POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER II

SELF LEARNING MATERIAL

PAPER : COR 2.1
(Pure & Applied Streams)

Real Analysis II

Complex Analysis II

Functional Analysis II

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

Content Writers

Block - I : Real Analysis II	Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani
Block - II : Complex Analysis II	Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani
Block - III : Functional Analysis II	Dr. Animesh Biswas Professor, Department of Mathematics, University of Kalyani

July, 2022

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 Message

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, 2020 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.) Manas Kumar Sanyal, 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 well 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 coordinated 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 written and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Board of Studies Members of Department of Mathematics, Directorate of Open and Distance Learning (DODL), University of Kalyani

Sl No.	Name & Designation	Role
1	Dr. Animesh Biswas, Professor & Head, Dept. of Mathematics, KU	Chairperson
2	Dr. Pulak Sahoo, Professor, Dept. of Mathematics, KU	Member
3	Dr. Sahidul Islam, Assistant Professor, Dept. of Mathematics, KU	Member
4	Dr. Sushanta Kumar Mohanta, Professor, Dept. of Mathematics, West Bengal State University	External Nominated Member
5	Dr. Biswajit Mallick, Assistant Professor (Cont.), Department of Mathematics, DODL, KU	Member
6	Ms. Audrija Choudhury, Assistant Professor (Cont), Department of Mathematics, DODL, KU	Member
7	Director, DODL, KU	Convener

CORE COURSE

COR 2.1
Marks : 100 (SEE : 80; IA : 20); Credit : 6

Real Analysis II (Marks : 32 (SEE: 25; IA: 07))
Complex Analysis II (Marks : 32 (SEE: 25; IA: 07))
Functional Analysis II (Marks : 36 (SEE: 30; IA: 06))

Syllabus

Block I

- **Unit 1:** The Lebesgue measure: Definition of the Lebesgue outer measure on the power set of R, countable subadditivity, countable additivity, Carathéodory's definition of the Lebesgue measure and basic properties. Measurability of an interval (finite or infinite)
- **Unit 2:** Characterizations of measurable sets by open sets, G_δ sets, closed sets and F_σ sets. Measurability of Borel sets, Existence of non-measurable sets.
- **Unit 3:** Measurable functions : Definition on a measurable set in $\mathbb{R}$ and basic properties, Sequences of measurable functions
- **Unit 4:** Simple functions, Measurable functions as the limits of sequences of simple functions
- **Unit 5:** Lusin's theorem on restricted continuity of measurable functions, Egoroff's theorem, Convergence in measure
- **Unit 6:** The Lebesgue integral : Integrals of non-negative simple functions, The integral of non-negative measurable functions on arbitrary measurable sets in $\mathbb{R}$ using integrals of non-negative simple functions, Monotone convergence theorem and Fatou's lemma.
- **Unit 7:** The integral of Measurable functions and basic properties, Absolute character of the integral, Dominated convergence theorem,
- **Unit 8:** Inclusion of the Riemann integral, Riesz-Fischer theorem on the completeness of the space of Lebesgue integrable functions.
- **Unit 9:** Lebesgue integrability of the derivative of a function of bounded variation on an interval. Descriptive characterization of the Lebesgue integral on intervals by absolutely continuous functions.

Block II

- **Unit 10:** Contour integration. Conformal mapping, Bilinear transformation.
- **Unit 11:** Idea of analytic continuation. Multivalued functions – branch point. Idea of winding number.
- **Unit 12:** Zeros of an analytic function. Singularities and their classification.
- **Unit 13:** Limit points of zeros and poles. Riemann's theorem. Weierstrass-Casorati theorem, Behaviour of a function at the point at infinity.
- **Unit 14:** Theory of residues. Argument principle. Rouche's theorem. Maximum modulus theorem. Schwarz lemma.

Block III

- **Unit 15:** Linear operators, Linear operators on normed linear spaces, continuity
- **Unit 16:** Bounded linear operators, norm of an operator, various expressions for the norm. Spaces of bounded linear operators. Inverse of an operator.
- **Unit 17:** Linear functionals. Hahn-Banach theorem (without proof), simple applications. Normed conjugate space and separability of the space. Uniform boundedness principle, simple application.
- **Unit 18:** Inner product spaces, Cauchy Schwarz's inequality, the induced norm, polarization identity, parallelogram law. Orthogonality, Pythagoras Theorem, orthonormality, Bessel's inequality and its generalisation.
- **Unit 19:** Hilbert spaces, orthogonal complement, projection theorem.
- **Unit 20:** The Riesz's representation theorem. Convergence of series corresponding to orthogonal sequence, Fourier coefficient, Perseval's identity.

Contents

Unit 1

Course Structure

- Definition of the Lebesgue outer measure on the power set of $\mathbb{R}$
- Properties of outer measure
- Carathéodory's definition of the Lebesgue measure and basic properties.
- Measurability of an interval (finite or infinite)

1.1 Introduction

The mathematical concept of measure is a generalisation of length in $\mathbb{R}$, area in $\mathbb{R}^{\nVdash}$ and volume in $\mathbb{R}^{\nVDash}$. We will develop measure theory for the subsets of $\mathbb{R}$, however it can be generalised for any arbitrary sets. Let us try to have some idea of the length of an interval, say $[0, 1]$. The most intuitive idea of length is to just find the difference between the extreme points of the interval, that is in this case $1-0 = 1$. Also, adding or subtracting the point 0 or 1 does not make any difference, that is, the length of $(0, 1]$, $[0, 1)$ and $(0, 1)$ are essentially the same. But what will happen if any one of the end points are infinite? Then intuitively, the length should also be infinite (in this case we will define the algebraic operations related to the points at infinity, viz., $-\infty$ and $+\infty$). The degenerate intervals, that is, the intervals where the end points are equal, deserve mention here. Say we have the interval $[a, a]$, or this can also be written as $\{a\}$. Then the length of this interval will be $a - a = 0$. Also, taking cue from the idea of length in the physical perspective, it is clear that any subset of an interval should have length less than that of the parent interval and translations of intervals don't change the length of intervals. Now, one can also question the length of the union of intervals. The case of $[0, 1]$ and $[3, 4]$ is easy and straightforward. We need to just sum up the lengths of each intervals and the resulting length will be $1 + 1 = 2$. What happens when we consider the union of non-disjoint intervals, say for example $[2, 4]$ and $[3, 5]$. The intervals are overlapping and its union is simply $[2, 5]$ having length 3 which is not equal to the sum of the summands that is $2 + 2 = 4$. This idea can be extended to countable union of intervals as well. So, if we consider the length as a function from the set of all intervals over $\mathbb{R}$ to the set of extended reals(definition in the next section), and denote the length of an interval I as $l(I)$, then the properties of length are straightforward and given below. It can also be observed that this function actually resembles the idea of length in the physical world.

1. $l(I) \geq 0$ for every interval I;

2. $l(\emptyset) = l([a,a]) = l(\{a\}) = 0$ for any $a \in \mathbb{R}$;
3. $l((a,b)) = l([a,b]) = l((a,b]) = l([a,b)] = b - a$ for $b \geq a$;
4. $I \subset J \Rightarrow l(I) \leq l(J)$;
5. $l(I + x) = l(I)$, for $x \in \mathbb{R}$;
6. If I and J are two disjoint intervals, then $l(I \cup J) = l(I) + l(J)$;
7. For any mutually disjoint sequence of intervals $\{I_n\}$, we have

$$l\left(\bigcup_{n=1}^{\infty} I_n\right) = \sum_{n=1}^{\infty} l(I_n).$$

We shall construct a measure function for any general set $A \subset \mathbb{R}$ that generalises the idea of length, that is, the measure function that we will define should satisfy the properties of length.

Objectives

After reading this unit, you will be able to

- define outer measure of a set using the idea of length of intervals and state its basic properties
- define the measure of a set with the help of outer measure and state its basic properties
- see that the idea of measure is actually a generalisation of the idea of length of intervals (finite or infinite)

1.2 Measure Theory

Before starting off with the technical stuff, let us be equipped with the basic definitions that we shall use henceforth.

1.2.1 Extended Real Number System

The Real numbers along with the two infinite numbers $-\infty$ and ∞ constitute the extended real numbers. It is denoted by $\mathbb{R}^*$ and is equal to $\mathbb{R} \cup \{-\infty, +\infty\}$. Having adjoined the two numbers, we also need to define the "interaction" of them with the other real numbers. The definitions/axioms are given below:

1. For any $x \in \mathbb{R}$, $x + \infty = \infty + x = x - (-\infty) = \infty$. Also, $x + (-\infty) = -\infty + x = x - \infty = -\infty$;
2. If $x > 0$, then $\infty.x = x.\infty = \infty$ and $x.(-\infty) = (-\infty).x = -\infty$;
3. If $x < 0$, then $\infty.x = x.\infty = -\infty$ and $x.(-\infty) = (-\infty).x = \infty$;
4. $\infty + \infty = \infty$ and $(-\infty) + (-\infty) = -\infty$;
5. $\infty.\infty = \infty$;
6. $\infty.0 = 0.\infty = (-\infty).0 = 0.(-\infty) = 0$.

The operations such as $\infty + (-\infty)$, or $-\infty + \infty$, or $\infty.(-\infty)$, $(-\infty).\infty$, or $(-\infty).(-\infty)$ are not defined. Also division by ∞ or $-\infty$ is also not defined.

1.2.2 Algebra and σ algebra of sets

Let $\mathscr{S}$ be a collection of subsets of $\mathbb{R}$.

Definition 1.2.1. $\mathscr{S}$ is said to form an *Algebra* over $\mathbb{R}$ if it satisfies the following conditions:

1. $\emptyset \in \mathscr{S}$;
2. If $A \in \mathscr{S}$, the $\mathbb{R} \setminus A \in \mathscr{S}$;
3. If $A, B \in \mathscr{S}$, then $A \cup B \in \mathscr{S}$.

It is easy to check that $\mathscr{S}$ is an algebra if it is closed under finite union and complement of sets. Thus, $\mathscr{S}$ is closed under finite intersection also.

Definition 1.2.2. $\mathscr{S}$ is said to form an σ-*Algebra* over $\mathbb{R}$ if it satisfies the following conditions:

1. $\emptyset \in \mathscr{S}$;
2. If $A \in \mathscr{S}$, the $\mathbb{R} \setminus A \in \mathscr{S}$;
3. If $\{A_n\}$ be a sequence of sets in $\mathscr{S}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{S}$.

It is evident that a σ-algebra is always an algebra but the converse is not true always.

Example 1.2.3.
1. The power set $\mathscr{P}(\mathbb{R})$ forms both an algebra and σ-algebra over $\mathbb{R}$.
2. For any $a, b, c \in \mathbb{R}$, the collection $\{\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{a, c\}, \{a, b, c\}\}$ is also both algebra and σ-algebra over $\mathbb{R}$.
3. Let $\mathscr{S}$ be the collection of all sets that are either finite or have finite complements. Then $\mathscr{S}$ forms an algebra but not a σ-algebra over $\mathbb{R}$.

Theorem 1.2.4. Let $\mathscr{S}_1$ and $\mathscr{S}_2$ are two σ-algebras over $\mathbb{R}$. Then $\mathscr{S}_1 \cap \mathscr{S}_2$ forms a σ-algebra over $\mathbb{R}$.

Proof. Left as exercise. □

It is needless to say that any finite intersection of σ-algebras will also be a σ-algebra. What about the arbitrary intersection of -algebras over $\mathbb{R}$? The next definition gives an idea in this direction.

Definition 1.2.5. Let $\mathscr{C}$ be a family of sets in $\mathbb{R}$ and consider all the σ-algebras over $\mathbb{R}$ containing $\mathscr{C}$. Then the intersections of all such σ-algebras will also be a σ-algebra over $\mathbb{R}$ containing $\mathscr{C}$. Such a σ-algebra is called the σ-algebra generated by $\mathscr{C}$ and is the smallest σ-algebra containing $\mathscr{C}$ (Check!).

We are now equipped with the basic definitions and terminologies to finally attempt to define the measure of a set.

1.2.3 Lebesgue Outer Measure

We want to define measure in such a way so that it generalises the definition length of intervals that we discussed previously. So, we will make use of the length to define the following.

Definition 1.2.6. Let $\mathscr{P}$ be the collection of all subsets of $\mathbb{R}$. We define a function $m^* : \mathscr{P} \Rightarrow \mathbb{R}^*$ as follows:

$$m^*(A) = \inf \left\{ \sum_{k=1}^{\infty} l(I_k) : \ A \subset \bigcup_{k=1}^{\infty} I_k, \ \ I_k \ \text{ are open intervals} \right\},$$

for any $A \in \mathscr{P}$. Then the function m^* is called *Lebesgue outer measure* and is define for any subset of $\mathbb{R}$.

The idea is that, if A is any interval, then the outer measure will simply be the length of the interval. Otherwise, we try to cover A by intervals (any kind) and take the infimum of the sum of the intervals that cover A. It can be seen that the length of the intervals (a, b), $[a, b)$, $(a, b]$ and $[a, b]$ are equal and hence, we can use any kind of intervals to cover A and the definition of the outer measure can be correspondingly defined. Let us check whether this outer measure serves our purpose or not. Let us list out the properties of m^* as the following theorem.

Theorem 1.2.7. Let m^* be the Lebesgue outer measure defined on $\mathscr{P}$. Then it satisfies the following properties.

1. $0 \leq m*(A) \leq \infty$ for every $A \in \mathscr{P}$; (non-negative extended real valued)
2. $m^*(A) \leq m^*(B)$ for $A \subseteq B$ for every $A, B \in \mathscr{P}$; (monotonicity)
3. $m^*(\emptyset) = 0$;
4. $m^*(\{a\}) = 0$; (points are dimensionless)
5. $m^*(I) = l(I)$ for any interval I;
6. $m^*(A + x) = m^*(A)$ for every $A \in \mathscr{P}$; (translation invariance)
7. $m^* \left(\bigcup_{k=1}^{\infty} A_k \right) \leq \sum_{k=1}^{\infty} m^*(A_k)$ for any sequence of sets $\{A_k\}$ in $\mathscr{P}$; (countable subadditivity)

Proof. 1. Follows from definition.

2. Let $A \subseteq B$ and $\{I_n\}$ be a collection of open intervals such that

$$B \subset \bigcup_{n=1}^{\infty} I_n.$$

Then $\{I_n\}$ also covers A. By the definition of outer measure,

$$\sum_{n=1}^{\infty} l(I_n) \geq m^*(A).$$

The above equation is true for all covers (by means of open intervals) of B and hence,

$$m^*(A) \leq \inf \left\{ \sum_{n=1}^{\infty} l(I_n) : \ B \subset \bigcup_{n=1}^{\infty} I_n, \ \ I_k \ \text{ are open intervals} \right\} = m^*(B).$$

3. We know that $\emptyset \subset I$ for every open interval I. Let $\epsilon > 0$ be an arbitrary real number. Consider the sequence of intervals $\{I_n\}$, where $I_n = \left(-\frac{\epsilon}{2^{n+2}}, \frac{\epsilon}{2^{n+2}}\right)$. Then $\{I_n\}$ is a collection of open intervals covering $\emptyset$. Thus,

$$\begin{aligned} 0 \leq m^*(\emptyset) &\leq \sum_{n=1}^{\infty} l(I_n) \\ &= \sum_{n=1}^{\infty} \frac{\epsilon}{2^{n+1}} \\ &= \frac{\epsilon}{2^2} \sum_{n=0}^{\infty} \frac{1}{2^n} = \frac{\epsilon}{2^2}.2 = \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

Since ϵ is arbitrary, so $m^*(\emptyset) = 0$.

4. Let $\epsilon > 0$ be arbitrary. Consider the sequence $I_n = \left(a - \frac{\epsilon}{2^{n+2}}, a + \frac{\epsilon}{2^{n+2}}\right)$. Then the sequence $\{I_n\}$ covers $\{a\}$. Now do it similarly as the previous one.

5. First suppose that $I = [a, b]$. Let $\epsilon > 0$. Then $\left(a - \frac{\epsilon}{2}, b + \frac{\epsilon}{2}\right)$ covers I and

$$l\left(a - \frac{\epsilon}{2^{n+2}}, a + \frac{\epsilon}{2^{n+2}}\right) = b - a + \epsilon.$$

Since ϵ is arbitrary, so

$$m^*(I) \leq b - a = l(I). \tag{1.2.1}$$

Next, let $\{I_n\}$ be a covering of $[a, b]$ by bounded open intervals. By the Heine-Borel Theorem, there exists a finite subset A of I_n's covering $[a, b]$. So, $a \in I_1$ for some $I_1 = (a_1, b_1) \in A$. Also, if $b_1 \leq b$, then $b_1 \in I_2$ for some $I_2 = (a_2, b_2) \in A$. Similarly we can construct $I_1, I_2, \ldots, I_k$ such that $a_i < b_{i-1} < b_i$. Then

$$\begin{aligned} \sum_{n=1}^{\infty} l(I_n) &\geq \sum_{i=1}^{k} l(I_i) \\ &= \sum_{i=1}^{k} (b_i - a_i) \\ &= (b_k - a_k) + (b_{k-1} - a_{k-1}) + \ldots + (b_1 - a_1) \\ &= b_k - (a_k - b_{k-1}) - \ldots - (a_2 - b_1) - a_1 \\ &> b_k - a_1. \end{aligned}$$

Since $a_1 < a$ and $b_k > b$ so

$$\sum_{n=1}^{\infty} l(I_n) > b - a. \tag{1.2.2}$$

Since $\{I_n\}$ is arbitrary, so by the definition of outer measure,

$$m^*(I) \geq b - a. \tag{1.2.3}$$

From equation (1.2.1) and (1.2.3), we get the desired result.

Next consider I as any arbitrary bounded interval. Then for any $\epsilon > 0$, there is a closed interval $J \subset I$ such that $l(J) > l(I) - \epsilon$. Notice that $m^*(I) \leq m^*(\overline{I})$ by monotonicity. So,

$$\begin{aligned} l(I) - \epsilon < l(J) &= m^*(J) \text{ [by previous argument]} \\ &\leq m^*(I) \text{ [by monotonicity]} \\ &\leq m^*(\overline{I}) \text{ [by monotonicity]} \\ &= l(\overline{I}) \text{ [by previous argument]} \\ &= l(I) \text{ [since } I \text{ is a bounded interval]}. \end{aligned}$$

Hence, $l(I) - \epsilon < m^*(I) \leq l(I)$. Since $\epsilon > 0$ is arbitrary, so $l(I) = m^*(I)$.

Finally let I be an unbounded interval. Then then given any natural number $n \in \mathbb{N}$, there is a closed interval $J \subset I$ such that $l(J) = n$. Hence, $m^*(I) \geq m^*(J) = l(J) = n$. Since $m^*(I) \geq n$ and $n \in \mathbb{N}$ is arbitrary, so $m^*(I) = \infty = l(I)$.

6. Let $m^*(A) = M < \infty$. Then for all $\epsilon > 0$, there exists a sequence $\{I_n\}$ bounded open intervals such that $A \subset \bigcup_n^\infty I_n$ and

$$\sum_{n=1}^{\infty} l(I_n) < M + \epsilon.$$

Hence, for $x \in \mathbb{R}$, $\{I_n + x\}$ is a covering of $A + x$ and so

$$m^*(A + x) \leq \sum_{n=1}^{\infty} l(I_n + x) = \sum_{n=1}^{\infty} l(I_n) < M + \epsilon.$$

Hence,

$$m^*(A + x) \leq M. \tag{1.2.4}$$

Now, let $\{J_n\}$ be a collection of bounded open intervals such that $A + x \subset \bigcup_n^\infty J_n$. Assume that $\sum_{n=1}^{\infty} l(J_n) < M$. Then $\{J_n - x\}$ is a covering of A and $\sum_{n=1}^{\infty} l(J_n - x) = \sum_{n=1}^{\infty} l(J_n) < M$, a contradiction. So, $\sum_{n=1}^{\infty} l(J_n) \geq M$ and hence

$$m^*(A + x) \geq M. \tag{1.2.5}$$

From equations (1.2.4) and (1.2.5), we get the desired result.

Next, let $m^*(A) = \infty$. The for any sequence $\{I_n\}$ bounded open intervals such that $A \subset \bigcup_n^\infty I_n$, we must have

$$\sum_{n=1}^{\infty} l(I_n) = \infty.$$

Consider $A + x$. For any sequence $\{J_n\}$ of bounded open intervals such that $A + x \subset \bigcup_n^\infty J_n$, the collection $\{J_n - x\}$ is a set of bounded open intervals such that $A \subset \bigcup_{n=1}^\infty (J_n - x)$. So, $\sum_{n=1}^{\infty} l(J_n - x) = \infty$.

But, $l(J_n) = l(J_n - x)$, so we must have $\sum_{n=1}^{\infty} l(J_n) = \infty$. Since $\{J_n\}$ is an arbitrary collection of bounded open intervals covering $A + x$, we must have $m^*(A + x) = m^*(A)$.

7. Let $\{A_k\}$ be any countable collection of sets. The result holds trivially if $m^*(A_k) = \infty$ for some k. So without loss of generality, we assume that $m^*(A_k) < \infty$ for each k. Then for all $\epsilon > 0$ and for each $k \in \mathbb{N}$, there is a countable set of open intervals $\{I_{k,m}\}$ such that $A_k \subset \bigcup_{m=1}^{\infty} I_{k,m}$ and $\sum_{m=1}^{\infty} l(I_{k,m}) < m^*(A_k) + \frac{\epsilon}{2^k}$. Then $\{I_{k,m}\}$ for $k, m \in \mathbb{N}$ a countable collection of open intervals that cover $\bigcup_{k=1}^{\infty} A_k$. So,

$$\begin{aligned} m^*\left(\bigcup_{k=1}^{\infty} A_k\right) &\leq \sum_{k,m} l(I_{k,m}) \\ &= \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} l(I_{k,m}) \\ &< \sum_{k=1}^{\infty}\left(m^*(A_k) + \frac{\epsilon}{2^k}\right) \\ &= \sum_{k=1}^{\infty} m^*(A_k) + \epsilon. \end{aligned}$$

Hence, $m^*\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} m^*(A_k)$.

□

The above properties are nearly exhaustive list of properties satisfied by the outer measure. Now, let us check whether it is our desired measure function. We see that nearly all the properties that we sought for are satisfied by the outer measure but for the countable additivity, that is, for mutually disjoint collection of sets $\{A_k\}$ of sets in $\mathscr{P}$, $m^*\left(\bigcup_{k=1}^{\infty}\right) \neq \sum_{k=1}^{\infty} m^*(A_k)$. But the outer measure does not actually satisfy additivity. But since we are so close, we might need to refine the definition using the outer measure a bit so that this problem is solved. In the year 1914, Caratheodory formulated a measurability criteria using the outer measure that is formally accepted as the definition of measure. We will discuss more on this in the next section.

Example 1.2.8. The Cantor set C is uncountable with outer measure zero. Let C_n denote the union of the intervals left at the nth stage while constructing the Cantor set. One may note that C_n consists of 2^n closed intervals, each of length 3^{-n}. Thus,

$$m^*(C_n) \leq 2^n.3^{-n}.$$

But, any point of C must be in one of the intervals comprising the union C_n for each $n \in \mathbb{N}$ and as such $C \subset C_n$ for each $n \in \mathbb{N}$. Hence,

$$m^*(C) \leq \left(\frac{2}{3}\right)^n.$$

This being true for each $n \in \mathbb{N}$, letting $n \to \infty$ gives $m^*(C) = 0$.

Exercise 1.2.9. 1. Show that any countable set has outer measure zero. Hence show that $[0, 1]$ is uncountable. Is the converse true? Justify.

2. If $m^*(A) = 0$, then show that $m^*(A \cup B) = m^*(A) + m^*(B) = m^*(B)$.

1.2.4 Lebesgue Measure

We begin by defining the measurability of a set due to Caratheodory.

Definition 1.2.10. Let $A \subset \mathbb{R}$. Then A is said to be Lebesgue measurable if for any subset E of $\mathbb{R}$,

$$m^*(E) = m^*(E \cap A) + m^*(E \setminus A).$$

If A is measurable, then the outer measure is called the measure of A and is denoted by $m(A)$.

This above definition can be understood as the additive "interaction" of A with every subset of $\mathbb{R}$, or A "splits" every subset of $\mathbb{R}$ in an additive manner. It is to be noted that for any set E,

$$E = E \cap \mathbb{R} = E \cap (A \cup (\mathbb{R} \setminus A)) = (E \cap A) \cup (E \setminus A).$$

So, by the countable subadditivity of outer measure,

$$m^*(E) \leq m^*(E \cap A) + m^*(E \setminus A).$$

So, in order to satisfy the measurability condition, it is sufficient if we prove only the reverse inequality, that is,

$$\boxed{m^*(E) \geq m^*(E \cap A) + m^*(E \setminus A)}. \tag{1.2.6}$$

Further, if $m^*(E) = \infty$, then obviously the above inequality will hold. So, we have to further add the criteria $m^*(E) < \infty$ to check the measurability of A. Now, let us check the measurability of certain sets.

Example 1.2.11. 1. The sets $\emptyset$ and $\mathbb{R}$ are measurable (Prove it!).

2. Any set A with outer measure zero is measurable. Indeed, for any subset E with finite outer measure, $m^*(E \cap A) = 0$. Also, $m^*(E) \geq m^*(E \cap (\mathbb{R} \setminus A))$. Hence, equation (1.2.6) is satisfied for every $E \subset \mathbb{R}$. Hence, A is measurable.

3. From the definition of measurability, the complement of any measurable set is measurable.

Theorem 1.2.12. Let A and B be two measurable sets. Then $A \cup B$ and $A \cap B$ are measurable.

Proof. Let E be any set in $\mathbb{R}$. Since A, B are measurable, so

$$\begin{aligned} m^*(E) &\geq m^*(E \cap A) + m^*(E \cap (\mathbb{R} \setminus A)) \\ &= m^*(E \cap A) + m^*(E \cap (\mathbb{R} \setminus A) \cap B) + m^*(E \cap (\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B)). \end{aligned} \tag{1.2.7}$$

But, $E \cap (A \cup B) = (E \cap A) \cup (E \cap (\mathbb{R} \setminus A) \cap B)$ and $(\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B) = \mathbb{R} \setminus (A \cup B)$. So,

$$m^*(E \cap (\mathbb{R} \setminus A) \cap (\mathbb{R} \setminus B)) = m^*(E \cap (\mathbb{R} \setminus (A \cap B)) = m^*(E \setminus (A \cup B)). \tag{1.2.8}$$

Also, by sub-additivity of outer measure,

$$m^*(E \cap (\mathbb{R} \setminus A) \cap B) \geq m^*(E \cap (A \cup B)) - m^*(E \cap A). \tag{1.2.9}$$

Using (1.2.8) and (1.2.9) in (1.2.7), we get,

$$m^*(E) \geq m^*(E \cap (A \cup B)) + m^*(E \setminus (A \cup B))$$

which implies that $A \cup B$ is measurable. Now, since A and B are measurable, their complements are also and hence their union $(\mathbb{R} \setminus A) \cup (\mathbb{R} \setminus B)$ is also measurable. Since

$$A \cap B = \mathbb{R} \setminus [(\mathbb{R} \setminus A) \cup (\mathbb{R} \setminus B)],$$

so, $A \cap B$ is measurable as the complement of a measurable set is measurable. □

Using the principle of mathematical induction, one can easily show that the above theorem is true for any finite collection of measurable sets. Also, the following result easily follows from the above theorem.

Corollary 1.2.13. If A and B are measurable, then $A \setminus B$ is measurable.

Proof. Since $A \setminus B = A \cap (\mathbb{R} \setminus B)$, and both the sets being measurable, $A \setminus B$ is measurable. □

Theorem 1.2.14. If E is measurable, then $E + x$ is measurable for any real x.

Proof. Since E is measurable, so for any $A \subseteq \mathbb{R}$ and $x \in \mathbb{R}$,

$$\begin{aligned} m^*(A) &= m^*(A \cap E) + m^*(A \cap (\mathbb{R} \setminus E)) \\ &= m^*[(A \cap E) + x] + m^*[(A \cap (\mathbb{R} \setminus E)) + x] \text{ [outer measure is translation invariant]} \\ &= m^*[(A + x) \cap (E + x)] + m^*[(A + x) \cap ((\mathbb{R} \setminus E) + x)] \\ &= m^*[(A + x) \cap (E + x)] + m^*[(A + x) \cap (\mathbb{R} \setminus (E + x))] \end{aligned}$$

We replace A by $A - x$ in the above and find that

$$m^*(A) = m^*(A - x) = m^*[A \cap (E + x)] + m^*[A \cap (\mathbb{R} \setminus (E + x))]$$

which is the definition of measurability. Hence the result. □

If we denote the set of all measurable sets as $\mathscr{M}$, then we have seen that

1. $\emptyset, \mathbb{R} \in \mathscr{M}$;
2. $A \in \mathscr{M} \Rightarrow \mathbb{R} \setminus A \in \mathscr{M}$;
3. $A, B \in \mathscr{M} \Rightarrow A \cup B \in \mathscr{M}$.

Hence, $\mathscr{M}$ forms an algebra over $\mathbb{R}$. Does it form a σ-algebra? The following theorem points in that direction.

Theorem 1.2.15. If $\{E_i\}$ is any sequence of measurable sets, then $\bigcup_{i=1}^{\infty} E_i$ and $\bigcap_{i=1}^{\infty} E_i$ are measurable.

To prove the theorem, we will need the following lemma:

Lemma 1.2.16. Let $\{E_i\}_{i=1}^n$ be a finite collection of disjoint measurable sets. If $A \subseteq \mathbb{R}$, then

$$m^*\left(\bigcup_{i=1}^{n}(A \cap E_i)\right) = m^*\left(A \cap \left(\bigcup_{i=1}^{n} E_i\right)\right) = \sum_{i=1}^{n} m^*(A \cap E_i).$$

In particular, if $A = \mathbb{R}$, then $m\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} m(E_i)$. This shows that the measure function is finitely additive.

Proof. We prove this by induction. The result is obvious when $n = 1$. Suppose that it holds for some n. So,

$$m^*\left(A \cap \left(\bigcup_{i=1}^{n} E_i\right)\right) = \sum_{i=1}^{n} m^*(A \cap E_i)$$

Consider $n+1$ disjoint measurable sets E_i. Since E_{n+1} is measurable,

$$\begin{aligned} m^*\left(A \cap \left(\bigcup_{i=1}^{n+1} E_i\right)\right) &= m^*\left(A \cap \left(\bigcup_{i=1}^{n+1} E_i\right) \cap E_{n+1}\right) + m^*\left(A \cap \left(\bigcup_{i=1}^{n+1} E_i\right) \cap (\mathbb{R} \setminus E_{n+1})\right) \\ &= m^*(A \cap E_{n+1}) + m^*\left(A \cap \left(\bigcup_{i=1}^{n} E_i\right)\right) \\ &= m^*(A \cap E_{n+1}) + \sum_{i=1}^{n} m^*(A \cap E_i) \text{ [by induction hypothesis]} \\ &= \sum_{i=1}^{n+1} m^*(A \cap E_i). \end{aligned}$$

Hence the result. □

We will now prove the original theorem for countable sets.

Proof. Let $E = \bigcup_i E_i$, and let

$$\begin{aligned} H_1 &= E_1 \\ H_2 &= E_2 \setminus E_1 \\ H_3 &= E_3 \setminus (E_1 \cup E_2) \\ &\vdots \\ H_n &= E_n \setminus \left(\bigcup_{i=1}^{n-1} E_i\right). \end{aligned}$$

Then, $\{H_i\}$ is a sequence of disjoint measurable sets such that $E = \bigcup_{i=1}^{\infty} H_i$. Let $A \subseteq \mathbb{R}$. Then by the previous lemma, we see that

$$m^*(A) = m^*\left(A \cap \left(\bigcup_{i=1}^{n} H_i\right)\right) + m^*\left(A \cap \left(\mathbb{R} \setminus \left(\bigcup_{i=1}^{n} H_i\right)\right)\right) \geq \sum_{i=1}^{n} m^*(A \cap H_i) + m^*(A \cap (\mathbb{R} \setminus E)),$$

since $\mathbb{R} \setminus E \subseteq \mathbb{R} \setminus \left(\bigcup_{i=1}^{n} H_i\right)$. Letting $n \to \infty$, we have

$$m^*(A) \geq \sum_{i=1}^{\infty} m^*(A \cap H_i) + m^*(A \cap (\mathbb{R} \setminus E)).$$

Since $A \cap E = \bigcup_{i=1}^{\infty}(A \cap H_i)$, we have by the countable sub-additivity of the outer measure, $m^*(A \cap E) \leq \sum_{i=1}^{\infty} m^*(A \cap H_i)$. This, and the above gives us

$$m^*(A \cap E) + m^*(A \cap (\mathbb{R} \setminus E)) \leq m^*(A).$$

Hence E is measurable. Again, since $\bigcap_{i=1}^{\infty} E_i = \mathbb{R} \setminus \left(\bigcup_{i=1}^{\infty} (\mathbb{R} \setminus E_i) \right)$, so $\bigcap_{i=1}^{\infty} E_i$ is also measurable. Hence the proof is complete. □

Hence, $\mathscr{M}$ forms a σ-algebra over $\mathbb{R}$.

Theorem 1.2.17. If A and B are measurable, such that $A \subseteq B$, then $m(A) \leq m(B)$. Further, if $m(A) < \infty$, then $m(B \setminus A) = m(B) - m(A)$.

Proof. Since $A \subseteq B$, so we have

$$B = (B \setminus A) \cup A$$

Since $B \setminus A$ and A are disjoint, so

$$\begin{aligned} m(B) &= m(B \setminus A) + m(A) \\ &\geq m(A) \end{aligned} \tag{1.2.10}$$

Hence,

$$m(A) \leq m(B)$$

Also, by (1.2.10), we get, since $m(A) < \infty$, so

$$m(B \setminus A) = m(B) - m(A).$$

□

If $m^*(A) = \infty$, then the result can't be true. Consider $B = \mathbb{R}$ and $A = \mathbb{R} \setminus \mathbb{N}$. Then clearly, $A \subseteq B$. We know that $m^*(\mathbb{N}) = 0 < \infty$. So, by the previous theorem, $m^*(A) = m^*(\mathbb{R}) - m^*(\mathbb{N}) = \infty - 0 = \infty$. And $m^*(\mathbb{R}) = \infty$. To find $m^*(B \setminus A)$, if we apply the above theorem, then we see that, $m^*(B \setminus A) = \infty - \infty$, which is undefined. Hence, we can't apply the above theorem in this case.

Theorem 1.2.18. Every interval is measurable.

Proof. We will prove only for the open interval of type (a, ∞), $a \in \mathbb{R}$. For this, we need to show that

$$m^*(A) \geq m^*(A \cap (a, \infty)) + m^*(A \cap (-\infty, a])$$

for any subset A of $\mathbb{R}$. If $m^*(A) = \infty$, then the result is obvious. So, we assume that $m^*(A) < \infty$. Let $\epsilon > 0$ be arbitrary. Then there exists a sequence $\{I_k\}$ of open intervals such that $A \subseteq \bigcup_{k=1}^{\infty} I_k$ such that

$$\sum_k l(I_k) < m^*(A) + \epsilon.$$

For each k, let

$$I_k^1 = I_k \cap (a, \infty),\ I_k^2 = I_k \cap (-\infty, a]$$

Then, $\{I_k^1\}$ and $\{I_k^2\}$ are sequence of open intervals covering $A \cap (a, \infty)$ and $A \cap (-\infty, a]$ respectively, and $m^*(I_k^1) + m^*(I_k^2) = l(I_k^1) + l(I_k^2) = l(I_k)$. Hence,

$$m^*(A \cap (a, \infty)) + m^*(A \cap (-\infty, a]) \leq \sum_k l(I_k) < m^*(A) + \epsilon$$

Since ϵ is arbitrary, so we get

$$m^*(A \cap (a, \infty)) + m^*(A \cap (-\infty, a]) \leq m^*(A).$$

We can similarly show that $(-\infty, a)$ is measurable. Hence, for any arbitrary interval (a, b), we have

$$(a, b) = (-\infty, b) \cap (a, \infty).$$

Showing the measurability of other types of intervals are elementary and left as exercise. □

We will now attempt to show that the measure thus defined by Caratheodory, actually is the function that we wished to seek. We are left to show the additivity of measure function. We have already shown the finite additivity of measure previously. We will use that to show the case for countable number of measurable sets.

Theorem 1.2.19. Let $\{E_k\}$ be countable collection of disjoint measurable sets. Then

$$m\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} m(E_k).$$

Proof. Since $\bigcup_{k=1}^{\infty} E_k$ is measurable and by the sub-additivity of outer measure,

$$m\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} m(E_k).$$

We only need to show the opposite inequality. By monotonicity of measure, for any finite subcollection of $\{E_k\}$ we have,

$$\sum_{k=1}^{n} m(E_k) \leq m\left(\bigcup_{k=1}^{\infty} E_k\right)$$

for each n. Now, since the right hand side of the inequality is independent of n it follows that

$$\sum_{k=1}^{\infty} m(E_k) \leq m\left(\bigcup_{k=1}^{\infty} E_k\right).$$

This along with the first equation of this proof yields the desired result. □

Now, if we summarize the properties of the measure, then we see that

1. $0 \leq m(A) \leq \infty$ for any $A \subset \mathbb{R}$;
2. $m(\emptyset) = 0 = m(C)$, for any countable set C;
3. For $A \subset B$, $m(A) \leq m(B)$;
4. Any interval I is measurable and $m(I) = l(I)$;

5. m is translation invariant;

6. m is countably additive.

Hence, we can consider m as a function $m : \mathscr{M} \to \mathbb{R}^*$ which is a direct generalisation of the length of intervals that we were seeking throughout the unit. We will pose a significant question: "Is $\mathscr{M} = \mathscr{P}$?" We will discuss it in the next unit.

Let us give few other definitions and properties of measurable sets.

Definition 1.2.20. Let $\{A_n\}$ be a sequence of sets. Using union and intersection, we define the limit superior and limit inferior as follows:

$$\liminf_{n\to\infty} A_n = \bigcup_{n\geq 1} \bigcap_{j\geq n} A_j$$

and

$$\limsup_{n\to\infty} A_n = \bigcap_{n\geq 1} \bigcup_{j\geq n} A_j.$$

It can be easily seen from the definition that $\liminf A_n \subseteq \limsup A_n$. If they are equal, then the resulting set is denoted by $\lim A_n$. It is clear from the definition that $\limsup A_n$ is the set of points belonging to infinitely many of the sets A_n. It is also immediate that if $A_1 \subseteq A_2 \subseteq \ldots$, then $\lim A_n = \bigcup A_n$ and if $A_1 \supseteq A_2 \supseteq \ldots$, then $\lim A_n = \bigcap A_n$. Now, we have the following theorem:

Theorem 1.2.21. Let $\{A_i\}$ be a sequence of measurable sets. Then

1. if $A_1 \subseteq A_2 \subseteq \ldots$, we have
$$m(\lim A_i) = \lim m(A_i).$$

2. if $A_1 \supseteq A_2 \supseteq \ldots$, and $m(A_i) < \infty$ for at least one i, then
$$m(\lim A_i) = \lim m(A_i).$$

Proof. 1. We write

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= A_2 \setminus A_1 \\ B_3 &= A_3 \setminus A_2 \\ &\vdots \\ B_i &= A_i \setminus A_{i-1}. \end{aligned}$$

Then clearly,

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i.$$

and the sets B_i are measurable and disjoint. Also note that

$$\bigcup_{i=1}^{n} B_i = A_n.$$

So,

$$\begin{aligned} m(\lim A_n) &= m\left(\bigcup_{i=1}^{\infty} A_i\right) \\ &= \sum_{i=1}^{\infty} m(B_i) \\ &= \lim \sum_{i=1}^{n} m(B_i) \\ &= \lim m\left(\bigcup_{i=1}^{n} B_i\right) \\ &= \lim m(A_n) \end{aligned}$$

which was required to be proved.

2. Let p be the least integer such that $m(A_p) < \infty$. Then $m(A_i) < \infty$ for all $i \geq p$. We set, $B_i = A_i \backslash A_{i+1}$ and $A = \bigcap_{i=1}^{\infty} A_i$. Then B_i are pairwise disjoint sets and

$$A_p \setminus A = \bigcup_{i=p}^{\infty} B_i.$$

Therefore,

$$m(A_p \setminus A) = \sum_{i=p}^{\infty} m(B_i) = \sum_{i=p}^{\infty} m(A_i A_{i+1}).$$

But, $m(A_p) = m(A) + m(A_p \setminus A)$ and $m(A_i) = m(A_{i+1}) + m(A_i - A_{i+1})$, for all $i \geq p$, since $A \subset A_p$ and $A_{i+1} \subset A_i$. Further, using the fact that $m(A_i) < \infty$, for all $i \geq p$, it follows that

$$\begin{aligned} m(A_p \setminus A) &= m(A_p) - m(A) \\ m(A_i \setminus A_{i+1}) &= m(A_i) - m(A_{i+1}), \ \ \forall i \geq p. \end{aligned}$$

Hence,

$$\begin{aligned} m(A_p) - m(A) &= \sum_{i=p}^{\infty} (m(A_i) - m(A_{i+1})) \\ &= \lim_{n\to\infty} \sum_{i=p}^{n} (m(A_i) - m(A_{i+1})) \\ &= \lim_{n\to\infty} \{m(A_p) - m(A_n)\} \\ &= m(A_p) - \lim_{n\to\infty} m(A_n). \end{aligned}$$

Since $m(A_p) < \infty$, it gives,

$$m(A) = \lim_{n\to\infty} m(A_n).$$

This proves the result.

□

Exercise 1.2.22.

1. For $k > 0$ and $A \subseteq \mathbb{R}$, let $kA = \{x : k^{-1}x \in A\}$. Show that
   i. $m^*(kA) = km^*(A)$,
   ii. A is measurable iff kA is measurable.
2. For $A \subseteq \mathbb{R}$, let $-A = \{x : -x \in A\}$. Show that
   i. $m^*(A) = m^*(-A)$,
   ii. A is measurable iff $-A$ is measurable.
3. Let A and B be two measurable sets. Prove that

$$m(A \cup B) + m(A \cap B) = m(A) + m(B).$$

Sample Questions

1. Show that set set of all measurable sets form an algebra.
2. Show that set set of all measurable sets form a σ-algebra.
3. Show that m is translation invariant.
4. Show that any open set is measurable. Hence show that all closed sets are also measurable.
5. Show that the condition of $m(A_i)$ can not be dropped in theorem 1.2.21.

Unit 2

Course Structure

- Characterizations of measurable sets by open sets, closed sets
- Characterizations of measurable sets by G_δ and F_σ sets
- Measurability of Borel sets
- Existence of non-measurable sets.

2.1 Introduction

We are most acquainted with the intervals so far and have seen sets being approximated by open intervals in the definition of outer measure. We have also seen that intervals and hence, open sets, closed sets, are all measurable (exercise). Open sets and closed sets are the sets we come across quite regularly. So, if we can understand measurability in terms of open and closed sets, we can get more insight into the structure of a measurable set. In this unit, that is what we are going to do. Approximating a measurable set by open set, closed set, G_δ set, F_σ sets and compact sets. Another important aspect of the study of measurable sets is the existence of non-measurable sets. The question that we come across while studying measurable sets which has also been pointed out in the previous unit is whether all sets are measurable? Most of the sets that we are acquainted with are measurable. So we might think that all sets are measurable. But that is not so. And in this unit, we will turn our attention towards those non-measurable sets.

Objectives

After reading this unit, you will be able to:

- learn the approximation of measurable set by open sets
- learn the approximation of measurable set by closed sets
- learn the approximation of measurable set by G_δ sets
- learn the approximation of measurable set by F_σ sets

- learn what are Borel sets
- learn about the measurability of the Borel sets
- learn about the existence of non-measurable sets

2.2 Characterization by open and closed sets

We begin this section by the following theorem:

Theorem 2.2.1. For every $\epsilon > 0$, there exists an open set $U \subseteq \mathbb{R}$ such that $m(U) \leq \epsilon$ and U contains the set $\mathbb{Q}$ or rational numbers.

