14 there exist $V_1, \ldots, V_n \in \mathcal{O}$ such that $U = \bigcap_{k=1}^{n} UV_k$. Then for $V := U \cap \bigcap_{k=1}^{n} V_k$ one has $UV = U$. This implies also $VV \subseteq U$, $VVV \subseteq U$ etc. Since V is symmetric, the subgroup $H = \langle V \rangle$ is contained in U as well. From $V \subseteq H$ one can deduce that H is open (cf. 3.16). In case G is compact, note that the heart $H_G = \bigcap_{x \in G} x^{-1}Hx$ of H is an open normal subgroup as the number of all conjugates $x^{-1}Hx$ of H is finite (being equal to $[G : N_G(H)] \leq [G : H] < \infty$). Hence H_G is an open normal subgroup of G contained in H, hence also in U. □

Corollary 4.20. *The quotient of a locally compact totally disconnected group is totally disconnected.*

Proof. Let G be a locally compact totally disconnected group and let N be a closed normal subgroup of G. It follows from the above theorem that G has a linear topology. This yields that the quotient G/N has a linear topology too. Thus G/N is totally disconnected. □

Corollary 4.21. *The continuous homomorphic images of compact totally disconnected groups are totally disconnected.*

Proof. Follows from the above corollary and the open mapping theorem. □

According to Example 3.46 none of the items (a) and (b) of the theorem remain true without the hypothesis "locally compact".

Corollary 4.22. *Let G be a locally compact group. Then $Q(G) = c(G)$.*

Proof. By item (a) of the above theorema $C(G)$ is an intersection of open subgroups, that are clopen being open subgroups (cf. Proposizione 3.16). Hence $c(G)$ contains $Q(G)$ which in turn coincides with the intersection of *all* clopen sets of G containing 1. The inclusion $C(G) \subseteq Q(G)$ is always true. □

4.2 Subgroups of the compact groups

For a subset E of an abelian group G we set $E_{(2)} = E - E$, $E_{(4)} = E - E + E - E$, $E_{(6)} = E - E + E - E + E - E$ and so on.

A subset X of an abelian group $(G, +)$ is *big*[10] if there exists a finite subset F of G such that $G = X + F$. Obviously, every non-empty set of a finite group is big; on the other hand, every big set in an infinite group is necessarily infinite.

Exercise 4.23. *Let B be an infinite subset of $\mathbb{Z}$. Show that B is big iff the following two conditions hold:*

(a) B is unbounded from above and from below;

(b) if $B = \{b_n\}_{n=-\infty}^{\infty}$ is a one-to-one monotone enumeration of B then the differences $b_{n+1} - b_n$ are bounded.

Exercise 4.24. *(a) Assume B_ν is a big set of the group G_ν for $\nu = 1, 2, \ldots, n$. Prove that $B_1 \times \ldots \times B_n$ is a big set of $G_1 \times \ldots \times G_n$.*

(b) if $f : G \to H$ is a surjective group homomorphism and B is a big subset of H, then $f^{-1}(B)$ is a big subset of G.

Definition 4.25. A topological group G is *totally bounded* if every open non-empty subset U of G is big. A Hausdorff totally bounded group will be called *precompact*.

Clearly, compact groups are precompact.

Note that if f in item (b) of 4.24 is not surjective, then the property may fail. The next proposition gives an easy remedy to this.

Proposition 4.26. *Let A be an abelian group and let B be a big subset of A. Then $(B - B) \cap H$ is big with respect to H for every subgroup H of A.*

If $a \in A$ then there exists a sufficiently large positive integer n such that $na \in B - B$.

Proof. There exists a finite subset F of A such that $B + F = A$. For every $f \in F$, if $(B + f) \cap H$ is not empty, choose $a_f \in (B + f) \cap H$, and if $(B + f) \cap H$ is empty, choose an arbitrary $a_f \in H$. On the other hand, for every $x \in H$ there exists $f \in F$ such that $x \in B + f$; since $a_f \in B + f$, we have $x - a_f \in B - B$ and so $H \subseteq (B - B) \cap H + \{a_f : f \in F\}$, that is $(B - B) \cap H$ is big in H.

For the last assertion it suffices to take $H = \langle a \rangle$. If H is finite, then there is nothing to prove as $0 \in B - B$. Otherwise $H \cong \mathbb{Z}$ so the first item of Exercise 4.23 applies. □

Combining this proposition with item (b) of 4.24 we get:

Corollary 4.27. *For every group homomorphism $f : G \to H$ and every big subset B of H, the subset $f^{-1}(B-B)$ of G is a big.*

Here comes the most important consequence of the above proposition.

Corollary 4.28. *Subgroups of precompact groups are precompact. In particular, all subgroups of compact groups are precompact.*

One can show that the precompact groups are precisely the subgroups of the compact groups. This requires two steps as the next theorem shows:

Theorem 4.29. *(a) A group having a dense precompact subgroup is necessarily precompact.*

(b) The compact groups are precisely the complete precompact groups.

Proof. (a) Indeed, assume that H is a dense precompact subgroup of a group G. Then for every $U \in \mathcal{V}_G(0)$ choose an open $V \in \mathcal{V}_G(0)$ with $V + V \subseteq U$. By the precompactness of H there exists a finite set $F \subseteq H$ such that $H = F + V \cap H$. Then

$$G = V + H \subseteq V + F + V \cap H \subseteq F + V + V \subseteq F + U.$$

[10] Some authors use also the terminology *large*, *relatively dense*, or *syndetically dense*. This notion can be given for non-abelian groups as well, but then both versions, *left large* and *right large*, do not coincide. This creates some technical difficulties that we prefer to avoid since the second part of this section is relevant only for abelian groups. The first half, including the characterization 4.29, remains valid in the non-abelian case as well (since, fortunately, the "left" and "right" versions of total boundedness coincide).

(b) Compact groups are complete and precompact. To prove the other implication take a complete precompact group G. To prove that G is compact it sufficies to prove that every ultrafilre on G converges. Assume $\mathcal{U}$ is such an ultrafiler. We show first that it is a Cauchy filter. Indeed, if $U \in \mathcal{V}_G(0)$, then U is a big set of G so there exists $g_1, g_2, \dots, g_n \in G$ such that $G = \bigcup_{i=1}^n g_i + U$. Since $\mathcal{U}$ is an ultrafilter, $g_i + U \in \mathcal{U}$ for some i. Hence $\mathcal{U}$ is a is a Cauchy filter. According to Exercise 3.86 $\mathcal{U}$ converges. □

In this way we have described the precompact groups internally (as the Hausdorff topological groups having big non-empty open sets), or externally (as the subgroups of the compact groups).

Now we adopt a different approach to describe the precompact groups, based on the use of characters. Our first aim will be to see that the topologies induced by characters are always totally bounded.

Proposition 4.30. *If A is an abelian group, $\delta > 0$ and $\chi_1, \dots, \chi_s \in A^*$ ($s \in \mathbb{N}_+$), then $U(\chi_1, \dots, \chi_s; \delta)$ is big in A. Moreover for every $a \in A$ there exists a sufficiently large positive integer n such that $na \in U(\chi_1, \dots, \chi_s; \delta)$.*

Proof. Define $h : A \to \mathbb{T}^s$ such that $h(x) = (\chi_1(x), \dots, \chi_s(x))$ and

$$B = \left\{(z_1, \dots, z_n) \in \mathbb{S}^s : |\operatorname{Arg} z_i| < \frac{\delta}{2} \text{ for } i = 1, \dots, s\right\} = \left\{z \in \mathbb{S} : |\operatorname{Arg} z| < \frac{\delta}{2}\right\}^s.$$

Then B is big in $\mathbb{S}^s$ and by Proposition 4.26 the set $(B - B) \cap h(A)$ is big with respect to $h(A)$. Since

$$B - B \subseteq C = \{(z_1, \dots, z_s) \in \mathbb{S}^s : \|\operatorname{Arg} z_i\| < \delta \text{ for } i = 1, \dots, s\},$$

we have that $C \cap h(A)$ is big with respect to $h(A)$. Therefore $U(\chi_1, \dots, \chi_s; \delta) = h^{-1}(C)$ is big in A.

The second statement follows from Proposition 4.26, since

$$U\left(\chi_1, \dots, \chi_s; \frac{\delta}{2}\right) - U\left(\chi_1, \dots, \chi_s; \frac{\delta}{2}\right) \subseteq U(\chi_1, \dots, \chi_s; \delta).$$

□

Corollary 4.31. *For an abelian group G all topologies of the form $\mathcal{T}_H$, where $H \leq G^*$, are totally bounded. Moreover, $\mathcal{T}_H$ is precompact iff H separates the points of G.*

It requires a considerable effort to prove that, conversely, every totally bounded group topology has the form $\mathcal{T}_H$ for some H (see Remark 6.3).

It follows easily from Corollary 4.31 and Proposition 4.30 that for every neighborhood E of 0 in the Bohr topology (namely, a set E containing a subset of the form $U(\chi_1, \dots, \chi_n; \varepsilon)$ with characters $\chi_i : G \to \mathbb{S}$, $i = 1, 2, \dots, n$, and $\varepsilon > 0$) there exists a big set B of G such that $B_{(8)} \subseteq E$ (just take $B = U(\chi_1, \dots, \chi_n; \varepsilon/8)$). Surprisingly, the converse is also true. Namely, we shall obtain as a corollary of Følner's lemma that every set E satisfying $B_{(8)} \subseteq E$ for some big set B of G must be a neighborhood of 0 in the Bohr topology of G (see Corollary 5.8).

Exercise 4.32. *If G is a countably infinite Hausdorff abelian group, then for every compact set K in G the set $K_{(2n)}$ is big for no $n \in \mathbb{N}$.*

(Hint. By Lemma 4.3 every set $K_{(2n)}$ is compact. So if $K_{(2n)}$ were big for some n, then G itself would be compact. Now Lemma 4.8 applies.)

Exercise 4.33. *Call a subset S of an infinite abelian group G* small *if there exist (necessarily distinct) elements $g_1, g_2, \dots, g_n, \dots$ of G such that $(g_n + S) \cap (g_m + S) = \emptyset$ whenever $m \neq n$.*

(a) Show that a subset S of G such that $S - S$ is not big is necessarily small.

(b) Show that every finite subset is small.

(c) Show that the group $\mathbb{Z}$ is not a finite union of small sets.

*(d) * Show that no infinite abelian group G is a finite union of small sets.*

(e) If $S = (a_n)$ is a one-to-one T-sequence in an abelian group G, then for every $n \in \mathbb{N}$ the set $S_{(2n)}$ is small in G.

(f) Show that the sequence (p_n) of prime numbers in $\mathbb{Z}$ is not a T-sequence.

(Hint. (d) Use a finitely additive invariant (Banach) measure on G. For (e) consider the (countable) subgroup generated by S and note that if $a_n \to 0$ in some Hausdorff group topology τ on G, then the set $S \cup \{0\}$ would be compact in τ, so item (a) and Exercise 4.32 apply. For (f) use (e) and the fact that there exists a constant m such that every integer number is a sum of at most m[11] prime numbers.)

Exercise 4.34. *Show that for an infinite abelian group G and a subgroup H of G the following are equivalent:*

(a) H has infinite index;

(b) H is not big;

(c) H is small.

Exercise 4.35. * *Every infinite abelian group has a small set of generators.*

This can be extended to arbitrary groups [30]. One can find in the literature also different (weaker) forms of smallness ([4, 11]).

Exercise 4.36. *(a) If $f : G \to H$ is a continuous surjective homomorphism of topological groups, then H is totally bounded whenever G is totally bounded.*

(b) Prove that a topological group G is totally bounded iff $G/\overline{\{1\}}$ is totally bounded.

(c) If $\{G_i : i \in I\}$ is a family of topological groups, then $\prod_i G_i$ is totally bounded iff each G_i is totally bounded.

(d) Prove that every topological abelian group G admits a "universal" totally bounded continuous surjective homomorphic image $q : G \to q(G)$ (i.e., every continuous homomorphsm $G \to P$, where P is a totally bounded group, factors through q[12]).

4.3 Subgroups of $\mathbb{R}^n$

Our main goal here is to prove that every closed subgroup of $\mathbb{R}^n$ is topologically isomorphic to $\mathbb{R}^s \times \mathbb{Z}^m$, with $s, m \in \mathbb{N}$ and $s + m \leq n$. More precisely:

Theorem 4.37. *Let $n \in \mathbb{N}_+$ and let H be a closed subgroup of $\mathbb{R}^n$. Then there exist closed subgroups V and D of $\mathbb{R}^n$ such that $H = V + D \cong V \times D$, $V \cong \mathbb{R}^s$, $D \cong \mathbb{Z}^m$ and $s + m \leq n$.*

The proof is split in several steps. Before starting it, we note the following curious dichotomy hidden in this theorem:

- the closed connected subgroups of $\mathbb{R}^n$ are always isomorphic to some $\mathbb{R}^s$ with $s \leq n$;
- the totally disconnected closed subgroups D of $\mathbb{R}^n$ must be free and have free-rank $r_0(D) \leq n$; in particular they are discrete.

In the general case, for every closed subgroup H of $\mathbb{R}^n$ the connected component $c(H)$ is open in H and isomorphic to $\mathbb{R}^s$ for some $s \leq n$. Therefore, by the divisibility of $\mathbb{R}^s$ one can write $H = c(H) \times D$ for some discrete subgroup D of H (see Corollary 2.8). Necessarily $r_0(D) \leq n - s$ as $c(H) \cong \mathbb{R}^s$ contains a discrete subgroup D_1 of rank s, so that $D_1 \times D$ will be a discrete subgroup of $\mathbb{R}^n$.

It is not hard to see that every discrete subgroup of $\mathbb{R}$ is cyclic (Exercise 3.20). The first part of the proof consists in appropriately extending this property to discrete subgroups of $\mathbb{R}^n$ (see Proposition 4.39). The first step is to see that the free-rank $r_0(H)$ of a discrete subgroup H of $\mathbb{R}^n$ coincides with the dimension of the subspace of $\mathbb{R}^n$ generated by H.

Lemma 4.38. *Let H be a discrete subgroup of $\mathbb{R}^n$. If the elements $v_1, \ldots, v_m$ of H are $\mathbb{Q}$-linearly independent, then they are also $\mathbb{R}$-linearly indipendent.*

[11] use the fact that according to the positive solution of the ternary Goldbach's conjecture a there exists a constant $C > 0$ such that every odd integer $\geq C$ is a sum of three primes (see [96] for further details on Goldbach's conjecture).

[12] in other words, the subcategory of all totally bounded groups forms an epireflective subcategory of the category of all topological groups.

Proof. Let $V \cong \mathbb{R}^k$ be the subspace of $\mathbb{R}^n$ generated by H. We can assume wlog that $V = \mathbb{R}^n$, i.e., $k = n$. Hence we have to prove that the free-rank $m = r_0(H)$ of H coincides with n. Obviously $m \geq n$. We need to prove that $m \leq n$. Let us fix n $\mathbb{R}$-linearly independent vectors $v_1, \ldots, v_n$ in H. It is enough to see that for every $h \in H$ the vectors $v_1, \ldots, v_n, h$ are not $\mathbb{Q}$-linearly independent. This would imply $m \leq n$. Let us note first that we can assume wlog that $H \supseteq \mathbb{Z}^n$. Indeed, as $v_1, \ldots, v_n$ are $\mathbb{R}$-linearly independent, there exists a linear isomorphism $\alpha : \mathbb{R}^n \to \mathbb{R}^n$ with $\alpha(v_i) = e_i$ for $i = 1, 2, \ldots, n$, where $e_1, \ldots, e_n$ is the canonical base of $\mathbb{R}^n$. Clearly, $\alpha(H)$ is still a discrete subgroup of $\mathbb{R}^n$ and the vectors $v_1, \ldots, v_n, h$ are $\mathbb{Q}$-linearly independent iff the vectors $e_1 = \alpha(v_1), \ldots, e_n = \alpha(v_n), \alpha(h)$ are. The latter fact is equivalent to $\alpha(h) \notin \mathbb{Q}^n$. Therefore, arguing for a contradiction, assume for simplicity that $H \supseteq \mathbb{Z}^n$ and there exists $h = (h_1, \ldots, h_n) \in H$ such that

$$h \notin \mathbb{Q}^n. \tag{4}$$

By the discreteness of H there exists an $\varepsilon > 0$ with $\max\{|h_i| : i = 1, 2, \ldots, n\} \geq \varepsilon$ for every $0 \neq h = (h_1, \ldots, h_n) \in H$. Represent the cube $C = [0, 1]^n$ as a finite union $\bigcup_i C_i$ of cubes C_i of diameter $< \varepsilon$ (e.g., take them with faces parallel to the coordinate axes, although their precise position is completely irrelevant). For a real number r denote by $\{r\}$ the unique number $0 \leq x < 1$ such that $r - x \in \mathbb{Z}$. Then $(\{mv_1\}, \ldots, \{mv_n\}) \neq (\{lh_1\}, \ldots, \{lh_n\})$ for every positive $l \neq m$, since otherwise, $(m - l)h \in \mathbb{Z}^n$ with $m - l \neq 0$ in contradiction with (4). Among the infinitely many points $a_m = (\{mh_1\}, \ldots, \{mh_n\}) \in C$ there exist two $a_m \neq a_l$ belonging to the same cube C_i. Hence, $|\{mh_j\} - \{lh_j\}| < \varepsilon$ for every $j = 1, 2, \ldots, n$. So there exists a $z = (z_1, \ldots, z_n) \in \mathbb{Z}^n$, such that $0 \neq (m - l)h - z \in H$ and $|(m - l)h_j - z_j| < \varepsilon$ for every $j = 1, 2, \ldots, n$, this contradicts the choice of ε. □

The aim of the next step is to see that the discrete subgroups of $\mathbb{R}^n$ are free.

Proposition 4.39. *If H is a discrete subgroup of $\mathbb{R}^n$, then H is free and $r(H) \leq n$.*

Proof. In fact, let $m = r(H)$. By the definition of $r(H)$ there exist m $\mathbb{Q}$-linearly independent vectors $v_1, \ldots, v_m$ of H. By the previous lemma the vectors $v_1, \ldots, v_m$ are also $\mathbb{R}$-linearly independent. Hence, $m \leq n$. Let $V \cong \mathbb{R}^m$ be the subspace of $\mathbb{R}^n$ generated by $v_1, \ldots, v_m$. Obviously, $H \subseteq V$, since H is contained in the $\mathbb{Q}$-subspace of $\mathbb{R}^n$ generated by the free subgroup $F = \langle v_1, \ldots, v_m \rangle$ di H. Since H is a discrete subgroup of V too, we can argue with V in place of $\mathbb{R}^n$. So, we can assume wlog that $m = n$ and $V = \mathbb{R}^n$. It suffices to see that H/F is finite. Then H will be finitely generated and torsion-free, hence H must be free.

Since the vectors $v_1, \ldots, v_n$ are linearly independent on $\mathbb{R}$ we can assume wlog that $H \supseteq \mathbb{Z}^n$. In fact, let $\alpha : \mathbb{R}^n \to \mathbb{R}^n$ be the linear isomorphism with $\alpha(v_i) = e_i$ for $i = 1, 2, \ldots, n$, where $e_1, \ldots, e_n$ is the canonical base of $\mathbb{R}^n$. Then $\alpha(H)$ is still a discrete subgroup of $\mathbb{R}^n$, $\mathbb{Z}^n = \alpha(F) \subseteq \alpha(H)$ and H/F is finite iff $\alpha(H)/\alpha(F) \cong H/F$ is finite.

In the sequel we assume $H \supseteq \mathbb{Z}^n = F$ for the sake of simplicity. To check that H/F is finite consider the canonical homomorphism $q : \mathbb{R}^n \to \mathbb{R}^n/\mathbb{Z}^n \cong \mathbb{T}^n$. According to Theorem 3.23, q sends the closed subgroup H onto a closed (hence compact) subgroup $q(H)$ of $\mathbb{T}^n$; moreover $H = q^{-1}(q(H))$, hence the restriction of q to H is open and $q(H)$ is discrete. Thus $q(H) \cong H/F$ is both compact and discrete, so $q(H)$ must be finite. □

Now we are in position to prove Theorem 4.37. We advise the reader to review the warming exercise 3.20.

Proof of Theorem 4.37. Let $H \neq 0$ a closed subgroup of $\mathbb{R}^n$. If H discrete, then H is free and generates a linear subspace of $\mathbb{R}^n$ of dimension $r(H) \leq n$ by Proposition 4.39, so the assertion is true with $s = 0$.

In case H is not discrete we argue by induction on n. The case $n = 1$ is Exercise 3.20. Let $n > 1$ and assume the theorem is true for $n - 1$. Consider the subset

$$M = \{u \in \mathbb{R}^n : \|u\| = 1 \text{ and } \exists \lambda \in (0, 1) \text{ with } \lambda u \in H\}$$

of the unitary sphere S in $\mathbb{R}^n$. For $u \in S$ let $N_u = \{r \in \mathbb{R} : ru \in H\}$. Then N_u is a closed subgroup of $\mathbb{R}$ and $H \cap \mathbb{R}u = N_u u$. Our aim will be to find some $u \in S$ such that the whole line $\mathbb{R}u$ is contained in H. This will allow us to use our inductive hypothesis. Since the proper closed subgroups of $\mathbb{R}$ are cyclic (see Exercise 3.20), it suffices to find some $u \in S$ such that N_u is not cyclic.

Case 1. If $M = \{u_1, \ldots, u_n\}$ is finite, then there exists an index i such that $\lambda u_i \in H$ for infinitely many $\lambda \in (0, 1)$. Then the closed subgroup N_{u_i} cannot be cyclic, so H contains to line $\mathbb{R}u_i$ and we are done.

Case 2. Assume M is infinite. By the assumption H is not discrete there exists a sequence $u_n \in M$ such that the corresponding λ_n, with $\lambda_n u_n \in H$, converge to 0. By the compactness of S there exists a limit point $u_0 \in S$ for the sequence $u_n \in M$. We can assume wlog that $u_n \to u_0$. Let $\varepsilon > 0$ and let Δ_ε be the open interval $(\varepsilon, 2\varepsilon)$. As $\lambda_n \to 0$, there exists n_0 such that $\lambda_n < \varepsilon$ for every $n \geq n_0$. Hence for every $n \geq n_0$ there exists an

appropriate $k_n \in \mathbb{N}$ with $\eta_n = k_n\lambda_n \in \Delta_\varepsilon$. Taking again a subsequence we can assume wlog that there exists some $\xi_\varepsilon \in \overline{\Delta_\varepsilon}$ such that $\eta_n \to \xi_\varepsilon$. Hence $\xi_\varepsilon u_0 = \lim_n k_n\lambda_n u_0 \in H$. This argument shows that N_{u_0} contains $\xi_\varepsilon \in \overline{\Delta_\varepsilon}$ with arbitrarily small ε. Therefore, N_{u_0} cannot be cyclic. Hence H contains the line $\mathbb{R}u_0$.

We proved in all cases that our assumption of non-discreteness of H yields the existence of a line $L \cong \mathbb{R}$ as a subgroup of H. Let $L' \cong \mathbb{R}^{n-1}$ be a subspace of $\mathbb{R}^n$ complementing L. Then $\mathbb{R}^n = L \times L'$ and the projection $\mathbb{R}^n \to L'$ sends H to a closed subgroup H_1 of L' as $L \leq H$ (cf. 3.23 (b)). Moreover, $H = L \times H_1$ in view of $L \leq H$ again. Now proceed by induction with the subgroup H_1 of $L' \cong \mathbb{R}^{n-1}$. This proves Theorem 4.37.

The next corollary describes the quotients of $\mathbb{R}^n$.

Corollary 4.40. *A quotient of $\mathbb{R}^n$ is isomorphic to $\mathbb{R}^k \times \mathbb{T}^m$, where $k + m \leq n$. In particular, a compact quotient of $\mathbb{R}^n$ is isomorphic to $\mathbb{T}^m$ for some $m \leq n$.*

Proof. Let H be a closed subgroup of $\mathbb{R}^n$. Then $H = V + D$, where V, D are as in Theorem 4.37. If $s = \dim V$ and $m = r_0(D)$, then $s + m \leq n$. Let V_1 be the linear subspace of $\mathbb{R}^n$ spanned by D. Pick a complementing subspace V_2 for the subspace $V + V_1$ and let $k = n - (s + m)$. Then $\mathbb{R}^n = V + V_1 + V_2$ is a factorization in direct product. Therefore $\mathbb{R}^n/H \cong (V_1/D) \times V_2$. Since $\dim V_1 = r_0(D) = m$, one has $V_1/D \cong \mathbb{T}^m$. Therefore, $\mathbb{R}^n/H \cong \mathbb{T}^m \times \mathbb{R}^k$. □

Let us denote by $(x|y)$ the usual scalar product in $\mathbb{R}^n$. Recall that every base $v_1, \ldots, v_n$ di $\mathbb{R}^n$ admits a dual base $v'_1, \ldots, v'_n$ defined by the relations $(v_i|v'_j) = \delta_{ij}$. For a subgroup H of $\mathbb{R}^n$ define the *associated subgroup* $H^\dagger$ setting

$$H^\dagger := \{u \in \mathbb{R}^n : (\forall x \in H)(x|u) \in \mathbb{Z}\}.$$

Then obviously $(\mathbb{Z}^n)^\dagger = \mathbb{Z}^n$.

Lemma 4.41. *Let H be a subgroup di $\mathbb{R}^n$. Then:*

1. *$H^\dagger$ is a closed subgroup of $\mathbb{R}^n$ and the correspondence $H \mapsto H^\dagger$ is decreasing;*
2. *$(\overline{H})^\dagger = H^\dagger$.*

Proof. The map $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by $(x, y) \mapsto (x|u)$ is continuous.

(a) Let $q : \mathbb{R} \to \mathbb{T} = \mathbb{R}/\mathbb{Z}$ be the canonical homomorphism.Then for every fixed $a \in \mathbb{R}^n$ the assignment $x \mapsto (a|x) \mapsto f((a|x))$ is a continuous homomorphism $\chi_a : \mathbb{R}^n \to \mathbb{T}$. Hence the set $\chi_h^{-1}(0) = \{u \in \mathbb{R}^n : (\forall h \in H)(h|u) \in \mathbb{Z}\}$ is closed in $\mathbb{R}^n$. Therefore $H^\dagger = \bigcap_{h\in H} \chi_h^{-1}(0)$ is closed. The same equality proves that the correspondence $H \mapsto H^*$ is decreasing.

(b) From the second part of (a) one has $(\overline{H})^\dagger \subseteq H^\dagger$. Suppose that $u \in H^\dagger$ e $x \in \overline{H}$. By the continuity of the map $\chi_x(u) = \chi_u(x)$, as a function of x, one can deduce that $\chi_x(u) = 0$, being $\chi_u(h) = 0$ for every $h \in H$. □

We study in the sequel the subgroup $H^\dagger$ associated to a closed subgroup H of $\mathbb{R}^n$. According to Theorem 4.37 there exist a base $v_1, \ldots, v_n$ of $\mathbb{R}^n$ and $k \leq n$, such that for some $0 \leq s \leq k$ $H = V \oplus L$ where V is the linear subspace generated by $v_1, \ldots, v_s$ and $L = \langle v_{s+1}, \ldots, v_k\rangle$. Let $v'_1, \ldots, v'_n$ be the dual base of $v_1, \ldots, v_n$.

Lemma 4.42. *In the above notation the subgroup $H^\dagger$ coincides with $\langle v'_{s+1}, \ldots, v'_k\rangle + W$, where W is the linear subspace generated by $v'_{k+1}, \ldots, v'_n$.*

Proof. Let V' be the linear subspace generated by $v'_1, \ldots, v'_s$, V'' the linear subspace generated by $v'_{s+1}, \ldots, v'_k$ and $L' = \langle v'_{s+1}, \ldots, v'_k\rangle$. Then $L^\dagger = V' + L' + W$, while $V^\dagger = V'' + W$. Hence $H^\dagger \leq L^\dagger \cap V^\dagger = L' + W$. On the other hand, obviously $L' + W \leq H^\dagger$. □

Corollario 4.43. *$\overline{H} = (H^\dagger)^\dagger$ for every subgroup H of $\mathbb{R}^n$.*

Proof. If H is closed of the form $V + L$ in the notation of the previous lemma, then $H^\dagger = L' + W$ with $v'_1, \ldots, v'_n$, L' and W defined as above. Now $H^\dagger = L' + W$ is a closed subgroup of $\mathbb{R}^n$ by Lemma 4.41 and $v_1, \ldots, v_n$ is a dual base of $v'_1, \ldots, v'_n$. Therefore, $H = V + L$ coincides with $(H^\dagger)^\dagger$. □

Lemma 4.44. *Let V be a hyperplain in $\mathbb{R}^n$ determined by the equation $\sum_{i=1}^n a_i x_i = 0$ such that there exists at least one coefficient $a_i = 1$. Then the subgroup $H = V + \mathbb{Z}^n$ of $\mathbb{R}^n$ is not dense iff all the coefficients a_i are rational.*

Proof. We can assume wlog that $i = n$. Suppose that H is not dense in $\mathbb{R}^n$. Then $H^\dagger \neq 0$ by Corollary 4.43. Let $0 \neq z \in H^\dagger$. Since $\mathbb{Z}^n \leq H$, one has $H^\dagger \leq \mathbb{Z}^n = (\mathbb{Z}^n)^\dagger$, so $z \in \mathbb{Z}^n$. If $j < n$, then $a_j \in \mathbb{Q}$ as $v = (0, \ldots, 0, 1, 0, \ldots, 0, -a_j) \in V$ □

Unit 5

5 Følner's theorem

This section is entirely dedicated to Følner's theorem.

5.1 Fourier theory for finite abelian groups

In the sequel G will be a finite abelian group, so $G^* \cong G$, so in particular $|G^*| = |G|$.

Here we recall some well known properties of the scalar product in finite-dimensional complex spaces $V = \mathbb{C}^n$. Since our space will be "spanned" by a finite abelian group G of size n (i.e., $V = \mathbb{C}^G$), we have also an action of G on V. We normalize the scalar product in a such way to let the vector $(1, 1, \ldots, 1)$ (i. e., the constant function 1) to have norm 1. The reader familiar with Haar integration may easily recognize in this the Haar integral on G.

Define the scalar product by

$$(f, g) = \frac{1}{|G|} \sum_{x \in G} f(x)\overline{g(x)}.$$

Let us see first that the elements of the subset G^* of V are pairwise orthogonal and have norm 1:

Proposition 5.1. *Let G be an abelian finite group and $\varphi, \chi \in G^*$, $x, y \in G$. Then:*

(a) $\frac{1}{|G|}\sum_{x\in G}\varphi(x)\overline{\chi(x)} = \begin{cases}1 & \text{if } \varphi=\chi\\ 0 & \text{if } \varphi\neq\chi\end{cases}$;

(b) $\frac{1}{|G^*|}\sum_{\chi\in G^*}\chi(x)\overline{\chi(y)} = \begin{cases}1 & \text{if } x=y\\ 0 & \text{if } x\neq y.\end{cases}$.

Proof. (a) If $\varphi=\chi$ then $\chi(x)\overline{\chi(x)} = \chi(x)\chi(x)^{-1} = 1$.

If $\varphi\neq\chi$ there exists $z\in G$ such that $\varphi(z)\neq\chi(z)$. Therefore the following equalities

$$\sum_{x\in G}\varphi(x)\overline{\chi(x)} = \frac{\varphi(z)}{\chi(z)}\sum_{x\in G}\varphi(x-z)\overline{\chi(x-z)} = \frac{\varphi(z)}{\chi(z)}\sum_{x\in G}\varphi(x)\overline{\chi(x)}$$

imply that $\sum_{x\in F}\varphi(x)\overline{\chi(x)} = 0$.

(b) If $x=y$ then $\chi(x)\overline{\chi(x)} = \chi(x)\chi(x)^{-1} = 1$.

If $x\neq y$, by Corollary 2.7 there exists $\chi_0\in G^*$ such that $\chi_0(x)\neq\chi_0(y)$. Now we can proceed as before, that is

$$\sum_{\chi\in G^*}\chi(x)\overline{\chi(y)} = \frac{\chi_0(x)}{\chi_0(y)}\sum_{\chi\in G^*}(\chi\chi_0)(x)\overline{(\chi\chi_0)(y)} = \frac{\chi_0(x)}{\chi_0(y)}\sum_{\chi\in G^*}\chi(x)\overline{\chi(y)}$$

yields $\sum_{\chi\in G^*}\chi(x)\overline{\chi(y)} = 0$. □

If G is a finite abelian group and f is a complex valued function on G, then for every $\chi\in G^*$ we can define

$$c_\chi = (f,\chi) = \frac{1}{|G|}\sum_{x\in G}f(x)\overline{\chi(x)},$$

that is the *Fourier coefficient* of f corresponding to χ.

For complex valued functions f, g on a finite abelian group G define the *convolution* $f*g$ by $(f*g)(x) = \frac{1}{|G|}\sum_{y\in G}\overline{f(y)}g(x+y)$.

Proposition 5.2. *Let G be an abelian finite group and f a complex valued function on G with Fourier coefficients c_χ where $\chi\in G^*$. Then:*

(a) $f(x) = \sum_{\chi\in G^*}c_\chi\chi(x)$ *for every* $x\in G$;

(b) if $\{a_\chi\}_{\chi\in G^*}$ *is such that* $f(x) = \sum_{\chi\in G^*}a_\chi\chi(x)$, *then* $a_\chi = c_\chi$ *for every* $\chi\in G^*$;

(c) $\frac{1}{|G|}\sum_{x\in G}|f(x)|^2 = \sum_{\chi\in G^*}|c_\chi|^2$;

(d) if g is an other complex valued function on G with Fourier coefficients $(d_\chi)_{\chi\in G^*}$, then $f*g$ has Fourier coefficients $(\overline{c_\chi}d_\chi)_{\chi\in G^*}$.

Proof. (a) The definition of the coefficients c_χ yields

$$\sum_{\chi\in G^*}c_\chi\chi(x) = \sum_{\chi\in G^*}\frac{1}{|G|}\sum_{y\in G}f(y)\overline{\chi(y)}\chi(x).$$

Computing $\sum_{\chi\in G^*}\overline{\chi(y)}\chi(x)$ with Proposition 5.1(b) we get $\sum_{\chi\in G^*}c_\chi\chi(x) = \frac{|G^*|}{|G|}f(x)$ for every $x\in G$. Now $|G| = |G^*|$ gives $f(x) = \sum_{\chi\in G^*}c_\chi\chi(x)$ for every $x\in G$.

(b) By Proposition 5.1 the definition of the coefficients c_χ and the relation $f(x) = \sum_{\chi\in G^*}a_\chi\chi(x)$

$$c_\chi = \frac{1}{|G|}\sum_{\varphi\in G^*}a_\varphi\sum_{x\in G}\varphi(x)\overline{\chi(x)} = a_\chi.$$

(d) By item (a) $g(x) = \sum_{\varphi\in G^*}d_\varphi\varphi(x)$ for every $x\in G$. Therefore

$$\begin{aligned}\sum_{y\in G}\overline{f(y)}g(x+y) &= \sum_{y\in G}\Big(\sum_{\chi\in G^*}\overline{c_\chi}\overline{\chi(y)}\Big)\Big(\sum_{\varphi\in G^*}d_\varphi\varphi(x)\varphi(y)\Big) =\\ &= \sum_{\chi\in G^*}\sum_{\varphi\in G^*}\overline{c_\chi}d_\varphi\varphi(x)\sum_{y\in G}\overline{\chi(y)}\varphi(y) = |G|\sum_{\chi\in G^*}\overline{c_\chi}d_\chi\chi(x).\end{aligned}$$

(c) It is sufficient to apply (d) with $g=f$ and let $x=0$. □

Corollary 5.3. *Let G be a finite abelian group, E be a non-empty subset of G and let f be the characteristic function of E. Then for the convolution $g = f * f$ one has*

(a) $g(x) > 0$ *iff* $x \in E_{(2)}$;

(b) $g(x) = \sum_{\chi \in G^*} |c_\chi|^2 \chi(x)$.

Proof. (a) $g(x) > 0$ if and only if there exists $y \in E$ with $x + y \in E$, that is $x \in E - E = E_{(2)}$.
(b) follows obviously from Proposition 5.2(d). □

5.2 Bogolioubоff and Følner Lemmas

Lemma 5.4 (Bogolioubоff lemma). *If F is a finite abelian group and E is a non-empty subset of F, then there exist $\chi_1, \ldots, \chi_m \in F^*$, where $m = \left[\left(\frac{|F|}{|E|}\right)^2\right]$, such that $U(\chi_1, \ldots, \chi_m; \frac{\pi}{2}) \subseteq E_{(4)}$.*

Proof. Let f be the characteristic function of E. By Proposition 5.2(a) we have

$$f(x) = \sum_{\chi \in F^*} c_\chi \chi(x), \text{ with } c_\chi = \frac{1}{|F|} \sum_{x \in F} f(x)\overline{\chi(x)}. \tag{1}$$

Define $g = f * f$ and $h = g * g$. The functions f and $g = f * f$ have real values and by Corollary 5.3

$$g(x) = \sum_{\chi \in F^*} |c_\chi|^2 \chi(x) \text{ and } h(x) = \sum_{\chi \in F^*} |c_\chi|^4 \chi(x) \text{ for } x \in F. \tag{2}$$

Moreover, $g(x) > 0$ if and only if $x \in E - E = E_{(2)}$. Analogously $h(x) > 0$ if and only if $x \in E_{(4)}$.

By Proposition 5.2(c) $\sum_{\chi \in F'} |c_\chi|^2 = \frac{|E|}{|F|}$. Set $a = \frac{|E|}{|F|}$ and order the Fourier coefficients of f so that

$$|c_{\chi_0}| \geq |c_{\chi_1}| \geq \ldots \geq |c_{\chi_k}| \geq \ldots$$

(note that they are finitely many). Thus $\chi_0 = 1$ and $c_{\chi_0} = a$ by (1). Then $\sum_{i=0}^{k} |c_{\chi_i}|^2 \leq \sum_{\chi \in F^*} |c_\chi|^2 = a$ for every $k \geq 0$. Consequently $(k+1)|c_{\chi_k}|^2 \leq a$, so

$$|c_{\chi_k}|^4 \leq \frac{a^2}{(k+1)^2}. \tag{3}$$

Now let $m = \left[\frac{1}{a^2}\right]$. We are going to show now that with these $\chi_1, \ldots, \chi_m \in F^*$ one has

$$h(x) > 0 \text{ for every } x \in U(\chi_1 \ldots, \chi_m; \frac{\pi}{2}). \tag{4}$$

Clearly $Re\chi_k(x) \geq 0$ for $k = 1, 2, ..., m$ whenever $x \in U(\chi_1 \ldots, \chi_m; \frac{\pi}{2})$ thus

$$|a^4 + \sum_{k=0}^{m} |c_{\chi_k}|^4 \chi_k(x)| \geq Re(a^4 + \sum_{k=1}^{m} |c_{\chi_k}|^4 \chi_k(x)) \geq a^4. \tag{5}$$

On the other hand, (3) yields

$$\sum_{k \geq m+1} |c_{\chi_k}|^4 \leq \sum_{k \geq m+1} \frac{a^2}{(k+1)^2} < a^2 \sum_{k \geq m+1} \frac{1}{k(k+1)} \leq \frac{a^2}{m+1}. \tag{6}$$

Since h has real values, (2), (5) and (6) give

$$h(x) = |h(x)| = |a^4 + |c_{\chi_1}|^4 \chi_1(x) + \ldots| \geq \left|a^4 + \sum_{k=1}^{m} |c_{\chi_k}|^4 \chi_k(x)\right| - \sum_{k \geq m+1} |c_{\chi_k}|^4 \geq a^4 - \frac{a^2}{m+1} \geq a^2(a^2 - \frac{1}{m+1}) > 0.$$

This proves (4). Therefore $U(\chi_1 \ldots, \chi_m; \frac{\pi}{2}) \subseteq E_{(4)}$. □

Let us note that the estimate for the number m of characters is certainly non-optimal when E is too small. For example, when E is just the singleton $\{0\}$, the upper bound given by the lemma is just $|F|^2$, while one can certainly find at most $m = |F| - 1$ characters $\chi_1, \ldots, \chi_m$ (namely, all non-trivial $\chi_i \in F^*$) such that $U(\chi_1, \ldots, \chi_m; \frac{\pi}{2}) = \{0\}$. For certain groups (e.g., $F = \mathbb{Z}_2^k$) one can find even a much smaller number (say $m = \log_2 |F|$). Nevertheless, in the cases relevant for the proof of Følner's theorem, namely when the subset E is relatively large with respect to F, this estimate seems more reasonable.

The next lemma will be needed in the following proofs.

Lemma 5.5. *Let A be an abelian group and $\{A_n\}_{n=1}^{\infty}$ be a sequence of finite subsets of A such that*

$$\lim_{n\to\infty} \frac{|(A_n - a) \cap A_n|}{|A_n|} = 1$$

for every $a \in A$. If k is a positive integer and V is a subset of A such that k translates of V cover A, then for every $\varepsilon > 0$ there exists $N > 0$ such that

$$|V \cap A_n| > \left(\frac{1}{k} - \varepsilon\right) |A_n|$$

for every $n \geq N$.

Proof. Let $a_1, \ldots, a_k \in A$ be such that $\bigcup_{i=1}^{k}(a_i + V) = A$. If $\varepsilon > 0$, then there exists $N_1 > 0$ such that for every $n \geq N_1$

$$|(A_n - a_i) \cap A_n| > (1 - \varepsilon)|A_n|$$

and consequently,

$$|(A_n - a_i) \setminus A_n| < \varepsilon |A_n| \tag{7}$$

for every $i = 1, \ldots, k$. Since $A_n = \bigcup_{i=1}^{k}(a_i + V) \cap A_n$, for every n there exists $i_n \in \{1, \ldots, k\}$ such that

$$\frac{1}{k}|A_n| \leq |(a_{i_n} + V) \cap A_n| = |V \cap (A_n - a_{i_n})|.$$

Since $V \cap (A_n - a_{i_n}) \subseteq (V \cap A_n) \cup ((A_n - a_{i_n}) \setminus A_n)$, (7) yields

$$\frac{1}{k}|A_n| \leq |V \cap (A_n - a_{i_n})| \leq |V \cap A_n| + |(A_n - a_{i_n}) \setminus A_n| < |V \cap A_n| + \varepsilon |A_n|.$$

□

Lemma 5.6 (Bogoliouboff-Følner lemma). *Let A be a finitely generated abelian group and let $r = r_0(A)$. If k is a positive integer and V is a subset of A such that k translates of V cover A, then there exist $\rho_1, \ldots, \rho_s \in A^*$, where $s = 3^{2r}k^2$, such that $U_A(\rho_1, \ldots, \rho_s; \frac{\pi}{2}) \subseteq V_{(4)}$.*

Proof. By Theorem 2.1 we have $A \cong \mathbb{Z}^r \times F$, where F is a finite abelian group; so we can identify A with the group $\mathbb{Z}^r \times F$. Define $A_n = (-n, n]^r \times F$, let $a = (a_1, \ldots, a_r; f) \in \mathbb{Z}^r \times F$. Then $J_{ni} = (-n, n] \cap (-n - a_i, n - a_i]$ satisfies $|J_{ni}| \geq 2n - |a_i|$. In particular, $J_{ni} \neq \emptyset$ for every $n > n_0 = \max\{|a_i| : i = 1, 2, \ldots, n\}$. As $(A_n - a) \cap A_n = \prod_{i=1}^{r} J_{ni} \times F$, we have

$$|(A_n - a) \cap A_n| \geq |F| \cdot \prod_{i=1}^{r}(2n - |a_i|)$$

or all $n > n_0$. Since $|A_n| = |F|(2n)^r$, we can apply Lemma 5.5. Thus for every $\varepsilon > 0$ we have

$$|V \cap A_n| > \left(\frac{1}{k} - \varepsilon\right) |A_n|. \tag{8}$$

for every sufficiently large n. For n with (8) define $G = A/(6n\mathbb{Z}^r)$ and $E = q(V \cap A_n)$ where q is the canonical projection of A onto G. Observe that $q \restriction_{A_n}$ is injective, as $(A_n - A_n) \cap \ker q = \{0\}$. Then (8) gives

$$|E| = |V \cap A_n| > \left(\frac{1}{k} - \varepsilon\right)|A_n| = \left(\frac{1}{k} - \varepsilon\right)(2n)^r |F|$$

and so

$$\frac{|G|}{|E|} \leq \frac{(6n)^r |F|}{(\frac{1}{k} - \varepsilon)(2n)^r |F|} = \frac{3^r k}{1 - k\varepsilon}.$$

Fix $\varepsilon > 0$ sufficiently small to have $\left[\frac{3^{2r}k^2}{(1-k\varepsilon)^2}\right] = 3^{2r}k^2$ and pick sufficiently large n to have (8). Now apply the Bogoliouboff Lemma 5.4 to find $s = 3^{2r}k^2$ characters $\xi_{1n}, \ldots, \xi_{sn} \in G^*$ such that $U_G(\xi_{1n}, \ldots, \xi_{sn}; \frac{\pi}{2}) \subseteq E_{(4)}$. For $j = 1, \ldots, s$ define $\varrho_{jn} = \xi_{jn} \circ \pi \in A^*$. If $a \in A_n \cap U_A(\varrho_{1n}, \ldots, \varrho_{sn}; \frac{\pi}{2})$ then $q(a) \in U_A(\xi_{1n}, \ldots, \xi_{sn}; \frac{\pi}{2}) \subseteq E_{(4)}$ and so there exist $b_1, b_2, b_3, b_4 \in V \cap A_n$ and $c = (c_i) \in 6n\mathbb{Z}^r$ such that $a = b_1 - b_2 + b_3 - b_4 + c$. Now

$$c = a - b_1 + b_2 - b_3 + b_4 \in (A_n)_{(4)} + A_n$$

implies $|c_i| \leq 5n$ for each i. So $c = 0$ as $6n$ divides c_i for each i. Thus $a \in V_{(4)}$ and so

$$A_n \cap U_A\left(\varrho_{1n}, \ldots, \varrho_{sn}; \frac{\pi}{2}\right) \subseteq V_{(4)} \tag{9}$$

for all n satisfying (8).

By Lemma 4.2 there exist $\varrho_1, \ldots, \varrho_s \in A^*$ and a subsequence $\{n_l\}_l$ of $\{n\}_{n\in\mathbb{N}_+}$ such that $\varrho_i(a) = \lim_l \varrho_{in_l}(a)$ for every $i = 1, \ldots, s$ and $a \in A$. We are going to prove now that

$$U_A\left(\varrho_1, \ldots \varrho_s; \frac{\pi}{2}\right) \subseteq V_{(4)}. \tag{10}$$

Take $a \in U_A(\varrho_1, \ldots, \varrho_s; \frac{\pi}{2})$. Since $A = \bigcup_{l=k}^{\infty} A_{n_l}$ for every $k \in \mathbb{N}_+$, there exists n_0 satisfying (8) and $a \in A_{n_0}$. As $\varrho_i(a) = \lim_l \varrho_{in_l}(a)$ for every $i = 1, \ldots, s$, we can pick l to have $n_l \geq n_0$ and $|Arg(\varrho_{in_l}(a))| < \pi/2$ for every $i = 1, \ldots, s$, i.e., $a \in U_A(\varrho_{1n_l}, \ldots, \varrho_{sn_l}; \frac{\pi}{2}) \cap A_{n_l}$. Now (9), applied to n_l, yields $a \in V_{(4)}$. This proves (10). □

Our next aim is to eliminate the dependence of the number m of characters on the free rank of the group A in Bogoliouboff - Følner's lemma. The price to pay for this is taking $V_{(8)}$ instead of $V_{(4)}$.