Proof. Let $\epsilon > 0$ and let $q_1, q_2, \ldots$ be an enumeration of the rational numbers. Construct U as

$$U = \bigcup_{n \in \mathbb{N}} \left(q_n - \frac{\epsilon}{2^{n+1}}, q_n + \frac{\epsilon}{2^{n+1}} \right)$$

Clearly, U is open and contains $\mathbb{Q}$, and

$$m(U) \leq \sum_{n \in \mathbb{N}} m\left(q_n - \frac{\epsilon}{2^{n+1}}, q_n + \frac{\epsilon}{2^{n+1}} \right) = \sum_{n \in \mathbb{N}} \frac{\epsilon}{2^n} = \epsilon$$

Hence the result. □

We begin by describing Lebesgue outer measure in terms of open sets as follows:

Theorem 2.2.2. If $S \subseteq \mathbb{R}$, then

$$m^*(S) = \inf\{m(U) | U \text{ is open and } S \subseteq U\}$$

Proof. Let x be the value of the infimum. Clearly, $m^*(S) \leq m(U)$ for every open set U containing S, and hence, $m^*(S) \leq x$. For the opposite inequality, let $\epsilon > 0$, and let $\mathcal{C}$ be a cover of S by open intervals so that

$$\sum_{I \in \mathcal{C}} l(I) \leq m^*(S) + \epsilon$$

Then $U = \bigcup \mathcal{C}$ is an open set that contains S, so

$$x \leq m(U) \leq \sum_{I \in \mathcal{C}} m(I) = \sum_{I \in \mathcal{C}} l(I) \leq m^*(S) + \epsilon$$

Since $\epsilon > 0$ is arbitrary, so $x \leq m^*(S)$. Combining the two inequalities, we get the required result. □

We will now use open sets to give a nice characterization of measurability. It says that every measurable set can be approximated by an open set from the exterior up-to any extent according to our wish. The following theorem can also be called the exterior approximation by open sets.

Theorem 2.2.3. A set $S \subseteq \mathbb{R}$ is Lebesgue measurable if and only if for every $\epsilon > 0$, there exists an open set U containing S such that $m^*(U \setminus S) < \epsilon$.

Proof. Let S be measurable, and $\epsilon > 0$. Let $m(S) < \infty$. From the definition of Lebesgue outer measure, we have a sequence $\{I_n\}$ of open intervals such that $S \subset \bigcup_{n=1}^{\infty} I_n$ and

$$\sum_{n=1}^{\infty} l(I_n) < m^*(S) + \epsilon.$$

Set $U = \bigcup_{n=1}^{\infty} I_n$. Then U is an open set containing S and

$$m^*(U) \leq \sum_{n=1}^{\infty} l(I_n) < m^*(S) + \epsilon$$

which implies that

$$m^*(U \setminus S) = m^*(U) - m^*(S) < \epsilon, \text{ since } m(S) < \infty.$$

Further, if $m(S) = \infty$, let $S_k = S \cap [-k, k]$. Then each S_k is measurable and $m(S_k) < \infty$ for all k. By the preceding argument, for each k we can find an open set U_k containing S_k such that $m^*(U_k \setminus S_k) < \frac{\epsilon}{2^k}$. Since $S = \bigcup_{k=1}^{\infty} S_k$ and $S \subset \bigcup_{k=1}^{\infty} U_k = U$, it follows that

$$\begin{aligned} m^*(U \setminus S) &= m(U \setminus S) \\ &\leq m\left(\bigcup_{k=1}^{\infty}(U_k \setminus S_k)\right) \\ &\leq \sum_{k=1}^{\infty} m(U_k \setminus S_k) < \epsilon. \end{aligned}$$

Conversely, let $S \subseteq \mathbb{R}$, and suppose that for every n, there exists an open set U_n containing S such that $m^*(U_n \setminus S) < \frac{1}{n}$. Let $E = \bigcap_{n\in\mathbb{N}} U_n$, and note that E is a measurable set containing S. But $E \setminus S \subseteq U_n \setminus S$ for each n. So,

$$m^*(E \setminus S) \leq m^*(U_n \setminus S) < \frac{1}{n}$$

for each n. We conclude that $m^*(E \setminus S) = 0$, and hence $E \setminus S$ is measurable. Then $S = E \setminus (E \setminus S)$ is measurable as well. □

We know that any arbitrary union of open sets is open. And finite intersection of open sets is open. But can we replace finite intersection by arbitrary intersection? Consider the sets $(-1/n, 1/n)$, $n \in \mathbb{N}$. Then,

$$\bigcap_{n\in\mathbb{N}} \left(-\frac{1}{n}, \frac{1}{n}\right) = \{0\}$$

which is not open. Hence, we can't replace finite intersection by arbitrary intersection always. So if we consider countable intersection of open sets, then it is called a G_δ set. The set in the above example is a G_δ set. And those are definitely measurable. We can characterize measurability by G_δ sets.

Theorem 2.2.4. A set S is measurable if and only if there exists a G_δ set G containing S such that $m^*(G \setminus S) = 0$.

Proof. Let S be measurable. Then, for each $n \in \mathbb{N}$, there exists an open set O_n containing S such that $m^*(O_n \setminus S) < 1/n$. Take $G = \bigcap_{n\in\mathbb{N}} O_n$. Then G is a G_δ set and $S \subseteq G$. Thus

$$m^*(G \setminus S) \leq m^*(O_n \setminus S) < \frac{1}{n}$$

for each n. Hence, $m^*(G \setminus S) = 0$.

Conversely, let there exist a G_δ set such that $m^*(G \setminus S) = 0$. Then $G \setminus S$ is measurable. Also, G is measurable since it is the countable intersection of measurable sets. Thus, $S = G\setminus(G\setminus S)$ is also measurable. Hence the proof. □

Note that the complement of any open set is a closed set. So the characterization of measurability by closed set is closely linked with that by open sets with a slight variation as we shall see. The theorem below is known as the inner approximation of measurable sets by closed sets.

Theorem 2.2.5. A set $S \subseteq \mathbb{R}$ is measurable if and only if for every $\epsilon > 0$, there exists a closed set F contained in S such that $m^*(S \setminus F) < \epsilon$.

Proof. Let S be measurable. Then $\mathbb{R} \setminus S$ is also measurable. Thus for each $\epsilon > 0$, there exists an open set containing $\mathbb{R} \setminus S$ such that $m^*(U \setminus (\mathbb{R} \setminus S)) < \epsilon$. Take $F = \mathbb{R} \setminus U$. Then F is closed. Also, F is contained in S. Since $m^*(U \setminus (\mathbb{R} \setminus S)) < \epsilon$, so $m^*(S \setminus F) < \epsilon$.

Converse can also be similarly shown. □

Similar to the G_δ sets, we have something in case of closed sets. We know that, arbitrary intersection of closed sets are closed and finite union of closed sets are closed. In a similar manner, the arbitrary union of closed sets may not be closed (give counter example!). Countable union of closed sets are called F_σ sets. These sets are not closed in general. However, they are also measurable. With the help of the above theorem, we can characterize measurable sets with the help of F_σ sets as follows:

Theorem 2.2.6. A set S is measurable if and only if there exists an F_σ set F contained in S such that $m^*(S \setminus F) = 0$.

This theorem can be proved directly and also using the theorem for G_δ sets.

Proof. Left as an exercise. □

From all the above theorems, we conclude the following corollary which gives a beautiful and important structure of a measurable set:

Corollary 2.2.7. Let $S \subseteq \mathbb{R}$. Then S is Lebesgue measurable if and only if for every $\epsilon > 0$ there exists a closed set F and an open set U such that $F \subseteq S \subseteq U$ and $m(U \setminus F) < \epsilon$.(see figure 2.2.1)

Proof. Left as exercise. □

We can also define measure with respect to closed sets.

Definition 2.2.8. If $S \subseteq \mathbb{R}$, the Lebesgue inner measure of S is defined as

$$m_*(S) = \sup\{m(F) | F \text{ is closed and } F \subseteq S\}$$

From the above corollary, it is clear that $m_*(E) = m(E)$ for any measurable set E. It is also apparent that $m_*(S) \leq m^*(S)$ for any set $S \subseteq \mathbb{R}$. The following theorem gives a nice characterization of measurability for sets of finite measure.

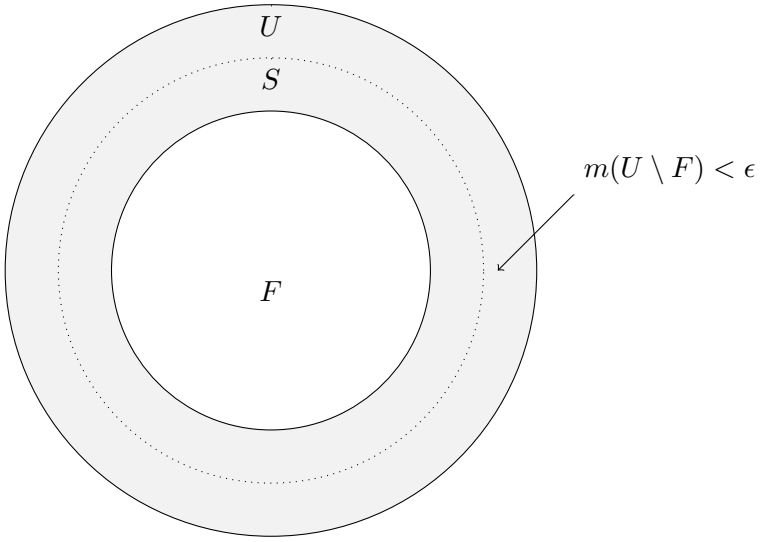


Figure 2.2.1: Structure of measurable set S

Theorem 2.2.9. Let $S \subseteq \mathbb{R}$, and suppose that $m^*(S) < \infty$. Then S is measurable if and only if

$$m_*(S) = m^*(S)$$

Proof. If S is measurable, then $m_*(S) = m(S) = m^*(S)$. Conversely, suppose that $m^*(S) < \infty$ and $m_*(S) = m^*(S)$. Let $\epsilon > 0$, and let $F \subseteq S$ be a closed set and $U \subseteq \mathbb{R}$ and open set containing S so that

$$m_*(S) \leq m(F) + \frac{\epsilon}{2} \text{ and } m(U) \leq m^*(S) + \frac{\epsilon}{2}$$

Then

$$m(U \setminus F) = m(U) - m(F) \leq \left(m^*(S) + \frac{\epsilon}{2}\right) - \left(m_*(S) - \frac{\epsilon}{2}\right) = \epsilon$$

Since ϵ is arbitrary, it follows from the previous corollary, that S is measurable. □

Theorem 2.2.10. Every compact subset of $\mathbb{R}$ is measurable.

Proof. By Heine-Borel theorem, every compact set on $\mathbb{R}$ is closed and bounded. So every compact set is measurable. □

Exercise 2.2.11. 1. Let E be Lebesgue measurable with $m(E) < \infty$. Show that

   (a) there exists a decreasing sequence of open sets U_k such that $\lim_{k\to\infty} m(U_k) = m(E)$.

   (b) there exists an increasing sequence of closed sets F_k such that $\lim_{k\to\infty} m(F_k) = m(E)$.

2. Show that for any set A and $\epsilon > 0$, there is an open set O containing A such that $m^*(O) \leq m^*(A) + \epsilon$.

3. Show that for any set A, there exists a measurable set E containing A such that $m^*(A) = m(E)$.

4. Let E be a set with $m^*(E) < \infty$. Show that E is measurable if and only if for any $\epsilon > 0$ there is a finite union B of open intervals such that

$$m^*(E \Delta B) < \epsilon.$$

5. Show that if $m^*(E) = \infty$ and for each $\epsilon > 0$, there are intervals $I_1, I_2, \ldots, I_n$ such that

$$m^* \left(E \Delta \left(\bigcup_{i=1}^{n} I_i \right) \right) < \epsilon,$$

then at least one of the intervals I_i is finite.

2.3 Borel Sets

Recall that a σ-algebra on X is any non-empty collection of subsets of X that is closed under taking complements and countable unions. For example, the Lebesgue measurable subsets of $\mathbb{R}$ forms a σ-algebra on $\mathbb{R}$. We also recall the following result.

Theorem 2.3.1. Let $\mathcal{C}$ be any collection of σ-algebras on $\mathbb{R}$. Then the intersection $\bigcap \mathcal{C}$ is also a σ-algebra on $\mathbb{R}$.

This is the smallest σ-algebra containing $\mathcal{C}$ and is known as the σ-algebra generated by $\mathcal{C}$. It is important for this definition that there is always at least one σ-algebra containing $\mathcal{C}$, namely the power set $\mathscr{P}(\mathbb{R})$ of all subsets of $\mathbb{R}$.

Definition 2.3.2. The Borel algebra $\mathcal{B}$ is the σ-algebra on $\mathbb{R}$ generated by the collection of all open sets. A set $B \subseteq \mathbb{R}$ is called a Borel set if $B \in \mathcal{B}$.

By definition, every open set is a Borel set. Moreover, since the Borel sets are a σ-algebra, the complement of any Borel set is a Borel set, and any countable union of Borel sets is a Borel set. The Borel σ-algebra can also be described as the σ-algebra generated by these family of subsets of $\mathbb{R}$.

1. Open intervals;
2. Open sets;
3. Closed intervals;
4. Closed sets;
5. Compact sets;
6. Left open, right closed intervals;
7. Right open, left closed intervals;
8. All intervals.

Theorem 2.3.3. Every Borel set is measurable.

Proof. Observe that the collection $\mathcal{M}$ of all Lebesgue measurable sets is a σ-algebra that contains the open sets. Since $\mathcal{B}$ is the intersection of all such σ-algebras, it follows that $\mathcal{B} \subseteq \mathcal{M}$. □

Theorem 2.3.4. Every open set, closed set, F_σ set, or G_δ set is a Borel set.

Proof. Every open set lies in $\mathcal{B}$ by definition. Since $\mathcal{B}$ is a σ-algebra, it follows immediately that closed sets, F_σ sets, G_δ sets are all Borel sets as well. □

Theorem 2.3.5. The Borel algebra is generated by the collection of all open intervals.

Proof. Let $\mathcal{A}$ be a σ-algebra generated by the open intervals. Since $\mathcal{B}$ contains the open intervals, we know that $\mathcal{A} \subseteq \mathcal{B}$. But, since every open set is the countable union of disjoint open intervals, $\mathcal{A}$ contains every open set, and hence $\mathcal{B} \subseteq \mathcal{A}$. □

We see that open sets, closed sets, F_σ sets and G_δ sets are among the simplest of the Borel sets. The next result shows the relation between Lebesgue measurable sets and Borel sets.

Theorem 2.3.6. Every Lebesgue measurable set is the union of a Borel set and a set of Lebesgue measure zero.

Proof. Let E be a Lebesgue measurable set. Then we have an F_σ (Borel set) set B such that $B \subset E$ and $m(E \setminus B) = 0$. But $E = B \cup (E \setminus B)$. Hence B is our desired Borel set and $E \setminus B$ is the set with zero Lebesgue measure. □

Exercise 2.3.7.

1. Show that not all measurable sets are Borel sets.
2. Show that $E \subset \mathbb{R}$ is measurable if and only if there are Borel sets B_1, B_2 satisfying $B_2 \subset E \subset B_1$ and $m^*(B_1 \setminus B_2) = 0$.

2.4 Non-measurable Sets

There are sets which are non-measurable. For this, we first state *Axiom of Choice*.

Definition 2.4.1. Let C be a collection of non-empty sets. Then we can choose a member from each set in that collection. In other words, there exists a function f defined on C with the property that, for each set S in the collection, $f(S)$ is a member of S.

The function f is called the *choice function*. Let us consider a few examples.

1. If C is the collection of subsets of $\{1, 2, \ldots\}$, then we can define f a follows: $f(S)$ is the least element of S.
2. If C is the collection of all intervals of real numbers with positive, finite lengths, then we can define $f(S)$ to be the midpoint of the interval S.

Axiom of Choice cannot be derived from the rest of set theory but must be introduced as an additional axiom. We will now construct a non-measurable set in the following way.

Theorem 2.4.2. There exists a non-measurable set.

Proof. Let $x, y \in [0, 1]$. Define a relation $'\sim'$ as

$$x \sim y \quad iff \quad y - x \in Q_1 = \mathbb{Q} \cap [-1, 1]$$

Verify that '$\sim$' is an equivalence relation on $[0, 1]$. Then, $[0, 1]$ gets partitioned into disjoint equivalence classes, say E_α, where x and y can be in E_α if and only if $x \sim y$ holds. Since Q_1 is countable, each E_α is countable. Since $[0, 1]$ is uncountable, there are uncountably many sets E_α. Using Axiom of Choice, we consider a set V in $[0, 1]$, containing just one element x_α from each set E_α. Such set is called Vitali set. Let $\{r_i\}$ be an enumeration of Q_1, and for each n, write $V_n = V + r_n$. We claim that V_n are disjoint. If $y \in V_n \cap V_m$, there exist x_α and $x_\beta \in V$ such that

$$y = x_\alpha + r_n \quad y = x_\beta + r_m$$

But then $x_\beta - x_\alpha \in Q_1$, so $x_\beta = x_\alpha$ by definition of V and we have $n = m$. So, $V_n \cap V_m = \emptyset$ for $n \neq m$. Also $[0,1] \subseteq \bigcup_{n=1}^{\infty} V_n \subseteq [-1,2]$, since for all $x \in [0,1]$, $x \in E_\alpha$ for some α. Then

$$x = x_\alpha + r_n$$

giving $x \in V_n$. The second inclusion is obvious.

If V is measurable, then V_n are also measurable and by translation invariance of measure, $m(V) = m(V_n)$. Using the measurability of the sets V_n, we have

$$1 = m([0,1]) \leq \sum_{1}^{\infty} m(V_n) = m(V) + m(V) + \cdots \leq 3$$

But the sum can only be 0 or ∞. So, our assumption is false and V can't be measurable. □

It is generally not necessary for a measurable set to contain interval. For example, if we consider the set of rational numbers, the we see that it does not contain any interval and its measure is zero. But, if the measure of a set is positive, then we see a rather interesting feature of such sets. Let us state the following theorem:

Theorem 2.4.3. If T is a measurable set with positive measure. Define

$$T^* = \{x - y : x \in T \;\; y \in T\}$$

Then T^* contains an interval $(-\alpha, \alpha)$ for some $\alpha > 0$.

Proof. Since T is measurable, there exists a closed set C contained i T, of positive measure. Since $m(C) = \lim m(C \cap [-n,n])$, we may assume that C is a bounded set. Also, since T is measurable, there exists an open set $U \supset C$ such that $m(U \setminus C) < m(C)$. Define the distance between two sets A and B as

$$d(A,B) = \inf\{|x-y| : x \in A,\ y \in B\}$$

Since $|x-y|$ is a continuous function of x and y, the distance between A and B is positive if A and B are disjoint closed sets one of which is bounded. Let α be the distance between the closed sets C and $\mathbb{R} \setminus U$, so that $\alpha > 0$. Let x be any point of $(-\alpha, \alpha)$. We wish to show that $C \cap (C - x) \neq \emptyset$> For, otherwise, since $C - x = \{y : y + x \in C\}$, we have that $\forall x \in (-\alpha, \alpha), \exists z \in C$ such that

$$z' = z + x \in C$$

and so that

$$x = z' - z \in T^*$$

Since $|x| < \alpha$, we have $C - x \subset U$ from the definition of α. So,

$$\begin{aligned} m(C \setminus (C-x)) &\leq m(U \setminus (C-x)) \\ &= m(U) - m(C-x) \\ &= m(U) - m(C) \\ &< m(U) \end{aligned}$$

Hence $m(C \cap (C-x)) > 0$ and so we must have $C \cap (C-x) \neq \emptyset$, as required. □

Exercise 2.4.4. Show that if A is any set with $m^*(A) > 0$, then there is a non-measurable set E contained in A.

Sample Questions

1. Show that a set S is measurable if and only if for any $\epsilon > 0$, there exists an open set U such that $S \subset U$ and $m(U \setminus S) < \epsilon$.
2. Show that a set S is measurable if and only if for any $\epsilon > 0$, there exists a closed set F such that $F \subset S$ and $m(S \setminus F) < \epsilon$.
3. Show that Borel sets are measurable.
4. Show that there exists a non-measurable set.

Unit 3

Course Structure

- Measurable functions : Definition on a measurable set in $\mathbb{R}$ and basic properties,
- Sequences of measurable functions,
- Simple functions, Measurable functions as the limits of sequences of simple functions,

---

3.1 Introduction

Measurable functions in measure theory are analogous to continuous functions in topology. A continuous function pulls back open sets to open sets, while a measurable function pulls back measurable sets to measurable sets. Lebesgue measurable functions play an important role in the Lebesgue theory of integration. It plays the same role in the theory as those bounded functions play in the theory of Riemann integration, which are continuous almost everywhere. We begin this unit defining measurable functions and discuss a few examples of measurable and non-measurable sets.

Objectives

After reading this unit, you will be able to

- differentiate between measurable and non-measurable functions and some basic examples
- learn about the sequence of measurable functions and their properties
- learn to approximate measurable functions with sequence of simple functions

3.2 Measurable Functions

Definition 3.2.1. An extended real valued function f defined on a Lebesgue measurable set E is said to be Lebesgue measurable function if for each $a\mathbb{R}$, the set $f^{-1}((a,\infty]) = \{x : f(x) > a\}$ is measurable.

In practice, the domain of definition of f will usually be either $\mathbb{R}$ or $\mathbb{R} \setminus F$, where $m(F) = 0$. The motivation behind this definition is in the fact that whether we are able to “measure" the inverse of intervals of type $(a,\infty]$. What about the other kind of intervals.

Theorem 3.2.2. The following statements are equivalent:

(a.) f is a measurable function,

(b.) $\forall a \in \mathbb{R}$, the set $f^{-1}([a, \infty]) = \{x : f(x) \geq a\}$,

(c.) $\forall a \in \mathbb{R}$, the set $f^{-1}([-\infty, a)) = \{x : f(x) < a\}$,

(d.) $\forall a \in \mathbb{R}$, the set $f^{-1}([-\infty, a]) = \{x : f(x) \leq a\}$.

Proof. Let f be measurable. Then for every $a \in \mathbb{R}$, the set $\{x : f(x) > a\}$ is measurable. Now,

$$\{x : f(x) \geq a\} = \bigcap_{n=1}^{\infty} \left\{x : f(x) > a - \frac{1}{n}\right\}.$$

Since the set $\left\{x : f(x) > a - \frac{1}{n}\right\}$ is measurable for each n, the set $\{x : f(x) \geq a\}$ is measurable for each $a \in \mathbb{R}$. Thus, $(a) \Rightarrow (b)$.

Now, let us assume (b) is true. So for each $a \in \mathbb{R}$, the set $\{x : f(x) \geq a\}$ is measurable. Thus, the complement of the set is also measurable for each $a \in \mathbb{R}$. Since

$$\{x : f(x) < a\} = \mathbb{R} \setminus \{x : f(x) \geq a\}$$

so, (c) is true. Hence, $(b) \Rightarrow (c)$.

Let us assume that (c) is true. We show that (d) is true. Since (c) is true, so for each $n \in \mathbb{N}$, the set $\left\{x : f(x) < a + \frac{1}{n}\right\}$ is measurable. Now,

$$\{x : f(x) \leq a\} = \bigcap_{n=1}^{\infty} \left\{x : f(x) < a + \frac{1}{n}\right\}.$$

Since the countable intersection of measurable sets is measurable, so the set $\{x : f(x) \leq a\}$ is measurable. Thus, $(c) \Rightarrow (d)$.

Lastly, we show that $(d) \Rightarrow (a)$. Let us assume that $(d.)$ is true. We know that for any $a \in \mathbb{R}$,

$$\{x : f(x) > a\} = \mathbb{R} \setminus \{x : f(x) \leq a\}$$

Since the complement of measurable set is measurable, hence, $\{x : f(x) > a\}$ is measurable. So, f is measurable. □

We also have an equivalent theorem as

Theorem 3.2.3. If an extended real function f is measurable then for every extended real number a, the set $\{x : f(x) = a\}$ is measurable.

Proof. Let f be measurable. Then, for every finite a,

$$\{x : f(x) = a\} = \{x : f(x) \geq a\} \cap \{x : f(x) \leq a\}.$$

So, $\{x : f(x) = a\}$ is measurable. For $a = \infty$,

$$\{x : f(x) = \infty\} = \bigcap_{n=1}^{\infty} \{x : f(x) > n\}$$

Since for each $n \in \mathbb{N}$, the set $\{x : f(x) > n\}$ is measurable, so $\{x : f(x) = \infty\}$ is measurable. Similarly, for $a = -\infty$, we see that

$$\{x : f(x) = -\infty\} = \bigcap_{n=1}^{\infty} \{x : f(x) < -n\}$$

Since each $\{x : f(x) < -n\}$ is measurable, so the set $\{x : f(x) = -\infty\}$ is measurable. This completes the proof. □

Exercise 3.2.4. 1. Suppose f is a measurable function defined on a measurable set E. Show that

(a) $f^{-1}([a, b))$ is measurable;
(b) $f^{-1}([a, b])$ is measurable;
(c) $f^{-1}\{\infty\}$ and $f^{-1}\{-\infty\}$ are measurable;
(d) $f^{-1}\{c\}$ is measurable;
(e) $f^{-1}(G)$ is measurable for any open set G in $\mathbb{R}$.

2. Check whether the following functions are measurable in their respective domains.

(a) $f(x) = \dfrac{1}{x}$, on $(0, 1)$.
(b) $f : \mathbb{R} \to \mathbb{R}$ defined as

$$\begin{aligned} f(x) &= x^2, \quad x < 1 \\ &= 2, \quad x = 1 \\ &= 2 - x, \quad x > 1. \end{aligned}$$

Let us see an important example.

Example 3.2.5. Let us define a function f on the closed interval $[0, 1]$ as

$$\begin{aligned} f(x) &= 1; \text{ if } x \in [0, 1] \cap \mathbb{Q} \\ &= 0; \text{ if } x \in [0, 1] \setminus \mathbb{Q}. \end{aligned}$$

Consider $a \in \mathbb{R}$. Then we have

$$\begin{aligned} \{x : f(x) > a\} &= \emptyset, \text{ if } a \geq 1 \\ \{x : f(x) > a\} &= [0, 1] \cap \mathbb{Q}, \text{ if } 0 \leq a < 1 \\ \{x : f(x) > a\} &= [0, 1], \text{ if } a < 0 \end{aligned}$$

In each case, the resulting sets are measurable.

The above function is called a characteristic function or indicator function of rational numbers in $[0, 1]$. We formally define indicator function of any subset E of $\mathbb{R}$ as follows:

Definition 3.2.6. Let E be a subset of $\mathbb{R}$. Then the function $f : \mathbb{R} \to \mathbb{R}$ is called the indicator function of E if

$$\begin{aligned} f(x) &= 1; \text{ if } x \in E \\ &= 0; \text{ if } x \notin E \end{aligned}$$

The indicator function of a particular set E is denoted by χ_E.

We have the following theorem in connection with the indicator functions:

Theorem 3.2.7. Indicator function of a set E is measurable if and only if E is measurable.

Proof. Let $a \in \mathbb{R}$. Then we see that

$$\begin{aligned} \{x : f(x) > a\} &= \emptyset, \text{ if } a \geq 1 \\ \{x : f(x) > a\} &= E, \text{ if } 0 \leq a < 1 \\ \{x : f(x) > a\} &= \mathbb{R}, \text{ if } a < 0 \end{aligned}$$

Each of the sets $\emptyset$ and $\mathbb{R}$ is measurable. If the set E is measurable, then the function f is measurable. Conversely, if the function f is measurable, then the set E is measurable. This completes the proof. □

Theorem 3.2.8. Any constant function is measurable.

Proof. Let $f(x) = c$ be a constant function on $\mathbb{R}$. Also, let $a \in \mathbb{R}$ be arbitrary. Then

$$\begin{aligned} \{x : f(x) > a\} &= \emptyset, \text{ if } a \geq c \\ \{x : f(x) > a\} &= \mathbb{R}, \text{ if } a < c \end{aligned}$$

Since both the sets $\emptyset$ and $\mathbb{R}$ are measurable, so f is measurable. □

Theorem 3.2.9. Every continuous function is measurable.

Proof. Let f be a continuous function on $\mathbb{R}$. Then for each $a \in \mathbb{R}$, the set

$$\{x : f(x) > a\} = f^{-1}(a, \infty)$$

Since f is continuous and the set (a, ∞) is open, so $f^{-1}(a, \infty)$ is open and hence measurable. Hence the theorem. □

Theorem 3.2.10. Let f and g be two measurable functions. Then:

1. $f \pm c$ is measurable.
2. $f \pm g$ is measurable.
3. cf is measurable for each $c \in \mathbb{R}$.
4. f^2 is measurable.
5. fg is measurable.

Proof.

1. Since f is measurable. So, for each $a \in \mathbb{R}$, the set $\{x : f(x) > a\}$ is measurable. Now,

$$\{x : f(x) \pm c > a\} = \{x : f(x) > a \mp c\}$$

Since f is measurable, so the set $\{x : f(x) > a \mp c\}$ is measurable for every $a \in \mathbb{R}$. Hence the result.

2. Since f and g are measurable, so for each $a \in \mathbb{R}$, the sets $\{x : f(x) > a\}$ and $\{x : g(x) > a\}$ are measurable. To show that $f + g$ is measurable, let $A = \{x : f(x) + g(x) > a\}$. Now,

$$x \in A \text{ only if } f(X) > a - g(x)$$

that is, only if there exists a rational number r_i such that $f(x) > r_i > a - g(x)$, where $\{r_1, r_2, \ldots\}$ is the enumeration of rational numbers. But then $g(x) > a - r_i$ and so

$$x \in \{x : f(x) > r_i\} \cap \{x : g(x) > a - r_i\}$$

Hence,

$$A \subseteq B = \bigcup_{i=1}^{\infty} (\{x : f(x) > r_i\} \cap \{x : g(x) > a - r_i\})$$

which is a measurable set. Also, since A contains B, we have $A = B$. So, $f + g$ is measurable. Then $f - g = f + (-g)$ is measurable.

3. Since f is measurable, so the set $\{x : f(x) > a\}$ is measurable for every $a \in \mathbb{R}$. Now, for $c > 0$,

$$\{x : cf(x) > a\} = \{x : f(x) > c^{-1}a\}$$

which is a measurable set. Similarly we can show that $\{x : cf(x) > a\}$ is measurable for $c < 0$.

4. For $a \in \mathbb{R}$, the set

$$\begin{aligned} \{x : f^2(x) > a\} &= 0 \text{ if } a < 0 \\ &= \{x : f(x) > \sqrt{a}\} \cap \{x : f(x) < -\sqrt{a}\} \text{ if } a \geq 0 \end{aligned}$$

Since the sets in both the cases are measurable, hence f^2 is measurable.

5. Finally

$$fg = \frac{1}{4}((f+g)^2 - (f-g)^2)$$

Since each $f + g$, $f - g$ are measurable, so their square is also measurable. And hence the above set is measurable.

□

Note 3.2.11. The results hold for extended real numbers except when $f + g$ is not defined. For example when $f = \infty$ and $g = -\infty$ or vice versa. Similarly for the case $f - g$.

3.3 Sequence of Measurable Functions

In fact, the above results are true for any finite sequence of sets $\{f_i\}$. But, what happens when we replace finite sequence by infinite sequence? Actually, we have the following theorem for infinite sequence of functions:

Theorem 3.3.1. Let $\{f_n\}$ be a sequence of measurable functions defined on the same measurable set. Then

1. $\sup_{1 \leq i \leq n} f_i$ is measurable for each n.
2. $\inf_{1 \leq i \leq n} f_i$ is measurable for each n.
3. $\sup f_n$ is measurable.
4. $\inf f_n$ is measurable.
5. $\limsup f_n$ is measurable.

6. $\liminf f_n$ is measurable.

7. $\lim f_n$ (finite or infinite) is measurable on the same set, if it exists on every point of the set.

Proof. 1. Each f_i is measurable. Let $a \in \mathbb{R}$. Since,

$$\left\{x : \sup_{1\le i\le n} f_i(x) > a\right\} = \bigcup_{i=1}^{\infty}\{x : f_i(x) > a\},$$

Since each $\{x : f_i(x) > a\}$ is measurable, so $\left\{x : \sup_{1\le i\le n} f_i(x) > a\right\}$ is measurable.

2. We know that

$$\inf_{1\le i\le n} f_i = -\sup_{1\le i\le n}(-f_i)$$

Since f_i is measurable for each i, so $-f_i$ is measurable for each i. Thus, by the previous subpart, we arrive at the required conclusion.

3. Since each f_i is measurable, for each $a \in \mathbb{R}$, the set $\{x : f_i(x) > a\}$ is measurable. We also know that

$$\{x : \sup f_n(x) > a\} = \bigcup_{n=1}^{\infty}\{x : f_n(x) > a\}$$

Hence, we arrive at the desired result.

4. We know that

$$\inf f_n = -\sup(-f_n)$$

and so $\inf f_n$ is measurable.

5. Since

$$\limsup f_n = \inf\left(\sup_{i\ge n} f_i\right)$$

a measurable function by the previous subparts.

6. Since

$$\liminf f_n = -\limsup(-f_n)$$

and so the required result.

7. Let $\lim f_n$ exists for all points on the measurable set. Then

$$\limsup f_n = \lim f_n = \liminf f_n$$

and by previous subparts, $\limsup f_n$ and $\liminf f_n$ are measurable and hence, $\lim f_n$ is also so.

□

Definition 3.3.2. A property is said to hold almost everywhere (abbreviated as a.e), if it holds everywhere except on a set of measure zero.

Example 3.3.3. Suppose f and g are extended real valued functions defined on a measurable set E. If f is measurable and $f = g$ a.e on E, then g is also measurable on E. To show this, let us assume any real number c. We must show that the set $\{x \in E : \ g(x) > c\}$ is a measurable subset of E. Define $A = \{x \in E : \ f \neq g\}$. Then by assumption, A is measurable with measure zero. Then $g = f$ on the measurable set $E \setminus A$, and

$$\begin{aligned}\{x \in E : \ g(x) > c\} &= \{x \in E \setminus A : \ g(x) > c\} \cup \{x \in A : \ g(x) > c\} \\ &= \{x \in E \setminus A : \ f(x) > c\} \cup \{x \in A : \ g(x) > c\} \\ &= (\{x \in E : \ f(x) > c\} \cap (E \setminus A)) \cup \{x \in A : \ g(x) > c\}.\end{aligned}$$

The set $\{x \in A : \ g(x) > c\}$ is measurable since it is a subset of a set of measure zero. Also, since f is a measurable function on E, $\{x \in E : \ f(x) > c\}$ is a measurable subset of E and hence its intersection with $E \setminus A$. Hence $\{x \in E : \ g(x) > c\}$ is a measurable subset of E.

Example 3.3.4. Every Riemann integrable function defined on $[a, b]$ is a measurable function on $[a, b]$. One may recall that a bounded function f on $[a, b]$ is Riemann integrable if and only if the set D of its discontinuities has measure zero. Then f is continuous on $[a, b] \setminus D$, hence measurable on $[a, b] \setminus D$. Define g to be f on $[a, b] \setminus D$ and zero on D. Then

$$\{x \in [a, b] : \ g(x) > c\} = \{x \in [a, b] \setminus D : \ g(x) > c\} \cup \{x \in D : \ g(x) > c\}.$$

And,

$$\begin{aligned}\{x \in D : \ g(x) > c\} &= \emptyset, \ \text{ if } c \geq 0 \\ &= D, \ \text{ if } c < 0 \ \text{ since } \ g = 0 \ \text{ on } \ D.\end{aligned}$$

Thus g is a measurable function on $[a, b]$. Also, $f = g$ except on a set of measure zero. By previous example, f is measurable function. The converse of the statement is not true in general, that is, there are measurable functions that are not Riemann integrable (Find an example!).

Example 3.3.5. Suppose f is a function defined on some measurable set E and $\{f_n\}$ is a sequence of measurable functions defined on E such that $f = \lim f_n$ a.e on E. Then f is measurable on E. Indeed, let us assume that $A = \{x \in E : \ \lim f_n(x) \ \text{ is not defined of } \lim f_n(x) \neq f(x)\}$. Then the set A has measure zero. Define a new sequence of functions $\{g_n\}$ on E by

$$\begin{aligned}g_n(x) &= f_n(x), \ \ x \notin A \\ &= 0, \ \ x \in A,\end{aligned}$$

and let g be given by

$$\begin{aligned}g(x) &= f(x), \ \ x \notin A \\ &= 0, \ \ x \in A.\end{aligned}$$

Since each g_n equals a measurable function f_n a.e on E, g_n is measurable. If $x \in A$, $\lim g_n(x) = 0 = g(x)$. If $x \notin A$, $\lim g_n(x) = \lim f_n(x) = f(x) = g(x)$. Hence, $\lim g_n(x) = g(x)$ on E. Since the pointwise limit of a sequence of measurable functions is measurable, so g is measurable on E. Now, $f = g$ a.e on E. Hence, f is also measurable on E.

So far, we have only seen examples of measurable functions. It might seem that all functions are measurable. But that is not necessarily true.

Example 3.3.6. Let V be a non-measurable subset of $[0,1]$ and define

$$\begin{aligned} f(x) &= 1, \quad x \in V \\ &= -1, \quad x \in [0,1] \setminus V. \end{aligned}$$

Then f is non-measurable function on $[0,1]$. (Prove it!)

Exercise 3.3.7.

1. Show that any function defined on a set of measure zero is measurable.
2. If f and g are measurable functions on a common measurable set E, show that the sets $\{x \in E : f(x) \leq g(x)\}$ and $\{x \in E : f(x) = g(x)\}$ are measurable.
3. If f is a measurable function on a measurable set E and if A is any measurable subset of E, then show that f is a measurable function on A.
4. Let f and g be measurable functions on a measurable set E. Show that $\max\{f,g\}$ and $\min\{f,g\}$ are measurable.
5. If f is a measurable function defined on a measurable set E, then show that $f^+ = \max\{f,0\}$, $f^- = -\min\{f,0\}$, and $|f|$ are measurable.

Sample Questions

1. Show that a function f defined on a measurable set E is measurable if and only if for every real number a, the set $\{x \in E : f(x) \geq a\}$ is measurable.
2. Define measurable function. Show that every continuous function is measurable.
3. Show that the characteristic function on some set E is measurable if and only if E is measurable.
4. Show that the pointwise limit of a sequence of measurable functions defined on a measurable set E is measurable.

Unit 4

Course Structure

- Simple functions,
- Measurable functions as the limits of sequences of simple functions

---

4.1 Introduction

The most common example of measurable functions are the simple functions. So, we will learn to approximate any measurable function with sequence of simple functions.

Objectives

After reading this unit, you will be able to

-

4.2 Simple Functions

Simple functions are really the “simplest" functions in certain respects. The easiest functions to deal with are the constant functions. However, there are certain functions which are piecewise constant. Such functions are called the simple functions.

We are acquainted with the definition of characteristic function of any set A, denoted by χ_A. We will define simple functions using them.

Definition 4.2.1. Suppose

$$E = \bigcup_{k=1}^{n} E_n,$$

where the sets E_k are measurable, mutually disjoint subsets of $\mathbb{R}$ and $c_1, c_2, \ldots, c_n$ are real numbers. Then a function ϕ defined on E by

$$\phi(x) = \sum_{k=1}^{n} c_k \chi_{E_k}(x),$$

is called a simple function.

A simple function assumes a finite number of real values and assumes each of these on a measurable set. $\phi(x) = c_k$ on E_k, $1 \leq k \leq n$. A simple function can also be called a linear combination of characteristic functions. Characteristic functions are also simple functions with only two sets of values 0 and 1 on two disjoint sets. From the definition of simple functions, it can be evidently said that simple functions defined on the measurable set E is measurable. Indeed, f is measurable if and only if all the sets E_k are measurable.

Example 4.2.2. 1. Each step function is a simple function.

2. Each characteristic function on a measurable set is a simple function.

4.3 Simple Approximation Theorem

We will now show that any measurable function defined on a measurable set can be approximated by simple functions. This is known as the Simple Approximation theorem. The statement is as follows.

Theorem 4.3.1. Let f be a measurable function defined on a measurable set E. Then there exists a sequence of simple functions $\{\phi_k\}$ on E such that

$$\lim \phi_k = f \quad \text{(finite or infinite)}$$

for all $x \in E$. If f is bounded on E, then

$$\lim \phi_k = f \quad \text{(uniformly)}$$

on E. If f is non-negative, the sequence $\{\phi_k\}$ may be constructed so that it is a monotonically increasing sequence.

Proof. Suppose f is non-negative on E. We will construct a monotonically increasing sequence $\{\phi_k\}$ with $\lim \phi_k = f$. The idea is to divide the range of f and approximate by level curves. Since $f(E) \subset [0, \infty]$, we partition $[0, \infty]$ as follows:

Step 1. $[0, \infty] = [0, 1) \cup [1, \infty] = \left[0, \frac{1}{2}\right) \cup \left[\frac{1}{2}, 1\right) \cup [1, \infty]$. Define $E_{11} = f^{-1}\left(\left[0, \frac{1}{2}\right)\right)$, $E_{12} = f^{-1}\left(\left[\frac{1}{2}, 1\right)\right)$, $E_1 = f^{-1}([1, \infty])$, and

$$\phi_1(x) = 0.\chi_{E_{11}} + \frac{1}{2}.\chi_{E_{12}} + 1.\chi_{E_1}.$$

Clearly, $\phi_1 \leq f$ on E.

Step 2.

$$\begin{aligned} [0, \infty] &= [1, 0) \cup [1, 2) \cup [2, \infty] \\ &= \left[0, \frac{1}{4}\right) \cup \left[\frac{1}{4}, \frac{1}{2}\right) \cup \left[\frac{1}{2}, \frac{3}{4}\right) \cup \left[\frac{3}{4}, 1\right) \cup \left[1, \frac{5}{4}\right) \cup \left[\frac{5}{4}, \frac{6}{4}\right) \cup \left[\frac{6}{4}, \frac{7}{4}\right) \cup \left[\frac{7}{4}, \frac{8}{4}\right) \cup [2, \infty]. \end{aligned}$$

We have decomposed $[0, \infty]$ into $2^2 + 2^2 + 1$ subintervals at the second step. We define the inverse images:

$$E_{21} = f^{-1}\left(\left[0, \frac{1}{4}\right)\right), \quad E_{22} = f^{-1}\left(\left[\frac{1}{4}, \frac{1}{2}\right)\right), \ldots,$$
$$E_{28} = f^{-1}\left(\left[\frac{7}{4}, \frac{8}{4}\right)\right), \quad E_2 = f^{-1}([2, \infty]).$$

Define

$$\begin{aligned}\phi_2 &= 0.\chi_{E_{21}} + \frac{1}{4}.\chi_{E_{22}} + \ldots + \frac{7}{4}.\chi_{E_{28}} + 2\chi_{E_2} \\ &= \sum_{i=1}^{2.2^2} \frac{i-1}{2^2}.\chi_{E_{2i}} + 2\chi_{E_2}.\end{aligned}$$

Hence,

$$E_{1i} = E_{22i-1} \cup E_{22i} \text{ for } i = 1, 2.$$

.

.

.

Step k. $[0, \infty] = [0, 1) \cup [1, 2) \cup [2, 3) \cup \cdots \cup [k-1, k) \cup [k, \infty]$ and partition into $2^k + 2^k + \ldots + 2^k + 1$ subintervals $= k.2^k + 1$ disjoint subintervals and form inverse images. Thus,

$$\phi_k = \sum_{i=1}^{k.2^k} \frac{i-1}{2^k}.\chi_{E_{ki}} + k\chi_{E_k}.$$

Note that $E_{ki} = E_{k+1\ 2i-1} \cup E_{k+1\ 2i}$. To construct ϕ_{k+1}, divide the intervals $\left[\frac{i-1}{2^k}, \frac{i}{2^k}\right)$ in half, and then ϕ_k to ϕ_{k+1} at those x's where ϕ_k changes.

Certainly, ϕ_k are non-negative simple functions. We must show that $\phi_k \leq \phi_{k+1}$ and $\lim \phi_k = f$ on E.