Lemma 5.7 (Følner lemma). *Let A be an abelian group. If k is a positive integer and V be a subset of A such that k translates of V cover A, then there exist $\chi_1, \ldots, \chi_m \in A^*$, where $m = k^2$, such that $U_A(\chi_1, \ldots, \chi_m; \frac{\pi}{2}) \subseteq V_{(8)}$.*

Proof. We consider first the case when A is finitely generated. Let $r = r_0(A)$. By Lemma 5.6 there exist $\varrho_1, \ldots, \varrho_s \in A^*$, where $s = 3^{2r}k^2$, such that

$$U_A\left(\varrho_1, \ldots, \varrho_s; \frac{\pi}{2}\right) \subseteq V_{(4)}.$$

Since it is finitely generated, we can identify A with $\mathbb{Z}^r \times F$, where F is a finite abelian group. For $t \in \{1, \ldots, r\}$ define a monomorphism $i_t : \mathbb{Z} \hookrightarrow A$ by letting

$$i_t(n) = (\underbrace{0, \ldots, 0, n}_{t}, 0, \ldots, 0; 0) \in A.$$

Then each $\kappa_{jt} = \varrho_j \circ i_t$, where $j \in \{1, \ldots, s\}, t \in \{1, \ldots, r\}$, is a character of $\mathbb{Z}$. By Proposition 4.30 the subset

$$L = U_{\mathbb{Z}}\left(\{\kappa_{jt} : j = 1, \ldots, s, t = 1, \ldots, r\}; \frac{\pi}{8r}\right)$$

of $\mathbb{Z}$ is infinite. Let $L^0 = \bigcup_{t=1}^{r} i_t(L)$, i.e., this is the set of all elements of A of the form $\pm i_t(n)$ with $n \in L$ and $t \in \{1, \ldots, r\}$. Then obviously $L^0 = -L^0 \subseteq U_A\left(\varrho_1, \ldots, \varrho_s; \frac{\pi}{8r}\right)$, therefore,

$$L^0_{(4r)} \subseteq U_A\left(\varrho_1, \ldots, \varrho_s; \frac{\pi}{2}\right) \subseteq V_{(4)}. \tag{λ}$$

Define $A_n = (-n, n]^r \times F$ and pick $\varepsilon > 0$ such that $\left[\left(\frac{k}{1-k\varepsilon}\right)^2\right] = k^2$. As in Lemma 5.6 A_n satisfies the hypotheses of Lemma 5.5 and so $|V \cap A_n| > (\frac{1}{k} - \varepsilon)|A_n|$ for sufficiently large n. Moreover, we choose this sufficiently large n from L. Let $G_n = A/(2n\mathbb{Z}^r) \cong \mathbb{Z}_{2n}^r \times F$ and $E = q(A_n \cap V)$ where q is the canonical projection $A \to G_n$. Then $q \restriction_{A_n}$ is injective as $(A_n - A_n) \cap \ker q = 0$. So q induces a bijection between A_n and G_n on one hand, and between $V \cap A_n$ and E. Thus $|A_n| = |G_n| = (2n)^r|F|$, $|E| > (\frac{1}{k} - \varepsilon)|A_n|$ and so

$$\left(\frac{|G_n|}{|E|}\right)^2 \leq \left(\frac{k}{\varepsilon k - 1}\right)^2 \leq k^2.$$

To the finite group G_n apply the Bogoliouboff Lemma 5.4 to get $\xi_{1n}, \ldots, \xi_{mn} \in G_n^*$, where $m = k^2$, such that

$$U_{G_n}\left(\xi_{1n}, \ldots, \xi_{mn}; \frac{\pi}{2}\right) \subseteq E_{(4)}.$$

Let $\chi_{jn} = \xi_{jn} \circ q \in A^*$. If $a \in A_n \cap U_A(\chi_{1n}, \ldots, \chi_{mn}; \frac{\pi}{2})$, then $q(a) \in U_{G_n}(\xi_{1n}, \ldots, \xi_{mn}; \frac{\pi}{2}) \subseteq E_{(4)}$. Therefore there exist $b_1, b_2, b_3, b_4 \in A_n \cap V$ and $c = (c_i) \in 2n\mathbb{Z}^r$ such that $a = b_1 - b_2 + b_3 - b_4 + c$. Since $2n$ divides c_i for every i and $|c_i| \leq 5n$, we conclude that $c_i \in \{0, \pm 2n \pm 4n\}$ for $i = 1, 2, \ldots, r$. This means that c can be written as a sum of at most $4r$ elements of L^0. This gives $c \in L^0_{(4r)} \subseteq V_{(4)}$ by (λ), consequently $a \in V_{(8)}$. Therefore

$$A_n \cap U_A\left(\chi_{1n}, \ldots, \chi_{mn}; \frac{\pi}{2}\right) \subseteq V_{(8)}$$

for $n \in L$ sufficiently large n. By Lemma 4.2 there exist $\chi_1, \dots, \chi_m \in A^*$ and a subsequence $\{n_l\}_l$ of $\{n\}_{n\in\mathbb{N}_+}$ such that $\chi_j(a) = \lim_l \chi_{jn_l}(a)$ for every $j = 1, \dots, m$ and for every $a \in A$. Being $A = \bigcup\{A_n : l > k, n_l \in L\}$ for every $k \in \mathbb{N}_+$ we can conclude as above that $U_A\left(\chi_1, \dots, \chi_m; \frac{\pi}{2}\right) \subseteq V_{(8)}$.

Consider now the general case. Let $g_1, \dots, g_k \in A$ be such that $A = \bigcup_{i=1}^k (g_i + V)$. Suppose that G is a finitely generated subgroup of A containing $g_1, \dots, g_k$. Then $G = \bigcup_{1=1}^k (g_i + V \cap G)$ and so k translates of $V \cap G$ cover G. By the above argument and by Theorem 2.5 there exist $\varphi_{1G}, \dots, \varphi_{mG} \in G^*$, where $m = k^2$, such that

$$U_G\left(\varphi_{1G}, \dots, \varphi_{mG}; \frac{\pi}{2}\right) \subseteq (V \cap G)_{(8)} \subseteq V_{(8)}.$$

By Corollary 2.6 we can extend each φ_{iG} to a character of A, so that we assume from now on $\varphi_{1G}, \dots, \varphi_{mG} \in A^*$ and

$$G \cap U_A\left(\varphi_{1G}, \dots, \varphi_{mG}; \frac{\pi}{2}\right) = U_G\left(\varphi_{1G}, \dots, \varphi_{mG}; \frac{\pi}{2}\right) \subseteq V_{(8)}. \tag{11}$$

Let $\mathcal{G}$ be the family of all finitely generated subgroups G of A containing $g_1, \dots, g_k$. It is a directed set under inclusion. So we get m nets $\{\varphi_{jG}\}_{G\in\mathcal{G}}$ in A^* for $j = 1, \dots, m$. By Lemma 4.2 there exist subnets $\{\varphi_{jG_\beta}\}_\beta$ and $\chi_1, \dots, \chi_m \in A^*$ such that

$$\varphi_j(x) = \lim_\beta \varphi_{jG_\beta}(x) \text{ for every } x \in A \text{ and } j = 1, \dots, m. \tag{12}$$

From (11) and (12) we conclude as before that $U_A(\chi_1, \dots, \chi_m; \frac{\pi}{2}) \subseteq V_{(8)}$. □

As a corollary of Følner's lemma we obtain the following description of the neighborhoods of 0 in the Bohr topology of A.

Corollary 5.8. *For a subset E of an abelian group A the following are equivalent:*

(a) E contains $V_{(8)}$ for some big subset V of A;

(b) for every $n \in \mathbb{N}_+$ E contains $V_{(2n)}$ for some big subset V of A;

(c) E is a neighborhood of 0 in the Bohr topology of A.

Proof. The implication (c) ⇒ (b) follows from Følner's lemma. The implication (c) ⇒ (b) follows from Corollary 4.31 and Proposition 4.30. □

It follows from results of Følner [45] obtained by less elementary tools, that (b) can be replaced by the weaker assumption $V_{(4)} \subseteq E$ (see also Ellis and Keynes [43] or Cotlar and Ricabarra [24] for further improvements). Nevertheless the following old problems concerning the group $\mathbb{Z}$ is still open (see Cotlar and Ricabarra [24], Ellis and Keynes [43], Følner [45], Glasner [54], Pestov [80, Question 1025] or Veech [94]):

Question 5.9. Does there exist a big set $V \subseteq \mathbb{Z}$ such that $V - V$ is not a neighborhood of 0 in the Bohr topology of G?

It is known that every infinite abelian group G admits a big set with empty interior with respect to the Bohr topology [4] (more precisely, these authors prove that every totally bounded group has a big subset with empty interior).

5.3 Prodanov's lemma and proof of Følner's theorem

Let C be a set in a real or complex vector space. Then C is said to be *convex* if, for all $x, y \in C$ and all $t \in [0, 1]$, the point $(1 - t)x + ty \in C$.

The next lemma, due to Prodanov [84], allows us to eliminate the discontinuous characters in uniform approximations of continuous functions via linear combinations of characters.

Lemma 5.10 (Prodanov's lemma). *Let G be a topological abelian group, let U be an open subset of G, f a complex valued continuous function on G and M a convex closed subset of $\mathbb{C}$. Let $k \in \mathbb{N}_+$ and $\chi_1, \dots, \chi_k \in G'$. Suppose that $c_1, \dots, c_k \in \mathbb{C}$ are such that $\sum_{j=1}^k c_j\chi_j(x) - f(x) \in M$ for every $x \in U$. If $\chi_{m_1}, \dots, \chi_{m_s}$, with $m_1 < \cdots < m_s, s \in \mathbb{N}, \{m_1, \dots, m_s\} \subseteq \{1, \dots, k\}$, are the continuous among $\chi_1, \dots, \chi_k$, then $\sum_{i=1}^s c_{m_i}\chi_{m_i}(x) - f(x) \in M$ for every $x \in U$.*

Proof. Let $\chi_k \in G^*$ be discontinuous. Then it is discontinuous at 0. Consequently there exists a net $\{x_\gamma\}_\gamma$ in G such that $\lim_\gamma x_\gamma = 0$ and there exist $y_j = \lim_\gamma \chi_j(x_\gamma)$ for all $j = 1, \dots, k$, but $y_k \neq 1$. Notice that always $|y_j| = 1$. Moreover, $y_j = 1$ when χ_j is continuous because $x_\gamma \to 0$, so $y_j = \lim \chi_j(x_\gamma) = 1$.

Consider $\sum_{j=1}^k c_j\chi_j(x + tx_\gamma) - f(x + tx_\gamma)$, where $t \in \mathbb{Z}$. Since $\lim_\gamma x_\gamma = 0$, we have $x + tx_\gamma \in U$ for every $x \in U$ and for every sufficiently large γ. Thus $\sum_{j=1}^k c_j\chi_j(x)\chi_j(x_\gamma)^t - f(x + tx_\gamma) \in M$ and so passing to the limit $\sum_{j=1}^k c_j\chi_j(x)y_j^t - f(x) \in M$, because f is continuous and M is closed.

Take an arbitrary $n \in \mathbb{N}$. By the convexity of M and the relation above for $t = 0, \dots, n$, we obtain

$$\frac{1}{n+1}\sum_{t=0}^{n}\left(\sum_{j=1}^{k} c_j\chi_j(x)y_j^t - f(x)\right) \in M.$$

Note that $\sum_{t=0}^n y_k^t = \frac{y_k^{n+1}-1}{y_k-1}$ because $y_k \neq 1$. Hence we get

$$\sum_{j=1}^{k-1} c_{jn}\chi_j(x) + \frac{c_k}{1+n}\frac{1-y_k^{n+1}}{1-y_k}\chi_k(x) - f(x) \in M$$

for every $x \in U$, where $c_{jn} = \frac{\sum_{t=0}^n c_j y_j^t}{n+1}$. Now for every $j = 1, 2, \dots, k-1$

- $|c_{jn}| \leq |c_j|\frac{\sum_{t=0}^n |y_j|^t}{n+1} = |c_j|$ (because $|y_j| = 1$), and
- if $y_j = 1$ then $c_{jn} = c_j$.

By the boundedness of the sequences $\{c_{jn}\}_{n=1}^\infty$ for $j = 1, \dots, k-1$, there exists a subsequence $\{n_m\}_{m=1}^\infty$ such that all limits $c'_j = \lim_m c_{jn_m}$ exist for $j = 1, \dots, k-1$. On the other hand, $|y_k| = 1$, so

$$\lim_n \frac{c_k}{n+1}\frac{1-y_k^{n+1}}{1-y_k} = 0.$$

Taking the limit for $m \to \infty$ in

$$\sum_{j=1}^{k-1} c_{jn_m}\chi_j(x) + \frac{c_k}{1+n_m}\frac{1-y_k^{n_m+1}}{1-y_k}\chi_k(x) - f(x) \in M$$

gives

$$\sum_{j=1}^{k-1} c'_j\chi_j(x) - f(x) \in M \quad \text{for } x \in U; \tag{13}$$

moreover $c'_j = c_j$ for every $j = 1, \dots, k-1$ such that χ_j is continuous.

The condition (13) is obtained by the hypothesis, removing the discontinuous character χ_k in such a way that the coefficients of the continuous characters remain the same. Iterating this procedure, we can remove all discontinuous characters among $\chi_1, \dots, \chi_k$. □

Now we give an (apparently) topology-free form of the local version of the Stone-Weierstraß theorem 2.19.

Proposition 5.11. *Let G be an abelian group and H be a group of characters of G. If X is a subset of G and f is a complex valued bounded function on X then the following conditions are equivalent:*

(a) f can be uniformly approximated on X by a linear combination of elements of H with complex coefficients;

(b) for every $\varepsilon > 0$ there exist $\delta > 0$ and $\chi_1, \dots, \chi_m \in H$ such that $x - y \in U_G(\chi_1, \dots, \chi_m; \delta)$ yields $|f(x) - f(y)| < \varepsilon$ for every $x, y \in X$.

Proof. (a)⇒(b) Let $\varepsilon > 0$. By (a) there exist $c_1, \dots, c_m \in \mathbb{C}$ and $\chi_1, \dots, \chi_m \in H$ such that $\| \sum_{i=1}^m c_1\chi_i - f\|_\infty < \frac{\varepsilon}{4}$, that is $|\sum_{i=1}^m c_1\chi_i(x) - f(x)| < \frac{\varepsilon}{4}$ for every $x \in X$.

On the other hand note that $|\sum_{i=1}^m c_i\chi_i(x) - \sum_{i=1}^m c_i\chi_i(y)| \leq \sum_{i=1}^m |c_i| \cdot |\chi_i(x) - \chi_i(y)|$ and that $|\chi_i(x - y) - 1| = |\chi_i(x)\chi_i(y)^{-1} - 1| = |\chi_i(x) - \chi_i(y)|$. If we take

$$\delta = \frac{\varepsilon}{2m \max_{i=1,\dots,m} |c_i|}$$

then $x-y \in U(\chi_1,\dots,\chi_m;\delta)$ implies $\sum_{i=1}^m |c_i|\cdot|\chi_i(x)-\chi_i(y)| < \frac{\varepsilon}{2}$ and so also $|\sum_{i=1}^m c_i\chi_i(x)-\sum_{i=1}^m c_i\chi_i(y)| < \frac{\varepsilon}{2}$. Consequently,

$$|f(x)-f(y)| \le \left|f(x)-\sum_{i=1}^m c_i\chi_i(x)\right| + \left|\sum_{i=1}^m c_i\chi_i(x)-\sum_{i=1}^m c_i\chi_i(y)\right| + \left|\sum_{i=1}^m c_i\chi_i(y)-f(y)\right| < \varepsilon.$$

(b)⇒(a) Let βX be the Čech-Stone compactification of X endowed with the discrete topology. If $F : X \to \mathbb{C}$ is bounded, there exists a unique continuous extension F^β of F to βX. Let $\mathcal{S}$ be the collection of all continuous functions g on βX such that $g = \sum_{j=1}^n c_j\chi_j^\beta$ with $\chi_j \in H$, $c_j \in \mathbb{C}$ and $n \in \mathbb{N}_+$. Then $\mathcal{S}$ is a subalgebra of $\mathcal{C}(\beta X,\mathbb{C})$ closed under conjugation and contains all constants. In fact in $\mathcal{S}$ we have $\chi_k^\beta\chi_j^\beta = (\chi_k\chi_j)^\beta$ by definition and $\overline{\chi^\beta} = (\overline{\chi})^\beta$ because $\chi\overline{\chi} = 1$ and so $(\chi\overline{\chi})^\beta = \chi^\beta(\overline{\chi})^\beta = 1$, that is $(\overline{\chi})^\beta = (\chi^{-1})^\beta = \overline{\chi^\beta}$.

Now we will see that $\mathcal{S}$ separates the points of βX separated by f^β, to apply the local Stone-Weierstraß Theorem 2.19. Let $x, y \in \beta X$ and $f^\beta(x) \neq f^\beta(y)$. Consider two nets $\{x_i\}_i$ and $\{y_i\}_i$ in X such that $x_i \to x$ and $y_i \to y$. Since f^β is continuous, we have $f^\beta(x) = \lim f(x_i)$ and $f^\beta(y) = \lim f(y_i)$. Along with $f^\beta(x) \neq f^\beta(y)$ this implies that there exists $\varepsilon > 0$ such that $|f(x_i)-f(y_i)| \ge \varepsilon$ for every sufficiently large i. By the hypothesis there exist $\delta > 0$ and $\chi_1,\dots,\chi_k \in H$ such that for every $u, v \in X$ if $u - v \in U_G(\chi_1,\dots,\chi_k;\delta)$ then $|f(u)-f(v)| < \varepsilon$. Assume $\chi_j^\beta(x) = \chi_j^\beta(y)$ holds true for every $j = 1,\dots,k$. Then $x_i - y_i \in U_G(\chi_1,\dots,\chi_k;\delta)$ for every sufficiently large i, this contradicts (a). So each pair of points of βX separated by f^β is also separated by $\mathcal{S}$. Since βX is compact, one can apply the local version of the Stone-Weierstraß Theorem 2.19 to $\mathcal{S}$ and f^β and so f^β can be uniformly approximated by $\mathcal{S}$. To conclude note that if $g = \sum c_j\chi_j^\beta$ on βX then $g\restriction_X = \sum c_j\chi_j$. □

The reader familiar with uniform spaces will note that item (b) is nothing else but uniform continuity of f w.r.t. the uniformity on X induced by the uniformity of the whole group G determined by the topology $\mathcal{T}_H$.

Theorem 5.12 (Følner theorem). *Let G be a topological abelian group. If k is a positive integer and E is a subset of G such that k translates of E cover G, then for every neighborhood U of 0 in G there exist $\chi_1,\dots,\chi_m \in \widehat{G}$, where $m = k^2$, and $\delta > 0$ such that $U_G(\chi_1,\dots,\chi_m;\delta) \subseteq U - U + E_{(8)}$.*

Proof. By Følnerś lemma 5.7 there exist $\varphi_1,\dots,\varphi_m \in G^*$ such that $U_G(\varphi_1,\dots,\varphi_m;\frac{\pi}{2}) \subseteq E_{(8)}$, where the characters φ_j can be discontinuous. Our aim will be to replace these characters by continuous ones "enlarging" $E_{(8)}$ to $U - U + E_{(8)}$.

It follows from Lemma 3.18 that $C := \overline{E_{(8)} + U} \subseteq E_{(8)} + U - U$. Consider the open set $X = U \cup (G \setminus C)$ and the function $f : X \to \mathbb{C}$ defined by

$$f(x) = \begin{cases} 0 & \text{if } x \in U \\ 1 & \text{if } x \in G \setminus C \end{cases}$$

Then f is continuous as $X = U \cup (G \setminus C)$ is a clopen partition of X.

Let H be the group generated by $\varphi_1,\dots,\varphi_m$. Take $x, y \in X$ with $x - y \in U_G(\varphi_1,\dots,\varphi_m;\frac{\pi}{2}) \subseteq E_{(8)}$. So if $y \in U$ then $x \in E_{(8)} + U$ and consequently $x \notin G \setminus \overline{E_{(8)} + U}$, that is $x \in U$. In the same way it can be showed that $x \in U$ yields $y \in U$. This gives $f(x) = f(y)$ by the definition of f. So by Proposition 5.11 one can uniformly approximate f on X by characters of H. Hence one can find a finite number of m-uples $\tilde{j} = (j_1,\dots,j_m)$ of integers and $c_{\tilde{j}} \in \mathbb{C}$ such that

$$\left|\sum_{\tilde{j}} c_{\tilde{j}}\varphi_1^{j_1}(x)\cdot\ldots\cdot\varphi_m^{j_m}(x) - f(x)\right| \le \frac{1}{3} \tag{13}$$

holds for every $x \in X$. Since X is open and f is continuous, we can apply Lemma 5.10 to the convex closed set $M = \{z \in \mathbb{C} : |z| \le \frac{1}{3}\}$ and this permits us to assume that all products $\varphi_1^{j_1}\cdot\ldots\cdot\varphi_m^{j_m}$ are continuous. Letting $x = 0$ in (13) one gets $|\sum_{\tilde{j}} c_{\tilde{j}}| \le \frac{1}{3}$, and consequently,

$$\frac{2}{3} \le \left|\sum_{\tilde{j}} c_{\tilde{j}} - 1\right|. \tag{14}$$

Let now Φ be the subgroup of H consisting of all *continuous* characters of H, i.e., $\Phi = H \cap \widehat{G}$. By Theorem 2.1 there exist $\chi_1,\dots,\chi_m \in \Phi$ that generate Φ. Choose $\varepsilon > 0$ with $\varepsilon\sum_{\tilde{j}} |c_{\tilde{j}}| < \frac{1}{3}$. By the continuity of $\chi_1,\dots,\chi_m \in \Phi$ there exists $\delta > 0$ such that $x \in U_G(\chi_1,\dots,\chi_m;\delta)$ implies $|\varphi_1^{j_1}(x)\cdot\ldots\cdot\varphi_m^{j_m}(x) - 1| \le \varepsilon$ for all summands $\varphi_1^{j_1}\cdot\ldots\cdot\varphi_m^{j_m}$ in (13).

Unit 6

6 Peter-Weyl's theorem and other applications of Følner's theorem

In this section we prove Peter-Weyl's theorem using Følner's theorem.

6.1 Precompact group topologies on abelian groups

Let us recall here that for an abelian group G and a subgroup H of G^*, the group topology $\mathcal{T}_H$ generated by H is the coarsest group topology on G that makes every character from H continuous. We recall its description and properties in the next proposition:

Proposition 6.1. *Let G be an abelian group and let H be a group of characters of G. A base of the neighborhoods of 0 in $(G, \mathcal{T}_H)$ is given by the sets $U(\chi_1, \dots, \chi_m; \delta)$, where $\chi_1, \dots, \chi_m \in H$ and $\delta > 0$. Moreover $(G, \mathcal{T}_H)$ is a Hausdorff if and only if H separates the points of G.*

Now we can characterize the precompact topologies on abelian groups.

Theorem 6.2. *Let (G, τ) be an abelian group. The following conditions are equivalent:*

(a) τ is precompact;

(b) τ is Hausdorff on G and the neighborhoods of 0 in G are big subsets;

(c) there exists a group H of continuous characters of G that separates the points of G and such that $\tau = \mathcal{T}_H$.

Proof. (a)⇒(b) is the definition of precompact topology.

(b)⇒(c) If $H = \widehat{(G, \tau)}$ then $\mathcal{T}_H \subseteq \tau$. Let U and V be open neighborhoods of 0 in (G, τ) such that $V_{(10)} \subseteq U$. Then V is big and by Følner's Theorem 5.12 there exist continuous characters $\chi_1, \dots, \chi_m$ of G such that $U_G(\chi_1, \dots, \chi_m; \delta) \subseteq V_{(10)} \subseteq U$ for some $\delta > 0$. Thus $U \in \mathcal{T}_H$ and $\tau \subseteq \mathcal{T}_H$.

(c)⇒(a) Even if this implication is contained in Corollary 4.31, we give a direct proof here. Let $i : G \to \mathbb{S}^H$ be defined by $i(g) = i_g : H \to \mathbb{S}$ (if $g \in G$) with $i_g(\chi) = \chi(g)$ for every $\chi \in H$. Since H separates the points of G, the function i is injective. The product $\mathbb{S}^H$ endowed with the product topology is compact and so i is a topological immersion by Proposition 6.1. The closure of $i(G)$ in $\mathbb{S}^H$ is compact and $\widetilde{G}$ is isomorphic to it, hence $\widetilde{G}$ is compact. □

Remark 6.3. The above theorem essentially belongs to Comfort and Ross [23]. It can be given in the following simpler "Hausdorff-free" version: τ is totally bounded iff $\tau = \mathcal{T}_H$ for some group H of continuous characters of G.

Corollary 6.4 (Peter-Weyl's theorem). *If G is a compact abelian group, then $\widehat{G}$ separates the points of G.*

Proof. Let τ be the topology of G. By Theorem 6.2 there exists a group H of continuous characters of G (i.e., $H \subseteq \widehat{G}$) such that $\tau = \mathcal{T}_H$. Since $\tau \supseteq \mathcal{T}_{\widehat{G}}$ and $H \subseteq \widehat{G}$ we conclude that $H = \widehat{G}$ separates the points of G. □

The next theorem will allow us to sharpen this property (see Corollary 6.6).

Theorem 6.5. *Let G be an abelian group. Let $\mathcal{H}$ be the set of all groups of characters of G separating the points of G and $\mathcal{P}$ be the set of all precompact group topologies on G. Then the function $T : \mathcal{H} \to \mathcal{P}$ which associates to $H \in \mathcal{H}$ the topology $\mathcal{T}_H \in \mathcal{P}$ is an order preserving bijection (if $H_1, H_2 \in \mathcal{H}$ then $\mathcal{T}_{H_1} \subseteq \mathcal{T}_{H_2}$ if and only if $H_1 \subseteq H_2$).*

Proof. The equivalence (a)⇔(c) of Theorem 6.2 yields that $\mathcal{T}_H \in \mathcal{P}$ for every $H \in \mathcal{H}$ and that T is surjective.

Let $H \in \mathcal{H}$ and suppose that $\chi \in (\widehat{G, \mathcal{T}_H})$. To show that $\chi \in H$ let $\varepsilon > 0$. Since χ is continuous in 0, by Proposition 6.1 there exist $\chi_1, \dots, \chi_m \in H$ and $\delta > 0$ such that $|\chi(x) - 1| < \varepsilon$ for $x \in U(\chi_1, \dots, \chi_m; \delta)$. Therefore for every $x, y \in G$ with $x - y \in U(\chi_1, \dots, \chi_m; \delta)$ we get $|\chi(x - y) - 1| < \varepsilon$, that is $|\chi(x) - \chi(y)| < \varepsilon$. Apply now Proposition 5.11 to find $\chi_1, \dots, \chi_m \in H$ and $c_i, \dots, c_m \in \mathbb{C}$ such that $\left| \sum_{j=1}^m c_j \chi_j(x) - \chi(x) \right| \leq \frac{1}{2}$ for every $x \in G$. This yields $\left| \sum_{j=1}^m c_j \chi_j(x) \chi^{-1}(x) - 1 \right| \leq \frac{1}{2}$.

Suppose now that $\chi \notin H$. Then each $\chi_j \chi^{-1}$ in the previous condition is non-constant. Equip G with the indiscrete topology. Then each character $\chi_j \chi^{-1}$ is discontinuous. Applying Lemma 5.10 we get the inequality $1 \leq \frac{1}{2}$, which is a contradiction. Therefore $\chi \in H$ and so $H = (\widehat{G, \tau_H})$ for every $H \in \mathcal{H}$.

If $H_1, H_2 \in \mathcal{H}$ and $\mathcal{T}_{H_1} = \mathcal{T}_{H_2}$ then $H_1 = H_2$, so T is a bijection.

The last statement of the theorem is obvious. □

As a corollary of Theorem 6.5 we obtain the following important fact that completes Corollary 6.4. It will be essentially used in the proof of the duality theorem.

Corollary 6.6. *If (G, τ) is a compact abelian group and $H \leq \widehat{G}$ separates the points of G, then $H = \widehat{G}$.*

Proof. By Theorem 6.2 it holds $\tau = \mathcal{T}_{\widehat{G}}$. Since $\mathcal{T}_H \subseteq \mathcal{T}_{\widehat{G}}$ by Theorem 6.5 and $\mathcal{T}_H$ is Hausdorff, then $\mathcal{T}_H = \mathcal{T}_{\widehat{G}}$. Now again Theorem 6.5 yields $H = \widehat{G}$. □

Definition 6.7. An abelian topological group is *elementary compact* if it is topologically isomorphic to $\mathbb{T}^s \times F$, where n is a positive integer and F is a finite abelian group.

Proposition 6.8. *Let G be a compact abelian group and let U be an open neighborhood of 0 in G. Then there exists a closed subgroup C of G such that $C \subseteq U$ and G/C is an elementary compact abelian group. In particular, G is an inverse limit of elementary compact abelian groups.*

Proof. By the Peter-Weyl Theorem 6.4 $\bigcap_{\chi \in \widehat{G}} \ker \chi = \{0\}$ and each $\ker \chi$ is a closed subgroup of G. By the compactness of G there exists a finite subset F of $\widehat{G}$ such that $C = \bigcap_{\chi \in F} \ker \chi \subseteq U$. Define now $g = \prod_{\chi \in F} \chi : G \to \mathbb{T}^F$. Thus $\ker g = C$ and G/C is topologically isomorphic to the closed subgroup $g(G)$ of $\mathbb{T}^F$ by the compactness of G. So G/C is elementary compact abelian by Lemma 4.47.

To prove the last statement, fix for every open neighborhood U_i of 0 in G a closed subgroup C_i of G with $C_i \subseteq U$ and such that G/C_i is elementary compact abelian. Note that for C_i and C_j obtained in this way the subgroup $C_i \cap C_j$ has the same property as $G/C_i \cap C_j$ is isomorphic to a closed subgroup of the product $G/C_i \times G/C_j$ which is again an elementary compact abelian group. Enlarging the family (C_i) with all finite intersections we obtain an inverse system of elementary compact abelian groups G/C_i where the connecting homomorphisms $G/C_i \to G/C_j$, when $C_i \leq C_j$, are simply the induced homomorphisms. Then the inverse limit G' of this inverse system is a compact abelian group together with a continuous homomorphism $f : G \to G'$ induced by the projections $p_i : G \to G/C_i$. Assume $x \in G$ is non-zero. Pick on open neighborhood U of 0. By the first part of the proof, there exists $C_i \subseteq U$, hence $x \notin C_i$. Therefore, $p_i(x) \neq 0$, so $f(x) \neq 0$ as well. This proves that f is injective. To check surjectivity of f take an element $x' = (x_i + C_i)$ of the inverse limit G'. Then the family of closed cosets $x_i + C_i$ in G has the finite intersection property, so has a non-empty intersection. For every element x of that intersection one has $f(x) = x'$. Finally, the continuous isomorphism $f : G \to G'$ must be open by the compactness of G. □

For a topological abelian group G we say that G has *no small subgroups*, or shortly, G is *NSS*, if there exists a neighborhood U of 0 such that U contain no non-trivial subgroups of G. It follows immediately from the above proposition that the compact abelian group G has no small subgroups precisely when G is an elementary compact abelian group.

6.2 Precompact group topologies determined by sequences

Large and lacunary sets (mainly in $\mathbb{Z}$ or elsewhere) are largely studied in number theory, harmonic analysis and dynamical systems ([43], [24], [80], [52], [53], [54], [55], [59]).

Let us consider a specific problem. For a strictly increasing sequence $\underline{u} = (u_n)_{n \geq 1}$ of integers, the interest in the distribution of the multiples $\{u_n \alpha : n \in \mathbb{N}\}$ of a non-torsion element α of the group $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ has roots in number theory (Weyl's theorem of uniform distribution modulo 1) and in ergodic theory (Sturmian sequences and Hartman sets [99]). According to Weyl's theorem, the set $\{u_n \alpha : n \in \mathbb{N}\}$ will be uniformly dense in $\mathbb{T}$ for almost all $\alpha \in \mathbb{T}$. One can consider the subset $t_{\underline{u}}(\mathbb{T})$ of all elements $\alpha \in \mathbb{T}$ such that $\lim_n u_n \alpha = 0$ in $\mathbb{T}$. Clearly

it will have measure zero. Moreover, it is a subgroup of $\mathbb{T}$ as well as a Borel set, so it is either countable or has size $\mathfrak{c}$. It was observed by Armacost [3] that when $u_n = p^n$ for all n and some prime p, then $t_{\underline{u}}(\mathbb{T}) = \mathbb{Z}(p^\infty)$. He posed the question of describing the subgroup $t_{\underline{u}}(\mathbb{T})$ for the sequence $u_n = n!$, this was done by Borel [19] (see also [36] and [31] for the more general problem concerning sequences $\underline{u}$ with $u_{n-1}|u_n$ for every n).

Another motivation for the study of the subgroups of the form $t_{\underline{u}}(\mathbb{T})$ come from the fact that they lead to the description of precompact group topologies on $\mathbb{Z}$ that make the sequence u_n converge to 0 in $\mathbb{Z}$ (see the comment after proposition 6.9). Let us start by an easy to prove general fact:

Proposition 6.9. [7] *A sequence $A = \{a_n\}_n$ in a precompact abelian group G converges to 0 in G iff $\chi(a_n) \to 0$ in $\mathbb{T}$ for every continuous character of G.*

In the case of $G = \mathbb{Z}$ the characters of G are simply simply elements of $\mathbb{T}$, i.e., a precompact group topology on $\mathbb{Z}$ has the form $\mathcal{T}_H$ for some subgroup H of $\mathbb{T}$. Thus the above proprosition for $G = \mathbb{Z}$ can be reformulated as: a sequence $A = \{a_n\}_n$ in $(\mathbb{Z}, \mathcal{T}_H)$ converges to 0 iff $a_n x \to 0$ for every $x \in H$, i.e., simply $H \subseteq t_{\underline{a}}(\mathbb{T})$.

Now we can discuss a counterpart of the notion of T-sequences (introduced in §3.5), defined with respect to topologies induced by characters, i.e., precompact topologies.

Definition 6.10. [7, 9] A sequence $A = \{a_n\}_n$ in an abelian group G is called a *TB-sequence* is there exists a precompct group topology on G such that $a_n \to 0$.

Clearly, every TB-sequence is a T-sequence (see Example 6.12 for a T-sequence in $\mathbb{Z}$ that is not a TB-sequence). The advantage of TB-sequences over the T-sequences is in the easier way of determining sufficient condition for a sequence to be a TB-sequence [7, 9]. For example, a a sequence (a_n) in $\mathbb{Z}$ is a TB-sequence iff the subgroup $t_{\underline{a}}(\mathbb{T})$ of $\mathbb{T}$ is infinite.

Egglestone [42] proved that the asymptotic behavior of the sequence of ratios $q_n = \frac{u_{n+1}}{u_n}$ may have an impact on the size of the subgroup $t_{\underline{u}}(\mathbb{T})$ in the following remarkable dichotomy:

Theorem 6.11. *Let (a_n) be a sequence in $\mathbb{Z}$.*

- *If $\lim_n \frac{a_{n+1}}{a_n} = +\infty$, then (a_n) is a TB-sequence and $|t_{\underline{a}}(\mathbb{T})| = \mathfrak{c}$.*
- *If $\frac{a_{n+1}}{a_n}$ is bounded, then $t_{\underline{a}}(\mathbb{T})$ is countable.*

Example 6.12. *[9] There exists a TB-sequence (a_n) in $\mathbb{Z}$ with $\lim_n \frac{a_{n+1}}{a_n} = 1$.*

Here is an example of a T-sequence in $\mathbb{Z}$ that is not a TB-sequence.

Example 6.13. For every TB-sequence $A = \{a_n\}$ in $\mathbb{Z}$ such that $t_{\underline{a}}(\mathbb{T})$ is countable, there exists a sequence $\{c_n\}$ in $\mathbb{Z}$ such that the sequence q_n defined by $q_{2n} = c_n$ and $q_{2n-1} = a_n$, is a T-sequence, but not a TB-sequence.

Proof. Let $\{z_1, \dots, z_n, \dots\}$ be an enumeration of $t_{\underline{a}}(\mathbb{T})$.

According to Lemma 3.51 there exists a sequence b_n in $\mathbb{Z}$ such that for every choice of the sequence (e_n), where $e_n \in \{0, 1\}$, the sequence q_n defined by $q_{2n} = b_n + e_n$ and $q_{2n-1} = a_n$, is a T-sequence. Now we define the sequence q_n with $q_{2m-1} = a_m$ and $q_{2m} = b_m$ when m is not a prime power. Let $p_1, \dots, p_n, \dots$ be all prime numbers enumerated one-to-one. Now fix k and define $e_k \in \{0, 1\}$ depending on $\lim_n b_{p_k^n} z_k$ as follows:

- if $\lim_n b_{p_k^n} z_k = 0$, let $e_k = 1$,
- if $\lim_n b_{p_k^n} z_k \neq 0$ (in particular, if the limit does not exists) let $e_k = 0$.

Now let $q_{2p_k^n} = b_{p_k^n} + e_k$ for $n \in \mathbb{N}$. Hence for every $k \in \mathbb{N}$

$$\lim_n q_{2p_k^n} z_k = 0 \implies e_k = 1. \qquad (*)$$

To see that (q_n) is not a TB-sequence assume that $\chi : \mathbb{Z} \to \mathbb{T}$ a character such that $\chi(q_n) \to 0$ in $\mathbb{T}$. Then $x = \chi(1) \in \mathbb{T}$ satisfies $q_n x \to 0$, so $x \in t_{\underline{q}}(\mathbb{T}) \subseteq t_{\underline{a}}(\mathbb{T})$. So there exists $k \in \mathbb{N}$ with $x = z_k$. By (*) $e_k = 1$. Hence $q_{2p_k^n} = b_{p_k^n} + 1$ and $\lim_n b_{p_k^n} z_k = 0$, so $x \in t_{\underline{q}}(\mathbb{T})$ yields $0 = \lim_n q_{2p_k^n} x = 0 + x$, i.e., $x = 0$. This proves that every character $\chi : \mathbb{Z} \to \mathbb{T}$ such that $\chi(q_n) \to 0$ in $\mathbb{T}$ is trivial. In particular, (q_n) not a TB-sequence. □

Let us note that the above proof gives much more. Since $q_n \to 0$ in $\tau_{(q_n)}$, it shows that every $\tau_{(q_n)}$-continuous character of $\mathbb{Z}$ is trivial, i.e., $\widehat{(\mathbb{Z}, \tau_{(q_n)}} = 0$.

The information accumulated on the properties of the subgroups $t_{\underline{u}}(\mathbb{T})$ of $\mathbb{T}$ motivated the problem of describing those subgroups H of $\mathbb{T}$ that can be characterized as $H = t_{\underline{u}}(\mathbb{T})$ for some sequence $\underline{u}$. As already

mentioned, such an H can be only countable or can have size $\mathfrak{c}$ being of measure zero. A measure zero subgroup H of $\mathbb{T}$ of size $\mathfrak{c}$ that is not even contained in any proper subgroup of $\mathbb{T}$ of the form $t_{\underline{u}}(\mathbb{T})$ was built in [7] (under the assumption of Martin Axiom) and in later in [61, 62] (in ZFC). Much earlier Borel [19] had already resolved in the positive the remaining part of the problem showing that every countable subgroup of $\mathbb{T}$ can be characterized (in the above sense). Unaware of his result, Larcher [74], and later Kraaikamp and Liardet [71], proved that some cyclic subgroups of $\mathbb{T}$ are characterizable (see also [16, 15, 12, 14, 13] for related results). The paper [9] describes the algebraic structure of the subgroup $t_{\underline{u}}(\mathbb{T})$ when the sequence $\underline{u} := (u_n)$ verifies a linear recurrence relation of order $\leq k$,

$$u_n = a_n^{(1)} u_{n-1} + a_n^{(2)} u_{n-2} + \ldots + a_n^{(k)} u_{n-k}$$

for every $n > k$ with $a_n^{(i)} \in \mathbb{Z}$ for $i = 1, \ldots, k$.

Three proofs of Borel's theorem of characterizability of the countable subgroups of $\mathbb{T}$ were given in [13]. These author mentioned that the theorem can be extended to compact abelian groups in place of $\mathbb{T}$, without giving any precise formulation. There is a natural way to extend the definition of $t_{\underline{u}}(\mathbb{T})$ to an arbitrary topological abelian group G by letting $t_{\underline{u}}(G) = \{x \in G : \lim_n u_n x = 0 \text{ in } G\}$. Actually, for the sequence $u_n = p^n$ (resp., $u_n = n!$) an element x satisfying $\lim_n u_n x = 0$ has been called *topologically p-torsion* (resp., *topologically torsion*) by Braconnier and Vilenkin in the forties of the last century and these notions played a prominent role in the development of the theory of locally compact abelian groups. One can easily reduce the computation of $t_{\underline{u}}(G)$ for an arbitrary locally compact abelian group to that of $t_{\underline{u}}(\mathbb{T})$ [26]. Independently on their relevance in other questions, the subgroups $t_{\underline{u}}(G)$ turned out to be of no help in the characterization of countable subgroups of the compact abelian groups. Indeed, a much weaker condition, turned out the characterize the circle group $\mathbb{T}$ in the class of all locally compact abelian groups:

Theorem 6.14. [31] *In a locally compact abelian group G every cyclic subgroup of the group G is an intersection of subgroups of the form $t_{\underline{u}}(G)$ iff $G \cong \mathbb{T}$.*

Actually, one can remove the "abelian" restraint in the theorem remembering that in the non-abelian case $t_{\underline{u}}(G)$ is just a subset of G, not a subgroup in general [31].

The above theorem suggested to use in [35] a different approach to the problem, replacing the sequence of integers u_n (characters of $\mathbb{T}$!) by a sequence u_n in the Pontryagin-van Kampen dual $\widehat{G}$. Then the subgroup $s_{\underline{u}}(G) = \{x \in G : \lim_n u_n(x) = 0 \text{ in } \mathbb{T}\}$ of G really can be used for such a characterization of all countable subgroups of the compact metrizable groups (see [35, 33, 17] for major detail).

6.3 On the structure of compactly generated locally compact abelian groups

From now on all groups are Hausdorff; quotients are taken for closed subgroups and so they are still Hausdorff.

An abelian topological group is *elementary locally compact* if it is topologically isomorphic to $\mathbb{R}^n \times \mathbb{Z}^m \times \mathbb{T}^s \times F$, where n, m, s are positive integers and F is a finite abelian group. Observe that the class of elementary locally compact abelian groups is closed under taking quotient, closed subgroups and finite products (see Theorem 4.37 and Corollary 4.47).

Lemma 6.15. *Let G be a locally compact monothetic group. Then G is either compact or is discrete.*

Proof. If G is finite, then G is both compact and discrete. So we can suppose without loss of generality that $\langle x \rangle \cong \mathbb{Z}$ is infinite and so also that $\mathbb{Z}$ is a subgroup of G.

If G induces the discrete topology on $\mathbb{Z}$, then $\mathbb{Z}$ is closed and so $G = \mathbb{Z}$ is discrete.

Suppose now that G induces on $\mathbb{Z}$ a non-discrete topology. Our aim is to show that it is totally bounded. Then the density of $\mathbb{Z}$ in G yields that $G = \widetilde{\mathbb{Z}} = \overline{\mathbb{Z}}$ is compact, as G is locally compact and so complete (see Lemma 4.7).

Every open subset of G has no maximal element. Indeed, if U is an open subset of $\mathbb{Z}$ that contains 0 and it has a maximal element, then $-U$ is an open subset of $\mathbb{Z}$ that contains 0 and it has a minimal element and $U \cap -U$ is an open finite neighborhood of 0 in $\mathbb{Z}$; thus $\mathbb{Z}$ is discrete against the assumption. Consequently every open subset of $\mathbb{Z}$ contains positive elements.

Let U be a compact neighborhood of 0 in G and V a symmetric neighborhood of 0 in G such that $V + V \subseteq U$. There exist $g_1, \ldots, g_m \in G$ such that $U \subseteq \bigcup_{i=1}^m (g_i + V)$. Let $n_1, \ldots, n_m \in \mathbb{Z}$ be positive integers such that $n_i \in g_i + V$ for every $i = 1, \ldots, m$. Equivalently $g_i \in n_i - V = n_i + V$. Thus

$$U \subseteq \bigcup_{i=1}^m (g_i + V) \subseteq \bigcup_{i=1}^m (n_i + V + V) \subseteq \bigcup_{i=1}^m (n_i + U)$$

implies

$$U \cap \mathbb{Z} \subseteq \bigcup_{i=1}^{m} (n_i + U \cap \mathbb{Z}). \tag{1}$$

We show that $U \cap \mathbb{Z}$ is big with respect to $\mathbb{Z}$. Let $t \in \mathbb{Z}$; since $U \cap \mathbb{Z}$ has no maximal element, then there exists $s \in U \cap \mathbb{Z}$ such that $s \geq t$. Define $s_t = \min\{s \in U \cap \mathbb{Z} : s \geq t\}$. By (1) $s_t = n_i + u_t$ for some $i \leq m$ and $u_t \in U \cap \mathbb{Z}$. Since $n_i > 0$, then $u_t < s_t$ and so $u_t < t \leq s_t$. Now put $N = \max\{n_1, \ldots, n_m\}$ and $F = \{1, \ldots, N\}$. Hence $U \cap \mathbb{Z} + F = \mathbb{Z}$. This proves that the topology induced on $\mathbb{Z}$ by G is totally bounded. □

Corollary 6.16. *Let G be a locally compact abelian group and $x \in G$. Then $\overline{\langle x \rangle}$ is either compact or discrete.*

Proposition 6.17. *Let G be a compactly generated locally compact abelian group. Then there exists a discrete subgroup H of G such that $H \cong \mathbb{Z}^n$ for some $n \in \mathbb{N}$ and G/H is compact.*

Proof. Suppose first that there exist $g_1, \ldots, g_m \in G$ such that $G = \overline{\langle g_1, \ldots, g_m \rangle}$. We proceed by induction. For $m = 1$ apply Lemma 6.15: if G is infinite and discrete take $H = G$ and if G is compact $H = \{0\}$. Suppose now that the property holds for $m \geq 1$ and $G = \overline{\langle g_1, \ldots, g_{m+1} \rangle}$. If every $\overline{\langle g_i \rangle}$ is compact, then so is G and $H = \{0\}$. If $\langle g_{m+1} \rangle$ is discrete, consider the canonical projection $\pi : G \to G_1 = G/\langle g_{m+1} \rangle$. Since G_1 has a dense subgroup generated by m elements, by the inductive hypothesis there exists a discrete subgroup H_1 of G_1 such that $H_1 \cong \mathbb{Z}^n$ and G_1/H_1 is compact. Therefore $H = \pi^{-1}(H_1)$ is a closed countable subgroup of G. Thus H is locally compact and countable, hence discrete by Lemma 4.8.