If $f(x_0) = \infty$, then $\phi_k(x_0) = k$ for all k and $\lim \phi_k(x_0) = \infty$. If $f(x_0) < \infty$, then for $k > f(x_0)$, $0 \leq f(x_0) - \phi_k(x_0) < \frac{1}{2^k}$ and $\lim \phi_k(x_0) = f(x_0)$. All that is left is monotonicity. We know that, $E_{ki} = E_{k+1\ 2i-1} \cup E_{k+1\ 2i}$, if $x_0 \in E_{ki}$, for some i, then $\phi_k(x_0) = \frac{i-1}{2^k}$ and $\phi_{k+1}(x_0) = \frac{2i-2}{2^{k+1}} = \frac{i-1}{2^k}$ or $\frac{2i-1}{2^{k+1}}$, and $\phi_k(x_0) \leq \phi_{k+1}(x_0)$. If $x_0 \notin E_{ki}$, $i = 1, 2, \ldots, k.2^k$, then $x_0 \in E_k = f^{-1}([k, \infty]) = f^{-1}([k, k+1)) \cup f^{-1}([k+1, \infty])$. So, x_0 is either in $f^{-1}([k, k+1))$, in which case, $\phi_k(x_0) = k$ and $\phi_{k+1} = \frac{j}{2^{k+1}} > \frac{2k.2^k}{2^{k+1}} = k = \phi_k(x_0)$, or $x_0 \in f^{-1}([k+1, \infty])$, and then $\phi_{k+1}(x_0) = k+1 > k = \phi_k(x_0)$. Thus, we have shown that for $f \geq 0$,

$$0 \leq \phi_1 \leq \ldots \leq \phi_k \leq \phi_{k+1} \leq \ldots$$

and $\lim \phi_k = f$ on E.

If f is non-negative and bounded on E, say $0 \leq f \leq M$ on E, then for all $k > M$, $0 \leq f(x) - \phi_k(x) < \frac{1}{2^k}$ for all $x \in E$, that is, $\lim \phi_k = f$ uniformly on E.

In the general case (f may be negative), recall that $f^+ = \max\{f, 0\}$ and $f^- = -\min\{f, 0\}$. This implies that $f = f^+ - f^-$, where f^+ and f^- are non-negative measurable functions on E. Applying the above arguments on f^+ and f^- and since the difference of two simple functions is again a simple function, so we get the theorem. □

Exercise 4.3.2. 1. Prove or disprove: The sum and difference of two simple functions is a simple function.

2. If ϕ is a simple function on a measurable set E, then show that ϕ is measurable.

Sample Questions

1. State and prove the simple approximation theorem.

Unit 5

Course Structure

- Lusin's theorem on restricted continuity of measurable functions,
- Egoroff's theorem,
- Convergence in measure

5.1 Introduction

When studying about Lusin's theorem and Egoroff's theorem, it is necessary to know about the Littlewood's three principles. It gives intuitive knowledge about measure theory. The three celebrated Littlewood's Principles for Lebesgue Measure on $\mathbb{R}$ are, roughly speaking:

1. Every measurable set of finite measure is nearly a finite union of intervals.
2. Every measurable function is nearly continuous.
3. Every convergent sequence of functions is nearly convergent.

The first principle can be proved using the concepts of unit 2 and has been left as exercise. The second and third principles are known as Lusin's theorem and Egoroff's theorem respectively. The idea of convergence in measure is also introduced in this unit. We are well aware with pointwise as well as the uniform convergence of functions in analysis. A more general notion is the convergence almost everywhere (convergence except on a measure zero set). But the concept of a.e convergence is almost same as that of pointwise convergence. The convergence in measure is a more general concept, which was introduced by Riesz and Fischer in early twientieth century. We start with the Egoroff's theorem and later advance through the unit.

Objectives

After reading this unit, you will be able to

- state Lusin's and Egoroff's theorems and apply them appropriately;
- define the converegnce in measure and discuss its implications.

5.2 Lusin's and Egoroff's Theorems

We will first prove the Egoroff's theorem and Lusin's theorem thereafter.

Theorem 5.2.1. Let E be a measurable set with $m(E) < \infty$. If $\{f_n\}$ is a sequence of measurable functions that converges pointwise to a function $f(x)$ on E, then for all $\epsilon > 0$, there exists a closed set $F \subseteq E$ such that $m(E \setminus F) < \epsilon$ and $\{f_n(x)\}$ converges to $f(x)$ uniformly on F.

Proof. Let E be a measurable set with $m(E) < \infty$ and let $\{f_n(x)\}$ be a sequence of measurable functions converging pointwise to $f(x)$ on E. Let $\epsilon > 0$. For each $n \in \mathbb{N}$, define a set $A_n(\epsilon)$ to be the set of elements in E such that $|f_k(x) - f(x)| < \epsilon$ where $k \in \{n, n+1, \ldots\}$. That is

$$A_n(\epsilon) = \{x \in E : \ |f_k(x) - f(x)| < \epsilon, \ k \in \{n, n+1, \ldots\}\}$$

Consider the condition that $\{f_n(x)\}$ converges uniformly to $f(x)$ on any set $A \subseteq E$ means that for all $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that if $x \in A$, then $x \in A_N(\epsilon)$. Equivalently, $\{f_n(x)\}$ converges uniformly to $f(x)$ on A if and only if for all $\epsilon > 0$, there exists an $N \in \mathbb{N}$ such that $A \subseteq A_N(\epsilon)$. Furthermore, we note that, for a fixed $\epsilon > 0$, the collection of sets $A_1(\epsilon), A_2(\epsilon), \ldots, A_N(\epsilon), A_{N+1}(\epsilon), \ldots$ is an ascending sequence of sets. That is,

$$A_1(\epsilon) \subseteq A_2(\epsilon) \subseteq \ldots \subseteq A_N(\epsilon) \subseteq A_{N+1}(\epsilon) \subseteq \ldots.$$

Let $\epsilon > 0$. Since $\{f_n(x)\}$ converges pointwise to $f(x)$ on E, we have for this given ϵ, for each $x \in E$, there exists an $N(\epsilon, x) \in \mathbb{N}$ such that if $n \geq N(\epsilon, x)$, then

$$|f_n(x) - f(x)| < \epsilon$$

So, for each $x \in E$, there exists an $N(\epsilon, x) \in \mathbb{N}$ such that $x \in A_{N(\epsilon,x)}(\epsilon)$ and hence

$$E = \bigcup_{n=1}^{\infty} A_n(\epsilon)$$

We will now use these to prove the main theorem.

For each $k \in \mathbb{N}$, let $\epsilon_k = \dfrac{1}{k} > 0$. For each $n \in \mathbb{N}$ consider the sets $A_n(\epsilon_k) = a_n\left(\dfrac{1}{k}\right)$. Then

$$E = \bigcup_{n=1}^{\infty} A_n\left(\frac{1}{k}\right)$$

Additionally, since $\left\{A_n\left(\dfrac{1}{k}\right)\right\}$ is a sequence of ascending sets that converge to E, for each k, $\dfrac{\epsilon}{2^{k+1}} > 0$, there exists an $N_k \in \mathbb{N}$ such that

$$m\left(E \setminus A_{N_k}\left(\frac{1}{k}\right)\right) < \frac{\epsilon}{2^{k+1}}$$

Now, consider

$$A = \bigcap_{k=1}^{\infty} A_{N_k}\left(\frac{1}{k}\right)$$

Consider the Lebesgue measure of $E \setminus A$.

$$\begin{aligned} m(E \setminus A) &= m\left(E \setminus \bigcap_{k=1}^{\infty} A_{N_k}\left(\frac{1}{k}\right)\right) \\ &= m\left(\bigcup_{k=1}^{\infty}\left(E \setminus A_{N_k}\left(\frac{1}{k}\right)\right)\right) \\ &\leq \sum_{k=1}^{\infty} m\left(E \setminus A_{N_k}\left(\frac{1}{k}\right)\right) \\ &\leq \sum_{k=1}^{\infty} \frac{\epsilon}{2^{k+1}} \\ &\leq \frac{\epsilon}{2} \end{aligned}$$

So for every $\epsilon > 0$, let $k \in \mathbb{N}$ be such that $\frac{1}{k} < \epsilon$. Then

$$A = \bigcap_{k=1}^{\infty} A_{N_k}\left(\frac{1}{k}\right) \subseteq A_{N_k}\left(\frac{1}{k}\right) \subseteq A_{N_k}(\epsilon_k)$$

So, $\{f_n(x)\}$ converges uniformly to $f(x)$ on A. Since $A \subseteq E$ and $m(E) < \infty$, we have that $m(A) < \infty$. So there exists a closed set $F \subseteq A$ such that $m(A \setminus F) < \frac{\epsilon}{2}$ and hence

$$m(E \setminus F) = m(E \setminus A) + m(A \setminus F) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and since $\{f_n(x)\}$ converges uniformly to $f(x)$ on A we also have that $\{f_n(x)\}$ converges uniformly to $f(x)$ on F. □

Lusin's Theorem enables us to approximate any measurable function with a continuous function. We will first prove an analogous theorem for simple measurable functions as follows:

Theorem 5.2.2. Let f be a simple function defined on a measurable set E. Then for each $\epsilon > 0$, there is a continuous function g on $\mathbb{R}$ and a closed set F contained in E for which

$$f \equiv g \text{ on } F \text{ and } m(E \setminus F) < \epsilon.$$

Proof. Let $a_1, a_2, \ldots, a_n$ be the finite number of distinct values taken by f and let the values be taken on the sets $E_1, E_2, \ldots, E_n$ respectively. Since the $a_k's$ are distinct, the sets E_k are disjoint. Then, by the theorem on approximation by closed sets, we get closed sets $F_1, F_2, \ldots, F_n$ such that $F_k \subset E_k$ for each k and

$$m(E_k \setminus F_k) < \epsilon/n.$$

Also, the set $F = \bigcup_{k=1}^{n} F_k$ is closed. Now, since the sets E_k are disjoint, we have by countable additivity,

$$\begin{aligned} m(E \setminus F) &= m\left(\left(\bigcup_{k=1}^{n} E_k\right) \setminus \left(\bigcup_{k=1}^{n} F_k\right)\right) \\ &= m\left(\bigcup_{k=1}^{n} (E_k \setminus F_k)\right) \\ &= \sum_{k=1}^{n} m(E_k \setminus F_k) \\ &< \sum_{k=1}^{n} \frac{\epsilon}{n} \\ &= \epsilon. \end{aligned}$$

Now we define g on F as $g(x) = a_k$ for $x \in F_k$. Since $F_k's$ are disjoint, so g is well-defined. We also see that g is continuous on F (for $x \in F_k$, there is an open interval containing x which is disjoint from the other $F_k's$, so g is constant on this open interval intersecting F). Then by Tietz Extension theorem, g can be extended to the $\mathbb{R}$. This extension is the desired function. □

We will now prove the Lusin's Theorem as follows:

Theorem 5.2.3. (**Lusin's Theorem**) Let f be a real-valued measurable function on E. Then for each $\epsilon > 0$, there is a continuous function g on $\mathbb{R}$ and a closed set F contained in E for which

$$f \equiv g \text{ on } F \text{ and } m(E \setminus F) < \epsilon.$$

Proof. By the previous theorem, we see that the result is true for simple functions. Let f be any arbitrary positive measurable function. First let $m(E) < \infty$. By the Simple Approximation Theorem, there is a sequence $\{f_n\}$ of simple functions defined on E that converges pointwise to f on E. Let $n \in \mathbb{N}$. So, by the previous theorem, for $\epsilon > 0$, there is a closed set F_n and a continuous function g_n defined on $\mathbb{R}$ such that

$$f_n = g_n \text{ on } F_n \text{ and } m(E \setminus F_n) < \epsilon/2^{n+1}.$$

By Egoroff's Theorem, there is a closed set F_0 contained in E such that $\{f_n\}$ converges uniformly to f on F_0 and $m(E \setminus F_0) < \epsilon/2$. Now, define $F = \bigcap_{n=0}^{\infty} F_n$. Then

$$\begin{aligned} m(E \setminus F) &= m\left(E \setminus \bigcap_{n=0}^{\infty} F_n\right) \\ &= m\left(\bigcup_{n=0}^{\infty} (E \setminus F_n)\right) \\ &= m((E \setminus F_0) \cup \left(\bigcup_{n=1}^{\infty} (E \setminus F_n)\right) \\ &< \frac{\epsilon}{2} + \sum_{n=1}^{\infty} \frac{\epsilon}{2^{n+1}} = \epsilon. \end{aligned}$$

The set F is closed. Each f_n is continuous on F since $F \subset F_n$ and $f_n = g_n$ on F_n and g_n is continuous on $\mathbb{R}$. Finally, $\{f_n\}$ converges to f uniformly on F since $F \subset F_0$ and $\{f_n\}$ converges uniformly to f on F_0. However, the uniform limit of continuous functions is continuous, so the restriction of f to the set F is continuous. By Tietz Extension theorem, there is a continuous function g defined on all of $\mathbb{R}$ such that $f = g$ on F. g is the desired function. □

Exercise 5.2.4.

1. Show that Egoroff's theorem is also true if the sequence $\{f_n\}$ converges pointwise to f almost everywhere on a measurable set E of finite measure.
2. Does the Egoroff's theorem hold if the condition $m(E) < \infty$ is removed? Justify your answer.
3. For a measurable function defined on a measurable set E, show that for every $\epsilon > 0$, there is a continuous function g defined on $\mathbb{R}$ such that $m(x \in E : \ f(x) \neq g(x)) < \epsilon$. Hence, show that there exists a sequence $\{g_n\}$ of continuous functions on $\mathbb{R}$ such that $g_n \to f$ a.e on E.

5.3 Convergence in Measure

We are already familiar with convergence of a sequence of functions, and in particular, we have also seen certain theorems on the limits of sequence of measurable functions. Here in this section, we will study a new kind of convergence of a sequence of functions on a set. This concept generalizes pointwise convergence of sequence of functions.

Definition 5.3.1. Let $\{f_n\}$ be a sequence of measurable functions and f, a measurable function defined on a measurable set E. Then the sequence $\{f_n\}$ converges in measure to f in E, if for each $\epsilon > 0$,

$$\lim_{n\to\infty} m(\{x \in E : \ |f_n(x) - f(x)| \geq \epsilon\}) = 0.$$

If $\{f_n\}$ converges in measure to f in measure, then it is also denoted as $f_n \xrightarrow{m} f$.

The concept of $f_n \xrightarrow{m} f$ on E means that for all sufficiently large n, the functions f_n in the sequence $\{f_n\}$ differ from the limit function f by a small quantity with the exception of a set of points whose measure is zero. The above definition can also be equivalently given as follows.

Definition 5.3.2. A sequence $\{f_n\}$ of measurable functions is said to converge in measure to a measurable function f on a set E if for each $\delta > 0$ and $\epsilon > 0$, there exists a positive integer N such that

$$m(\{x \in E : \ |f_n(x) - f(x)| \geq \epsilon\}) < \delta, \ \ \forall n > N.$$

Theorem 5.3.3. If a sequence of measurable functions converges in measure, then the limit function is unique a.e.

Proof. Let $f_n \xrightarrow{m} f$, and $f_n \xrightarrow{m} g$ on E. Now, since

$$|f - g| \leq |f - f_n| + |g - f_n|,$$

we must have, for any $\epsilon > 0$,

$$\{x : \ |f(x) - g(x)| > 2\epsilon\} \subseteq \{x : \ |f(x) - f_n(x)| \geq \epsilon\} \cup \{x : \ |g(x) - f_n(x)| \geq \epsilon\}.$$

But the measure of the set on the right hand side tends to zero as $n \to \infty$. So $f = g$ a.e. on E. □

Theorem 5.3.4. (Riesz Theorem) If a sequence $\{f_n\}$ converges in measure to f on E, then there exists a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ which converges to f a.e on E.

Proof. Let us consider two sequence $\{\epsilon_n\}$ and $\{\delta_n\}$ of positive real numbers such that $\epsilon_n \to 0$ as $n \to \infty$ and $\sum_{n=1}^{\infty} \delta_n < \infty$. We now choose a strictly increasing sequence $\{n_k\}$ of positive integers as follows.

Let n_1 be a positive integer such that

$$m(\{x \in E : |f_{n_1}(x) - f(x)| \geq \epsilon_1\}) < \delta_1.$$

Such a number n_1 must exist since $f_n \xrightarrow{m} f$ on E. Similarly, let n_2 be a positive number such that

$$m(\{x \in E : |f_{n_2}(x) - f(x)| \geq \epsilon_2\}) < \delta_2,$$

and $n_2 \geq n_1$. Continuing in this process, we get a sequence $\{n_k\}$ such that

$$m(\{x \in E : |f_{n_k}(x) - f(x)| \geq \epsilon_k\}) < \delta_k,$$

and $n_k \geq n_{k-1}$ for all k. We shall now show that the subsequence $\{f_{n_k}\}$ converges to f a.e.

Define

$$A_k = \bigcup_{i=k}^{\infty} \{x \in E : |f_{n_i}(x) - f(x)| \geq \epsilon_i\}, \quad k \in \mathbb{N}$$

and

$$A = \bigcap_{k=1}^{\infty} A_k.$$

Clearly, $\{A_k\}$ is a decreasing sequence of measurable sets and hence,

$$m(A) = \lim_{k \to \infty} m(A_k).$$

But,

$$m(A_k) \leq \sum_{i=k}^{\infty} \delta_i \to 0 \text{ as } k \to \infty.$$

Hence, $m(A) = 0$. It remains to be verified that $\{f_{n_k}\}$ converges to f on $E \setminus A$. Indeed, for $x_0 \in E \setminus A$, $x_0 \notin A_{k_0}$ for some positive integer k_0. That means,

$$x_0 \notin \{x \in E : |f_{n_k}(x) - f(x)| \geq \epsilon_k\}, \ k \geq k_0.$$

This gives

$$|f_{n_k}(x_0) - f(x_0)| < \epsilon_k, \ k \geq k_0.$$

But $\epsilon_k \to 0$ as $k \to 0$. Hence,

$$\lim_{k \to \infty} f_{n_k}(x_0) = f(x_0).$$

Hence $\{f_{n_k}\}$ converges a.e to f on E. □

Theorem 5.3.5. Let $f_n \xrightarrow{m} f$ and $g_n \xrightarrow{m} g$ on E. Then

1. $f_n + g_n \xrightarrow{m} f + g$;
2. $\alpha f_n \xrightarrow{m} \alpha f$, α is a real number;

3. $f_n^+ \xrightarrow{m} f^+$, $f_n^- \xrightarrow{m} f^-$ and $|f_n| \xrightarrow{m} |f|$.

Further, if $m(E) < \infty$, then

a $f_n^2 \xrightarrow{m} f^2$;

b $f_n.g_n \xrightarrow{m} f.g$.

Proof. Since $f_n \xrightarrow{m} f$ and $g_n \xrightarrow{m} g$, so for each $\epsilon > 0$,

$$\lim_{n\to\infty} m(\{x \in E : \ |f_n(x) - f(x)| \geq \epsilon\}) = 0,$$

and

$$\lim_{n\to\infty} m(\{x \in E : \ |g_n(x) - g(x)| \geq \epsilon\}) = 0.$$

1. Now,
$$|(f_n + g_n)(x) - (f + g)(x)| \leq |f_n(x) - f(x)| + |g_n(x) - g(x)|.$$
Also,
$$\{x \in E : \ |(f_n+g_n)(x)-(f+g)(x)| \geq 2\epsilon\} \subseteq \{x \in E : \ |f_n(x)-f(x)| \geq \epsilon\}\cup\{x \in E : \ |g_n(x)-g(x)| \geq \epsilon\}.$$
Since the measure of both the sets on the right hand side tends to zero as $n \to \infty$, and since $\epsilon > 0$ is arbitrary, so the result follows.

2. If $\alpha = 0$, it follows trivially. Let $\alpha \neq 0$. Then,
$$\{x \in E : \ |\alpha f_n(x) - \alpha f(x)| \geq \epsilon\} = \{x \in E : \ |f_n(x) - f(x)| \geq \frac{\epsilon}{|\alpha|}.$$
Since the set on the right hand side has measure zero for $n \to \infty$, the result follows.

3. We know that,
$$\begin{aligned} \left|f_n^+ - f^+\right| &\leq |f_n - f| \\ \left|f_n^- - f^-\right| &\leq |f_n - f|. \end{aligned}$$
Also,
$$||f_n| - |f|| \leq |f_n - f|.$$
The rest can be similarly done as the preceding part and is left as exercise.

Now, let $m(E) < \infty$.

a Left as exercise.

b Left as exercise.

□

The condition $m(E) < \infty$ in the last two results can not be removed as can be seen from the following example.

Example 5.3.6. Take $E = (0, \infty)$. Consider for each n, $f_n(x) = x$, $x \in E$. Then $f(x) = x$, $x \in E$. Let $g_n(x) = c_n$, where $\{c_n\}$ is a sequence of positive real numbers such that $c_n \to 0$ as $n \to \infty$.

Here, $m(E) = \infty$, and we note that

$$m(\{x : |f_n(x)g_n(x) - f(x)g(x)| \geq \epsilon\}) = m(\{x : |c_n x| \geq \epsilon\}) = \infty, \ \ \forall n.$$

This shows that $f_n.g_n \xrightarrow{m} f.g$ on E.

Definition 5.3.7. A sequence of functions is said to be **fundamental** with respect to a particular kind of convergence if it forms a Cauchy sequence in that sense. Thus a sequence $\{f_n\}$ is fundamental in measure if for any $\epsilon > 0$,

$$\lim_{m,n\to\infty} m(\{x : |f_n(x) - f_m(x)| > \epsilon\}) = 0.$$

Theorem 5.3.8. If a sequence $\{f_n\}$ converges in measure to f, then $\{f_n\}$ is fundamental in measure.

Proof. It follows from the relation

$$\{x : |f_n(x) - f_p(x)| \geq \epsilon\} \subseteq \left\{x : |f_n(x) - f(x)| \geq \frac{\epsilon}{2}\right\} \cup \left\{x : |f_p(x) - f(x)| \geq \frac{\epsilon}{2}\right\}.$$

□

We now prove a 'completeness' theorem for convergence in measure.

Theorem 5.3.9. If $\{f_n\}$ is a sequence of measurable functions which is fundamental in measure, then there exists a measurable function f such that $f_n \xrightarrow{m} f$.

Proof. For every integer k, we can find n_k such that $n, m \geq n_k$,

$$m\left(\left\{x : |f_n(x) - f_m(x)| \geq \frac{1}{2^k}\right\}\right) < \frac{1}{2^k},$$

and we may assume that for each k, $n_{k+1} > n_k$. Let

$$E_k = \left\{x : |f_{n_k}(x) - f_{n_{k+1}}(x)| \geq \frac{1}{2^k}\right\}$$

Then if $x \notin \bigcup_{k=m}^{\infty} E_k$, we have for $r > s \geq m$

$$|f_{n_r}(x) - f_{n_s}(x)| \leq \sum_{i=s+1}^{r} |f_{n_i}(x) - f_{n_{i-1}}(x)| < \sum_{i=s+1}^{r} \frac{1}{2^i} = \frac{1}{2^s}$$

So, $\{f_{n_k}(x)\}$ is a Cauchy sequence for each $x \notin \limsup E_k = \bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} E_k$. But, for all m,

$$m(\limsup E_k) \leq m\left(\bigcup_{k=m}^{\infty} E_k\right) \leq \sum_{k=m}^{\infty} \frac{1}{2^k} = \frac{1}{2^{m-1}}. \tag{5.3.1}$$

So $\{f_{n_k}\}$ converges a.e to some measurable function f. Also from (5.3.1), we have that $\{f_{n_k}\}$ is uniformly fundamental in $\mathbb{R} \setminus \bigcup_{k=m}^{\infty} E_k$, for each m. So, $f_{n_k} \to f$ uniformly on $\mathbb{R} \setminus \bigcup_{k=m}^{\infty} E_k$, and hence, for every positive ϵ,

$$m(\{x : |f_{n_k}(x) - f(x)| > \epsilon/2\}) \to 0 \text{ as } k \to \infty. \tag{5.3.2}$$

But,

$$\{x : |f_n(x) - f(x)| > \epsilon\} \subseteq \{x : |f_n(x) - f_{n_k}(x)| > \epsilon/2\} \cup \{x : |f(x) - f_{n_k}(x)| > \epsilon/2\}.$$

If n and n_k are sufficiently large, the measure of the first set on the right is arbitrarily small, as $\{f_n\}$ is fundamental in measure. But the second set has been shown to have arbitrarily small measure by (5.3.2) and the result follows. □

The Riesz theorem can also be proved as a corollary of the above theorem.

Corollary 5.3.10. Let $f_n \xrightarrow{m} f$, where each f_n and f are measurable functions. Then there exists a subsequence $\{n_l\}$ such that $f_{n_l} \to f$ a.e.

Proof. Clearly, $\{f_n\}$ is fundamental in measure, so from the proof of the previous theorem, we can find a subsequence $\{f_{n_l}\}$ and a measurable function g such that $f_{n_l} \to g$ a.e and in measure. But $f_{n_l} \xrightarrow{m} f$ so by a previous theorem, $f = g$, a.e. Hence proved. □

Exercise 5.3.11.

1. Show that the a.e limit f of a sequence of measurable functions $\{f_n\}$ on a measurable set E implies that $f_n \xrightarrow{m} f$ on E. Is the converse true? Justify.
2. Show that if $f_n \to f$ in measure and $g_n \to g$ in measure, then $f_n - g_n \to f - g$ in measure.
3. Prove that almost uniform convergence implies convergence in measure.
4. Prove that every subsequence of a sequence fundamental in measure is again a fundamental in measure.

Sample Questions

1. State and prove Egoroff's theorem.
2. State and prove Lusin's theorem.
3. Define convergence in measure. Show that the limit of convergence in measure of a sequence $\{f_n\}$ of measurable functions is unique a.e.
4. State and prove Riesz theorem for convergence in measure.
5. Show that the property of convergence in measure is closed under function addition. Is the same true for function multiplication? Justify your answer.
6. When is a sequence of measurable functions said to be fundamental in measure? Show that pointwise convergent sequence of measurable functions is fundamental in measure.
7. Let $\{f_n\}$ be a sequence of measurable functions which is fundamental in measure, then there exists a measurable function to which $\{f_n\}$ convereges in measure.

Unit 6

Course Structure

- The Riemann Integral
- The Lebesgue integral : Integrals of simple functions and bounded function defined one a measurable set with finite measure

6.1 Introduction

The theory of Riemann integration though very thoroughly useful and adequate for solving various problems, in both pure and applied streams, has its own drawbacks. It does not meet the needs of a number of important branches of mathematics and physics of comparatively recent development. First of all, the Riemann integral of a function is defined on a closed interval and cannot be defined on an arbitrary set. Investigations in probability theory, partial differential equations, hydromechanics and quantum mechanics often pose problems which require integration over sets. Second and more important is the fact that the Riemann integrability depends upon the continuity of the function. Of course, there are functions which are discontinuous and yet Riemann-integrable, but these functions are continuous almost everywhere. Again, given a sequence of Riemann integrable functions converging to some function in a domain, the limit of the sequence of integrated functions may not be the Riemann integral of the limit function. In fact, the Riemann integral of the limit function may not even exist. This is a major drawback of the Riemann theory of integration, apart from the fact that even relatively simple functions are not integrable in the sense of Riemann. H. Lebesgue in his classical work, introduced the concept of an integral, known after his name the Lebesgue integral, based on the measure theory that generalizes the Riemann integral. It has the advantage that it takes care of both bounded and unbounded functions and simultaneously allows their domains to be more general sets and thereby enlarges the class of functions for which the Lebesgue integral is defined. Also, it gives more powerful and useful convergence theorems relating to the interchange of the limit and integral valid under less restrictive conditions required for the Riemann integral. We shall start with recollecting the Riemann integral and then gradually develop the theory of Lebesgue integral and see that in fact, Lebesgue integral is a generalisation of Riemann integral.

Objectives

After reading this unit, you will be able to

- revise the idea of Riemann integrals from a new perspective
- define the integral of simple functions on measurable sets
- use the definition of simple function to define the integral of bounded measurable functions on measurable sets on $\mathbb{R}$

6.2 Riemann Integral: A short recapitulation

We shall define the Riemann integral of a step function and then will extend it to more general bounded functions f on $[a, b]$ via approximation from above and below by step functions.

Definition 6.2.1. A real-valued function ϕ on $[a, b]$ is called a step function if there is a partition

$$a = x_0 < x_1 < \ldots < x_n = b$$

of the interval such that ϕ is constant on each subinterval $I_k = (x_{k-1}, x_k)$; that is,

$$\phi_k(x) = c_k, \text{ for } x \in I_k, \quad k = 1, 2, \ldots, n,$$

with $\phi_k(x_k) = d_k$, $k = 0, 1, 2, \ldots, n$.

Definition 6.2.2. Let ϕ be a step function on $[a, b]$:

$$\phi(x) = \begin{cases} c_k, & x_{k-1} < x < x_k, \quad k = 1, 2, \ldots, n \\ d_k, & k = 0, 1, \ldots, n, \quad x = x_k. \end{cases}$$

The Riemann integral of ϕ on $[a, b]$, denoted by $\int_a^b \phi(x)dx$, is

$$\int_a^b \phi(x)dx = \sum_{k=1}^{n} c_k \left(x_k - x_{k-1}\right).$$

We could write $\phi = \sum_{1}^{n} c_k \mathcal{X}_{(x_{k-1}+x_k)} + \sum_{0}^{n} d_k \mathcal{X}_{\{x_k\}}$, and $\int_a^b \phi(x)dx$

$= \sum_{k=1}^{\pi} c_k m\left((x_{k-1}, x_k)\right) + \sum_{k=0}^{n} d_k m\left(\{d_k\}\right) = \sum_{k=1}^{n} c_k \left(x_k - x_{k-1}\right)$.

The step function's values at the endpoints of the subintervals have no bearing on the existence or value of the Riemann integral of a step function (d_k does not appear in the definition of the integral).

$$\phi_1(x) = \begin{cases} 1, & 0 \le x < 1 \\ 3, & x = 1 \\ 2, & 1 < x \le 2 \end{cases} \quad \text{and} \quad \phi_2(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 2, & 1 < x \le 2 \end{cases}$$

$\phi_1 = \phi_2$ for $0 < x < 1$ and $1 < x < 2, \phi_1(1) \neq \phi_2(1)$, but

$$\int_0^2 \phi_1(x)dx = 1 \cdot 1 + 2 \cdot 1 = \int_0^2 \phi_2(x)dx.$$

Also, it should be noted that the value of the Riemann integral of a step function is independent of the choice of the partition of $[a, b]$ as long as the step function is constant on the open subintervals of the partition, for example,

$$\phi(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 2, & 1 < x \le 2 \end{cases}$$

Partition: $\{0,1,2\}$, $\int_0^2 \phi(x)dx = 1\cdot 1 + 2\cdot 1 = 3$, Partition:

$$\{0,1/2,1,7/4,2\}, \int_0^2 \phi(x)dx = 1\cdot 1/2 + 1\cdot 1/2 + 2\cdot 3/4 +$$
$$2\cdot 1/4 = 3$$

More formally, the Riemann integral of a step function is well defined; it is independent of the particular representation of ϕ. For example, if $\phi(x) = c_k, x_{k-1} < x < x_k$ and we add another partition point x^*, $x_{k-1} < x^* < x_k$, we have

$$c_k\left(x_k - x_{k-1}\right) = c_k\left(x_k - x^* + x^* - x_{k-1}\right) = c_k\left(x_k - x^*\right) + c_k\left(x^* - x_{k-1}\right).$$

We will now state few obvious properties of step functions and their corresponding integrals in the following.

Theorem 6.2.3. If ϕ and ψ are step functions on $[a,b]$, and k is any real number, then

1. $(k\phi)$ is a step function on $[a,b]$, and $\int_a^b (k\phi)(x)dx = k\int_a^b \phi(x)dx$ (homogeneous);
2. $(\phi+\psi)$ is a step function on $[a,b]$, and
$$\int_a^b (\phi+\psi)(x)dx = \int_a^b \phi(x)dx + \int_a^b \psi(x)dx \text{ (additivity);}$$
3. $\int_a^b \phi(x)dx \le \int_a^b \psi(x)dx$ if $\phi \le \psi$ on $[a,b]$ (monotone);
4. If $a < c < b$, the integrals $\int_a^c \phi(x)dx$, $\int_c^b \phi(x)dx$ exist and
$$\int_a^c \phi(x)dx + \int_c^b \phi(x)dx = \int_a^b \phi(x)dx. \text{ (additive on the domain)}$$

We now define the Riemann integral for more general bounded function f on $[a,b]$.

Definition 6.2.4. Let f be a bounded function on $[a,b]$, say $\alpha \le f \le \beta$, for $x \in [a,b]$. Let $\phi,\ \psi$ denote arbitrary step functions on $[a,b]$ such that $\phi \le f \le \psi$. The lower Riemann integral of f on $[a,b]$, $\underline{\int_a^b} f(x)dx$, is defined as

$$\underline{\int_a^b} f(x)dx = \sup\left\{\int_a^b \phi(x)dx |\ \phi \le f,\ \ \phi \text{ is a step function}\right\}.$$

The upper Riemann integral of f on $[a,b]$, $\overline{\int_a^b} f(x)dx$, is defined as

$$\overline{\int_a^b} f(x)dx = \inf\left\{\int_a^b \psi(x)dx |\ f \le \psi,\ \ \psi \text{ is a step function}\right\}.$$

Since $\alpha \le f$, the set $\left\{\int_a^b \phi(x)dx |\ \phi \le f,\ \ \phi \text{ is a step function}\right\}$ is not empty and since $\alpha \le f \le \beta$ implies

$$\int_a^b \phi(x)dx \le \int_a^b \beta dx = \beta(b-a),$$

the set is bounded above. The least upper bound is a real number. The lower Riemann integral on a closed bounded interval is well-defined. Similarly, for the upper Riemann integral.

Since $\phi \leq \psi$, $\int_a^b \phi(x)dx \leq \int_a^b \psi(x)dx$ by monotonicity of step functions. Since ϕ is arbitrary, we may interpret this inequality as saying $\int_a^b \psi(x)dx$ is an upper bound for the set

$$\left\{ \int_a^b \phi(x)dx |\ \phi \leq f,\ \ \phi \text{ is a step function} \right\}.$$

But, $\underline{\int_a^b} f(x)dx$ is the smallest upper bound. So, we have,

$$\underline{\int_a^b} f(x)dx \leq \int_a^b \psi(x)dx.$$

Again, we can say that $\underline{\int_a^b} f(x)dx$ is a lower bound of the set

$$\left\{ \int_a^b \psi(x)dx |\ f \leq \psi,\ \ \psi \text{ is a step function} \right\}.$$

Since the upper Riemann integral $\overline{\int_a^b} f(x)dx$ is the greatest lower bound,

$$\underline{\int_a^b} f(x)dx \leq \overline{\int_a^b} f(x)dx.$$

It follows that a bounded function f on $[a.b]$ satisfies

$$\int_a^b \phi(x)dx \leq \underline{\int_a^b} f(x)dx \leq \overline{\int_a^b} f(x)dx \leq \int_a^b \psi(x)dx$$

for any step functions $\phi \leq f \leq \psi$ on $[a, b]$.

When this approximation from above and below approach a common value, then f will be Riemann integrable.

Definition 6.2.5. A bounded function f on $[a, b]$ is Riemann integrable on $[a, b]$ whenever $\underline{\int_a^b} f(x)dx = \overline{\int_a^b} f(x)dx$. We denote the common value by $\int_a^b f(x)dx$.

We state a necessary and sufficient condition for a bounded function f on $[a, b]$ to be Riemann integrable.

Theorem 6.2.6. A bounded function f on $[a, b]$ is Riemann integrable if and only if for every $\epsilon > 0$, we have step functions ϕ and ψ, $\phi \leq f \leq \psi$ on $[a, b]$, so that

$$0 \leq \int_a^b \psi(x)dx - \int_a^b \phi(x)dx = \int_a^b [\psi(x) - \phi(x)]dx < \epsilon.$$

We have known from previous knowledge that continuous functions are Riemann integrable. But what about discontinuous functions? We have seen that functions having finite number of discontinuities are continuous (as an example, the function $[x]$ on $[0,3]$ is Riemann integrable). However, functions like the following

$$\begin{aligned} f(x) &= 1, \ x \in [0,1] \cap \mathbb{Q} \\ &= 0, \ x \in [0,1] \setminus \mathbb{Q} \end{aligned}$$

are not Riemann integrable. In fact, the discontinuities of f here is the whole set $[0,1]$. Then, one must be thinking that for f to be Riemann integrable, the set of its discontinuities must be "small". Here, this "small"-ness is measured using Lebesgue measure. The following theorem gives us an answer for the relationship between Riemann integrability and continuity for any f on $[a,b]$.

Theorem 6.2.7. A bounded function on a closed bounded interval is Riemann integrable if and only if the function is continuous a.e on the interval.

Before concluding this section, we will state the following theorem that shows that the integral properties of step functions are retained by Riemann integrable functions.

Theorem 6.2.8. If bounded functions f and g are Riemann integrable on $[a,b]$, and k is any real number, the

1. (kf) is Riemann integrable on $[a,b]$, and

$$\int_a^b kf(x)dx = k\int_a^b f(x)dx \quad \text{(homogeneous);}$$

2. $(f+g)$ is Riemann integrable on $[a,b]$, and

$$\int_a^b (f+g)(x)dx = \int_a^b f(x)dx + \int_a^b g(x)dx \quad \text{(additive);}$$

3. $\int_a^b f(x)dx \leq \int_a^b g(x)dx$ if $f \leq g$ on $[a,b]$ (monotone);

4. If $a < c < b$, f is Riemann integrable on $[a,c]$ and $[c,b]$, and

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx \quad \text{(additive on the domain);}$$

5. If $\alpha \leq f \leq \beta$ on $[a,b]$,

$$\alpha(b-a) \leq \int_a^b f(x)dx \leq \beta(b-a) \quad \text{(mean value).}$$

Exercise 6.2.9. Show that the function f defined on $[0,1]$ as

$$\begin{aligned} f(x) &= 1; \ x \in [0,1] \cap \mathbb{Q} \\ &= 0; \ x \in [0,1] \setminus \mathbb{Q} \end{aligned}$$

is not Riemann integrable.

6.3 Lebesgue Integral

We develop the theory of Lebesgue integral for a bounded function f on a set E of finite Lebesgue measure. The treatment is parallel to that of Riemann integral, replacing step functions with simple functions. For that, we first need to define the integral of simple functions.

Definition 6.3.1. Suppose ϕ is a simple function defined on a measurable set E, that is

$$\phi(x) = \sum_{k=1}^{n} c_k \chi_{E_k}(x)$$

with $\bigcup_{k=1}^{n} E_k = E$, E_k are mutually disjoint with $m(E) < \infty$, c_k are real numbers. The Lebesgue integral of ϕ on E, $\int_E \phi$ is defined as

$$\int_E \phi = \sum_{k=1}^{n} c_k m(E_k).$$

We would like to remark on a few things.

1. If ϕ is a step function on $[a, b]$, ϕ is a simple function, and

$$\begin{aligned}\int_E \phi &= \sum_{k=1}^{n} c_k m\left(E_k\right) = \sum_{k=1}^{n} c_k m\left(\left(x_{k-1}, x_k\right)\right) + \sum_{k=1}^{n+1} d_k m\left(x_{k-1}\right) \\ &= \sum_{k=1}^{n} c_k \left(x_k - x_{k-1}\right) = \int_a^b \phi(x) dx.\end{aligned}$$

   The Lebesgue integral of a step function agrees with the Riemann integral of a step function.

2. Because of the many possible representations of ϕ we must check to see that our definition is not ambiguous. Suppose

$$E = \bigcup_{i=1}^{n} E_i = \bigcup_{j=1}^{m} F_j.$$

   Somehow we want to refine both of these decompositions of E into a "common" decomposition. Each set in this "common" decomposition would be a subset of E_i and F_j, * And" suggests intersection:

$$E = \bigcup_j \left(E_i \cap F_j\right) = \bigcup_j \left(E_i \cap F_j\right).$$

   Now suppose $\phi = \sum_i c_i X_{E_i} = \sum_j d_j X_{F_j}$, $\{E_i\}$ and $\{F_j\}$ mutually disjoint collections of Lebesgue measurable sets, $m(E) < \infty$. Then we claim

$$\sum_i c_i m\left(E_i\right) = \sum_j d_j m\left(F_j\right),$$

   the Lebesgue integral is independent of the representation. We know the nonempty $E_1 \sqcap F_j$ are measurable and mutually disjoint. Because

$$\bigcup_{i=1}^{n} E_i = E = \bigcup_{j=1}^{m} F_j,$$

$$\begin{aligned}\sum_i c_i m(E_i) &= \sum_i c_i \sum_j m(E_i \cap F_j) = \sum_j \sum_j c_i m(E_i \cap F_j) \\ &= \sum_j \sum_i d_j m(E_i \cap F_j) = \sum_j d_j \sum_j m(E_i \cap F_j) \\ &= \sum_j d_j m(F_j)\end{aligned}$$

since if $E_i \cap F_j \neq \emptyset$, then $c_i = c_i \chi_{E_1} = \phi = d_j \chi_{E_j} = d_j$, and the argument is complete.

Example 6.3.2. We calculate some Lebesgue integrals of simple functions.

1.
$$\phi(x) = \begin{cases} -1, & 0 < x \leq 1 \\ 2, & 1 < x \leq 2 \\ 0, & 2 < x \leq 3. \end{cases}$$

$$\begin{aligned}\phi &= -1\chi_{(0,1]} + 2X_{(1,2]} \\ \int_{(0,3]} \phi &= -1 \cdot 1 + 2 \cdot 1 = 1 = \int_0^3 \phi(x) dx.\end{aligned}$$

2.
$$\phi(x) = \begin{cases} 0, & x \text{ rational}, \ 0 \leq x \leq 1 \\ 1, & x \text{ irrational}, 0 \leq x \leq 1. \end{cases}$$

$$\begin{aligned}\phi &= -1\chi_{[0,1]\setminus\mathbb{Q}} \\ \int_{[0,1]} \phi &= 1m([0,1] \setminus \mathbb{Q}) = 1.\end{aligned}$$

Theorem 6.3.3. If ϕ and ψ are simple functions defined on a set E with finite measure, and k is any real number, then

1. $(k\phi)$ is a simple function on E, and $\int_E (k\phi) = k \int_E \phi$ (homogeneous);
2. $\phi + \psi$ is a simple function on E, and $\int_E (\phi + \psi) = \int_E \phi + \int_E \psi$ (additive);
3. $\int_E \phi \leq \int_E \psi$ if $\phi \leq \psi$ on E (monotone);
4. If A and B are disjoint measurable subsets of E with $E = A \cup B$, the integrals $\int_A \phi$ and $\int_B \phi$ exist and $\int_E \phi = \int_A \phi + \int_B \phi$ (additive on the domain).

Proof. Suppose $\phi = \sum_{i=1}^n c_i \chi_{E_i}$, and $\psi = \sum_{j=1}^m d_j \chi_{F_j}$, where

$$E = \bigcup_{i=1}^n E_i = \bigcup_{j=1}^m F_j,$$

$\{E_i\}$ and $\{F_j\}$ are mutually disjoint collections of measurable subsets of E.

1. $\displaystyle\int_E k\phi = \sum_{i=1}^{n}(kc_i)\chi_{E_i} = k\sum_{i=1}^{n} c_i\chi_{E_i} = k\int_E \phi.$

2. Let $A_{ij} = E_i \cap F_j$. The non-empty sets in the collection of A_{ij}, $1 \le i \le n$, $1 \le j \le m$, are mutually disjoint measurable sets whose union is E. Then

$$\phi + \psi = \sum_{i=1}^{n}\sum_{j=1}^{m}(c_i + d_j)\chi_{A_{ij}},$$

and

$$\begin{aligned}
\int_E(\phi+\psi) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(c_i+d_j)m(A_{ij}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} c_i m(E_i\cap F_j) + \sum_{j=1}^{m}\sum_{i=1}^{n} d_j m(E_i\cap F_j) \\
&= \sum_{i=1}^{n} c_i \sum_{j=1}^{m} m(E_i\cap F_j) + \sum_{j=1}^{m} d_j \sum_{i=1}^{n} m(E_i\cap F_j) \\
&= \sum_{i=1}^{n} c_i m(E_i) + \sum_{j=1}^{m} d_j m(F_j) \\
&= \int_E \phi + \int_E \psi.
\end{aligned}$$

3. If $\phi \le \psi$, then $\psi - \phi$ is a non-negative simple function on E, whose integral will be non-negative by the definition of the integral, and then from parts 1 and 2, we have

$$0 \le \int_E(\psi-\phi) = \int_E \psi + \int_E(-\phi) = \int_E \psi - \int_E \phi$$

and thus

$$\int_E \phi \le \int_E \psi.$$

4. We first observe that $\chi_E = \chi_A + \chi_B$, $E = A \cup B$, $A \cap B = \emptyset$.

$$\begin{aligned}
\int_E \phi = &\quad \sum_{i=1}^{n} c_i m(E_i) = \sum_{i=1}^{n} c_i m\left((E_i\cap A)\cup(E_i\cap B)\right) \\
= &\quad \sum_{i=1}^{n} c_i m\left(E_i\cap A\right) + \sum_{i=1}^{n} c_i m\left(E_i\cap B\right).
\end{aligned}$$

But, $\{E_i \cap A\}$ and $\{E_i \cap B\}$ are collections of mutually disjoint measurable subsets of A, B respectively, with

$$A = \bigcup_{i=1}^{n}(E_i\cap A), \quad B = \bigcup_{i=1}^{n}(E_i\cap B),$$

and since the integral is independent of the decomposition used, we have,

$$\int_A \phi = \sum_{i=1}^{n} c_i m\left(E_i\cap A\right), \quad \int_B \phi = \sum_{i=1}^{n} c_i m\left(E_i\cap B\right),$$

and thus,

$$\int_E \phi = \int_A \phi + \int_B \phi.$$

□

We will now define the Lebesgue integral of any bounded function defined on a measurable set E with finite measure.