Since H is finitely generated, it is isomorphic to $H_2 \times F$, where $H_2 \cong \mathbb{Z}^s$ for some $s \in \mathbb{N}$ and F is a finite abelian group (see Theorem 2.1). Now G/H is isomorphic to G_1/H_1 and H/H_2 is finite, so G/H_2 is compact thanks to Lemma 4.5.

Now consider the general case. There exists a compact subset K of G that generates G. By Lemma 4.14 we can assume wlog that $K = \overline{U}$, where U is a symmetric neighborhood of 0 in G with compact closure. We show now that there exists a finite subset F of G such that

$$K + K \subseteq K + \langle F \rangle. \tag{2}$$

In fact, pick a symmetric neighborhood V of 0 in G such that $V + V \subseteq U$. For the compact set K satisfying $K \subseteq \bigcup_{x \in K} (x + V)$ there exists a finite subset F of K such that $K \subseteq \bigcup_{x \in F} (x + V) = F + V$. Then

$$K + K \subseteq F + F + V + V \subseteq \langle F \rangle + U \subseteq \langle F \rangle + K.$$

gives (2). An easy inductive argument shows that $\langle K \rangle = G$ and (2) imply $G = \langle K \rangle \subseteq K + \langle F \rangle$.

Let $G_1 = \langle F \rangle$. By $G = \langle F \rangle + K$ the quotient $\pi(K) = G/G_1$ is compact. By the first part of the proof there exists a discrete subgroup H of the locally compact subgroup G_1 of G, such that $H \cong \mathbb{Z}^n$ for some $n \in \mathbb{N}$ and G_1/H is compact. Since G_1/H is a compact subgroup of G/H such that $(G/H)/(G_1/H) \cong G/G_1$ is compact, we conclude that also G/H is compact. □

Proposition 6.18. *Let G be a compactly generated locally compact abelian group. Then there exists a compact subgroup K of G such that G/K is elementary locally compact abelian.*

Proof. By Proposition 6.17 there exists a discrete subgroup H of G such that the quotient G/H is compact. Consider the canonical projection π of G onto G/H. Let U be a compact symmetric neighborhood of 0 in G such that $(U + U + U) \cap H = \{0\}$. So $\pi(U)$ is a neighborhood of 0 in G/H and applying Lemma 6.8 we find a closed subgroup $L \supseteq H$ of G such that the closed subgroup L/H of G/H satisfies

$$L/H \subseteq \pi(U) \text{ and } (G/H)/(L/H) = G/L \cong \mathbb{T}^t \times F, \tag{4}$$

where F is a finite abelian group and $t \in \mathbb{N}$, i.e., G/L is elementary compact abelian.

The set $K = L \cap U$ is compact being closed in the compact neighborhood U. Let us see now that K is a subgroup of G. To this end take $x, y \in K$. Then $x - y \in L$ and $\pi(x - y) \in C \subseteq \pi(U)$. Thus $\pi(x - y) = \pi(u)$ for some $u \in U$. As $\pi(x - y - u) = 0$ in G/H, one has $x - y - u \in (U + U + U) \cap H = \{0\}$. Hence $x - y = u \in L \cap U = K$.

Now take $x \in L$; consequently $\pi(x) \in C \subseteq \pi(U)$ so $\pi(x) = \pi(u)$ for some $u \in U$. Clearly, $u \in L \cap U = K$, hence $\pi(L) = \pi(K)$. Thus $L = K + H$ and $K \cap H = \{0\}$ yields that the canonical projection $l : G \to G/K$ restricted to H is a continuous isomorphism of H onto $l(H) = l(L)$. Let us see now that $l(H)$ is discrete. The compact set K is contained in the open set $W_1 = G \setminus (H \setminus \{0\}) = G \setminus H \cup \{0\}$ (H is discrete). By Lemma 4.3 (c) there exists an open neighborhood V of 0 in G such that $K + V \subseteq W_1$. This implies that $(K + V) \cap H = \{0\}$ and so $(K + V) \cap (K + H) = K$, that gives $l(V) \cap l(H) = \{0\}$ in G/K. Thus

$$l(L) = l(H) \cong H \cong \mathbb{Z}^s$$

is discrete in G/K.

Observe that (4) yields the following isomorphisms:

$$(G/K)/l(L) = (G/K)/(L/K) \cong G/L \cong \mathbb{T}^t \times F.$$

Denote by ϱ the composition $G/K \to G/L \to \mathbb{T}^t \times F$. Let W be a compact neighborhood of 0 in G/K such that $W + W \subseteq l(V)$ and $\varrho(W) \subseteq \mathbb{T}^t \times \{0\}$. Then $\varrho \restriction_W$ is injective because $l(V) \cap l(L) = \{0\}$. In particular, ϱ is a local homeomorphism.

Consider now the canonical projection $q : \mathbb{R}^t \to \mathbb{T}^t$. Our aim is to lift it to a continuous homomorphism $f : \mathbb{R}^t \to G/K$ such that $\varrho \circ f = q$. The existence of such a lifting is immediate from the facts that both q and ϱ are covering homomorphisms and $\mathbb{R}^t$ simply connected. In particular, $D = f(\mathbb{R}^t)$ is an open subgroup of G/K as has a non-empty interior (as q and ϱ are local homeomorphisms). Since $\mathbb{R}^t$ is divisible, by Lemma 2.8 $G/K = D \times B$ where B is a discrete subgroup of G/K because $D \cap B = \{0\}$ and D is open. Moreover B is compactly generated as it is a quotient of G. Since it is also discrete, B is finitely generated. Then $f : \mathbb{R}^t \to D$ is open by Theorem 4.9 and so D is isomorphic to a quotient of $\mathbb{R}^t$, which is elementary locally compact abelian.

For the reader who is not familiar with covering maps we provide now a self-contained proof.

Take an open neighborhood U of 0 in $\mathbb{R}^t$ such that $U - U \cap \ker q = \{0\}$ and let $U_0 = U \cap q^{-1}(\varrho(W))$. Then U_0 is an open neighborhood of 0 in $\mathbb{R}^t$ such that $q \restriction_{U_0}$ is one-to-one from U_0 to $\varrho(W)$. Pick a symmetric neighborhood U_1 of 0 in $\mathbb{R}^t$ such that $U_1 + U_1 \subseteq U_0$. Define a map $f : \mathbb{R}^t \to G/K$ as follows: $f \restriction_{U_0}$ is simply the composition $\varrho^{-1} \circ q$. So f maps U_0 onto the open subset $\varrho^{-1}(q(U_0))$ of G/K. If $x \in \mathbb{R}^t$ there exists $n \in \mathbb{N}_+$ such that $\frac{1}{n}x \in U_0$. We put $f(x) = nf(\frac{1}{n}x)$ and we note that this definition does not depend on n. Moreover, $f(x_1 + x_2) = f(x_1) + f(x_2)$ for every $x_1, x_2 \in U_1$.

We can prove now that f is a homomorphism. First of all we note that for every $x \in \mathbb{R}^t$ $f \restriction_{\langle x \rangle}$ is a homomorphism, i.e., $f(kx) = kf(x)$ for every $k \in \mathbb{Z}$. Now take $x, y \in \mathbb{R}^t$. There exists an integer $n > 0$ such that $\frac{1}{n}x, \frac{1}{n}y \in U_1$ and so $\frac{1}{n}x + \frac{1}{n}y \in U_0$. By the the previous step

$$f(x+y) = nf\left(\frac{1}{n}(x+y)\right) = nf\left(\frac{1}{n}x + \frac{1}{n}y\right) = nf\left(\frac{1}{n}x\right) + nf\left(\frac{1}{n}y\right) = f(x) + f(y),$$

for all $x, y \in \mathbb{R}^t$.

So f is continuous and also a local homeomorphism on $\mathbb{R}^t$ because it is the composition of local homeomorphisms: restricted to the open subset U_0, f is the composition of q and ϱ^{-1} (note that both $\varrho \restriction_W$ and $q \restriction_{U_0}$ are continuous and open). □

To prove the Pontryagin-van Kampen duality theorem in the general case (for $G \in \mathcal{L}$), we need Theorem 6.19, which generalizes the Peter-Weyl Theorem 6.4.

Theorem 6.19. *If G is a locally compact abelian group, then $\widehat{G}$ separates the points of G.*

Proof. Let V be a compact neighborhood of 0 in G. Take $x \in G \setminus \{0\}$. Then $G_1 = \langle V \cup \{x\} \rangle$ is an open (it has non-void interior) compactly generated subgroup of G. In particular G_1 is locally compact. By Proposition 6.17 there exists a discrete subgroup H of G_1 such that $H \cong \mathbb{Z}^m$ for some $m \in \mathbb{N}$ and G_1/H is compact. Thus $\bigcap_{n \in \mathbb{N}_+} nH = \{0\}$ and so there exists $n \in \mathbb{N}_+$ such that $x \notin nH$. Since H/nH is finite, the quotient $G_2 = G_1/nH$ is compact by Lemma 4.5. Consider the canonical projection $\pi : G_1 \to G_2$ and note that $\pi(x) = y \neq 0$ in G_2. By the Peter-Weyl Theorem 6.4 there exists $\xi \in \widehat{G}$ such that $\xi(y) \neq 0$. Consequently $\chi = \xi \circ \pi \in \widehat{G_1}$ and $\chi(x) \neq 0$. By Theorem 2.5 there exists $\overline{\chi} \in \widehat{G}$ such that $\overline{\chi} \restriction_{G_1} = \chi$. □

It follows from Theorem 6.19 and Remark 7.24 that ω_G is a continuous monomorphism for every locally compact abelian group G.

Corollary 6.20. *Let G be a locally compact abelian group and K a compact subgroup of G. Then for every $\chi \in \widehat{K}$ there exists $\xi \in \widehat{G}$ such that $\xi \restriction_K = \chi$.*

Proof. Define $H = \{\chi \in \widehat{K} : \text{there exists } \xi \in \widehat{G} \text{ with } \xi \restriction_K = \chi\}$. By Theorem 6.19 the continuous characters of G separate the points of G. Therefore H separate the points of K. Now apply Corollary 6.6 to conclude that $H = \widehat{K}$. □

Here is another corollary of Theorem 6.19:

Corollary 6.21. *A σ-compact and locally compact abelian group is totally disconnected iff for every continuous character χ of G the image $\chi(G)$ is a proper subgroup of $\mathbb{T}$.*

Units 7, 8

7 Pontryagin-van Kampen duality

7.1 The dual group

In the sequel we shall write the circle additively as $(\mathbb{T}, +)$ and we denote by $q_0 : \mathbb{R} \to \mathbb{T} = \mathbb{R}/\mathbb{Z}$ the canonical projection. For every $k \in \mathbb{N}_+$ let $\Lambda_k = q_0((-\frac{1}{3k}, \frac{1}{3k}))$. Then $\{\Lambda_k : k \in \mathbb{N}_+\}$ is a base of the neighborhoods of 0 in $\mathbb{T}$, because $\{(-\frac{1}{3k}, \frac{1}{3k}) : k \in \mathbb{N}_+\}$ is a base of the neighborhoods of 0 in $\mathbb{R}$.

For every abelian group $G^* = \mathrm{Hom}\,(G, \mathbb{T})$. For a subset K of G and a subset U of $\mathbb{T}$ let

$$W_{G^*}(K, U) = \{\chi \in G : \chi(K) \subseteq U\}.$$

For any subgroup H of G^* we abbreviate $H \cap W(K, U)$ to $W_H(K, U)$. When there is no danger of confusion we shall write only $W(K, U)$ in place of $W_{G^*}(K, U)$. The group G^* will be considered only with one topology, namely the induced from $\mathbb{T}^G$ compact topology (see Remark 4.1).

If G is a topological abelian group, $\widehat{G}$ will denote the subgroup of G^* consisting of continuous characters.

The group $\widehat{G}$ will carry the *compact open topology* that has as basic neighborhoods of 0 the sets $W_{\widehat{G}}(K, U)$, where K is a compact subset of G and U is neighborhood of 0 in $\mathbb{T}$. We shall see below that when $U \subseteq \Lambda_1$, then $W_{\widehat{G}}(K, U)$ coincides with $W_{G^*}(K, U)$ in case K is a neighborhood of 0 in G. Therefore we shall use mainly the notation $W(K, U)$ when the group G is clear from the context.

Let us start with an easy example.

Example 7.1. Let G be an abelian topological group.

(1) If G is compact, then $\widehat{G}$ is discrete.

(2) If G is discrete, then $\widehat{G}$ is compact.

Indeed, to prove (1) it is sufficient to note that $W_{\widehat{G}}(G, \Lambda_1) = \{0\}$ as Λ_1 contains no subgroup of $\mathbb{T}$ beyond 0.

(2) Suppose that G is discrete. Then $\widehat{G} = \mathrm{Hom}\,(G, \mathbb{T})$ is a subgroup of the compact group $\mathbb{T}^G$. The compact-open topology of $\widehat{G}$ coincides with the topology inherited from $\mathbb{T}^G$: let F be a finite subset of G and U an open neighborhood of 0 in $\mathbb{T}$, then

$$\begin{aligned}\bigcap_{x \in F} \pi_x^{-1}(U) \cap \mathrm{Hom}\,(G, \mathbb{T}) &= \{\chi \in \mathrm{Hom}\,(G, \mathbb{T}) : \pi_x \in U \text{ for every } x \in F\} \\ &= \{\chi \in \mathrm{Hom}\,(G, \mathbb{T}) : \chi(x) \in U \text{ for every } x \in F\} = W(F, U).\end{aligned}$$

Moreover $\mathrm{Hom}\,(G, \mathbb{T})$ is closed in the compact product $\mathbb{T}^G$ by Remark 4.1 and we can conclude that $\widehat{G}$ is compact.

Now we prove that the dual group is always a topological group. If the group G is locally compact, then its dual is locally compact too. This is the first step of the Pontryagin-van Kampen duality theorem.

Theorem 7.2. *For an abelian topological group G the following assertions hold true:*

(a) if $x \in \mathbb{T}$ and $k \in \mathbb{N}_+$, then $x \in \Lambda_k$ if and only if $x, 2x, \dots, kx \in \Lambda_1$;

(b) $\chi \in Hom(G, \mathbb{T})$ is continuous if and only if $\chi^{-1}(\Lambda_1)$ is a neighborhood of 0 in G;

(c) $\{W_{\widehat{G}}(K, \Lambda_1) : K \text{ compact } \subseteq G\}$ is a base of the neighborhoods of 0 in $\widehat{G}$, in particular $\widehat{G}$ is a topological group.

(d) $W_{\widehat{G}}(A, \Lambda_s) + W_{\widehat{G}}(A, \Lambda_s) \subseteq W_{\widehat{G}}(A, \Lambda_{s-1})$ and $W_{\widehat{G}}(\overline{A}, \Lambda_s) + W_{\widehat{G}}(\overline{A}, \Lambda_s) \subseteq W_{\widehat{G}}(\overline{A}, \Lambda_{s-1})$ for every $A \subseteq G$ and $s > 1$.

(e) if F is a closed subset of $\mathbb{T}$, then for every $K \subseteq G$ the subset $W_{G^*}(K, F)$ of G^* is closed (hence, compact);

(f) if U is neighborhoodof 0 in G, then

(f_1) $W_{\widehat{G}}(\overline{U}, V) = W_{G^*}(\overline{U}, V)$ for every neighborhood of 0 $V \subseteq \Lambda_1$ in $\mathbb{T}$;

(f_2) $W(\overline{U}, \Lambda_4)$ has compact closure;

(f_3) if U has compact closure, then $W(\overline{U}, \Lambda_4)$ is a neighborhood of 0 in $\widehat{G}$ with compact closure, so $\widehat{G}$ is locally compact.

Proof. (a) Note that for $s \in \mathbb{N}$, $sx \in \Lambda_1$ if and only if $x \in A_{s,t} = \Lambda_s + \pi_{\mathbb{T}}(\frac{t}{s})$ for some integer t with $0 \le t \le s$. On the other hand, $A_{s,0} = \Lambda_s$ and $\Lambda_s \cap A_{s+1,t}$ is non-empty if and only if $t = 0$. Hence, if $x \in \Lambda_s$ and $(s+1)x \in \Lambda_1$, then $x \in \Lambda_{s+1}$ and this holds in particular for $1 \le s < k$. This proves that $sx \in \Lambda_1$ for $s = 1, \dots, k$ if and only if $x \in \Lambda_k$.

(b) Suppose that $\chi^{-1}(\Lambda_1)$ is a neighborhood of 0 in G. So there exists an open neighborhood U of 0 in G such that $U \subseteq \chi^{-1}(\Lambda_1)$. Moreover, there exists an other neighborhood V of 0 in G with $\underbrace{V + \dots + V}_{k} \subseteq U$ where $k \in \mathbb{N}_+$. Now $s\chi(y) \in \Lambda_1$ for every $y \in V$ and $s = 1, \dots, k$. By item (a) $\chi(y) \in \Lambda_k$ and so $\chi(V) \subseteq \Lambda_k$.

(c) Let $k \in \mathbb{N}_+$ and K be a compact subset of G. Define $L = \underbrace{K + \dots + K}_{k}$, which is a compact subset of G because it is a continuous image of the compact subset K^k of G^k. Take $\chi \in W(L, \Lambda_1)$. For every $x \in K$ we have $s\chi(x) \in \Lambda_1$ for $s = 1, \dots, k$ and so $\chi(x) \in \Lambda_k$ by item (a). Hence $W(L, \Lambda_1) \subseteq W(K, \Lambda_k)$.

(d) obvious.

(e) If $\pi_x : \mathbb{T}^G \to \mathbb{T}$ is the projection defined by the evaluation at x, for $x \in G$, then obviously

$$W_{G^*}(K, F) = \bigcap_{x \in K} \{\chi \in G^* : \chi(x) \in F\} = \bigcap_{x \in K} \pi_x^{-1}(F)$$

is cloased as each $\pi_x^{-1}(F)$ is closed in G^*.

(f_1) follows immediately from item (c).

(f_2) To prove that the closure of $W_0 = W(\overline{U}, \Lambda_4)$ is compact it is sufficient to note that $W_0 \subseteq W_1 := W(\overline{U}, \overline{\Lambda_4})$ and prove that W_1 is compact. Let τ_s denote the subspace topology of W_1 in $\widehat{G}$. We prove in the sequel that (W_1, τ_s) is compact.

Consider on the set W_1 also the weaker topology τ induced from G^* and consequently from $\mathbb{T}^G$. By (e) (W_1, τ) is compact.

It remains to show that both topologies τ_s and τ of W_1 coincide. Since τ_s is finer than τ, it suffices to show that if $\alpha \in W_1$ and K is a compact subset of G, then $(\alpha + W(K, \Lambda_1)) \cap W_1$ is also a neighborhood of α in (W_1, τ).

Since $\bigcup\{a + U : a \in K\} \supseteq K$ and K is compact, $K \subseteq F + U$, where F is a finite subset of K. We prove now that

$$(\alpha + W(F, \Lambda_2)) \cap W_1 \subseteq (\alpha + W(K, \Lambda_1)) \cap W_1. \qquad (*)$$

Let $\xi \in W(F, \Lambda_2)$, so that $\alpha + \xi' \in W_1 = W(\overline{U}, \overline{\Lambda_4})$. As $\alpha \in W_1$ as well, we deduce from items (c) and (d) that $\xi = (\alpha + \xi') - \alpha \in W_1 - W_1$. Hence $\xi(\overline{U}) \subseteq \overline{\Lambda_2}$ and consequently

$$\xi(K) \subseteq \xi(F + U) \subseteq \Lambda_2 + \overline{\Lambda_2} \subseteq \Lambda_1.$$

This proves $\xi \in W(K, \Lambda_1)$ and (*).

(f_3) Follows obviously from (f_2) and the definition of the compact open topology. □

The above proof shows another relevant fact. The neighborhood $W(\overline{U}, \Lambda_4)$ of 0 in the dual group $\widehat{G}$ carries the same topology in $\widehat{G}$ and G^*, nevertheless the inclusion map $j : \widehat{G} \hookrightarrow G^*$ need not be an embedding:

Corollary 7.3. *For a locally compact abelian group G the following are equivalent:*

(a) the inclusion map $j : \widehat{G} \hookrightarrow G^*$ is an embedding;

(b) G is discrete;

(c) $\widehat{G} = G^*$ is compact.

Proof. Since G^* is compact, j can be an embedding iff $\widehat{G}$ itself is compact. According to Example 7.1 this occurs precisely when G is discrete. In that case $\widehat{G} = G^*$ is compact. □

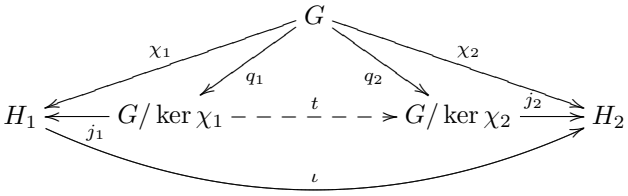
Actually, it can be proved, once the duality theorem is available, that $j : \widehat{G} \hookrightarrow G^*$ need not be even a local homeomorphism. (If j is a local homeomorphism, then the topological subgroup $j(\widehat{G})$ of G^* will be locally compact, hence closed in G^*. This would yield that $j(\widehat{G})$ is compact. On the other hand, the topology of $j(\widehat{G})$ is precisely the initial topology of all projections p_x restricted to $\widehat{G}$. By the Pontryagin duality theorem, these projections form the group of all continuous characters of $\widehat{G}$. So this topology coincides with $\mathcal{T}_{\widehat{\widehat{G}}}$. By a general theorem of Glicksberg, a locally compact abelian groups H and $(H, \mathcal{T}_{\widehat{H}})$ have the same compact sets. In particular, compactness of $(H, \mathcal{T}_{\widehat{H}})$ yields compactness of H. This proves that if $j : \widehat{G} \hookrightarrow G^*$ is a local homeomorphism, then $\widehat{G}$ is compact and consequently G is discrete.)

7.2 Computation of some dual groups

In the next proprosition we show, roughly speaking, that the projective order between continuous surjective open homomorphisms with the same domain corresponds to the order by inclusion of their kernels.

Proposition 7.4. *Let G, H_1 and H_2 be topological abelian groups and let $\chi_i : G \to H_i$, $i = 1, 2$, be continuous surjective open homomorphisms. Then there exists a continuous homomorphism $\iota : H_1 \to H_2$ such that $\chi_2 = \iota \circ \chi_1$ iff $\ker \chi_1 \leq \ker \chi_2$. If $\ker \chi_1 = \ker \chi_2$ then ι will be a topological isomorphism.*

Proof. The necessity is obvious. So assume that $\ker \chi_1 \leq \ker \chi_2$ holds. By the homomorphism theorem applied to χ_i there exists a topological isomorphisms$j_i : G/\ker \chi_i \to H_i$ such that $\chi_i = j_i \circ q_i$, where $q_i : G \to G/\ker \chi_i$ is the canonical homomorphism for $i = 1, 2$. As $\ker \chi_1 \leq \ker \chi_2$ we get a continuous homomorphism t that makes commutative the following diagram



Obviously $\iota = j_2 \circ t \circ j_1^{-1}$ works. If $\ker \chi_1 = \ker \chi_2$, then t is a topological isomorphism, hence ι will be a topological isomorphism as well. □

In the sequel we denote by $k \cdot id_G$ the endomorphism of an abelian group G obtained by the map $x \mapsto kx$, for a fixed $k \in \mathbb{Z}$. The next lemma will be used for the computation of the dual groups in Example 7.7.

Lemma 7.5. *Every continuous homomorphism $\chi : \mathbb{T} \to \mathbb{T}$ has the form $k \cdot id_{\mathbb{T}}$, for some $k \in \mathbb{Z}$. In particular, the only topological isomorphisms $\chi : \mathbb{T} \to \mathbb{T}$ are $\pm id_{\mathbb{T}}$.*

Proof. We prove first that the only topological isomorphisms $\chi : \mathbb{T} \to \mathbb{T}$ are $\pm id_{\mathbb{T}}$. The proof will exploit the fact that the arcs are the only connected sets of $\mathbb{T}$. Hence χ sends any arc of $\mathbb{T}$ to an arc, sending end points to end points. Denote by φ the canonical homomorphism $\mathbb{R} \to \mathbb{T}$ and for $n \in \mathbb{N}$ let $c_n = \varphi(1/2^n)$ be the generators of the Prüfer subgroup $\mathbb{Z}(2^\infty)$ of $\mathbb{T}$. Then, c_1 is the only element of $\mathbb{T}$ of order 2, hence $g(c_1) = c_1$. Therefore, the arc $A_1 = \varphi([0, 1/2])$ either goes onto itself, or goes onto its symmetric image $-A_1$. Let us consider the first case. Clearly, either $g(c_2) = c_2$ or $g(c_2) = -c_2$ as $o(g(c_2)) = 4$ and being $\pm c_2$ the only elements of order 4 of $\mathbb{T}$. By our assumption $g(A_1) = A_1$ we have $g(c_2) = c_2$ since c_2 is the only element of order 4 on the arc A_1. Now the arc $A_2 = [0, c_2]$ goes onto itself, hence for c_3 we must have $g(c_3) = c_3$ as the only element of order 8 on the

arc A_2, etc. We see in the same way that $g(c_n) = c_n$. Hence g is identical on the whole subgroup $\mathbb{Z}(2^\infty)$. As this subgroup is dense in $\mathbb{T}$, we conclude that g coincides with $id_\mathbb{T}$. In the case $g(A_1) = -A_1$ we replace g by $-g$ and the previous proof gives $-g = id_\mathbb{T}$, i.e., $g = -id_\mathbb{T}$.

For $k \in \mathbb{N}_+$ let $\pi_k = k \cdot id_\mathbb{T}$. Then $\ker \pi_k = \mathbb{Z}_k$ and π_k is surjective. Let now $\chi : \mathbb{T} \to \mathbb{T}$ be a non-trivial continuous homomorphism. Then $\ker \chi$ is a closed proper subgroup of $\mathbb{T}$, hence $\ker \chi = \mathbb{Z}_k$ for some $k \in \mathbb{N}_+$. Moreover, $\chi(\mathbb{T})$ is a connected non-trivial subgroup of $\mathbb{T}$, hence $\chi(\mathbb{T}) = \mathbb{T}$. By Proposition 7.4 $\chi = \pm\pi_k$. □

Obviously, $\chi = \pm\xi$ for characters $\chi, \xi : G \to \mathbb{T}$ implies $\ker \chi = \ker \xi$ and $\chi(G) = \xi(G)$. More generally, if $\chi = k \cdot \xi$ for some $k \in \mathbb{Z}$, then $\ker \chi \geq \ker \xi$ and $\chi(G) \leq \xi(G)$. Now we see that this implication can be (partially) inverted under appropriate hypotheses.

Corollary 7.6. *Let G be a σ-compact locally compact abelian group and let $\chi, \xi : G \to \mathbb{T}$ be continuous characters such that $\ker \chi \geq \ker \xi$ and $\chi(G) \leq \xi(G)$.*

(a) If $\chi(G) = \xi(G) = \mathbb{T}$ then $\chi = k \cdot \xi$ for some $k \in \mathbb{Z}$; moreover, $\ker \chi = \ker \xi$ iff $\chi = \pm\xi$.

(b) If G is compact and $|\xi(G)| = m$ for some $m \in \mathbb{N}_+$, then $\chi = k\xi$ for some $k \in \mathbb{Z}$; moreover, $\ker \chi = \ker \xi$ iff $\chi(G) = \xi(G)$, in such a case k must be coprime to m.

(c) If $\ker \xi = \ker \xi$ is open and $H = \chi(G) = \xi(G)$, then $\chi = \iota \circ \xi$, where $\iota : H \to H$ is an arbitrary automorphism of the subgroup H of $\mathbb{T}$ equipped with the discrete topology.

Proof. (a) As $\chi(G) = \xi(G) = \mathbb{T}$ and G is σ-compact, we can apply Lemma 7.4 and observe that the only ι given by the lemma can be $k \cdot id_\mathbb{T}$ for some $k \in \mathbb{Z}$ in view of the previous lemma. The same lemma yields $k = \pm 1$ when $\ker \chi = \ker \xi$.

(b) If G is compact and $|\xi(G)| = m$ for some $m \in \mathbb{N}_+$, $\xi(G)$ is a cyclic subgroup of $\mathbb{T}$ of order m. Note that $\mathbb{T}$ has a unique such cyclic subgroup. By Proposition 7.4 there exists a homomorphism $\iota : \xi(G) \to \chi(G)$ such that $\chi = \iota \circ \xi$. The hypothesis $\chi(G) \leq \xi(G)$ implies that there such a ι must by the multiplication by some $k \in \mathbb{Z}$. In case $\chi(G) = \xi(G)$ this k is coprime to m.

(c) Obvious. □

Example 7.7. Let p be a prime. Then $\widehat{Z(p^\infty)} \cong \mathbb{J}_p$, $\widehat{\mathbb{J}_p} \cong \mathbb{Z}(p^\infty)$, $\widehat{\mathbb{T}} \cong \mathbb{Z}$, $\widehat{\mathbb{Z}} \cong \mathbb{T}$ and $\widehat{\mathbb{R}} \cong \mathbb{R}$.

Proof. The first isomorphism $\widehat{\mathbb{Z}(p^\infty)} = \mathbb{J}_p$ follows from our definition $\mathbb{J}_p = \mathrm{End}(\mathbb{Z}(p^\infty)) = \mathrm{Hom}(\mathbb{Z}(p^\infty), \mathbb{T}) = \widehat{\mathbb{Z}(p^\infty)}$.

To verify the isomorphism $\widehat{\mathbb{J}_p} \cong \mathbb{Z}(p^\infty)$ consider first the quotient homomorphism $\eta_n : \mathbb{J}_p \to \mathbb{J}_p/p^n\mathbb{J}_p \cong \mathbb{Z}_{p^n} \leq \mathbb{T}$. With this identifications we consider $\eta_n \in \widehat{\mathbb{J}_p}$. It is easy to see that under this identification $p\eta_n = \eta_{n-1}$. Therefore, the subgroup H of $\widehat{\mathbb{J}_p}$ generated by the characters η_n is isomorphic to $\mathbb{Z}(p^\infty)$. Let us see that $H = \widehat{\mathbb{J}_p}$. Indeed, take any non-trivial character $\chi : \mathbb{J}_p \to \mathbb{T}$. Then $N = \ker \chi$ is a closed proper subgroup of $\mathbb{J}_p$. Moreover, $N \neq 0$ as $\mathbb{J}_p$ is not isomorphic to a subgroup of $\mathbb{T}$ by Exercise 4.49. Thus $N = p^n\mathbb{J}_p$ for some $n \in \mathbb{N}_+$. Since $N = \ker \eta_n$, we conclude with (b) of Corollary 7.6 that $\chi = k\eta_n$ for some $k \in \mathbb{Z}$. This proves that $\chi = H$ and consequently $\widehat{\mathbb{J}_p} \cong \mathbb{Z}(p^\infty)$.

The isomorphism $g : \widehat{\mathbb{Z}} \to \mathbb{T}$ is obtained by setting $g(\chi) := \chi(1)$ for every $\chi : \mathbb{Z} \to \mathbb{T}$. It is easy to check that this isomorphism is topological.

According to 7.5 every $\chi \in \widehat{\mathbb{T}}$ has the form $\chi = k \cdot id_\mathbb{T}$ for some $k \in \mathbb{Z}$. This gives a homomorphism $\widehat{\mathbb{T}} \to \mathbb{Z}$ assigning $\chi \mapsto k$. It is obviously injective and surjective. This proves $\widehat{\mathbb{T}} \cong \mathbb{Z}$ since both groups are discrete.

To prove $\widehat{\mathbb{R}} \cong \mathbb{R}$ consider the character $\chi_1 : \mathbb{R} \to \mathbb{T}$ obtained simply by the canonical map $\mathbb{R} \to \mathbb{R}/\mathbb{Z}$. For every non-zero $r \in \mathbb{R}$ consider the map $\rho_r : \mathbb{R} \to \mathbb{R}$ defined by $\rho_r(x) = rx$. Then its composition $\chi_r = \chi_1 \circ \rho_r$ with χ_1 gives a continuous character of $\mathbb{R}$ that is surjective and $\ker \chi_r = \langle 1/r \rangle$. Now consider any continuous non-trivial character $\chi \in \widehat{\mathbb{R}}$. Then χ is surjective and $N = \ker \chi$ is a proper closed subgroup of $\mathbb{R}$. Hence N is cyclic by Exercise 3.20. Let $N = \langle 1/r \rangle$. Then $\ker \chi = \ker \chi_r$, so that Corollary 7.6 yields $\chi = \pm\chi_r$. The assignment $\chi \mapsto \pm r$ defines a homomorphism $\widehat{\mathbb{R}} \to \mathbb{R}$ that is obviously injective and surjective. Its continuity immediately follows from the definition of the compact-open topology of $\widehat{\mathbb{R}}$. As $\mathbb{R}$ is σ-compact, this isomorphism is also open by the open mapping theorem. □

Exercise 7.8. *Let G be an abelian group and p be a prime. Prove that*

(a) $\chi \in p\widehat{G}$ iff $\chi(G[p]) = 0$.

(b) $p\chi = 0$ in $\widehat{G}$ iff $\chi(pG) = 0$.

Conclude that

(i) a discrete abelian group G is divisible (resp., torsion-free) iff $\widehat{G}$ is torsion-free (resp., divisible).

(ii) the groups $\widehat{\mathbb{Q}}$ and $\widehat{\mathbb{Q}_p}$ are torsion-free and divisible.

Exercise 7.9. *Let G be a totally disconnected locally compact abelian group. Prove that* $\ker\chi$ *is an open subgroup of G for every $\chi \in \widehat{G}$.*

(Hint. Use the fact that by the continuity of χ and the total disconnectedness of G there exists an open subgroup O of G such that $\chi(O) \subseteq \Lambda_1$.)

Exercise 7.10. *Let p be a prime. Prove that $\widehat{\mathbb{Q}_p} \cong \mathbb{Q}_p$, where $\mathbb{Q}_p$ denotes the field of all p-adic numbers.*

(Hint. Fix $N = \{\chi \in \widehat{\mathbb{Q}_p} : \ker\chi \geq \mathbb{J}_p\}$. By the compactness of $\mathbb{J}_p$, conclude that N is an open subgroup of $\widehat{\mathbb{Q}_p}$ topologically isomorphic to $\mathbb{J}_p$ using Exercise 7.9 and Corollary 7.6 (c). For every $n \in \mathbb{N}_+$ let $\xi_n : \mathbb{Q}_p \to \mathbb{Q}_p/p^n\mathbb{J}_p$ be the canonical homomorphism. As $\mathbb{Q}_p/p^{-n}\mathbb{J}_p \cong \mathbb{Z}(p^\infty) \leq \mathbb{T}$, we can consider $\xi_n \in \widehat{\mathbb{Q}_p}$. Show that $p\xi_{n+1} = \xi_n$ for $n \in \mathbb{N}_+$ and $p\xi_1 \in N$. The subgroup of $\widehat{\mathbb{Q}_p}$ generated by N and (ξ_n) is isomorphic to $\mathbb{Q}_p$. Using Corollary 7.6 (c) and Exercise 7.9 deduce that it coincides with the whole group $\widehat{\mathbb{Q}_p}$.)

Exercise 7.11. *Let H be a subgroup of $\mathbb{R}^n$. Prove that every $\chi \in \widehat{H}$ extends to a continuous character of $\mathbb{R}^n$.*

7.3 Some general properties of the dual

We prove next that the dual group of a finite product of abelian topological groups is the product of the dual groups of each group.

Lemma 7.12. *If G and H are topological abelian groups, then $\widehat{G \times H}$ is isomorphic to $\widehat{G} \times \widehat{H}$.*

Proof. Define $\Phi : \widehat{G} \times \widehat{H} \to \widehat{G \times H}$ by $\Phi(\chi_1, \chi_2)(x_1, x_2) = \chi_1(x_1) + \chi_2(x_2)$ for every $(\chi_1, \chi_2) \in \widehat{G} \times \widehat{H}$ and $(x_1, x_2) \in G \times H$. Then Φ is a homomorphism, in fact $\Phi(\chi_1+\psi_1, \chi_2+\psi_2)(x_1, x_2) = (\chi_1+\psi_1)(x_1)+(\chi_2+\psi_2)(x_2) = \chi_1(x_1) + \psi_1(x_1) + \chi_2(x_2) + \psi_2(x_2) = \Phi(\chi_1, \chi_2)(x_1, x_2) + \Phi(\psi_1, \psi_2)(x_1, x_2)$.

Moreover Φ is injective, because

$$\begin{aligned}
\ker\Phi &= \{(\chi, \psi) \in \widehat{G} \times \widehat{H} : \Phi(\chi, \psi) = 0\} \\
&= \{(\chi, \psi) \in \widehat{G} \times \widehat{H} : \Phi(\chi, \psi)(x, y) = 0 \text{ for every } (x, y) \in G \times H\} \\
&= \{(\chi, \psi) \in \widehat{G} \times \widehat{H} : \chi(x) + \psi(y) = 0 \text{ for every } (x, y) \in G \times H\} \\
&= \{(\chi, \psi) \in \widehat{G} \times \widehat{H} : \chi(x) = 0 \text{ and } \psi(x) = 0 \text{ for every } (x, y) \in G \times H\} \\
&= \{(0, 0)\}.
\end{aligned}$$

To prove that Φ is surjective, take $\psi \in \widehat{G \times H}$ and note that $\psi(x_1, x_2) = \psi(x_1, 0) + \psi(0, x_2)$. Now define $\psi_1(x_1) = \psi(x_1, 0)$ for every $x_1 \in G$ and $\psi_2(x_2) = \psi(0, x_2)$ for every $x_2 \in H$. Hence $\psi_1 \in \widehat{G}$, $\psi_2 \in \widehat{H}$ and $\psi = \Phi(\psi_1, \psi_2)$.

Now we show that Φ is continuous. Let $W(K, U)$ be an open neighborhood of 0 in $\widehat{G \times H}$ (K is a compact subset of $G \times H$ and U is an open neighborhood of 0 in $\mathbb{T}$). Since the projections π_G and π_H of $G \times H$ onto G and H are continuous, $K_G = \pi_G(K)$ and $K_H = \pi_H(K)$ are compact in G and in H respectively. Taking an open symmetric neighborhood V of 0 in $\mathbb{T}$, it follows $\Phi(W(K_G, V) \times W(K_H, V)) \subseteq W(K, U)$.

It remains to prove that Φ is open. Consider two open neighborhoods $W(K_G, U_G)$ of 0 in $\widehat{G}$ and $W(K_H, U_H)$ of 0 in $\widehat{H}$, where $K_G \subseteq G$ and $K_H \subseteq H$ are compact and U_G, U_H are open neighborhoods of 0 in $\mathbb{T}$. Then $K = (K_G \cup \{0\}) \times (K_H \cup \{0\})$ is a compact subset of $G \times H$ and $U = U_G \cap U_H$ is an open neighborhood of 0 in $\mathbb{T}$. Thus $W(K, U) \subseteq \Phi(W(K_G, U_G) \times W(K_H, U_H))$, because if $\chi \in W(K, U)$ then $\chi = \Phi(\chi_1, \chi_2)$, where $\chi_1(x_1) = \chi(x_1, 0) \in U \subseteq U_G$ for every $x_1 \in G$ and $\chi_2(x_2) = \chi(0, x_2) \in U \subseteq U_H$ for every $x_2 \in H$. □

It follows from Proposition 7.7 that the groups $\mathbb{T}$, $\mathbb{Z}$, $\mathbb{Z}(p^\infty)$, $\mathbb{J}_p$ e $\mathbb{R}$ satisfy $\widehat{\widehat{G}} \cong G$, namely the Pontryagin-van Kampen duality theorem. Using the next theorem this propertiy extends to all finite direct products of these groups.

Call a topological abelian group G *autodual*, if G satisfies $\widehat{G} \cong G$. We have seen already that $\mathbb{R}$ and $\mathbb{Q}_p$ are autodual. By Lemma 7.12 finite direct products of autodual groups are autodual. Now using this observation and Lemma 7.12 we provide a large supply of groups for which the Pontryagin-van Kampen duality holds true.

Proposition 7.13. *Let P_1, P_2 and P_3 be finite sets of primes, $m, n, k, k_p \in \mathbb{N}$ $(p \in P_3)$ and $n_p, m_p \in \mathbb{N}_+$ $(p \in P_1 \cup P_2)$. Then every group of the form*

$$G = \mathbb{T}^n \times \mathbb{Z}^m \times \mathbb{R}^k \times F \times \prod_{p \in P_1} \mathbb{Z}(p^\infty)^{n_p} \times \prod_{p \in P_2} \mathbb{J}_p^{m_p} \times \prod_{j \in P_3} \mathbb{Q}_p^{k_p},$$

where F is a finite abelian group, satisfies $\widehat{\widehat{G}} \cong G$.

Moreover, such a group is autodual iff $n = m$, $P_1 = P_2$ and $n_p = m_p$ for all $p \in P_1 = P_2$. In particular, $\widehat{\widehat{G}} \cong G$ holds true for all elementary locally compact abelian groups.

Proof. Let us start by proving $\widehat{F} = F^* \cong F$. Recall that F has the form $F \cong \mathbb{Z}_{n_1} \times \ldots \times \mathbb{Z}_{n_m}$. So applying Theorem 7.14 we are left with the proof of the isomorphism $\mathbb{Z}_n^* \cong \mathbb{Z}_n$ for every $n \in \mathbb{N}_+$. The elements x of $\mathbb{T}$ satisfying $nx = 0$ are precisely those of the unique cyclic subgroup of order n of $\mathbb{T}$, we shall denote that subgroup by $\mathbb{Z}_n$. Therefore, the group $Hom(\mathbb{Z}_n, \mathbb{Z}_n)$ of all homomorphisms $\mathbb{Z}_n \to \mathbb{Z}_n$ is isomorphic to $\mathbb{Z}_n$.

It follows easily from Lemma 7.12 that if $\widehat{\widehat{G_i}} \cong G_i$ (resp., $\widehat{G_i} \cong G_i$) for a finite family $\{G_i\}_{i=1}^n$ of topological abelian groups, then also $G = \prod_{i=1}^n G_i$ satisfies $\widehat{\widehat{G}} \cong G$ (resp., $\widehat{G} \cong G$). Therefore, it suffices to verify that the groups $\mathbb{T}$, $\mathbb{Z}$, $\mathbb{Z}(p^\infty)$, and $\mathbb{J}_p$ e satisfy $\widehat{\widehat{G}} \cong G$, while $\widehat{\mathbb{R}} \cong \mathbb{R}$, $\widehat{\mathbb{Q}_p} \cong \mathbb{Q}_p$ were already checked.

It follows from Proposition 7.7 that $\widehat{\mathbb{Z}} \cong \mathbb{T}$ and $\widehat{\mathbb{T}} \cong \mathbb{Z}$, hence $\mathbb{Z} \cong \widehat{\widehat{\mathbb{Z}}}$ and $\mathbb{T} \cong \widehat{\widehat{\mathbb{T}}}$. Analogously, $\widehat{\mathbb{Z}(p^\infty)} \cong \mathbb{J}_p$ and $\widehat{\mathbb{J}_p} \cong \mathbb{Z}(p^\infty)$ yield $\mathbb{Z}(p^\infty) \cong \widehat{\widehat{\mathbb{Z}(p^\infty)}}$ and $\mathbb{J}_p \cong \widehat{\widehat{\mathbb{J}_p}}$. □

The problem of characterizing all autodual locally compact abelian groups is still open [47, 48].

Theorem 7.14. *Let $\{D_i\}_{i \in I}$ be a family of discrete abelian groups and let $\{G_i\}_{i \in I}$ be a family of compact abelian groups. Then*

$$\widehat{\bigoplus_{i \in I} D_i} \cong \prod_{i \in I} \widehat{D_i} \quad \text{and} \quad \widehat{\prod_{i \in I} G_i} \cong \bigoplus_{i \in I} \widehat{G_i}. \tag{5}$$

Proof. Let $\chi : \bigoplus_{i \in I} D_i \to \mathbb{T}$ be a character and let $\chi_i : D_i \to \mathbb{T}$ be its restriction to D_i. Then $\chi \mapsto (\chi_i) \in \prod_{i \in I} \widehat{D_i}$ is the first isomorphism in (5).

Let $\chi : \prod_{i \in I} G_i \to \mathbb{T}$ be a continuous character. Pick a neighborhood U of 0 containing no non-trivial subgroups of $\mathbb{T}$. Then there exists a neighborhood V of 0 in $G = \prod_{i \in I} G_i$ with $\chi(V) \subseteq U$. By the definition of the Tychonov topology there exists a finite subset $F \subseteq I$ such that V contains the subproduct $B = \prod_{i \in I \setminus F} G_i$. Being $\chi(B)$ a subgroup of $\mathbb{T}$, we conclude that $\chi(B) = 0$ by the choice of U. Hence χ factorizes through the projection $p : G \to \prod_{i \in F} G_i = G/B$; so there exists a character $\chi' : \prod_{i \in F} G_i \to \mathbb{T}$ such that $\chi = \chi' \circ p$. Obviously, $\chi' \in \oplus_{i \in I} \widehat{G_i}$. Then $\chi \mapsto \chi'$ is the second isomorphism in (5). □

In order to extend the isomorphism (5) to the general case of locally compact abelian groups one has to consider a specific topology on the direct sum.

Algebraic properties of the dual group $\widehat{G}$ of a compact abelian group G can be described in terms of topological properties of the group G. We saw in Corollary 6.22 that $\widehat{G}$ is torsion precisely when G is totally disconnected. Here is the counterpart of this property in the connected case:

Proposition 7.15. *Let G be a topological abelian group.*

(a) If G is connected, then the dual group $\widehat{G}$ is torsion-free.

(b) If G is compact, then the dual group $\widehat{G}$ is torsion-free iff G is connected.

Proof. (a) Since for every non-zero continuous character $\chi : G \to \mathbb{T}$ the image $\chi(G)$ is a non-trivial connected subgroup of $\mathbb{T}$, we deduce that $\chi(G) = \mathbb{T}$ for every non-zero $\chi \in \widehat{G}$. Hence $\widehat{G}$ is torsion-free.

(b) If the group G is compact and disconnected, then by Theorem 4.19 there exists a proper open subgroup N of G. Take any non-zero character ξ of the finite group G/N. Then $m\xi = 0$ for some positive integer m. Now the composition χ of ξ and the canonical homomorphism $G \to G/N$ satisfies $m\chi = 0$ as well. So $\widehat{G}$ haas a non-zero torsion character. This proves the implication left open by item (a). □

Let G and H be abelian topological groups. If $f : G \to H$ is a continuous homomorphism, define $\widehat{f} : \widehat{H} \to \widehat{G}$ putting $\widehat{f}(\chi) = \chi \circ f$ for every $\chi \in \widehat{H}$.

Lemma 7.16. *If $f : G \to H$ is a continuous homomorphism of topological abelian group, then $\widehat{f}(\chi) = \chi \circ f$ is a continuous homomorphism as well.*

(a) If $f(G)$ is dense in H, then $\widehat{f}$ is injective.

(b) If f is injective and $f(G)$ is either open or dense in H, then $\widehat{f}$ is surjective.

(c) if f is a surjective homomorphism, such that every compact subset of H is covered by some compact subset of G, then $\widehat{f}$ is an embedding.

(d) if f is a quotient homomorphism and G is locally compact, then $\widehat{f}$ is an embedding.

(e) If f is a topological isomorphism, then $\widehat{f}$ is a topological isomorphism too.

Proof. Assume K is a compact subset of G and U a neighborhood of 0 in $\mathbb{T}$. Then $f(K)$ is a compact set in H, so $W = W_{\widehat{G}}(f(K), U)$ is a neighborhood of 0 in $\widehat{H}$ and $\widehat{f}(W) \subseteq W(K, U)$. This proves the continuity of $\widehat{f}$.

(a) If $\widehat{f}(\chi) = 0$, then $\chi \circ f = 0$. By the density of $f(G)$ in H this yields $\chi = 0$.

(b) Let $\chi \in \widehat{G}$. If $f(G)$ is open in H, then any extension $\xi : H \to \mathbb{T}$ of χ will be continuous on $f(G)$. There exists at least one such extension ξ by Corollary 2.6. Hence $\xi \in \widehat{H}$ and $\chi = \widehat{f}(\xi)$. Now consider the case when $f(G)$ is dense in H. Then $\widetilde{H} = \widetilde{G}$ and the characters of H can be extended to characters of G (see Theorem 3.79).

(c) Assume L is a compact subset of G/H and U a neighborhood of 0 in $\mathbb{T}$. Let K be a compact set in G such that $f(K) = L$. Then $\widehat{f}(W_{\widehat{H}}(L, U)) = \mathrm{Im}\widehat{f} \cap W_{\widehat{G}}(K, U)$, so $\widehat{f}$ is an embedding.

(d) Follows from (c) and Lemma 4.6.

(e) Obvious.

□

Exercise 7.17. *Prove that $\widehat{\mathbb{Q}/\mathbb{Z}} \cong \prod_p \mathbb{J}_p$.*

(Hint. Use the isomorphism $\mathbb{Q}/\mathbb{Z} \cong \bigoplus_p \mathbb{Z}(p^\infty)$, Example 7.7 and Theorem 7.14.)
Now we shall see that the group $\mathbb{Q}$ satisfies the duality theorem (see item (b) below).

Example 7.18. Let K denote the compact group $\widehat{\mathbb{Q}}$. Then:

(a) K contains a closed subgroup H isomorphic to $\widehat{\mathbb{Q}/\mathbb{Z}}$ such that $K/H \cong \mathbb{T}$;

(ii) $\widehat{K} \cong \mathbb{Q}$.

(a) Denote by H the subgroup of all $\chi \in K$ such that $\chi(\mathbb{Z}) = 0$. To prove that H is a closed subgroup of K such that K/H is isomorphic to $\mathbb{T}$. To this end consider the continuous map $\rho : K \to \widehat{\mathbb{Z}}$ obtained by the restriction to $\mathbb{Z}$ of every $\chi \in K$. Obviously, $\ker\rho = H$, so $\mathbb{T} \cong \widehat{\mathbb{Z}} \cong K/H$. To see that $H \cong \widehat{\mathbb{Q}/\mathbb{Z}}$ note that the characters of $\mathbb{Q}/\mathbb{Z}$ correspond precisely to those characters of $\mathbb{Q}$ that vanish on $\mathbb{Z}$, i.e., precisely H.

(b) By Exercise 7.8 K is a divisible torsion-free group, every non-zero $r \in \mathbb{Q}$ defines a continuous automorphism λ_r of K by setting $\lambda_r(x) = rx$ for every $x \in K$. Then the composition $\rho \circ \lambda_r : K \to \mathbb{T}$ defines a character $\chi_r \in \widehat{K}$ with $\ker\chi_r = r^{-1}H$. For the sake of completeness let $\chi_0 = 0$. By Exercise 7.17 $\widehat{\mathbb{Q}/\mathbb{Z}} \cong \prod_p \mathbb{J}_p$ is totally disconnected, so by Corollary 6.21 H has no surjective characters $\chi : H \to \mathbb{T}$. Now let $\chi \in \widehat{K}$ be non-zero. Then $\chi(K)$ will be a non-zero closed divisible subgroup of $\mathbb{T}$, hence $\chi(K) = \mathbb{T}$. On the other hand, $N = \ker\chi$ is a proper closed subgroup of K such that $N + H \neq \mathbb{T}$, as $\chi(H)$ is a proper closed subgroup of $\mathbb{T}$ by the previous argument. Hence, $\chi(H)$ is finite, say of order m. Then $N + H$ contains N is a finite-index subgroup, more precisely $[H : (N \cap H)] = [(N + H) : N] = m$. Then $mH \leq N$. Consider the character $\chi_{m^{-1}}$ of K having $\ker\chi_{m^{-1}} = mH \leq N$. By Corollary there exists $k \in \mathbb{Z}$ such that $\chi = k\chi_{m^{-1}} = \chi_r$, where $r = km^{-1} \in \mathbb{Q}$. This shows that $\widehat{K} = \{\chi_r : r \in \mathbb{Q}\} \cong \mathbb{Q}$.)

The compact group $\widehat{\mathbb{Q}}$ is closely related to the adele rings of the field $\mathbb{Q}$, more detail can be found in [34, 38, 75, 97].

Exercise 7.19. *Prove that a discrete abelian group G satisfies $\widehat{\widehat{G}} \cong G$ whenever*

(a) G is divisible;

(b) G is free;

(c) G is of finite exponent;

(d) G is torsion and every primary component of G is of finite exponent.

(Hint. (a) Use Examples 7.7 and 7.18 (b) and the fact that every divisible group is a direct sum of copies of $\mathbb{Q}$ and the groups $\mathbb{Z}(p^\infty)$.

(c) and (d) Use that fact that every abelian group of finite exponent is a direct sum of cyclic subgroups (i.e., Prüfer's theorem, see (d) of Example 2.3).

Exercise 7.20. *Prove that every torsion compact abelian group G is bounded. More precisely, there exists natural numbers $m_1, \dots, m_n$ and cardinals $\alpha_1, \dots, \alpha_n$ such that $G \cong \prod_{i=1}^n \mathbb{Z}(m_i)^{\alpha_i}$.*

(Hint. Use the Baire category theorem for the union $G = \bigcup_{n=1}^\infty G[n!]$ of closed subgroups. Conclude that $G[n!]$ is open for some n, so must have finite index by the compactness of G. This yields $mG = 0$ for some m. Show that this yields also $m\widehat{G} = 0$. Now apply Prüfer's theorem to $\widehat{G}$ and the fact that $G \cong \widehat{\widehat{G}}$.)

7.4 The natural transformation ω

Let G be a topological abelian group. Define $\omega_G : G \to \widehat{\widehat{G}}$ such that $\omega_G(x)(\chi) = \chi(x)$, for every $x \in G$ and for every $\chi \in \widehat{G}$. We show now that $\omega_G(x) \in \widehat{\widehat{G}}$.

Proposition 7.21. *If G is a topological abelian group. Then $\omega_G(x) \in \widehat{\widehat{G}}$ and $\omega_G : G \to \widehat{\widehat{G}}$ is a homomorphism.*
If G is locally compact, then the homomorphism ω_G is a continuous.

Proof. In fact,

$$\omega_G(x)(\chi + \psi) = (\chi + \psi)(x) = \chi(x) + \psi(x) = \omega_G(x)(\chi) + \omega_G(x)(\psi),$$

for every $\chi, \psi \in \widehat{G}$. Moreover, if U is an open neighborhood of 0 in $\mathbb{T}$, then $\omega_G(x)(W(\{x\}, U)) \subseteq U$. This proves that $\omega_G(x)$ is a character of $\widehat{G}$, i.e., $\omega_G(x) \in \widehat{\widehat{G}}$. For every $x, y \in G$ and for every $\chi \in \widehat{G}$ we have $\omega_G(x + y)(\chi) = (\chi)(x + y) = \chi(x) + \chi(y) = \omega_G(\chi)(x) + \omega_G(\chi)(y)$ and so ω_G is a homomorphism.

Now assume G is locally compact. To prove that ω_G is continuous, pick an open neighborhood A of 0 in $\mathbb{T}$ and a compact subset K of $\widehat{G}$. Then $W(K, A)$ is an open neighborhood of 0 in $\widehat{\widehat{G}}$. Let U be an open neighborhood of 0 in G with compact closure. Take an open symmetric neighborhood B of 0 in $\mathbb{T}$ with $B+B \subseteq A$. Thus $W(\overline{U}, B)$ is an open neighborhood of 0 in $\widehat{G}$. Since K is compact, there exist finitely many characters $\chi_1, \dots, \chi_m$ of G such that $K \subseteq (\chi_1 + W(\overline{U}, B)) \cup \dots \cup (\chi_m + W(\overline{U}, B))$. For every $i = 1, \dots, m$ there is an open neighborhood V_i of 0 in G such that $\chi_i(V_i) \subseteq B$. Define $V = U \cap V_1 \cap \dots \cap V_m \subseteq U$ and note that $\chi_i(V) \subseteq B$ for every $i = 1, \dots, m$. Thus $\omega_G(V) \subseteq W(K, A)$. Indeed, if $x \in V$ and $\chi \in K$, then $\chi_i(x) \in B$ for every $i = 1, \dots, m$ and there exists $i_0 \in \{1, \dots, m\}$ such that $\chi \in \chi_{i_0} + W(\overline{U}, B)$; so $\chi(x) = \chi_{i_0}(x) + \psi(x)$ with $\psi \in W(\overline{U}, B)$ and then $\omega_G(x)(\chi) = \chi(x) \in B + B \subseteq A$. □

In this chapter we shall have a precise approach, by saying that a group G satisfies the Pontryagin-van Kampen duality theorem when ω_G is a topological isomorphism.

Lemma 7.22. *If the topological abelian groups G_i satisfy Pontryagin-van Kampen duality theorem for $i = 1, 2, \dots, n$, then also $G = \prod_{i=1}^n G_i$ satisfies Pontryagin-van Kampen duality theorem.*

Proof. Apply Lemma 7.12 twice to obtain an isomorphism $j : \prod_{i=1}^n \widehat{\widehat{G}}_i \to \widehat{\widehat{G}}$. It remains to verify that the product $\pi : G \to \prod_{i=1}^n \widehat{\widehat{G}}_i$ of the isomorphisms $\omega_{G_i} : G_i \to \prod_{i=1}^n \widehat{\widehat{G}}_i$ given by our hypothesis composed with the isomorphism j gives precisely ω_G. □

Consider two categories $\mathcal{A}$ and $\mathcal{B}$. A *covariant [contravariant] functor* $F : \mathcal{A} \to \mathcal{B}$ assigns to each object $A \in \mathcal{A}$ an object $FA \in \mathcal{B}$ and to each arrow $f : A \to A'$ in $\mathcal{A}$ an arrow $Ff : FA \to FA'$ $[Ff : FA' \to FA]$ such that $Fid_A = id_{FA}$ and $F(g \circ f) = Fg \circ Ff$ $[F(g \circ f) = Ff \circ Fg]$ for every arrow $f : A \to A'$ and $g : A' \to A''$ in $\mathcal{A}$.

Let $F, F' : \mathcal{A} \to \mathcal{B}$ be covariant functors. A *natural transformation* γ from F to F' assigns to each $A \in \mathcal{A}$ an arrow $\gamma_A : FA \to F'A$ such that for every arrow $f : A \to A'$ in $\mathcal{A}$ the following diagram is commutative

$$\begin{array}{ccc} FA & \xrightarrow{Ff} & FA' \\ {\scriptstyle \gamma_A}\downarrow & & \downarrow{\scriptstyle \gamma_{A'}} \\ F'A & \xrightarrow[F'f]{} & F'A' \end{array}$$

A *natural equivalence* is a natural transformation γ such that each γ_A is an isomorphism.

If $\mathcal{H}$ denote the category of all Hausdorff abelian topological groups, the *Pontryagin-van Kampen duality functor*, defined by

$$G \mapsto \widehat{G} \text{ and } f \mapsto \widehat{f}$$

for objects G and morphisms f of $\mathcal{H}$, is a contravariant functor $\widehat{\ }: \mathcal{H} \to \mathcal{H}$. Let $\mathcal{L}$ be the full subcategory of $\mathcal{H}$ having as objects all locally compact abelian groups. According to Proposition 7.2, the functor $\widehat{\ }$ sends $\mathcal{L}$ to itself, i.e., defines a functor $\widehat{\ }: \mathcal{L} \to \mathcal{L}$. The Pontryagin-van Kampen duality theorem states that ω is a natural equivalence from $id_{\mathcal{L}}$ to $\widehat{\widehat{\ }}: \mathcal{L} \to \mathcal{L}$. We start by proving that ω is a natural transformation.

Proposition 7.23. *ω is a natural transformation from $id_{\mathcal{L}}$ to $\widehat{\widehat{\ }}: \mathcal{L} \to \mathcal{L}$.*

Proof. By Proposition 7.21 ω_G is continuous for every $G \in \mathcal{L}$. Moreover for every continuous homomorphism $f : G \to H$ the following diagram commutes:

$$\begin{array}{ccc} G & \xrightarrow{f} & H \\ {\scriptstyle \omega_G}\downarrow & & \downarrow{\scriptstyle \omega_H} \\ \widehat{\widehat{G}} & \xrightarrow[\widehat{\widehat{f}}]{} & \widehat{\widehat{H}} \end{array}$$

In fact, if $x \in G$ and $\xi \in \widehat{H}$, then $\omega_H(f(x))(\xi) = \xi(f(x))$. On the other hand,

$$(\widehat{\widehat{f}}(\omega_G(x)))(\xi) = (\omega_G(x) \circ \widehat{f})(\xi) = \omega_G(x)(\widehat{f}(\xi)) = \omega_G(x)(\xi \circ f) = \xi(f(x)).$$

Hence $\omega_H(f(x)) = \widehat{\widehat{f}}(\omega_G(x))$ for every $x \in G$. □

Remark 7.24. Note that ω_G is a monomorphism if and only if $\widehat{G}$ separates the points of G. Moreover, $\omega_G(G)$ is a subgroup of $\widehat{\widehat{G}}$ that separates the points of $\widehat{G}$.

Now we can prove the Pontryagin-van Kampen duality theorem in the case when G is either compact or discrete.

Theorem 7.25. *If the abelian topological group G is either compact or discrete, then ω_G is a topological isomorphism.*

Proof. If G is discrete, then $\widehat{G}$ separates the points of G by Corollary 2.7 and if G is compact, then $\widehat{G}$ separates the points of G by the Peter-Weyl Theorem 6.4. Therefore ω_G is injective by Remark 7.24. If G is discrete, then $\widehat{G}$ is compact and $\omega_G(G) = \widehat{\widehat{G}}$ by Corollary 6.6. Since $\widehat{\widehat{G}}$ is discrete, ω_G is a topological isomorphism.

Let now G be compact. Then ω_G is open thanks to Theorem 4.9. Suppose that $\omega_G(G)$ is a proper subgroup of $\widehat{\widehat{G}}$. By the compactness of G, $\widehat{\widehat{G}}$ is compact, hence closed in $\widehat{\widehat{G}}$. By the Peter-Weyl Theorem 6.4 applied to $\widehat{\widehat{G}}/\omega_G(G)$, there exists $\xi \in \widehat{\widehat{G}} \setminus \{0\}$ such that $\xi(\omega_G(G)) = \{0\}$. Since $\widehat{G}$ is discrete, $\omega_{\widehat{G}}$ is a topological isomorphism and so there exists $\chi \in \widehat{G}$ such that $\omega_{\widehat{G}}(\chi) = \xi$. Thus for every $x \in G$ we have $0 = \xi(\omega_G(x)) = \omega_{\widehat{G}}(\chi)(\omega_G(x)) = \omega_G(x)(\chi) = \chi(x)$. It follows that $\chi \equiv 0$ and so that also $\xi \equiv 0$, a contradiction. □

Our next step is to prove the Pontryagin-van Kampen duality theorem when G is elementary locally compact abelian:

Theorem 7.26. *If G is an elementary locally compact abelian group, then ω_G is a topological isomorphism of G onto $\widehat{\widehat{G}}$.*

Proof. According to Lemma 7.22 and Theorem 7.25 it suffices to prove that $\omega_{\mathbb{R}}$ is a topologically isomorphism. Of course, by the fact that $\widehat{\mathbb{R}}$ is topologically isomorphic to $\mathbb{R}$, one concludes immediately that also $\mathbb{R}$ and $\widehat{\widehat{\mathbb{R}}}$ are topologically isomorphism. A more careful analysis of the dual $\widehat{\mathbb{R}}$ shows the crucial role of the ($\mathbb{Z}$-)bilnear map $\lambda : \mathbb{R} \times \mathbb{R} \to \mathbb{T}$ defined by $\lambda(x, y) = \chi_1(xy)$, where $\chi_1 : \mathbb{R} \to \mathbb{T}$ is the character determined by the canonical quotient map $\mathbb{R} \to \mathbb{T} = \mathbb{R}/\mathbb{Z}$. Indeed, for every $y \in \mathbb{R}$ the map $\chi_y : \mathbb{R} \to \mathbb{T}$ defined by $x \mapsto \lambda(x, y)$ is an element of $\widehat{\mathbb{R}}$. Hence the second copy $\{0\} \times \mathbb{R}$ of $\mathbb{R}$ in $\mathbb{R} \times \mathbb{R}$ can be identified with $\widehat{\mathbb{R}}$. On the other hand, every element $x \in \mathbb{R}$ gives a continuous character$\mathbb{R} \to \mathbb{T}$ defined by $y \mapsto \lambda(x, y)$, so can be considered as the element $\omega_{\mathbb{R}}(x)$ of $\widehat{\widehat{\mathbb{R}}}$. We have seen that every $\xi \in \widehat{\widehat{\mathbb{R}}}$ has this form. This means that $\omega_{\mathbb{R}}$ is surjective. Since continuity of $\omega_{\mathbb{R}}$, as well as local compactness of $\widehat{\widehat{\mathbb{R}}}$ are already established, $\omega_{\mathbb{R}}$ is a topological isomorphism by the open mapping theorem. □

For a subset X of G the *annihilator* of X in $\widehat{G}$ is $A_{\widehat{G}}(X) = \{\chi \in \widehat{G} : \chi(A) = \{0\}\}$ and for a subset Y of $\widehat{G}$ the *annihilator* of Y in G is $A_G(Y) = \{x \in G : \chi(x) = 0 \text{ for every } x \in Y\}$. When no confusion is possible we shall omit the subscripts $_{\widehat{G}}$ and $_G$.

The next lemma will help us in computing the dual of a subgroup and a quotient group.

Lemma 7.27. *Let G be a locally compact abelian group. If M is a subset of G, then $A_{\widehat{G}}(M)$ is a closed subgroup of $\widehat{G}$.*

Proof. It suffices to note that

$$A_{\widehat{G}}(M) = \bigcap_{x \in M} \{\chi \in \widehat{G} : \chi(x)\} = \bigcap \{\ker \omega_G(x) : x \in M\},$$

where each $\ker \omega(x)$ is a closed subgroup of $\widehat{G}$. □

Call a continuous homomorphism $f : G \to H$ of topological groups *proper* if $f : G \to f(G)$ is open, whenever $f(G)$ carries the topology inherited from H. In particular, a surjective continuous homomorphism is proper iff it is open.

A short sequence $0 \to G_1 \xrightarrow{f} G \xrightarrow{h} G_2 \to 0$ in $\mathcal{L}$, where f and h are continuous homomorphisms, is *exact* if f is injective, h is surjective and $\operatorname{im} f = \ker h$. It is *proper* if f and h are proper.

Lemma 7.28. *Let G be a locally compact abelian group, H a subgroup of G and $i : H \to G$ the canonical inclusion of H in G. Then*

(a) $\widehat{i} : \widehat{G} \to \widehat{H}$ is surjective if H is dense or open or compact;

(b) $\widehat{i}$ is injective if and only if H is dense in G;

(c) if H is closed and $\pi : G \to G/H$ is the canonical projection, then the sequence

$$0 \to \widehat{G/H} \xrightarrow{\widehat{\pi}} \widehat{G} \xrightarrow{\widehat{i}} \widehat{H}$$

is exact, $\widehat{\pi}$ is proper and $\operatorname{im} \widehat{\pi} = A_{\widehat{G}}(H)$. If H is open or compact, then $\widehat{i}$ is open and surjective.

Proof. (a) Note that $\widehat{i}$ is surjective if and only if for every $\chi \in \widehat{H}$ there exists $\xi \in \widehat{G}$ such that $\xi \restriction_H = \chi$. If H is compact apply Corollary 6.20. Otherwise Lemma 7.16 applies.

(b) If H is dense, then $\widehat{i}$ is injective by Lemma 7.16. Conversely, assume that $\overline{H}$ is a proper subgroup of G and let $q : G \to G/\overline{H}$ be the canonical projection. By Theorem 6.19 there exists $\chi \in \widehat{G/\overline{H}}$ not identically zero. Then $\xi = \chi \circ q \in \widehat{G}$ is non-zero and satisfies $\xi(H) = \{0\}$, i.e., $\widehat{i}(\xi) = 0$. This implies that $\widehat{i}$ is not injective.

(c) According to Lemma 7.16 $\widehat{\pi}$ is a monomorphism, since π is surjective. We have that $\widehat{i} \circ \widehat{\pi} = \widehat{\pi \circ i} = 0$. If $\xi \in \ker \widehat{i} = \{\chi \in \widehat{G} : \chi(H) = \{0\}\}$, then $\xi(H) = \{0\}$. So there exists $\xi_1 \in \widehat{G/H}$ such that $\xi = \xi_1 \circ \pi$ (i.e. $\xi = \widehat{\pi}(\xi_1)$) and we can conclude that $\ker \widehat{i} = \operatorname{im} \widehat{\pi}$. So the sequence is exact and $\operatorname{im} \widehat{\pi} = \ker \widehat{i} = A_{\widehat{G}}(H)$.

To show that $\widehat{\pi}$ is proper it suffices to apply Lemma 7.16.

If H is open or compact, (a) implies that $\widehat{i}$ is surjective. It remains to show that $\widehat{i}$ is open. If H is compact then $\widehat{H}$ is discrete by Example 7.1(2), so $\widehat{i}$ is obviously open. If H is open, let K be a compact neighborhood of 0 in G such that $K \subseteq H$. Then $W = W_{\widehat{G}}(K, \overline{\Lambda_4})$ is a compact neighborhood of 0 in $\widehat{G}$. Since $\widehat{i}$ is surjective, $V = \widehat{i}(W) = W_{\widehat{H}}(K, \overline{\Lambda_4})$ is a neighborhood of 0 in $\widehat{H}$. Now $M = \langle W \rangle$ and $M_1 = \langle V \rangle$ are open compactly generated subgroups respectively of $\widehat{G}$ and $\widehat{H}$, and $\widehat{i}(M) = M_1$. Since M is σ-compact by Lemma 4.12, we can apply Theorem 4.9 to the continuous surjective homomorphism $\widehat{i} \restriction_M : M \to M_1$ and so also $\widehat{i}$ is open. □

The lemma gives these immediate corollaries:

Corollary 7.29. *Let G be a locally compact abelian group and let H be a closed subgroup of G. Then $\widehat{G/H} \cong A_{\widehat{G}}(H)$. Moreover, if H is open or compact, then $\widehat{H} \cong \widehat{G}/A_{\widehat{G}}(H)$.*

The next corollary says that the duality functor preserves proper exactness for some sequences.

Corollary 7.30. *If the sequence $0 \to G_1 \xrightarrow{f} G \xrightarrow{h} G_2 \to 0$ in $\mathcal{L}$ is proper exact, with G_1 compact or G_2 discrete, then $0 \to \widehat{G_2} \xrightarrow{\hat{h}} \widehat{G} \xrightarrow{\hat{f}} \widehat{G_1} \to 0$ is proper exact with the same property.*

Now we can prove prove the Pontryagin-van Kampen duality theorem, namely ω is a natural equivalence from $id_{\mathcal{L}}$ to $\widehat{\widehat{\ }}\colon \mathcal{L} \to \mathcal{L}$.

Theorem 7.31. *If G is a locally compact abelian group, then ω_G is a topological isomorphism of G onto $\widehat{\widehat{G}}$.*

Proof. We know by Proposition 7.23 that ω is a natural transformation from $id_{\mathcal{L}}$ to $\widehat{\ }\colon \mathcal{L} \to \mathcal{L}$. Our plan is to chase the given locally compact abelian group G into an appropriately chosen proper exact sequence

$$0 \to G_1 \xrightarrow{f} G \xrightarrow{h} G_2 \to 0$$

in $\mathcal{L}$, with G_1 compact or G_2 discrete, such that G_1 and G_2 satisfy the duality theorem. By Corollary 7.30 the sequences

$$0 \to \widehat{G}_2 \xrightarrow{\hat{h}} \widehat{G} \xrightarrow{\hat{f}} \widehat{G}_2 \to 0 \quad \text{and} \quad 0 \to \widehat{\widehat{G}}_1 \xrightarrow{\hat{\hat{f}}} \widehat{\widehat{G}} \xrightarrow{\hat{\hat{h}}} \widehat{\widehat{G}}_2 \to 0$$

are proper exact. According to Proposition 7.23 the following diagram commutes:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & G_1 & \xrightarrow{f} & G & \xrightarrow{h} & G_2 & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \omega_{G_1}} & & \downarrow{\scriptstyle \omega_G} & & \downarrow{\scriptstyle \omega_{G_2}} & & \\ 0 & \longrightarrow & \widehat{\widehat{G}}_1 & \xrightarrow[\hat{\hat{f}}]{} & \widehat{\widehat{G}} & \xrightarrow[\hat{\hat{h}}]{} & \widehat{\widehat{G}}_2 & \longrightarrow & 0 \end{array}$$

According to Theorem 6.19, ω_{G_1}, ω_G, ω_{G_2} are injective. Moreover, ω_{G_1} and ω_{G_2} are surjective by our choice of G_1 and G_2. Then ω_G must be surjective too. (Indeed, if $x \in \ker \hat{\hat{h}}$, then there exists $y \in \omega_G(G)$ with $\hat{\hat{h}}(x) = \hat{\hat{h}}(y)$, because $\hat{\hat{h}}(\omega_G(G)) = \widehat{\widehat{G}}_2$. Now $y - x \in \ker \hat{\hat{h}} \subseteq \omega_G(G)$ and so $x \in y + \omega_G(G) = \omega_G(G)$.)

If G is locally compact abelian and compactly generated, by Proposition 6.18 we can choose G_1 compact and G_2 elementary locally compact abelian. Then G_1 and G_2 satisfy the duality theorem by Theorems 7.25 and 7.26, hence ω_G is surjective. Since ω_G is a continuous isomorphism and G is σ-compact, we conclude with Theorem 4.9 that ω_G is a topological isomorphism.

In the general case of locally compact abelian group G, we can take an open compactly generated subgroup G_1 of G. This will produce a proper exact sequence $0 \to G_1 \xrightarrow{f} G \xrightarrow{h} G_2 \to 0$ with G_1 compactly generated and $G_2 \cong G/G_1$ discrete. By the previous case ω_{G_1} is a topological isomorphism and ω_{G_2} is an isomorphism thanks to Theorem 7.25. Therefore ω_G is a continuous isomorphism.

Moreover $\omega_G \restriction_{f(G_1)}\colon f(G_1) \to \hat{\hat{f}}(\widehat{\widehat{G}}_1)$ is a topological isomorphism (as ω_{G_1}, $f : G_1 \to f(G_1)$ and $\hat{\hat{f}} : \widehat{\widehat{G}}_1 \to \hat{\hat{f}}(\widehat{\widehat{G}}_1)$ are topological isomorphisms) and $f(G_1)$ and $\hat{\hat{f}}(\widehat{\widehat{G}}_1)$ are open subgroups respectively of G and $\widehat{\widehat{G}}$. Thus ω_G is a topological isomorphism. □

Our last aim is to prove that the annihilators define an inclusion-inverting bijection between the family of all closed subgroups of a locally compact group G and the family of all closed subgroups of $\widehat{G}$. We use that fact that one can identify G and $\widehat{\widehat{G}}$ by the topological isomorphism ω_G. In more precise terms:

Exercise 7.32. *Let G be a locally compact abelian group and Y be a subset of $\widehat{G}$. Then $A_{\widehat{\widehat{G}}}(Y) = \omega_G(A_G(Y))$.*

Lemma 7.33. *Let G be a locally compact abelian group and H a closed subgroup of G. If $a \in G \setminus H$ then there exists $\chi \in A(H)$ such that $\chi(x) \neq 0$.*

Proof. Let $\rho : \widehat{G/H} \to A(H)$ be the topological isomorphism of Corollary 7.29. By Theorem 6.19 there exists $\psi \in \widehat{G/H}$ such that $\psi(a + H) \neq 0$. Therefore $\chi = \rho(\psi) \in A(H)$ and $\chi(a) = \rho(\psi)(a) = \psi(a + H) \neq 0$. □

Corollary 7.34. *If G is a locally compact abelian group and H a closed subgroup of G, then*

$$H = A_G(A_{\widehat{G}}(H))) = \omega_G^{-1}(A_{\widehat{\widehat{G}}}(A_{\widehat{G}}(H))).$$

Proof. The first equality follows immediately from the above lemma.
The last equality follows from the equality $H = A_G(A_{\widehat{G}}(H)))$ and Exercise 7.32. □

By Lemma 7.29 the equality $H = A_G(A_{\widehat{G}}(H)))$ holds if and only if H is a closed subgroup of G.

Proposition 7.35. *Let G be a locally compact abelian group and H a closed subgroup of G. Then $\widehat{H} \cong \widehat{G}/A(H)$.*

Proof. Since $H = \omega_G^{-1}(A_{\widehat{\widehat{G}}}(A_{\widehat{G}}(H)))$ by Lemma 7.34 we have a topological isomorphism ϕ from H to $\widehat{\widehat{G}/A(H)}$ given by $\phi(h)(\alpha + A(H)) = \alpha(h)$ for every $h \in H$ and $\alpha \in \widehat{G}$. This gives rise to another topological isomorphism $\widehat{\phi} : \widehat{\widehat{G}/A(H)} \to \widehat{H}$. By Pontryagin's duality theorem 7.31 $\omega_{\widehat{G}/A(H)}$ is a topological isomorphism from $\widehat{G}/A(H)$ to $\widehat{\widehat{G}/A(H)}$. The composition gives the desired isomorphism. □

Finally, let us resume for reader's benefit some of the most relevant points of Pontryagin-van Kampen duality theorem established so far:

Theorem 7.36. *Let G be a locally compact abelian group. Then $\widehat{G}$ is a locally compact abelian group and:*

(a) the correspondence $H \mapsto A_{\widehat{G}}(H)$, $N \mapsto A_G(N)$, where H is a closed subgroup of G and N is a closed subgroup of $\widehat{G}$, defines an order-inverting bijection between the family of all closed subgroups of G and the family of all closed subgroups of $\widehat{G}$;

(b) for every closed subgroup H of G the dual group $\widehat{H}$ is isomorphic to $\widehat{G}/A(H)$, while $A(H)$ is isomorphic to the dual $\widehat{G/H}$;

(c) $\omega_G : G \to \widehat{\widehat{G}}$ is a topological isomorphism;

(d) G is compact (resp., discrete) if and only if $\widehat{G}$ is discrete (resp., compact);

Proof. The first sentence is proved in Theorem 7.2. (a) is Corollary 7.34 while (b) is Proposition 7.35. (c) is Theorem 7.31. To prove (d) apply Theorem 7.31 and Lemma 7.1. □

Using the full power of the duality theorem one can prove the following structure theorem on compactly generated locally compact abelian groups.

Theorem 7.37. *Let G be a locally compact compactly generated abelian group. Prove that $G \cong \mathbb{R}^n \times \mathbb{Z}^m \times K$, where $n, m \in \mathbb{N}$ and K is a compact abelian group.*

Proof. According to Theorem 6.18 there exists a compact subgroup K of G such that G/K is an elementary locally compact abelian group. Taking a bigger compact subgroup one can get the quotient G/K to be of the form $\mathbb{R}^n \times \mathbb{Z}^m$ for some $n, m \in \mathbb{N}$. Now the dual group $\widehat{G}$ has an open subgroup $A(K) \cong \widehat{G/K} \cong \mathbb{R}^n \times \mathbb{T}^m$. Since this subgroup is divisible, one has $\widehat{G} \cong \mathbb{R}^n \times \mathbb{T}^m \times D$, where $D \cong \widehat{G}/A(K)$ is discrete and $D \cong \widehat{K}$. Taking duals gives $G \cong \widehat{\widehat{G}} \cong \mathbb{R}^n \times \mathbb{Z}^m \times K$. □

Making sharp use of the annihilators one can prove the structure theorem on locally compact abelian groups (see [67, 36] for a proof).

Theorem 7.38. *Let G be a locally compact abelian group. Then $G \cong \mathbb{R}^n \times G_0$, where G_0 is a closed subgroup of G containing an open compact subgroup K.*

As a corollary one can prove:

Corollary 7.39. *Every locally compact abelian group is isomorphic to a subgroup of a group of the form $\mathbb{R}^n \times D \times K$, where $n \in \mathbb{N}$, D is a discrete abelian group and K is a compact abelian group.*

Exercise 7.40. *Let G be a locally compact abelian group. Prove that for $\chi_1, \ldots, \chi_n \in \widehat{G}$ and $\delta > 0$ one has*

$$U_G(\chi_1, \ldots, \chi_n; \delta) = \omega_G^{-1}(W_{\widehat{\widehat{G}}}(\{\chi_1, \ldots, \chi_n\}, U),$$

where U is the neighborhood of 0 in $\mathbb{T} \cong \mathbb{S}$ determined by $|\mathrm{Arg} z| < \delta$.

Exercise 7.41. *Let G be a compact connected abelian group. Prove that $t(G)$ is dense in G iff $\widehat{G}$ is reduced. Deduce that every compact connected abelian group G has the form $G \cong G_1 \times \mathbb{Q}^\alpha$ for some cardinal α, where the compact subgroup G_1 coincides with the closure of the subgroup $t(G)$ of G.*

(Hint. Note first that $\widehat{G}$ is torsion-free. Deduce that $\widehat{G}$ is reduced iff $\bigcap_{n=1}^{\infty} n\widehat{G} = 0$. Show that this equality is equivalent to density of $t(G) = \bigcup_{n=1}^{\infty} G[n]$ in G. To prove the second assertion consider the torsion-free dual $\widehat{G}$ and its decomposition $\widehat{G} = d(\widehat{G}) \times R$, where R is a reduced subgroup of $\widehat{G}$. Now apply the first part and the isomorphism $G \cong \widehat{\widehat{G}}$.)

Exercise 7.42. *Give example of a reduced abelian group G such that $\bigcap_{n=1}^{\infty} nG \neq 0$.*

(Hint. Fix a prime number p and take an appropriate quotient of the group $\bigoplus_{n=1}^{\infty} \mathbb{Z}(p^n)$.

8 Appendix

8.1 Uniqueness of Pontryagin-van Kampen duality

For topological abelian groups G, H denote by $Chom(G, H)$ the group of all continuous homomorphisms $G \to H$ equipped with the compact-open topology. It was pointed out already by Pontryagin that the group $\mathbb{T}$ is the unique locally compact group L with the property $Chom(Chom(\mathbb{T}, L), L) \cong \mathbb{T}$ (note that this is much weaker than asking $Chom(-, L)$ to define a duality of $\mathcal{L}$). Much later Roeder [91] proved that Pontryagin-van Kampen dualityis the unique functorial duality of $\mathcal{L}$, i.e., the unique involutive contravariant endofunctor $\mathcal{L} \to \mathcal{L}$. Several years later Prodanov [85] rediscovered this result in the following much more general setting. Let R be a locally compact commutative ring and $\mathcal{L}_R$ be the category of locally compact topological R-modules. A *functorial duality* $^{\#} : \mathcal{L}_R \to \mathcal{L}_R$ is a contravariant functor such that $^{\#} \cdot ^{\#}$ is naturally equivalent to the identity of $\mathcal{L}_R$ and for each morphism $f : M \to N$ in $\mathcal{L}_R$ and $r \in R$ $(rf)^{\#} = rf^{\#}$ (where, as usual, rf is the morphism $M \to N$ defined by $(rf)(x) = rf(x)$). It is easy to see that the restriction of the Pontryagin-van Kampen duality functor on $\mathcal{L}_R$ is a functorial duality, since the Pontryagin-van Kampen dual $\widehat{M}$ of an $M \in \mathcal{L}_R$ has a natural structure of an R-module. So *there is always a functorial duality in* $\mathcal{L}_R$. This stimulated Prodanov to raise the question *how many* functorial dualities can carry $\mathcal{L}_R$ and extend this question to other well known dualities and adjunctions, such as Stone duality[13], the spectrum of a commutative rings [86], etc. at his Seminar on dualities (Sofia University, 1979/83). Uniqueness of the functorial duality was obtained by L. Stoyanov [93] in the case of a compact commutative ring R. In 1988 Gregorio [56] extended this result to the general case of compact (not necessarily commutative) ring R (here left and right R-modules should be distinguished, so that the dualities are no more *endo*functors). Later Gregorio jointly with Orsatti [58] offered another approach to this phenomenon.

Surprisingly the case of a discrete ring R turned out to be more complicated. For each functorial duality $^{\#} : \mathcal{L}_R \to \mathcal{L}_R$ the module $T = R^{\#}$ (the *torus* of the duality $^{\#}$) is compact and for every $X \in \mathcal{L}_R$ the module $\Delta_T(X) := Chom_R(X, T)$ of all continuous R-module homomorphisms $X \to T$, equipped with the compact-open topology, is algebraically isomorphic to $X^{\#}$. The duality $^{\#}$ is called *continuous* if for each X this isomorphism is also topological, otherwise $^{\#}$ is *discontinuous*. Clearly, continuous dualities are classified by their tori, which in turn can be classified by means of the Picard group $Pic(R)$ of R. In particular, the unique continuous functorial duality on $\mathcal{L}_R$ is the Pontryagin-van Kampen duality if and only if $Pic(R) = 0$ ([29, Theorem 5.17]). Prodanov [85] (see also [36, §3.4]) proved that every functorial duality on $\mathcal{L} = \mathcal{L}_{\mathbb{Z}}$ is continuous, which in view of $Pic(\mathbb{Z}) = 0$ gives another proof of Roeder's theorem of uniqueness. Continuous dualities were studied in the non-commutative context by Gregorio [57]. While the Picard group provides a good tool to measure the failure of uniqueness for continuous dualities, there is still no efficient way to capture it for discontinuous ones. The first example of a discontinuous duality was given in [29, Theorem 11.1]. Discontinuous dualities of $\mathcal{L}_{\mathbb{Q}}$ and its subcategories are discussed in [34]. It was conjectured by Prodanov that in case R is an algebraic number ring uniqueness of dualities is available if and only if R is a principal ideal domain. This conjecture was proved to be true for real algebraic number rings, but Prodanov's conjecture was shown to fail in case R is an order in an imaginary quadratic number field [25].

We will not touch other well-known dualities for module categories such as Morita duality (see [76]) or more general setting of dualities of (representable dualities, adjunctions rather than involutions, etc. [40], [41] and [83]).

[13] his conjecture that the Stone duality is the unique functorial duality between compact totally disconnected Hausdorff spaces and Boolean algebras was proved to be true by Dimov [39].

8.2 Non-abelian or non-locally compact groups

The Pontryagin-van Kampen duality theorem was extended to some non-locally compact abelian topological groups (e.g., infinite powers of the reals, the underlying additive groups of certain linear topological spaces, etc.). A characterization of the abelian topological groups admitting duality were proposed by Venkatamaran [95] and Kye [73], but they contained flaws. These gaps were removed in the recent paper of Hernández [63]. An important class of abelian groups (nuclear groups) were introduced and studied in the monograph [6] (see also [5]) in relation to the duality theorem. Further reference can be found also in [21, 51, 65]

We do not discuss here non-commutative versions of duality for locally compact groups. The difficulties arise already in the compact case – there is no appropriate (or at least, comfortable) structure on the set of irreducible unitary representations of a compact non-abelian group. The reader is referred to [67] for a historical panorama of this trend (Tanaka-Keĭn duality, etc.). In the locally compact case one should see the pioneering paper of H. Chu [22], as well as the monograph of Heyer [68] (see also [69]). In the recent survey of Galindo, Hernández, and Wu [53] the reader can find the last achievements in this field (see also [64]).

8.3 Relations to the topological theory of topological groups

The Pontryagin-van Kampen dual of a compact abelian group K carries a lot of useful information about the topology of H. For example,

- $w(K) = |\widehat{K}|$,
- $d(K) = \log|\widehat{K}| = \min\{\kappa : 2^\kappa \geq |\widehat{K}|\}$,
- K is connected iff $\widehat{K}$ is torsion-free,
- K is totally connected iff $\widehat{K}$ is torsion,
- $c(K) = A(t(\widehat{K}))$, where $t(\widehat{K})$ is the torsion subgroup of $\widehat{K}$,
- $\dim K = r_0(\widehat{K})$,
- $H^1(K, \mathbb{Z}) \cong \widehat{K}$ if K is connected (here $H^1(K, \mathbb{Z})$ denotes the first cohomology group),
- for two compact connected abelian groups K_1 and K_2 the following are equivalent: (i) K_1 and K_2 are homotopically equivalent as topological spaces; (ii) K_1 and K_2 are homeomorphic as topological spaces; (iii) $\widehat{K_1} \cong \widehat{K_2}$; (iv) $K_1 \cong K_2$ as topological groups.

The first equality can be generalized to $w(K) = w(\widehat{K})$ for all locally compact abelian groups K.

The Pontryagin-van Kampen duality can be used to easily build the *Bohr compactification* bG of a locally compact abelian group G (this is the reflection of G into the subcategory of compact abelian groups). In the case when G is discrete, bG is simply the completion of $G^\#$, the group G equipped with its Bohr topology. One can prove that $bG \cong \widehat{\widehat{G}_d}$, where $\widehat{G}_d$ denotes the group $\widehat{G}$ equipped with the discrete topology. For a comment on the non-abelian case see [28, 53].

Many nice properties of $\mathbb{Z}^\#$ can be found in Kunnen and Rudin [72]. For a fast growing sequence (a_n) in $\mathbb{Z}^\#$ the range is a closed discrete set of $\mathbb{Z}^\#$ (see [53] for further properties of the lacunary sets in $\mathbb{Z}^\#$), whereas for a polynomial function $n \mapsto a_n = P(n)$ the range has no isolated points [72, 44, Theorem 5.4]. Moreover, the range $P(\mathbb{Z})$ is closed when $P(x) = x^k$ is a monomial. For quadratic polynomials $P(x) = ax^2 + bx + c$, $(a, b, c, \in \mathbb{Z}, a \neq 0)$ the situation is already more complicated: the range $P(\mathbb{Z})$ is closed iff there is at most one prime that divides a, but does not divide b [72, 44, Theorem 5.6]. This leaves open the general question [26, Problem 954].

Problem 8.1. *Characterize the polynomials $P(x) \in \mathbb{Z}[x]$ such that $P(\mathbb{Z})$ is closed in $Z^\#$.*

8.4 Relations to dynamical systems

Among the known facts relating the dynamical systems with the topic of these notes let us mention just two.

- A compact group G admits ergodic translations $T_a(x) = ax$ iff G is monothetic. The ergodic rotations T_a of G are precisely those defined by a topological generator a of G.
- A continuous surjective endomorphism $T : K \to K$ of a compact abelian group is ergodic iff the dual $\widehat{T} : \widehat{K} \to \widehat{K}$ has no periodic points except $x = 0$.

The Pontryagin-van Kampen duality has an important impact also on the computation of the entropy of endomorphisms of (topological) abelian groups. Adler, Konheim, and McAndrew introduced the notion of topological entropy of continuous self-maps of compact topological spaces in the pioneering paper [1]. In 1975 Weiss [98] developed the definition of entropy for endomorphisms of abelian groups briefly sketched in [1]. He

called it "algebraic entropy", and gave detailed proofs of its basic properties. His main result was that the topological entropy of a continuous endomorphism ϕ of a profinite abelian group coincides with the algebraic entropy of the adjoint map $\widehat{\phi}$ of ϕ (note that pro-finite abelian groups are precisely the Pontryagin duals of the torsion abelian groups).

In 1979 Peters [82] extended Weiss's definition of entropy for automorphisms of a discrete abelian group G. He generalized Weiss's main result to metrizable compact abelian groups, relating again the opological entropy of a continuous automorphism of such a group G to the entropy of the adjoint automorphism of the dual group $\widehat{G}$. The definition of entropy of automorphisms given by Peters is easily adaptable to endomorphisms of Abelian groups, but it remains unclear whether his theorem can be extended to the computation of the topological entropy of a continuous endomorphism of compact abelian groups.

Elective Paper

MATP 3.4

Block - II

Marks : 50 (SSE : 40; IA : 10)

Measure Theory (Pure Stream)

Unit 9

Introduction

The Riemann integral, dealt with in calculus courses, is well suited for computations but less suited for dealing with limit processes. In this course we will introduce the so called Lebesgue integral, which keeps the advantages of the Riemann integral and eliminates its drawbacks. At the same time we will develop a general measure theory which serves as the basis of contemporary analysis and probability.

In this introductory chapter we set forth some basic concepts of measure theory, which will open for abstract Lebesgue integration.

1.1. σ-Algebras and Measures

Throughout this course

$\mathbf{N} = \{0, 1, 2, ...\}$ (the set of natural numbers)
$\mathbf{Z} = \{..., -2, -1, 0, 1, , 2, ...\}$ (the set of integers)
$\mathbf{Q}$ = the set of rational numbers
$\mathbf{R}$ = the set of real numbers
$\mathbf{C}$ = the set of complex numbers.

If $A \subseteq \mathbf{R}$, A_+ is the set of all strictly positive elements in A.

If f is a function from a set A into a set B, this means that to every $x \in A$ there corresponds a point $f(x) \in B$ and we write $f : A \to B$. A function is often called a map or a mapping. The function f is injective if

$$(x \neq y) \Rightarrow (f(x) \neq f(y))$$

and surjective if to each $y \in B$, there exists an $x \in A$ such that $f(x) = y$. An injective and surjective function is said to be bijective.

A set A is finite if either A is empty or there exist an $n \in \mathbf{N}_+$ and a bijection $f : \{1, ..., n\} \to A$. The empty set is denoted by ϕ. A set A is said to be denumerable if there exists a bijection $f : \mathbf{N}_+ \to A$. A subset of a denumerable set is said to be at most denumerable.