Definition 6.3.4. Suppose f is a bounded function defined on a measurable set E with finite measure, say $\alpha \leq f \leq \beta$ on E, $m(E) < \infty$. Let ϕ and ψ denote simple functions such that $\phi \leq f \leq \psi$ on E. The lower Lebesgue integral of f on E, $\underline{\int_E} f$ is given by

$$\underline{\int_E} f = \sup\left\{\int_E \phi |\ \ \phi \leq f,\ \ \phi \text{ is a simple function}\right\}.$$

The upper Lebesgue integral of f on E, $\overline{\int_E} f$ is given by

$$\overline{\int_E} f = \inf\left\{\int_E \psi |\ \ f \leq \psi,\ \ \psi \text{ is a simple function}\right\}.$$

The constant simple functions α, β assure us that the lower and upper Lebesgue integrals are well-defined, since the appropriate sets will be non-empty and bounded above and below. Now, by monotonicity, if ϕ and ψ are simple functions such that $\phi \leq f \leq \psi$, then

$$\int_E \phi \leq \int_E \psi.$$

Because ψ is arbitrary, $\int_E \phi$ is a lower bound of the set $\left\{\int_E \psi |\ \ f \leq \psi,\ \ \psi \text{ is a simple function}\right\}$. Since $\overline{\int_E} f$ is the greatest lower bound, $\int_E \phi \leq \overline{\int_E} f$. Similarly, $\int_E \phi \leq \overline{\int_E} f$ for every $\phi \leq f$. Thus, $\overline{\int_E} f$ is an upper bound of the set $\left\{\int_E \phi |\ \ \phi \leq f,\ \ \phi \text{ is a simple function}\right\}$. Since $\underline{\int_E} f$ is the least upper bound, so $\underline{\int_E} f \leq \overline{\int_E} f$.

We have established that a bounded function f on a set E of finite measure has lower as well as upper integrals satisfying

$$\int_E \phi \leq \underline{\int_E} f \leq \overline{\int_E} f \leq \int_E \psi$$

for any simple function ϕ and ψ satisfying $\phi \leq f \leq \psi$ on E.

Similar to the notion of Riemann integrability, it is natural to guess that f, as above, will be Lebesgue integrable when the two lower and upper integrals agree. We formally define it as follows.

Definition 6.3.5. A bounded function f, defined on a measurable set E with finite measure is Lebesgue integrable on E whenever the lower and upper integrals are the same. Denote the common value by $\int_E f$. Thus

$$\underline{\int_E} f = \int_E f = \overline{\int_E} f.$$

In this case, $\phi \leq f \leq \psi$ on E implies $\int_E \phi \leq \int_E f \leq \int_E \psi$ for any simple functions ϕ, ψ.

Theorem 6.3.6. Let f be a bounded function on the interval $[a,b]$. If f is Riemann integrable on $[a,b]$, then f is Lebesgue integrable on $[a,b]$ and

$$\int_a^b f(x)dx = \int_{[a,b]} f.$$

Proof.

$$\begin{aligned}
\underline{\int_a^b} f(x)dx &= \sup\left\{\int_a^b \phi(x)dx |\ \phi \leq f,\ \ \phi \text{ is a step function}\right\} \\
&= \sup\left\{\int_{[a,b]} \phi |\ \phi \leq f,\ \ \phi \text{ is a step function}\right\} \\
&\leq \sup\left\{\int_{[a,b]} \phi |\ \phi \leq f,\ \ \phi \text{ is a simple function}\right\} \\
&= \underline{\int_{[a,b]}} f \\
&\leq \overline{\int_{[a,b]}} f \\
&\leq \inf\left\{\int_{[a,b]} \psi |\ f \leq \psi,\ \ \psi \text{ is a simple function}\right\} \\
&\leq \inf\left\{\int_{[a,b]} \psi |\ f \leq \psi,\ \ \psi \text{ is a step function}\right\} \\
&= \inf\left\{\int_{[a,b]} \psi(x)dx |\ f \leq \psi,\ \ \psi \text{ is a step function}\right\} \\
&= \overline{\int_a^b} f(x)dx.
\end{aligned}$$

Since f is Riemann integrable, $\underline{\int_a^b} f(x)dx = \overline{\int_a^b} f(x)dx$ and the conclusion follows. □

Thus, Riemann integrability implies Lebesgue integrability. Hence, Lebesgue integrability is a generalised idea. The converse is not true in general as can be seen from the example below.

Example 6.3.7. Consider the function f on $[0,1]$ as follows:

$$\begin{aligned}
f(x) &= 1,\ \ x \in [0,1] \cap \mathbb{Q} \\
&= 0,\ \ x \in [0,1] \setminus \mathbb{Q}.
\end{aligned}$$

It can be checked that $\underline{\int_0^1} f(x)dx = 0$, $\overline{\int_0^1} f(x)dx = 1$. Thus, f is not Riemann integrable. However, it is trivially integrable. It is a simple function and

$$\int_{[0,1]} f = 1m\left([0,1] \cap \mathbb{Q}\right) = 0.$$

Next we will prove a necessary and sufficient condition for a bounded function f on a set with finite measure to be Lebesgue integrable.

Theorem 6.3.8. A bounded function f, defined on a set E with finite measure, is Lebesgue integrable if and only if for every $\epsilon > 0$, we have simple functions ϕ and ψ such that $\phi \leq f \leq \psi$ on E, so that

$$0 \leq \int_E \psi - \int_E \phi = \int_E (\psi - \phi) < \epsilon.$$

Proof. Assume the bounded function f is Lebesgue integrable on the measurable set $E, m(E) < \infty$, and let $\epsilon > 0$. From the definitions of greatest lower bound and least upper bound we have simple functions $\hat{\phi}$ and $\hat{\psi}, \hat{\phi} \leq f \leq \hat{\psi}$ on E, so that

$$\begin{aligned}\int_E f - \frac{\epsilon}{2} = \underline{\int_E} f - \frac{\epsilon}{2} < \int_E \hat{\phi} \leq \underline{\int_E} f \\ \leq \overline{\int_E} f \leq \int_E \hat{\psi} < \overline{\int_E} f + \frac{\epsilon}{2} = \int_E f + \frac{\epsilon}{2}.\end{aligned}$$

Thus $0 \leq \int_E \hat{\psi} - \int_E \hat{\phi} = \int_E (\hat{\psi} - \hat{\phi}) < \epsilon$.

For the other direction, let $\epsilon > 0$ be given along with simple functions ϕ and ψ, $\phi \leq f \leq \psi$, so that $0 \leq \int_E \psi - \int_E \phi = \int_E (\psi - \phi) < \epsilon$. Then again, from greatest lower bound and least upper bound properties

$$\int_E \phi \leq \underline{\int_E} f \leq \overline{\int_E} f \leq \int_E \psi.$$

Hence $0 \leq \overline{\int_E} f - \underline{\int_E} f < \epsilon$ and the conclusion follows from the arbitrary nature of ϵ. □

Note 6.3.9. When the bounded function f is Lebesgue integrable on E, $m(E) < \infty$, we have $\phi \leq f \leq \psi$ on E, $\int_E \phi \leq \int_E f \leq \int_E \psi$, and $\int_E (\psi - \phi) < \epsilon$, for some simple functions ϕ and ψ.

We will conclude this unit by stating the following theorem that gives us an idea of the relationship between measurability and integrability of a bounded function f.

Theorem 6.3.10. Let f be a bounded function on a set E with finite measure. Then f is Lebesgue integrable if and only if f is measurable on E.

Proof. Let $|f| \leq M$ on E and assume that f is measurable on E. We will show f is Lebesgue integrable by constructing simple functions $\hat{\phi}$ and $\hat{\psi}$ such that $\hat{\phi} \leq f \leq \hat{\psi}$ on E, so that

$$0 \leq \int_E (\hat{\psi} - \hat{\phi}) < \epsilon.$$

Let $E_k = \left\{ x \frac{k-1}{n} M < f(x) \leq \frac{k}{n} M \right\}, -n \leq k \leq n$. Then $E = \bigcup E_k, E_k$ mutually disjoint measurable sets. We define $\hat{\phi}, \hat{\psi}$ as follows.

$$\hat{\phi} = \frac{M}{n} \sum_{k=-n}^{n} (k-1) \chi_{E_k} \quad \text{and} \quad \hat{\psi} = \frac{M}{n} \sum_{k=-n}^{n} k \chi_{E_k}.$$

Obviously $\hat{\phi} \leq f \leq \hat{\psi}$ and $0 \leq \int_E (\hat{\psi} - \hat{\phi}) = \frac{M}{n} m(E)$. Thus f is Lebesgue integrable on E by the previous theorem.

Conversely suppose f is Lebesgue integrable and bounded on the set E, $m(E) < \infty$. We will show that f is a measurable function on E by showing that f equals almost everywhere the "inf" of a sequence of simple functions.

Since f is bounded and Lebesgue integrable on E, we have simple functions ϕ_n and ψ_n so that $\phi_n \leq f \leq \psi_n$ on E, $\int_E \phi_n \leq \int_E f \leq \int_E \psi_n$, and

$$\int_E (\psi_n - \phi_n) < \frac{1}{n}, \quad n = 1, 2, 3, \ldots.$$

Define two measurable functions:

$$\phi^* = \sup \{\phi_1, \phi_2, \ldots\} \quad \text{and} \quad \psi^* = \inf \{\psi_1, \psi_2, \ldots\}.$$

Certainly $\phi_n \leq \phi^* \leq f \leq \psi^* \leq \psi_n$ on E for all $n \geq 1$. We want to show $\phi^* = \psi^*$ a.e on E and thus conclude $f = \psi^*$ a.e. Hence, f will be measurable on E. Consider the set

$$\begin{aligned} \{x \in E \mid \psi^*(x) - \phi^*(x) > 0\} &= \bigcup_k \left\{ x \in E \mid \psi^*(x) - \phi^*(x) > \frac{1}{k} \right\} \\ &\subset \bigcup_k \left\{ x \in E \mid \psi_n(x) - \phi_n(x) > \frac{1}{k} \right\} \end{aligned}$$

for all $n \geq 1$. If we show $\left\{ x \in E \mid \psi^*(x) - \phi^*(x) > \frac{1}{k} \right\}$ has measure zero we would be finished. We know this set is measurable because $\psi^* - \phi^*$ is a measurable function.

By construction, $\int_E (\psi_n - \phi_n) < \frac{1}{n}$. But then $\frac{1}{n} > \int_E (\psi_n - \phi_n) > \frac{1}{k} m\left(\left\{ x |\ \psi_n(x) - \phi_n(x) > \frac{1}{k} \right\} \right)$ with $E_1 = \left\{ x |\ \psi_n - \phi_n > \frac{1}{k} \right\}$ and $E_2 = \left\{ x |\ \psi_n - \phi_n \leq \frac{1}{k} \right\}$. Thus

$$m\left(\left\{ x |\ \psi_n(x) - \phi_n(x) > \frac{1}{k} \right\} \right) < \frac{k}{n} \quad \text{for all } n \geq 1.$$

i.e., $m\left(\left\{ x |\ \psi^*(x) - \phi^*(x) > \frac{1}{k} \right\} \right) = 0$ and the proof is completed by recalling that a countable union of sets of measure zero is a measurable set of measure zero. □

Example 6.3.11. If f is a bounded, Lebesgue integrable function on a set E of finite measure, and g is a bounded function on E such that $g = f$ a.e. on E, then g is Lebesgue integrable on E and

$$\int_E g = \int_E f$$

The function f is measurable by the previous theorem and since $f = g$ a.e on E so g is also measurable and hence integrable. Let $A = \{x |\ f(x) \neq g(x)\}$. The set A has measure zero, thus $E - A$ is measurable and, by the third assignment of exercise 6.3.12,

$$\int_E f = \int_{E \setminus A} f + \int_A f = \int_{E \setminus A} g = \int_{E \setminus A} g + \int_A g = \int_E g.$$

Exercise 6.3.12. 1. Find the Lebesgue integral of the following functions in the domain as indicated.

(a) $f(x) = \sqrt{x}$ on $0 \leq x \leq 1$

(b) $f(x) = \mathrm{e}^x$ on $0 \leq x \leq \ln 2$

(c) $f(x)$ is defined on $[0, 2]$ as follows:

$$\begin{aligned} f(x) &= x^{1/3},\ 0 \leq x \leq 1 \\ &= 2,\ \ 1 < x \leq 2. \end{aligned}$$

2. Show that if f and g are measurable on a set of finite measure and k is any real number, then $f + g$ is Lebesgue integrable on E and $\int_E (f + g) = \int_E f + \int_E g$.

3. For the above f and E with $E = E_1 \cup E_2$, show that f is measurable on E_1 and E_2 and

$$\int_E f = \int_{E_1} f + \int_{E_2} f.$$

Sample Questions

1. Show that for two simple functions ϕ and ψ defined on a set E with finite measure, $\int_E (\phi + \psi) = \int_E \phi + \int_E \psi$.

2. When is a bounded function f defined on a measurable set E with finite measure, said to be Lebesgue integrable? Show that such a function is Riemann integrable. Is the converse true? Justify your answer.

3. Deduce a necessary and sufficient condition for a bounded function f defined on a set E with finite measure, to be Lebesgue integrable.

4. Show that a bounded function on a set E of finite measure is Lebesgue integrable if and only if it is measurable on E.

Unit 7

Course Structure

- The integral of non-negative simple functions
- The integral of non-negative measurable functions on arbitrary measurable sets in $\mathbb{R}$ using integrals of non-negative simple functions

---

7.1 Introduction

In the previous unit, we have mainly focused on introducing the idea of Lebesgue measurability of bounded functions on sets of finite measure with the use of simple functions. As we have previously pointed out, Riemann integrals were defined only on closed and bounded intervals and we have also defined the integrability of functions on any arbitrary measurable set of finite measure. However, the question of unbounded sets still remain open. In this unit, we will try to generalise the idea of integrability on any arbitrary measurable set (finite or infinite measure). We first start by defining integral of simple measurable functions defined on a measurable set which assumes only non-negative values. Using this, we will define measurability of any non-negative measurable function defined any measurable set. Finally, the case of an arbitrary measurable function on an arbitrary measurable set is addressed in the next unit.

Objectives

After reading this unit, you will be able to

- define the Lebesgue integral of non-negative simple functions on arbitrary measurable sets;
- define the Lebesgue integral of non-negative measurable functions on arbitrary measurable sets and learn their properties.

7.2 The Lebesgue integral for non-negative measurable functions

Let us see a function and try to integrate it according to the idea of integration we learnt in the previous unit.

Example 7.2.1. Let f be ddfined on the interval $(0, 2)$ as follows:

$$\begin{aligned} f(x) &= \frac{1}{x}; \quad 0 < x \leq 1 \\ &= \frac{1}{x-2}; \quad 1 < x < 2. \end{aligned}$$

Then f is an unbounded function. Let us try to find its integral using the previous definitions. Let

$$\begin{aligned} \int_{(0,2)} f &= \int_{(0,1]} f + \int_{(1,2)} f \\ &= \lim_{\epsilon \to 0} \int_{\epsilon}^{1} \frac{1}{x} dx + \lim_{\epsilon \to 0} \int_{1}^{2-\epsilon} \frac{1}{x-2} dx \\ &= \lim_{\epsilon \to 0} \int_{(\epsilon,1]} \frac{1}{x} dx + \lim_{\epsilon \to 0} \int_{(1,2-\epsilon)} \frac{1}{x-2} dx \\ &= \infty - \infty, \end{aligned}$$

which is not defined. Does it mean that we won't be able to define the integrals of such kind of functions? We will exactly see this now.

In the above example, the function being unbounded, which is both positive and negative, causes problem with the integral interpretations. One thing at a time. Let us try to picturise it in this way: any measurable function can be written as the difference of two nonnegative measurable functions: $f = f^+ - f^-$. As a result of this decomposition, it is natural to define the Lebesgue integral of f, whether f is positive, negative, or both, as

$$\int_E f = \int_E f^+ - \int_E f^-,$$

and this will be meaningful if we have meaning for $\int_E f^+$and $\int_E f^-$in some sense.

Then, we will proceed with the nonnegative measurable functions whose domains need not have finite measure, and then in the next section remove the condition that f be nonnegative.

Example 7.2.2. $f(t) = \begin{cases} t^{-1}, & 0 < t \leq 1 \\ (2-t)^{-1/2}, & 1 \leq t < 2. \end{cases}$

$$\int_{(0,2)} f = \int_{(0,1]} f + \int_{(1,2)} f = \infty + 2 = \infty$$

In the above example, f nonnegative and unbounded. But, how do we define the Lebesgue integral for a nonnegative, not necessarily bounded, measurable function? The idea is inherent in the simple approximation theorem. We can approximate any nonnegative measurable function f with a monotone increasing sequence of nonnegative simple functions; $0 \leq \phi_n \leq \phi_{n+1} \leq \cdots, \lim \phi_n = f$ on E. It is natural to define the Lebesgue integral of f on E, $\int_E f$, by

$$\int_E f = \int_E (\lim \phi_n) = \lim \int_E \phi_n.$$

Roughly speaking, we must make sense of $\int_E \phi_n$, that is, the Lebesgue integral of a nonnegative simple function. We will start with the definition of integral of any nonnegative measurable function.

Definition 7.2.3. Let ϕ be a nonnegative simple function on $\mathbb{R}$, that is,

$$\phi(x) = \sum_{k=1}^{n} c_k \chi_{E_k}(x),$$

where E_k are mutually disjoint Lebesgue measurable subsets of $\mathbb{R}$,

$$\mathbb{R} = \bigcup_{k=1}^{n} E_k,$$

and c_k are nonnegative real numbers (Note: Nothing lost in assuming

$$\bigcup_{k=1}^{n} E_k = \mathbb{R},$$

for if not, then

$$E_0 \equiv \mathbb{R} \setminus \bigcup_{k=1}^{n} E_k$$

is a measurable set and

$$\mathbb{R} = \bigcup_{k=0}^{\pi} E_k, \Bigg)$$

Thus $m(E_k)$ will be infinite for at least one k, $1 \leq k \leq n$. Obviously a simple function ϕ has many representations as linear combinations of characteristic functions.

Definition 7.2.4. The Lebesgue integral of a nonnegative simple function ϕ, on a measurable set E, written $\int_E \phi$, is defined by

$$\int_E \phi = \sum_{1}^{n} c_k m(E \cap E_k),$$

where $\phi = \sum_{1}^{n} c_k \chi_{E_k}$, E_k mutually disjoint, $R = \bigcup_{1}^{n} E_k, c_k \geq 0$.

By convention, we define $\int_E \phi = 0$ whenever $\phi = 0$ on E, even if $m(E) = 0$. Also, the definition of nonnegative simple function on E, with $m(E) < \infty$ agrees with the previous definition. It is to be noted that despite more than one representations of a non-negative simple function ϕ, the Lebesgue integral is well-defined. Suppose

$$\phi = \sum_{1}^{n} c_k \chi_{E_k}, \;\; c_k \geq 0$$

and

$$\phi = \sum_{1}^{m} d_j \chi_{D_j}, \;\; d_j \geq 0$$

with

$$\mathbb{R} = \bigcup_{1}^{n} E_k = \bigcup_{1}^{m} D_j$$

E_k and D_j mutually disjoint measurable subsets of $\mathbb{R}$. We show that

$$\sum_1^n c_k m\left(E \cap E_k\right) = \sum_1^m d_j m\left(E \cap D_j\right),$$

that is, the integral as we have defined it is independent of the representation of ϕ. Note that

$$E_k = E_k \cap \left(\bigcup_1^m F_j\right) = \bigcup_1^m (E_k \cap F_j)$$

and

$$F_j = F_j \cap \left(\bigcup_1^n E_k\right) = \bigcup_1^n \left(E_k \cap F_j\right).$$

Thus,

$$\begin{aligned}
\sum_1^n c_k m\left(E \cap E_k\right) &= \sum_1^n c_k m\left(E \cap \left(\bigcup_1^m (E_k \cap F_j)\right)\right) \\
&= \sum_1^n c_k m\left(\bigcup_1^m (E \cap E_k \cap F_j)\right) \\
&= \sum_1^n c_k \sum_1^m m\left(E \cap E_k \cap F_j\right) \\
&= \sum_1^n \sum_1^m d_j m\left(E \cap E_k \cap F_j\right) \\
&= \sum_1^m \sum_1^n d_j m\left(E \cap E_k \cap F_j\right) \\
&= \sum_1^m d_j \sum_1^n m\left(E \cap E_k \cap F_j\right) \\
&= \sum_1^m d_j m\left(E \cap \left(\bigcup_1^n (E_k \cap F_j)\right)\right) \\
&= \sum_1^m d_j m\left(E \cap F_j\right)
\end{aligned}$$

since, for $E_k \cap E \cap F_j \neq \emptyset$, $c_k = c_k \chi_{E_k} = \phi = d_j \chi_{F_j} = d_j$, and if $E_k \cap E \cap F_j = \emptyset$, no contribution due to $m(\emptyset) = 0$. Finally, the Lebesgue integral of a nonnegative simple function is a nonnegative real number or ∞.

Theorem 7.2.5. If ϕ, ψ are nonnegative simple functions on $\mathbb{R}$, if E is any measurable subset of $\mathbb{R}$, and k is any nonnegative real number, then

1. $k\phi$ is a nonnegative simple function on E, and

$$\int_E (k\phi) = k \int_E \phi \quad \text{(homogeneous);}$$

2. $(\phi + \psi)$ is a nonnegative simple function on E, and

$$\int_E (\phi + \psi) = \int_E \phi + \int_E \psi \quad \text{(additive);}$$

3.

$$\int_E \phi \leq \int_E \psi \text{ if } 0 \leq \phi \leq \psi \text{ on } E \quad \text{(monotone);}$$

4. If E_1 and E_2 are disjoint measurable subsets of E with $E = E_1 \cup E_2$, the integrals $\int_{E_1} \psi$ and $\int_{E_2} \psi$ exist, and

$$\int_E \psi = \int_{E_1} \psi + \int_{E_2} \psi \quad \text{(additive on the domain).}$$

Proof. 1. Suppose $\phi = \sum_1^n c_i \chi_{E_i}$. Then $k\phi = \sum kc_i \chi_{E_i}$ and

$$\int_E (k\phi) = \sum_1^n (kc_i)\, m(E \cap E_i) = k \sum_1^n c_i m(E \cap E_i) = k \int_E \phi.$$

2. Let $\phi = \sum_1^n c_k \chi_{E_k}$ and $\psi = \sum_1^m d_j \chi_{F_j}$, $0 \leq c_k, d_j$. The idea is to form the $n \cdot m$ sets;

$$\begin{aligned} &E_1 \cap F_1, E_1 \cap F_2, \ldots, E_1 \cap F_m \\ &E_2 \cap F_1, E_2 \cap F_2, \ldots, E_2 \cap F_m \\ &\vdots \\ &E_n \cap F_1, E_n \cap F_2, \ldots, E_n \cap F_m. \end{aligned}$$

If $E_k \cap F_j \neq \emptyset$, define $\phi + \psi$ as $c_k + d_j$. The nonempty $E_k \cap F_j$ are mutually disjoint measurable subsets of $\mathbb{R}$,

$$\mathbb{R} = \bigcup_{k,j} (E_k \cap F_j),$$

and

$$\phi + \psi = \sum_{k,j} (c_k + d_j)\, \chi_{E_k \cap F_j}.$$

Thus

$$\begin{aligned} \int_E (\phi + \psi) &= \sum_{k,j} (c_k + d_j)\, m(E_k \cap F_j \cap E) \\ &= \sum_{k=1}^n \sum_{j=1}^m (c_k + d_j)\, m(E_k \cap F_j \cap E) \\ &= \sum_{k=1}^n c_k \sum_{j=1}^m m(E_k \cap F_j \cap E) + \sum_{j=1}^m d_j \sum_{k=1}^n m(E_k \cap F_j \cap E) \\ &= \sum_{k=1}^n c_k m(E_k \cap E) + \sum_{j=1}^m d_j m(F_j \cap E) \\ &= \int_E \phi + \int_E \psi. \end{aligned}$$

3. Suppose $\phi = \sum_{k=1}^{n} c_k \chi_{E_k}$, E_k mutually disjoint, and $\psi = \sum_{j=1}^{m} d_j \chi_{F_j}$, F_j mutually disjoint, where

$$\bigcup_{1}^{n} E_k = \mathbb{R} = \bigcup_{1}^{m} F_j.$$

Since $0 \leq \phi \leq \psi$, $0 \leq c_k \leq d_j$ on nonempty $E_k \cap F_j$ and thus

$$\begin{aligned}\int_E \phi &= \sum_{k=1}^{n} c_k m\left(E_k \cap E\right) = \sum_{k=1}^{n} c_k \sum_{j=1}^{m} m\left(E_k \cap F_j \cap E\right) \\ &\leq \sum_{j=1}^{m} d_j \sum_{k=1}^{n} m\left(E_k \cap F_j \cap E\right) = \sum_{j=1}^{m} d_j m\left(F_j \cap E\right) \\ &= \int_E \psi.\end{aligned}$$

4.

$$\begin{aligned}\int_E \psi &= \sum d_j m\left(F_j \cap E\right) = \sum d_j m\left(F_j \cap \left(E_1 \cup E_2\right)\right) \\ &= \sum d_j \left[m\left(F_j \cap E_1\right) + m\left(F_j \cap E_2\right)\right] \\ &= \sum d_j m\left(F_j \cap E_1\right) + \sum d_j m\left(F_j \cap E_2\right) \\ &= \int_{E_1} \psi + \int_{E_2} \psi.\end{aligned}$$

□

We will now define the Lebesgue integral of a nonnegative measurable function.

Definition 7.2.6. Definition A: If f is a nonnegative, measurable function, defined on a measurable set E, the Lebesgue integral of f over E, $\int_E f$, is given by

$$\int_E f \equiv \sup\left\{\int_E \phi |\ \phi \leq f, \quad \phi \text{ is nonnegative and simple }\right\}.$$

The definition can also be given in terms of the simple approximation theorem as follows.

Definition 7.2.7. Definition B: If f is a nonnegative, measurable function, defined on a Lebesgue measurable set E, and ϕ_n is a nonnegative monotone sequence of simple functions, $0 \leq \phi_n \leq \phi_{n+1}$ on E, with

$$\lim \phi_n(x) = f(x) \quad \text{(finite or infinite)}$$

for all $x \in E$, the Lebesgue integral of f over E, $\int_E f$, is given by

$$\int_E f \equiv \lim \int_E \phi_n = \int_E \left(\lim \phi_n\right).$$

Some comments are in order before we show the equivalence of these definitions, hereafter referred to as A and B.

We need to keep few things in mind.

1. Suppose f is nonnegative, bounded and measurable on a set E of finite measure. Then definition A agrees with the previous definition of integral. By our previous discussions, we can say that $\int_E f = \bar{\int}_E f$. But then,

$$\begin{aligned}\underline{\int}_E f &= \sup\left\{\int_E \phi |\ \phi \le f, \quad \phi \text{ simple } \right\} \\ &= \sup\left\{\int_E \phi |\ \phi \le f, \quad \phi \text{ simple and nonnegative } \right\},\end{aligned}$$

since f is nonnegative, and this is definition A.

2. Because f is nonnegative, we always have simple functions below $f (\phi = 0)$. Thus the set $\left\{\int_E \phi \mid \phi \le f\right\}$ is nonempty and the "sup" is a nonnegative member of the extended reals.

3. If f is nonnegative and simple, say $f = \hat{\phi}$, then $\int_E \hat{\phi} \in \left\{\int_E \phi \mid \phi \le \hat{\phi}\right\}$ and $\int_E \phi \le \int_E \hat{\phi}$; $\int_E \hat{\phi}$ is a member of and an upper bound of the set $\left\{\int_E \phi \mid \phi \le f\right\}$. Thus $\int_E \hat{\phi} = \sup\left\{\int_E \phi \mid \phi \le \bar{\phi}\right\}$. Definition 5.3.2 and A are in agreement.

4. By the simple approximation theorem, we have a monotone sequence $\left\{\hat{\phi}_m\right\}$ of nonnegative simple functions, $0 \le \hat{\phi}_m \le \hat{\phi}_{m+1}$ on E with

$$\lim \hat{\phi}_m = f \quad \text{ on } E.$$

The sequence $\left\{\int_E \hat{\phi}_m\right\}$ is a nondecreasing sequence in the extended reals, so the limit is defined in the extended reals: $\lim \int_E \hat{\phi}_m$ is a nonnegative real number or ∞.

Theorem 7.2.8. The definitions A and B of the Lebesgue integral of a nonnegative measurable functions are equivalent.

Proof. Left as exercise. □

We have the familiar properties of the integral.

Theorem 7.2.9. If f and g are nonnegative measurable functions, defined on a measurable set E, and k is any nonnegative real number, then

1. (kf) is nonnegative, measurable, and $\int_E (kf) = k \int_E f$ (homogeneous);
2. $(f + g)$ is nonnegative, measurable, and $\int_E (f + g) = \int_E f + \int_E g$ (additive);
3. $\int_E f \le \int_E g$ if $0 \le f \le g$ (monotone);
4. If E_1 and E_2 are disjoint measurable subsets of E with $E = E_1 \cup E_2$, the integrals $\int_{E_1} f$ and $\int_{E_2} f$ exist in the extended reals, and

$$\int_E f = \int_{E_1} f + \int_{E_2} f \quad \text{(additive on domain).}$$

Proof. Measurability of the appropriate functions follows from the preceding units.

1. From the simple approximation theorem, we have a sequence $\left\{\hat{\phi}_n\right\}$ of simple functions satisfying $0 \leq \hat{\phi}_n \leq \hat{\phi}_{n+1}$ with $\lim \hat{\phi}_n = f$. But then, $0 \leq k\hat{\phi}_n \leq k\hat{\phi}_{n+1}$ and $\lim_n \left(k\hat{\phi}_n\right) = (kf)$. Using Definition B,
$$k\int_E f = k\lim \int_E \hat{\phi}_n = \lim \int_E k\hat{\phi}_n = \int_E kf.$$
2. $\lim \phi_n = f, \quad \lim \psi_n = g$ implies $\lim (\phi_n + \psi_n) = f+g$. Thus, $\lim \int_E \phi_n = \int_E f, \quad \lim \int_E \psi_n = \int_E g$ implies $\lim \int_E (\phi_n + \psi_n) = \int_E (f+g)$, etc.
3. If $0 \leq \phi \leq f$, then $\phi \leq$ g. Thus $\{\phi \mid \phi \leq f\} \subset \{\phi \mid \phi \leq g\}$. Hence
$$\sup\left\{\int_E \phi \mid \phi \leq f\right\} \leq \sup\left\{\int_E \phi \mid \phi \leq g\right\},$$
that is, $\int_E f \leq \int_E g$.
4. $\int_E \phi_n = \int_{E_1} \phi_n + \int_{E_2} \phi_n$. The sequences $\left\{\int_E \phi_n\right\}$, $\left\{\int_{E_1} \phi_n\right\}$, and $\left\{\int_{E_2} \phi_n\right\}$ are monotonically increasing, limits are defined and nonnegative, possibly in the extended reals. The result follows by taking limits.

   This completes the proof.

□

Exercise 7.2.10. 1. For $f(x) = \dfrac{1}{x^2}$, calculate $\displaystyle\int_{[1,\infty)} f(x)$.

2. Calculate $\displaystyle\int_{(0,1]} \frac{1}{\sqrt{x}}$.

Sample Questions

1. Show that the Lebesgue integral of a nonnegative simple function is independent of its representation.
2. For two nonnegative simple functions ϕ and ψ defined on $\mathbb{R}$, show that
$$\int_E (\phi + \psi) = \int_E \phi + \int_E \psi$$
where E is a measurable subset of $\mathbb{R}$.
3. Show that for two disjoint subsets E_1 and E_2 of a measurable set E, if f is a nonnegative measurable function defined on E such that $\int_{E_1} f$ and $\int_{E_2} f$ exists in the extended reals, then
$$\int_E f = \int_{E_1} f + \int_{E_2} f.$$

4. Prove or disprove: If f and g are two nonnegative measurable functions defined on a measurable set E and $f = g$ a.e. on E, then $\int_E f = \int_E g$.

Unit 8

Course Structure

- Monotone convergence theorem and Fatou's lemma.
- The integral of Measurable functions and basic properties, Absolute character of the integral
- Dominated convergence theorem

8.1 Introduction

Given a sequence of functions $\{f_n\}$ that converges pointwise to some limit function f, it is not always true that

$$\int \lim_{n\to\infty} f_n = \lim_{n\to\infty} \int f_n.$$

The Monotone Convergence Theorem (MCT), the Dominated Convergence Theorem (DCT), and Fatou's Lemma are three major results in the theory of Lebesgue integration that answer the question, "When does the integral and limit operator commute?" Fatou's Lemma is somewhat idea in this direction. However, the MCT and DCT tell us that if certain restrictions are imposed on both the f_n and f then the interchange of the limit and integral is indeed possible.

Objectives

After reading this unit, you will be able to

- define the Lebesgue integral of arbitrary measurable functions on arbitrary measurable sets from the idea of the preceding integral and learn their properties;
- state and apply monotone convergence theorem and Fatou's lemma;
- state and apply the Lebesgue dominated convergence theorem.

8.2 Fatou's Lemma and Lebesgue Monotone converegnce theorem

Theorem 8.2.1. (Fatou's Lemma) Let $\{f_n\}$ be a sequence of non-negative measurable functions. Then,

$$\liminf \int f_n dx \geq \int \liminf f_n dx$$

Proof. Let $f = \liminf f_n$. Then f is a non-negative measurable function. From the definition of integration of a non-negative measurable function, the result follows for each simple measurable function ϕ with $\phi \leq f$ we have

$$\int \phi dx \leq \liminf \int f_n dx$$

Case I When $\int \phi dx = \infty$. Then from the definition of the integration of simple measurable function, for some measurable set A, we have $m(A) = \infty$ and $\phi > a > 0$ on A. We write $g_k(x) = \inf_{j \geq k} f_j(x)$, and $A_n = \{x : g_k(x) > a,\ \forall k \geq n\}$, a measurable set. Then $A_n \subseteq A_{n+1}$, each n. But, for each x, $\{g_k(x)\}$ is monotonically increasing and

$$\lim_{k \to \infty} g_k(x) = f(x) \geq \phi(x)$$

So, $A \subseteq \bigcup_{n=1}^{\infty} A_n$. Hence, $\lim(A_n) = \infty$. But for each n,

$$\int f_n dx \geq \int g_n dx > am(A_n)$$

So, $\liminf \int f_n dx = \infty$ and the result is true.

Case II When $\int \phi dx < \infty$. Write $B = \{x : \phi(x) > 0\}$. Then $m(B) < \infty$. Let M be the largest value of ϕ, and if $0 < \epsilon < 1$, write $B_n = \{x : g_k(x) > (1-\epsilon)\phi(x), k \geq n\}$, where g_k is as defined above. Then the sets B_n are measurable, $B_n \subseteq B_{n+1}$ for each n, and $\bigcup_{n=1}^{\infty}(B \setminus B_n) = \emptyset$. As $m(B) < \infty$, so there exists N such that $m(B \setminus B_n) < \epsilon$ for all $n \geq N$. So, if $n \geq N$

$$\begin{aligned}
\int g_n dx &\geq \int_{B_n} g_n dx \\
&\geq (1-\epsilon) \int_{B_n} \phi dx \\
&= (1-\epsilon) \left(\int_B \phi dx - \int_{B \setminus B_n} \phi dx \right) \\
&\geq (1-\epsilon) \int \phi dx - \int_{B \setminus B_n} \phi dx \\
&\geq \int \phi dx - \epsilon \int \phi dx - \epsilon M
\end{aligned}$$

Since ϵ is arbitrary, $\liminf \int g_n dx \geq \int \phi dx$, and since $f_n \geq g_n$, we get the desired result.

□

Example 8.2.2. 1. Let $f_n = \frac{1}{n}\chi_{[0,n]}$. Then

$$\int_{[0,\infty)} f_n = 1 \neq 0 = \int_{[0,\infty)} \lim f_n.$$

But $\int_{[0,\infty)} \liminf f_n = \int_{[0,\infty)} \lim f_n = 0 \leq 1 = \liminf \int_{[0,\infty)} f_n.$

2. Nonnegativity is necessary. Let us see with this example of $f_n = -\frac{1}{n}\chi_{[0,n]}$. Then $f_n \to 0$ (uniformly) on $[0, \infty)$.

$$\lim \int_{[0,\infty)} f_n = -1 \neq \int_{[0,\infty)} \lim f_n = 0,$$

but

$$\int_E \liminf f_n = 0 > -1 = \liminf \int_E f_n$$

for any measurable set E.

3. The inequality may be strict. Let $f_n = \chi_{[n,n+1]}$. Then

$$\int_E \liminf f_n = 0 < 1 = \liminf \int_E f_n.$$

for any measurable set E.

Theorem 8.2.3. (Lebesgue Monotone Convergence Theorem - LMCT) If $\{f_n\}$ is a monotonically increasing sequence of non-negative measurable functions converging to a function f, then

$$\int f dx = \lim \int f_n dx.$$

Proof. We have by Fatou's Lemma,

$$\int f dx = \int \liminf f_n dx \leq \liminf \int f_n dx$$

By hypothesis, we have, $f \geq f_n$. So, we have by the property of integration of non-negative measurable functions,

$$\int f dx \geq \int f_n dx$$

and hence,

$$\int f dx \geq \limsup \int f_n dx$$

Combining, we have the desired result. □

Theorem 8.2.4. Let f and g be two non-negative measurable functions. Then

$$\int (f + g) dx = \int f dx + \int g dx$$

Proof. We first consider the theorem for simple measurable functions ϕ and ψ. Let the values of ϕ be $a_1, a_2, \ldots, a_n$ on the sets $A_1, A_2, \ldots, A_n$ and let the values of ψ be $b_1, b_2, \ldots, n_m$ on the sets $B_1, B_2, \ldots, B_m$. Then the simple function $\phi + \psi$ has the value $a_j + b_j$ on the measurable set $A_i \cap B_j$. So, from the properties of the integral of simple measurable functions we can say that

$$\int_{A_i \cap B_j} (\phi + \psi) dx = \int_{A_i \cap B_j} \phi dx + \int_{A_i \cap B_j} \psi dx$$

But the union of the disjoint sets $A_i \cap B_j$ is $\mathbb{R}$. So we have by the properties of integration of simple measurable functions,

$$\int (\phi + \psi) dx = \int \phi dx + \int \psi dx$$

Let f and g be any non-negative measurable functions. Let $\{\phi_n\}$, $\{\psi_n\}$ be sequences of measurable simple functions ϕ_n converging to f and ψ_n converging to g. Then $\phi_n + \psi_n$ converges to $f + g$. But, by the properties of the integral of simple measurable functions, we have

$$\int (\phi_n + \psi_n) dx = \int \phi_n dx + \int \psi_n dx$$

Letting the limit as $n \to \infty$, we get the desired result. □

Theorem 8.2.5. Let $\{f_n\}$ be a sequence of non-negative measurable functions. Then

$$\int \sum_{n=1}^{\infty} f_n dx = \sum_{n=1}^{\infty} \int f_n dx$$

Proof. By induction, we can show that the previous theorem applies to the sum of n functions. So, if

$$S_n = \sum_{i=1}^{n} f_i$$

then

$$\int S_n dx = \sum_{i=1}^{n} \int f_i dx.$$

But, $\{S_n\}$ is a monotonically increasing sequence of functions converging to $f = \sum_{i=1}^{\infty} f_i$. So, the result follows from Monotone Convergence Theorem. □

Exercise 8.2.6. 1. Show that if f is a non-negative measurable function, then

$$f = 0 \text{ a.e iff } \int f dx = 0.$$

2. Calculate the following integrals.

   (a) $\int_{(0,1]} t^{-1/2}$.

   (b) $\int_{[0,1)} (1 - t^2)^{-1/2}$.

   (c) $\int_{(0,2)} f(t)$ if $f(t)$ is given by

$$\begin{aligned} f(t) &= t^{-1} \;\; 0 < t \leq 1 \\ &= (2-t)^{-1/2}, \;\; 1 \leq t \leq 2. \end{aligned}$$

   (d) $\int_{[0,\infty)} \mathrm{e}^{-t}$.

8.3 Lebesgue Integral and Lebesgue Integrability

Recall that for a measurable function f defined on a measurable set E,

$$f^+(x) = \max\{f(x), 0\}, \quad f^-(x) = \max\{-f(x), 0\},$$

are the positive and negative parts of f, respectively. Also, both are nonnegative measurable functions and thus, $\int_E f^+$ and $\int_E f^-$ can be calculated using the definitions of integration of nonnegative measurable function in the previous unit. Further, we have

$$f = f^+ - f^-.$$

We are now in a position to define the integrability of a general measurable function on a measurable set E.

Definition 8.3.1. If f is a measurable function defined on a measurable set E and $\int_E f^+ < \infty$ and $\int_E f^- < \infty$. Then we say that f is integrable, and its integral is given by

$$\int f dx = \int f^+ dx - \int f^- dx.$$

It should be noted here that if one of the two integrals $\int_E f^+$ and $\int_E f^-$ is finite, then also the integral of f is not defined in $\mathbb{R}$. In case both the integrals are infinite, $\int_E f^+ - \int_E f^-$ would yield $\infty - \infty$, which is not defined.

This new definition is consistent with the definition of the integral of nonnegative function. Indeed, if f is nonnegative, then $f^+ = f$ and $f^- = 0$. Thus $\int_E f^- = 0$ and hence the integral of f exists and is equal to that of f^+.

Example 8.3.2. 1. Let f be defined as

$$\begin{aligned} f(x) &= \frac{1}{x}, \quad 0 < x \leq 1 \\ &= \frac{1}{x-2}, \quad 1 < x < 2. \end{aligned}$$

Then

$$\int_{(0,2)} f^+ = \int_{(0,1]} \frac{1}{x} = \infty \quad \text{and} \quad \int_{(0,2)} f^- = \int_{(1,2)} -\frac{1}{x-2} = \infty.$$

Hence the Lebesgue integral of f is not defined on $(0, 2)$.

2. Let f be defined as

$$\begin{aligned} f(x) &= x^{-1/2}, \quad 0 < x \leq 1 \\ &= \frac{1}{x-2}, \quad 1 < x < 2. \end{aligned}$$

Then

$$\int_{(0,2)} f^+ = \int_{(0,1]} x^{-1/2} = 2 \quad \text{and} \quad \int_{(0,2)} f^- = \int_{(1,2)} -\frac{1}{x-2} = \infty.$$

Since one of the integrals is infinite, the integral of f is not defined in $\mathbb{R}$. Let us now check the next example.

3.

$$\begin{aligned} f(x) &= x^{-1/2}, \quad 0 < x \leq 1 \\ &= -(2-x)^{-1/2}, \quad 1 < x < 2. \end{aligned}$$

Then

$$\int_{(0,2)} f^+ = \int_{(0,1]} x^{-1/2} = 2 \quad \text{and} \quad \int_{(0,2)} f^- = \int_{(1,2)} (2-x)^{-1/2} = 2.$$

The Lebesgue integral of f is defined and $\int_{(0,2)} f = 2 - 2 = 0$, and hence f is Lebesgue integrable on $(0, 2)$.

Henceforth all Lebesgue integrable functions will be called integrable for simplicity.

Theorem 8.3.3. Suppose f is a measurable function defined on a measurable set E. Then f is integrable on E if and only if $|f|$ is integrable on E. Furthermore,

$$\left| \int_E f \right| \leq \int_E |f|.$$

Proof. Assume f is integrable on E. We want to show that $|f|$ is measurable and $\int_E |f|^+$, $\int_E |f|^- < \infty$. But since f is measurable, $|f|$ is measurable, $\int_E |f|^- = 0$, and $\int_E |f|^+ = \int_E |f| = \int_E (f^+ + f^-) = \int_E f^+ + \int_E f^- < \infty$ because f is integrable on E. Thus $|f|$ is integrable on E.