Let X be a set. For any $A \subseteq X$, the indicator function χ_A of A relative to X is defined by the equation

$$\chi_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \in A^c. \end{cases}$$

The indicator function χ_A is sometimes written 1_A. We have the following relations:

$$\chi_{A^c} = 1 - \chi_A$$

$$\chi_{A \cap B} = \min(\chi_A, \chi_B) = \chi_A \chi_B$$

and

$$\chi_{A \cup B} = \max(\chi_A, \chi_B) = \chi_A + \chi_B - \chi_A \chi_B.$$

Definition 1.1.1. Let X be a set.

a) A collection $\mathcal{A}$ of subsets of X is said to be an algebra in X if $\mathcal{A}$ has the following properties:

(i) $X \in \mathcal{A}$.
(ii) $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$, where A^c is the complement of A relative to X.
(iii) If $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$.

(b) A collection $\mathcal{M}$ of subsets of X is said to be a σ-algebra in X if $\mathcal{M}$ is an algebra with the following property:

If $A_n \in \mathcal{M}$ for all $n \in \mathbf{N}_+$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{M}$.

If $\mathcal{M}$ is a σ-algebra in X, $(X, \mathcal{M})$ is called a measurable space and the members of $\mathcal{M}$ are called measurable sets. The so called power set $\mathcal{P}(X)$, that is the collection of all subsets of X, is a σ-algebra in X. It is simple to prove that the intersection of any family of σ-algebras in X is a σ-algebra. It follows that if $\mathcal{E}$ is any subset of $\mathcal{P}(X)$, there is a unique smallest σ-algebra $\sigma(\mathcal{E})$ containing $\mathcal{E}$, namely the intersection of all σ-algebras containing $\mathcal{E}$.

The σ-algebra $\sigma(\mathcal{E})$ is called the σ-algebra generated by $\mathcal{E}$. The σ-algebra generated by all open intervals in $\mathbf{R}$ is denoted by $\mathcal{R}$. It is readily seen that the σ-algebra $\mathcal{R}$ contains every subinterval of $\mathbf{R}$. Before we proceed, recall that a subset E of $\mathbf{R}$ is open if to each $x \in E$ there exists an open subinterval of $\mathbf{R}$ contained in E and containing x; the complement of an open set is said to be closed. We claim that $\mathcal{R}$ contains every open subset U of $\mathbf{R}$. To see this suppose $x \in U$ and let $x \in]a, b[\subseteq U$, where $-\infty < a < b < \infty$. Now pick $r, s \in \mathbf{Q}$ such that $a < r < x < s < b$. Then $x \in]r, s[\subseteq U$ and it follows that U is the union of all bounded open intervals with rational boundary points contained in U. Since this family of intervals is at most denumberable we conclude that $U \in \mathcal{R}$. In addition, any closed set belongs to $\mathcal{R}$ since its complements is open. It is by no means simple to grasp the definition of $\mathcal{R}$ at this stage but the reader will successively see that the σ-algebra $\mathcal{R}$ has very nice properties. At the very end of Section 1.3, using the so called Axiom of Choice, we will exemplify a subset of the real line which does not belong to $\mathcal{R}$. In fact, an example of this type can be constructed without the Axiom of Choice (see Dudley's book [D]).

In measure theory, inevitably one encounters ∞. For example the real line has infinite length. Below $[0, \infty] = [0, \infty[\cup \{\infty\}$. The inequalities $x \leq y$ and $x < y$ have their usual meanings if $x, y \in [0, \infty[$. Furthermore, $x \leq \infty$ if $x \in [0, \infty]$ and $x < \infty$ if $x \in [0, \infty[$. We define $x + \infty = \infty + x = \infty$ if $x, y \in [0, \infty]$, and

$$x \cdot \infty = \infty \cdot x = \begin{cases} 0 \text{ if } x = 0 \\ \infty \text{ if } 0 < x \leq \infty. \end{cases}$$

Sums and multiplications of real numbers are defined in the usual way.

If $A_n \subseteq X$, $n \in \mathbf{N}_+$, and $A_k \cap A_n = \phi$ if $k \neq n$, the sequence $(A_n)_{n \in \mathbf{N}_+}$ is called a disjoint denumerable collection. If $(X, \mathcal{M})$ is a measurable space, the collection is called a denumerable measurable partition of A if $A = \cup_{n=1}^{\infty} A_n$ and $A_n \in \mathcal{M}$ for every $n \in \mathbf{N}_+$. Some authors call a denumerable collection of sets a countable collection of sets.

Definition 1.1.2. (a) Let $\mathcal{A}$ be an algebra of subsets of X. A function $\mu : \mathcal{A} \to [0, \infty]$ is called a content if

(i) $\mu(\phi) = 0$
(ii) $\mu(A \cup B) = \mu(A) + \mu(B)$ if $A, B \in \mathcal{A}$ and $A \cap B = \phi$.

(b) If $(X, \mathcal{M})$ is a measurable space a content μ defined on the σ-algebra $\mathcal{M}$ is called a positive measure if it has the following property:

For any disjoint denumerable collection $(A_n)_{n \in \mathbf{N}_+}$ of members of $\mathcal{M}$

$$\mu(\cup_{n=1}^{\infty} A_n) = \Sigma_{n=1}^{\infty} \mu(A_n).$$

If $(X, \mathcal{M})$ is a measurable space and the function $\mu : \mathcal{M} \to [0, \infty]$ is a positive measure, $(X, \mathcal{M}, \mu)$ is called a positive measure space. The quantity $\mu(A)$ is called the μ-measure of A or simply the measure of A if there is no ambiguity. Here $(X, \mathcal{M}, \mu)$ is called a probability space if $\mu(X) = 1$, a finite positive measure space if $\mu(X) < \infty$, and a σ-finite positive measure space if X is a denumerable union of measurable sets with finite μ-measure. The measure μ is called a probability measure, finite measure, and σ-finite measure, if $(X, \mathcal{M}, \mu)$ is a probability space, a finite positive measure space, and a σ-finite positive measure space, respectively. A probability space is often denoted by $(\Omega, \mathcal{F}, P)$. A member A of $\mathcal{F}$ is called an event.

As soon as we have a positive measure space $(X, \mathcal{M}, \mu)$, it turns out to be a fairly simple task to define a so called μ-integral

$$\int_X f(x) d\mu(x)$$

as will be seen in Chapter 2.

The class of all finite unions of subintervals of $\mathbf{R}$ is an algebra which is denoted by $\mathcal{R}_0$. If $A \in \mathcal{R}_0$ we denote by $l(A)$ the Riemann integral

$$\int_{-\infty}^{\infty} \chi_A(x)dx$$

and it follows from courses in calculus that the function $l : \mathcal{R}_0 \to [0, \infty]$ is a content. The algebra $\mathcal{R}_0$ is called the Riemann algebra and l the Riemann content. If I is a subinterval of $\mathbf{R}$, $l(I)$ is called the length of I. Below we follow the convention that the empty set is an interval.

If $A \in \mathcal{P}(X)$, $c_X(A)$ equals the number of elements in A, when A is a finite set, and $c_X(A) = \infty$ otherwise. Clearly, c_X is a positive measure. The measure c_X is called the counting measure on X.

Given $a \in X$, the probability measure δ_a defined by the equation $\delta_a(A) = \chi_A(a)$, if $A \in \mathcal{P}(X)$, is called the Dirac measure at the point a. Sometimes we write $\delta_a = \delta_{X,a}$ to emphasize the set X.

If μ and ν are positive measures defined on the same σ-algebra $\mathcal{M}$, the sum $\mu + \nu$ is a positive measure on $\mathcal{M}$. More generally, $\alpha\mu + \beta\nu$ is a positive measure for all real α, $\beta \geq 0$. Furthermore, if $E \in \mathcal{M}$, the function $\lambda(A) = \mu(A \cap E)$, $A \in \mathcal{M}$, is a positive measure. Below this measure λ will be denoted by μ^E and we say that μ^E is concentrated on E. If $E \in \mathcal{M}$, the class $\mathcal{M}_E = \{A \in \mathcal{M};\ A \subseteq E\}$ is a σ-algebra of subsets of E and the function $\theta(A) = \mu(A)$, $A \in \mathcal{M}_E$, is a positive measure. Below this measure θ will be denoted by $\mu_{|E}$ and is called the restriction of μ to $\mathcal{M}_E$.

Let $I_1, ..., I_n$ be subintervals of the real line. The set

$$I_1 \times ... \times I_n = \{(x_1, ..., x_n) \in \mathbf{R}^n;\ x_k \in I_k,\ k = 1, ..., n\}$$

is called an n-cell in $\mathbf{R}^n$; its volume $\mathrm{vol}(I_1 \times ... \times I_n)$ is, by definition, equal to

$$\mathrm{vol}(I_1 \times ... \times I_n) = \Pi_{k=1}^n l(I_k).$$

If $I_1, ..., I_n$ are open subintervals of the real line, the n-cell $I_1 \times ... \times I_n$ is called an open n-cell. The σ-algebra generated by all open n-cells in $\mathbf{R}^n$ is denoted by $\mathcal{R}_n$. In particular, $\mathcal{R}_1 = \mathcal{R}$. A basic theorem in measure theory states that there exists a unique positive measure v_n defined on $\mathcal{R}_n$ such that the measure of any n-cell is equal to its volume. The measure v_n is called the volume measure on $\mathcal{R}_n$ or the volume measure on $\mathbf{R}^n$. Clearly, v_n is σ-finite. The measure v_2 is called the area measure on $\mathbf{R}^2$ and v_1 the linear measure on $\mathbf{R}$.

Theorem 1.1.1. *The volume measure on* $\mathbf{R}^n$ *exists.*

Theorem 1.1.1 will be proved in Section 1.5 in the special case $n = 1$. The general case then follows from the existence of product measures in Section 3.4. An alternative proof of Theorem 1.1.1 will be given in Section 3.2. As soon as the existence of volume measure is established a variety of interesting measures can be introduced.

Next we prove some results of general interest for positive measures.

Theorem 1.1.2. *Let* $\mathcal{A}$ *be an algebra of subsets of* X *and* μ *a content defined on* $\mathcal{A}$*. Then,*

(a) μ *is finitely additive, that is*

$$\mu(A_1 \cup ... \cup A_n) = \mu(A_1) + ... + \mu(A_n)$$

if $A_1, ..., A_n$ *are pairwise disjoint members of* $\mathcal{A}$*.*

(b) *if* $A, B \in \mathcal{A}$,

$$\mu(A) = \mu(A \setminus B) + \mu(A \cap B).$$

Moreover, if $\mu(A \cap B) < \infty$, *then*

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$$

(c) $A \subseteq B$ *implies* $\mu(A) \leq \mu(B)$ *if* $A, B \in \mathcal{A}$*.*

(d) μ *finitely sub-additive, that is*

$$\mu(A_1 \cup ... \cup A_n) \leq \mu(A_1) + ... + \mu(A_n)$$

if $A_1, ..., A_n$ *are members of* $\mathcal{A}$*.*

If $(X, \mathcal{M}, \mu)$ *is a positive measure space*

(e) $\mu(A_n) \to \mu(A)$ *if* $A = \cup_{n\in\mathbf{N}_+} A_n$, $A_n \in \mathcal{M}$, *and*

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq ... \ .$$

(f) $\mu(A_n) \to \mu(A)$ *if* $A = \cap_{n\in\mathbf{N}_+} A_n$, $A_n \in \mathcal{M}$,

$$A_1 \supseteq A_2 \supseteq A_3 \supseteq ...$$

and $\mu(A_1) < \infty$.
(g) μ *is sub-additive, that is for any denumerable collection* $(A_n)_{n\in\mathbf{N}_+}$ *of members of* $\mathcal{M}$,

$$\mu(\cup_{n=1}^{\infty} A_n) \leq \Sigma_{n=1}^{\infty} \mu(A_n).$$

PROOF (a) If $A_1, ..., A_n$ are pairwise disjoint members of $\mathcal{A}$,

$$\mu(\cup_{k=1}^{n} A_k) = \mu(A_1 \cup (\cup_{k=2}^{n} A_k))$$
$$= \mu(A_1) + \mu(\cup_{k=2}^{n} A_k)$$

and, by induction, we conclude that μ is finitely additive.

(b) Recall that

$$A \setminus B = A \cap B^c.$$

Now $A = (A \setminus B) \cup (A \cap B)$ and we get

$$\mu(A) = \mu(A \setminus B) + \mu(A \cap B).$$

Moreover, since $A \cup B = (A \setminus B) \cup B$,

$$\mu(A \cup B) = \mu(A \setminus B) + \mu(B)$$

and, if $\mu(A \cap B) < \infty$, we have

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B).$$

(c) Part (b) yields $\mu(B) = \mu(B \setminus A) + \mu(A \cap B) = \mu(B \setminus A) + \mu(A)$, where the last member does not fall below $\mu(A)$.

(d) If $(A_i)_{i=1}^n$is a sequence of members of $\mathcal{A}$ define the so called disjunction $(B_k)_{k=1}^n$ of the sequence $(A_i)_{i=1}^n$ as

$$B_1 = A_1 \text{ and } B_k = A_k \setminus \cup_{i=1}^{k-1} A_i \text{ for } 2 \leq k \leq n.$$

Then $B_k \subseteq A_k$, $\cup_{i=1}^k A_i = \cup_{i=1}^k B_i$, $k = 1, .., n$, and $B_i \cap B_j = \phi$ if $i \neq j$. Hence, by Parts (a) and (c),

$$\mu(\cup_{k=1}^n A_k) = \Sigma_{k=1}^n \mu(B_k) \leq \Sigma_{k=1}^n \mu(A_k).$$

(e) Set $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n \geq 2$. Then $A_n = B_1 \cup .... \cup B_n$, $B_i \cap B_j = \phi$ if $i \neq j$ and $A = \cup_{k=1}^\infty B_k$. Hence

$$\mu(A_n) = \Sigma_{k=1}^n \mu(B_k)$$

and

$$\mu(A) = \Sigma_{k=1}^\infty \mu(B_k).$$

Now e) follows, by the definition of the sum of an infinite series.

(f) Put $C_n = A_1 \setminus A_n$, $n \geq 1$. Then $C_1 \subseteq C_2 \subseteq C_3 \subseteq ...$,

$$A_1 \setminus A = \cup_{n=1}^\infty C_n$$

and $\mu(A) \leq \mu(A_n) \leq \mu(A_1) < \infty$. Thus

$$\mu(C_n) = \mu(A_1) - \mu(A_n)$$

and Part (e) shows that

$$\mu(A_1) - \mu(A) = \mu(A_1 \setminus A) = \lim_{n\to\infty} \mu(C_n) = \mu(A_1) - \lim_{n\to\infty} \mu(A_n).$$

This proves (f).

(g) The result follows from Parts d) and e).
This completes the proof of Theorem 1.1.2.

The hypothesis "$\mu(A_1) < \infty$ " in Theorem 1.1.2 (f) is not superfluous. If $c_{\mathbf{N}_+}$ is the counting measure on $\mathbf{N}_+$ and $A_n = \{n, n+1, ...\}$, then $c_{\mathbf{N}_+}(A_n) = \infty$ for all n but $A_1 \supseteq A_2 \supseteq ....$ and $c_{\mathbf{N}_+}(\cap_{n=1}^{\infty} A_n) = 0$ since $\cap_{n=1}^{\infty} A_n = \phi$.

If $A, B \subseteq X$, the symmetric difference $A\Delta B$ is defined by the equation

$$A\Delta B =_{def} (A \setminus B) \cup (B \setminus A).$$

Note that

$$\chi_{A\Delta B} =\mid \chi_A - \chi_B \mid .$$

Moreover, we have

$$A\Delta B = A^c \Delta B^c$$

and

$$(\cup_{i=1}^{\infty} A_i)\Delta(\cup_{i=1}^{\infty} B_i) \subseteq \cup_{i=1}^{\infty}(A_i \Delta B_i).$$

Unit 10

1.2. Measure Determining Classes

Suppose μ and ν are probability measures defined on the same σ-algebra $\mathcal{M}$, which is generated by a class $\mathcal{E}$. If μ and ν agree on $\mathcal{E}$, is it then true that μ and ν agree on $\mathcal{M}$? The answer is in general no. To show this, let

$$X = \{1, 2, 3, 4\}$$

and

$$\mathcal{E} = \{\{1, 2\}, \{1, 3\}\}.$$

Then $\sigma(\mathcal{E}) = \mathcal{P}(X)$. If $\mu = \frac{1}{4}c_X$ and

$$\nu = \frac{1}{6}\delta_{X,1} + \frac{1}{3}\delta_{X,2} + \frac{1}{3}\delta_{X,3} + \frac{1}{6}\delta_{X,4}$$

then $\mu = \nu$ on $\mathcal{E}$ and $\mu \neq \nu$.

In this section we will prove a basic result on measure determining classes for σ-finite measures. In this context we will introduce so called π-systems and σ-additive classes, which will also be of great value later in connection with the construction of so called product measures in Chapter 3.

Definition 1.2.1. A class $\mathcal{G}$ of subsets of X is a π-system if $A \cap B \in \mathcal{G}$ for all $A, B \in \mathcal{G}$.

The class of all open n-cells in $\mathbf{R}^n$ is a π-system.

Definition 1.2.2. A class $\mathcal{D}$ of subsets of X is called a σ-additive class if the following properties hold:

(a) $X \in \mathcal{D}$.

(b) If $A, B \in \mathcal{D}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{D}$.

(c) If $(A_n)_{n\in\mathbf{N}_+}$ is a disjoint denumerable collection of members of the class $\mathcal{D}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{D}$.

Theorem 1.2.1. *If a σ-additive class $\mathcal{M}$ is a π-system, then $\mathcal{M}$ is a σ-algebra.*

PROOF. If $A \in \mathcal{M}$, then $A^c = X \setminus A \in \mathcal{M}$ since $X \in \mathcal{M}$ and $\mathcal{M}$ is a σ-additive class. Moreover, if $(A_n)_{n\in\mathbf{N}_+}$ is a denumerable collection of members of $\mathcal{M}$,

$$A_1 \cup ... \cup A_n = (A_1^c \cap ... \cap A_n^c)^c \in \mathcal{M}$$

for each n, since $\mathcal{M}$ is a σ-additive class and a π-system. Let $(B_n)_{n=1}^{\infty}$ be the disjunction of $(A_n)_{n=1}^{\infty}$. Then $(B_n)_{n\in\mathbf{N}_+}$ is a disjoint denumerable collection of members of $\mathcal{M}$ and Definition 1.2.2(c) implies that $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} B_n \in \mathcal{M}$.

Theorem 1.2.2. *Let $\mathcal{G}$ be a π-system and $\mathcal{D}$ a σ-additive class such that $\mathcal{G} \subseteq \mathcal{D}$. Then $\sigma(\mathcal{G}) \subseteq \mathcal{D}$.*

PROOF. Let $\mathcal{M}$ be the intersection of all σ-additive classes containing $\mathcal{G}$. The class $\mathcal{M}$ is a σ-additive class and $\mathcal{G} \subseteq \mathcal{M} \subseteq \mathcal{D}$. In view of Theorem 1.2.1 $\mathcal{M}$ is a σ-algebra, if $\mathcal{M}$ is a π-system and in that case $\sigma(\mathcal{G}) \subseteq \mathcal{M}$. Thus the theorem follows if we show that $\mathcal{M}$ is a π-system.

Given $C \subseteq X$, denote by $\mathcal{D}_C$ be the class of all $D \subseteq X$ such that $D \cap C \in \mathcal{M}$.

CLAIM 1. If $C \in \mathcal{M}$, then $\mathcal{D}_C$ is a σ-additive class.

PROOF OF CLAIM 1. First $X \in \mathcal{D}_C$ since $X \cap C = C \in \mathcal{M}$. Moreover, if $A, B \in \mathcal{D}_C$ and $A \subseteq B$, then $A \cap C, B \cap C \in \mathcal{M}$ and

$$(B \setminus A) \cap C = (B \cap C) \setminus (A \cap C) \in \mathcal{M}.$$

Accordingly from this, $B \setminus A \in \mathcal{D}_C$. Finally, if $(A_n)_{n \in \mathbf{N}_+}$ is a disjoint denumerable collection of members of $\mathcal{D}_C$, then $(A_n \cap C)_{n \in \mathbf{N}_+}$ is disjoint denumerable collection of members of $\mathcal{M}$ and

$$(\cup_{n \in \mathbf{N}_+} A_n) \cap C = \cup_{n \in \mathbf{N}_+} (A_n \cap C) \in \mathcal{M}.$$

Thus $\cup_{n \in \mathbf{N}_+} A_n \in \mathcal{D}_C$.

CLAIM 2. If $A \in \mathcal{G}$, then $\mathcal{M} \subseteq \mathcal{D}_A$.

PROOF OF CLAIM 2. If $B \in \mathcal{G}$, $A \cap B \in \mathcal{G} \subseteq \mathcal{M}$. Thus $B \in \mathcal{D}_A$. We have proved that $\mathcal{G} \subseteq \mathcal{D}_A$ and remembering that $\mathcal{M}$ is the intersection of all σ-additive classes containing $\mathcal{G}$ Claim 2 follows since $\mathcal{D}_A$ is a σ-additive class.

To complete the proof of Theorem 1.2.2, observe that $B \in \mathcal{D}_A$ if and only if $A \in \mathcal{D}_B$. By Claim 2, if $A \in \mathcal{G}$ and $B \in \mathcal{M}$, then $B \in \mathcal{D}_A$ that is $A \in \mathcal{D}_B$. Thus $\mathcal{G} \subseteq \mathcal{D}_B$ if $B \in \mathcal{M}$. Now the definition of $\mathcal{M}$ implies that $\mathcal{M} \subseteq \mathcal{D}_B$ if $B \in \mathcal{M}$. The proof is almost finished. In fact, if $A, B \in \mathcal{M}$ then $A \in \mathcal{D}_B$ that is $A \cap B \in \mathcal{M}$. Theorem 1.2.2 now follows from Theorem 1.2.1.

Theorem 1.2.3. *Let μ and ν be positive measures on $\mathcal{M} = \sigma(\mathcal{G})$, where $\mathcal{G}$ is a π-system, and suppose $\mu(A) = \nu(A)$ for every $A \in \mathcal{G}$.*

(a) *If μ and ν are probability measures, then $\mu = \nu$.*

(b) *Suppose there exist $E_n \in \mathcal{G}$, $n \in \mathbf{N}_+$, such that $X = \cup_{n=1}^{\infty} E_n$,*

$E_1 \subseteq E_2 \subseteq ...$, *and*

$$\mu(E_n) = \nu(E_n) < \infty, \text{ all } n \in \mathbf{N}_+.$$

Then $\mu = \nu$.

PROOF. (a) Let

$$\mathcal{D} = \{A \in \mathcal{M};\ \ \mu(A) = \ \nu(A)\}.$$

It is immediate that $\mathcal{D}$ is a σ-additive class and Theorem 1.2.2 implies that $\mathcal{M} = \sigma(\mathcal{G}) \subseteq \mathcal{D}$ since $\mathcal{G} \subseteq \mathcal{D}$ and $\mathcal{G}$ is a π-system.

(b) If $\mu(E_n) = \nu(E_n) = 0$ for all all $n \in \mathbf{N}_+$, then

$$\mu(X) = \lim_{n\to\infty} \mu(E_n) = 0$$

and, in a similar way, $\nu(X) = 0$. Thus $\mu = \nu$. If $\mu(E_n) = \nu(E_n) > 0$, set

$$\mu_n(A) = \frac{1}{\mu(E_n)}\mu(A \cap E_n) \text{ and } \nu_n(A) = \frac{1}{\nu(E_n)}\nu(A \cap E_n)$$

for each $A \in \mathcal{M}$. By Part (a) $\mu_n = \nu_n$ and we get

$$\mu(A \cap E_n) = \nu(A \cap E_n)$$

for each $A \in \mathcal{M}$. Theorem 1.1.2(e) now proves that $\mu = \nu$.

Theorem 1.2.3 implies that there is at most one positive measure defined on $\mathcal{R}_n$ such that the measure of any open n-cell in $\mathbf{R}^n$ equals its volume.

Next suppose $f : X \to Y$ and let $A \subseteq X$ and $B \subseteq Y$. The image of A and the inverse image of B are

$$f(A) = \{y;\ y = f(x) \text{ for some } x \in A\}$$

and

$$f^{-1}(B) = \{x;\ f(x) \in B\}$$

respectively. Note that

$$f^{-1}(Y) = X$$

and

$$f^{-1}(Y \setminus B) = X \setminus f^{-1}(B).$$

Moreover, if $(A_i)_{i\in I}$ is a collection of subsets of X and $(B_i)_{i\in I}$ is a collection of subsets of Y

$$f(\cup_{i\in I} A_i) = \cup_{i\in I} f(A_i)$$

and

$$f^{-1}(\cup_{i\in I} B_i) = \cup_{i\in I} f^{-1}(B_i).$$

Given a class $\mathcal{E}$ of subsets of Y, set

$$f^{-1}(\mathcal{E}) = \left\{ f^{-1}(B);\ B \in \mathcal{E} \right\}.$$

If $(Y, \mathcal{N})$ is a measurable space, it follows that the class $f^{-1}(\mathcal{N})$ is a σ-algebra in X. If $(X, \mathcal{M})$ is a measurable space

$$\left\{ B \in \mathcal{P}(Y);\ f^{-1}(B) \in \mathcal{M} \right\}$$

is a σ-algebra in Y. Thus, given a class $\mathcal{E}$ of subsets of Y,

$$\sigma(f^{-1}(\mathcal{E})) = f^{-1}(\sigma(\mathcal{E})).$$

Definition 1.2.3. Let $(X, \mathcal{M})$ and $(Y, \mathcal{N})$ be measurable spaces. The function $f : X \to Y$ is said to be $(\mathcal{M}, \mathcal{N})$-measurable if $f^{-1}(\mathcal{N}) \subseteq \mathcal{M}$. If we say that $f : (X, \mathcal{M}) \to (Y, \mathcal{N})$ is measurable this means that $f : X \to Y$ is an $(\mathcal{M}, \mathcal{N})$-measurable function.

Theorem 1.2.4. *Let $(X, \mathcal{M})$ and $(Y, \mathcal{N})$ be measurable spaces and suppose $\mathcal{E}$ generates $\mathcal{N}$. The function $f : X \to Y$ is $(\mathcal{M}, \mathcal{N})$-measurable if*

$$f^{-1}(\mathcal{E}) \subseteq \mathcal{M}.$$

PROOF. The assumptions yield

$$\sigma(f^{-1}(\mathcal{E})) \subseteq \mathcal{M}.$$

Since

$$\sigma(f^{-1}(\mathcal{E})) = f^{-1}(\sigma(\mathcal{E})) = f^{-1}(\mathcal{N})$$

we are done.

Corollary 1.2.1. *A function $f : X \to \mathbf{R}$ is $(\mathcal{M}, \mathcal{R})$-measurable if and only if the set $f^{-1}(]\alpha, \infty[) \in \mathcal{M}$ for all $\alpha \in \mathbf{R}$.*

If $f : X \to Y$ is $(\mathcal{M}, \mathcal{N})$-measurable and μ is a positive measure on $\mathcal{M}$, the equation

$$\nu(B) = \mu(f^{-1}(B)),\ B \in \mathcal{N}$$

defines a positive measure ν on $\mathcal{N}$. We will write $\nu = \mu f^{-1}$, $\nu = f(\mu)$ or $\nu = \mu_f$. The measure ν is called the image measure of μ under f and f is said to transport μ to ν. Two $(\mathcal{M}, \mathcal{N})$-measurable functions $f : X \to Y$ and $g : X \to Y$ are said to be μ-equimeasurable if $f(\mu) = g(\mu)$.

As an example, let $a \in \mathbf{R}^n$ and define $f(x) = x + a$ if $x \in \mathbf{R}^n$. If $B \subseteq \mathbf{R}^n$,

$$f^{-1}(B) = \{x;\ x + a \in B\} = B - a.$$

Thus $f^{-1}(B)$ is an open n-cell if B is, and Theorem 1.2.4 proves that f is $(\mathcal{R}_n, \mathcal{R}_n)$-measurable. Now, granted the existence of volume measure v_n, for every $B \in \mathcal{R}_n$ define

$$\mu(B) = f(v_n)(B) = v_n(B - a).$$

Then $\mu(B) = v_n(B)$ if B is an open n-cell and Theorem 1.2.3 implies that $\mu = v_n$. We have thus proved the following

Theorem 1.2.5. *For any $A \in \mathcal{R}_n$ and $x \in \mathbf{R}^n$*

$$A + x \in \mathcal{R}_n$$

and

$$v_n(A + x) = v_n(A).$$

Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. A measurable function ξ defined on Ω is called a random variable and the image measure P_ξ is called the probability law of ξ. We sometimes write

$$\mathcal{L}(\xi) = P_\xi.$$

Here are two simple examples.

If the range of a random variable ξ consists of n points $S = \{s_1, ..., s_n\}$ $(n \geq 1)$ and $P_\xi = \frac{1}{n} c_S$, ξ is said to have a uniform distribution in S. Note that

$$P_\xi = \frac{1}{n} \Sigma_{k=1}^{n} \delta_{s_k}.$$

Suppose $\lambda > 0$ is a constant. If a random variable ξ has its range in $\mathbf{N}$ and

$$P_\xi = \Sigma_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} \delta_n$$

then ξ is said to have a Poisson distribution with parameter λ.

Exercises

1. Let $f : X \to Y$, $A \subseteq X$, and $B \subseteq Y$. Show that

$$f(f^{-1}(B)) \subseteq B \text{ and } f^{-1}(f(A)) \supseteq A.$$

2. Let $(X, \mathcal{M})$ be a measurable space and suppose $A \subseteq X$. Show that the function χ_A is $(\mathcal{M}, \mathcal{R})$-measurable if and only if $A \in \mathcal{M}$.

3. Suppose $(X, \mathcal{M})$ is a measurable space and $f_n : X \to \mathbf{R}$, $n \in \mathbf{N}$, a sequence of $(\mathcal{M}, \mathcal{R})$-measurable functions such that

$$\lim_{n \to \infty} f_n(x) \text{ exists and } = f(x) \in \mathbf{R}$$

for each $x \in X$. Prove that f is $(\mathcal{M}, \mathcal{R})$-measurable.

4. Suppose $f : (X, \mathcal{M}) \to (Y, \mathcal{N})$ and $g : (Y, \mathcal{N}) \to (Z, \mathcal{S})$ are measurable. Prove that $g \circ f$ is $(\mathcal{M}, \mathcal{S})$-measurable.

5. Granted the existence of volume measure v_n, show that $v_n(rA) = r^n v_n(A)$ if $r \geq 0$ and $A \in \mathcal{R}_n$.

6. Let μ be the counting measure on $\mathbf{Z}^2$ and $f(x, y) = x$, $(x, y) \in \mathbf{Z}^2$. The positive measure μ is σ-finite. Prove that the image measure $f(\mu)$ is not a σ-finite positive measure.

7. Let $\mu, \nu : \mathcal{R} \to [0, \infty]$ be two positive measures such that $\mu(I) = \nu(I) < \infty$ for each open subinterval of $\mathbf{R}$. Prove that $\mu = \nu$.

8. Let $f : \mathbf{R}^n \to \mathbf{R}^k$ be continuous. Prove that f is $(\mathcal{R}_n, \mathcal{R}_k)$-measurable.

9. Suppose ξ has a Poisson distribution with parameter λ. Show that $P_\xi [2\mathbf{N}] = e^{-\lambda} \cosh \lambda$.

9. Find a σ-additive class which is not a σ-algebra.

1.3. Lebesgue Measure

Once the problem about the existence of volume measure is solved the existence of the so called Lebesgue measure is simple to establish as will be seen in this section. We start with some concepts of general interest.

If $(X, \mathcal{M}, \mu)$ is a positive measure space, the zero set $\mathcal{Z}_\mu$ of μ is, by definition, the set at all $A \in \mathcal{M}$ such that $\mu(A) = 0$. An element of $\mathcal{Z}_\mu$ is called a null set or μ-null set. If

$$(A \in \mathcal{Z}_\mu \text{ and } B \subseteq A) \Rightarrow B \in \mathcal{M}$$

the measure space $(X, \mathcal{M}, \mu)$ is said to be complete. In this case the measure μ is also said to be complete. The positive measure space $(X, \{\phi, X\}, \mu)$, where $X = \{0, 1\}$ and $\mu = 0$, is not complete since $X \in \mathcal{Z}_\mu$ and $\{0\} \notin \{\phi, X\}$.

Theorem 1.3.1 *If $(E_n)_{n=1}^\infty$ is a denumerable collection of members of $\mathcal{Z}_\mu$ then $\cup_{n=1}^\infty E_n \in \mathcal{Z}_\mu$.*

PROOF We have

$$0 \leq \mu(\cup_{n=1}^\infty E_n) \leq \Sigma_{n=1}^\infty \mu(E_n) = 0$$

which proves the result.

Granted the existence of linear measure v_1 it follows from Theorem 1.3.1 that $\mathbf{Q} \in \mathcal{Z}_{v_1}$ since $\mathbf{Q}$ is countable and $\{a\} \in \mathcal{Z}_{v_1}$ for each real number a.

Suppose $(X, \mathcal{M}, \mu)$ is an arbitrary positive measure space. It turns out that μ is the restriction to $\mathcal{M}$ of a complete measure. To see this suppose $\mathcal{M}^-$ is the class of all $E \subseteq X$ is such that there exist sets $A, B \in \mathcal{M}$ such that $A \subseteq E \subseteq B$ and $B \setminus A \in \mathcal{Z}_\mu$. It is obvious that $X \in \mathcal{M}^-$ since $\mathcal{M} \subseteq \mathcal{M}^-$. If $E \in \mathcal{M}^-$, choose $A, B \in \mathcal{M}$ such that $A \subseteq E \subseteq B$ and $B \setminus A \in \mathcal{Z}_\mu$. Then $B^c \subseteq E^c \subseteq A^c$ and $A^c \setminus B^c = B \setminus A \in \mathcal{Z}_\mu$ and we conclude that $E^c \in \mathcal{M}^-$. If $(E_i)_{i=1}^\infty$ is a denumerable collection of members of $\mathcal{M}^-$, for each i there exist sets $A_i, B_i \in \mathcal{M}$ such that $A_i \subseteq E \subseteq B_i$ and $B_i \setminus A_i \in \mathcal{Z}_\mu$. But then

$$\cup_{i=1}^\infty A_i \subseteq \cup_{i=1}^\infty E_i \subseteq \cup_{i=1}^\infty B_i$$

where $\cup_{i=1}^\infty A_i, \cup_{i=1}^\infty B_i \in \mathcal{M}$. Moreover, $(\cup_{i=1}^\infty B_i) \setminus (\cup_{i=1}^\infty A_i) \in \mathcal{Z}_\mu$ since

$$(\cup_{i=1}^\infty B_i) \setminus (\cup_{i=1}^\infty A_i) \subseteq \cup_{i=1}^\infty (B_i \setminus A_i).$$

Thus $\cup_{i=1}^\infty E_i \in \mathcal{M}^-$ and $\mathcal{M}^-$ is a σ-algebra.

If $E \in \mathcal{M}$, suppose $A_i, B_i \in \mathcal{M}$ are such that $A_i \subseteq E \subseteq B_i$ and $B_i \setminus A_i \in \mathcal{Z}_\mu$ for $i = 1, 2$. Then for each i, $(B_1 \cap B_2) \setminus A_i \in \mathcal{Z}_\mu$ and

$$\mu(B_1 \cap B_2) = \mu((B_1 \cap B_2) \setminus A_i) + \mu(A_i) = \mu(A_i).$$

Thus the real numbers $\mu(A_1)$ and $\mu(A_2)$ are the same and we define $\bar{\mu}(E)$ to be equal to this common number. Note also that $\mu(B_1) = \bar{\mu}(E)$. It is plain

that $\bar{\mu}(\phi) = 0$. If $(E_i)_{i=1}^{\infty}$ is a disjoint denumerable collection of members of $\mathcal{M}$, for each i there exist sets $A_i, B_i \in \mathcal{M}$ such that $A_i \subseteq E_i \subseteq B_i$ and $B_i \setminus A_i \in \mathcal{Z}_\mu$. From the above it follows that

$$\bar{\mu}(\cup_{i=1}^{\infty} E_i) = \mu(\cup_{i=1}^{\infty} A_i) = \Sigma_{n=1}^{\infty} \mu(A_i) = \Sigma_{n=1}^{\infty} \bar{\mu}(E_i).$$

We have proved that $\bar{\mu}$ is a positive measure on $\mathcal{M}^-$. If $E \in \mathcal{Z}_{\bar{\mu}}$ the definition of $\bar{\mu}$ shows that any set $A \subseteq E$ belongs to the σ-algebra $\mathcal{M}^-$. It follows that the measure $\bar{\mu}$ is complete and its restriction to $\mathcal{M}$ equals μ.

The measure $\bar{\mu}$ is called the completion of μ and $\mathcal{M}^-$ is called the completion of $\mathcal{M}$ with respect to μ.

Definition 1.3.1 The completion of volume measure v_n on $\mathbf{R}^n$ is called Lebesgue measure on $\mathbf{R}^n$ and is denoted by m_n. The completion of $\mathcal{R}_n$ with respect to v_n is called the Lebesgue σ-algebra in $\mathbf{R}^n$ and is denoted by $\mathcal{R}_n^-$. A member of the class $\mathcal{R}_n^-$ is called a Lebesgue measurable set in $\mathbf{R}^n$ or a Lebesgue set in $\mathbf{R}^n$. A function $f : \mathbf{R}^n \to \mathbf{R}$ is said to be Lebesgue measurable if it is $(\mathcal{R}_n^-, \mathcal{R})$-measurable. Below, m_1 is written m if this notation will not lead to misunderstanding. Furthermore, $\mathcal{R}_1^-$ is written $\mathcal{R}^-$.

Theorem 1.3.2. *Suppose $E \in \mathcal{R}_n^-$ and $x \in \mathbf{R}^n$. Then $E + x \in \mathcal{R}_n^-$ and $m_n(E + x) = m_n(E)$.*

PROOF. Choose $A, B \in \mathcal{R}_n$ such that $A \subseteq E \subseteq B$ and $B \setminus A \in \mathcal{Z}_{v_n}$. Then, by Theorem 1.2.5, $A + x, B + x \in \mathcal{R}_n$, $v_n(A + x) = v_n(A) = m_n(E)$, and $(B + x) \setminus (A + x) = (B \setminus A) + x \in \mathcal{Z}_{v_n}$. Since $A + x \subseteq E + x \subseteq B + x$ the theorem is proved.

The Lebesgue σ-algebra in $\mathbf{R}^n$ is very large and contains each set of interest in analysis and probability. In fact, in most cases, the σ-algebra $\mathcal{R}_n$ is sufficiently large but there are some exceptions. For example, if $f : \mathbf{R}^n \to \mathbf{R}^n$ is continuous and $A \in \mathcal{R}_n$, the image set $f(A)$ need not belong to the class $\mathcal{R}_n$ (see e.g. the Dudley book [D]). To prove the existence of a subset of the real line, which is not Lebesgue measurable we will use the so called Axiom of Choice.

Axiom of Choice. If $(A_i)_{i\in I}$ is a non-empty collection of non-empty sets, there exists a function $f : I \to \cup_{i\in I} A_i$ such that $f(i) \in A_i$ for every $i \in I$.

Let X and Y be sets. The set of all ordered pairs (x, y), where $x \in X$ and $y \in Y$ is denoted by $X \times Y$. An arbitrary subset R of $X \times Y$ is called a relation. If $(x, y) \in R$, we write $x \sim y$. A relation is said to be an equivalence relation on X if $X = Y$ and

(i) $x \sim x$ (reflexivity)
(ii) $x \sim y \Rightarrow y \sim x$ (symmetry)
(iii) ($x \sim y$ and $y \sim z$) $\Rightarrow x \sim z$ (transitivity)

The equivalence class $R(x) =_{def} \{y;\ y \sim x\}$. The definition of the equivalence relation $\sim$ implies the following:

(a) $x \in R(x)$
(b) $R(x) \cap R(y) \neq \phi \Rightarrow R(x) = R(y)$
(c) $\cup_{x\in X} R(x) = X$.

An equivalence relation leads to a partition of X into a disjoint collection of subsets of X.

Let $X = \left[-\frac{1}{2}, \frac{1}{2}\right]$ and define an equivalence relation for numbers x, y in X by stating that $x \sim y$ if $x - y$ is a rational number. By the Axiom of Choice it is possible to pick exactly one element from each equivalence class. Thus there exists a subset NL of X which contains exactly one element from each equivalence class.

If we assume that $NL \in \mathcal{R}^-$ we get a contradiction as follows. Let $(r_i)_{i=1}^{\infty}$ be an enumeration of the rational numbers in $[-1, 1]$. Then

$$X \subseteq \cup_{i=1}^{\infty}(r_i + NL)$$

and it follows from Theorem 1.3.1 that $r_i + NL \notin \mathcal{Z}_m$ for some i. Thus, by Theorem 1.3.2, $NL \notin \mathcal{Z}_m$.

Now assume $(r_i + NL) \cap (r_j + NL) \neq \phi$. Then there exist $a', a'' \in NL$ such that $r_i + a' = r_j + a''$ or $a' - a'' = r_j - r_i$. Hence $a' \sim a''$ and it follows that a' and a'' belong to the same equivalence class. But then $a' = a''$. Thus $r_i = r_j$ and we conclude that $(r_i + NL)_{i \in \mathbf{N}_+}$ is a disjoint enumeration of Lebesgue sets. Now, since

$$\cup_{i=1}^{\infty}(r_i + NL) \subseteq \left[-\frac{3}{2}, \frac{3}{2}\right]$$

it follows that

$$3 \geq m(\cup_{i=1}^{\infty}(r_i + NL)) = \Sigma_{n=1}^{\infty} m(NL).$$

But then $NL \in \mathcal{Z}_m$, which is a contradiction. Thus $NL \notin \mathcal{R}^-$.

In the early 1970' Solovay [S] proved that it is consistent with the usual axioms of Set Theory, excluding the Axiom of Choice, that every subset of $\mathbf{R}$ is Lebesgue measurable.

From the above we conclude that the Axiom of Choice implies the existence of a subset of the set of real numbers which does not belong to the class $\mathcal{R}$. Interestingly enough, such an example can be given without any use of the Axiom of Choice and follows naturally from the theory of analytic sets. The interested reader may consult the Dudley book [D].

Exercises

1. $(X, \mathcal{M}, \mu)$ is a positive measure space. Prove or disprove: If $A \subseteq E \subseteq B$ and $\mu(A) = \mu(B)$ then E belongs to the domain of the completion $\bar{\mu}$.

2. Prove or disprove: If A and B are not Lebesgue measurable subsets of $\mathbf{R}$, then $A \cup B$ is not Lebesgue measurable.

3. Let $(X, \mathcal{M}, \mu)$ be a complete positive measure space and suppose $A, B \in \mathcal{M}$, where $B \setminus A$ is a μ-null set. Prove that $E \in \mathcal{M}$ if $A \subseteq E \subseteq B$ (stated otherwise $\mathcal{M}^- = \mathcal{M}$).

4. Suppose $E \subseteq \mathbf{R}$ and $E \notin \mathcal{R}^-$. Show there is an $\varepsilon > 0$ such that

$$m(B \setminus A) \geq \varepsilon$$

for all $A, B \in \mathcal{R}^-$ such that $A \subseteq E \subseteq B$.

5. Suppose $(X, \mathcal{M}, \mu)$ is a positive measure space and $(Y, \mathcal{N})$ a measurable space. Furthermore, suppose $f : X \to Y$ is $(\mathcal{M}, \mathcal{N})$-measurable and let $\nu = \mu f^{-1}$, that is $\nu(B) = \mu(f^{-1}(B))$, $B \in \mathcal{N}$. Show that f is $(\mathcal{M}^-, \mathcal{N}^-)$-measurable, where $\mathcal{M}^-$ denotes the completion of $\mathcal{M}$ with respect to μ and $\mathcal{N}^-$ the completion of $\mathcal{N}$ with respect to ν.

1.4. Carathéodory's Theorem

In these notes we exhibit two famous approaches to Lebesgue measure. One is based on the Carathéodory Theorem, which we present in this section, and the other one, due to F. Riesz, is a representation theorem of positive linear functionals on spaces of continuous functions in terms of positive measures. The latter approach, is presented in Chapter 3. Both methods depend on topological concepts such as compactness.

Definition 1.4.1. A function $\theta : \mathcal{P}(X) \to [0, \infty]$ is said to be an outer measure if the following properties are satisfied:

(i) $\theta(\phi) = 0$.
(ii) $\theta(A) \leq \theta(B)$ if $A \subseteq B$.
(iii) for any denumerable collection $(A_n)_{n=1}^{\infty}$ of subsets of X

$$\theta(\cup_{n=1}^{\infty} A_n) \leq \Sigma_{n=1}^{\infty} \theta(A_n).$$

Since

$$E = (E \cap A) \cup (E \cap A^c)$$

an outer measure θ satisfies the inequality

$$\theta(E) \leq \theta(E \cap A) + \theta(E \cap A^c).$$

If θ is an outer measure on X we define $\mathcal{M}(\theta)$ as the set of all $A \subseteq X$ such that

$$\theta(E) = \theta(E \cap A) + \theta(E \cap A^c) \text{ for all } E \subseteq X$$

or, what amounts to the same thing,

$$\theta(E) \geq \theta(E \cap A) + \theta(E \cap A^c) \text{ for all } E \subseteq X.$$

The next theorem is one of the most important in measure theory.

Theorem 1.4.1. (Carathéodory's Theorem) *Suppose θ is an outer measure. The class $\mathcal{M}(\theta)$ is a σ-algebra and the restriction of θ to $\mathcal{M}(\theta)$ is a complete measure.*

PROOF. Clearly, $\phi \in \mathcal{M}(\theta)$ and $A^c \in \mathcal{M}(\theta)$ if $A \in \mathcal{M}(\theta)$. Moreover, if $A, B \in \mathcal{M}(\theta)$ and $E \subseteq X$,

$$\begin{aligned}\theta(E) &= \theta(E \cap A) + \theta(E \cap A^c) \\ &= \theta(E \cap A \cap B) + \theta(E \cap A \cap B^c) \\ &+\theta(E \cap A^c \cap B) + \theta(E \cap A^c \cap B^c).\end{aligned}$$

But

$$A \cup B = (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$$

and

$$A^c \cap B^c = (A \cup B)^c$$

and we get

$$\theta(E) \geq \theta(E \cap (A \cup B)) + \theta(E \cap (A \cup B)^c).$$

It follows that $A \cup B \in \mathcal{M}(\theta)$ and we have proved that the class $\mathcal{M}(\theta)$ is an algebra. Now if $A, B \in \mathcal{M}(\theta)$ are disjoint

$$\theta(A \cup B) = \theta((A \cup B) \cap A) + \theta((A \cup B) \cap A^c) = \theta(A) + \theta(B)$$

and therefore the restriction of θ to $\mathcal{M}(\theta)$ is a content.