Assume $|f|$ is integrable on E. We show f is integrable on E. Since f is measurable by hypothesis, and $\int_E f^+ + \int_E f^- = \int_E (f^+ + f^-) = \int_E |f| = \int_E |f|^+ < \infty$, the nonnegative integrals $\int_E f^+, \int_E f^-$ are both finite. Consequently, f is integrable on E.

$$\begin{aligned} \left| \int_E f \right| = \left| \int_E f^+ - \int_E f^- \right| &\leq \left| \int_E f^+ \right| + \left| \int_E f^- \right| \\ &= \int_E f^+ + \int_E f^- = \int_E (f^+ + f^-) = \int_E |f|. \end{aligned}$$

□

Theorem 8.3.4. If f is a measurable function defined on a measurable set E, and g is integrable on E with $|f| \leq |g|$, then $\int_E |f| \leq \int_E |g|$ and f is integrable on E.

Proof. We have $\int_E |f| \leq \int_E |g| < \infty$. To show that f is integrable on E requires f to be measurable on E (which is already given) and $\int_E f^+$ and $\int_E f^-$ are both finite.

But $0 \leq \int_E f^+ + \int_E f^- = \int_E |f| < \infty$, and the argument is complete. □

The next result shows that measure zero sets do not affect integrability.

Theorem 8.3.5. If $f = g$ a.e. on a measurable set E, and if g is integrable on E, then f is integrable on E and

$$\int_E f = \int_E g.$$

Proof. The function g is measurable on E by the assumption of being Lebesgue integrable on E. Since f is equal almost everywhere to a measurable function g, f is measurable on E. Hence f^+ and f^- are measurable on E. Let $A = \{x \in E \mid f(x) \neq g(x)\}$. Then $f^+ = g^+$ and $f^- = g^-$ on $E \setminus A$, and $\int_{E\setminus A} f^+ = \int_{E\setminus A} g^+$and $\int_{E\setminus A} f^- = \int_{E\setminus A} g^-$, that is, f is measurable on $E \setminus A$, $\int_{E\setminus A} f^+$, $\int_{E\setminus A} f^- < \infty : f$ is integrable on $E \setminus A$. Because A is a measurable subset of E, f is measurable on A, $m(A) = 0$, and hence $\int_A f^+$, $\int_A f^- = 0$. But then $\int_E f^+ = \int_{E\setminus A} f^+ + \int_A f^+ < \infty$ and $\int_E f^- = \int_{E\setminus A} f^- + \int_A f^- < \infty :$ The function f is integrable on E. Then

$$\begin{aligned}\int_E g &= \int_E g^+ - \int_E g^- = \int_{E-A} g^+ + \int_A g^+ - \int_{E-A} g^- - \int_A g^- \\ &= \int_{E-A} f^+ + \int_A f^+ - \int_{E-A} f^- - \int_A f^- = \int_E f^+ - \int_E f^- \\ &= \int_E f.\end{aligned}$$

□

Theorem 8.3.6. If f, g are integrable on a measurable set E, and k is any real number, then

1. (kf) is integrable on E, and $\int_E (kf) = k \int_E f$ (homogeneous);
2. $(f + g)$ is integrable on E, and $\int_E (f + g) = \int_E f + \int_E g$ (additive);
3. $\int_E f \leq \int_E g$ if $f \leq g$ on E (monotone);
4. If E_1 and E_2 are disjoint measurable subsets of E with $E = E_1 \cup E_2$, f is integrable on E_1 and E_2, and $\int_E f = \int_{E_1} f + \int_{E_2} f$ (additive on the domain).

Proof. 1. If $k \geq 0$, $\int_E (kf)^+ = \int_E kf^+ = k \int_E f^+ < \infty$ and $\int_E (kf)^- = k \int_E f^- < \infty$ because $kf^+, kf^-$are nonnegative measurable functions. By definition, (kf) is integrable on E. Furthermore, $\int_E (kf) = \int_E (kf)^+ - \int_E (kf)^- = k \int_E f^+ - k \int_E f^- = k \int_E f$, where the last equality is the definition of f being integrable on E. If $k < 0, (kf)^+ = (-k)f^-, (kf)^- = (-k)f^+, \int_E (kf)^+ = -k \int_E f^- < \infty$, and $\int_E (kf)^- = -k \int_E f^+ < \infty$, that is, (kf) is integrable on E. Again,

$$\begin{aligned}\int_E (kf) &= \int_E (kf)^+ - \int_E (kf)^- = \int_E (-k)f^- - \int_E (-k)f^+ \\ &= k \left[\int_E f^+ - \int_E f^- \right] = k \int_E f.\end{aligned}$$

2. Since f, g are integrable on E, $|f|$, $|g|$ are integrable on E. Because $\int_E |f| = \int_E |f|^+ < \infty$, $\int_E |g| = \int_E |g|^+ < \infty$, and $|f+g| \leq |f| + |g|$, $\int_E |f+g| \leq \int_E (|f| + |g|) \leq \int_E |f| + \int_E |g| < \infty$. But $|f+g|^+ = |f+g|$ and $|f+g|^- = 0$, $\int_E |f+g|^+ < \infty$, that is, $|f+g|$ is integrable on E, but then $f+g$ is integrable on E.

   Now, $f + g = (f^+ + g^+) - (f^- + g^-)$, that is, the integrable function $(f+g)$ has been written as the difference of two nonnegative measurable functions, $(f^+ + g^+)$and $(f^- + g^-)$, whose integrals are finite. Comment 5.4.2 reveals

$$\begin{aligned}\int_E (f+g) &= \int_E \left(f^+ + g^+\right) - \int_E \left(f^- + g^-\right) \\ &= \int_E f^+ + \int_E g^+ - \int_E f^- - \int_E g^- \\ &= \int_E f + \int_E g.\end{aligned}$$

3. Since $f \leq g$ on E, $f^+ - f^- \leq g^+ - g^-$, i.e., $f^+ + g^- \leq g^+ + f^-$. Because $\left(f^+ + g^-\right), \left(g^+ + f^-\right)$are nonnegative measurable functions we may apply Proposition 5.7 to conclude

$$\int_E f^+ + \int_E g^- = \int_E \left(f^+ + g^-\right) \leq \int_E \left(g^+ + f^-\right) = \int_E g^+ + \int_E f^-$$

   Because all terms are finite, we may subtract: $\int_E f \leq \int_E g$.

4.

$$\begin{aligned}\int_E f &= \int_E f^+ - \int_E f^- \\ &= \int_{E_1} f^+ + \int_{E_2} f^+ - \int_{E_1} f^- - \int_{E_2} f^- \\ &= \int_{E_1} f + \int_{E_2} f.\end{aligned}$$

□

Theorem 8.3.7. (**Lebesgue Dominated Convergence Theorem**) Let $\{f_n\}$ be a sequence of measurable functions such that $|f_n| \leq g$, where g is integrable, and let $\lim f_n = f$ a.e. Then f is integrable and

$$\lim \int f_n dx = \int f dx.$$

Proof. Since for each n, $|f_n| \leq g$, we have $|f| \leq g$ a.e and so f_n and f are integrable. Also, $\{g + f_n\}$ is a sequence of non-negative measurable functions, so by Fatou's Lemma

$$\liminf \int (g + f_n) dx \geq \int \liminf (g + f_n) dx.$$

So, $\int g dx + \liminf \int f_n dx \geq \int g dx + \int f dx$. But, $\int g dx$ is finite so

$$\liminf \int f_n dx \geq \int f dx.$$

Again, $\{g - f_n\}$ is also a sequence of non-negative measurable functions, so

$$\liminf \int (g - f_n)dx \leq \int \liminf(g - f_n)dx.$$

So, $\int gdx - \limsup \int f_n dx \geq \int gdx - \int fdx$. So, $\limsup \int f_n dx \leq \int fdx \leq \liminf \int f_n dx$. So, the result follows. □

Example 8.3.8. With the same hypothesis as the above theorem, we have $\lim \int |f_n - f|dx = 0$. In fact, $|f_n - f| \leq 2g$, for each n. So applying the Lebesgue Dominated Convergence theorem, we get the desired result.

Theorem 8.3.9. Let $\{f_n\}$ be a sequence of integrable functions such that

$$\sum_{n=1}^{\infty} \int |f_n|dx < \infty.$$

Then the series $\sum_{n=1}^{\infty} f_n(x)$ converges a.e., its sum $f(x)$ is integrable and

$$\int fdx = \sum_{n=1}^{\infty} \int f_n dx.$$

Proof. Let $\phi(x) = \sum_{i=1}^{\infty} |f_n|$. Then by the given condition, $\int \phi dx < \infty$, so ϕ is finite-valued a.e. □

Exercise 8.3.10.

1. Show that $\lim \int_{[0,1]} \frac{kx}{1 + k^2x^2} = 0$, where k is any nonnegative integer.
2. Show that if f and g are measurable , $|f| \leq |g|$ a.e., and g is integrable , then f is integrable.
3. Show that if f is an integrable function, then $|\int fdx| \leq \int |f|dx$.
4. Show that if f is integrable , then f is finite-valued a.e .

Sample Questions

1. State and prove LMCT.
2. State and prove Fatou's lemma. Is the statement of Fatou's lemma valid for any arbitrary sequence of functions? Justify your answer.
3. If f is a measurable function defined on a measurable set E, and g is integrable on E with $|f| < |g|$, then show that f is integrable on E.
4. Let f and g are equal a.e on e measurable set E. If g is integrable on E, show that f is also so.
5. If $f,\ g$ are integrable on a measurable set E, and $f < g$ on E, show that $\int_E f < \int_E g$.
6. State and prove Lebesgue dominated convergence theorem.

Unit 9

Course Structure

- Lebesgue integrability of the derivative of a function of bounded variation on an interval.
- Descriptive characterization of the Lebesgue integral on intervals by absolutely continuous functions.
- Riesz-Fischer theorem on the completeness of the space of Lebesgue integrable functions.

9.1 Introduction

In Riemann theory of integration it is known that differentiation and integration are inverse operations of each other in the following sense:

1. If f is a Riemann integrable function over $[a, b]$, then its indefinite integral

$$F(x) = \mathscr{R}\int_a^x f(t)dt$$

defines a continuous function on $[a, b]$. Furthermore, if f is continuous at a point $x_0 \in [a, b]$, then F is differentiable there at x_0, and $F'(x_0) = f(x_0)$.

2. If f is Riemann integrable over $[a, b]$ and if there is a differentiable function F on $[a, b]$ such that $F'(x) = f(x)$ for $x \in [a, b]$, then

$$\mathscr{R}\int_a^x f(t)dt = F(x) - F(a), \quad \forall x \in [a, b]$$

(This result is usually referred to as the Fundamental Theorem of Calculus.)

The present unit deals with similar types of interrelation between differentiation and Lebesgue integration.

Objectives

After reading this unit, you will be able to

- know the relationship between the differentiation and Lebesgue integration
- the relationship between absolute continuity and integration

9.2 Differentiation of an integral

If f is an integrable function on an interval $[a, b]$, then f is integrable on any subinterval $[a, x] \subset [a, b]$. The function F given by

$$F(x) = \int_a^x f(t)dt + c$$

, where c is a constant, is called the indefinite integral of f.

Theorem 9.2.1. Let f be an integrable function on $[a, b]$. Then the indefinite integral of f is a continuous function of bounded variation on $[a, b]$.

Proof. Let x_0 be any point in $[a, b]$. Then

$$\begin{aligned} |F(x) - F(x_0)| &= \left| \int_{x_0}^x f(t)dt \right| \\ &\leq \left| \int_{x_0}^x |f(t)|dt \right|. \end{aligned}$$

But f being integrable, the function $|f|$ is integrable over $[a, b]$. Therefore, given $\epsilon > 0$, there is a $\delta > 0$ such that for every measurable set $A \subset [a, b]$ with $m(A) < \delta$, we have

$$\int_A |f| < \epsilon.$$

In particular,

$$\left| \int_{x_0}^x |f(t)|dt \right| < \epsilon, \text{ for } |x - x_0| < \delta.$$

Hence $|F(x) - F(x_0)| < \epsilon$, whenever $|x - x_0| < \delta$. This proves the continuity of F at x_0, and hence in $[a, b]$.

In order to establish that F is a function of bounded variation, let

$$P = \{a = x_0 < x_1 < x_2 \ldots < x_n = b\}$$

be a partition of $[a, b]$. Then

$$\begin{aligned} \sum_{i=1}^n |F(x_i) - F(x_{i-1})| &= \sum_{i=1}^n \left| \int_{x_{i-1}}^{x_i} f(t)dt \right| \\ &\leqslant \sum_{i=1}^n \int_{x_{t-1}}^{x_i} |f(t)|dt \\ &= \int_a^b |f(t)|dt \\ \Rightarrow \quad T_a^b(F) &= \int_a^b |f(t)|dt < \infty \end{aligned}$$

where $T_a^b(F)$ is the total variation of F on $[a, b]$. Hence the result follows. □

Theorem 9.2.2. Let f be an integrable function on [a,b]. If

$$\int_a^x f(t)dt = 0,$$

for all $x \in [a, b]$, then $f = 0$ a.e. in $[a, b]$

Proof. If possible, let $f \neq 0$ a.e. in $[a, b]$. Suppose $f(t) > 0$ on a set E of positive measure. Then there exists a closed set $F \subset E$ with $m(F) > 0$. Put $A = (a, b) \setminus F$. Then A is an open set and we have

$$\begin{aligned} 0 &= \int_a^b f(t)dt = \int_{A\cup F} f(t)dt = \int_A f(t)dt + \int_F f(t)dt \\ \Rightarrow \quad & \int_A f(t)dt = -\int_F f(t)dt. \end{aligned}$$

But $f(t) > 0$ on F with $m(F) > 0$ implies

$$\int_F f(t)dt \neq 0.$$

Therefore,

$$\int_A f(t)dt \neq 0.$$

Now, A being an open set, it can be expressed as a union of countable collection $\{(a_m, b_n)\}$ of disjoint open intervals. Thus

$$\begin{aligned} & 0 \neq \int_A f(t)dt = \sum_n \int_{a_n}^{b_n} f(t)dt \\ \Rightarrow \quad & \int_{a_n}^{b_n} f(t)dt \neq 0 \quad \text{for some } n \\ \Rightarrow \quad & \text{either } \int_a^{a_n} f(t)dt \neq 0 \quad \text{or} \quad \int_a^{b_n} f(t)dt \neq 0. \end{aligned}$$

In either case, we see that if f is positive on a set of positive measure, then for some $x \in [a, b]$ we have

$$\int_a^x f(t)dt \neq 0.$$

Similar assertion is obtained if f is negative on a set of positive measure. Hence the result follows by contradiction.

□

Theorem 9.2.3. Let f be a bounded and measurable function defined on $[a, b]$. If

$$F(x) = \int_a^x f(t)dt + F(a),$$

then $F'(x) = f(x)$ a.e. in $[a, b]$.

Proof. Since f is bounded and measurable, it is integrable (cf. IV2.1). Therefore, in view of Theorem 4.1, F is a continuous function of bounded variation on $[a, b]$ and hence F' exists a.e. in $[a, b]$, cf. Corollary 3.3. Let $|f| \leq K$. Set

$$f_n(x) = \frac{F(x + h) - F(x)}{h},$$

with $h = \frac{1}{n}$. Then

$$\begin{aligned} & f_n(x) = \frac{1}{h}\int_x^{x+h} f(t)dt \\ \Rightarrow \quad & |f_n| \leq K. \end{aligned}$$

Also, $f_n \to F'$ a.e. If $c \in [a,b]$ is arbitrary, then the Bounded convergence theorem implies that

$$\begin{aligned}\int_a^c F'(x)dx &= \lim_{n\to\infty}\int_a^c f_n(x)dx \\ &= \lim_{h\to 0}\frac{1}{h}\int_a^c [F(x+h)-F(x)]dx \\ &= \lim_{h\to 0}\left[\frac{1}{h}\int_c^{c+h} F(x)dx - \frac{1}{h}\int_a^{a+h} F(x)dx\right].\end{aligned}$$

But F being a continuous function, we note that

$$\begin{aligned}\lim_{h\to 0}\frac{1}{h}\int_c^{c+h} F(x)dx &= \lim_{h\to 0}\frac{1}{h}\mathscr{R}\int_c^{c+h} F(x)dx \\ &= \lim_{h\to 0}\frac{1}{h}F(c+\theta h), \quad 0<\theta<1 \\ &= F(c);\end{aligned}$$

and similarly

$$\lim_{h\to 0}\frac{1}{h}\int_a^{a+h} F(x)dx = F(a).$$

Thus

$$\begin{aligned}&\int_a^c F'(x)dx = F(c)-F(a) = \int_a^c f(x)dx \\ \Rightarrow \quad &\int_a^c \big[F'(x)-f(x)\big] = 0.\end{aligned}$$

This is true for all c in $[a,b]$. Hence, by the previous theorem, we have $F' = f$ a.e. □

Note that in the above theorem, the function considered is bounded and measurable. We extend it, in the following theorem, to any measurable function which of course is integrable.

Theorem 9.2.4. Let f be an integrable function on $[a,b]$, and suppose

$$F(x) = \int_a^x f(t)dt + F(a).$$

Then $F'(x) = f(x)$ a.e. in $[a,b]$.

Proof. Without any loss of generality, we may assume that $f \geq 0$. Define a sequence $\{f_n\}$ of functions $f_n : [a,b] \to \mathbb{R}$, where

$$\begin{aligned}f_n(x) &= f(x) \quad \text{if } f(x) \leq n, \\ &= n \quad \text{if } f(x) > n.\end{aligned}$$

Clearly, each f_n is a bounded measurable function and so by the previous theorem, we have

$$\frac{d}{dx}\int_a^x f_n = f_n(x) \text{ a.e.}$$

Also, $f - f_n \geq 0$ for all n, and hence the function G_n defined by

$$G_n(x) = \int_a^x (f - f_n)$$

is an increasing function of x, which must have a derivative almost everywhere, (since it is a function of bounded variation); and this derivative would, clearly, be nonnegative. Thus, from the relation

$$F(x) = \int_a^x f(t)dt + F(a) = G_n(x) + \int_a^x f_n(t)dt + F(a),$$

it follows that

$$\begin{aligned} F'(x) &= G_n'(x) + f_n(x) \text{ a.e.} \\ &\geqslant f_n(x) \text{ a.e.}, \quad \forall n. \end{aligned}$$

Since n is arbitrary, we have

$$\begin{aligned} &F'(x) \geqslant f(x) \text{ a.e.} \\ &\int_a^b F'(x)dx \geqslant \int_a^b f(x)dx = F(b) - F(a). \end{aligned}$$

Consequently, we get

$$\begin{aligned} \int_a^b F'(x)dx &= F(b) - F(a) \\ &= \int_a^b f(x)dx \\ \Rightarrow \quad \int_a^b \left\{F'(x) - f(x)\right\} dx &= 0. \end{aligned}$$

Since $F'(x) - f(x) \geqslant 0$, this gives that $F'(x) - f(x) = 0$ a.e., and so $F'(x) = f(x)$ a.e. □

An indefinite integral need not be differentiable everywhere, and even if it is differentiable, it need not follow that $F' = f$ everywhere.

Example 9.2.5. 1. 1. Consider the function $f : [0, 2] \to \mathbf{R}$ given by

$$f(x) = \begin{cases} 1 & \text{if } 0 \leqslant x \leqslant 1 \\ 2 & \text{if } \quad 1 < x \leqslant 2 \end{cases}$$

Then

$$F(x) = \int_0^x f(t)dt = \begin{cases} x & \text{if } 0 \leqslant x \leqslant 1, \\ 2x - 1 & \text{if } 1 < x \leqslant 2 \end{cases}$$

Here $\boldsymbol{F}$ defines a function which is continuous but not differentiable in $[0, 2]$. In fact, F is not differentiable at $x = 1$.

2. Consider the function $f : [0, 1] \to \mathbf{R}$ given by

$$f(x) = \begin{cases} \frac{1}{q} & \text{if } x = p/q \\ 0 & \text{if otherwise.} \end{cases}$$

Then

$$F(x) = \int_0^x f(t)dt = 0.$$

Here F defines a function differentiable in $[0, 1]$. However, $F(x) \neq f(x)$ for $x = p/q$ in $[0, 1]$.

Exercise 9.2.6. 1. If f is the greatest integer function and $F(x) = \int_0^x f(t)dt$, determine F on $[0, 5]$ and verify that $F'(x) = f(x)$ a.e. in $[0, 5]$.

2. Let

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is rational} \\ 0 & \text{if } x \text{ is irrational} \end{cases}$$

If F is as in the preceding exercise, then verify that $F'(x) = f(x)$ a.e. in $[0, 1]$.

9.3 Integral of the derivative

We are acquainted with the idea of absolute continuity of real functions. Let us try to find the relationship between absolute continuity and the integral of a function f.

Theorem 9.3.1. Let f be an integrable function on $[a, b]$. Then the indefinite integral F of f is absolutely continuous on $[a, b]$.

Proof. Let $\epsilon > 0$ be given. Then there is a $\delta > 0$ such that for every measurable set $A \subset [a, b]$ with $m(A) < \delta$, we have

$$\int_A |f| < \epsilon,$$

since the inetgrability of f implies that of $|f|$. Thus for any finite collection $\{(x_i, x_i')\}_{i=1}^N$ of pairwise disjoint open intervals in $[a, b]$ with $\sum_{i=1}^N (x_i' - x_i) < \delta$, we have

$$\sum_{i=1}^N \left| \int_{x_i}^{x_i'} f(t)dt \leq \int_{x_i}^{x_i'} |f(t)|dt < \epsilon \right.$$

$$\Rightarrow \quad \sum_{i=1}^N |F(x_i') - F(x_i)| < \epsilon.$$

Hence, it follows that F is an absolutely continuous function. □

We now prove the converse of the above theorem.

Theorem 9.3.2. If F is an absolutely continuous function on $[a, b]$, then F is an indefinite integral of its derivative; more precisely:

$$F(x) = \int_a^x f(t)dt + C,$$

where $f = F'$ and C is a constant.

Alternatively, we may state the theorem as: If F is absolutely continuous function on $[a, b]$, then F' is integrable over $[a, b]$, and

$$\int_a^x F'(t)dt = F(x) - F(a).$$

Proof. The function F, being absolutely continuous, is of bounded variation, and so we may write

$$F = F_1 - F_2,$$

where F_1 and F_2 are monotone increasing functions. Also, F' exists a.e. on $[a, b]$ and

$$\begin{aligned} & F' = F_1' - F_2' \\ \Rightarrow \quad & \left|F'\right| \leqslant \left|F_1'\right| + \left|F_2'\right|. \end{aligned}$$

This gives

$$\begin{aligned} \int_a^b \left|F'\right| &\leqslant \int_a^b \left|F_1'\right| + \int_a^b \left|F_2'\right| \\ &\leqslant F_1(b) - F_1(a) + F_2(b) - F_2(a) \\ &< \infty \end{aligned}$$

$\Rightarrow$ F' is integrable over $[a, b]$. Write

$$G(x) = \int_a^x F'(t)dt.$$

Then, by the previous theorem, G is an absolutely continuous function on $[a, b]$ and so it the function $H = F - G$. But, it may be noted that

$$H' = F' - G' = 0 \text{ a.e.}$$

Hence, H is a constant function, A (say). Hence,

$$F(x) = \int_a^x F'(t)dt + A.$$

Taking $x = a$, we get $A = F(a)$. This establishes the result. □

From the preceding two theorems we can conclude the following:

A function f is absolutely continuous on $[a, b]$ if and only if it is an indefinite integral of an integrable function on $[a, b]$.

It should also be noted that if the restriction of absolute continuity is removed, then F' need not be integrable.

Example 9.3.3. Let $F : [0, 1] \to \mathbf{R}$ be a function defined by

$$F(x) = \begin{cases} x^2 \sin\left(\frac{\pi}{x^2}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

The derivative F' exists on $[0, 1]$ while F' is not integrable on $[0, 1]$. In fact,

$$\int_0^1 \frac{1}{x} \left|\cos\left(\frac{\pi}{x^2}\right)\right| dx = \infty$$

Note 9.3.4. The class of absolutely continuous functions over $[a, b]$ is identical with the class of functions obtained by integrating Lebesgue integrable functions over $[a, b]$ except that the corresponding functions in two classes differ at the most by a constant.

As an application of Theorems 9.3.1 and 9.3.2, we prove a result on integration by parts which is similar to that for the Riemann integral.

Theorem 9.3.5. Let f and g be integrable functions over $[a, b]$. Suppose

$$F(x) = \int_a^x f(t)dt + F(a)$$

and

$$G(x) = \int_a^x g(t)dt + G(a),$$

for all $x \in [a, b]$. Then

$$\int_a^b F(t)g(t)dt + \int_a^b f(t)G(t)dt = F(b)G(b) - F(a)G(a).$$

Proof. By theorem 9.3.1, F and G are both absolutely continuous functions on $[a, b]$ and hence so is FG. Using theorem 9.3.2, we have

$$\begin{aligned} \int_a^b (FG)' &= (FG)(b) - (FG)(a) \\ &= F(b)G(b) - F(a)G(a). \end{aligned}$$

Also, by a previous result, $F' = f$ a.e. and $G' = g$ a.e. in $[a, b]$, and therefore $(FG)' = FG' + F'G = Fg + fG$ a.e. in $[a, b]$. Hence

$$\int_a^b F(t)g(t)dt + \int_a^b f(t)G(t)dt = F(b)G(b) - F(a)G(a).$$

□

Corollary 9.3.6. If f and g are absolutely continuous functions on $[a, b]$, then

$$\int_a^b f(t)g'(t)dt + \int_a^b f'(t)g(t)dt = f(b)g(b) - f(a)g(a).$$

Proof. Since f and g are absolutely continuous, f' and g' are integrable over $[a, b]$. Also

$$f(x) = \int_a^x f'(t)dt + f(a)$$

and

$$g(x) = \int_a^x g'(t)dt + g(a).$$

The result now follows from Theorem 9.3.5. □

Sample Questions

1. Show that indefinite integral of and integrable function f on $[a, b]$ is a continuous function of bounded variation on $[a, b]$.
2. If f is an integrable function on $[a, b]$ and $\int_a^x f(t)dt = 0$ for all $x \in [a, b]$, then show that $f = 0$ a.e. on $[a, b]$.
3. If f is a bounded measurable function on $[a, b]$ and $F(x) = \int_a^x f(t)dt + F(a)$, then show that $F'(x) = f(x)$ a.e. on $[a, b]$. What happens if f is integrable? Justify your answer.
4. If f is integrable on $[a, b]$, show that its definite integral is absolutely continuous on $[a, b]$.

Unit 10

Course Structure

- Contour Integration
- Conformal mapping, Bilinear transformation.

10.1 Introduction

This unit deals mainly with recollection of the basic idea of contour integration. Thereafter, the idea of conformal mappings have been introduced. Conformal mappings are quite a new concept and has numerous applications in various physical situations. They are precisely the mappings that preserves the angle and shape of objects but not necessarily their size. The idea of conformal equivalences comes from the conformal mappings and are quite interesting to study. Let us start one by one.

Objectives

After reading this unit, you will be able to

- recall the idea of contour integration
- understand the idea of conformal mappings
- study a special kind of conformal mappings called the bilinear transformation and their basic principles

10.1.1 Contour Integration

We already have had a basic idea of integration of a complex variable along a contour. To start, let us first recall the definition of an arc in the complex plane and their types.

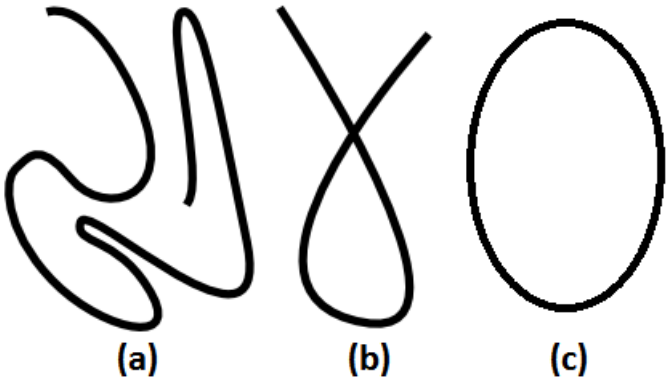
Definition 10.1.1. A set of points $z = (x, y)$ in a complex plane is said to be an arc if

$$x = x(t), \;\; y = t(t) \;\; a \leq t \leq b$$

where $x(t)$ and $y(t)$ are continuous functions of the real parameter t. This definition establishes a continuous mapping of the interval $a \leq t \leq b$ into the z-plane; and the image points are ordered according to increasing values of t. The points of an arc C is given by means of the equation

$$\begin{aligned} z &= z(t) \\ &= x(t) + iy(t). \end{aligned}$$

Definition 10.1.2. The arc C is called a simple arc, or a Jordan arc, if it does not cross itself; that is, C is said to be simple if $z(t_1) \neq z(t_2)$ when $t_1 \neq t_2$. And when the arc is simple except at the end points, that is, $z(a) = z(b)$, then C is a simple closed curve. The figure below gives an idea of (a) simple curve, (b) non-simple curve and (c) simple closed curve.



Example 10.1.3. The unit circle $z = e^{i\theta}$, $0 \leq \theta \leq 2\pi$ is a simple closed closed curve about the origin, oriented in the anti-clockwise sense while $z = e^{-i\theta}$ describes a simple closed curve in the clockwise sense.

Example 10.1.4. The points on the arc

$$z = e^{i2\theta} \ \ (0 \leq \theta \leq 2\pi)$$

are the same as those making up the arcs in the previous example. The arc here differs, however, from each of those arcs since the circle is traversed twice in the anticlockwise sense.

We now turn to integrals of complex-valued function f of the complex variable z. Such an integral is defined in terms of the values $f(z)$ along a given contour C, extending from a point $z = z_1$ to a point $z = z_2$ in the complex plane. It is therefore, a line integral; and its value depends on the contour C as well as on the function f. It is written as $\int_C f(z)dz$ or $\int_{z_1}^{z_2} f(z)dz$, the latter notation often being used when the value of the integral is independent of the choice of the contour taken between the two fixed points. If $f(z) = u(x, y) + iv(x, y)$ then the value of the integral is

$$\int_C f(z)dz = \int_C (u + iv)(dx + idy) = \int_C (udx - vdy) + i\int_C (udy + vdx).$$

The integral can be calculated in another way. Let the equation

$$z = z(t), \ \ a \leq t \leq b \tag{10.1.1}$$

represents the contour C, extending from a point $z_1 = z(a)$ to a point $z_2 = z(b)$. Let the function f be piecewise continuous on C. We define the line integral, or *contour integral* of f along C as follows:

$$\int_C f(z)dz = \int_a^b f[z(t)]z'(t)dt \tag{10.1.2}$$

Note that, since C is a contour, $z'(t)$ is also piecewise continuous on the interval $a \leq t \leq b$; so the existence of the integral (10.1.2) is ensured.

It follows immediately from the definition (10.1.2) that for any constant z_0,

$$\int_C z_0 f(z)dz = z_0 \int_C f(z)dz$$

and,

$$\int_C [f(z) + g(z)]dz = \int_C f(z)dz + \int_C g(z)dz$$

Associated with the contour C used in the integral (10.1.2), is the contour $-C$, consisting of the same set of points but with the order reversed so that the new contour extends from the point z_2 to z_1. The contour $-C$ has the parametric representation

$$z = z(-t), \quad -b \leq t \leq -a$$

so, we have

$$\begin{aligned} \int_{-C} f(z)dz &= \int_{-b}^{-a} f[z(-t)]\frac{d}{dt}z(-t)dt \\ &= -\int_{-b}^{-a} f[z(-t)]z'(-t)dt \end{aligned}$$

where $z'(-t)$ denotes the derivative of $z(t)$ with respect to t, evaluated at $-t$. Making the substitution $\tau = -t$ in the last integral, we obtain

$$\int_{-C} f(z)dz = -\int_a^b f[z(\tau)]z'(\tau)d\tau$$

which is the same as

$$\int_{-C} f(z)dz = -\int_C f(z)dz \tag{10.1.3}$$

Now, consider a path C, denoted as $z = z(t)$, $a \leq t \leq b$, that consists of a contour C_1 from z_1 to z_2 followed by a contour z_2 to z_3, the initial point of C_2 being the final point of C_1. There is a value c of t, $a < t < b$, such that $z(c) = z_2$. Also,

$$C = C_1 + C_2$$

Also, we can represent C_1 as

$$z = z(t), \quad a \leq t \leq c$$

and C_2 as

$$z = z(t), \quad c \leq t \leq b.$$

By a rule of the integrals of the functions $w(t)$, we have,

$$\int_a^b f[z(t)]z'(t)dt = \int_a^c f[z(t)]z'(t)dt + \int_c^b f[z(t)]z'(t)dt$$

which means that,

$$\int_C f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz$$

Sometimes the contour C is called the sum of its legs C_1 and C_2 and is denoted by $C_1 + C_2$.

Definite integrals in calculus can be interpreted as areas, and they have other interpretations as well. Except in special cases , no corresponding helpful interpretation, geometric or physical, is available for integrals in the complex plane. Let us look into a few examples for contour integrals.

Example 10.1.5. Let us find the value of the integral

$$I = \int_C \bar{z} dz$$

where C is the right-hand half

$$z = e^{2i\theta} \quad (-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2})$$

of the circle $|z| = 2$, from $z = -2i$ to $z = 2i$. By the definition above, we have

$$I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2\bar{e^{i\theta}}(2e^{i\theta})' d\theta;$$

and since $\overline{e^{i\theta}} = e^{-i\theta}$ and $(e^{i\theta})' = ie^{i\theta}$, this means that

$$I = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} 2e^{-i\theta} 2ie^{i\theta} d\theta = 4 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta.$$

Note that when a point z is on the circle $|z| = 2$, it follows that $z\bar{z} =$, or $\bar{z} = 4/z$. Hence the result $I = 4\pi i$ can also be written as

$$\int_C \frac{dz}{z} = \pi i$$

Example 10.1.6. Let C_1 denote the contour OAB which joins the origin $O(0, 0)$ to $A(0, 1)$ and $B(1, 1)$; and C_2 be the contour joining B to O. Suppose we are to find the integral

$$\int_C f(z) dz$$

where, $C = C_1 + C_2$ and $f(x, y) = y - x - i3x^2$. Then,

$$\int_C f(z) dz = \int_{C_1} f(z) dz + \int_{C_2} f(z) dz$$

Now,

$$\int_{C_1} f(z) dz = \int_{OA} f(z) dz + \int_{AB} f(z) dz \tag{10.1.4}$$

The leg OA can be represented as $z = 0 + iy$, $0 \leq y \leq 1$; and since $x = 0$ at points on that leg, the values of f there vary with the parameter y according to the equation $f(z) = y$, $0 \leq y \leq 1$. As a result,

$$\begin{aligned} \int_{OA} f(z) dz &= \int_0^1 yi dy \\ &= i \int_0^1 y dy \\ &= \frac{i}{2}. \end{aligned}$$

On the leg AB, $z = x + i$, $0 \leq x \leq 1$. So,

$$\begin{aligned} \int_{AB} f(z) dz &= \int_0^1 (1 - x - i3x^2) dx \\ &= \int_0^1 (1 - x) dx - 3i \int_0^1 x^2 dx \\ &= \frac{1}{2} - i. \end{aligned}$$

So, by equation (10.1.4), we have,

$$\int_{C_1} f(z)dz = \frac{1-i}{2}.$$

Again, the leg BO has parametric representation $z = x + ix$, $0 \le x \le 1$. So,

$$\begin{aligned}\int_{C_2} f(z)dz &= \int_1^0 -i3x^2(1+i)dx \\ &= 3(1-i)\int_1^0 x^2dx \\ &= i-1.\end{aligned}$$

Then, the integrals of f along the paths C_1 and C_2 have different values even though those paths have the same initial and the same final points. Now, we have

$$\int_C f(z)dz = \frac{i-1}{2}.$$

Exercise 10.1.7. Evaluate each of the following integrals using contour integration.

1. $\displaystyle\int_{1-i}^{1+i} z^3dz.$
2. $\displaystyle\int_{-3i}^{2i} (z^3 - z)dz$
3. $\displaystyle\int_C \frac{1}{z}dz$, where C is the
   (a) arc of the circle $4\,\mathrm{e}^i\,t$, $-\pi/2 \le t \le \pi/2$.
   (b) line segment between $1+i$ and $4+4i$.
4. $\displaystyle\int_1^2 (t^2+i)^2dt$
5. $\displaystyle\int_0^{\pi/2} \mathrm{e}^{-2it}dt$
6. $\displaystyle\int_C (z^2 - 3|z| + \mathrm{Im}z)dz$ where $C = 2\,\mathrm{e}^{it}$, $0 \le t \le \pi/2$.
7. $\displaystyle\int_C f(z)dz$ where $f(z) = \mathrm{e}^y\,\mathrm{e}^{1-ix}$ and $C = 2\,\mathrm{e}^{it}$, $0 \le t \le \pi/2$.

10.1.2 Conformal Mappings

Let C be a smooth arc represented by the equation

$$z = z(t), \quad a \le t \le b$$

and let $f(z)$ be a function defined at all points z on C. The equation

$$w = f(z(t)), \quad a \leq t \leq b$$

is a parametric representation of the image γ of C under the transformation $w = f(z)$.

Suppose that C passes through a point $z_0 = z(t_0)$, at which f is analytic and that $f'(t_0) \neq 0$. According to the chain rule of differentiation, we get

$$w'(t_0) = f'(z_0)z'(t_0)$$

and this gives

$$\arg w'(t_0) = \arg f'(z_0) + \arg z'(t_0) \tag{10.1.5}$$

Now, let ψ_0 be a particular value of $\arg f'(z_0)$, and let θ_0 be the angle of inclination of a directed line tangent to γ at z_0. Then from equation (10.1.5), we get that

$$\phi_0 = \psi_0 + \theta_0,$$

which is the value of $\arg w'(t_0)$ and is hence the angle of inclination of a directed line tangent to γ at the point $w_0 = f(z_0)$. Hence the angle of inclination of the directed line at w_0 differs from the angle of inclination of the directed line at z_0 by the angle of rotation

$$\psi_0 = \arg f'(z_0).$$

Now, let C_1 and C_2 be two smooth arcs passing through z_0, and let θ_1 and θ_2 be angles of inclination of directed lines tangent to C_1 and C_2 respectively, at z_0. We know from the preceding paragraph that the quantities

$$\begin{aligned} \phi_1 &= \psi_0 + \theta_1 \\ \phi_2 &= \psi_0 + \theta_2 \end{aligned}$$

are the angles of inclination of directed lines tangent to the image curves γ_1 and γ_2, respectively, at the point $w_0 = f(z_0)$. Thus, $\phi_2 - \phi_1 = \theta_2 - \theta_1$, that is, the angle $\phi_2 - \phi_1$ from γ_1 to γ_2 is the same in magnitude and sense as the angle $\theta_2 - \theta_1$ from C_1 to C_2.

Because of this angle preserving property, a transformation $w = f(z)$ is said to be $conformal$ at a point if f is analytic there and $f'(z_0) \neq 0$.

Example 10.1.8. The mapping $w = e^z$ is conformal throughout the entire complex plane since $(e^z) = e^z \neq 0$ for each z.

A mapping that preserves the magnitude of the angle between two smooth arcs but not necessarily the sense is called an $isogonal\ mapping$.

Example 10.1.9. The transformation $w = \overline{z}$, which is a reflection in the real axis, is isogonal but not conformal.

Suppose that f is not a constant function and is analytic at a point z_0. If, in addition, $f'(z_0) = 0$, then z_0 is called the *critical point* of the transformation f.

Example 10.1.10. The point $z = 0$ is a critical point of the transformation $w = 1 + z^2$.

The behavior of an analytic function in a neighborhood of critical point is given by the following theorem

Theorem 10.1.11. Let f be analytic at z_0. If $f'(z_0) = \ldots = f^{(k-1)}(z_0) = 0$, and $f^{(k)}(z_0) \neq 0$ then the mapping $w = f(z)$ magnifies the angle at z_0 by a factor k.

Proof. Since f is analytic at z_0, so

$$\begin{aligned} f(z) &= f(z_0) + (z - z_0)f'(z_0) + \frac{(z - z_0)^2}{2!}f''(z_0) + \ldots \\ &= f(z_0) + (z - z_0)^k \left[\frac{1}{k!}f^{(k)}(z_0) + \frac{1}{(k+1)!}(z - z_0)f^{(k+1)}(z_0) + \ldots\right] \\ &= f(z_0) + (z - z_0)^k g(z) \text{ (say)}. \end{aligned}$$

Thus,

$$f(z) - f(z_0) = (z - z_0)^k g(z),$$

and hence

$$\arg(w - w_0) = k \arg(z - z_0) + \arg g(z).$$

□

We now come across an important example of conformal maps of the unit disk.

Example 10.1.12. Consider a map $f : D \mapsto D$ of the form

$$f(z) = \frac{a - z}{1 - \overline{a}z}$$

where D is the open unit disk and a is a point inside D. We note that f is analytic in the unit disk since

$$\left|\frac{1}{\overline{a}}\right| = \frac{1}{|a|} > 1$$

since $|a| < 1$. Now,

$$f'(z) = -\frac{1 + |a|^2}{(1 - \overline{a}z)^2} \neq 0$$

in D. Hence f is conformal in D. Also, we can check that,

$$(f \circ f)(z) = z$$

This shows that f is a bijective conformal map from D onto D.

The above example is in fact a conformal map moving the origin to the point a. In fact, there are many such bijective conformal maps from the unit circle onto itself. We will discuss them shortly. For now, let us see the following definition:

Definition 10.1.13. Let U and V be any two subsets of the complex plane. Then U and V are said to be conformally equivalent if there exists a bijective analytic map $f : U \to V$. An important fact is that, given such a function f, its inverse is automatically analytic.

We have the following theorem in this context:

Theorem 10.1.14. If $f : U \to V$ is analytic and injective, then $f'(z_0) \neq 0$ for all $z_0 \in U$. In particular, the inverse of f defined on its range is analytic, and thus the inverse of a conformal map is also analytic.

The Unit Disk and the Upper Half Plane

The upper half plane, denoted by $\mathbb{H}$, consists of the complex numbers with positive imaginary part, that is

$$\mathbb{H} = \{z \in \mathbb{C} : \text{Im}(z) > 0\}$$

A remarkable fact, which at first seems surprising, is that the unbounded set $\mathbb{H}$ is conformally equivalent to the unit disc. Moreover, an explicit formula giving this equivalence exists. Indeed, let

$$F(z) = \frac{i-z}{i+z},\ G(w) = i\frac{1-w}{1+w}$$

Theorem 10.1.15. The map $F : \mathbb{H} \to D$ is a conformal map with inverse $F : D \to \mathbb{H}$.

Proof. First we observe that both the maps are analytic in their respective domains. Then, notice that any point on the upper half plane is closer to i than to $-i$. So, $|F(z)| < 1$ and F maps $\mathbb{H}$ into D. To prove that G maps into the upper half-plane, we must compute $\text{Im}(G(w))$ for $w \in D$. For this, let $w = u + iv$. Then

$$\begin{aligned}
\text{Im}(G(w)) &= \text{Re}\left(\frac{1-u-iv}{1+u+iv}\right) \\
&= \text{Re}\left(\frac{(1-u-iv)(1+u-iv)}{(1+u)^2+v^2}\right) \\
&= \frac{1-u^2-v^2}{(1+u)^2+v^2} > 0
\end{aligned}$$

since $|w| < 1$. Hence G maps the D to the upper half plane. Finally,

$$\begin{aligned}
F(G(w)) &= \frac{i - i\frac{1-w}{1+w}}{i + i\frac{1-w}{1+w}} \\
&= \frac{1+w-1+w}{1+w+1-w} \\
&= w
\end{aligned}$$

We can similarly show that $G(F(z)) = z$. This proves the theorem. □

An interesting aspect of these functions is their behaviour on the boundaries of our open sets. Observe that F is analytic everywhere on $\mathbb{C}$ except at $z = -i$, and in particular it is continuous everywhere on the boundary of $\mathbb{H}$, namely the real line. If we take $z = x$ real, then the distance from x to i is the same as the distance from i to $-i$, therefore $|F(x)| = 1$. Thus F maps $\mathbb{R}$ onto the boundary of D.

Example 10.1.16. Translations and dilations provide simple examples. If $h \in \mathbb{C}$, the translation $z \mapsto z + h$ is a conformal map from $\mathbb{C}$ to itself whose inverse is $w \mapsto w - h$.