Next we prove that $\mathcal{M}(\theta)$ is a σ-algebra. Let $(A_i)_{i=1}^{\infty}$ be a disjoint denumerable collection of members of $\mathcal{M}(\theta)$ and set for each $n \in \mathbf{N}$

$$B_n = \cup_{1 \leq i \leq n} A_i \text{ and } B = \cup_{i=1}^{\infty} A_i$$

(here $B_0 = \phi$). Then for any $E \subseteq X$,

$$\theta(E \cap B_n) = \theta(E \cap B_n \cap A_n) + \theta(E \cap B_n \cap A_n^c)$$

$$= \theta(E \cap A_n) + \theta(E \cap B_{n-1})$$

and, by induction,

$$\theta(E \cap B_n) = \Sigma_{i=1}^{n} \theta(E \cap A_i).$$

But then

$$\theta(E) = \theta(E \cap B_n) + \theta(E \cap B_n^c)$$

$$\geq \Sigma_{i=1}^{n} \theta(E \cap A_i) + \theta(E \cap B^c)$$

and letting $n \to \infty$,

$$\theta(E) \geq \Sigma_{i=1}^{\infty} \theta(E \cap A_i) + \theta(E \cap B^c)$$

$$\geq \theta(\cup_{i=1}^{\infty}(E \cap A_i)) + \theta(E \cap B^c)$$

$$= \theta(E \cap B) + \theta(E \cap B^c) \geq \theta(E).$$

All the inequalities in the last calculation must be equalities and we conclude that $B \in \mathcal{M}(\theta)$ and, choosing $E = B$, results in

$$\theta(B) = \Sigma_{i=1}^{\infty} \theta(A_i).$$

Thus $\mathcal{M}(\theta)$ is a σ-algebra and the restriction of θ to $\mathcal{M}(\theta)$ is a positive measure.

Finally we prove that the the restriction of θ to $\mathcal{M}(\theta)$ is a complete measure. Suppose $B \subseteq A \in \mathcal{M}(\theta)$ and $\theta(A) = 0$. If $E \subseteq X$,

$$\theta(E) \leq \theta(E \cap B) + \theta(E \cap B^c) \leq \theta(E \cap B^c) \leq \theta(E)$$

and so $B \in \mathcal{M}(\theta)$. The theorem is proved.

Unit 11

1.5. Existence of Linear Measure

The purpose of this section is to show the existence of linear measure on $\mathbf{R}$ using the Carathéodory Theorem and a minimum of topology.

First let us recall the definition of infimum and supremum of a non-empty subset of the extended real line. Suppose A is a non-empty subset of $[-\infty, \infty] = \mathbf{R} \cup \{-\infty, \infty\}$. We define $-\infty \leq x$ and $x \leq \infty$ for all $x \in [-\infty, \infty]$. An element $b \in [-\infty, \infty]$ is called a majorant of A if $x \leq b$ for all $x \in A$ and a minorant if $x \geq b$ for all $x \in A$. The Supremum Axiom states that A possesses a least majorant, which is denoted by $\sup A$. From this follows that if A is non-empty, then A possesses a greatest minorant, which is denoted by $\inf A$. (Actually, the Supremum Axiom is a theorem in courses where time is spent on the definition of real numbers.)

Theorem 1.5.1. (The Heine-Borel Theorem; weak form) *Let $[a, b]$ be a closed bounded interval and $(U_i)_{i \in I}$ a collection of open sets such that*

$$\cup_{i \in I} U_i \supseteq [a, b].$$

Then

$$\cup_{i \in J} U_i \supseteq [a, b]$$

for some finite subset J of I.

PROOF. Let A be the set of all $x \in [a, b]$ such that

$$\cup_{i \in J} U_i \supseteq [a, x]$$

for some finite subset J of I. Clearly, $a \in A$ since $a \in U_i$ for some i. Let $c = \sup A$. There exists an i_0 such that $c \in U_{i_0}$. Let $c \in]a_0, b_0[\subseteq U_{i_0}$, where $a_0 < b_0$. Furthermore, by the very definition of least upper bound, there exists a finite set J such that

$$\cup_{i \in J} U_i \supseteq [a, (a_0 + c)/2] .$$

Hence

$$\cup_{i \in J \cup \{i_0\}} U_k \supseteq [a, (c + b_0)/2]$$

and it follows that $c \in A$ and $c = b$. The lemma is proved.

A subset K of $\mathbf{R}$ is called compact if for every family of open subsets U_i, $i \in I$, with $\cup_{i \in I} U_i \supseteq K$ we have $\cup_{i \in J} U_i \supseteq K$ for some finite subset J of I. The Heine-Borel Theorem shows that a closed bounded interval is compact.

If $x, y \in \mathbf{R}$ and E, $F \subseteq \mathbf{R}$, let

$$d(x, y) = \mid x - y \mid$$

be the distance between x and y, let

$$d(x, E) = \inf_{u \in E} d(x, u)$$

be the distance from x to E, and let

$$d(E, F) = \inf_{u \in E, v \in F} d(u, v)$$

be the distance between E and F (here the infimum of the emty set equals ∞). Note that for any $u \in E$,

$$d(x, u) \leq d(x, y) + d(y, u)$$

and, hence

$$d(x, E) \leq d(x, y) + d(y, u)$$

and

$$d(x, E) \leq d(x, y) + d(y, E).$$

By interchanging the roles of x and y and assuming that $E \neq \phi$, we get

$$| d(x, E) - d(y, E) | \leq d(x, y).$$

Note that if $F \subseteq \mathbf{R}$ is closed and $x \notin F$, then $d(x, F) > 0$.
An outer measure $\theta : \mathcal{P}(\mathbf{R}) \to [0, \infty]$ is called a metric outer measure if

$$\theta(A \cup B) = \theta(A) + \theta(B)$$

for all $A, B \in \mathcal{P}(\mathbf{R})$ such that $d(A, B) > 0$.

Theorem 1.5.2. *If* $\theta : \mathcal{P}(\mathbf{R}) \to [0, \infty]$ *is a metric outer measure, then* $\mathcal{R} \subseteq \mathcal{M}(\theta)$.

PROOF. Let $F \in \mathcal{P}(\mathbf{R})$ be closed. It is enough to show that $F \in \mathcal{M}(\theta)$. To this end we choose $E \subseteq X$ with $\theta(E) < \infty$ and prove that

$$\theta(E) \geq \theta(E \cap F) + \theta(E \cap F^c).$$

Let $n \geq 1$ be an integer and define

$$A_n = \left\{ x \in E \cap F^c;\ d(x, F) \geq \frac{1}{n} \right\}.$$

Note that $A_n \subseteq A_{n+1}$ and

$$E \cap F^c = \cup_{n=1}^{\infty} A_n.$$

Moreover, since θ is a metric outer measure

$$\theta(E) \geq \theta((E \cap F) \cup A_n) = \theta(E \cap F) + \theta(A_n)$$

and, hence, proving

$$\theta(E \cap F^c) = \lim_{n \to \infty} \theta(A_n)$$

we are done.

Let $B_n = A_{n+1} \cap A_n^c$. It is readily seen that

$$d(B_{n+1}, A_n) \geq \frac{1}{n(n+1)}$$

since if $x \in B_{n+1}$ and

$$d(x, y) < \frac{1}{n(n+1)}$$

then

$$d(y, F) \leq d(y, x) + d(x, F) < \frac{1}{n(n+1)} + \frac{1}{n+1} = \frac{1}{n}.$$

Now

$$\theta(A_{2k+1}) \geq \theta(B_{2k} \cup A_{2k-1}) = \theta(B_{2k}) + \theta(A_{2k-1})$$
$$\geq ... \geq \Sigma_{i=1}^{k} \theta(B_{2i})$$

and in a similar way

$$\theta(A_{2k}) \geq \Sigma_{i=1}^{k} \theta(B_{2i-1}).$$

But $\theta(A_n) \leq \theta(E) < \infty$ and we conclude that

$$\Sigma_{i=1}^{\infty} \theta(B_i) < \infty.$$

We now use that

$$E \cap F^c = A_n \cup (\cup_{i=n}^{\infty} B_i)$$

to obtain

$$\theta(E \cap F^c) \leq \theta(A_n) + \Sigma_{i=n}^{\infty} \theta(B_i).$$

Now, since $\theta(E \cap F^c) \geq \theta(A_n)$,

$$\theta(E \cap F^c) = \lim_{n \to \infty} \theta(A_n)$$

and the theorem is proved.

PROOF OF THEOREM 1.1.1 IN ONE DIMENSION. Suppose $\delta > 0$. If $A \subseteq \mathbf{R}$, define

$$\theta_\delta(A) = \inf \Sigma_{k=1}^{\infty} l(I_k)$$

the infimum being taken over all open intervals I_k with $l(I_k) < \delta$ such that

$$A \subseteq \cup_{k=1}^{\infty} I_k.$$

Obviously, $\theta_\delta(\phi) = 0$ and $\theta_\delta(A) \leq \theta_\delta(B)$ if $A \subseteq B$. Suppose $(A_n)_{n=1}^\infty$ is a denumerable collection of subsets of $\mathbf{R}$ and let $\varepsilon > 0$. For each n there exist open intervals $I_{kn}, k \in \mathbf{N}_+$, such that $l(I_{kn}) < \delta$,

$$A_n \subseteq \cup_{k=1}^\infty I_{kn}$$

and

$$\Sigma_{k=1}^\infty l(I_{kn}) \leq \theta_\delta(A_n) + \varepsilon 2^{-n}.$$

Then

$$A =_{def} \cup_{n=1}^\infty A_n \subseteq \cup_{k,n=1}^\infty I_{kn}$$

and

$$\Sigma_{k,n=1}^\infty l(I_{kn}) \leq \Sigma_{n=1}^\infty \theta_\delta(A_n) + \varepsilon.$$

Thus

$$\theta_\delta(A) \leq \Sigma_{n=1}^\infty \theta_\delta(A_n) + \varepsilon$$

and, since $\varepsilon > 0$ is arbitrary,

$$\theta_\delta(A) \leq \Sigma_{n=1}^\infty \theta_\delta(A_n).$$

It follows that θ_δ is an outer measure.

If I is an open interval it is simple to see that

$$\theta_\delta(I) \leq l(I).$$

To prove the reverse inequality, choose a closed bounded interval $J \subseteq I$. Now, if

$$I \subseteq \cup_{k=1}^\infty I_k$$

where each I_k is an open interval of $l(I_k) < \delta$, it follows from the Heine-Borel Theorem that

$$J \subseteq \cup_{k=1}^n I_k$$

for some n. Hence

$$l(J) \leq \Sigma_{k=1}^n l(I_k) \leq \Sigma_{k=1}^\infty l(I_k)$$

and it follows that

$$l(J) \leq \theta_\delta(I)$$

and, accordingly from this,

$$l(I) \leq \theta_\delta(I).$$

Thus, if I is an open interval, then

$$\theta_\delta(I) = l(I).$$

Note that $\theta_{\delta_1} \geq \theta_{\delta_2}$ if $0 < \delta_1 \leq \delta_2$. We define

$$\theta_0(A) = \lim_{\delta \to 0} \theta_\delta(A) \text{ if } A \subseteq \mathbf{R}.$$

It obvious that θ_0 is an outer measure such that $\theta_0(I) = l(I)$, if I is an open interval.

To complete the proof we show that θ_0 is a metric outer measure. To this end let $A, B \subseteq \mathbf{R}$ and $d(A, B) > 0$. Suppose $0 < \delta < d(A, B)$ and

$$A \cup B \subseteq \cup_{k=1}^{\infty} I_k$$

where each I_k is an open interval with $l(I_k) < \delta$. Let

$$\alpha = \{k;\ I_k \cap A \neq \phi\}$$

and

$$\beta = \{k;\ I_k \cap B \neq \phi\}.$$

Then $\alpha \cap \beta = \phi$,

$$A \subseteq \cup_{k \in \alpha} I_k$$

and

$$B \subseteq \cup_{k \in \beta} I_k$$

and it follows that

$$\Sigma_{k=1}^{\infty} l(I_k) \geq \Sigma_{k \in \alpha} l(I_k) + \Sigma_{k \in \beta} l(I_k)$$

$$\geq \theta_\delta(A) + \theta_\delta(B).$$

Thus

$$\theta_\delta(A \cup B) \geq \theta_\delta(A) + \theta_\delta(B)$$

and by letting $\delta \to 0$ we have

$$\theta_0(A \cup B) \geq \theta_0(A) + \theta_0(B)$$

and

$$\theta_0(A \cup B) = \theta_0(A) + \theta_0(B).$$

Finally by applying the Carathéodory Theorem and Theorem 1.5.2 it follows that the restriction of θ_0 to $\mathcal{R}$ equals v_1.

We end this section with some additional results of great interest.

Theorem 1.5.3. *For any $\delta > 0$, $\theta_\delta = \theta_0$. Moreover, if $A \subseteq \mathbf{R}$*

$$\theta_0(A) = \inf \Sigma_{k=1}^{\infty} l(I_k)$$

the infimum being taken over all open intervals I_k, $k \in \mathbf{N}_+$, such that $\cup_{k=1}^{\infty} I_k \supseteq A$.

PROOF. It follows from the definition of θ_0 that $\theta_\delta \leq \theta_0$. To prove the reverse inequality let $A \subseteq \mathbf{R}$ and choose open intervals I_k, $k \in \mathbf{N}_+$, such that $\cup_{k=1}^{\infty} I_k \supseteq A$. Then

$$\theta_0(A) \leq \theta_0(\cup_{k=1}^{\infty} I_k) \leq \Sigma_{k=1}^{\infty} \theta_0(I_k)$$

$$= \Sigma_{k=1}^{\infty} l(I_k).$$

Hence

$$\theta_0(A) \leq \inf \Sigma_{k=1}^{\infty} l(I_k)$$

the infimum being taken over all open intervals I_k, $k \in \mathbf{N}_+$, such that $\cup_{k=1}^{\infty} I_k \supseteq A$. Thus $\theta_0(A) \leq \theta_\delta(A)$, which completes the proof of Theorem 1.5.3.

Theorem 1.5.4. *If $A \subseteq \mathbf{R}$,*

$$\theta_0(A) = \inf_{\substack{U \supseteq A \\ U \text{ open}}} \theta_0(U).$$

Moreover, if $A \in \mathcal{M}(\theta_0)$,

$$\theta_0(A) = \sup_{\substack{K \subseteq A \\ K \text{ closed bounded}}} \theta_0(K).$$

PROOF. If $A \subseteq U$, $\theta_0(A) \leq \theta_0(U)$. Hence

$$\theta_0(A) \leq \inf_{\substack{U \supseteq A \\ U \text{ open}}} \theta_0(U).$$

Next let $\varepsilon > 0$ be fixed and choose open intervals I_k, $k \in \mathbf{N}_+$, such that $\cup_{k=1}^{\infty} I_k \supseteq A$ and

$$\Sigma_{k=1}^{\infty} l(I_k) \leq \theta_0(A) + \varepsilon$$

(here observe that it may happen that $\theta_0(A) = \infty$). Then the set $U =_{def} \cup_{k=1}^{\infty} I_k$ is open and

$$\theta_0(U) \leq \Sigma_{k=1}^{\infty} \theta_0(I_k) = \Sigma_{k=1}^{\infty} l(I_k) \leq \theta_0(A) + \varepsilon.$$

Thus

$$\inf_{\substack{U \supseteq A \\ U \text{ open}}} \theta_0(U) \leq \theta_0(A)$$

and we have proved that

$$\theta_0(A) = \inf_{\substack{U \supseteq A \\ U \text{ open}}} \theta_0(U).$$

If $K \subseteq A$, $\theta_0(K) \leq \theta_0(A)$ and, accordingly from this,

$$\sup_{\substack{K \subseteq A \\ K \text{ closed bounded}}} \theta_0(K) \leq \theta_0(A).$$

To prove the reverse inequality we first assume that $A \in \mathcal{M}(\theta_0)$ is bounded. Let $\varepsilon > 0$ be fixed and suppose J is a closed bounded interval containing A. Then we know from the first part of Theorem 1.5.4 already proved that there exists an open set $U \supseteq J \setminus A$ such that

$$\theta_0(U) < \theta_0(J \setminus A) + \varepsilon.$$

But then

$$\theta_0(J) \leq \theta_0(J \setminus U) + \theta_0(U) < \theta_0(J \setminus U) + \theta_0(J \setminus A) + \varepsilon$$

and it follows that

$$\theta_0(A) - \varepsilon < \theta_0(J \setminus U).$$

Since $J \setminus U$ is a closed bounded set contained in A we conclude that

$$\theta_0(A) \leq \sup_{\substack{K \subseteq A \\ K \text{ closed bounded}}} \theta_0(K).$$

If $A \in \mathcal{M}(\theta_0)$ let $A_n = A \cap [-n, n]$, $n \in \mathbf{N}_+$. Then given $\varepsilon > 0$ and $n \in \mathbf{N}_+$, let K_n be a closed bounded subset of A_n such that $\theta_0(K_n) > \theta_0(A_n) - \varepsilon$. Clearly, there is no loss of generality to assume that $K_1 \subseteq K_2 \subseteq K_3 \subseteq ...$ and by letting n tend to plus infinity we get

$$\lim_{n \to \infty} \theta_0(K_n) \geq \theta_0(A) - \varepsilon.$$

Hence

$$\theta_0(A) = \sup_{\substack{K \subseteq A \\ K \text{ compact}}} \theta_0(K).$$

and Theorem 1.5.4 is completely proved.

Unit 12

Introduction

In this chapter Lebesgue integration in abstract positive measure spaces is introduced. A series of famous theorems and lemmas will be proved.

2.1. Integration of Functions with Values in $[0, \infty]$

Recall that $[0, \infty] = [0, \infty[\cup \{\infty\}$. A subinterval of $[0, \infty]$ is defined in the natural way. We denote by $\mathcal{R}_{0,\infty}$ the σ-algebra generated by all subintervals of $[0, \infty]$. The class of all intervals of the type $]\alpha, \infty]$, $0 \leq \alpha < \infty$, (or of the type $[\alpha, \infty]$, $0 \leq \alpha < \infty$) generates the σ-algebra $\mathcal{R}_{0,\infty}$ and we get the following

Theorem 2.1.1. *Let $(X, \mathcal{M})$ be a measurable space and suppose $f : X \to [0, \infty]$.*

(a) *The function f is $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable if $f^{-1}(]\alpha, \infty]) \in \mathcal{M}$ for every $0 \leq \alpha < \infty$.*

(b) *The function f is $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable if $f^{-1}([\alpha, \infty]) \in \mathcal{M}$ for every $0 \leq \alpha < \infty$.*

Note that the set $\{f > \alpha\} \in \mathcal{M}$ for all real α if f is $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable.

If $f, g : X \to [0, \infty]$ are $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable, then $\min(f, g)$, $\max(f, g)$, and $f + g$ are $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable, since, for each $\alpha \in [0, \infty[$,

$$\min(f, g) \geq \alpha \Leftrightarrow (f \geq \alpha \text{ and } g \geq \alpha)$$

$$\max(f, g) \geq \alpha \Leftrightarrow (f \geq \alpha \text{ or } g \geq \alpha)$$

and

$$\{f+g>\alpha\}=\bigcup_{q\in\mathbf{Q}}(\{f>\alpha-q\}\cap\{g>q\}).$$

Given functions $f_n : X \to [0,\infty]$, $n = 1, 2, ...$, $f = \sup_{n\geq 1} f_n$ is defined by the equation

$$f(x)=\sup\{f_n(x);\ n=1,2,...\}.$$

Note that

$$f^{-1}(]\alpha,\infty])=\cup_{n=1}^{\infty}f_n^{-1}(]\alpha,\infty])$$

for every real $\alpha \geq 0$ and, accordingly from this, the function $\sup_{n\geq 1} f_n$ is $(\mathcal{M},\mathcal{R}_{0,\infty})$-measurable if each f_n is $(\mathcal{M},\mathcal{R}_{0,\infty})$-measurable. Moreover, $f = \inf_{n\geq 1} f_n$ is given by

$$f(x)=\inf\{f_n(x);\ n=1,2,...\}.$$

Since

$$f^{-1}([0,\alpha[)=\cup_{n=1}^{\infty}f_n^{-1}([0,\alpha[)$$

for every real $\alpha \geq 0$ we conclude that the function $f = \inf_{n\geq 1} f_n$ is $(\mathcal{M},\mathcal{R}_{0,\infty})$-measurable if each f_n is $(\mathcal{M},\mathcal{R}_{0,\infty})$-measurable.

Below we write

$$f_n\uparrow f$$

if f_n, $n = 1, 2, ...$, and f are functions from X into $[0,\infty]$ such that $f_n \leq f_{n+1}$ for each n and $f_n(x) \to f(x)$ for each $x \in X$ as $n \to \infty$.

An $(\mathcal{M},\mathcal{R}_{0,\infty})$-measurable function $\varphi : X \to [0,\infty]$ is called a simple measurable function if $\varphi(X)$ is a finite subset of $[0,\infty[$. If it is neccessary to be more precise, we say that φ is a simple $\mathcal{M}$-measurable function.

Theorem 2.1.2. *Let $f : X \to [0,\infty]$ be $(\mathcal{M},\mathcal{R}_{0,\infty})$-measurable. There exist simple measurable functions φ_n, $n \in \mathbf{N}_+$, on X such that $\varphi_n \uparrow f$.*

PROOF. Given $n \in \mathbf{N}_+$, set

$$E_{in}=f^{-1}(\left[\frac{i-1}{2^n},\frac{i}{2^n}\right[),\ i\in\mathbf{N}_+$$

and

$$\rho_n = \sum_{i=1}^{\infty} \frac{i-1}{2^n} \chi_{E_{in}} + \infty \chi_{f^{-1}(\{\infty\})}.$$

It is obvious that $\rho_n \leq f$ and that $\rho_n \leq \rho_{n+1}$. Now set $\varphi_n = \min(n, \rho_n)$ and we are done.

Let $(X, \mathcal{M}, \mu)$ be a positive measure space and $\varphi : X \to [0, \infty[$ a simple measurable function. If $\alpha_1, ..., \alpha_n$ are the distinct values of the simple function φ, and if $E_i = \varphi^{-1}(\{\alpha_i\})$, $i = 1, ..., n$, then

$$\varphi = \Sigma_{i=1}^{n} \alpha_i \chi_{E_i}.$$

Furthermore, if $A \in \mathcal{M}$ we define

$$\nu(A) = \int_A \varphi d\mu = \Sigma_{i=1}^{n} \alpha_i \mu(E_i \cap A) = \Sigma_{k=1}^{n} \alpha_i \mu^{E_i}(A).$$

Note that this formula still holds if $(E_i)_1^n$ is a measurable partition of X and $\varphi = \alpha_i$ on E_i for each $i = 1, ..., n$. Clearly, ν is a positive measure since each term in the right side is a positive measure as a function of A. Note that

$$\int_A \alpha\varphi d\mu = \alpha \int_A \varphi d\mu \text{ if } 0 \leq \alpha < \infty$$

and

$$\int_A \varphi d\mu = a\mu(A)$$

if $a \in [0, \infty[$ and φ is a simple measurable function such that $\varphi = a$ on A.

If ψ is another simple measurable function and $\varphi \leq \psi$,

$$\int_A \varphi d\mu \leq \int_A \psi d\mu.$$

To see this, let $\beta_1, ..., \beta_p$ be the distinct values of ψ and $F_j = \psi^{-1}(\{\beta_j\})$, $j = 1, ..., p$. Now, putting $B_{ij} = E_i \cap F_j$,

$$\int_A \varphi d\mu = \nu(\cup_{ij}(A \cap B_{ij}))$$

$$= \Sigma_{ij} \nu(A \cap B_{ij}) = \Sigma_{ij} \int_{A \cap B_{ij}} \varphi d\mu = \Sigma_{ij} \int_{A \cap B_{ij}} \alpha_i d\mu$$

$$\leq \Sigma_{ij} \int_{A\cap B_{ij}} \beta_j d\mu = \int_A \psi d\mu.$$

In a similar way one proves that

$$\int_A (\varphi + \psi) d\mu = \int_A \varphi d\mu + \int_A \psi d\mu.$$

From the above it follows that

$$\int_A \varphi\chi_A d\mu = \int_A \Sigma_{i=1}^n \alpha_i \chi_{E_i\cap A} d\mu$$

$$= \Sigma_{i=1}^n \alpha_i \int_A \chi_{E_i\cap A} d\mu = \Sigma_{i=1}^n \alpha_i \mu(E_i \cap A)$$

and

$$\int_A \varphi\chi_A d\mu = \int_A \varphi d\mu.$$

If $f : X \to [0, \infty]$ is an $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable function and $A \in \mathcal{M}$, we define

$$\int_A f d\mu = \sup \left\{ \int_A \varphi d\mu;\ 0 \leq \varphi \leq f,\ \varphi \text{ simple measurable} \right\}$$

$$= \sup \left\{ \int_A \varphi d\mu;\ 0 \leq \varphi \leq f,\ \varphi \text{ simple measurable and } \varphi = 0 \text{ on } A^c \right\}.$$

The left member in this equation is called the Lebesgue integral of f over A with respect to the measure μ. Sometimes we also speek of the μ-integral of f over A. The two definitions of the μ-integral of a simple measurable function $\varphi : X \to [0, \infty[$ over A agree.

From now on in this section, an $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable function $f : X \to [0, \infty]$ is simply called measurable.

The following properties are immediate consequences of the definitions. The functions and sets occurring in the equations are assumed to be measurable.

(a) If f, $g \geq 0$ and $f \leq g$ on A, then $\int_A f d\mu \leq \int_A g d\mu$.

(b) $\int_A f d\mu = \int_X \chi_A f d\mu$.

(c) If $f \geq 0$ and $\alpha \in [0, \infty[$, then $\int_A \alpha f d\mu = \alpha \int_A f d\mu$.

(d) $\int_A f d\mu = 0$ if $f = 0$ and $\mu(A) = \infty$.

(e) $\int_A f d\mu = 0$ if $f = \infty$ and $\mu(A) = 0$.

If $f : X \to [0, \infty]$ is measurable and $0 < \alpha < \infty$, then $f \geq \alpha \chi_{f^{-1}([\alpha,\infty])} = \alpha \chi_{\{f \geq \alpha\}}$and

$$\int_X f d\mu \geq \int_X \alpha \chi_{\{f \geq \alpha\}} d\mu = \alpha \int_X \chi_{\{f \geq \alpha\}} d\mu.$$

This proves the so called Markov Inequality

$$\mu(f \geq \alpha) \leq \frac{1}{\alpha} \int_X f d\mu$$

where we write $\mu(f \geq \alpha)$ instead of the more precise expression $\mu(\{f \geq \alpha\})$.

Example 2.1.1. Suppose $f : X \to [0, \infty]$ is measurable and

$$\int_X f d\mu < \infty.$$

We claim that

$$\{f = \infty\} = f^{-1}(\{\infty\}) \in \mathcal{Z}_\mu.$$

To prove this we use the Markov Inequality and have

$$\mu(f = \infty) \leq \mu(f \geq \alpha) \leq \frac{1}{\alpha} \int_X f d\mu$$

for each $\alpha \in]0, \infty[$. Thus $\mu(f = \infty) = 0$.

Example 2.1.2. Suppose $f : X \to [0, \infty]$ is measurable and

$$\int_X f d\mu = 0.$$

We claim that

$$\{f > 0\} = f^{-1}(]0, \infty]) \in \mathcal{Z}_\mu.$$

To see this, note that

$$f^{-1}(]0, \infty]) = \cup_{n=1}^{\infty} f^{-1}(\left]\frac{1}{n}, \infty\right])$$

Furthermore, for every fixed $n \in \mathbf{N}_+$, the Markov Inequality yields

$$\mu(f > \frac{1}{n}) \leq n \int_X f d\mu = 0$$

and we get $\{f > 0\} \in \mathcal{Z}_\mu$ since a countable union of null sets is a null set.

We now come to one of the most important results in the theory.

Theorem 2.1.3. (Monotone Convergence Theorem) *Let* $f_n : X \to [0, \infty]$, $n = 1, 2, 3, ...$, *be a sequence of measurable functions and suppose that* $f_n \uparrow f$, that is $0 \leq f_1 \leq f_2 \leq ...$ *and*

$$f_n(x) \to f(x) \text{ as } n \to \infty, \text{ for every } x \in X.$$

Then f *is measurable and*

$$\int_X f_n d\mu \to \int_X f d\mu \text{ as } n \to \infty.$$

PROOF. The function f is measurable since $f = \sup_{n \geq 1} f_n$.

The inequalities $f_n \leq f_{n+1} \leq f$ yield $\int_X f_n d\mu \leq \int_X f_{n+1} d\mu \leq \int_X f d\mu$ and we conclude that there exists an $\alpha \in [0, \infty]$ such that

$$\int_X f_n d\mu \to \alpha \text{ as } n \to \infty$$

and

$$\alpha \leq \int_X f d\mu.$$

To prove the reverse inequality, let φ be any simple measurable function such that $0 \leq \varphi \leq f$, let $0 < \theta < 1$ be a constant, and define, for fixed $n \in \mathbf{N}_+$,

$$A_n = \{x \in X;\ f_n(x) \geq \theta \varphi(x)\}.$$

If $\alpha_1, ..., \alpha_p$ are the distinct values of φ,

$$A_n = \cup_{k=1}^{p} (\{x \in X;\ f_n(x) \geq \theta \alpha_k\} \cap \{\varphi = \alpha_k\})$$

and it follows that A_n is measurable. Clearly, $A_1 \subseteq A_2 \subseteq ...$. Moreover, if $f(x) = 0$, then $x \in A_1$ and if $f(x) > 0$, then $\theta \varphi(x) < f(x)$ and $x \in A_n$ for all sufficiently large n. Thus $\cup_{n=1}^{\infty} A_n = X$. Now

$$\alpha \geq \int_{A_n} f_n d\mu \geq \theta \int_{A_n} \varphi d\mu$$

and we get

$$\alpha \geq \theta \int_X \varphi d\mu$$

since the map $A \to \int_A \varphi d\mu$ is a positive measure on $\mathcal{M}$. By letting $\theta \uparrow 1$,

$$\alpha \geq \int_X \varphi d\mu$$

and, hence

$$\alpha \geq \int_X f d\mu.$$

The theorem follows.

Theorem 2.1.4. (a) *Let $f, g : X \to [0, \infty]$ be measurable functions. Then*

$$\int_X (f + g) d\mu = \int_X f d\mu + \int_X g d\mu.$$

(b) **(Beppo Levi's Theorem)** *If* $f_k : X \to [0,\infty]$, $k = 1, 2, ...$ *are measurable,*

$$\int_X \Sigma_{k=1}^{\infty} f_k d\mu = \Sigma_{k=1}^{\infty} \int_X f_k d\mu$$

PROOF. (a) Let $(\varphi_n)_{n=1}^{\infty}$ and $(\psi_n)_{n=1}^{\infty}$ be sequences of simple and measurable functions such that $0 \leq \varphi_n \uparrow f$ and $0 \leq \psi_n \uparrow g$. We proved above that

$$\int_X (\varphi_n + \psi_n) d\mu = \int_X \varphi_n d\mu + \int_X \psi_n d\mu$$

and, by letting $n \to \infty$, Part (a) follows from the Monotone Convergence Theorem.

(b) Part (a) and induction imply that

$$\int_X \Sigma_{k=1}^{n} f_k d\mu = \Sigma_{k=1}^{n} \int_X f_k d\mu$$

and the result follows from monotone convergence.

Theorem 2.1.5. *Suppose* $w : X \to [0,\infty]$ *is a measurable function and define*

$$\nu(A) = \int_A w d\mu, \ A \in \mathcal{M}.$$

Then ν *is a positive measure and*

$$\int_A f d\nu = \int_A f w d\mu, \ A \in \mathcal{M}$$

for every measurable function $f : X \to [0,\infty]$.

PROOF. Clearly, $\nu(\phi) = 0$. Suppose $(E_k)_{k=1}^{\infty}$ is a disjoint denumerable collection of members of $\mathcal{M}$ and set $E = \cup_{k=1}^{\infty} E_k$. Then

$$\nu(\cup_{k=1}^{\infty} E_k) = \int_E w d\mu = \int_X \chi_E w d\mu = \int_X \Sigma_{k=1}^{\infty} \chi_{E_k} w d\mu$$

where, by the Beppo Levi Theorem, the right member equals

$$\Sigma_{k=1}^{\infty}\int_X \chi_{E_k} w d\mu = \Sigma_{k=1}^{\infty}\int_{E_k} w d\mu = \Sigma_{k=1}^{\infty}\nu(E_k).$$

This proves that ν is a positive measure.

Let $A \in \mathcal{M}$. To prove the last part in Theorem 2.1.5 we introduce the class $\mathcal{C}$ of all measurable functions $f : X \to [0, \infty]$ such that

$$\int_A f d\nu = \int_A f w d\mu.$$

The indicator function of a measurable set belongs to $\mathcal{C}$ and from this we conclude that every simple measurable function belongs to $\mathcal{C}$. Furthermore, if $f_n \in \mathcal{C}$, $n \in \mathbf{N}$, and $f_n \uparrow f$, the Monotone Convergence Theorem proves that $f \in \mathcal{C}$. Thus in view of Theorem 2.1.2 the class $\mathcal{C}$ contains every measurable function $f : X \to [0, \infty]$. This completes the proof of Theorem 2.1.5.

The measure ν in Theorem 2.1.5 is written

$$\nu = w\mu$$

or

$$d\nu = w d\mu.$$

Let $(\alpha_n)_{n=1}^{\infty}$ be a sequence in $[-\infty, \infty]$. First put $\beta_k = \inf\{\alpha_k, \alpha_{k+1}, \alpha_{k+2}, ...\}$ and $\gamma = \sup\{\beta_1, \beta_2, \beta_3, ..\} = \lim_{n\to\infty} \beta_n$. We call γ the lower limit of $(\alpha_n)_{n=1}^{\infty}$ and write

$$\gamma = \liminf_{n\to\infty} \alpha_n.$$

Note that

$$\gamma = \lim_{n\to\infty} \alpha_n$$

if the limit exists. Now put $\beta_k = \sup\{\alpha_k, \alpha_{k+1}, \alpha_{k+2}, ...\}$ and $\gamma = \inf\{\beta_1, \beta_2, \beta_3, ..\} = \lim_{n\to\infty} \beta_n$. We call γ the upper limit of $(\alpha_n)_{n=1}^{\infty}$ and write

$$\gamma = \limsup_{n\to\infty} \alpha_n.$$

Note that

$$\gamma = \lim_{n\to\infty} \alpha_n$$

if the limit exists.

Given measurable functions $f_n : X \to [0,\infty]$, $n = 1, 2, ...$, the function $\liminf_{n\to\infty} f_n$ is measurable. In particular, if

$$f(x) = \lim_{n\to\infty} f_n(x)$$

exists for every $x \in X$, then f is measurable.

Theorem 2.1.6. (Fatou's Lemma) *If* $f_n : X \to [0,\infty]$, $n = 1, 2, ...,$ *are measurable*

$$\int_X \liminf_{n\to\infty} f_n d\mu \le \liminf_{n\to\infty} \int_X f_n d\mu.$$

PROOF. Introduce

$$g_k = \inf_{n\ge k} f_n.$$

The definition gives that $g_k \uparrow \liminf_{n\to\infty} f_n$ and, moreover,

$$\int_X g_k d\mu \le \int_X f_n d\mu,\ n \ge k$$

and

$$\int_X g_k d\mu \le \inf_{n\ge k} \int_X f_n d\mu.$$

The Fatou Lemma now follows by monotone convergence.

Below we often write

$$\int_E f(x) d\mu(x)$$

instead of

$$\int_E f d\mu.$$

Example 2.1.3. Suppose $a \in \mathbf{R}$ and $f : (\mathbf{R},\mathcal{R}^-) \to ([0,\infty], \mathcal{R}_{0,\infty})$ is measurable. We claim that

$$\int_{\mathbf{R}} f(x+a) dm(x) = \int_{\mathbf{R}} f(x) dm(x).$$

First if $f = \chi_A$, where $A \in \mathcal{R}^-$,

$$\int_{\mathbf{R}} f(x+a)dm(x) = \int_{\mathbf{R}} \chi_{A-a}(x)dm(x) = m(A-a) =$$

$$m(A) = \int_{\mathbf{R}} f(x)dm(x).$$

Next it is clear that the relation we want to prove is true for simple measurable functions and finally, we use the Monotone Convergence Theorem to deduce the general case.

Example 2.1.3, Suppose $\Sigma_1^\infty a_n$ is a positive convergent series and let E be the set of all $x \in [0,1]$ such that

$$\min_{p\in\{0,\ldots,n\}} \mid x - \frac{p}{n} \mid < \frac{a_n}{n}$$

for infinitely many $n \in \mathbf{N}_+$. We claim that E is a Lebesgue null set.

To prove this claim for fixed $n \in \mathbf{N}_+$, let E_n be the set of all $x \in [0,1]$ such that

$$\min_{p\in\mathbf{N}_+} \mid x - \frac{p}{n} \mid < \frac{a_n}{n}.$$

Then if $B(x,r) =]x-r, x+r[\,,\ x \in [0,1]\,,\ r > 0$, we have

$$E_n \subseteq \bigcup_{p=0}^{n} B(\frac{p}{n}, \frac{a_n}{n})$$

and

$$m(E_n) \leq (n+1)\frac{2a_n}{n} \leq 4a_n.$$

Hence

$$\sum_1^\infty m(E_n) < \infty$$

and by the Beppo Levi theorem

$$\int_0^1 \sum_1^\infty \chi_{E_n} dm < \infty.$$

Accordingly from this the set

$$F = \left\{ x \in [0,1]\,;\ \sum_{1}^{\infty} \chi_{E_n}(x) < \infty \right\}$$

is of Lebesgue measure 1. Since $E \subseteq [0,1] \setminus F$ we have $m(E) = 0$.

Exercises

1. Suppose $f_n : X \to [0,\infty]$, $n = 1, 2, ...$, are measurable and

$$\Sigma_{n=1}^{\infty} \mu(f_n > 1) < \infty.$$

Prove that

$$\left\{ \limsup_{n\to\infty} f_n > 1 \right\} \in \mathcal{Z}_\mu\ .$$

2. Set $f_n = n^2 \chi_{\left[0,\frac{1}{n}\right]}$, $n \in \mathbf{N}_+$. Prove that

$$\int_{\mathbf{R}} \liminf_{n\to\infty} f_n dm = 0 < \infty = \liminf_{n\to\infty} \int_{\mathbf{R}} f_n dm$$

(the inequality in the Fatou Lemma may be strict).

3. Suppose $f : (\mathbf{R}, \mathcal{R}^-) \to ([0,\infty], \mathcal{R}_{0,\infty})$ is measurable and set

$$g(x) = \Sigma_{k=1}^{\infty} f(x+k),\ x \in \mathbf{R}.$$

Show that

$$\int_{\mathbf{R}} g dm < \infty \text{ if and only if } \{f > 0\} \in \mathcal{Z}_m.$$

4. Let $(X, \mathcal{M}, \mu)$ be a positive measure space and $f : X \to [0,\infty]$ an $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable function such that

$$f(X) \subseteq \mathbf{N}$$

and

$$\int_X f d\mu < \infty.$$

For every $t \geq 0$, set

$$F(t) = \mu(f > t) \text{ and } G(t) = \mu(f \geq t).$$

Prove that

$$\int_X f d\mu = \Sigma_{n=0}^{\infty} F(n) = \Sigma_{n=1}^{\infty} G(n).$$

Unit 13

Introduction

In the first section of this chapter we collect some basic results on metric spaces, which every mathematician must know about. Section 3.2 gives a version of the Riesz Representation Theorem, which leads to another and perhaps simpler approach to Lebesgue measure than the Carathéodory Theorem. A reader can skip Section 3.2 without losing the continuity in this paper. The chapter also treats so called product measures and Stieltjes integrals.

3.1. Metric Spaces

The construction of our most important measures requires topological concepts. For our purpose it will be enough to restrict ourselves to so called metric spaces.

A metric d on a set X is a mapping $d : X \times X \to [0, \infty[$ such that

(a) $d(x, y) = 0$ if and only if $x = y$
(b) $d(x, y) = d(y, x)$ (symmetry)
(c) $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality).

Here recall, if $A_1, ..., A_n$ are sets,

$$A_1 \times ... \times A_n = \{(x_1, ..., x_n);\ x_i \in A_i \text{ for all } i = 1, ..., n\}$$

A set X equipped with a metric d is called a metric space. Sometimes we write $X = (X, d)$ to emphasize the metric d. If E is a subset of the metric

space (X,d), the function $d_{|E\times E}(x,y)=d(x,y)$, if $x,y\in E$, is a metric on E. Thus $(E,d_{|E\times E})$ is a metric space.

The function $\varphi(t)=\min(1,t)$, $t\geq 0$, satisfies the inequality

$$\varphi(s+t)\leq\varphi(s)+\varphi(t).$$

Therefore, if d is a metric on X, $\min(1,d)$ is a metric on X. The metric $\min(1,d)$ is a bounded metric.

The set $\mathbf{R}$ equipped with the metric $d_1(x,y)=\mid x-y\mid$ is a metric space. More generally, $\mathbf{R}^n$ equipped with the metric

$$d_n(x,y)=d_n((x_1,...,x_n),(y_1,...,y_n))=\max_{1\leq k\leq n}\mid x_k-y_k\mid$$

is a metric space. If not otherwise stated, it will always be assumed that $\mathbf{R}^n$ is equipped with this metric.

Let $C[0,T]$ denote the vector space of all real-valued continuous functions on the interval $[0,T]$, where $T>0$. Then

$$d_\infty(x,y)=\max_{0\leq t\leq T}\mid x(t)-y(t)\mid$$

is a metric on $C[0,T]$.

If (X_k,e_k), $k=1,...,n$, are metric spaces,

$$d(x,y)=\max_{1\leq k\leq n}e_k(x_k,y_k),\ x=(x_1,...,x_n)\ ,y=(y_1,...,y_n)$$

is a metric on $X_1\times...\times X_n$. The metric d is called the product metric on $X_1\times...\times X_n$.

If $X=(X,d)$ is a metric space and $x\in X$ and $r>0$, the open ball with centre at x and radius r is the set $B(x,r)=\{y\in X;d(y,x)<r\}$. If $E\subseteq X$ and E is contained in an appropriate open ball in X it is said to be bounded. The diameter of E is, by definition,

$$\text{diam } E=\sup_{x,y\in E}d(x,y)$$

and it follows that E is bounded if and only if diam $E<\infty$. A subset of X which is a union of open balls in X is called open. In particular, an open ball is an open set. The empty set is open since the union of an empty family of sets is empty. An arbitrary union of open sets is open. The class of all

open subsets of X is called the topology of X. The metrics d and $\min(1,d)$ determine the same topology. A subset E of X is said to be closed if its complement E^c relative to X is open. An intersection of closed subsets of X is closed. If $E \subseteq X$, E° denotes the largest open set contained in E and E^- (or $\bar{E}$) the smallest closed set containing E. E° is the interior of E and E^- its closure. The σ-algebra generated by the open sets in X is called the Borel σ-algebra in X and is denoted by $\mathcal{B}(X)$. A positive measure on $\mathcal{B}(X)$ is called a positive Borel measure.

A sequence $(x_n)_{n=1}^\infty$ in X converges to $x \in X$ if

$$\lim_{n\to\infty} d(x_n, x) = 0.$$

If, in addition, the sequence $(x_n)_{n=1}^\infty$ converges to $y \in X$, the inequalities

$$0 \leq d(x,y) \leq d(x_n,x) + d(x_n,y)$$

imply that $y = x$ and the limit point x is unique.

If $E \subseteq X$ and $x \in X$, the following properties are equivalent:

(i) $x \in E^-$.
(ii) $B(x,r) \cap E \neq \phi$, all $r > 0$.
(iii) There is a sequence $(x_n)_{n=1}^\infty$ in E which converges to x.

If $B(x,r) \cap E = \phi$, then $B(x,r)^c$ is a closed set containing E but not x. Thus $x \notin E^-$. This proves that (i)⇒(ii). Conversely, if $x \notin E^-$, since $\bar{E}^c$ is open there exists an open ball $B(y,s)$ such that $x \in B(y,s) \subseteq \bar{E}^c \subseteq E^c$. Now choose $r = s - d(x,y) > 0$ so that $B(x,r) \subseteq B(y,s)$. Then $B(x,r) \cap E = \phi$. This proves (ii)⇒(i).

If (ii) holds choose for each $n \in \mathbf{N}_+$ a point $x_n \in E$ with $d(x_n,x) < \frac{1}{n}$ and (iii) follows. If there exists an $r > 0$ such that $B(x,r) \cap E = \phi$, then (iii) cannot hold. Thus (iii)⇒(ii).

If $E \subseteq X$, the set $E^- \setminus E^\circ$ is called the boundary of E and is denoted by ∂E.

A set $A \subseteq X$ is said to be dense in X if $A^- = X$. The metric space X is called separable if there is an at most denumerable dense subset of X. For example, $\mathbf{Q}^n$ is a dense subset of $\mathbf{R}^n$. The space $\mathbf{R}^n$ is separable.

Theorem 3.1.1. $\mathcal{B}(\mathbf{R}^n) = \mathcal{R}_n$.

PROOF. The σ-algebra $\mathcal{R}_n$ is generated by the open n-cells in $\mathbf{R}^n$ and an open n-cell is an open subset of $\mathbf{R}^n$. Hence $\mathcal{R}_n \subseteq \mathcal{B}(\mathbf{R}^n)$. Let U be an open subset in $\mathbf{R}^n$ and note that an open ball in $\mathbf{R}^n = (\mathbf{R}^n, d_n)$ is an open n-cell. If $x \in U$ there exist an $a \in \mathbf{Q}^n \cap U$ and a rational number $r > 0$ such that $x \in B(a, r) \subseteq U$. Thus U is an at most denumerable union of open n-cells and it follows that $U \in \mathcal{R}_n$. Thus $\mathcal{B}(\mathbf{R}^n) \subseteq \mathcal{R}_n$ and the theorem is proved.

Let $X = (X, d)$ and $Y = (Y, e)$ be two metric spaces. A mapping $f : X \to Y$ (or $f : (X, d) \to (Y, e)$ to emphasize the underlying metrics) is said to be continuous at the point $a \in X$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$x \in B(a, \delta) \Rightarrow f(x) \in B(f(a), \varepsilon).$$

Equivalently this means that for any sequence $(x_n)_{n=1}^{\infty}$ in X which converges to a in X, the sequence $(f(x_n))_{n=1}^{\infty}$ converges to $f(a)$ in Y. If f is continuous at each point of X, the mapping f is called continuous. Stated otherwise this means that

$$f^{-1}(V) \text{ is open if } V \text{ is open}$$

or

$$f^{-1}(F) \text{ is closed if } F \text{ is closed.}$$

The mapping f is said to be Borel measurable if

$$f^{-1}(B) \in \mathcal{B}(X) \text{ if } B \in \mathcal{B}(Y)$$

or, what amounts to the same thing,

$$f^{-1}(V) \in \mathcal{B}(X) \text{ if } V \text{ is open.}$$

A Borel measurable function is sometimes called a Borel function. A continuous function is a Borel function.

Example 3.1.1. Let $f : (\mathbf{R}, d_1) \to (\mathbf{R}, d_1)$ be a continuous strictly increasing function and set $\rho(x, y) = \mid f(x) - f(y) \mid$, $x, y \in \mathbf{R}$. Then ρ is a metric on $\mathbf{R}$.