For any non-zero complex number c, the map $z \mapsto cz$ is a map from $\mathbb{C}$ to itself, whose inverse is $w \mapsto c^{-1}w$. If $|c| = 1$, then $c = e^{i\psi}$ for some real ψ, then the map is a rotation by angle ψ. If $c > 0$ then the map is a dilation. Finally, if $c < 0$ the map consists of a dilation by the factor $|c|$, followed by a rotation of π.

Example 10.1.17. The map

$$f(z) = \frac{1+z}{1-z}$$

takes the upper half disc $\{z \in \mathbb{C} : |z| < 1 \text{ and } \text{Im}(z) > 0\}$ conformally to the first quadrant $\{w = u + iv : u > 0 \text{ and } v > 0\}$. Indeed, if $z = x + iy$, then

$$f(z) = \frac{1 - (x^2 + y^2)}{(1-x)^2 + y^2} + i\frac{2y}{(1-x)^2 + y^2}$$

The inverse map, given by

$$g(w) = \frac{w-1}{w+1}$$

is clearly analytic in the first quadrant. Moreover, for all w in the first quadrant, $|w+1| > |w-1|$, since the distance of w to -1 is greater than the distance of w from 1. Thus g maps into the disc D. An easy calculation also shows that the imaginary part of $g(w)$ is positive whenever w is in the first quadrant. So we conclude that f is conformal since g is the inverse of f.

To examine the action of f on the boundary, note that if $z = e^{i\theta}$ belongs to the upper half circle, then

$$\begin{aligned} f(z) &= \frac{1 + e^{i\theta}}{1 - e^{i\theta}} \\ &= \frac{e^{-i\theta/2} + e^{i\theta/2}}{e^{-i\theta/2} - e^{i\theta/2}} \\ &= \frac{i}{\tan(\theta/2)} \end{aligned}$$

As θ travels from 0 to π, we see that $f(e^{i\theta})$ travels along the imaginary axis from infinity to 0. Moreover, if $z = x$ is real, then

$$f(z) = \frac{1+x}{1-x}$$

is also real. So, we see that f is actually a bijection from $(-1, 1)$ to the positive real axis, with $f(x)$ increasing from 0 to infinity as x travels from -1 to 1. Note also that $f(0) = 1$.

10.1.3 Bilinear Transformations

Bilinear Transformations are a special kind of conformal maps. We formally define them as follows:

Definition 10.1.18. A mapping of the form

$$S(z) = \frac{az + b}{cz + d},$$

where a, b, c, and d are complex constants satisfying $ad - bc \neq 0$, is called a Bilinear transformation or a linear fractional transformation (LFT) or Möbius Transformation. If $ad - bc = 0$, then $S(z)$ is a constant map (Prove it!).

Example 10.1.19. Let S be a Möbius Transformation on the upper half plane $\mathbb{H}$ where,

$$S(z) = \frac{az + b}{cz + d}$$

where $ad - bc > 0$. Then S maps $\mathbb{H}$ conformally onto itself.

If S is a Möbius Transformation, then

$$S^{-1}(z) = \frac{dz - b}{-cz + a},$$

satisfies $S(S^{-1}(z)) = S^{-1}(S(z)) = z$. Then, S^{-1} is the inverse of S. Now, if S and T are both linear fractional transformations, then it follows that $S \circ T$ is also so. S can be defined on the extended complex plane $\mathbb{C}_\infty$ with $S(\infty) = a/c$ and $S(-d/c) = \infty$. Since S has inverse, it maps $\mathbb{C}_\infty$ to $\mathbb{C}_\infty$. It is worth saying that every Möbius transformation as given in the definition is analytic in $\mathbb{C} \setminus \{-d/c\}$.

If $S(z) = z + a$, it is called Translation; if $S(z) = az$ with $a \neq 0$, it is called dilation; if $S(z) = \exp^{i\theta}$, it is called rotation; if $S(z) = 1/z$, it is called inversion.

Theorem 10.1.20. If S is a Möbius Transformation, then S is the composition of translations, dilations, rotations, and inversions.

Proof. First, let us suppose that $c = 0$. Hence, $S(z) = \frac{a}{d}z + \frac{b}{d}$. Hence, $S_1(z) = \frac{a}{d}z$, then $S_2 \circ S_1 = S$ and we are done.

Now, let $c \neq 0$ and put $S_1(z) = z + \frac{d}{c}$, $S_2(z) = \frac{1}{z}$, $S_3(z) = \frac{ad - bc}{c^2}z$, $S_4(z) = z + \frac{a}{c}$. Then,

$$S \equiv S_4 \circ S_3 \circ S_2 \circ S_1.$$

Hence the result. □

Example 10.1.21. Let $S(z) = az$. If a is real, then it scales the plane whereas, if $a = \mathrm{e}^{i\theta}$, then it scales the plane. So, if $a = r\,\mathrm{e}^{i\theta}$, then it does both.

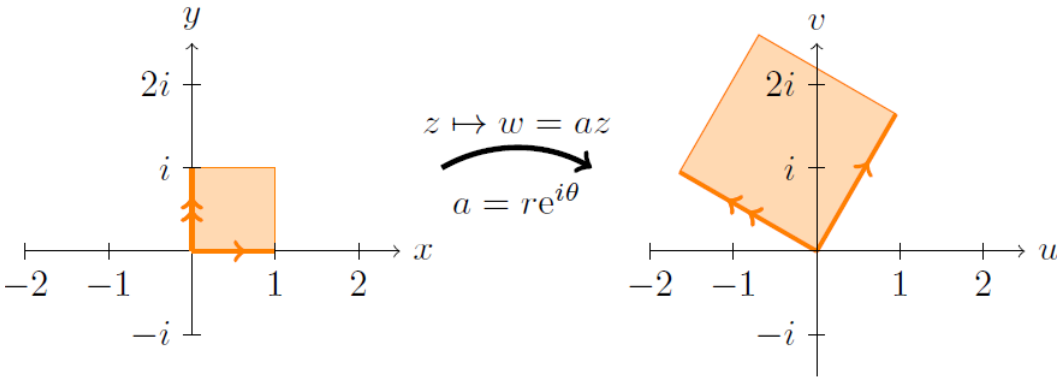


Figure 10.1.1: Scale and rotate

Example 10.1.22. Let $S(z) = az + b$. Adding the term b adds translation to the previous example. The representation is given in figure 10.1.2.

Example 10.1.23. Let $S(z) = \frac{1}{z}$. It turns the unit circle inside out. Note that $S(0) = \infty$ and $S(\infty) = 0$. In the figure 10.1.3, the circle that is outside the unit circle in the z plane is inside the unit circle in the w plane and vice-versa. Note that the arrows on the curves are reversed.

Theorem 10.1.24. A linear fractional transformation maps lines and circles to lines and circles.

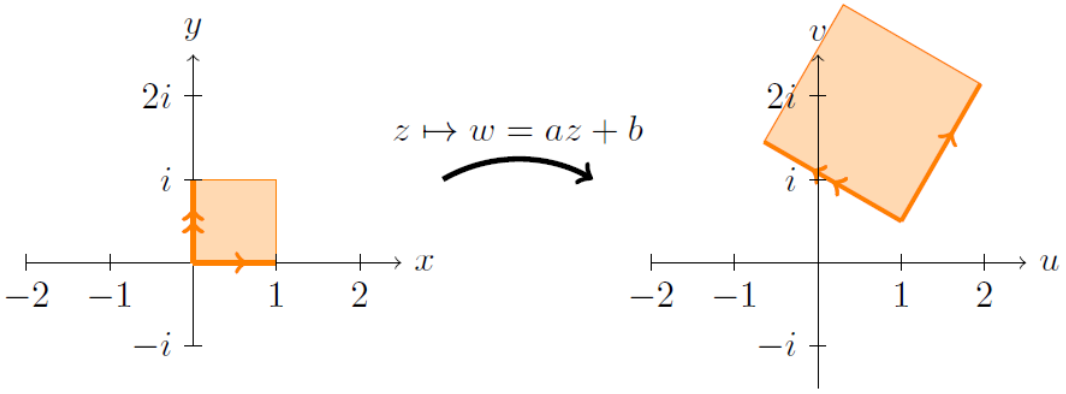


Figure 10.1.2: Scale, rotate and translate

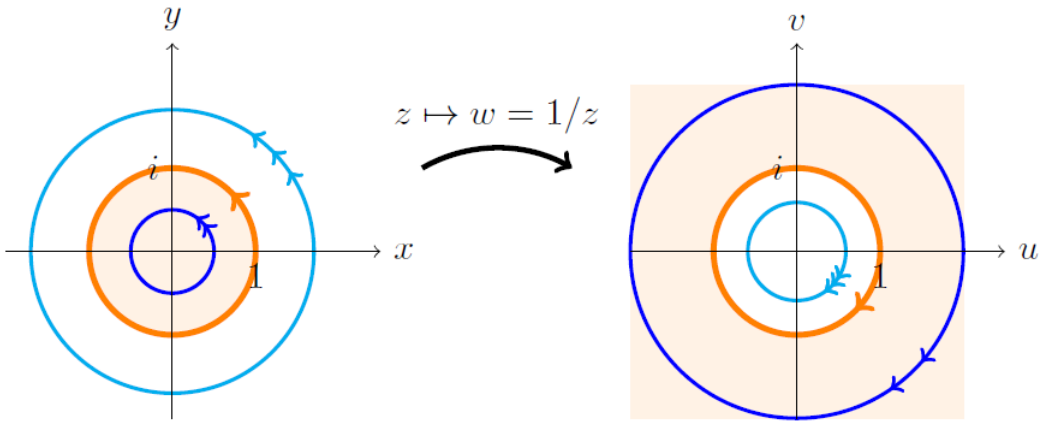


Figure 10.1.3: Inversion

Proof. We start by showing that inversion maps lines and circles to lines and circles. Given z and $w = 1/z$, we define x, y, u and v as

$$z = x + iy, \quad \text{and} \quad w = \frac{1}{z} = \frac{x - iy}{x^2 + y^2} = u + iv,$$

so,

$$u = \frac{x}{x^2 + y^2}, \quad \text{and} \quad v = \frac{-y}{x^2 + y^2}.$$

Now, every circle or line can be described by the equation

$$Ax + By + C(x^2 + y^2) = D.$$

If $C = 0$ it describes a line, otherwise a circle. We convert this to an equation in u, v as follows.

$$\begin{aligned} & Ax + By + C(x^2 + y^2) = D \\ \Rightarrow \quad & \frac{A}{x^2 + y^2} + \frac{B}{x^2 + y^2} + C = \frac{D}{x^2 + y^2} \\ \Rightarrow \quad & Au - Bv + C = D(u^2 + v^2). \end{aligned}$$

We have shown that a line or circle in x, y is transformed to a line or circle in u, v. This shows that inversion maps lines and circles to lines and circles.

We note that for the inversion $w = 1/z$,

1. Any line not through the origin is mapped to a circle through the origin.
2. Any line through the origin is mapped to a line through the origin.
3. Any circle not through the origin is mapped to a circle not through the origin.
4. Any circle through the origin is mapped to a line not through the origin.

Now, to prove that an arbitrary fractional linear transformation maps lines and circles to lines and circles, we factor it into a sequence of simpler transformations.

First suppose that $c = 0$. So,

$$S(z) = \frac{az + b}{d}.$$

Since this is just translation, scaling and rotating, it is clear it maps circles to circles and lines to lines.

Now suppose that $c \neq 0$. Then

$$S(z) = \frac{az + b}{cz + d} = \frac{a}{c} + \frac{b - ad/c}{cz + d}.$$

So, $w = S(z)$ can be computed as a composition of translations, dilations, rotations, and inversions. We know that each of the transforms in this sequence maps lines and circles to lines and circles. Therefore the entire sequence also follows the same. □

Now let us investigate about the fixed points of S. Fixed points of S are precisely the roots of the equation

$$\begin{aligned} & S(z) = z \\ \Rightarrow \quad & cz^2 + (d - a)z - b = 0. \end{aligned}$$

Hence, a Möbius Transformation can have at most two fixed points unless it is the identity transformation.

Now, let S_1 be a Möbius Transformation and let a, b, c be distinct points in $\mathbb{C}_\infty$ with, $S_1(a) = \alpha$, $S_1(b) = \beta$, $S_1(c) = \gamma$. Let S_2 be another Möbius Transformation with the same property. Then, $S_2^{-1} \circ S_1$ has a, b, c as fixed points. So, $S_2^{-1} \circ S_1$ is identically equal to the identity transformation. Thus, $S_1 \equiv S_2$. Hence, a Möbius Transformation is uniquely determined by its action on three distinct points in $\mathbb{C}_\infty$.

Let z_1, z_2, z_3 be points in $\mathbb{C}_\infty$. Define $S : \mathbb{C}_\infty \mapsto \mathbb{C}_\infty$ by

$$\begin{aligned} S(z) &= \left(\frac{z - z_3}{z - z_4}\right)\left(\frac{z_2 - z_4}{z_2 - z_3}\right) \; if \; z_2, z_3, z_4 \in \mathbb{C} \\ &= \frac{z - z_3}{z - z_4} \; if \; z_2 = \infty \\ &= \frac{z_2 - z_4}{z - z_4} \; if \; z_3 = \infty \\ &= \frac{z - z_3}{z_2 - z_3} \; if \; z_4 = \infty \end{aligned}$$

In all the cases, $S(z_2) = 1$, $S(z_3) = 0$, and $S(z_4) = \infty$ and S is the only transformation having the property.

Definition 10.1.25. If $z_1 \in \mathbb{C}_\infty$, then (z_1, z_2, z_3, z_4), called the *cross ratio* of z_1, z_2, z_3 and z_4, is the image of z_1 under the unique Möbius Transformation that takes z_2 to 1, z_3 to 0 and z_4 to ∞.

Exercise 10.1.26. 1. Show that an LFT S given by

$$S(z) = \frac{az+b}{cz+d}$$

where $ad - bc = 0$, is a constant.

2. Show that the map

$$S(z) = \frac{z-i}{z+i}$$

maps the x-axis to the unit circle and the upper half-plane to the unit disk.

Sample Questions

1. Show that the Möbius transformation $S(z) = \dfrac{az+b}{cz+d}$, where $ad - bc > 0$ maps the upper half plane conformally to itself.
2. Show that a Möbius Transformation is the composition of translations, dilations, rotations, and inversions.
3. Show that an LFT maps lines and circles to lines and circles.

Unit 11

Course Structure

- Idea of analytic continuation.
- Multivalued functions
- Branch cuts, branch point.

11.1 Introduction

This unit is divided into two sections. The first section deals with the analytic continuation of functions and the second section deals with the idea of multivalued functions. First the idea of analytic continuation deals with extending the domain of an analytic function to some “larger" one while preserving the analyticity. It is primarily based on the uniqueness of entire functions (more famously called the interior uniqueness theorem) which is discussed in unit 12 12.1.9. Next comes the multivalued functions. As the name suggests, a “function" having multiple values is called multivalued. However, the definition is not that simple. From basic set theory, we know that a relation, where a point of the domain has more than one images, fails to be a function. As opposed to the real line, the numbers in the complex plane has two aspects, one being the magnitude and the other the amplitude. The multivalued-ness of such functions arises due to this amplitude as we shall later see.

Objectives

After reading this unit, you will be able to

- define analytic continuation of function f and discuss certain examples and consequences
- define and understand the multivalued functions and related terminology

11.1.1 Analytic Continuation

Our main aim in this section is whether a given domain of an analytic function can be extended to a larger one. We now formally define analytic continuation of a function analytic in a domain D.

Definition 11.1.1. Let the function f be analytic in a domain D. If there exists another function g analytic in a domain D_1 such that

1. $D \cap D_1 \neq \phi$,
2. $f(z) = g(z), \ \forall z \in D \cap D_1$,

then we say that g is a direct analytic continuation of f and vice-versa.

Theorem 11.1.2. (Uniqueness of analytic continuation) Let f be analytic in domain D. If f_1, f_2 be two direct analytic continuations of f in the domain D_1, then $f_1(z) = f_2(z), \forall z \in D_1$.

Proof. By hypothesis, $f(z) = f_1(z) = f_2(z), \forall z \in D \cap D_1$, where $D \cap D_1 \neq \phi$. Hence, f_1, f_2 which are analytic in D_1 coincide in the subdomain $D \cap D_1$. Hence, by the interior uniqueness theorem for analytic functions we have, $f_1(z) = f_2(z), \forall z \in D_1$. □

The function F defined by

$$\begin{aligned} F(z) &= f(z), \quad z \in D \\ &= f_1(z), \quad z \in D_1 \end{aligned}$$

is analytic in the union domain $D \cup D_1$.

Example 11.1.3. Let $f(z) = \sum_{n=0}^{\infty} z^n$ and $g(z) = \dfrac{1}{1-z}$. Then, f is defined and analytic only in the disc $D : |z| < 1$ and $f(z) = \dfrac{1}{1-z}$ in D. f is not defined when $|z| \geq 1$. But the function g is analytic in the domain $D_1 : \mathbb{C} \setminus \{1\}$. Hence, $f(z) = g(z), \forall z \in D \cap D_1 (= D)$. Hence, g is a direct analytic continuation of f from D to D_1.

Example 11.1.4. Let $f(z) = \sum_{n=0}^{\infty} z^n$, $g(z) = \dfrac{1}{1-i} \sum_{n=0}^{\infty} \left(\dfrac{z-i}{1-i} \right)^n$. Then f is defined and analytic only in the domain $D : |z| < 1$ and $f(z) = \dfrac{1}{1-z}$ for all $z \in D$. f is not defined when $|z| \geq 1$. We now find the domain of analyticity of the function g. Let $t = \dfrac{z-i}{1-i}$. Then

$$g(z) = \frac{1}{1-i} \sum_{n=0}^{\infty} t^n$$

and it is analytic in the domain $D_1 : |t| < 1$, that is, in $|z-i| < |1-i| = \sqrt{2}$. Thus, D_1 is the domain $|z-i| < \sqrt{2}$. Clearly $D \cap D_1 \neq \emptyset$. Now, in D_1,

$$g(z) = \frac{1}{1-i} \frac{1}{1-t} = \frac{1}{1-z}.$$

Hence $f(z) = g(z)$ for all $z \in D \cap D_1$. Thus, g is a direct analytic continuation of f and vice versa.

11.1.2 Natural Boundary

Let γ be a closed curve which is the boundary of a region D and f is analytic in D. Then γ is called a natural boundary of f if the function f can not be analytically continued beyond any point of γ. If γ is dense with singularities of f, then γ is a natural boundary of f.

11.2 Multivalued Functions

A *multivalued function* $f(z)$ has two or more distinct values for each value of z. Important multivalued functions are $\sqrt{z}$, $z^{1/n}$, $\sqrt{(z-a)(z-b)}$, $\log z$, z^{α}, for any non-integral α, and the inverse trigonometric functions. In order to use the multi-valued functions , we must use a single branch to ensure single-valuedness. This is done by introducing cuts in the convenient positions of the complex plane, which are commonly called the *branch cuts*. In order to illustrate this, let us use the simplest example, say $w = \sqrt{z}$. If $z = re^{i\theta}$, then the square root has two branches: $w_1 = r^{1/2}e^{i\theta/2}$ and $w_2 = -r^{1/2}e^{i\theta/2}$.

Let us go around $z = 0$. If we start at some point $z_0 = re^{i\theta}$ with the branch w_1, and go along any closed curve around $z = 0$, then r remains the same while θ changes to $\theta + 2\pi$. As z moves, w_1 changes continuously, and when z returns to the original point, w_1 has changed to w_2. Going around $z = 0$ gain changes w_2 again to w_1. Any other path which excludes $z = 0$ has ordinary single-valued behaviour. The point $z = 0$, in this case acts as a singularity called *branch point* for the function and is very different from pole.The function $\sqrt{z}$ is double valued in any region which includes $z = 0$ as an interior point. If z goes on a circuit around $z = 0$, w changes to $-w$. If we stay in a region that does not include the branch point, then while z moves around w_1 does not change to w_2. This is achieved by restricting the movement of z to a region R which excludes the branch point. For this purpose, we just introduce a "cut" in the complex plane around the branch point and extending to infinity. This is known as the *branch cut*. The function remains single-valued as long as z does not cross over the branch cut. If it crosses the cut, then w moves on to the next branch.

Example 11.2.1. The function $\log z$ is the inverse of the exponential. So, if

$$z = e^w$$

then,

$$w = \log z$$

The logarithm has infinite number of branches. We can write from $z = e^w$,

$$\begin{aligned} z &= re^{i\theta} \\ &= re^{i\theta + i2n\pi} \\ &= e^{\ln r + i\theta + i2n\pi} \end{aligned}$$

where, $n = 0, \pm 1, \pm 2, \cdots$, any integer. So, the logarithm is

$$\log z = \ln r + i\theta + i2n\pi$$

Each time z moves in a closed curve around the branch point 0, θ increases by 2π and we go from one branch to another, and this can go on forever. θ goes through multiples of 2π and the infinite set of values of logarithm differ by $2n\pi i$. We can get a single-valued function by picking a cut, such as the negative real axis

$$\log z = \ln r + i\theta, \ -\pi \leq \theta < \pi$$

At two points z_1, z_2, on either side of the cut,

$$\begin{aligned} \log z_1 &= \ln r - i\pi \\ \log z_2 &= \ln r + i\pi \end{aligned}$$

Sample questions

1. Show that the direct analytic continuation of an analytic function f over a domain D is unique.

2. Show that the power series
$$z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \dots$$
may be continued analytically to a wider region by means of the series
$$\log 2 - \frac{1-z}{2} - \frac{(1-z)^2}{2.2^2} - \frac{(1-z)^3}{3.2^3} - \dots$$

3. Show that the function $f(z) = \sum_{n=0}^{\infty} z^{2^n}$ has the unit circle as its natural boundary.

Unit 12

Course Structure

- Zeros of an analytic function
- Singularities and their classification

12.1 Introduction

From previous knowledge, we know that if a function f is analytic at some point α, then α is called a regular point of f. Any point other than the regular points are called singularities. Singularities are of various kinds as we shall see here in this unit. Firstly, we shall start with the zeros of f. The zeros are a particular regular points of f, where the functional value is equal to 0. The zeros and the singularities are somewhat connected and gives an idea of the number of zeros an analytic function can possibly have in a given region. The properties that we shall see here are however true for any a-points of f, that is, the points of f in the domain where $f(z) = a$ for some complex number a. Indeed, if we define another analytic function $g(z) = f(z) - a$, then the properties of the zeros of g and the properties of the a-points of f are equivalent. The singularities are classified in this unit and in the next unit, we shall deal with the characteristics of them.

Objectives

After reading this unit, you will be able to

- define the zeros of analytic functions
- tell about the zeros of analytic functions by looking at its Taylor series expansion about some point
- know the singularities of complex functions and differentiate between their types
- tell about the poles of a complex function by looking at their Laurent series expansion about some point

12.1.1 Zeros of an analytic function

If a function f is analytic at a point α, then α is called a zero of f if $f(\alpha) = 0$. Recall that if a function f is analytic at a point α then f has a Taylor series expansion

$$f(z) = \sum_{n=0}^{\infty} a_n (z-\alpha)^n,$$

where

$$a_n = \frac{f^{(n)}(\alpha)}{n!}.$$

If α is a zero of f then $a_0 = 0$. Other than a_0, if $a_1 = a_2 = \cdots = a_{m-1} = 0$, and $a_m \neq 0$, then α is called a zero of f of order m. Zeros of order one and two are called simple and double zero respectively. α is called a zero of *infinite order* if $f^{(k)}(\alpha) = 0$ for all $k \geq 0$. Our first result shows that non-constant functions analytic in a domain can not have zero of infinite order.

Theorem 12.1.1. Let f be analytic in a domain D. If f has a zero of infinite order at $\alpha \in D$, then $f \equiv 0$.

Proof. Let $Z = \{z_0 \in D : \ f \text{ has an infinite order zero at } z_0\}$. We show that Z is clopen in D. Let $Z_k = \{z_0 \in D : \ f^{(k)}(z_0) = 0\}$. Then each Z_k is closed and so

$$Z = \cap_k Z_k$$

is closed. To show that Z is open, $z_0 \in Z$. We know that f agrees with its Taylor series on an open disk D' centered at z_0

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z-z_0)^k = 0$$

on D'. Since $f \equiv 0$, so $D' \subset Z$. Thus, Z is open. Bur since D is connected, Z is either empty or $Z = D$. But, we have $\alpha \in Z$. So $Z = D$. Hence $f \equiv 0$. □

Corollary 12.1.2. A nonconstant analytic function on a domain D only has zeros of finite order in D.

So, in general, the zeros of an analytic function is finite ordered. Motivation can be had from the structure of polynomials. We know that any complex polynomial can be factorised completely. The following result gives a general idea to factorise any general analytic function f by just looking at the Taylor's series expansion of f near their zeros.

Theorem 12.1.3. Let α be a zero of f of order m. Then, near α, $f(z) = (z-\alpha)^m \phi(z)$, where ϕ is analytic at α and $\phi(\alpha) \neq 0$.

Proof. Since f is analytic at α and α is a zero of f of order m, so near α, f has a Taylor series expansion

$$\begin{aligned} f(z) &= a_m(z-\alpha)^m + a_{m+1}(z-\alpha)^{m+1} + \cdots \\ &= (z-\alpha)^m(a_m + a_{m+1}(z-\alpha) + \cdots), \end{aligned}$$

where $a_m \neq 0$. Now, let

$$\phi(z) = a_m + a_{m+1}(z-\alpha) + \cdots .$$

Then ϕ is analytic at α with $\phi(\alpha) = a_m \neq 0$. □

The converse of the above theorem is also true. That is, if near a point α, f is of the form

$$f(z) = (z-\alpha)^m \phi(z),$$

where ϕ is analytic at α and $\phi(\alpha) = a_m \neq 0$, then f has a zero of order m at α. (Prove it!)

Exercise 12.1.4. 1. Prove the above statement.

2. Find the zeros of each of the following functions. Also mention the order of each of the zeros with proper justifications:

   a. $e^z - 1$; b. $\sin z$;

   c. $\dfrac{\sin z}{z}$; d. $\dfrac{(z+3)}{z^2+z-6}$;

   e. $(z-3)^2 \cos z$ f. $\cos z - e^z + z$

3. Let f and g has zero of order m and n respectively at z_0. What can be said about the functions $f \pm g$ and fg at z_0?

12.1.2 Singularities and their classification

For a function f, a point α is called a regular point of f if f is analytic at α. The points where f is not analytic are called the *singular points* or *singularities* of f. Let a point α be a singular point of f. If f is analytic in a deleted neighbourhood of α, then α is called an *isolated singularity* of f. If α is not an isolated singularity of f, then it is called *non-isolated singularity* of f.

Let α be an isolated singularity of f. Then we have a positive real number r, such that f has a *Laurent series* expansion of the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z-\alpha)^n + \sum_{n=1}^{\infty} b_n (z-\alpha)^{-n}$$

valid in $0 < |z-\alpha| < r$. a_n and b_n are given as follows:

$$a_n = \frac{1}{2\pi i} \int_\gamma \frac{f(z)dz}{(z-\alpha)^{n+1}} \text{ and } b_n = \frac{1}{2\pi i} \int_\gamma \frac{f(z)dz}{(z-\alpha)^{-n+1}},$$

where γ is a simple closed curve in $0 < |z-\alpha| < r$. The sum

$$\sum_{n=0}^{\infty} a_n (z-\alpha)^n$$

is called the analytic part and the sum

$$\sum_{n=1}^{\infty} b_n (z-\alpha)^{-n}$$

is called the *principal part* of f at the isolated singularity α. If the principal part is a terminating series, then the point α is called the pole of f. α is called the pole of f of order m if $b_m \neq 0$ and $b_{m+1} = b_{m+2} = \cdots = 0$. Poles of order one and two are called *simple pole* and *double pole* respectively. If the principal part is non-terminating, then α is called an *essential singularity* of f. And if all the coefficients b_n of the principal part are zero then α is called removable singularity of f.

Theorem 12.1.5. α is a pole of a function f of order m if and only if f can be written in the form

$$f(z) = \frac{\phi(z)}{(z-\alpha)^m}$$

near α, where, ϕ is analytic at α and $\phi(\alpha) \neq 0$.

Proof. If α is a pole of order m, then

$$f(z) = \sum_{n=0}^{\infty} a_n (z-\alpha)^n + \sum_{n=1}^{m} b_n (z-\alpha)^{-n},$$

where $b_m \neq 0$ and $b_k = 0$ for $k > m$. This gives,

$$\begin{aligned} f(z) &= \left(\sum_{n=0}^{\infty} a_n (z-\alpha)^{n+m} + b_1 (z-\alpha)^{m-1} + \cdots + b_m \right) (z-\alpha)^{-m} \\ &= \frac{\phi(z)}{(z-\alpha)^m} \end{aligned}$$

where

$$\phi(z) = \sum_{n=0}^{\infty} a_n (z-\alpha)^{n+m} + b_1 (z-\alpha)^{m-1} + \cdots + b_m$$

and $\phi(\alpha) = b_m \neq 0$.

Conversely, if

$$f(z) = \frac{\phi(z)}{(z-\alpha)^m},$$

where ϕ satisfies the given criteria, then, due to analyticity of ϕ at α,

$$\begin{aligned} \phi(z) &= \frac{\sum_{n=0}^{\infty} a_n (z-\alpha)^n}{(z-\alpha)^m} \\ &= \frac{a_0}{(z-\alpha)^m} + \frac{a_1}{(z-\alpha)^{m-1}} + \cdots + a_m + a_{m+1}(z-\alpha) + a_{m+2}(z-\alpha)^2 + \cdots \\ &= \frac{a_0}{(z-\alpha)^m} + \frac{a_1}{(z-\alpha)^{m-1}} + \cdots + \frac{a_{m-1}}{(z-\alpha)} + \sum_{n=0}^{\infty} a_{m+n} (z-\alpha)^n. \end{aligned}$$

Hence, α is a pole of f of order m, since $a_0 \neq 0$. □

Exercise 12.1.6. 1. Find the singularities of each of the following functions and classify them. For poles, specify their orders.

a. $\dfrac{e^z - 1}{z^2}$

b. $\dfrac{\cos z}{z^2}$

c. $\dfrac{\sinh z}{z^4}$

d. $\dfrac{1 - \cos z}{z^2}$

e. $\dfrac{\sin z}{z}$

f. $\dfrac{z}{e^z - 1}$

g. $\exp\left(\dfrac{1}{z}\right)$

h. $(z-1)\cos\left(\dfrac{1}{z}\right)$.

2. If f, g have poles of order m, n respectively at z_0, then their pointwise product fg has a pole of order $m+n$ at z_0.

3. Suppose a function f has a pole of order m at z_0 and a function g has a zero at z_0 of order n. What can be said about the functions $f \pm g$, fg and f/g at z_0?

4. Give an example of a function holomorphic in all of $\mathbb{C}$ except for essential singularities at the two points 0 and 1.

5. Prove or disprove: If f and g have a pole and an essential singularity respectively at the point z_0, then fg has an essential singularity at z_0.

Theorem 12.1.7. The zeros of an analytic function $f(\not\equiv 0)$ are isolated points, that is, if α is a zero of f, then there exists a neighbourhood of α which contains no other zeros of f unless $f \equiv 0$.

OR

If f is an analytic function on a domain D and $f(\alpha) = 0$ for some $\alpha \in D$, then there exists a neighbourhood of α where $f(z) \neq 0$.

Proof. Let α be a zero of f of order m. Then $f(z) = (z-\alpha)^m \phi(z)$, where ϕ is analytic at α and $\phi(\alpha) \neq 0$. Take $\epsilon = \frac{1}{2}|\phi(\alpha)|$. Since ϕ is continuous at α, there exists a $\delta > 0$ such that

$$|\phi(z) - \phi(\alpha)| < \epsilon \text{ for } |z-\alpha| < \delta.$$

Hence,

$$\begin{aligned} |\phi(z)| &= |\phi(\alpha) + \phi(z) - \phi(\alpha)| \\ &\geq |\phi(\alpha)| - |\phi(z) - \phi(\alpha)| \\ &> |\phi(\alpha)| - \frac{1}{2}|\phi(\alpha)| = \frac{1}{2}|\phi(\alpha)| \end{aligned}$$

for $|z-\alpha| < \delta$. Hence, $\phi(z) \neq 0$ in $|z-\alpha| < \delta$. Since $f(z) = (z-\alpha)^m \phi(z)$, it follows that $f(z) \neq 0$ in $0 < |z-\alpha| < \delta$. Hence, α is an isolated zero of f. □

The above theorem is important since it shows that the number of zeros of f in a "small" area can not be "numerous". Also notice that in the course of proving the above theorem, we have shown the following.

Theorem 12.1.8. If f is analytic at α and $f(\alpha) \neq 0$, then there exists a neighbourhood of α in which $f(z) \neq 0$.

Proof. Same as done in the previous theorem. □

One of the most important applications of theorem 12.1.7 is perhaps the interior uniqueness theorem or identity theorem, which completely characterizes an analytic function in a **domain** D by its behaviour on a subset of D. We have highlighted the term domain because the domain, which is an open connected set, plays an important role in proving the theorem. The statement is given as follows.

Theorem 12.1.9. Let f and g be analytic in a domain D such that $f(z) = g(z)$ on a set $S \subset D$ having a limit point $z_0 \in D$. Then $f(z) \equiv g(z)$, for all $z \in D$.

Proof. Let $\phi(z) = f(z) - g(z)$. Then ϕ is analytic in D and $\phi(z) = 0$ on S. Hence, $S \subset Z(\phi)$, where $Z(\phi)$ is the set of zeros of ϕ in D. Since z_0 is a limit point of S, then there exists a sequence $\{z_n\}$ of zeros of ϕ in D such that $z_n \to z_0$. Since $\phi(z_n) = 0$ for all n and ϕ is continuous at z_0, so we have $\phi(z_0) = \lim_{n\to\infty} \phi(z_n) = 0$ so that z_0 is a zero of ϕ in D. Since the zeros of a function are isolated points, this can happen only when $\phi(z) \equiv 0$ in a neighbourhood of z_0. We now split D into two sets A and B, where $A = \{\alpha \in D : \ \alpha \text{ is a limit point of } Z(\phi)\}$ and $B = D \setminus A$. Then $D = A \cup B$, where $A \cap B = \emptyset$ and A is non-empty since $z_0 \in A$. Let α be an arbitrary point of A. Then α is a limit point of $Z(\alpha)$ and hence $\phi(z) \equiv 0$ in a neighbourhood of α. Hence, α is an interior point of A and so A is open. We now show that B is also open. Let $\beta \in B$. Then β is not a limit point of $Z(\phi)$. Since ϕ is continuous at β, there exists a $\delta > 0$ such that $\phi(z) \neq 0$ in $\{z : \ |z - \beta| < \delta\} \subset D$. Thus, β is an interior point of B and hence B is open. Since D is connected, it can not be written as the union of two disjoint non-empty open sets. Hence, we must have either $A = \emptyset$ or $B = \emptyset$. Since $A \neq \emptyset$ we must have $B = \emptyset$. Thus, $A = D$. So every point of D is a limit point of $Z(\phi)$ and hence $Z(\phi) = D$, that is $\phi(z) = 0$ for all $z \in D$, that is $f(z) \equiv g(z)$ on D. □

We see that the connectedness of D was used to prove the uniqueness of f. The following is an example to illustrates that this connectedness of D is necessary for the validity of the above theorem.

Example 12.1.10. Let $D = \mathbb{C} \setminus \{z : \ 1 \leq |z| \leq 3\}$ and let $f : D \to \mathbb{C}$ be defined by

$$\begin{aligned} f(z) &= 0 \text{ for } |z| < 1 \\ &= 2 \text{ for } |z| > 3. \end{aligned}$$

Also let g be another function defined on D such that $g(z) = 0$ for all $z \in D$. Then $f(z) = g(z)$ on $|z| < 1$ having closure $|z| \leq 1$. Of this, $\{z : \ |z| < 1\} \subset D$, but $f \not\equiv g$ on D.

A few applications of the above result can be listed as follows:

1. Let f and g be two analytic functions defined on a domain D and let $\{z_n\}$ be a sequence of points in D having a limit point in D. If $f(z_n) = g(z_n)$ for each $n \in \mathbb{N}$, then $f(z) \equiv g(z)$ on D. For example, let f and g be two functions defined on the domain $D = \{z : \ |z| < 1\}$ such that $f\left(\frac{1}{n}\right) = g\left(\frac{1}{n}\right)$. Then $f \equiv g$ on D.

2. Trigonometric identities like $\sin^2 x + \cos^2 x = 1$ can be extended to the complex plane using the identity theorem.

Example 12.1.11. Suppose f is entire and $f(\mathbb{R}) \subset \mathbb{R}$. Then $\overline{f(z)} = f(\overline{z})$ for all $z \in \mathbb{C}$. It is easy to show that $g(z) = \overline{f(\overline{z})}$ is analytic if f is so (Prove it!). If $x \in \mathbb{R}$, then

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

since x and $f(x)$ are real. So, $g\big|_{\mathbb{R}} = f\big|_{\mathbb{R}}$, and hence by identity theorem, $f \equiv g$.

Exercise 12.1.12.

1. Show that for any complex numbers z_1 and z_2, $\mathrm{e}^{z_1+z_2} = \mathrm{e}^{z_1}\,\dot{\mathrm{e}}^{z_2}$.

2. Let f be a non-constant analytic function defined on the domain $D = \{z : \ |z| \leq 2\}$ such that $f\left(\frac{n}{2n+1}\right) = 0$. Does such a function exist on D? Justify your answer.

3. Let f be an analytic function defined on the open unit circle such that $f\left(\frac{1}{n}\right) = \frac{1}{n^2}$. Find f.

Sample Questions

1. Show that an analytic function f defined on a domain D having an infinite ordered zero is identically equal to zero.
2. Show that a function f analytic on a domain D has a zero of order m at $\alpha \in D$ if and only if it can be represented as
$$f(z) = (z-\alpha)^m \phi(z),$$
near α where ϕ is analytic at α with $\phi(\alpha) \neq 0$.
3. Find the Laurent series expansion of $f(z) = (1-z)\cos\left(\frac{1}{z}\right)$ about the point $z = 0$. Hence show that f has an essential singularity at 0.
4. Show that a point α is a pole of a function f of order m if and only if f can be written in the form
$$f(z) = \frac{\phi(z)}{(z-\alpha)^m}$$
near α, where, ϕ is analytic at α and $\phi(\alpha) \neq 0$.
5. Show that the zeros of a non-constant analytic function f is an isolated point.
6. If f is analytic at α and $f(\alpha) \neq 0$, then show that exists a neighbourhood of α in which $f(z) \neq 0$.
7. State and prove the interior uniqueness theorem. Is the result true for any arbitrary domain? Justify your answer.
8. If possible, find a function f analytic in the unit disc $\{z : |z| < 1\}$ satisfying the following relation
$$f\left(\frac{1}{n}\right) = \frac{(-1)^n}{n},$$
for each positive integer n.
9. Let f, g be analytic functions in a domain D. Which of the following conditions imply $f \equiv g$ on D?
   (a) There is a sequence $\{z_n\}$ of distinct points in D such that $f(z_n) = g(z_n)$ for all $n \in \mathbb{N}$.
   (b) There is a convergent sequence $\{z_n\}$ of distinct points in D with its limit in D such that $f(z_n) = g(z_n)$ for all $n \in \mathbb{N}$.
   (c) γ is a smooth path in D joining distinct points $a,\ b \in D$ and $f = g$ on γ.
   (d) $w \in D$ is such that $f^{(k)}(w) = g^{(k)}(w)$ for all $n \geq 0$.
10. Suppose that f is an entire function, and that in every power series (that is, for every $z_0 \in \mathbb{C}$) $f(z) = \sum_{n=0}^{\infty} c_n (z-z_0)^n$, at least one coefficient is zero. Prove that f is a polynomial.

Unit 13

Course Structure

- Limit point of zeros and poles
- Characteristics of the singularities
- Behaviour of a function at the point at infinity

13.1 Introduction

The previous unit gave us an introduction to the zeros and the types of singularities of a function f. In this unit, we will be mainly concerned with the behaviour of f near the singularities. Also, from the previous unit, we have got the idea that the zeros and poles of f are isolated. This means that the zeros of any analytic function can not be limit points of zeros of f. Also, the poles can also not be a limit point of poles. So, what do we call the limit points of the zeros and poles of f? We will explore these concepts in this unit.

Objectives

After reading this unit, you will be able to

- identify the limit points of zeros and poles of a function as essential singularities
- know the type of singularity of a function by just analysing its behaviour near that point
- classify functions depending upon its behaviour at the point at infinity

13.1.1 Limit points of Zeros and poles

Theorem 13.1.1. If α is a pole of f, then $f(z) \to \infty$ as $z \to \alpha$.

Proof. Let α be a pole of f of order m. Then,

$$f(z) = \frac{\phi(z)}{(z-\alpha)^m},$$

where, ϕ is analytic at α and $\phi(\alpha) \neq 0$. So,

$$|f(z)| = \frac{\phi(z)}{|z-\alpha|^m}.$$

Since ϕ is continuous at α, so for

$$\epsilon = \frac{1}{2}|\phi(\alpha)| > 0,$$

we can find a $\delta > 0$, such that

$$|\phi(z) - \phi(\alpha)| < \epsilon$$

for $|z-\alpha| < \delta$. Hence,

$$\begin{aligned} |\phi(z)| &= |\phi(z) - \phi(\alpha) + \phi(\alpha)| \\ &\geq |\phi(\alpha)| - \frac{1}{2}|\phi(\alpha)| \\ &= \frac{1}{2}|\phi(\alpha)|, \end{aligned}$$

for $|z-\alpha| < \delta$. Hence,

$$|f(z)| > \frac{\frac{1}{2}|\phi(\alpha)|}{|z-\alpha|^m}$$

for $0 < |z-\alpha| < \delta$. Let $M > 0$ be a large number. Then $|f(z)| > M$ if

$$\begin{aligned} &\frac{\frac{1}{2}|\phi(\alpha)|}{|z-\alpha|^m} > M, \\ \text{or,} \quad &0 < |z-\alpha|^m < \frac{\frac{1}{2}|\phi(\alpha)|}{M} \\ \text{or,} \quad &0 < |z-\alpha| < \frac{\frac{1}{2}|\phi(\alpha)|^{\frac{1}{m}}}{M}. \end{aligned}$$

This shows that $f(z) \to \infty$ as $z \to \alpha$. □

Theorem 13.1.2. Limit point of the zeros of a function f which is not identically equal to zero is an essential singularity of the function.

Proof. Let α be a limit point of the zeros of f. Then, by definition, every neighbourhood of α contains infinitely many zeros of f. If possible, let f be analytic at α. Then f is continuous at the point α. So, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(z) - f(\alpha)| < \epsilon$ for $|z-\alpha| < \delta$. But, there is an infinite number of zeros of f in $|z-\alpha| < \delta$. For these zeros, we must have, $|f(\alpha)| < \epsilon$, that is, $f(\alpha) = 0$. Hence, α is a zero of f. This is impossible unless f is identically equal to zero since zeros are isolated points. Hence f is not analytic at α.

Now, if possible, let α be a pole of f. Then, $f(z) \to \infty$ as $z \to \alpha$, that is, given any number $M > 0$, we can find a $\delta > 0$ such that $|f(z)| > M$ is valid in $0 < |z-\alpha| < \delta$. But the deleted neighbourhood $0 < |z-\alpha| < \delta$ has infinite number of zeros of f so that $|f(z)| > M$ is not true for infinite number of points in $0 < |z-\alpha| < \delta$. Hence α cannot be a pole of f. Thus f has an essential singularity at α. □

Theorem 13.1.3. The limit point of the poles of a function f is a non-isolated essential singularity of f.

Proof. Let α be a limit point of the poles of f. Since every neighbourhood of α contains infinite number of poles of f, α cannot be a regular point of f. Hence α is a singularity of f which is non-isolated. Hence the result. □

The above results are very important since it gives and idea of the number of zeros of an analytic function over a bounded domain. Also, functions which are analytic except for poles can not have infinite number of zeros in a bounded domain. But, if we consider the whole complex plane then there are analytic functions which have infinite number of zeros in $\mathbb{C}$. Say, for example, the function $f(z) = \sin z$. The zeros of f are n, $n = 0, 1, 2, \ldots$. The limit point of the zeros is infinity. Hence, we can say that the point at infinity is an essential singularity of $\sin z$. The behaviour of functions at the point at infinity is discussed in details in the later section.