Define $j(x) = x$, $x \in \mathbf{R}$. The mapping $j : (\mathbf{R},d_1) \to (\mathbf{R},\rho)$ is continuous. We claim that the map $j : (\mathbf{R},\rho) \to (\mathbf{R},d_1)$ is continuous. To see this, let $a \in \mathbf{R}$ and suppose the sequence $(x_n)_{n=1}^{\infty}$ converges to a in the metric space $(\mathbf{R},\rho)$, that is $\mid f(x_n) - f(a) \mid \to 0$ as $n \to \infty$. Let $\varepsilon > 0$. Then

$$f(x_n) - f(a) \geq f(a+\varepsilon) - f(a) > 0 \text{ if } x_n \geq a + \varepsilon$$

and

$$f(a) - f(x_n) \geq f(a) - f(a-\varepsilon) > 0 \text{ if } x_n \leq a - \varepsilon.$$

Thus $x_n \in]a-\varepsilon, a+\varepsilon[$ if n is sufficiently large. This proves that he map $j : (\mathbf{R},\rho) \to (\mathbf{R},d_1)$ is continuous.

The metrics d_1 and ρ determine the same topology and Borel subsets of $\mathbf{R}$.

A mapping $f : (X,d) \to (Y,e)$ is said to be uniformly continuous if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that $e(f(x), f(y)) < \varepsilon$ as soon as $d(x,y) < \delta$.

If $x \in X$ and E, $F \subseteq X$, let

$$d(x,E) = \inf_{u \in E} d(x,u)$$

be the distance from x to E and let

$$d(E,F) = \inf_{u \in E, v \in F} d(u,v)$$

be the distance between E and F. Note that $d(x,E) = 0$ if and only if $x \in \bar{E}$.

If $x, y \in X$ and $u \in E$,

$$d(x,u) \leq d(x,y) + d(y,u)$$

and, hence

$$d(x,E) \leq d(x,y) + d(y,u)$$

and

$$d(x,E) \leq d(x,y) + d(y,E).$$

Next suppose $E \neq \phi$. Then by interchanging the roles of x and y, we get

$$\mid d(x,E) - d(y,E) \mid \leq d(x,y)$$

and conclude that the distance function $d(x, E)$, $x \in X$, is continuous. In fact, it is uniformly continuous. If $x \in X$ and $r > 0$, the so called closed ball $\bar{B}(x, r) = \{y \in X;\ d(y, x) \leq r\}$ is a closed set since the map $y \to d(y, x)$, $y \in X$, is continuous.

If $F \subseteq X$ is closed and $\varepsilon > 0$, the continuous function

$$\Pi^X_{F,\varepsilon} = \max(0, 1 - \frac{1}{\varepsilon} d(\cdot, F))$$

fulfils $0 \leq \Pi^X_{F,\varepsilon} \leq 1$ and $\Pi^X_{F,\varepsilon} = 1$ on F. Furthermore, $\Pi^X_{F,\varepsilon}(a) > 0$ if and only if $a \in F_\varepsilon =_{def} \{x \in X;\ d(x, F) < \varepsilon\}$. Thus

$$\chi_F \leq \Pi^X_{F,\varepsilon} \leq \chi_{F_\varepsilon}.$$

Let $X = (X, d)$ be a metric space. A sequence $(x_n)_{n=1}^\infty$ in X is called a Cauchy sequence if to each $\varepsilon > 0$ there exists a positive integer p such that $d(x_n, x_m) < \varepsilon$ for all $n, m \geq p$. If a Cauchy sequence $(x_n)_{n=1}^\infty$ contains a convergent subsequence $(x_{n_k})_{k=1}^\infty$ it must be convergent. To prove this claim, suppose the subsequence $(x_{n_k})_{k=1}^\infty$ converges to a point $x \in X$. Then

$$d(x_m, x) \leq d(x_m, x_{n_k}) + d(x_{n_k}, x)$$

can be made arbitrarily small for all sufficiently large m by choosing k sufficiently large. Thus $(x_n)_{n=1}^\infty$ converges to x.

A subset E of X is said to be complete if every Cauchy sequence in E converges to a point in E. If $E \subseteq X$ is closed and X is complete it is clear that E is complete. Conversely, if X is a metric space and a subset E of X is complete, then E is closed.

It is important to know that $\mathbf{R}$ is complete equipped with its standard metric. To see this let $(x_n)_{n=1}^\infty$ be a Cauchy sequence. There exists a positive integer such that $\mid x_n - x_m \mid < 1$ if $n, m \geq p$. Therefore

$$\mid x_n \mid \leq \mid x_n - x_p \mid + \mid x_p \mid \leq 1 + \mid x_p \mid$$

for all $n \geq p$. We have proved that the sequence $(x_n)_{n=1}^\infty$ is bounded (the reader can check that every Cauchy sequence in a metric space has this property). Now define

$$a = \sup \{x \in \mathbf{R};\ \text{there are only finitely many } n \text{ with } x_n \leq x\}.$$

The definition implies that there exists a subsequence $(x_{n_k})_{k=1}^\infty$, which converges to a (since for any $r > 0$, $x_n \in B(a, r)$ for infinitely many n). The

original sequence is therefore convergent and we conclude that $\mathbf{R}$ is complete (equipped with its standard metric d_1). It is simple to prove that the product of n complete spaces is complete and we conclude that $\mathbf{R}^n$ is complete.

Let $E \subseteq X$. A family $(V_i)_{i\in I}$ of subsets of X is said to be a cover of E if $\cup_{i\in I} V_i \supseteq E$ and E is said to be covered by the $V_i's$. The cover $(V_i)_{i\in I}$ is said to be an open cover if each member V_i is open. The set E is said to be totally bounded if, for every $\varepsilon > 0$, E can be covered by finitely many open balls of radius ε. A subset of a totally bounded set is totally bounded.

The following definition is especially important.

Definition 3.1.1. A subset E of a metric space X is said to be compact if to every open cover $(V_i)_{i\in I}$ of E, there is a finite subcover of E, which means there is a finite subset J of I such that $(V_i)_{i\in J}$ is a cover of E.

If K is closed, $K \subseteq E$, and E is compact, then K is compact. To see this, let $(V_i)_{i\in I}$ be an open cover of K. This cover, augmented by the set $X \setminus K$ is an open cover of E and has a finite subcover since E is compact. Noting that $K \cap (X \setminus K) = \phi$, the assertion follows.

Theorem 3.1.2. *The following conditions are equivalent:*

(a) *E is complete and totally bounded.*

(b) *Every sequence in E contains a subsequence which converges to a point of E.*

(c) *E is compact.*

PROOF. (a)⇒(b). Suppose $(x_n)_{n=1}^{\infty}$ is a sequence in E. The set E can be covered by finitely many open balls of radius 2^{-1} and at least one of them must contain x_n for infinitely many $n \in \mathbf{N}_+$. Suppose $x_n \in B(a_1, 2^{-1})$ if $n \in N_1 \subseteq N_0 =_{def} \mathbf{N}_+$, where N_1 is infinite. Next $E \cap B(a_1, 2^{-1})$ can be covered by finitely many balls of radius 2^{-2} and at least one of them must contain x_n for infinitely many $n \in N_1$. Suppose $x_n \in B(a_2, 2^{-1})$ if $n \in N_2$, where $N_2 \subseteq N_1$ is infinite. By induction, we get open balls $B(a_j, 2^{-j})$ and infinite sets $N_j \subseteq N_{j-1}$ such that $x_n \in B(a_j, 2^{-j})$ for all $n \in N_j$ and $j \geq 1$.

Let $n_1 < n_2 < ...$, where $n_k \in N_k$, $k = 1, 2, ...$. The sequence $(x_{n_k})_{k=1}^{\infty}$ is a Cauchy sequence, and since E is complete it converges to a point of E .

(b)⇒(a). If E is not complete there is a Cauchy sequence in E with no limit in E. Therefore no subsequence can converge in E, which contradicts (b). On the other hand if E is not totally bounded, there is an $\varepsilon > 0$ such that E cannot be covered by finitely many balls of radius ε. Let $x_1 \in E$ be arbitrary. Having chosen $x_1, ..., x_{n-1}$, pick $x_n \in E \setminus \cup_{i=1}^{n-1} B(x_i, \varepsilon)$, and so on. The sequence $(x_n)_{n=1}^{\infty}$ cannot contain any convergent subsequence as $d(x_n, x_m) \geq \varepsilon$ if $n \neq m$, which contradicts (b).

{(a) and (b)} ⇒(c). Let $(V_i)_{i \in I}$ be an open cover of E. Since E is totally bounded it is enough to show that there is an $\varepsilon > 0$ such that any open ball of radius ε which intersects E is contained in some V_i. Suppose on the contrary that for every $n \in \mathbf{N}_+$ there is an open ball B_n of radius $\leq 2^{-n}$ which intersects E and is contained in no V_i. Choose $x_n \in B_n \cap E$ and assume without loss of generality that $(x_n)_{n=1}^{\infty}$ converges to some point x in E by eventually going to a subsequence. Suppose $x \in V_{i_0}$ and choose $r > 0$ such that $B(x, r) \subseteq V_{i_0}$. But then $B_n \subseteq B(x, r) \subseteq V_{i_0}$ for large n, which contradicts the assumption on B_n.

(c)⇒(b). If $(x_n)_{n=1}^{\infty}$ is a sequence in E with no convergent subsequence in E, then for every $x \in E$ there is an open ball $B(x, r_x)$ which contains x_n for only finitely many n. Then $(B(x, r_x))_{x \in E}$ is an open cover of E without a finite subcover.

Theorem 3.2.1. (The Riesz Representation Theorem) *Suppose X is a compact metric space and let T be a positive linear functional on $C(X)$. Then there exists a unique finite positive Borel measure μ in X with the following properties:*

(a)

$$Tf = \int_X f d\mu, \ f \in C(X).$$

(b) *For every $E \in \mathcal{B}(X)$*

$$\mu(E) = \sup_{\substack{K \subseteq E \\ K \text{ compact}}} \mu(K).$$

(c) *For every $E \in \mathcal{B}(X)$*

$$\mu(E) = \inf_{\substack{V \supseteq E \\ V \text{ open}}} \mu(V).$$

The property (c) is a consequence of (b), since for each $E \in \mathcal{B}(X)$ and $\varepsilon > 0$ there is a compact $K \subseteq X \setminus E$ such that

$$\mu(X \setminus E) < \mu(K) + \varepsilon.$$

But then

$$\mu(X \setminus K) < \mu(E) + \varepsilon$$

and $X \setminus K$ is open and contains E. In a similar way, (b) follows from (c) since X is compact.

The proof of the Riesz Representation Theorem depends on properties of continuous functions of independent interest. Suppose $K \subseteq X$ is compact and $V \subseteq X$ is open. If $f : X \to [0, 1]$ is a continuous function such that

$$f \leq \chi_V \text{ and } \text{supp} f \subseteq V$$

we write

$$f \prec V$$

and if

$$\chi_K \leq f \leq \chi_V \text{ and } \text{supp} f \subseteq V$$

we write

$$K \prec f \prec V.$$

Theorem 3.2.2. *Let K be compact subset X.*

(a) *Suppose $K \subseteq V$ where V is open. There exists a function f on X such that*

$$K \prec f \prec V.$$

(b) *Suppose X is compact and $K \subseteq V_1 \cup ... \cup V_n$, where K is compact and $V_1, ..., V_n$ are open. There exist functions $h_1, ..., h_n$ on X such that*

$$h_i \prec V_i, \ i = 1, ..., n$$

and

$$h_1 + ... + h_n = 1 \text{ on } K.$$

PROOF. (a) Suppose $\varepsilon = \frac{1}{2}\min_K d(\cdot, V^c)$. By Corollary 3.1.2, $\varepsilon > 0$. The continuous function $f = \Pi^X_{K,\varepsilon}$ satisfies $\chi_K \leq f \leq \chi_{K_\varepsilon}$, that is $K \prec f \prec K_\varepsilon$. Part (a) follows if we note that the closure $(K_\varepsilon)^-$ of K_ε is contained in V.

(b) For each $x \in K$ there exists an $r_x > 0$ such that $B(x, r_x) \subseteq V_i$ for some i. Let $U_x = B(x, \frac{1}{2}r_x)$. It is important to note that $(U_x)^- \subseteq V_i$ and $(U_x)^-$ is compact since X is compact. There exist points $x_1, ..., x_m \in K$ such that $\cup_{j=1}^m U_{x_i} \supseteq K$. If $1 \leq i \leq n$, let F_i denote the union of those $(U_{x_j})^-$ which are contained in V_i. By Part (a), there exist continuous functions f_i such that $F_i \prec f_i \prec V_i$, $i = 1, ..., n$. Define

$$\begin{aligned} h_1 &= f_1 \\ h_2 &= (1 - f_1)f_2 \\ &.... \\ h_n &= (1 - f_1)...(1 - f_{n-1})f_n. \end{aligned}$$

Clearly, $h_i \prec V_i$, $i = 1, ..., n$. Moreover, by induction, we get

$$h_1 + ... + h_n = 1 - (1 - f_1)...(1 - f_{n-1})(1 - f_n).$$

Since $\cup_{i=1}^n F_i \supseteq K$ we are done.

The uniqueness in Theorem 3.2.1 is simple to prove. Suppose μ_1 and μ_2 are two measures for which the theorem holds. Fix $\varepsilon > 0$ and compact $K \subseteq X$ and choose an open set V so that $\mu_2(V) \leq \mu_2(K) + \varepsilon$. If $K \prec f \prec V$,

$$\mu_1(K) = \int_X \chi_K d\mu_1 \leq \int_X f d\mu_1 = Tf$$

$$= \int_X f d\mu_2 \leq \int_X \chi_V d\mu_2 = \mu_2(V) \leq \mu_2(K) + \varepsilon.$$

Thus $\mu_1(K) \leq \mu_2(K)$. If we interchange the roles of the two measures, the opposite inequality is obtained, and the uniqueness of μ follows.

To prove the existence of the measure μ in Theorem 3.2.1, define for every open V in X,

$$\mu(V) = \sup_{f \prec V} Tf.$$

Here $\mu(\phi) = 0$ since the supremum over the empty set, by convention, equals 0. Note also that $\mu(X) = T1$. Moreover, $\mu(V_1) \leq \mu(V_2)$ if V_1 and V_2 are open and $V_1 \subseteq V_2$. Now set

$$\mu(E) = \inf_{\substack{V \supseteq E \\ V \text{ open}}} \mu(V) \text{ if } E \in \mathcal{B}(X).$$

Clearly, $\mu(E_1) \leq \mu(E_2)$, if $E_1 \subseteq E_2$ and $E_{1,}E_2 \in \mathcal{B}(X)$. We therefore say that μ is increasing.

Lemma 3.2.1. (a) *If $V_1, ..., V_n$ are open,*

$$\mu(\cup_{i=1}^n V_i) \leq \Sigma_{i=1}^n \mu(V_i).$$

(b) *If $E_1, E_2, ... \in \mathcal{B}(X)$,*

$$\mu(\cup_{i=1}^\infty E_i) \leq \Sigma_{i=1}^\infty \mu(E_i).$$

(c) *If $K_1, ..., K_n$ are compact and pairwise disjoint,*

$$\mu(\cup_{i=1}^n K_i) = \Sigma_{i=1}^n \mu(K_i).$$

PROOF. (a) It is enough to prove (a) for $n = 2$. To this end first choose $g \prec V_1 \cup V_2$ and then $h_i \prec V_i$, $i = 1, 2$, such that $h_1 + h_2 = 1$ on supp g. Then

$$g = h_1 g + h_2 g$$

and it follows that

$$Tg = T(h_1 g) + T(h_2 g) \leq \mu(V_1) + \mu(V_2).$$

Thus

$$\mu(V_1 \cup V_2) \leq \mu(V_1) + \mu(V_2).$$

(b) Choose $\varepsilon > 0$ and for each $i \in \mathbf{N}_+$, choose an open $V_i \supseteq E_i$ such $\mu(V_i) < \mu(E_i) + 2^{-i}\varepsilon$. Set $V = \cup_{i=1}^{\infty} V_i$ and choose $f \prec V$. Since supp$f$ is compact, $f \prec V_1 \cup ... \cup V_n$ for some n. Thus, by Part (a),

$$Tf \leq \mu(V_1 \cup ... \cup V_n) \leq \Sigma_{i=1}^{n} \mu(V_i) \leq \Sigma_{i=1}^{\infty} \mu(E_i) + \varepsilon$$

and we get

$$\mu(V) \leq \Sigma_{i=1}^{\infty} \mu(E_i)$$

since $\varepsilon > 0$ is arbitrary. But $\cup_{i=1}^{\infty} E_i \subseteq V$ and it follows that

$$\mu(\cup_{i=1}^{\infty} E_i) \leq \Sigma_{i=1}^{\infty} \mu(E_i).$$

(c) It is enough to treat the special case $n = 2$. Choose $\varepsilon > 0$. Set $\rho = d(K_1, K_2)$ and $V_1 = (K_1)_{\rho/2}$ and $V_2 = (K_2)_{\rho/2}$. There is an open set $U \supseteq K_1 \cup K_2$ such that $\mu(U) < \mu(K_1 \cup K_2) + \varepsilon$ and there are functions $f_i \prec U \cap V_i$ such that $Tf_i > \mu(U \cap V_i) - \varepsilon$ for $i = 1, 2$. Now, using that μ increases

$$\mu(K_1) + \mu(K_2) \leq \mu(U \cap V_1) + \mu(U \cap V_2)$$

$$\leq Tf_1 + Tf_2 + 2\varepsilon = T(f_1 + f_2) + 2\varepsilon.$$

Since $f_1 + f_2 \prec U$,

$$\mu(K_1) + \mu(K_2) \leq \mu(U) + 2\varepsilon \leq \mu(K_1 \cup K_2) + 3\varepsilon$$

and, by letting $\varepsilon \to 0$,

$$\mu(K_1) + \mu(K_2) \leq \mu(K_1 \cup K_2).$$

The reverse inequality follows from Part (b). The lemma is proved.

Next we introduce the class

$$\mathcal{M} = \left\{ E \in \mathcal{B}(X);\ \mu(E) = \sup_{\substack{K \subseteq E \\ K \text{ compact}}} \mu(K) \right\}$$

Since μ is increasing $\mathcal{M}$ contains every compact set. Recall that a closed set in X is compact, since X is compact. Especially, note that ϕ and $X \in \mathcal{M}$.

COMPLETION OF THE PROOF OF THEOREM 3.2.1:

CLAIM 1. $\mathcal{M}$ contains every open set.

PROOF OF CLAIM 1. Let V be open and suppose $\alpha < \mu(V)$. There exists an $f \prec V$ such that $\alpha < Tf$. If U is open and $U \supseteq K =_{def} \text{supp} f$, then $f \prec U$, and hence $Tf \leq \mu(U)$. But then $Tf \leq \mu(K)$. Thus $\alpha < \mu(K)$ and Claim 1 follows since K is compact and $K \subseteq V$.

CLAIM 2. Let $(E_i)_{i=1}^{\infty}$ be a disjoint denumerable collection of members of $\mathcal{M}$ and put $E = \cup_{i=1}^{\infty} E_i$. Then

$$\mu(E) = \Sigma_{i=1}^{\infty} \mu(E_i)$$

and $E \in \mathcal{M}$.

PROOF OF CLAIM 2. Choose $\varepsilon > 0$ and for each $i \in \mathbf{N}_+$, choose a compact $K_i \subseteq E_i$ such that $\mu(K_i) > \mu(E_i) - 2^{-i}\varepsilon$. Set $H_n = K_1 \cup ... \cup K_n$. Then, by Lemma 3.2.1 (c),

$$\mu(E) \geq \mu(H_n) = \Sigma_{i=1}^{n} \mu(K_i) > \Sigma_{i=1}^{n} \mu(E_i) - \varepsilon$$

and we get

$$\mu(E) \geq \Sigma_{i=1}^{\infty} \mu(E_i).$$

Thus, by Lemma 3.2.1 (b), $\mu(E) = \Sigma_{i=1}^{\infty} \mu(E_i)$. To prove that $E \in \mathcal{M}$, let ε be as in the very first part of the proof and choose n such that

$$\mu(E) \leq \Sigma_{i=1}^{n} \mu(E_i) + \varepsilon.$$

Then

$$\mu(E) < \mu(H_n) + 2\varepsilon$$

and this shows that $E \in \mathcal{M}$.

CLAIM 3. Suppose $E \in \mathcal{M}$ and $\varepsilon > 0$. Then there exist a compact K and an open V such that $K \subseteq E \subseteq V$ and $\mu(V \setminus K) < \varepsilon$.

PROOF OF CLAIM 3. The definitions show that there exist a compact K and an open V such that

$$\mu(V) - \frac{\varepsilon}{2} < \mu(E) < \mu(K) + \frac{\varepsilon}{2}.$$

The set $V \setminus K$ is open and $V \setminus K \in \mathcal{M}$ by Claim 1. Thus Claim 2 implies that

$$\mu(K) + \mu(V \setminus K) = \mu(V) < \mu(K) + \varepsilon$$

and we get $\mu(V \setminus K) < \varepsilon$.

CLAIM 4. If $A \in \mathcal{M}$, then $X \setminus A \in \mathcal{M}$.

PROOF OF CLAIM 4. Choose $\varepsilon > 0$. Furthermore, choose compact $K \subseteq A$ and open $V \supseteq A$ such that $\mu(V \setminus K) < \varepsilon$. Then

$$X \setminus A \subseteq (V \setminus K) \cup (X \setminus V).$$

Now, by Lemma 3.2.1 (b),

$$\mu(X \setminus A) \leq \varepsilon + \mu(X \setminus V).$$

Since $X \setminus V$ is a compact subset of $X \setminus A$, we conclude that $X \setminus A \in \mathcal{M}$.

Claims 1, 2 and 4 prove that $\mathcal{M}$ is a σ-algebra which contains all Borel sets. Thus $\mathcal{M} = \mathcal{B}(X)$.

We finally prove (a). It is enough to show that

$$Tf \leq \int_X f d\mu$$

for each $f \in C(X)$. For once this is known

$$-Tf = T(-f) \leq \int_X -f d\mu \leq -\int_X f d\mu$$

and (a) follows.

Choose $\varepsilon > 0$. Set $f(X) = [a, b]$ and choose $y_0 < y_1 < ... < y_n$ such that $y_1 = a$, $y_{n-1} = b$, and $y_i - y_{i-1} < \varepsilon$. The sets

$$E_i = f^{-1}([y_{i-1}, y_i[), \ i = 1, ..., n$$

constitute a disjoint collection of Borel sets with the union X. Now, for each i, pick an open set $V_i \supseteq E_i$ such that $\mu(V_i) \leq \mu(E_i) + \frac{\varepsilon}{n}$ and $V_i \subseteq f^{-1}(]-\infty, y_i[)$. By Theorem 3.2.2 there are functions $h_i \prec V_i$, $i = 1, ..., n$, such that $\Sigma_{i=1}^n h_i = 1$ on $\text{supp} f$ and $h_i f \prec y_i h_i$ for all i. From this we get

$$\begin{aligned} Tf = \Sigma_{i=1}^n T(h_i f) &\leq \Sigma_{i=1}^n y_i T h_i \leq \Sigma_{i=1}^n y_i \mu(V_i) \\ &\leq \Sigma_{i=1}^n y_i \mu(E_i) + \Sigma_{i=1}^n y_i \frac{\varepsilon}{n} \\ &\leq \Sigma_{i=1}^n (y_i - \varepsilon)\mu(E_i) + \varepsilon\mu(X) + (b + \varepsilon)\varepsilon \\ &\leq \Sigma_{i=1}^n \int_{E_i} f d\mu + \varepsilon\mu(X) + (b + \varepsilon)\varepsilon \\ &= \int_X f d\mu + \varepsilon\mu(X) + (b + \varepsilon)\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we get

$$Tf \leq \int_X f d\mu.$$

This proves Theorem 3.2.1.

It is now simple to show the existence of volume measure in $\mathbf{R}^n$. For pedagogical reasons we first discuss the so called volume measure in the unit cube $Q = [0, 1]^n$ in $\mathbf{R}^n$.

The Riemann integral

$$\int_Q f(x)dx,$$

is a positive linear functional as a function of $f \in C(Q)$. Moreover, $T1 = 1$ and the Riesz Representation Theorem gives us a Borel probability measure μ in Q such that

$$\int_Q f(x)dx = \int_Q f d\mu.$$

Suppose $A \subseteq Q$ is a closed n-cell and $i \in \mathbf{N}_+$. Then

$$\mathrm{vol}(A) \leq \int_Q \Pi^Q_{A,2^{-i}}(x)dx \leq \mathrm{vol}(A_{2^{-i}})$$

and

$$\Pi^Q_{A,2^{-i}}(x) \to \chi_A(x) \text{ as } i \to \infty$$

for every $x \in \mathbf{R}^n$. Thus

$$\mu(A) = \mathrm{vol}(A).$$

The measure μ is called the volume measure in the unit cube. In the special case $n = 2$ it is called the area measure in the unit square and if $n = 1$ it is called the linear measure in the unit interval.

PROOF OF THEOREM 1.1.1. Let $\hat{\mathbf{R}}$=$\mathbf{R}\cup\{-\infty, \infty\}$ be the two-point compactification of $\mathbf{R}$ introduced in Example 3.1.3 and let $\hat{\mathbf{R}}^n$ denote the product of n copies of the metric space $\hat{\mathbf{R}}$. Clearly,

$$\mathcal{B}(\mathbf{R}^n) = \left\{A \cap \mathbf{R}^n;\ A \in \mathcal{B}(\hat{\mathbf{R}}^n)\right\}.$$

Moreover, let $w : \mathbf{R}^n \to]0, \infty[$ be a continuous map such that

$$\int_{\mathbf{R}^n} w(x)dx = 1.$$

Now we define

$$Tf = \int_{\mathbf{R}^n} f(x)w(x)dx,\ f \in C(\hat{\mathbf{R}}^n).$$

Note that $T1 = 1$. The function T is a positive linear functional on $C(\hat{\mathbf{R}}^n)$ and the Riesz Representation Theorem gives us a Borel probability measure μ on $\hat{\mathbf{R}}^n$ such that

$$\int_{\mathbf{R}^n} f(x)w(x)dx = \int_{\hat{\mathbf{R}}^n} f d\mu,\ f \in C(\hat{\mathbf{R}}^n).$$

As above we get

$$\int_A w(x)dx = \mu(A)$$

for each compact n-cell in $\mathbf{R}^n$. Thus

$$\mu(\mathbf{R}^n) = \lim_{i\to\infty} \int_{[-i,i]^n} w(x)dx = 1$$

and we conclude that μ is concentrated on $\mathbf{R}^n$. Set $\mu_0(A) = \mu(A)$, $A \in \mathcal{B}(\mathbf{R}^n)$, and

$$dm_n = \frac{1}{w} d\mu_0.$$

Then, if $f \in C_c(\mathbf{R}^n)$,

$$\int_{\mathbf{R}^n} f(x)w(x)dx = \int_{\mathbf{R}^n} f d\mu_0$$

and by replacing f by f/w,

$$\int_{\mathbf{R}^n} f(x)dx = \int_{\mathbf{R}^n} f dm_n.$$

From this $m_n(A) =$vol(A) for every compact n-cell A and it follows that m_n is the volume measure on $\mathbf{R}^n$. Theorem 1.1.1 is proved.

3.4. Product Measures

Suppose $(X, \mathcal{M})$ and $(Y, \mathcal{N})$ are two measurable spaces. If $A \in \mathcal{M}$ and $B \in \mathcal{N}$, the set $A \times B$ is called a measurable rectangle in $X \times Y$. The product σ-algebra $\mathcal{M} \otimes \mathcal{N}$ is, by definition, the σ-algebra generated by all measurable rectangles in $X \times Y$. If we introduce the projections

$$\pi_X(x, y) = x, \ (x, y) \in X \times Y$$

and

$$\pi_Y(x, y) = y, \ (x, y) \in X \times Y,$$

the product σ-algebra $\mathcal{M} \otimes \mathcal{N}$ is the least σ-algebra $\mathcal{S}$ of subsets of $X \times Y$, which makes the maps $\pi_X : (X \times Y, \mathcal{S}) \to (X, \mathcal{M})$ and $\pi_Y : (X \times Y, \mathcal{S}) \to (Y, \mathcal{N})$ measurable, that is $\mathcal{M} \otimes \mathcal{N} = \sigma(\pi_X^{-1}(\mathcal{M}) \cup \pi_Y^{-1}(\mathcal{N}))$..

Suppose $\mathcal{E}$ generates $\mathcal{M}$, where $X \in \mathcal{E}$, and $\mathcal{F}$ generates $\mathcal{N}$, where $Y \in \mathcal{F}$. We claim that the class

$$\mathcal{E} \boxtimes \mathcal{F} = \{E \times F; E \in \mathcal{E} \text{ and } F \in \mathcal{F}\}$$

generates the σ-algebra $\mathcal{M} \otimes \mathcal{N}$. First it is clear that

$$\sigma(\mathcal{E} \boxtimes \mathcal{F}) \subseteq \mathcal{M} \otimes \mathcal{N}.$$

Moreover, the class

$$\{E \in \mathcal{M}; E \times Y \in \sigma(\mathcal{E} \boxtimes \mathcal{F}) \} = \mathcal{M} \cap \left\{E \subseteq X;\ \pi_X^{-1}(E) \in \sigma(\mathcal{E} \boxtimes \mathcal{F}) \right\}$$

is a σ-algebra, which contains $\mathcal{E}$ and therefore equals $\mathcal{M}$. Thus $A \times Y \in \sigma(\mathcal{E} \boxtimes \mathcal{F})$ for all $A \in \mathcal{M}$ and, in a similar way, $X \times B \in \sigma(\mathcal{E} \boxtimes \mathcal{F})$ for all $B \in \mathcal{N}$ and we conclude that $A \times B = (A \times Y) \cap (X \times B) \in \sigma(\mathcal{E} \boxtimes \mathcal{F})$ for all $A \in \mathcal{M}$ and all $B \in \mathcal{N}$. This proves that

$$\mathcal{M} \otimes \mathcal{N} \subseteq \sigma(\mathcal{E} \boxtimes \mathcal{F})$$

and it follows that

$$\sigma(\mathcal{E} \boxtimes \mathcal{F}) = \mathcal{M} \otimes \mathcal{N}.$$

Thus

$$\sigma(\mathcal{E} \boxtimes \mathcal{F}) = \sigma(\mathcal{E}) \otimes \sigma(\mathcal{F}) \text{ if } X \in \mathcal{E} \text{ and } Y \in \mathcal{F}.$$

Since the σ-algebra $\mathcal{R}_n$ is generated by all open n-cells in $\mathbf{R}^n$, we conclude that

$$\mathcal{R}_{k+n} = \mathcal{R}_k \otimes \mathcal{R}_n.$$

Given $E \subseteq X \times Y$, define

$$E_x = \{y;\ (x, y) \in E\} \text{ if } x \in X$$

and

$$E^y = \{x;\ (x, y) \in E\} \text{ if } y \in Y.$$

If $f : X \times Y \to Z$ is a function and $x \in X$, $y \in Y$, let

$$f_x(y) = f(x, y), \text{ if } y \in Y$$

and

$$f^y(x) = f(x, y), \text{ if } x \in X.$$

Theorem 3.4.1 (a) *If* $E \in \mathcal{M} \otimes \mathcal{N}$, *then* $E_x \in \mathcal{N}$ and $E^y \in \mathcal{M}$ *for every* $x \in X$ *and* $y \in Y$.

(b) *If* $f : (X \times Y, \mathcal{M} \otimes \mathcal{N}) \to (Z, \mathcal{O})$ *is measurable, then* f_x *is* $(\mathcal{N}, \mathcal{O})$*-measurable for each* $x \in X$ *and* f^y *is* $(\mathcal{M}, \mathcal{O})$*-measurable for each* $y \in Y$.

Proof. (a) Choose $y \in Y$ and define $\varphi : X \to X \times Y$ by $\varphi(x) = (x, y)$. Then

$$\mathcal{M} = \sigma(\varphi^{-1}(\mathcal{M} \boxtimes \mathcal{N})) = \varphi^{-1}(\sigma(\mathcal{M} \boxtimes \mathcal{N})) = \varphi^{-1}(\mathcal{M} \otimes \mathcal{N})$$

and it follows that $E^y \in \mathcal{M}$. In a similar way $E_x \in \mathcal{N}$ for every $x \in X$.

(b) For any set $V \in \mathcal{O}$,

$$(f^{-1}(V))_x = (f_x)^{-1}(V)$$

and

$$(f^{-1}(V))^y = (f^y)^{-1}(V).$$

Part (b) now follows from (a).

Below an $(\mathcal{M}, \mathcal{R}_{0,\infty})$-measurable or $(\mathcal{M}, \mathcal{R})$-measurable function is simply called $\mathcal{M}$-measurable.

Theorem 3.4.2. *Suppose* $(X, \mathcal{M}, \mu)$ *and* $(Y, \mathcal{N}, \nu)$ *are positive* σ*-finite measurable spaces and suppose* $E \in \mathcal{M} \otimes \mathcal{N}$. *If*

$$f(x) = \nu(E_x) \text{ and } g(y) = \mu(E^y)$$

for every $x \in X$ *and* $y \in Y$, *then* f *is* $\mathcal{M}$*-measurable,* g *is* $\mathcal{N}$*-measurable, and*

$$\int_X f d\mu = \int_Y g d\nu.$$

Proof. We first assume that $(X, \mathcal{M}, \mu)$ and $(Y, \mathcal{N}, \nu)$ are finite positive measure spaces.

Let $\mathcal{D}$ be the class of all sets $E \in \mathcal{M} \otimes \mathcal{N}$ for which the conclusion of the theorem holds. It is clear that the class $\mathcal{G}$ of all measurable rectangles in $X \times Y$ is a subset of $\mathcal{D}$ and $\mathcal{G}$ is a π-system. Furthermore, the Beppo Levi Theorem shows that $\mathcal{D}$ is a σ-additive class. Therefore, using Theorem 1.2.2, $\mathcal{M} \otimes \mathcal{N} = \sigma(\mathcal{G}) \subseteq \mathcal{D}$ and it follows that $\mathcal{D} = \mathcal{M} \otimes \mathcal{N}$.

In the general case, choose a denumerable disjoint collection $(X_k)_{k=1}^{\infty}$of members of $\mathcal{M}$ and a denumerable disjoint collection $(Y_n)_{n=1}^{\infty}$of members of $\mathcal{N}$ such that

$$\cup_{k=1}^{\infty} X_k = X \text{ and } \cup_{n=1}^{\infty} Y_n = Y.$$

Set

$$\mu_k = \chi_{X_k}\mu,\ k = 1, 2, ...$$

and

$$\nu_n = \chi_{Y_n}\nu,\ n = 1, 2, ... \ .$$

Then, by the Beppo Levi Theorem, the function

$$f(x) = \int_Y \Sigma_{n=1}^{\infty} \chi_E(x, y)\chi_{Y_n}(y)d\nu(y)$$

$$= \Sigma_{n=1}^{\infty} \int_Y \chi_E(x, y)\chi_{Y_n}(y)d\nu(y) = \Sigma_{n=1}^{\infty}\nu_n(E_x)$$

is $\mathcal{M}$-measurable. Again, by the Beppo Levi Theorem,

$$\int_X f d\mu = \Sigma_{k=1}^{\infty} \int_X f d\mu_k$$

and

$$\int_X f d\mu = \Sigma_{k=1}^{\infty}(\Sigma_{n=1}^{\infty} \int_X \nu_n(E_x)d\mu_k(x)) = \Sigma_{k,n=1}^{\infty} \int_X \nu_n(E_x)d\mu_k(x).$$

In a similar way, the function g is $\mathcal{N}$-measurable and

$$\int_Y g d\nu = \Sigma_{n=1}^{\infty}(\Sigma_{k=1}^{\infty} \int_Y \mu_k(E^y)d\nu_n(y)) = \Sigma_{k,n=1}^{\infty} \int_Y \mu_k(E^y)d\nu_n(y).$$

Since the theorem is true for finite positive measure spaces, the general case follows.

Unit 14

3.6. Independence in Probability

Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. The random variables $\xi_k : (\Omega, P) \to (S_k, \mathcal{S}_k)$, $k = 1, ..., n$ are said to be independent if

$$P_{(\xi_1,...,\xi_n)} = \times_{k=1}^{n} P_{\xi_k}.$$

A family $(\xi_i)_{i\in I}$ of random variables is said to be independent if $\xi_{i_1}, ..., \xi_{i_n}$ are independent for any $i_1, ...i_n \in I$ with $i_k \neq i_l$ if $k \neq l$. A family of events $(A_i)_{i\in I}$ is said to be independent if $(\chi_{A_i})_{i\in I}$ is a family of independent random variables. Finally a family $(\mathcal{A}_i)_{i\in I}$ of sub-σ-algebras of $\mathcal{F}$ is said to be independent if, for any $A_i \in \mathcal{A}_i$, $i \in I$, the family $(A_i)_{i\in I}$ is a family of independent events.

Example 3.6.1. Let $q \geq 2$ be an integer. A real number $\omega \in [0, 1[$ has a q-adic expansion

$$\omega = \Sigma_{k=1}^{\infty} \frac{\xi_k^{(q)}}{q^k}.$$

The construction of the Cantor set shows that $(\xi_k^{(q)})_{k=1}^{\infty}$ is a sequence of independent random variables based on the probability space

$$([0, 1[\,, v_{1|[0,1[}, \mathcal{B}([0, 1[)).$$

Theorem 3.6.1. *Suppose $\xi_1, ..., \xi_n$ are independent random variables and $\xi_k \in N(0,1)$, $k = 1, ..., n$. If $\alpha_1, ..., \alpha_n \in R$, then*

$$\Sigma_{k=1}^n \alpha_k \xi_k \in N(0, \Sigma_{k=1}^n \alpha_k^2)$$

PROOF. The case $\alpha_1, ..., \alpha_n = 0$ is trivial so assume $\alpha_k \neq 0$ for some k. We have for each open interval A,

$$P\left[\Sigma_{k=1}^n \alpha_k \xi_k \in A\right] = \int_{\Sigma_{k=1}^n \alpha_k x_k \in A} d\gamma_1(x_1)...d\gamma_1(x_n)$$

$$\int_{\Sigma_{k=1}^n \alpha_k x_k \in A} \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}(x_1^2 + ... + x_n^2)} dx_1...dx_n.$$

Set $\sigma = \sqrt{\alpha_1^2 + ... + \alpha_n^2}$ and let $y = Gx$ be an orthogonal transformation such that

$$y_1 = \frac{1}{\sigma}(\alpha_1 x_1 + ... + \alpha_n x_n).$$

Then, since $\det G = 1$,

$$P\left[\Sigma_{k=1}^\infty \alpha_k \xi_k \in A\right] = \int_{\sigma y_1 \in A} \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}(y_1^2 + ... + y_n^2)} dy_1...dy_n$$

$$= \int_{\sigma y_1 \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_1^2} dy_1$$

where we used Fubini's theorem in the last step. The theorem is proved.

Finally, in this section, we prove a basic result about the existence of infinite product measures. Let μ_k, $k \in \mathbf{N}_+$ be Borel probability measures in $\mathbf{R}$. The space $\mathbf{R}^{\mathbf{N}_+}$ is, by definition, the set of all sequences $x = (x_k)_{k=1}^\infty$ of real numbers. For each $k \in \mathbf{N}_+$, set $\pi_k(x) = x_k$. The σ-algebra $\mathcal{R}^{\mathbf{N}_+}$ is the least σ-algebra $\mathcal{S}$ of subsets of $\mathbf{R}^{\mathbf{N}_+}$ which makes all the projections $\pi_k : (\mathbf{R}^{\mathbf{N}_+}, \mathcal{S}) \to (\mathbf{R}, \mathcal{R})$, $k \in \mathbf{N}_+$, measurable. Below, $(\pi_1, ..., \pi_n)$ denotes the mapping of $\mathbf{R}^{\mathbf{N}_+}$ into $\mathbf{R}^n$ defined by the equation

$$(\pi_1, ..., \pi_n)(x) = (\pi_1(x), ..., \pi_n(x)).$$

Theorem 3.6.1. *There is a unique probability measure μ on $\mathcal{R}^{\mathbf{N}_+}$ such that*

$$\mu_{(\pi_1,...,\pi_n)} = \mu_1 \times ... \times \mu_n$$

for every $n \in \mathbf{N}_+$.

The measure μ in Theorem 3.6.1 is called the product of the measures μ_k, $k \in \mathbf{N}_+$, and is often denoted by

$$\times_{k=1}^{\infty} \mu_k.$$

PROOF OF THEOREM 3.6.1. Let $(\Omega, P, \mathcal{F}) = ([0,1[\,, v_{1|[0,1[}, \mathcal{B}([0,1[)$ and set

$$\eta(\omega) = \Sigma_{k=1}^{\infty} \frac{\xi_k^{(2)}(\omega)}{2^k},\ \omega \in \Omega.$$

We already know that $P_\eta = P$. Now suppose $(k_i)_{i=1}^{\infty}$ is a strictly increasing sequence of positive integers and introduce

$$\eta' = \Sigma_{i=1}^{\infty} \frac{\xi_{k_i}^{(2)}(\omega)}{2^i},\ \omega \in \Omega.$$

Note that for each fixed positive integer n, the $\mathbf{R}^n$-valued maps $(\xi_1^{(2)}, ..., \xi_n^{(2)})$ and $(\xi_{k_1}^{(2)}, ..., \xi_{k_n}^{(2)})$ are P-equimeasurable. Thus, if $f : \Omega \to \mathbf{R}$ is continuous,

$$\int_\Omega f(\eta) dP = \lim_{n\to\infty} \int_\Omega f(\Sigma_{k=1}^n \frac{\xi_k^{(2)}(\omega)}{2^k}) dP(\omega)$$

$$= \lim_{n\to\infty} \int_\Omega f(\Sigma_{i=1}^n \frac{\xi_{k_i}^{(2)}(\omega)}{2^i}) dP(\omega) = \int_\Omega f(\eta') dP$$

and it follows that $P_{\eta'} = P_\eta = P$.

By induction, we define for each $k \in \mathbf{N}_+$ an infinite subset N_k of the set $\mathbf{N}_+ \setminus \cup_{i=1}^{k-1} N_i$ such that the set $\mathbf{N}_+ \setminus \cup_{i=1}^{k} N_i$ contains infinitely many elements and define

$$\eta_k = \Sigma_{i=1}^{\infty} \frac{\xi_{n_{ik}}^{(2)}(\omega)}{2^i}$$

where $(n_{ik})_{i=1}^{\infty}$ is an enumeration of N_k. The map

$$\Psi(\omega) = (\eta_k(\omega))_{k=1}^{\infty}$$

is a measurable map of $(\Omega, \mathcal{F})$ into $(\mathbf{R}^{\mathbf{N}_+}, \mathcal{R}^{\mathbf{N}_+})$ and

$$P_\Psi = \times_{k=1}^{\infty} \lambda_i$$

where $\lambda_i = P$ for each $i \in \mathbf{N}_+$.

For each $i \in \mathbf{N}_+$ there exists a measurable map φ_i of $(\Omega, \mathcal{F})$ into $(\mathbf{R}, \mathcal{R})$ such that $P_{\varphi_i} = \mu_i$ (see Section 1.6). The map

$$\Gamma(x) = (\varphi_i(x_i))_{i=1}^{\infty}$$

is a measurable map of $(\mathbf{R}^{\mathbf{N}_+}, \mathcal{R}^{\mathbf{N}_+})$ into itself and we get $\mu = (P_\Psi)_\Gamma$. This completes the proof of Theorem 3.6.1.

Unit 15

Introduction

In this chapter we will treat a variety of different sorts of convergence notions in measure theory. So called L^2-convergence is of particular importance.

4.1. Convergence in Measure, in $L^1(\mu)$, and in $L^2(\mu)$

Let $(X, \mathcal{M}, \mu)$ be a positive measure space and denote by $\mathcal{F}(X)$ the class of measurable functions $f : (X, \mathcal{M}) \to (\mathbf{R}, \mathcal{R})$. For any $f \in \mathcal{F}(X)$, set

$$\| f \|_1 = \int_X | f(x) | d\mu(x)$$

and

$$\| f \|_2 = \sqrt{\int_X f^2(x) d\mu(x)}.$$

The Cauchy-Schwarz inequality states that

$$\int_X | fg | d\mu \leq \| f \|_2 \| g \|_2 \text{ if } f, g \in \mathcal{F}(X).$$

To prove this, without loss of generality, it can be assumed that

$$0 < \| f \|_2 < \infty \text{ and } 0 < \| g \|_2 < \infty.$$

We now use the inequality

$$\alpha\beta \leq \frac{1}{2}(\alpha^2 + \beta^2), \ \alpha, \beta \in \mathbf{R}$$

to obtain

$$\int_X \frac{|\ f\ |}{\|\ f\ \|_2}\frac{|\ g\ |}{\|\ g\ \|_2}d\mu \leq \int \frac{1}{2}(\frac{f^2}{\|\ f\ \|_2^2}+\frac{g^2}{\|\ g\ \|_2^2})d\mu = 1$$

and the Cauchy-Schwarz inequality is immediate.

If not otherwise stated, in this section p is a number equal to 1 or 2. If it is important to emphasize the underlying measure $\|\ f\ \|_p$ is written $\|\ f\ \|_{p,\mu}$.

We now define

$$\mathcal{L}^p(\mu) = \{f \in \mathcal{F}(X);\ \ \|\ f\ \|_p < \infty\}\,.$$

The special case $p = 1$ has been introduced earlier. We claim that the following so called triangle inequality holds, viz.

$$\|\ f+g\ \|_p \leq \|\ f\ \|_p + \|\ g\ \|_p \ \text{ if } f,g \in \mathcal{L}^p(\mu).$$

The case $p = 1$, follows by μ-integration of the relation

$$|\ f+g\ | \leq |\ f\ | + |\ g\ |\,.$$

To prove the case $p = 2$, we use the Cauchy-Schwarz inequality and have

$$\|\ f+g\ \|_2^2 \leq \|\,|\ f\ | + |\ g\ |\,\|_2^2$$

$$= \|\ f\ \|_2^2 + 2\int_X |\ fg\ |\ d\mu + \|\ g\ \|_2^2$$

$$\leq \|\ f\ \|_2^2 + 2\ \|\ f\ \|_2 \|\ g\ \|_2 + \|\ g\ \|_2^2 = (\|\ f\ \|_2 + \|\ g\ \|_2)^2$$

and the triangle inequality is immediate.

Suppose $f, g \in \mathcal{L}^p(\mu)$. The functions f and g are equal almost everywhere with respect to μ if $\{f \neq g\} \in \mathcal{Z}_\mu$. This is easily seen to be an equivalence relation and the set of all equivalence classes is denoted by $L^p(\mu)$. Below we consider the elements of $L^p(\mu)$ as members of $\mathcal{L}^p(\mu)$ and two members of $L^p(\mu)$ are identified if they are equal a.e. $[\mu]$. From this convention it is straight-forward to define $f+g$ and αf for all $f, g \in L^p(\mu)$ and $\alpha \in \mathbf{R}$ and the function $d^{(p)}(f,g) = \|\ f-g\ \|_p$ is a metric on $L^p(\mu)$. Convergence in the metric space $L^p(\mu) = (L^p(\mu), d^{(p)})$ is called convergence in $L^p(\mu)$. A sequence $(f_k)_{k=1}^\infty$ in $\mathcal{F}(X)$ converges in measure to a function $f \in \mathcal{F}(X)$ if

$$\lim_{k\to\infty} \mu(|\ f_k - f\ | > \varepsilon) = 0 \text{ all } \varepsilon > 0.$$

If the sequence $(f_k)_{k=1}^{\infty}$ in $\mathcal{F}(X)$ converges in measure to a function $f \in \mathcal{F}(X)$ as well as to a function $g \in \mathcal{F}(X)$, then $f = g$ a.e. $[\mu]$ since

$$\{| f - g | > \varepsilon\} \subseteq \left\{| f - f_k | > \frac{\varepsilon}{2}\right\} \cup \left\{| f_k - g | > \frac{\varepsilon}{2}\right\}$$

and

$$\mu(| f - g | > \varepsilon) \leq \mu(| f - f_k | > \frac{\varepsilon}{2}) + \mu(| f_k - g | > \frac{\varepsilon}{2})$$

for every $\varepsilon > 0$ and positive integer k. A sequence $(f_k)_{k=1}^{\infty}$ in $\mathcal{F}(X)$ is said to be Cauchy in measure if for every $\varepsilon > 0$,

$$\mu(| f_k - f_n | > \varepsilon) \to 0 \text{ as } k, n \to \infty.$$

By the Markov inequality, a Cauchy sequence in $L^p(\mu)$ is Cauchy in measure.

Example 4.1.1. (a) If $f_k = \sqrt{k}\chi_{\left[0,\frac{1}{k}\right]}$, $k \in \mathbf{N}_+$, then

$$\| f_k \|_{2,m} = 1 \text{ and } \| f_k \|_{1,m} = \frac{1}{\sqrt{k}}.$$

Thus $f_k \to 0$ in $L^1(m)$ as $k \to \infty$ but $f_k \nrightarrow 0$ in $L^2(m)$ as $k \to \infty$.

(b) $L^1(m) \nsubseteq L^2(m)$ since

$$\chi_{[1,\infty[}(x)\frac{1}{| x |} \in L^2(m) \setminus L^1(m)$$

and $L^2(m) \nsubseteq L^1(m)$ since

$$\chi_{]0,1]}(x)\frac{1}{\sqrt{| x |}} \in L^1(m) \setminus L^2(m).$$

Theorem 4.1.1. *Suppose $p = 1$ or 2.*

(a) *Convergence in $L^p(\mu)$ implies convergence in measure.*

(b) *If* $\mu(X) < \infty$, *then* $L^2(\mu) \subseteq L^1(\mu)$ *and convergence in* $L^2(\mu)$ *implies convergence in* $L^1(\mu)$.

Proof. (a) Suppose the sequence $(f_n)_{n=1}^{\infty}$ converges to f in $L^p(\mu)$ and let $\varepsilon > 0$. Then, by the Markov inequality,

$$\mu(\mid f_n - f \mid \geq \varepsilon) \leq \frac{1}{\varepsilon^p} \int_X \mid f_n - f \mid^p d\mu = \frac{1}{\varepsilon^p} \parallel f_n - f \parallel_p^p$$

and (a) follows at once.

(b) The Cauchy-Schwarz inequality gives for any $f \in \mathcal{F}(X)$,

$$(\int_X \mid f \mid \cdot 1 d\mu)^2 \leq \int_X f^2 d\mu \int_X 1 d\mu$$

or

$$\parallel f \parallel_1 \leq \parallel f \parallel_2 \sqrt{\mu(X)}$$

and Part (b) is immediate.

Theorem 4.1.2. *Suppose* $f_n \in \mathcal{F}(X)$, $n \in \mathbf{N}_+$.

(a) *If* $(f_n)_{n=1}^{\infty}$ *is Cauchy in measure, there is a measurable function* $f : X \to \mathbf{R}$ *such that* $f_n \to f$ *in measure as* $n \to \infty$ *and a strictly increasing sequence of positive integers* $(n_j)_{j=1}^{\infty}$ *such that* $f_{n_j} \to f$ *a.e.* $[\mu]$ *as* $j \to \infty$.

(b) *If* μ *is a finite positive measure and* $f_n \to f \in \mathcal{F}(X)$ *a.e.* $[\mu]$ *as* $n \to \infty$, *then* $f_n \to f$ *in measure.*

(c) **(Egoroff's Theorem)** *If* μ *is a finite positive measure and* $f_n \to f \in \mathcal{F}(X)$ *a.e.* $[\mu]$ *as* $n \to \infty$, *then for every* $\varepsilon > 0$ *there exists* $E \in \mathcal{M}$ *such that* $\mu(E) < \varepsilon$ and

$$\sup_{\substack{k \geq n \\ x \in E^c}} \mid f_k(x) - f(x) \mid \to 0 \text{ as } n \to \infty.$$

PROOF. (a) For each positive integer j, there is a positive integer n_j such that

$$\mu(\mid f_k - f_l \mid > 2^{-j}) < 2^{-j}, \text{ all } k, l \geq n_j.$$

There is no loss of generality to assume that $n_1 < n_2 < \ldots$. Set

$$E_j = \left\{ \mid f_{n_j} - f_{n_{j+1}} \mid > 2^{-j} \right\}$$

and

$$F_k = \cup_{j=k}^{\infty} E_j.$$

If $x \in F_k^c$ and $i \geq j \geq k$

$$\mid f_{n_i}(x) - f_{n_j}(x) \mid \leq \sum_{j \leq l < i} \mid f_{n_{l+1}}(x) - f_{n_l}(x) \mid$$

$$\leq \sum_{j \leq l < i} 2^{-l} < 2^{-j+1}$$

and we conclude that $(f_{n_j}(x))_{j=1}^{\infty}$ is a Cauchy sequence for every $x \in F_k^c$. Let $G = \cup_{k=1}^{\infty} F_k^c$ and note that for every fixed positive integer k,

$$\mu(G^c) \leq \mu(F_k) < \sum_{j=k}^{\infty} 2^{-j} = 2^{-k+1}.$$

Thus G^c is a μ-null set. We now define $f(x) = \lim_{j \to \infty} f_{n_j}(x)$ if $x \in G$ and $f(x) = 0$ if $x \notin G$.

We next prove that the sequence $(f_n)_{n=1}^{\infty}$ converges to f in measure. If $x \in F_k^c$ and $j \geq k$ we get

$$\mid f(x) - f_{n_j}(x) \mid \leq 2^{-j+1}.$$

Thus, if $j \geq k$

$$\mu(\mid f - f_{n_j} \mid > 2^{-j+1}) \leq \mu(F_k) < 2^{-k+1}.$$

Since

$$\mu(\mid f_n - f \mid > \varepsilon) \leq \mu(\mid f_n - f_{n_j} \mid > \frac{\varepsilon}{2}) + \mu(\mid f_{n_j} - f \mid > \frac{\varepsilon}{2})$$

if $\varepsilon > 0$, Part (a) follows at once.

(b) For each $\varepsilon > 0$,

$$\mu(\mid f_n - f \mid > \varepsilon) = \int_X \chi_{]\varepsilon, \infty[}(\mid f_n - f \mid) d\mu$$

and Part (c) follows from the Lebesgue Dominated Convergence Theorem.

(c) Set for fixed $k, n \in \mathbf{N}_+$,

$$E_{kn} = \cup_{j=n}^{\infty} \left\{ \mid f_j - f \mid > \frac{1}{k} \right\}.$$

We have

$$\cap_{n=1}^{\infty} E_{kn} \in Z_\mu$$

and since μ is a finite measure

$$\mu(E_{kn}) \to 0 \text{ as } n \to \infty.$$

Given $\varepsilon > 0$ pick $n_k \in \mathbf{N}_+$ such that $\mu(E_{kn_k}) < \varepsilon 2^{-k}$. Then, if $E = \cup_{k=1}^{\infty} E_{kn_k}$, $\mu(E) < \varepsilon$. Moreover, if $x \notin E$ and $j \geq n_k$

$$\mid f_j(x) - f(x) \mid \leq \frac{1}{k}.$$

The theorem is proved.

Corollary 4.1.1. *The spaces $L^1(\mu)$ and $L^2(\mu)$ are complete.*

PROOF. Suppose $p = 1$ or 2 and let $(f_n)_{n=1}^{\infty}$ be a Cauchy sequence in $L^p(\mu)$. We know from the previous theorem that there exists a subsequence $(f_{n_j})_{j=1}^{\infty}$ which converges pointwise to a function $f \in \mathcal{F}(X)$ a.e. $[\mu]$. Thus, by Fatou's Lemma,

$$\int_X \mid f - f_k \mid^p d\mu \leq \liminf_{j \to \infty} \int_X \mid f_{n_j} - f_k \mid^p d\mu$$

and it follows that $f - f_k \in L^p(\mu)$ and, hence $f = (f - f_k) + f_k \in L^p(\mu)$. Moreover, we have that $\| f - f_k \|_p \to 0$ as $k \to \infty$. This concludes the proof of the theorem.

Corollary 4.1.2. *Suppose $\xi_n \in N(0, \sigma_n^2)$, $n \in \mathbf{N}_+$, and $\xi_n \to \xi$ in $L^2(P)$ as $n \to \infty$. Then ξ is a centred Gaussian random variable.*

PROOF. We have that $\| \xi_n \|_2 = \sqrt{E\left[\xi_n^2\right]} = \sigma_n$ and $\| \xi_n \|_2 \to \| \xi \|_2 =_{def} \sigma$ as $n \to \infty$.

Suppose f is a bounded continuous function on $\mathbf{R}$. Then, by dominated convergence,

$$E\left[f(\xi_n)\right] = \int_{\mathbf{R}} f(\sigma_n x) d\gamma_1(x) \to \int_{\mathbf{R}} f(\sigma x) d\gamma_1(x)$$

as $n \to \infty$. Moreover, there exists a subsequence $(\xi_{n_k})_{k=1}^{\infty}$ which converges to ξ a.s. Hence, by dominated convergence

$$E\left[f(\xi_{n_k})\right] \to E\left[f(\xi)\right]$$

as $k \to \infty$ and it follows that

$$E\left[f(\xi)\right] = \int_{\mathbf{R}} f(\sigma x) d\gamma_1(x).$$

By using Corollary 3.1.3 the theorem follows at once.

Theorem 4.1.3. *Suppose X is a standard space and μ a positive σ-finite Borel measure on X. Then the spaces $L^1(\mu)$ and $L^2(\mu)$ are separable.*

PROOF. Let $(E_k)_{k=1}^{\infty}$ be a denumerable collection of Borel sets with finite μ-measures and such that $E_k \subseteq E_{k+1}$ and $\cup_{k=1}^{\infty} E_k = X$. Set $\mu_k = \chi_{E_k}\mu$ and first suppose that the set D_k is at most denumerable and dense in $L^p(\mu_k)$ for every $k \in \mathbf{N}_+$. Without loss of generality it can be assumed that each member of D_k vanishes off E_k. By monotone convergence

$$\int_X f d\mu = \lim_{k\to\infty} \int_X f d\mu_k,\ f \geq 0 \text{ measurable},$$

and it follows that the set $\cup_{k=1}^{\infty} D_k$ is at most denumerable and dense in $L^p(\mu)$.

From now on we can assume that μ is a finite positive measure. Let A be an at most denumerable dense subset of X and and suppose the subset $\{r_n;\ n \in \mathbf{N}_+,\}$ of $]0,\infty[$ is dense in $]0,\infty[$. Furthermore, denote by $\mathcal{U}$ the

class of all open sets which are finite unions of open balls of the type $B(a, r_n)$, $a \in A$, $n \in \mathbf{N}_+$. If U is any open subset of X

$$U = \cup [V : V \subseteq U \text{ and } V \in \mathcal{U}]$$

and, hence, by the Ulam Theorem

$$\mu(U) = \sup \{\mu(V);\ V \in \mathcal{U} \text{ and } V \subseteq U\}.$$

Let $\mathcal{K}$ be the class of all functions which are finite sums of functions of the type $\kappa\chi_U$, where κ is a positive rational number and $U \in \mathcal{U}$. It follows that $\mathcal{K}$ is at most denumerable.

Suppose $\varepsilon > 0$ and that $f \in L^p(\mu)$ is non-negative. There exists a sequence of simple measurable functions $(\varphi_i)_{i=1}^{\infty}$ such that

$$0 \leq \varphi_i \uparrow f \text{ a.e. } [\mu].$$

Since $\mid f - \varphi_i \mid^p \leq f^p$, the Lebesgue Dominated Convergence Theorem shows that $\parallel f - \varphi_k \parallel_p < \frac{\varepsilon}{2}$ for an appropriate k. Let $\alpha_1, ..., \alpha_l$ be the distinct positive values of φ_k and set

$$C = 1 + \Sigma_{k=1}^{l} \alpha_k.$$

Now for each fixed $j \in \{1, ..., l\}$ we use Theorem 3.1.3 to get an open $U_j \supseteq \varphi_k^{-1}(\{\alpha_j\})$ such that $\parallel \chi_{U_j} - \chi_{\varphi_k^{-1}(\{\alpha_j\})} \parallel_p < \frac{\varepsilon}{4C}$ and from the above we get a $V_j \in \mathcal{U}$ such that $V_j \subseteq U_j$ and $\parallel \chi_{U_j} - \chi_{V_j} \parallel_p < \frac{\varepsilon}{4C}$. Thus

$$\parallel \chi_{V_j} - \chi_{\varphi_k^{-1}(\{\alpha_j\})} \parallel_p < \frac{\varepsilon}{2C}$$

and

$$\parallel f - \Sigma_{k=1}^{l} \alpha_j \chi_{V_j} \parallel_p < \varepsilon$$

Now it is simple to find a $\psi \in \mathcal{K}$ such that $\parallel f - \psi \parallel_p < \varepsilon$. From this we deduce that the set

$$\mathcal{K} - \mathcal{K} = \{g - h;\ g, h \in \mathcal{K}\}$$

is at most denumerable and dense in $L^p(\mu)$.

The set of all real-valued and infinitely many times differentiable functions defined on $\mathbf{R}^n$ is denoted by $C^{(\infty)}(\mathbf{R}^n)$ and

$$C_c^{(\infty)}(\mathbf{R}^n) = \left\{ f \in C^{(\infty)}(\mathbf{R}^n);\ \text{supp} f \text{ compact} \right\}.$$

Recall that the support $\text{supp} f$ of a real-valued continuous function f defined on $\mathbf{R}^n$ is the closure of the set of all x where $f(x) \neq 0$. If

$$f(x) = \prod_{k=1}^{n} \left\{ \varphi(1+x_k)\varphi(1-x_k) \right\},\ x = (x_1, ..., x_n) \in \mathbf{R}^n$$

where $\varphi(t) = \exp(-t^{-1})$, if $t > 0$, and $\varphi(t) = 0$, if $t \leq 0$, then $f \in C_c^{\infty}(\mathbf{R}^n)$.

The proof of the previous theorem also gives Part (a) of the following

Theorem 4.1.4. *Suppose μ is a positive Borel measure in $\mathbf{R}^n$ such that $\mu(K) < \infty$ for every compact subset K of $\mathbf{R}^n$. The following sets are dense in $L^1(\mu)$, and $L^2(\mu)$:*

(a) *the linear span of the functions*

$$\chi_I,\ I \text{ open bounded } n\text{-cell in } \mathbf{R}^n,$$

(b) $C_c^{(\infty)}(\mathbf{R}^n)$.

PROOF. a) The proof is almost the same as the proof of Theorem 4.1.3. First the E_k:s can be chosen to be open balls with their centres at the origin since each bounded set in $\mathbf{R}^n$ has finite μ-measure. Moreover, as in the proof of Theorem 4.1.3 we can assume that μ is a finite measure. Now let A be an at most denumerable dense subset of $\mathbf{R}^n$ and for each $a \in A$ let

$$R(a) = \{r > 0;\ \mu(\{x \in X;\ \ | x_k - a_k | = r\}) > 0 \text{ for some } k = 1, ..., n\}.$$

Then $\cup_{a \in A} R(a)$ is at most denumerable and there is a subset $\{r_n;\ n \in \mathbf{N}_+\}$ of $]0, \infty[\setminus \cup_{a \in A} R(a)$ which is dense in $]0, \infty[$. Finally, let $\mathcal{U}$ denote the class of all open sets which are finite unions of open balls of the type $B(a, r_n)$, $a \in A$, $n \in \mathbf{N}_+$, and proceed as in the proof of Theorem 4.1.3. The result follows by observing that the characteristic function of any member of $\mathcal{U}$ equals a finite sum of characteristic functions of open bounded n-cells a.e. $[\mu]$.

Part (b) in Theorem 4.1.4 follows from Part (a) and the following

Lemma 4.1.1. *Suppose $K \subseteq U \subseteq \mathbf{R}^n$, where K is compact and U is open. Then there exists a function $f \in C_c^{(\infty)}(\mathbf{R}^n)$* such that

$$K \prec f \prec U$$

that is

$$\chi_K \leq f \leq \chi_U \text{ and } \text{supp} f \subseteq U.$$

PROOF. Suppose $\rho \in C_c^{\infty}(\mathbf{R}^n)$ is non-negative, supp $\rho \subseteq B(0,1)$, and

$$\int_{\mathbf{R}^n} \rho dm_n = 1.$$

Moreover, let $\varepsilon > 0$ be fixed. For any $g \in L^1(v_n)$ we define

$$f_\varepsilon(x) = \varepsilon^{-n} \int_{\mathbf{R}^n} g(y)\rho(\varepsilon^{-1}(x-y))dy.$$

Since

$$\mid g \mid \max_{\mathbf{R}^n} \mid \frac{\partial^{k_1+...+k_n}\rho}{\partial x_1^{k_1}...\partial x_n^{k_n}} \mid \in L^1(v_n), \text{ all } k_1,...,k_n \in \mathbf{N}$$

the Lebesgue Dominated Convergent Theorem shows that $f_\varepsilon \in C^{\infty}(\mathbf{R}^n)$. Here $f_\varepsilon \in C_c^{\infty}(\mathbf{R}^n)$ if g vanishes off a bounded subset of $\mathbf{R}^n$. In fact,

$$\text{supp } f_\varepsilon \subseteq (\text{supp } g)_\varepsilon.$$

Now choose a positive number $\varepsilon \leq \frac{1}{2}d(K, U^c)$ and define $g = \chi_{K_\varepsilon}$. Since

$$f_\varepsilon(x) = \int_{\mathbf{R}^n} g(x - \varepsilon y)\rho(y)dy$$

we also have that $f_\varepsilon(x) = 1$ if $x \in K$. The lemma is proved.

4.2 Orthogonality

Suppose $(X, \mathcal{M}, \mu)$ is a positive measure space. If $f, g \in L^2(\mu)$, let

$$\langle f, g\rangle =_{def} \int_X f g d\mu$$

be the so called scalar product of f and g. The Cauchy-Schwarz inequality

$$| \langle f, g\rangle | \leq \| f \|_2 \| g \|_2$$

shows that the map $f \to \langle f, g\rangle$ of $L^2(\mu)$ into $\mathbf{R}$ is continuous. Observe that

$$\| f + g \|_2^2 = \| f \|_2^2 + 2\langle f, g\rangle + \| g \|_2^2$$

and from this we get the so called Parallelogram Law

$$\| f + g \|_2^2 + \| f - g \|_2^2 = 2(\| f \|_2^2 + \| g \|_2^2).$$

We will say that f and g are orthogonal (abbr. $f \perp g$) if $\langle f, g\rangle = 0$. Note that

$$\| f + g \|_2^2 = \| f \|_2^2 + \| g \|_2^2 \text{ if and only if } f \perp g.$$

Since $f \perp g$ implies $g \perp f$, the relation $\perp$ is symmetric. Moreover, if $f \perp h$ and $g \perp h$ then $(\alpha f + \beta g) \perp h$ for all $\alpha, \beta \in \mathbf{R}$. Thus $h^\perp =_{def} \{f \in L^2(\mu);\ f \perp h\}$ is a subspace of $L^2(\mu)$, which is closed since the map $f \to \langle f, h\rangle$, $f \in L^2(\mu)$ is continuous. If M is a subspace of $L^2(\mu)$, the set

$$M^\perp =_{def} \cap_{h \in M} h^\perp$$

is a closed subspace of $L^2(\mu)$. The function $f = 0$ if and only if $f \perp f$.

If M is a subspace of $L^2(\mu)$ and $f \in L^2(\mu)$ there exists at most one point $g \in M$ such that $f - g \in M^\perp$. To see this, let $g_0, g_1 \in M$ be such that $f - g_k \in M^\perp$, $k = 0, 1$. Then $g_1 - g_0 = (f - g_0) - (f - g_1) \in M^\perp$ and hence $g_1 - g_0 \perp g_1 - g_0$ that is $g_0 = g_1$.

Theorem 4.2.1. *Let M be a closed subspace in $L^2(\mu)$ and suppose $f \in L^2(\mu)$. Then there exists a unique point $g \in M$ such that*

$$\| f - g \|_2 \leq \| f - h \|_2 \text{ all } h \in M.$$

Moreover,

$$f - g \in M^{\perp}.$$

The function g in Theorem 4.2.1 is called the projection of f on M and is denoted by $\mathrm{Proj}_M\ f$.

PROOF OF THEOREM 4.2.1. Set

$$d =_{def} d^{(2)}(f, M) = \inf_{g \in M} \| f - g \|_2 .$$

and let $(g_n)_{n=1}^{\infty}$ be a sequence in M such that

$$d = \lim_{n \to \infty} \| f - g_n \|_2 .$$

Then, by the Parallelogram Law

$$\| (f-g_k)+(f-g_n) \|_2^2 + \| (f-g_k)-(f-g_n) \|_2^2 = 2(\| f-g_k \|_2^2 + \| f-g_n \|_2^2)$$

that is

$$4 \| f - \frac{1}{2}(g_k + g_n) \|_2^2 + \| g_n - g_k \|_2^2 = 2(\| f - g_k \|_2^2 + \| f - g_n \|_2^2)$$

and, since $\frac{1}{2}(g_k + g_n) \in M$, we get

$$4d^2 + \| g_n - g_k \|_2^2 \leq 2(\| f - g_k \|_2^2 + \| f - g_n \|_2^2).$$

Here the right hand converges to $4d^2$ as k and n go to infinity and we conclude that $(g_n)_{n=1}^{\infty}$ is a Cauchy sequence. Since $L^2(\mu)$ is complete and M closed there exists a $g \in M$ such that $g_n \to g$ as $n \to \infty$. Moreover,

$$d = \| f - g \|_2 .$$

We claim that $f - g \in M^{\perp}$. To prove this choose $h \in M$ and $\alpha > 0$ arbitrarily and use the inequality

$$\| (f - g) + \alpha h \|_2^2 \geq \| f - g \|_2^2$$

to obtain

$$\| f - g \|_2^2 + 2\alpha \langle f - g, h \rangle + \alpha^2 \| h \|_2^2 \geq \| f - g \|_2^2$$

and

$$2\langle f-g,h\rangle+\alpha\parallel h\parallel_2^2\geq 0.$$

By letting $\alpha\to 0$, $\langle f-g,h\rangle\geq 0$ and replacing h by $-h$, $\langle f-g,h\rangle\leq 0$. Thus $f-g\in h^{\perp}$ and it follows that $f-g\in M^{\perp}$.

The uniqueness in Theorem 4.2.1 follows from the remark just before the formulation of Theorem 4.2.1. The theorem is proved.

A linear mapping $T:L^2(\mu)\to\mathbf{R}$ is called a linear functional on $L^2(\mu)$. If $h\in L^2(\mu)$, the map $h\to\langle f,h\rangle$ of $L^2(\mu)$ into $\mathbf{R}$ is a continuous linear functional on $L^2(\mu)$. It is a very important fact that every continuous linear functional on $L^2(\mu)$ is of this type.

Theorem 4.2.2. *Suppose T is a continuous linear functional on $L^2(\mu)$. Then there exists a unique $w\in L^2(\mu)$ such that*

$$Tf=\langle f,w\rangle \text{ all } f\in L^2(\mu).$$

PROOF. Uniqueness: If $w,w'\in L^2(\mu)$ and $\langle f,w\rangle=\langle f,w'\rangle$ for all $f\in L^2(\mu)$, then $\langle f,w-w'\rangle=0$ for all $f\in L^2(\mu)$. By choosing $f=w-w'$ we get $f\perp f$ that is $w=w'$.

Existence: The set $M=_{def}T^{-1}(\{0\})$ is closed since T is continuous and M is a linear subspace of $L^2(\mu)$ since T is linear. If $M=L^2(\mu)$ we choose $w=0$. Otherwise, pick a $g\in L^2(\mu)\setminus M$. Without loss of generality it can be assumed that $Tg=1$ by eventually multiplying g by a scalar. The previous theorem gives us a vector $h\in M$ such that $u=_{def}g-h\in M^{\perp}$. Note that $0<\parallel u\parallel_2^2=\langle u,g-h\rangle=\langle u,g\rangle$.

To conclude the proof, let fixed $f\in L^2(\mu)$ be fixed, and use that $(Tf)g-f\in M$ to obtain

$$\langle (Tf)g-f,u\rangle=0$$

or

$$(Tf)\langle g,u\rangle=\langle f,u\rangle.$$

By setting

$$w = \frac{1}{\| u \|_2^2} u$$

we are done.

Unit 16

Introduction

In this section a version of the fundamental theorem of calculus for Lebesgue integrals will be proved. Moreover, the concept of differentiating a measure with respect to another measure will be developped. A very important result in this chapter is the so called Radon-Nikodym Theorem.

5.1. Complex Measures

Let $(X, \mathcal{M})$ be a measurable space. Recall that if $A_n \subseteq X$, $n \in \mathbf{N}_+$, and $A_i \cap A_j = \phi$ if $i \neq j$, the sequence $(A_n)_{n\in\mathbf{N}_+}$ is called a disjoint denumerable collection. The collection is called a measurable partition of A if $A = \cup_{n=1}^{\infty} A_n$ and $A_n \in \mathcal{M}$ for every $n \in \mathbf{N}_+$.

A complex function μ on $\mathcal{M}$ is called a complex measure if

$$\mu(A) = \Sigma_{n=1}^{\infty} \mu(A_n)$$

for every $A \in \mathcal{M}$ and measurable partition $(A_n)_{n=1}^{\infty}$ of A. Note that $\mu(\phi) = 0$ if μ is a complex measure. A complex measure is said to be a real measure if it is a real function. The reader should note that a positive measure need not be a real measure since infinity is not a real number. If μ is a complex measure $\mu = \mu_{\mathrm{Re}} + i\mu_{\mathrm{Im}}$, where μ_{Re} =Re μ and μ_{Im} =Im μ are real measures.

If $(X, \mathcal{M}, \mu)$ is a positive measure and $f \in L^1(\mu)$ it follows that

$$\lambda(A) = \int_A f d\mu, \ A \in \mathcal{M}$$

is a real measure and we write $d\lambda = f d\mu$.

A function $\mu : \mathcal{M} \to [-\infty, \infty]$ is called a signed measure measure if

(a) $\mu : \mathcal{M} \to]-\infty, \infty]$ or $\mu : \mathcal{M} \to [-\infty, \infty[$
(b) $\mu(\phi) = 0$
and
(c) for every $A \in \mathcal{M}$ and measurable partition $(A_n)_{n=1}^{\infty}$ of A,

$$\mu(A) = \Sigma_{n=1}^{\infty} \mu(A_n)$$

where the latter sum converges absolutely if $\mu(A) \in \mathbf{R}$.

Here $-\infty - \infty = -\infty$ and $-\infty + x = -\infty$ if $x \in \mathbf{R}$. The sum of a positive measure and a real measure and the difference of a real measure and a positive measure are examples of signed measures and it can be proved that there are no other signed measures (see Folland [F]). Below we concentrate on positive, real, and complex measures and will not say more about signed measures here.

Suppose μ is a complex measure on $\mathcal{M}$ and define for every $A \in \mathcal{M}$

$$| \mu | (A) = \sup \Sigma_{n=1}^{\infty} | \mu(A_n) |,$$

where the supremum is taken over all measurable partitions $(A_n)_{n=1}^{\infty}$ of A. Note that $| \mu | (\phi) = 0$ and

$$| \mu | (A) \geq | \mu(B) | \text{ if } A, B \in \mathcal{M} \text{ and } A \supseteq B.$$

The set function $| \mu |$ is called the total variation of μ or the total variation measure of μ. It turns out that $| \mu |$ is a positive measure. In fact, as will shortly be seen, $| \mu |$ is a finite positive measure.

Theorem 5.1.1. *The total variation $| \mu |$ of a complex measure is a positive measure.*

PROOF. Let $(A_n)_{n=1}^{\infty}$ be a measurable partition of A.

For each n, suppose $a_n <\mid \mu \mid (A_n)$ and let $(E_{kn})_{k=1}^{\infty}$ be a measurable partition of A_n such that

$$a_n < \Sigma_{k=1}^{\infty} \mid \mu(E_{kn}) \mid .$$

Since $(E_{kn})_{k,n=1}^{\infty}$ is a partition of A it follows that

$$\Sigma_{n=1}^{\infty} a_n < \Sigma_{k,n=1}^{\infty} \mid \mu(E_{kn}) \mid \leq \mid \mu \mid (A).$$

Thus

$$\Sigma_{n=1}^{\infty} \mid \mu \mid (A_n) \leq \mid \mu \mid (A).$$

To prove the opposite inequality, let $(E_k)_{k=1}^{\infty}$ be a measurable partition of A. Then, since $(A_n \cap E_k)_{n=1}^{\infty}$ is a measurable partition of E_k and $(A_n \cap E_k)_{k=1}^{\infty}$ a measurable partition of A_n,

$$\Sigma_{k=1}^{\infty} \mid \mu(E_k) \mid = \Sigma_{k=1}^{\infty} \mid \Sigma_{n=1}^{\infty} \mu(A_n \cap E_k) \mid$$

$$\leq \Sigma_{k,n=1}^{\infty} \mid \mu(A_n \cap E_k) \mid \leq \Sigma_{n=1}^{\infty} \mid \mu \mid (A_n)$$

and we get

$$\mid \mu \mid (A) \leq \Sigma_{n=1}^{\infty} \mid \mu \mid (A_n).$$

Thus

$$\mid \mu \mid (A) = \Sigma_{n=1}^{\infty} \mid \mu \mid (A_n).$$

Since $\mid \mu \mid (\phi) = 0$, the theorem is proved.

Theorem 5.1.2. *The total variation $\mid \mu \mid$ of a complex measure μ is a finite positive measure.*

PROOF. Since

$$\mid \mu \mid \leq \mid \mu_{\mathrm{Re}} \mid + \mid \mu_{\mathrm{Im}} \mid$$

there is no loss of generality to assume that μ is a real measure.

Suppose $\mid \mu \mid (E) = \infty$ for some $E \in \mathcal{M}$. We first prove that there exist disjoint sets $A, B \in \mathcal{M}$ such that

$$A \cup B = E$$

and

$$| \mu(A) | > 1 \text{ and } | \mu | (B) = \infty.$$

To this end let $c = 2(1+ | \mu(E) |)$ and let $(E_k)_{k=1}^{\infty}$ be a measurable partition of E such that

$$\Sigma_{k=1}^{n} | \mu(E_k) | > c$$

for some sufficiently large n. There exists a subset N of $\{1, ..., n\}$ such that

$$| \Sigma_{k \in N} \mu(E_k) | > \frac{c}{2}.$$

Set $A = \cup_{k \in N} E_k$ and $B = E \setminus A$. Then $| \mu(A) | > \frac{c}{2} \geq 1$ and

$$| \mu(B) | = | \mu(E) - \mu(A) |$$

$$\geq | \mu(A) | - | \mu(E) | > \frac{c}{2} - | \mu(E) | = 1.$$

Since $\infty = | \mu | (E) = | \mu | (A) + | \mu | (B)$ we have $| \mu | (A) = \infty$ or $| \mu | (B) = \infty$. If $| \mu | (B) < \infty$ we interchange A and B and have $| \mu(A) | > 1$ and $| \mu | (B) = \infty$.

Suppose $| \mu | (X) = \infty$. Set $E_0 = X$ and choose disjoint sets $A_0, B_0 \in \mathcal{M}$ such that

$$A_0 \cup B_0 = E_0$$

and

$$| \mu(A_0) | > 1 \text{ and } | \mu | (B_0) = \infty.$$

Set $E_1 = B_0$ and choose disjoint sets $A_1, B_1 \in \mathcal{M}$ such that

$$A_1 \cup B_1 = E_1$$

and

$$| \mu(A_1) | > 1 \text{ and } | \mu | (B_1) = \infty.$$

By induction, we find a measurable partition $(A_n)_{n=0}^{\infty}$ of the set $A =_{def} \cup_{n=0}^{\infty} A_n$ such that $| \mu(A_n) | > 1$ for every n. Now, since μ is a complex measure,

$$\mu(A) = \Sigma_{n=0}^{\infty} \mu(A_n).$$

But this series cannot converge, since the general term does not tend to zero as $n \to \infty$. This contradiction shows that $| \mu |$ is a finite positive measure.

If μ is a real measure we define

$$\mu^+ = \frac{1}{2}(\mid \mu \mid +\mu)$$

and

$$\mu^- = \frac{1}{2}(\mid \mu \mid -\mu).$$

The measures μ^+ and μ^- are finite positive measures and are called the positive and negative variations of μ, respectively . The representation

$$\mu = \mu^+ - \mu^-$$

is called the Jordan decomposition of μ.

Exercises

1. Suppose $(X, \mathcal{M}, \mu)$ is a positive measure space and $d\lambda = f d\mu$, where $f \in L^1(\mu)$. Prove that $d \mid \lambda \mid = \mid f \mid d\mu$.

2. Suppose λ, μ, and ν are real measures defined on the same σ-algebra and $\lambda \leq \mu$ and $\lambda \leq \nu$. Prove that

$$\lambda \leq \min(\mu, \nu)$$

where

$$\min(\mu, \nu) = \frac{1}{2}(\mu + \nu - \mid \mu - \nu \mid).$$

3. Suppose $\mu : \mathcal{M} \to \mathbf{C}$ is a complex measure and $f, g : X \to \mathbf{R}$ measurable functions. Show that

$$\mid \mu(f \in A) - \mu(g \in A) \mid \leq \mid \mu \mid (f \neq g)$$

for every $A \in \mathcal{R}$.

5.2. The Lebesque Decomposition and the Radon-Nikodym Theorem

Let μ be a positive measure on $\mathcal{M}$ and λ a positive or complex measure on $\mathcal{M}$. The measure λ is said to be absolutely continuous with respect to μ (abbreviated $\lambda << \mu$) if $\lambda(A) = 0$ for every $A \in \mathcal{M}$ for which $\mu(A) = 0$. If we define

$$\mathcal{Z}_\lambda = \{A \in \mathcal{M};\ \lambda(A) = 0\}$$

it follows that $\lambda << \mu$ if and only if

$$\mathcal{Z}_\mu \subseteq \mathcal{Z}_\lambda.$$

For example, $\gamma_n << v_n$ and $v_n << \gamma_n$.

The measure λ is said to be concentrated on $E \in \mathcal{M}$ if $\lambda = \lambda^E$, where $\lambda^E(A) =_{def} \lambda(E \cap A)$ for every $A \in \mathcal{M}$. This is equivalent to the hypothesis that $A \in \mathcal{Z}_\lambda$ if $A \in \mathcal{M}$ and $A \cap E = \phi$. Thus if $E_1, E_2 \in \mathcal{M}$, where $E_1 \subseteq E_2$, and λ is concentrated on E_1, then λ is concentrated on E_2. Moreover, if $E_1, E_2 \in \mathcal{M}$ and λ is concentrated on both E_1 and E_2, then λ is concentrated on $E_1 \cap E_2$. Two measures λ_1 and λ_2 are said to be mutually singular (abbreviated $\lambda_1 \perp \lambda_2$) if there exist disjoint measurable sets E_1 and E_2 such that λ_1 is concentrated on E_1 and λ_2 is concentrated on E_2.

Theorem 5.2.1. *Let μ be a positive measure and λ, λ_1, and λ_2 complex measures.*

(i) *If $\lambda_1 << \mu$ and $\lambda_2 << \mu$, then $(\alpha_1\lambda_1 + \alpha_2\lambda_2) << \mu$ for all complex numbers α_1 and α_2.*

(ii) *If $\lambda_1 \perp \mu$ and $\lambda_2 \perp \mu$, then $(\alpha_1\lambda_1 + \alpha_2\lambda_2) \perp \mu$ for all complex numbers α_1 and α_2.*

(iii) *If $\lambda << \mu$ and $\lambda \perp \mu$, then $\lambda = 0$.*

(iv) *If $\lambda << \mu$, then $\mid \lambda \mid << \mu$.*

PROOF. The properties (i) and (ii) are simple to prove and are left as exercises.

To prove (iii) suppose $E \in \mathcal{M}$ is a μ-null set and $\lambda = \lambda^E$. If $A \in \mathcal{M}$, then $\lambda(A) = \lambda(A \cap E)$ and $A \cap E$ is a μ-null set. Since $\lambda << \mu$ it follows that $A \cap E \in Z_\lambda$ and, hence, $\lambda(A) = \lambda(A \cap E) = 0$. This proves (iii)

To prove (iv) suppose $A \in \mathcal{M}$ and $\mu(A) = 0$. If $(A_n)_{n=1}^{\infty}$ is measurable partition of A, then $\mu(A_n) = 0$ for every n. Since $\lambda << \mu$, $\lambda(A_n) = 0$ for every n and we conclude that $\mid \lambda \mid (A) = 0$. This proves (vi).

Theorem 5.2.2. *Let μ be a positive measure on $\mathcal{M}$ and λ a complex measure on $\mathcal{M}$. Then the following conditions are equivalent:*

(a) $\lambda << \mu$.

(b) *To every $\varepsilon > 0$ there corresponds a $\delta > 0$ such that $\mid \lambda(E) \mid < \varepsilon$ for all $E \in \mathcal{M}$ with $\mu(E) < \delta$.*

If λ is a positive measure, the implication $(a) \Rightarrow (b)$ in Theorem 5.2.2 is, in general, wrong. To see this take $\mu = \gamma_1$ and $\lambda = v_1$. Then $\lambda << \mu$ and if we choose $A_n = [n, \infty[$, $n \in \mathbf{N}_+$, then $\mu(A_n) \to 0$ as $n \to \infty$ but $\lambda(A_n) = \infty$ for each n.

PROOF. (a)$\Rightarrow$(b). If (b) is wrong there exist an $\varepsilon > 0$ and sets $E_n \in \mathcal{M}$, $n \in \mathbf{N}_+$, such that $\mid \lambda(E_n) \mid \geq \varepsilon$ and $\mu(E_n) < 2^{-n}$. Set

$$A_n = \cup_{k=n}^{\infty} E_k \text{ and } A = \cap_{n=1}^{\infty} A_n.$$

Since $A_n \supseteq A_{n+1} \supseteq A$ and $\mu(A_n) < 2^{-n+1}$, it follows that $\mu(A) = 0$ and using that $\mid \lambda \mid (A_n) \geq \mid \lambda(E_n) \mid$, Theorem 1.1.2 (f) implies that

$$\mid \lambda \mid (A) = \lim_{n \to \infty} \mid \lambda \mid (A_n) \geq \varepsilon.$$

This contradicts that $\mid \lambda \mid << \mu$.

(b)$\Rightarrow$(a). If $E \in \mathcal{M}$ and $\mu(E) = 0$ then to each $\varepsilon > 0$, $\mid \lambda(E) \mid < \varepsilon$, and we conclude that $\lambda(E) = 0$. The theorem is proved.

Theorem 5.2.3. *Let μ be a σ-finite positive measure and λ a real measure on $\mathcal{M}$.*

(a) **(The Lebesgue Decomposition of λ)** *There exists a unique pair of real measures λ_a and λ_s on$\mathcal{M}$ such that*

$$\lambda = \lambda_a + \lambda_s,\ \lambda_a << \mu, \text{ and } \lambda_s \perp \mu.$$

If λ is a finite positive measure, λ_a and λ_s are finite positive measures.

(b) **(The Radon-Nikodym Theorem)** *There exits a unique $g \in L^1(\mu)$ such that*

$$d\lambda_a = g d\mu.$$

If λ is a finite positive measure, $g \geq 0$ a.e. $[\mu]$.

The proof of Theorem 5.2.3 is based on the following

Lemma 5.2.1. *Let (X, M, μ) be a finite positive measure space and suppose $f \in L^1(\mu)$.*

(a) *If $a \in \mathbf{R}$ and*

$$\int_E f d\mu \leq a\mu(E),\ \text{all } E \in \mathcal{M}$$

then $f \leq a$ a.e. $[\mu]$.

(b) *If $b \in \mathbf{R}$ and*

$$\int_E f d\mu \geq b\mu(E),\ \text{all } E \in \mathcal{M}$$

then $f \geq b$ a.e. $[\mu]$.

PROOF. (a) Set $g = f - a$ so that

$$\int_E g d\mu \leq 0,\ \textit{all } E \in \mathcal{M}.$$

Now choose $E = \{g > 0\}$ to obtain

$$0 \geq \int_E g d\mu = \int_X \chi_E g d\mu \geq 0$$

as $\chi_E g \geq 0$ a.e. $[\mu]$. But then Example 2.1.2 yields $\chi_E g = 0$ a.e. $[\mu]$ and we get $E \in Z_\mu$. Thus $g \leq 0$ a.e. $[\mu]$ or $f \leq a$ a.e. $[\mu]$.

Part (b) follows in a similar way as Part (a) and the proof is omitted here.

PROOF. Uniqueness: (a) Suppose $\lambda_a^{(k)}$ and $\lambda_s^{(k)}$ are real measures on $\mathcal{M}$ such that

$$\lambda = \lambda_a^{(k)} + \lambda_s^{(k)},\ \lambda_a^{(k)} << \mu,\ \text{and}\ \lambda_s^{(k)} \perp \mu$$

for $k = 1, 2$. Then

$$\lambda_a^{(1)} - \lambda_a^{(2)} = \lambda_s^{(2)} - \lambda_s^{(1)}$$

and

$$\lambda_a^{(1)} - \lambda_a^{(2)} << \mu \text{ and } \lambda_a^{(1)} - \lambda_a^{(2)} \perp \mu.$$

Thus by applying Theorem 5.2.1, $\lambda_a^{(1)} - \lambda_a^{(2)} = 0$ and $\lambda_a^{(1)} = \lambda_a^{(2)}$. From this we conclude that $\lambda_s^{(1)} = \lambda_s^{(2)}$.

(b) Suppose $g_k \in L^1(\mu), k = 1, 2$, and

$$d\lambda_a = g_1 d\mu = g_2 d\mu.$$

Then $h d\mu = 0$ where $h = g_1 - g_2$. But then

$$\int_{\{h>0\}} h d\mu = 0$$

and it follows that $h \leq 0$ a.e. $[\mu]$. In a similar way we prove that $h \geq 0$ a.e. $[\mu]$. Thus $h = 0$ in $L^1(\mu)$, that is $g_1 = g_2$ in $L^1(\mu)$.

Existence: The beautiful proof that follows is due to von Neumann.

First suppose that μ and λ are finite positive measures and set $\nu = \lambda + \mu$. Clearly, $L^1(\lambda) \supseteq L^1(\nu) \supseteq L^2(\nu)$. Moreover, if $f : X \to \mathbf{R}$ is measurable

$$\int_X \mid f \mid d\lambda \leq \int_X \mid f \mid d\nu \leq \sqrt{\int_X f^2 d\nu}\sqrt{\nu(X)}$$

and from this we conclude that the map

$$f \to \int_X f d\lambda$$

is a continuous linear functional on $L^2(\nu)$. Therefore, in view of Theorem 4.2.2, there exists a $g \in L^2(\nu)$ such that

$$\int_X f d\lambda = \int_X f g d\nu \text{ all } f \in L^2(\nu).$$

Suppose $E \in \mathcal{M}$ and put $f = \chi_E$ to obtain

$$0 \leq \lambda(E) = \int_E g d\nu$$

and, since $\nu \geq \lambda$,

$$0 \leq \int_E g d\nu \leq \nu(E).$$

But then Lemma 5.2.1 implies that $0 \leq g \leq 1$ a.e. $[\nu]$. Therefore, without loss of generality we can assume that $0 \leq g(x) \leq 1$ for all $x \in X$ and, in addition, as above

$$\int_X f d\lambda = \int_X f g d\nu \text{ all } f \in L^2(\nu)$$

that is

$$\int_X f(1-g) d\lambda = \int_X f g d\mu \text{ all } f \in L^2(\nu).$$

Put $A = \{0 \leq g < 1\}$, $S = \{g = 1\}$, $\lambda_a = \lambda^A$, and $\lambda_s = \lambda^S$. Note that $\lambda = \lambda^A + \lambda^S$. The choice $f = \chi_S$ gives $\mu(S) = 0$ and hence $\lambda_s \perp \mu$. Moreover, the choice

$$f = (1 + ... + g^n)\chi_E$$

where $E \in \mathcal{M}$, gives

$$\int_E (1 - g^{n+1}) d\lambda = \int_E (1 + ... + g^n) g d\mu.$$

By letting $n \to \infty$ and using monotone convergence

$$\lambda(E \cap A) = \int_E h d\mu.$$

where

$$h = \lim_{n \to \infty} (1 + ... + g^n) g.$$

Since h is non-negative and

$$\lambda(A) = \int_X h d\mu$$

it follows that $h \in L^1(\mu)$. Moreover, the construction above shows that $\lambda = \lambda_a + \lambda_s$.

In the next step we assume that μ is a σ-finite positive measure and λ a finite positive measure. Let $(X_n)_{n=1}^{\infty}$ be a measurable partition of X such that $\mu(X_n) < \infty$ for every n. Let n be fixed and apply Part (a) to the pair μ^{X_n} and λ^{X_n} to obtain finite positive measures $(\lambda^{X_n})_a$ and $(\lambda^{X_n})_s$ such that

$$\lambda^{X_n} = (\lambda^{X_n})_a + (\lambda^{X_n})_s,\ (\lambda^{X_n})_a << \mu^{X_n},\ \text{and}\ (\lambda^{X_n})_s \perp \mu^{X_n}$$

and

$$d(\lambda^{X_n})_a = h_n d\mu^{X_n}\ (\text{or}\ (\lambda^{X_n})_a = h_n \mu^{X_n})$$

where $0 \leq h_n \in L^1(\mu^{X_n})$. Without loss of generality we can assume that $h_n = 0$ off X_n and that $(\lambda^{X_n})_s$ is concentrated on $A_n \subseteq X_n$ where $A_n \in \mathcal{Z}_\mu$. In particular, $(\lambda^{X_n})_a = h_n \mu$. Now

$$\lambda = h\mu + \Sigma_{n=1}^{\infty} (\lambda^{X_n})_s$$

where

$$h = \Sigma_{n=1}^{\infty} h_n$$

and

$$\int_X h d\mu \leq \lambda(X) < \infty.$$

Thus $h \in L^1(\mu)$. Moreover, $\lambda_s =_{def} \Sigma_{n=1}^{\infty} (\lambda^{X_n})_s$ is concentrated on $\cup_{n=1}^{\infty} A_n \in \mathcal{Z}_\mu$. Hence $\lambda_s \perp \mu$.

Finally if λ is a real measure we apply what we have already proved to the positive and negative variations of λ and we are done.