Theorem 13.1.4. Let f be an analytic function. Then α is a zero of f of order m if and only if α is a pole of $\dfrac{1}{f}$ of order m.

Proof. Let α be a zero of f of order m. Then

$$f(z) = (z-\alpha)^m \phi(z),$$

near α, where ϕ is a function analytic at α and $\phi(\alpha) \neq 0$. Then clearly,

$$\frac{1}{f}(z) = \frac{1}{f(z)} = \frac{1}{(z-\alpha)^m \phi(z)} = \frac{\psi(z)}{(z-\alpha)^m},$$

where

$$\psi(z) = \frac{1}{\phi(z)},$$

is analytic at α and $\psi(\alpha) = \dfrac{1}{\phi(\alpha)} \neq 0$. The converse also holds. □

Corollary 13.1.5. If α is an essential singularity of f, then it is also an essential singularity of $\dfrac{1}{f}$.

Proof. Left as exercise. □

Exercise 13.1.6. Locate the singularities of the following functions in the extended complex plane $\mathbb{C}^* = \mathbb{C} \cup \{\infty\}$

a. $\cos \dfrac{1}{z}$ b. $\dfrac{1}{\mathrm{e}^z - 1}$

c. $\dfrac{z}{\sin \frac{\pi}{z}}$ d. $\dfrac{\mathrm{e}^z}{1+z^2}$

13.1.2 Riemann's Theorem On Removable Singularity

Theorem 13.1.7. If a function f is bounded and analytic in a deleted neighbourhood $0 < |z-\alpha| < \delta$, of a point α, then either f is analytic at α or α is a removable singularity of f.

Proof. Since f is analytic in $0 < |z-\alpha| < \delta$, so there it has a Laurent expansion of the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z-\alpha)^n + \sum_{n=1}^{\infty} b_n (z-\alpha)^{-n} \tag{13.1.1}$$

If $\gamma : |z-\alpha| = \rho$, $\rho < \delta$ then the coefficients b_n are given by

$$b_n = \frac{1}{2\pi i} \int_\gamma \frac{f(z)dz}{(z-\alpha)^{-n+1}} \tag{13.1.2}$$

for $n = 1, 2, \cdots$ Since f is bounded in the deleted neighbourhood, there exists a $M > 0$ such that $|f(z)| \leq M$ in $0 < |z - \alpha| < \delta$. Hence,

$$\begin{aligned} |b_n| &= \frac{1}{2\pi} \left| \int_\gamma \frac{f(z)dz}{(z-\alpha)^{-n+1}} \right| \\ &\leq \frac{M}{2\pi} \rho^{n-1} 2\pi\rho \\ &= M\rho^n. \end{aligned}$$

Since $b_n's$ are constants and ρ can be chosen arbitrarily small, we conclude $b_n = 0$, $n = 1, 2, \cdots$. Hence, equation (13.1.1) reduces to

$$f(z) = \sum_{n=0}^{\infty} a_n (z - \alpha)^n$$

in $0 < |z - \alpha| < \delta$. If $f(\alpha) = a_0$, this power series representation of f is actually valid in $|z - \alpha| < \delta$, and in that case, f is analytic at α. Otherwise f can be made analytic at α by letting $f(\alpha) = a_0$. The point α is then a removable singularity of f. □

13.1.3 Casorati-Weierstrass Theorem

Theorem 13.1.8. Let α be an isolated essential singularity of a function$f$ and let z_0 be any given complex number. Then, for any $\epsilon > 0$, the inequality $|f(z) - z_0| < \epsilon$ is satisfied at some point z in each deleted neighbourhood of α.

Proof. Since α is isolated, there exists a deleted neighbourhood $0 < |z - \alpha| < \delta$ of α where f is analytic. If possible, let

$$|f(z) - z_0| < \epsilon \tag{13.1.3}$$

be not satisfied at any point in that deleted neighbourhood. Then $|f(z) - z_0| \geq \epsilon$ for all z in $0 < |z - \alpha| < \delta$. Let

$$g(z) = \frac{1}{f(z) - z_0},$$

in $0 < |z - \alpha| < \delta$. Then

$$|g(z)| = \left| \frac{1}{f(z) - z_0} \right| \leq \frac{1}{\epsilon}$$

in $0 < |z - \alpha| < \delta$. Hence g is bounded and analytic in $0 < |z - \alpha| < \delta$. By Riemann's Theorem on removable singularity, α is a removable singularity of g. Let $g(\alpha)$ be defined such that g is analytic at α. Since f cannot be a constant function, g is also not a constant function. Since g is analytic at α, it has a Taylor series representation at α. Hence, either $g(\alpha) \neq 0$ or g has a zero at α. So, its reciprocal

$$\frac{1}{g(z)} = f(z) - z_0$$

is either analytic at α or has a pole there which contradicts the hypothesis that α is an essential singularity of f. Hence the result. □

Corollary 13.1.9. If α is an isolated essential singularity of f, then for arbitrary positive numbers δ and M, there is a point z in $0 < |z - \alpha| < \delta$ at which $|f(z)| > M$.

Hence from all the above theorems , we can conclude that

1. If $f(z)$ tends to a finite limit as z tends to α, then α is either a regular point or a removable singularity of f.
2. If $f(z) \to \infty$ as $z \to \alpha$, then α is a pole of f.
3. If $f(z)$ does not tend to any definite limit, finite or infinite, then α is an essential singularity of f.

Example 13.1.10. The function $\mathrm{e}^{\frac{1}{z}}$ has an essential singularity at 0. We will show that it takes on every given non zero $w(= \rho \exp(i\theta)) \in \mathbb{C}$ in any arbitrarily small neighbourhood of 0. Setting $z = r\,\mathrm{e}^{it}$, we need to solve

$$\exp \frac{1}{z} = \exp\left(\frac{\cos t}{r} - i\frac{\sin t}{r}\right) = \rho\,\mathrm{e}^{i\theta}\,.$$

By equating the absolute values, we obtain

$$\frac{\cos t}{r} = \log \rho.$$

On the other hand, by looking at arguments, we see that a solution is given when

$$-\frac{\sin t}{r} = \theta.$$

Using $\cos^2 t + \sin^2 t = 1$, we have

$$r = \frac{1}{\sqrt{(\log \rho)^2 + \theta^2}}.$$

But we are allowed to increase θ by integral multiple of 2π, without changing w. Bearing this in mind, it is clear from the above expression for r that we can make r as small as we please.

Exercise 13.1.11.
1. Prove, using the Casorati-Weierstrass Theorem, that if f has an essential singularity at z_0, and if w is any complex value whatever, then there exists a sequence $\{z_n\}$ such that

$$\lim_{n\to\infty} z_n = z_0 \text{ and } \lim_{n\to\infty} f(z_n) = w.$$

2. Suppose f is analytic in the punctured disk $0 < |z| < 1$ except for poles $\{z_n\}$ converging to 0. Show that the range of f in the punctured disk is dense in the complex plane.
3. Let f be an entire function which is not a polynomial. If B is a bounded set then show that the image of $\mathbb{C} \setminus B$ is dense in $\mathbb{C}$.

13.1.4 Behaviour of a function at the point at infinity

As we have already seen before, the function $\sin z$ has an essential singularity at ∞. We have deduced this from the behaviour of the zeros of $\sin z$. However, what can be said about the zeros of any complex function f in general. Let us look at $\sin z$ and consider the point $z = 0$. Clearly, $z = 0$ is a zero of $\sin z$. Also, let us consider the Taylor's series expansion of $\sin z$ about $z = 0$.

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \dots$$

What if we change the variable z to a new variable w, where $w = \dfrac{1}{z}$. Then, the above equation will be transformed into

$$g(w) = \sin \frac{1}{w} = \frac{1}{w} - \frac{1}{3!.w^3} + \frac{1}{5!.w^5} - \frac{1}{7!.w^7} + \dots$$

As $z \to 0$, $w \to \infty$. So, the above equation give the expansion of the function g near infinity. Since the principal part of the series is non-terminating. This implies that 0 is an essential singularity of g and hence, ∞ is an essential singularity of $\sin z$. Thus, to understand the behaviour of a function f near infinity, we need to understand the behaviour of f near 0. We formalise this in the next paragraph.

Definition 13.1.12. The point $z = \infty$ is called an isolated singularity of a function f if f is analytic in the exterior of a disc $\{z \in \mathbb{C} : \ |z| > R\}$.

This is quite natural because through stereographic projection, the region $\{z \in \mathbb{C} : \ |z| > R\}$ corresponds to a punctured disc on the sphere centered at the north pole.

Also, $z = \infty$ is an isolated singularity of $f(z)$ if and only if $z = 0$ is an isolated singularity of $g(z) = f\left(\frac{1}{z}\right)$. We can use the following definitions to classify the singularities at $z = \infty$.

Definition 13.1.13. Let $z = \infty$ be an isolated singularity of $f(z)$.

1. $f(z)$ has removable singularity at $z = \infty$ if $f\left(\frac{1}{z}\right)$ has a removable singularity at $z = 0$;
2. $f(z)$ has a pole of order $m \geq 1$ at $z = \infty$ if $f\left(\frac{1}{z}\right)$ has a pole of order $m \geq 1$ at $z = 0$;
3. $f(z)$ has an essential singularity at $z = \infty$ if $f\left(\frac{1}{z}\right)$ has an essential singularity at $z = 0$.

The following result gives an idea of the series expansion of f near the point at infinity.

Theorem 13.1.14. Let $f(z)$ be a function of a complex variable z.

1. If $z = \infty$ is an isolated singularity of $f(z)$, then
$$f(z) = \sum_{n=-\infty}^{\infty} a_n z^n \ \ (|z| > R),$$
where R is a positive number.
2. If $z = \infty$ is a removable singularity of $f(z)$, then $a_n = 0$ for all $n > 0$:
$$f(z) = \sum_{n=-\infty}^{0} a_n z^n \ \ (|z| > R).$$
3. If $z = \infty$ is a pole of $f(z)$ of order $m \geq 1$, then $a_m \neq 0$ and $a_n = 0$ for all $n > m$:
$$f(z) = \sum_{n=-\infty}^{m} a_n z^n \ \ (|z| > R).$$
4. If $z = \infty$ is an essential singularity of $f(z)$, then $a_n \neq 0$ for infinitely many positive integers n.

Proof. Left as exercise. □

Example 13.1.15.
1. $f(z) = z^3$ has a pole of order 3 at infinity.
2. e^z has an essential singularity at infinity.

3. $e^{\frac{1}{z}}$ has a removable singularity at infinity.

Having discussed the singularities, let us now discuss when ∞ can be a zero of f. A motivation can be drawn by studying the behaviour of f near 0. Consider the example of $f(z) = \sin\frac{1}{z}$. The Laurent expansion of f about $z = 0$ is given by

$$\sin\frac{1}{z} = \frac{1}{z} - \frac{1}{3!.z^3} + \frac{1}{5!.z^5} - \frac{1}{7!.z^7} + \dots$$

Replacing $\frac{1}{z}$ by w we get

$$\sin w = w - \frac{w^3}{3!} + \frac{w^5}{5!} - \dots$$

which indicates a zero at $w = 0$. We now formalise the definition using the Laurent series expansion of f near ∞.

Definition 13.1.16. Let $f(z)$ be a function of complex variable and let the Laurent series expansion of f near $z = \infty$ be given as follows:

$$f(z) = \sum_{n=-\infty}^{0} a_n z^n \ \ (|z| > R).$$

Then $z = \infty$ will be called a zero of f of order m if $a_{-m} \neq 0$ and $a_n = 0$ for all $n > -m$. The Laurent series will then become

$$f(z) = \sum_{n=-\infty}^{-m} a_n z^n \ \ (|z| > R).$$

We will conclude this unit by considering the zeros and singularities of rational functions at infinity. We know that a function f is called a rational function if

$$f(z) = \frac{p(z)}{q(z)},$$

where both $p(z)$ and $q(z)$ are polynomials in $\mathbb{C}$ of degrees m and n respectively. Then $z = \infty$ is

1. a zero of order $n - m$ of $f(z)$ if $n > m$;
2. a removable singularity of $f(z)$ if $n = m$;
3. a pole of $f(z)$ of order $m - n$ if $n < m$.

Exercise 13.1.17.

1. Show that an entire function f has a removable singularity at infinity if and only if f is constant.
2. Show that an entire function f has a pole of order m at infinity if and only if f is a polynomial of degree m.
3. Characterise those rational functions which have removable singularity at infinity.
4. Characterise those rational functions which have a pole of order m at infinity.
5. Discuss about the essential singularity of a rational function at infinity.

Sample Questions

1. Show that if α is a pole of f, then $f(z) \to \infty$ as $z \to \alpha$.
2. Show that the limit points of the zeros of a function f is an essential singularity unless $f \equiv 0$.
3. State and prove Riemann's theorem on removable singularity.
4. State and prove Casorati Weierstrass theorem.

Unit 14

Course Structure

- Theory of Residues, Argument Principle
- Rouche's Theorem
- Maximum Modulus Theorem, Schwarz Lemma

14.1 Introduction

The inspiration behind this unit is a desire for an answer to the following question: if f has an isolated singularity at $z = \alpha$, what are the possible values for $\int_\gamma f$ where γ is a simple closed curve not passing through α? If the singularity is removable, then clearly the integral will be zero. If $z = \alpha$ is a pole or an essential singularity, then the answer is not always zero. We investigate the problem by introducing the theory of residues. Residue theory is also very instrumental in solving certain real integrals as we shall explore later.

Objectives

After reading this unit, you will be able to

- define the residue of a function and state the Cauchy's Residue theorem
- solve real integrals using the residue theorem
- state the maximum modulus theorem and study its applications

14.1.1 Theory of Residues

Definition 14.1.1. Let f has an isolated singularity at $z = \alpha$ and let

$$f(z) = \sum_{n=0}^{\infty} a_n(z-\alpha)^n + \sum_{n=1}^{\infty} b_n(z-\alpha)^{-n}$$

be its Laurent expansion about $z = \alpha$. Then the residue of f at $z = \alpha$ is the coefficient b_1. We denote it as $Res(f;\alpha) = b_1$.

We know that

$$b_n = \frac{1}{2\pi i}\int_\gamma \frac{f(z)dz}{(z-\alpha)^{-n+1}}$$

where γ is any simple closed curve lying in a deleted neighbourhood $0 < |z-\alpha| < r$ of α and $\alpha \in \text{Int}\gamma$. So for $n = 1$, we have

$$b_1 = \frac{1}{2\pi i}\int_\gamma f(z)dz$$

Hence,

$$\int_\gamma f(z)dz = 2\pi i b_1$$

Theorem 14.1.2. If α be a pole of f of order m, then

$$Res(f;\alpha) = \frac{1}{(m-1)!}\lim_{z\to\alpha}\frac{d^{m-1}}{dz^{m-1}}(z-\alpha)^m f(z)$$

Proof. Since α is a pole of f of order m,

$$f(z) = \phi(z) + \frac{b_1}{z-\alpha} + \ldots + \frac{b_m}{(z-\alpha)^m},$$

where ϕ is analytic at α and $b_m \neq 0$. Hence,

$$(z-\alpha)^m f(z) = (z-\alpha)^m\phi(z) + b_1(z-\alpha)^{m-1} + \ldots + b_m.$$

Differentiating the above equation with respect to z $m-1$ times we get,

$$\frac{d^{m-1}}{dz^{m-1}}\left\{(z-\alpha)^m f(z)\right\} = \frac{d^{m-1}}{dz^{m-1}}\left\{(z-\alpha)^m\phi(z)\right\} + b_1(m-1)!$$

Since $\phi(z)$ is analytic at α, so taking limit as $z \to \alpha$ to the above equation, we get

$$\lim_{z\to\alpha}\frac{d^{m-1}}{dz^{m-1}}\left\{(z-\alpha)^m f(z)\right\} = 0 + b_1(m-1)!$$

Hence, the result. □

Corollary 14.1.3. If α is a simple pole of f, then $Res(f;\alpha) = \lim_{z\to\alpha}(z-\alpha)f(z)$. Also, if $m = 2$, then

$$Res(f;\alpha) = \lim_{z\to\alpha}\frac{d}{dz}\left\{(z-\alpha)^2 f(z)\right\}.$$

Exercise 14.1.4. 1. Find the residue of the following functions at each of the poles:

a. e^z b. $e^{\frac{1}{z}}$ c. $\frac{\sin z}{z^2}$ d. $\frac{1}{z(z^2+1)(z-2)^2}$

d. $\cot z$

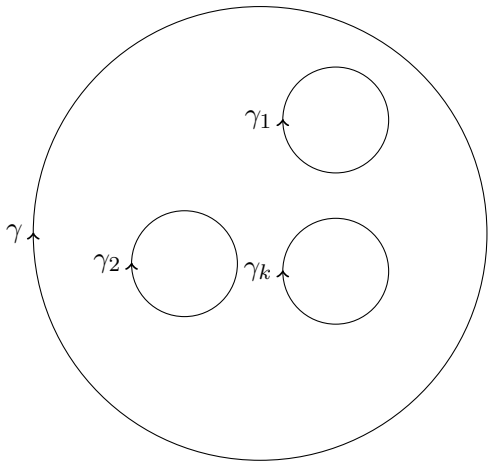
2. Let p, q be analytic at $z = z_0$. Assume $p(z_0) \neq 0$, $q(z_0) = 0$, $q'(z_0) \neq 0$. Find

$$Res\left(\frac{p(z)}{q(z)}; z_0\right).$$

3. Find the residue of f at $z = 0$ where
$$f(z) = \frac{\sinh z\, \mathrm{e}^z}{z^5}.$$

Theorem 14.1.5. (Cauchy's Residue Theorem) If f is analytic within and on a simple closed curve γ except for a finite number of singular points $a_1, a_2, \cdots, a_n$ within γ, then
$$\int_\gamma f(z)dz = 2\pi i \sum_{k=1}^{n} Res(f; a_k)$$

Proof. Since there are finite number of singularities, they must be isolated. Round each singularity a_k, $k = 1, 2, \cdots, n$, we draw a circle γ_k with radius so small that these n circles do not intersect each other and they all lie within γ. Then f becomes analytic within and on the multiply connected region bounded by γ, γ_1, ...,



γ_n. According to Cauchy Goursat Theorem, extended to such regions, we have,
$$\begin{aligned}\int_\gamma f(z)dz &= \sum_{k=1}^{n}\int_{\gamma_k} f(z)dz \\ &= 2\pi i.Res(f; a_1) + 2\pi i.Res(f; a_2) + \ldots + 2\pi i.Res(f; a_n) \\ &= 2\pi i \sum_{k=1}^{n} Res(f; a_k)\end{aligned}$$
all the integrals taken in the positive sense. □

The theory of residues can be used to evaluate improper real integrals. We consider the following cases:

1. When we evaluate the integral of the form
$$\int_0^{2\pi} f(\cos\theta, \sin\theta)d\theta,$$
where $f(\cos\theta, \sin\theta)$ is a real rational function of $\cos\theta$ and $\sin\theta$, we use the transformation $z = \mathrm{e}^{i\theta}$ and choose the contour C as the unit circle $|z| = 1$. Then,
$$\cos\theta = \frac{1}{2}\left(z + \frac{1}{z}\right), \quad \sin\theta = \frac{1}{2i}\left(z - \frac{1}{z}\right)$$
and,
$$dz = i\,\mathrm{e}^{i\theta}\, d\theta \Rightarrow d\theta = \frac{dz}{iz}.$$

Example 14.1.6. Let

$$\begin{aligned} I &= \int_0^{2\pi} \frac{\cos^2 3\theta}{5-4\cos 2\theta} d\theta \\ &= \frac{1}{2}\int_0^{2\pi} \frac{1+\cos 6\theta}{5-4\cos 2\theta} d\theta \\ &= \text{Real part of } \frac{1}{2}\int_0^{2\pi} \frac{1+e^{i6\theta}}{5-4\cos 2\theta} d\theta. \end{aligned} \tag{14.1.1}$$

Let

$$I_1 = \int_0^{2\pi} \frac{1+e^{i6\theta}}{5-4\cos 2\theta} d\theta.$$

Let $z = \mathrm{e}^{i\theta}$. Then, $dz = i\,\mathrm{e}^{i\theta}\,d\theta$. Also we have,

$$\begin{aligned} \cos 2\theta &= \cos^2\theta - \sin^2\theta \\ &= \frac{1}{4}\left(z+\frac{1}{z}\right)^2 + \frac{1}{4}\left(z-\frac{1}{z}\right)^2 \\ &= \frac{1}{2}\left(z^2+\frac{1}{z^2}\right) \end{aligned}$$

We take $C : |z| = 1$. Then we will get,

$$\begin{aligned} I_1 &= \int_C \frac{1+z^6}{5-2\left(z^2+\frac{1}{z^2}\right)} \frac{dz}{iz} \\ &= \frac{1}{i}\int_C \frac{z(1+z^6)}{5z^2-2z^4-2} dz \\ &= -\frac{1}{i}\int_C \frac{z(1+z^6)}{2z^4-4z^2-z^2+2} dz \\ &= -\frac{1}{i}\int_C \frac{z(1+z^6)}{(2z^2-1)(z^2-2)} dz. \end{aligned} \tag{14.1.2}$$

Let

$$f(z) = \frac{z(1+z^6)}{(2z^2-1)(z^2-2)}.$$

Then the poles of f are at $z = \pm\frac{1}{\sqrt{2}}$ and $z = \pm\sqrt{2}$, of which, $z = \pm\frac{1}{\sqrt{2}}$ lie inside C. Also, the poles are simple. Now,

$$\begin{aligned} Res\left(f; \frac{1}{\sqrt{2}}\right) &= \lim_{z\to\frac{1}{\sqrt{2}}} \frac{\left(z-\frac{1}{\sqrt{2}}\right) z(1+z^6)}{2\left(z+\frac{1}{\sqrt{2}}\right)\left(z-\frac{1}{\sqrt{2}}\right)(z^2-2)} \\ &= \lim_{z\to\frac{1}{\sqrt{2}}} \frac{z(1+z^6)}{2\left(z+\frac{1}{\sqrt{2}}\right)(z^2-2)} \\ &= \frac{\frac{1}{\sqrt{2}}\left(1+\frac{1}{8}\right)}{2.\frac{2}{\sqrt{2}}\left(\frac{1}{2}-2\right)} = -\frac{3}{16} \end{aligned}$$

and similarly,

$$Res\left(f;-\frac{1}{\sqrt{2}}\right)=\lim_{z\to-\frac{1}{\sqrt{2}}}\frac{\left(z+\frac{1}{\sqrt{2}}\right)z(1+z^6)}{2\left(z+\frac{1}{\sqrt{2}}\right)\left(z-\frac{1}{\sqrt{2}}\right)(z^2-2)}=-\frac{3}{16}.$$

Hence, from residue theorem we get from equation (14.1.2),

$$I_1=-\frac{1}{i}.2\pi i.(\text{sum of residues of } f \text{ at poles within } C)=-\frac{1}{i}.2\pi i\left(-\frac{3}{8}\right)=\frac{3\pi}{4}.$$

Hence from equation (14.1.1), we get,

$$I=\text{Real part of }\frac{1}{2}.I_1=\text{Real part of }\frac{1}{2}.\frac{3\pi}{4}=\frac{3\pi}{8}.$$

2. When we want to evaluate integrals of the form

$$\int_{-\infty}^{\infty}f(x)dx$$

where, f is a rational function of the real variable x. The improper integral converges if

(a) the degree of the denominator of f exceeds that of the numerator by more than one, and

(b) the denominator of f does not vanish on the real line.

We assume that f satisfies both the conditions. Also we have the following lemma:

Lemma 14.1.7. If $f(z)\to 0$ uniformly as $z\to\infty$, then

$$\lim_{R\to\infty}\int_{C_1}f(z)dz=0$$

where C_1 is the semi-circle $|z|=R$, $\text{Im}z>0$.

If

$$f(z)=\frac{a_0+a_1z+\cdots+a_nz^n}{b_0+b_1z+\cdots+b_mz^m},\ a_n\neq 0,\ b_m\neq 0$$

and $m>n+1$, then $zf(z)\to 0$ uniformly as $z\to\infty$. Hence we can apply the above lemma for a function of this type. Now we proceed with the evaluation of the integral $\int_{-\infty}^{\infty}f(x)dx$. We choose a curve C (see figure 14.1.1) consisting of

(a) the semi-circle $C_1:|z|=R$, $\text{Im}z>0$ and

(b) the line segment C_2 of the real axis from $-R$ to R.

Since f is a rational function function, it has a finite number of singularities which are poles, we choose R sufficiently large so that all the poles of f in the upper half plane lie within C. If α_k is a pole of f within C, then by residue theorem, we have

$$2\pi i.\sum_k Res(f;\alpha_k)=\int_C f(z)dz=\int_{C_1}f(z)dz+\int_{C_2}f(z)dz=\int_{C_1}f(z)dz+\int_{-R}^{R}f(x)dx. \tag{14.1.3}$$

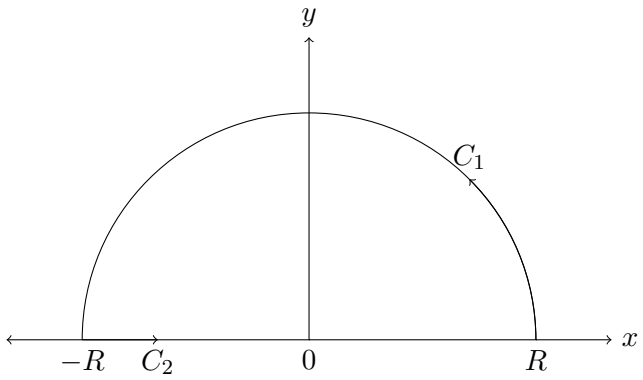


Figure 14.1.1: Structure of C

By the lemma,

$$\lim_{R\to\infty}\int_{C_1} f(z)dz = 0.$$

Hence, taking limit as $R \to \infty$ in equation (14.1.3), we get

$$\int_{-\infty}^{\infty} f(x)dx = 2\pi i.\sum_k Res(f;\alpha_k).$$

Example 14.1.8. Let

$$I = \int_{-\infty}^{\infty} \frac{x^2}{(x^2+1)(x^2+4)}dx.$$

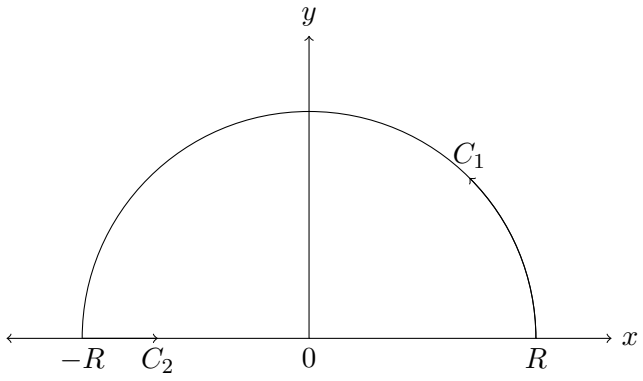
We take

$$f(z) = \frac{z^2}{(z^2+1)(z^2+4)}$$

and consider

$$\int_C f(z)dz$$

where, $C = C_1 + C_2$, $C_1 : z = Re^{i\theta}$, $0 \le \theta \le \pi$ and $C_2 : [-R, R]$, for some real R, sufficiently large.



Poles of f are at $z = \pm i$ and $z = \pm 2i$, of which i and $2i$ lie within C. Now,

$$Res(f;i) = \frac{i}{6}$$

and,

$$Res(f;2i) = -\frac{i}{3}$$

By Residue theorem,

$$\begin{aligned}\int_C f(z)dz &= 2\pi i(Res(f;i) + Res(f;2i)) \\ &= 2\pi i\left(\frac{i}{6} - \frac{i}{3}\right) \\ &= \frac{\pi}{3}. \end{aligned} \tag{14.1.4}$$

Now, since the degree of the denominator of f exceeds that of the numerator by more than one, $zf(z) \to 0$ uniformly as $z \to \infty$. Hence,

$$\lim_{R\to\infty}\int_{C_1} f(z)dz = 0.$$

Also,

$$\int_C f(z)dz = \int_{C_1} f(z)dz + \int_{C_2} f(z)dz = \int_{C_1} f(z)dz + \int_{-R}^{R} f(x)dx = \int_{-R}^{R} \frac{x^2}{(x^2+1)(x^2+4)}dx.$$

Thus, equation (14.1.4) becomes,

$$\int_{-R}^{R} \frac{x^2}{(x^2+1)(x^2+4)}dx = \frac{\pi}{3}$$

Taking $R \to \infty$ in the above equation, we get,

$$\int_{-\infty}^{\infty} \frac{x^2}{(x^2+1)(x^2+4)}dx = \frac{\pi}{3}.$$

3. Suppose we want to evaluate integrals of the form

$$\int_{-\infty}^{\infty} f(z)\cos mxdx \text{ or } \int_{-\infty}^{\infty} f(z)\sin mxdx, \tag{14.1.5}$$

where $m > 0$ and f is a rational function of x. We know that the improper integrals given in equation (14.1.5) would converge if

(a) the degree of the denominator of f exceeds that of the numerator and

(b) the denominator of f does not vanish on the real axis.

We assume that f satisfies both the conditions. We first state the lemma due to Jordan.

Lemma 14.1.9. If $f(z) \to 0$ uniformly as $z \to \infty$, then

$$\lim_{R\to\infty}\int_{C_1} \mathrm{e}^{imz} f(z)dz = 0,\ m > 0$$

where $C_1 : |z| = R$, $\mathrm{Im}z > 0$.

We now consider the evaluation of the integrals in equation (14.1.5). Let C be the closed contour consisting of $C_1 : |z| = R$, $\text{Im}z > 0$, and $C_2 : [-R, R]$ (figure is same as 14.1.1). We choose R sufficiently large so that all the poles of $e^{imz} f(z)$, that is, of $f(z)$, lying in the upper half plane , are contained inside C. Hence, if α_k is a pole of f within γ, by residue theorem,

$$\begin{aligned} 2\pi i. \sum_k Res(e^{imz} f(z); \alpha_k) &= \int_C e^{imz} f(z)dz \\ &= \int_{C_1} e^{imz} f(z)dz + \int_{C_2} e^{imz} f(z)dz \\ &= \int_{C_1} e^{imz} f(z)dz + \int_{-R}^{R} e^{imx} f(x)dx. \end{aligned}$$

Taking $R \to \infty$ and applying Jordan's lemma, we get,

$$\begin{aligned} 2\pi i. \sum_k Res(e^{imz} f(z); \alpha_k) &= \int_{-\infty}^{\infty} e^{imx} f(x)dx \\ &= \int_{-\infty}^{\infty} f(x) \cos mxdx + i \int_{-\infty}^{\infty} f(x) \sin mxdx. \end{aligned}$$

Equating the real and imaginary parts, we get the values of the given integrals.

Example 14.1.10. Let us evaluate the integral

$$\int_{-\infty}^{\infty} \frac{\cos x}{x^2 + a^2} dx, \quad a > 0.$$

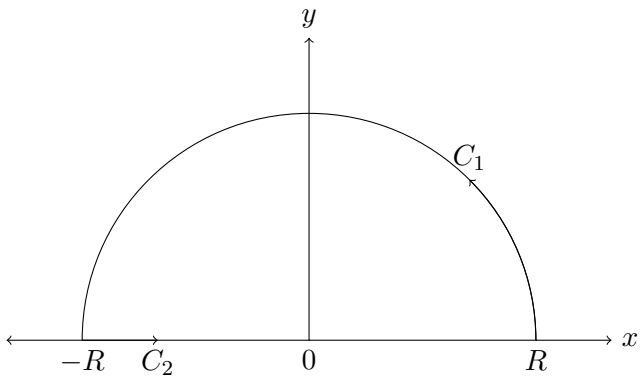
We take

$$f(z) = \frac{1}{x^2 + a^2}$$

and consider the integral

$$\int_C \frac{e^{iz}}{z^2 + a^2} dz,$$

where $C = C_1 + C_2$, $C_1 : |z| = R$, $\text{Im}z > 0$ and $C_2 : [-R, R]$, R being sufficiently large. The poles



of $\dfrac{e^{iz}}{z^2 + a^2}$ are precisely the poles of f. But the poles of f are $\pm ia$ of which only ia lies inside C and

the pole is a simple pole. By residue theorem,

$$\begin{aligned} 2\pi i.Res(\mathrm{e}^{iz} f(z); ia) &= \int_C \mathrm{e}^{iz} f(z)dz \\ &= \int_{C_1} \mathrm{e}^{iz} f(z)dz + \int_{C_2} \mathrm{e}^{iz} f(z)dz \\ &= \int_{C_1} \mathrm{e}^{iz} f(z)dz + \int_{-R}^{R} \mathrm{e}^{ix} f(x)dx. \end{aligned} \tag{14.1.6}$$

Since the degree of the denominator of f exceeds that of the numerator, $f(z) \to 0$ uniformly as $z \to \infty$. Hence, by Jordan's lemma,

$$\lim_{R\to\infty} \int_{C_1} \mathrm{e}^{iz} f(z)dz = 0.$$

Taking limit as $R \to \infty$ in equation (14.1.6), we get

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{\mathrm{e}^{ix}}{x^2+a^2}dx &= 2\pi i.Res(\mathrm{e}^{iz} f(z); ia) \\ &= 2\pi i \lim_{z\to ia} \frac{(z-ia)\,\mathrm{e}^{iz}}{(z-ia)(z+ia)} \\ &= 2\pi i.\frac{\mathrm{e}^{-a}}{2ia} = \frac{\pi\,\mathrm{e}^{-a}}{a}. \end{aligned}$$

Now,

$$\begin{aligned} \frac{\pi\,\mathrm{e}^{-a}}{a} &= \int_{-\infty}^{\infty} \frac{\mathrm{e}^{ix}}{x^2+a^2}dx \\ &= \int_{-\infty}^{\infty} \frac{\cos x}{x^2+a^2}dx + i\int_{-\infty}^{\infty} \frac{\sin x}{x^2+a^2}dx. \end{aligned}$$

Equating the real part, we have

$$\int_{-\infty}^{\infty} \frac{\cos x}{x^2+a^2}dx = \frac{\pi\,\mathrm{e}^{-a}}{a}.$$

4. Suppose we want to evaluate a real definite integral for which the corresponding complex integrand has a simple pole on the real axis. We consider the integral $\int_{-\infty}^{\infty} f(x)dx$, where $f(z)$ has a simple pole at $z = \alpha$ on the real axis. We take the integral $\int_C f(z)dz$, where $C = C_1 + C_2 + C_3 + C_4$, and $C_1 : |z| = R, \mathrm{Im}z > 0; C_2 : [-R, \alpha-\epsilon]; C_3 : |z-\alpha| = \epsilon, \mathrm{Im}z > 0; C_4 : [\alpha+\epsilon, R]$ (see figure 14.1.3). Here, C_1 is positively oriented and C_3 is negatively oriented and R is sufficiently large and ϵ is arbitrarily small. We have indented the contour C at α since α is a simple pole of f. Let k be the sum of the residues of f at its singularities within C. Then by residue theorem,

$$\begin{aligned} 2\pi ik &= \int_C f(z)dz \\ &= \int_{C_1} f(z)dz + \int_{C_2} f(z)dz + \int_{C_3} f(z)dz + \int_{C_4} f(z)dz \\ &= \int_{C_1} f(z)dz + \int_{-R}^{\alpha-\epsilon} f(z)dz + \int_{C_3} f(z)dz + \int_{\alpha+\epsilon}^{R} f(z)dz \\ &= \int_{C_1} f(z)dz + \int_{C_3} f(z)dz + \int_{-\infty}^{\infty} f(x)dx \end{aligned} \tag{14.1.7}$$

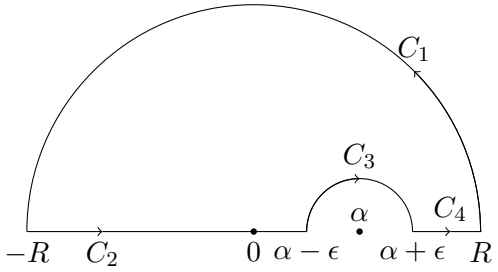


Figure 14.1.2: Structure of C

when $R \to \infty$ and $\epsilon \to 0$. We assume that

$$\lim_{R\to\infty} \int_{C_1} f(z)dz = 0.$$

Also, since α is a simple pole of f, near α f has a Laurent series expansion of the form

$$f(z) = \phi(z) + \frac{a}{z-\alpha},$$

where ϕ is analytic at α and $a = Res(f;\alpha)$. Hence,

$$\int_{C_3} f(z)dz = a\int_{C_3} \frac{dz}{z-\alpha} + \int_{C_3} \phi(z)dz.$$

On C_3, $z - \alpha = \epsilon\, e^{i\theta}$, $0 \le \theta \le \pi$. Hence,

$$\int_{C_3} f(z)dz = a\int_{\pi}^{0} \frac{i\epsilon\, e^{i\theta}\, d\theta}{\epsilon\, e^{i\theta}} + \int_{C_3} \phi(z)dz = -i\pi a + \int_{C_3} \phi(z)dz.$$

Since ϕ is analytic at α, there exists a positive number M such that in a small neighbourhood of α, $|\phi(z)| \le M$. We choose ϵ so small that C_3 lies in this neighbourhood. Hence, $|\phi(z)| \le M$, for all $z \in C_3$. Hence,

$$\left|\int_{C_3} \phi(z)dz\right| \le M\pi\epsilon \to 0 \text{ as } \epsilon \to 0.$$

Therefore,

$$\lim_{\epsilon\to 0} \int_{C_3} f(z)dz = -i\pi a = -i\pi Res(f;\alpha).$$

Now, from equation (14.1.7) we can evaluate the real definite integral $\int_{-\infty}^{\infty} f(x)dx$.

Example 14.1.11. Suppose we want to evaluate the integral

$$\int_0^{\infty} \frac{\sin x}{x} dx.$$

We consider the integral

$$\int_C \frac{e^{iz}}{z} dz = \int_C f(z)dz,$$

where $f(z) = \dfrac{e^{iz}}{z}$ and $C = C_1 + C_2 + C_3 + C_4$, and $C_1 : |z| = R$, $\text{Im}z > 0$; $C_2 : [-R, -\epsilon]$; $C_3 : |z| = \epsilon$, $\text{Im}z > 0$; $C_4 : [\epsilon, R]$. Here, C_1 is positively oriented whereas C_3 is negatively oriented.

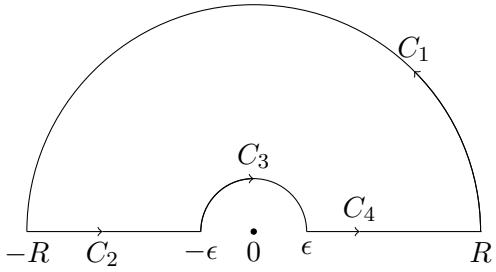


Figure 14.1.3: Structure of C

We have indented the contour C at the origin since the origin is a simple pole of f. Since f is analytic within and on C, by Cauchy Goursat theorem,

$$\begin{aligned} 0 &= \int_C f(z)dz \\ &= \int_{C_1} f(z)dz + \int_{C_2} f(z)dz + \int_{C_3} f(z)dz + \int_{C_4} f(z)dz \\ &= \int_{C_1} f(z)dz + \int_{-R}^{-\epsilon} f(z)dz + \int_{C_3} f(z)dz + \int_{\epsilon}^{R} f(z)dz \\ &= \int_{C_1} f(z)dz + \int_{C_3} f(z)dz + \int_{-\infty}^{\infty} f(x)dx \end{aligned} \tag{14.1.8}$$

when $R \to \infty$ and $\epsilon \to 0$. By Jordan's lemma,

$$\lim_{R\to\infty} \int_{C_1} f(z)dz = 0.$$

Also, we know that,

$$\begin{aligned} \lim_{\epsilon\to 0} \int_{C_3} f(z)dz &= -i\pi.Res(f;0) \\ &= -i\pi \lim_{z\to 0} \frac{z\, e^{i\theta}}{z} \\ &= -i\pi. \end{aligned}$$

Thus, when $R \to \infty$ and $\epsilon \to 0$, we have, from equation (14.1.8),

$$0 = -i\pi + \int_{-\infty}^{\infty} f(x)dx$$

$$\text{or,} \quad \int_{-\infty}^{\infty} \frac{e^{ix}}{x}dx = i\pi$$

$$\text{or,} \quad \int_{-\infty}^{\infty} \frac{\cos x + i \sin x}{x}dx = i\pi.$$

Equating the imaginary parts, we get

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}dx = \pi,$$

and since $\dfrac{\sin x}{x}$ is an even function, so

$$\int_{0}^{\infty} \frac{\sin x}{x}dx = \frac{\pi}{2}.$$

Exercise 14.1.12. 1. Evaluate the following integrals using residue theorem

a. $\displaystyle\int_{|z|=1} \frac{z+1}{z^2-2z}dz$ b. $\displaystyle\int_{|z|=3} \frac{z+1}{z^2-2z}dz$ c. $\displaystyle\int_{|z|=2} \frac{\mathrm{e}^{2z}}{(z-1)^2}dz$

d. $\displaystyle\int_{|z|=4} \left(\frac{z}{z-1}+\frac{z^2}{z+2}\right)dz$ e. $\displaystyle\int_{|z|=2} \frac{\cos z}{z^5}dz$ f. $\displaystyle\int_{|z|=1/2} \frac{1}{z\sin z}dz$

g. $\displaystyle\int_{|z|=1} \frac{\sin z}{z^4}dz$ h. $\displaystyle\int_{|z|=3/2} \frac{1}{z(z^2+1)(z-2)^2}dz$ i. $\displaystyle\int_{|z|=3} \frac{dz}{z(z^2+1)(z-2)^2}$

2. By evaluating $\dfrac{1}{2\pi i}\displaystyle\int_C \frac{dz}{(z-a)\left(z-\frac{1}{a}\right)}$, $C: |z|=1$, prove that

$$\int_0^{2\pi} \frac{d\theta}{1+a^2-2a\cos\theta} = \frac{2\pi}{1-a^2}, \text{ if } 0<a<1.$$

3. Evaluate the following real integrals by contour integration.

a. $\displaystyle\int_0^{2\pi} \frac{d\theta}{a+b\cos\theta}$, $a>b>0$ b. $\displaystyle\int_0^{\pi} \frac{1+2\cos\theta}{5+4\cos\theta}d\theta$

c. $\displaystyle\int_{-\infty}^{\infty} \frac{dx}{(x^2+1)^3}$ d. $\displaystyle\int_0^{\infty} \frac{x^2dx}{(x^2+1)^2}$ e. $\displaystyle\int_{-\infty}^{\infty} \frac{dx}{(x^4+a^4)}$, $a>0$

d. $\displaystyle\int_0^{\infty} \frac{\cos axdx}{(x^2+b^2)^2}$, $a,b>0$ e. $\displaystyle\int_{-\infty}^{\infty} \frac{\sin xdx}{(x^2+4x+5)}$ f. $\displaystyle\int_{-\infty}^{\infty} \frac{x\sin xdx}{(a^2x^2+b^2)}$, $a,b>0$

g. $\displaystyle\int_0^{\infty} \frac{\sin xdx}{x(x^2+a^2)}$, $a>0$

4. By integrating $\dfrac{\mathrm{e}^{iz}}{z-ia}$, $a>0$ over a suitable contour, show that

$$\int_{-\infty}^{\infty} \frac{a\cos x + x\sin x}{x^2+a^2}dx = 2\pi\,\mathrm{e}^{-a}.$$

14.1.2 Argument Principle

Suppose f is analytic and has a zero of order m at $z=a$. So, $f(z)=(z-a)^m g(z)$, where g is analytic at a and $g(a)\neq 0$. Hence,

$$\frac{f'(z)}{f(z)} = \frac{m}{z-a} + \frac{g'(z)}{g(z)} \tag{14.1.9}$$

and g'/g is analytic near $z=a$ since $g(z)\neq 0$. Now suppose, f has a pole of order m at a. Then, $f(z)=(z-a)^{-m}g(z)$ where g is analytic at a and $g(a)\neq 0$. This gives

$$\frac{f'(z)}{f(z)} = -\frac{m}{z-a} + \frac{g'(z)}{g(z)}. \tag{14.1.10}$$

Again g'/g is analytic near $z=a$.

Definition 14.1.13. If G is open and f is a function defined and analytic in G except for poles, then f is called a meromorphic function on G.

Theorem 14.1.14. Given a function f, meromorphic in a region G, suppose γ is a simple closed curve in G and $f(z) \neq 0$ on γ and f is analytic on γ. If, N and P denote the number of zeros and poles respectively of f within γ, multiplicities counted accordingly, then,

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz = N - P.$$

Proof. Since f is analytic on γ, and $f(z) \neq 0$ on γ, the function f'/f is also analytic on γ. Also, a point within γ which is neither a pole or zero of f is a regular point of f'/f. Thus, the only singularities of f'/f within γ are the zeros and poles of f within γ. Let a be a zero of f of order m within γ. Then, by equation (14.1.9)

$$\frac{f'(z)}{f(z)} = \frac{m}{z-a} + \frac{g'(z)}{g(z)}.$$

Hence, a is a simple pole of f'/f with residue m. Next let, b be a pole of f of order n. Then, by equation (14.1.10),

$$\frac{f'(z)}{f(z)} = -\frac{n}{z-b} + \frac{g'(z)}{g(z)}.$$

Hence, b is a simple pole of f'/f with residue $-n$. Now, let $a_1, a_2, \ldots, a_k$ be the zeros of f with respective orders $p_1, p_2, \ldots, p_k$ within γ and let $b_1, b_2, \ldots, b_l$ be the poles of f with respective orders $q_1, q_2, \ldots, q_l$ within γ. Then the only singularities of f'/f within γ are the points $a_1, a_2, \ldots, a_k, b_1, b_2, \ldots b_l$ with residues $p_1, p_2, \ldots, p_k, -q_1, -q_2, \ldots, -q_l$. By residue theorem,

$$\begin{aligned}\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz &= (p_1 + p_2 + \cdots + p_k) - (q_1 + q_2 + \cdots + q_l)\\ &= N - P.\end{aligned}$$

□

Corollary 14.1.15. Let a function f be analytic within and on a simple closed curve γ and $f(z) \neq 0$ on γ. If, N denotes the number of zeros of f within γ, multiplicities being counted, then,

$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz = N.$$

Note 14.1.16. The integral $\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz$ is called the *logarithmic residue* of f relative to γ.

Theorem 14.1.17. (Argument principle) Let f be analytic within and on a simple closed curve γ except for at most a finite number of poles within γ and $f(z) \neq 0$ on γ. Then

$$N - P = \frac{1}{2\pi}\left[\text{Arg} f\right]_\gamma,$$

where $\left[\text{Arg} f\right]_\gamma$ denotes the change of $\text{Arg} f(z)$ as z moves once round γ in the positive sense. N and P are respectively the number of zeros and poles of f within γ, counted according to their multiplicities.

Proof. We know that,

$$\begin{aligned}N - P &= \frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz\\ &= \frac{1}{2\pi i}\int_\gamma \frac{d}{dz}\text{Log} f(z)dz\\ &= \frac{1}{2\pi i}\left[\text{Log} f(z)\right]_\gamma. \qquad (14.1.11)\end{aligned}$$

Choosing any point $z_0 \in \gamma$ as the initial and terminal point of the path of integration, we make one circuit round γ in the positive sense. Then $\text{Log} f(z)$ varies continuously and in general, its value at z_0 after one circuit differs from its original value at z_0. In fact, since $\text{Log} f(z) = \log|f(z)| + i\text{Arg} f(z)$, the change in $\text{Log} f(z)$ is entirely due to the change in $\text{Arg} f(z)$ (since $\log|f(z)|$ is single-valued). From equation (14.1.11), we have

$$N - P = \frac{1}{2\pi i}\left[\log|f(z)| + i\text{Arg} f(z)\right]_\gamma = \frac{1}{2\pi}\left[\text{Arg} f\right]_\gamma .$$

□

Corollary 14.1.18. If f is analytic within and on γ and $f(z) \neq 0$ on γ, then the number of zeros of f within γ is given by

$$N = \frac{1}{2\pi}\left[\text{Arg} f\right]_\gamma .$$

The argument principle establishes a striking relationship between the number of zeros and the number of poles of f in a domain D and how f maps the boundary δD of the domain, namely, the number of times the image of δD winds around the origin.

Example 14.1.19. Let $f(z) = z^2 - 1$, on $C : |z - 1| = 1$. Then

$$w = f(z) = |z|^2 e^{2i\theta} - 1$$

The change of argument of w on C is 2π.

14.1.3 Rouche's Theorem

Theorem 14.1.20. If the functions f and g are analytic within and on a simple closed curve γ and if $|g(z)| < |f(z)|$ on γ, then f and $f + g$ have the same number of zeros inside γ.

Proof. Let N =number of zeros of f within γ, N' =number of zeros of $f+g$ within γ. Since $|g(z)| < |f(z)|$ on γ, $f(z) \neq 0$ on γ. Also, on γ, $|f(z) + g(z)| \geq |f(z)| - |g(z)| > 0$. Hence, $f(z) + g(z) \neq 0$ on γ. Hence,

$$N = \frac{1}{2\pi}\left[\text{Arg} f\right]_\gamma$$

and

$$\begin{aligned} N' &= \frac{1}{2\pi}\left[\text{Arg}(f+g)\right]_\gamma \\ &= \frac{1}{2\pi}\left[\text{Arg} f\left(1 + \frac{g}{f}\right)\right]_\gamma \\ &= \frac{1}{2\pi}\left[\text{Arg} f + \text{Arg}\left(1 + \frac{g}{f}\right)\right]_\gamma \\ &= \frac{1}{2\pi}\left[\text{Arg} f\right]_\gamma + \frac{1}{2\pi}\left[\text{Arg}\left(1 + \frac{g}{f}\right)\right]_\gamma \\ &= N + \frac{1}{2\pi}\left[\text{Arg}\left(1 + \frac{g}{f}\right)\right]_\gamma . \end{aligned} \tag{14.1.12}$$

Let

$$F(z) = 1 + \frac{g(z)}{f(z)}.$$

Then,

$$|w-1| = \frac{|g(z)|}{|f(z)|} < 1, \ \ \forall z \in \gamma$$

where $w = F(z)$. This shows that as z describes the closed contour γ, the variable w describes a closed curve which lies entirely to the right of the imaginary axis and hence the origin lies outside the curve. Hence Arg$w$, that is Arg$\left(1+\frac{g}{f}\right)$ returns to its original value as z describes γ. Hence, $\left[\text{Arg}\left(1+\frac{g}{f}\right)\right]_\gamma = 0$. From equation (14.1.12), we have, $N' = N$ and the theorem is proved. □

The Fundamental theorem of Algebra can be proved by using the Rouche's theorem.

Theorem 14.1.21. Every polynomial of degree n has n zeros in the complex plane.

Proof. Let

$$P(z) = a_0 + a_1 z + \ldots + a_n z^n, \ \ a_n \neq 0$$

be a polynomial of degree n. Let $f(z) = a_n z^n$ and $g(z) = a_0 + a_1 z + \ldots + a_{n-1} z^{n-1}$. Then $f(z) + g(z) = P(z)$. Let γ be the circle $|z| = R$, $R > 1$. Now, f has n zeros inside γ, all the zeros being at the origin. On γ,

$$|f(z)| = |a_n| R^n$$

and

$$\begin{aligned} |g(z)| &\leq |a_0| + |a_1||z| + \ldots + |a_{n-1}||z^{n-1}| \\ &= |a_0| + |a_1|R + \ldots + |a_{n-1}|R^{n-1} \\ &\leq R^{n-1}\left(|a_0| + |a_1| + \ldots + |a_{n-1}|\right). \end{aligned}$$

Hence, on γ,

$$|g(z)| < |f(z)| \ \text{ if } \ \frac{|a_0| + |a_1| + \ldots + |a_{n-1}|}{|a_n| R} < 1,$$

that is, if

$$R > \frac{|a_0| + |a_1| + \ldots + |a_{n-1}|}{|a_n|}.$$

We choose R_1 such that

$$R_1 > \max\left\{1, \frac{|a_0| + |a_1| + \ldots + |a_{n-1}|}{|a_n|}\right\}.$$

Then the functions f and g are analytic within and on the circle $\gamma_1 : |z| = R_1$ and $|g(z)| < |f(z)|$ for all $z \in \gamma_1$. Hence, by Rouche's theorem, f and $f+g$, that is, P will have the same number of zeros within γ_1. Hence P has n zeros within γ_1 and as such, P has n zeros in the entire complex plane. □

Example 14.1.22. Suppose we want to determine the number of zeros, including multiplicity, of the polynomial $2z^5 - 6z^2 + z + 1$ in $1 \leq |z| < 2$. Let $g(z) = 2z^5 - 6z^2 + z + 1$, $f_1(z) = -6z^2$ and $f_2(z) = 2z^5$, then

$$|f_1 - g| \leq 2 + 1 + 1 = 4 < 6 = |f_1|$$

on $|z| = 1$ and

$$|f_2 - g| \leq 24 + 2 + 1 = 27 < 64 = |f_2|$$

on $|z| = 2$. Hence, by Rouche's theorem, g has 2 zeros in $|z| < 1$ and 5 zeros in $|z| < 2$. Thus, g has 3 zeros in $1 \leq |z| < 2$.

Exercise 14.1.23.

1. Use Rouche's theorem to show that the equation $z^4 + 5z - 1 = 0$ has just one zero inside $|z| = 1$.
2. Determine the number of zeroes, including multiplicity, of the following polynomials in $|z| < 1$.
   (a) $z^6 - 5z^4 + z^3 - 2z$
   (b) $2z^4 - 2z^3 + 2z^2 - 2z + 9$.
3. Determine the number of zeroes, including multiplicity, of the following polynomials in $|z| < 2$.
   (a) $z^4 + 3z^3 + 6$
   (b) $z^4 - 2z^3 + 9z^2 + z - 1$.
4. Suppose $c \in \mathbb{C}$ is such that $|c| > \mathrm{e}$. Show that the number of solution, including multiplicity, of the equation $\mathrm{e}^z = cz^n$ in $|z| < 1$ is n.
5. If $k > 1$, show that the equation $z^n \, \mathrm{e}^{k-z} = 1$ has n roots inside $|z| = 1$, n being a positive integer.
6. Show that all the roots of the equation $z^5 + az + 1 = 0$ lie within the circle $|z| = r$ if $|a| < r^4 - \dfrac{1}{r}$.
7. Let f be analytic on $|z| \leq 1$ and $|f(z)| < 1$ on $|z| = 1$. Show that there is just one point α within the circle $|z| = 1$ such that $f(\alpha) = \alpha$.
8. Show that all the roots of $z^7 - 5z^3 + 12 = 0$ lie in the annulus bounded by the circles $|z| = 1$ and $|z| = 2$.
9. Find the number of roots of the equation $z^8 + 6z^5 - 3z^3 + 1 = 0$ in the annulus $1 < |z| < 2$.

14.1.4 Maximum Modulus Theorem

In this unit, we will state the maximum modulus theorem which gives us an idea on the size of an analytic function in some specific domains. Let us first state the following lemma.

Lemma 14.1.24. If $\phi(x)$ is continuous in $[a, b]$, $\phi(x) \leq k$ in $[a, b]$ and $\dfrac{1}{b-a}\displaystyle\int_a^b \phi(x)dx \geq k$, then $\phi(x) \equiv k$ throughout $[a, b]$.

Proof. If possible let $\phi(x) \neq k$ throughout the open interval (a, b). Then there exists a point $\alpha \in (a, b)$ such that $\phi(\alpha) < k$. Due to continuity of ϕ in (a, b), we can find an interval $(\alpha - \delta, \alpha + \delta)$ in which $\phi(x) < k - \epsilon$ for a chosen $\epsilon > 0$. Now,

$$\begin{aligned}\int_a^b \phi(x)dx &= \int_a^{\alpha-\delta} \phi(x)dx + \int_{\alpha-\delta}^{\alpha+\delta} \phi(x)dx + \int_{\alpha+\delta}^b \phi(x)dx \\ &\leq k(\alpha - \delta - a) + 2\delta(k - \epsilon) + k(b - \alpha - \delta) \\ &= k(b - a) - 2\delta\epsilon.\end{aligned}$$

Hence,

$$\frac{1}{b-a}\int_a^b \phi(x)dx \leq k - \frac{2\delta\epsilon}{b-a}$$

which contradicts the given condition. Hence $\phi(x) = k$ throughout (a, b). From the continuity of ϕ at a and b, it follows that $\phi(a) = k = \phi(b)$. Hence, $\phi(x) = k$ in $[a, b]$. □

Let us now state the maximum modulus theorem.

Theorem 14.1.25. If a function f is analytic in a bounded region G and continuous on $\overline{G}$ (closure of G), and $M = \max_{z\in\partial G} |f(z)|$, ∂G being the boundary of G, then $|f(z)| < M$ in G unless f is a constant function.

Proof. Let f be non-constant on $\overline{G}$. If possible, let the maximum value of $|f(z)|$ on $\overline{G}$ be attained at an interior point α of $\overline{G}$. Let $\gamma : |z-\alpha| = r$, where r is so small that $\gamma \subset G$. By Cauchy's integral formula,

$$f(\alpha) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-\alpha}dz. \tag{14.1.13}$$

Clearly, $|f(\alpha)| \neq 0$. For otherwise f will be a constant function since $|f(\alpha)|$ is assumed to be a maximum. Putting $z - \alpha = r\,e^{i\theta}$,

$$\frac{f(z)}{f(\alpha)} = \rho(\theta)\, e^{i\phi(\theta)}$$

so that ρ and ϕ are continuous functions pf θ, we get from equation (14.1.13),

$$\begin{aligned} 1 &= \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{f(\alpha)}\frac{dz}{z-\alpha} \\ &= \frac{1}{2\pi i}\int_0^{2\pi} \frac{\rho\, e^{i\phi}\, .ir\, e^{i\theta}\, d\theta}{r\, e^{i\theta}} \\ &= \frac{1}{2\pi}\int_0^{2\pi} \rho\, e^{i\phi}\, d\theta. \end{aligned} \tag{14.1.14}$$

Hence,

$$1 = \frac{1}{2\pi}\left|\int_0^{2\pi} \rho\, e^{i\phi}\, d\theta\right| \leq \frac{1}{2\pi}\int_0^{2\pi} |\rho\, e^{i\phi}\,|d\theta = \frac{1}{2\pi}\int_0^{2\pi} |\rho(\theta)|d\theta.$$

Also, $|\rho(\theta)| = \left|\dfrac{f(z)}{f(\alpha)}\right| < 1$ on $\overline{G}$, since $|f(\alpha)|$ is a maximum on $\overline{G}$. Hence, by lemma 14.1.24, $\rho = \rho(\theta) = 1$ in $0 \leq \theta \leq 2\pi$. Now, taking real part of equation (14.1.14) we get

$$1 = \frac{1}{2\pi}\int_0^{2\pi} \cos\phi(\theta)d\theta.$$

Since $\cos\phi(\theta) \leq 1$ in $0 \leq \theta \leq 2\pi$, we have, by the lemma 14.1.24, $\cos\phi(\theta) = 1$ in $0 \leq \theta \leq 2\pi$. Hence, $\sin\phi(\theta) = 0$. Thus, $\dfrac{f(z)}{f(\alpha)} = 1$ on γ, that is, $f(z) = f(\alpha)$ on γ and hence by the interior uniqueness theorem, everywhere on $\overline{G}$. This implies that f is a constant function which contradicts our assumption. Hence, the maximum value of $|f|$ can not be attained at an interior point of $\overline{G}$, unless f is a constant function. □

Corollary 14.1.26. Suppose that G is a bounded region with compact closure $\overline{G}$. . If f is analytic on G and continuous on $\overline{G}$ then

$$\sup_{z\in G}|f(z)| \leq \sup_{z\in\overline{G}\setminus G} |f(z)|$$

Note 14.1.27. It is impossible to drop the assumption that G is bounded in the maximum modulus theorem. For example, let $G = \{z : \text{Im}z > 0\}$ and $f(z) = e^{-iz}$. Then f is continuous on $\overline{G} = \{z : \text{Im}z \geq 0\}$ and analytic on G. If $z \in \partial G = \{z : \text{Im}z = 0\}$, then $|f(z)| = |\,e^{-ix}\,| = 1$. But for $z = x + iy \in G$, $|f(z)| = e^y \to \infty$ as $y \to \infty$. Thus, maximum modulus theorem is not true for f and G.

Theorem 14.1.28. (Minimum modulus theorem) Let f be a non-constant analytic function in a bounded region G and continuous on $\overline{G}$. If $f(z) \neq 0$ inside ∂G, ∂G being the boundary of G, then $|f(z)|$ must attain its minimum value on ∂G.

Proof. The theorem is clearly true when f vanishes at a point on ∂G. We therefore assume that $f(z) \neq 0$ on ∂G. It follows that $\frac{1}{f}$ is also analytic in G and continuous on $\overline{G}$ since $f(z) \neq 0$ on ∂G. By Maximum modulus theorem, $\frac{1}{|f|}$ cannot assume its maximum value inside ∂G and so $|f|$ can not attain its minimum value inside ∂G. Since $|f|$ has a minimum on $\overline{G}$ (since f is continuous on a compact set $\overline{G}$), this minimum must be attained on ∂G. □

Note 14.1.29. If f is analytic within and on a simple closed curve γ and $f(z) = 0$ at some point in the interior of γ, then $|f|$ need not assume its minimum value on γ. For example, let $f(z) = z$ for $|z| \leq 1$ and $\gamma : |z| = 1$. Then $f(0) = 0$. Also, for all $z \in \gamma$, $|f(z)| = |z| = 1$. Hence, the minimum value of $|f|$ does not occur on γ.

The next theorem is an important application of the maximum modulus theorem.

Theorem 14.1.30. (Schwarz Lemma) Let $D = \{z : |z| < 1\}$ and suppose that f is analytic on D with

1. $|f(z)| \leq 1$
2. $f(0) = 0$.

Then $|f'(0)| \leq 1$ and $|f(z)| \leq |z|$ for all z in the disk D. Moreover, if $|f'(0)| = 1$ or if $|f(z)| = |z|$ for some $z \neq 0$, then there exists a constant c, $|c| = 1$, such that $f(w) = cw$ for all w in D.

Proof. Define $g : D \to \mathbb{C}$ by

$$\begin{aligned} g(z) &= \frac{f(z)}{z}, z \neq 0 \\ &= f'(z), z = 0 \end{aligned}$$

Then, g is analytic in D. Using Maximum Modulus Theorem,

$$|g(z)| \leq \frac{1}{r}$$

for $|z| \leq r$ and $0 < r < 1$. Letting r approach 1, we get,

$$|g(z)| \leq 1$$

for all z in D. That is,

$$|f(z)| \leq |z|$$

and $|f'(0)| = |g(0)| \leq 1$.

If $|f(z)| = |z|$ for some $z \neq 0$, in D, or $|f'(0)| = 1$, then $|g|$ assumes its maximum value inside D. Thus, by Maximum Modulus Theorem, $g(z) \equiv c$, for some constant c with $|c| = 1$. Hence, $f(z) = cz$ and hence the result. □

Theorem 14.1.31. (Open Mapping theorem) Let G be a region and suppose that f is a non-constant analytic function on G. Then for any open set U in G, $f(U)$ is open.

Proof. Let $U \subset G$ be open. To show that $f(U)$ is open we show that for each $a \in U$, there exists a $\delta > 0$ such that the open ball $B(f(a);\delta) \subset f(U)$. Let $\phi(z) = f(z) - f(a)$. Then a is a zero of ϕ. Since zeros of an analytic function are isolated points, there exists an open ball $B(a;r)$ with $\overline{B(a;r)} \subset U$ such that $\phi(z) \neq 0$ in $0 < |z-\alpha| < r$. In particular, $\phi(\alpha) \neq 0$ for $\alpha \in \partial B(a;\rho)$ where $\rho < r$.

Let $2\delta = \min\limits_{\alpha\in\partial B(a;\rho)} |\phi(\alpha)|$. Then $\delta > 0$. Now, for any $w \in B(f(a);\delta)$ we have

$$\begin{aligned} |f(\alpha) - w| &\geq |f(\alpha) - f(a)| - |f(a) - w| \\ &= |\phi(\alpha)| - |f(a) - w| \\ &> 2\delta - \delta > |f(a) - w| \ \ \forall \alpha \in \partial B(a;\rho). \end{aligned}$$

This implies that

$$\min_{\alpha\in\partial B(a;\rho)} |f(\alpha) - w| > |f(a) - w|. \tag{14.1.15}$$

Let $F(z) = f(z) - w$. Then f has a zero in $B(a;\rho)$. For if $F(z) \neq 0$ in $B(a;\rho)$, there exists a neighbourhood $N(a)$ of a containing $\overline{B(a;\rho)}$ lying in G such that $F(z) \neq 0$ in $N(a)$. Then $\dfrac{1}{F(z)}$ will be analytic in $N(a)$ and

$$\left|\frac{1}{F(a)}\right| < \max_{\alpha\in\partial B(a;\rho)} \left|\frac{1}{F(\alpha)}\right| = \frac{1}{\min_{\alpha\in\partial B(a;\rho)} |F(\alpha)|},$$

that is,

$$\min_{\alpha\in\partial B(a;\rho)} |f(\alpha) - w| < |f(a) - w|$$

which contradicts equation (14.1.15). Hence, there exists $z_0 \in B(a;\rho)$ such that $f(z_0) = w$. Since w is an arbitrary point of $B(f(a);\delta)$ it follows that $B(f(a);\delta) \subset f(U)$ and the theorem is proved. □

Exercise 14.1.32.

1. Let $D = \{z : |z| < 1\}$ and $f : D \to D$ be analytic with $f(0) = 0$. Prove that $|f(z)| \leq |z|$ for all $z \in D$.

2. Let f be analytic in a bounded region D and continuous on $\overline{D}$. Let $u = \text{Re} f$. Show that u can not attain its maximum value at an interior point of $\overline{D}$.

3. Let f be a non-constant analytic function on a bounded domain D and continuous on $\overline{D}$ and suppose that $|f(z) \equiv$constant on ∂D. Prove that f has at least one zero in D.

4. Apply maximum modulus principle to prove the fundamental theorem of Algebra.

5. If f is analytic on the disk $|z| \leq 1$, $|f(z)| \leq M$ on $|z| = 1$ and $f(a) = 0$ where $|a| < 1$, show that

$$|f(z)| \leq M \left|\frac{z-a}{1-\overline{a}z}\right| \ \text{ on } \ |z| \leq 1.$$

Sample Questions

1. State and prove the Cauchy's Residue theorem.
2. State and prove the Open mapping theorem.
3. State and prove the maximum modulus theorem. Hence prove the Fundamental theorem of Algebra.
4. State and prove the Schwarz lemma.
5. Let $f(z) = \dfrac{P(z)}{Q(z)}$, where (i) P, Q are analytic at α, (ii) $P(\alpha) \neq 0$, and (iii) Q has a double zero at α. Show that
$$Res(f;\alpha) = \frac{2}{3}.\frac{3P'(\alpha)Q''(\alpha) - P(\alpha)Q'''(\alpha)}{(Q''(\alpha))^2}.$$
6. Let f be analytic within and on a simple closed curve γ except for a finite number of poles within γ and $f(z) \neq 0$ on γ. Also, let ϕ be analytic within and on γ. If $\alpha_1, \alpha_2, \ldots, \alpha_m$ be the zeros of f within γ of respective orders $p_1, p_2, \ldots, p_m$ and $\beta_1, \beta_2, \ldots, \beta_n$ be the poles of f within γ with respective orders $q_1, q_2, \ldots, q_n$, show that
$$\frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}\phi(z)dz = \sum_{k=1}^{m} p_k\phi(\alpha_k) - \sum_{k=1}^{n} q_k\phi(\beta_k).$$

Unit 15

Course Structure

- Introduction
- Objectives
- Linear Operators

15.1 Introduction

In this chapter the concept of linear operators and their properties would be briefly discussed. Conventionally, a mapping from one vector space to another vector space is called an operator.

15.2 Linear Operators

Definition 15.2.1. Let X and Y be any two linear spaces over the same scalar field Φ ($\mathbb{R} or \mathbb{C}$). A mapping $T : XY$ is called a linear operator if it preserves both addition and scalar multiplication. i.e., if for all $x, y \in X$ and all $\alpha \in \Phi$, we have

1. $T(x + y) = Tx + Ty$ (additive property)
2. $T(\alpha x) = \alpha Tx$ (homogeneity property).

Obviously, conditions (i) and (ii) are equivalent to the single condition $T(\alpha x + \beta y) = \alpha Tx + \beta Ty$, for all $x, y \in X$ and $\alpha, \beta \in \Phi$.

Definition 15.2.2. The Null space $N(T)$ of T is the set of all $x \in X$ such that $T(x) = 0$. The Null space of T is also called the Kernel of T.

Definition 15.2.3. The range space $R(T)$ is the set

$$R(T) = T(x) \in Y \; x \in X.$$

Example 15.2.4. Let X be a vector space.

1. The identity operator $I : XX$, defined as
$$I(x) = x$$
for all $x \in X$. This is a linear operator on X.
2. The zero operator $O : X \to X$ defined as
$$O(x) = x$$
for all $x \in X$ is a linear operator on X.
3. Let X be the space of all polynomial on $[a, b]$. Then the operator $T : X \to X$ defined as
$$\mathrm{T(x(t)) = x'(t)}$$
for every polynomial $x(t) \in X$, is linear.
4. A linear operator T on $C[a, b]$ into itself can be defined as
$$\mathrm{T(x(t)) = \int_a^t x(\xi)d\xi}$$
where $t \in [a, b]$.

We can define another linear operator on $C[a, b]$ as
$$T'(x(t)) = tx(t).$$

Some properties of linear operators are presented below.

Theorem 15.2.5. For every linear operators $T : X \to Y$, we have

1. $T0 = 0$
2. $T(-x) = -Tx$
3. $T(x - y) = Tx - Ty$
4. $T\left(\sum_{i=1}^n \alpha_i x_i\right) = \sum_{i=1}^n \alpha_i T(x_i)$.

Proof. 1. We have $T0 = T(0.0) = 0T0 = 0$

2. $T(-x) = T((-1)x) = -1Tx = -Tx$
3. $T(x - y) = T(x + (-y)) = Tx + T(-y) = Tx + (-Ty) = Tx - Ty$.
4. For $n = 1, T\left(\sum_{i=1}^n \alpha_i x_i\right) = T(\alpha_1 x_1) = \alpha_1 Tx_1 = \sum_{i=1}^n \alpha_i T(x_i)$ Now for $n > 1$,
$$\begin{aligned} T\left(\sum_{i=1}^n \alpha_i x_i\right) &= T(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n-1} x_{n-1} + \alpha_n x_n) \\ &= T(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n-1} x_{n-1}) + \alpha_n T x_n \\ &= T(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n-2} x_{n-2}) + \alpha_{n-1} T x_{n-1} + \alpha_n T x_n \\ &\vdots \\ &\vdots \\ &= \alpha_1 T x_1 + \alpha_2 T x_2 + \cdots + \alpha_{n-1} T x_{n-1} + \alpha_n T x_n \\ &= \sum_{i=1}^n \alpha_i T(x_i). \end{aligned}$$

□

Theorem 15.2.6. Let $T : X \to Y$ be a linear operator, then

1. The range $T(X)$ of T is a linear subspace of Y.
2. The inverse operator $T^{-1} : T(X) \to X$ exists if and only if $Tx = 0 \Rightarrow x = 0$.
3. If T^{-1} exists then T^{-1} is a linear operator on $T(X)$.

Proof. 1. Since $T0 = 0 \in Y$, so $0 \in T(X)$.

Let $u, v \in T(X)$ and α, β be any two scalars. Then

$$u = Tx \text{ and } v = Ty \text{ for some } x, y \in X.$$

So, $T(\alpha x + \beta y) = \alpha Tx + \beta Ty = \alpha u + \beta v \in T(X)$. Hence, $T(X)$ is a linear subspace of Y.

2. The inverse operator $T^{-1} : T(X) \to X$ exists if and only if T is one – to – one.

Now let T^{-1} exists. Then T is one – to – one and hence

$$Tx = 0 \Rightarrow Tx = T0 \Rightarrow x = 0.$$

Conversely, assume that $Tx = 0 \Rightarrow x = 0$. If then $Ty = Tz$ for some $y, z \in X$, we have $Ty - Tz = 0$ i.e., $T(y - z) = 0$ and hence by hypothesis, $y - z = 0$ i.e., $y = z$.

Hence T is one – to – one and, therefore, T^{-1} exists.

3. Suppose $T^{-1} : T(X) \to X$ exists. We know that $T(X)$ is a linear subspace of Y. So $T(X)$ and X are linear spaces over the same scalar field.

Let $u, v \in T(X)$ and let α, β be any two scalars. Then $T^{-1}(u)$ and $T^{-1}(v)$ are unique vectors of X. Since T is linear, we have

$$T\left(\alpha T^{-1}(u) + \beta T^{-1}(v)\right) = \alpha T\left(T^{-1}(u)\right) + \beta T\left(T^{-1}(v)\right) = \alpha u + \beta v.$$

Therefore by definition, $T^{-1}(\alpha u + \beta v) = \alpha T^{-1}(u) + \beta T^{-1}(v)$.

Hence T^{-1} is a linear operator on $T(X)$.

□

Theorem 15.2.7. Let $T : X \to Y$ be a linear operator. If X has finite dimension n then the linear subspace $T(X)$ of Y has some finite dimension $m \leq n$, equality holding if and only if T^{-1} exists. If X and Y are of the same finite dimension and if T^{-1} exists, then $T(X) = Y$ (i.e., T is onto Y).

Proof. Since $T : X \to Y$ be a linear operator, we know that $T(X)$ is a linear subspace of Y. Now X has finite dimension n. Let $y_1, y_2, \ldots, y_{n+1}$ be any $(n + 1)$ vectors in $T(X)$. Then there are vectors $x_1, x_2, \ldots, x_{n+1}$ in X such that $Tx_1 = y_1, Tx_2 = y_2, \ldots, Tx_{n+1} = y_{n+1}$.

Since $\dim X = n$, so the set of vectors $\{x_1, x_2, \ldots, x_{n+1}\}$ is linearly independent in X. Then there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_{n+1}$ not all zero such that

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n+1} x_{n+1} = 0.$$

Then

$$\begin{aligned} & T(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{n+1} x_{n+1}) = T0 = 0 \\ \Rightarrow \quad & \alpha_1 Tx_1 + \alpha_2 Tx_2 + \cdots + \alpha_{n+1} Tx_{n+1} = 0 \\ \Rightarrow \quad & \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_{n+1} y_{n+1} = 0, \text{ where } \alpha_1, \alpha_2, \ldots, \alpha_{n+1} \text{ not all zero.} \end{aligned}$$

So any $(n+1)$ vectors $y_1, y_2, \ldots, y_{n+1}$ in $T(X)$ are linearly dependent. Therefore, the maximum number of linearly independent vectors in $T(X)$ cannot exceed n. Hence $T(X)$ has some finite dimension $m \leq n$.

Next first suppose that $m = n$.

If $m = n = 0$, then $X = \{0\}$ and $T(X) = \{0\}$. So in that case T is one - to - one and hence T^{-1} exists. Assume that $m = n \geq 1$. Then $T(X)$ has a basis of n-vectors, $\{u_1, u_2, \ldots, u_n\}$, say. Then there are vectors $e_1, e_2, \ldots, e_n$ in X such that

$$u_1 = Te_1, u_2 = Te_2, \ldots, u_n = Te_n$$

Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be scalars such that $\alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_n e_n = 0$. Then

$$\begin{aligned} & T(\alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_n e_n) = T0 = 0 \\ \Rightarrow \quad & \alpha_1 Te_1 + \alpha_2 Te_2 + \cdots + \alpha_n Te_n = 0 \\ \Rightarrow \quad & \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n = 0. \end{aligned}$$

Since the basis $\{u_1, u_2, \ldots, u_n\}$ is linearly independent, we must have $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$. Hence the set of vectors $\{e_1, e_2, \ldots, e_n\}$ in X is linearly independent. Since $\dim X = n$, so $e_1, e_2, \ldots, e_n$ is a basis of X. Therefore, each $x \in X$ can be expressed uniquely as a linear combination $x = \beta_1 e_1 + \beta_2 e_2 + \ldots + \beta_n e_n$, say.

$$\begin{aligned} & \text{Then} \quad Tx = 0 \\ \Rightarrow \quad & T(\beta_1 e_1 + \beta_2 e_2 + \ldots + \beta_n e_n) = 0 \\ \Rightarrow \quad & \beta_1 Te_1 + \beta_2 Te_2 + \ldots + \beta_n Te_n = 0 \\ \Rightarrow \quad & \beta_1 u_1 + \beta_2 u_2 + \cdots + \beta_n u_n = 0 \\ \Rightarrow \quad & \beta_1 = \beta_2 = \ldots = \beta_n = 0 \end{aligned}$$

and so $x = 0$. Thus $Tx = 0 \Rightarrow x = 0$. Hence T^{-1} exists.

Conversely, assume that $T^{-1} : T(X) \to X$ exists. Then we know that T^{-1} is a linear operator. So by the first part of the proof, we have

$$\begin{aligned} & \dim T^{-1}(T(X)) \leq \dim T(X) = m \\ & \text{i.e., } \dim X \leq m, \text{ i.e., } n \leq m. \end{aligned}$$

Since $m \leq n$, it follows that $m = n$. Finally, if $\dim X = \dim Y = n$ and if T^{-1} exists, then by the above $\dim T(X) = m = n$. So $\dim T(X) = \dim Y$. But we know that if $T(X)$ is a proper subspace of Y, then $\dim T(X) < \dim Y$. Hence we must have $T(X) = Y$, i.e., T is onto or surjective.

□

Exercise 15.2.8. Prove that $N(T)$ is a linear subspace of X.

Lemma 15.2.9. Let $T : X \to Y$ and $S : Y \to Z$ be bijective linear operators, where X, Y, Z are vector spaces. Then the inverse $(ST)^{-1} : Z \to X$ of the product of ST exists, and

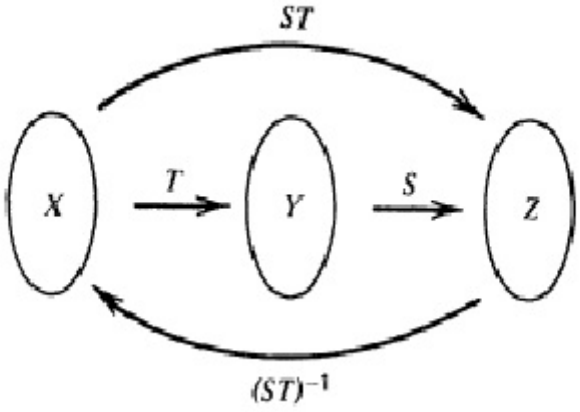
$$(ST)^{-1} = T^{-1}S^{-1}.$$

Proof. The operator $ST : X \to Z$ is bijective, so that $(ST)^{-1}$ exists. We thus have

$$ST(ST)^{-1} = I_Z$$

where I_Z is the identity operator on Z. Applying S^{-1} and using $S(-1)S = I_Y$ (identity operator on Y), we obtain

$$S^{-1}ST(ST)^{-1} = T(ST)^{-1} = S^{-1}I_Z = S^{-1}.$$



Applying T^{-1} and using $T^{-1}T = I_X$, we obtain

$$T^{-1}T(ST)^{-1} = T^{-1}S^{-1}$$
$$\text{i.e.,} \quad I_X(ST)^{-1} = (ST)^{-1} = T^{-1}S^{-1}.$$

Hence the result. □

Exercise 15.2.10. 1. If the product (the composite) of two linear operators exists, show that the product operator is also linear.

2. Let $T : X \to Y$ be a linear operator whose inverse exists. If $\{x_1, x_2, \ldots, x_n\}$ is linearly independent in X then show that the set $\{T(x_1), T(x_2), \ldots, T(x_n)\}$ is linearly independent in Y.

3. Let $T : C[0,1] \to C[0,1]$, where $C[0,1]$ is a Banach space under sup norm, such that $Tx = y$ where

$$y(t) = \int_a^t x(\xi)d\xi;\ x \in C[0,1] \text{ and } 0 \leq t \leq 1.$$

Finde the range of T, and obtain $T^{-1} : T(C[0,1]) \to C[0,1]$. Examine if T^{-1} is linear.

Unit 16

Course Structure

- Introduction
- Objectives
- Bounded Linear operators
- Continuity

16.1 Introduction

In this section the concept of bounded linear operators on normed linear spaces are introduced. The properties of those linear operators are also discussed.

Objectives

After reading this unit, the readers will be able to

- define bounded linear operators
- learns about their various characterizations
- learn various examples of bounded linear operators
- learn about boundedness in finite dimensional normed linear spaces
- learn the relationship between continuity and boundedness

16.2 Linear Operators on Normed Linear Spaces

Definition 16.2.1. Let X, Y be any two normed linear spaces over the same scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). A linear operator $T : X \to Y$ is said to be bounded if there is a number $c > 0$ such that

$$||Tx|| \leq c||x|| \text{ for all } x \in X.$$

Theorem 16.2.2. Let $T : X \to Y$ be a linear operator where X, Y be any two normed linear spaces over the same scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). Then the following conditions are equivalent:

1. T is continuous at some point of X
2. T is continuous on X
3. T is a bounded linear operator
4. The image $T(S)$ of the unit sphere $S = \{x \in X : \ ||x|| = 1\}$ is bounded
5. The image $T(E)$ of every bounded subset $E \subset X$ is bounded in Y.

Proof. 1 $\Rightarrow$ 2 Suppose that T is continuous at some point $x_0 \in X$. Then for any $\epsilon > 0$, there is a $\delta > 0$ such that

$$||Tx - Tx_0|| < \epsilon \ \text{ whenever } \ ||x - x_0|| < \delta. \tag{16.2.1}$$

Then for any two points $x_1, x_2 \in X$ with $||x_1 - x_2|| < \delta$ we have $||(x_1 - x_2 + x_0) - x_0|| < \delta$; and so by (16.2.1)

$$\begin{aligned} & ||T(x_1 - x_2 + x_0) - Tx_0|| < \epsilon \\ \text{i.e.,} \quad & ||Tx_1 - Tx_2 + Tx_0 Tx_0|| < \epsilon \\ \text{i.e.,} \quad & ||Tx_1 - Tx_2|| < \epsilon. \end{aligned}$$

Hence T is not only continuous but also uniformly continuous on X.

Thus, 1 implies 2.

2 $\Rightarrow$ 3 By 2 T is continuous on $0 \in X$. So for $\epsilon = 1$, there is a $\delta > 0$ such that

$$||Tx|| = ||Tx - T0|| < \epsilon = 1 \ \text{ whenever } \ ||x - 0|| < \delta.$$

Now if $x \in X$ and $x \neq 0$, then $\dfrac{\delta}{2||x||}.x \in X$ and $\left| |\dfrac{\delta}{2x}.||x|| \right| | = \dfrac{\delta}{2||x||}||x|| = \dfrac{\delta}{2} < \delta$. So by 2,

$$\begin{aligned} & \left| |T\left(\frac{\delta}{2||x||}.x\right) \right| | < 1 \\ \text{i.e.,} \quad & \left| |\frac{\delta}{2||x||}.Tx \right| | < 1 \\ \text{i.e.,} \quad & \frac{\delta}{2||x||}.||Tx|| < 1 \\ \text{i.e.,} \quad & ||Tx|| < \frac{2}{\delta}||x||. \end{aligned}$$

Hence for all $x \in X$ (with $x = 0$) we have $||Tx|| \leq \dfrac{2}{\delta}||x||$. Therefore, by definition, T is a bounded linear operator.

3 $\Rightarrow$ 4 Do yourself.

4 $\Rightarrow$ 5 Do yourself.

$\underline{5 \Rightarrow 1}$ The set $E = \{x \in X : \ ||x|| < 1\} \subset X$ is bounded. So by 5, $T(E)$ is bounded in Y. So there is a number $M > 0$ such that $||Tx|| \leq M$ for all $x \in E$, i.e., for $||x|| < 1$.

Now given $\epsilon > 0$, let $\delta = \dfrac{\epsilon}{2M}$. Then for all $x \in X$ with $||x - 0|| < \delta$, we have $\left||\dfrac{1}{\delta}x\right| = \dfrac{1}{\delta}||x|| < 1$ and hence by the above theorem,

$$\begin{aligned} & \left||T\left(\frac{1}{\delta}x\right)\right| \leq M \\ \text{or,} \quad & \frac{1}{\delta}||Tx|| \leq M \\ \text{or,} \quad & ||Tx|| \leq M\delta = \frac{\epsilon}{2} < \epsilon \\ \text{or,} \quad & ||Tx - T0|| < \epsilon. \end{aligned}$$

Hence T is continuous at $0 \in X$.

This completes the proof of the theorem. □

Example 16.2.3.

1. The identity operator $I : X \to X$ on a normed linear space $X = \{0\}$ is bounded and has norm $||I|| = 1$.
2. The zero operator $0 : X \to Y$ is bounded and has norm $||0|| = 0$.
3. Let X be the normed linear space of all polynomials on $J = [0, 1]$ with norm $||x|| = \max |x(t)|,\ t \in J$. Let us define a linear operator T on X as

$$T(x(t)) = x'(t).$$

Do it yourself.

Example 16.2.4. An additive operator $T : X \to Y$, where X, Y are any two normed linear spaces over the same scalar field is continuous if and only if it is bounded.

First suppose that T is continuous. By continuity of T at $0 \in X$, for $\epsilon = 1$, there is a $\delta > 0$ such that

$$||Tx - T0|| < \epsilon = 1, \ \text{ whenever } \ ||x - 0|| < \delta.$$

Now, $T0 = T(0 + 0) = T0 + T0$, so $T0 = 0$. For any x with $x \neq 0$, we chose a rational number $\dfrac{m}{n}$ ($m,\ n$ are positive integers) such that

$$\frac{\delta}{2||x||} < \frac{m}{n} < \frac{\delta}{||x||}.$$

Then $\left||\dfrac{m}{n}x\right| = \dfrac{m}{n}||x|| < \delta$ and so by the above $\left||T\left(\dfrac{m}{n}x\right)\right| < \epsilon = 1$. Since T is additive, we have

$$\begin{aligned} T(mx) &= T(x + x + \ldots + x - m \text{ times}) \\ &= mTx = mT\left(n.\frac{1}{n}x\right) = mnT\left(\frac{1}{n}x\right) = nT\left(\frac{m}{n}x\right) \\ \text{i.e.,} \quad \frac{m}{n}Tx &= T\left(\frac{m}{n}x\right). \end{aligned}$$

Therefore,

$$\left||T\left(\frac{m}{n}x\right)\right| = \left||\frac{m}{n}T\left(x\right)\right| = \frac{m}{n}||Tx||.$$

Therefore,

$$\frac{m}{n}||Tx|| < 1 \text{ i.e., } ||Tx|| < \frac{m}{n} < \frac{2}{\delta}||x||.$$

Therefore,

$$||Tx|| < \frac{2}{\delta}||x|| \text{ for all } x \in X \text{ (including } x = 0).$$

Thus T is a bounded linear operator.

Conversely, assume that the operator T is bounded. Then there is a number $c > 0$ such that $||Tx|| \leq c||x||$ for all $x \in X$. Then

$$||Tx - Ty|| = ||T(x - y + y) - Ty|| = ||T(x - y) + Ty + Ty|| = ||T(x - y)|| \leq c||x - y||.$$

This shows that T is continuous.

Theorem 16.2.5. Theorem 16.2.5. Let $T : X \to Y$ be a linear operator where X, Y are any two normed linear spaces over the same scalar field. Then $T^{-1} : T(X) \to X$ exists and is continuous if and only if there is a constant $c > 0$ such that $\|Tx\| \geq c\|x\|$ for all $x \in X$.

Proof. First suppose that $T^{-1} : T(X) \to X$ exists and is continuous on $T(X)$. We know that T^{-1} is a linear operator on the normed linear space $T(X)$ of Y (by Theorem 1.1.2). Since further here T^{-1} is continuous, so there is a number $d > 0$ such that $\left\|T^{-1}y\right\| \leq d \mid y\|$ for all $y \in T(X)$. Now for each $x \in X$ we have $Tx \in T(X)$ and so

$$\left\|T^{-1}(Tx)\right\| \leq d \mid Tx\|, \text{ i.e., } \|x\| \leq d\|Tx\|, \text{ i.e., } \|Tx\| \geq c\|x\|, \text{ where } c = \frac{1}{d} > 0.$$

Conversely, assume that there is a constant $c > 0$ such that $\|Tx\| \geq c\|x\|$ for all $x \in X$. Then for any $x \in X, Tx = 0$ implies $c||x\| \leq \|Tx\| = \|0\| = 0$

$$\begin{aligned} &\Rightarrow \quad \|x\| = 0 \\ &\Rightarrow \quad x = 0. \end{aligned}$$

Therefore, T is one - to - one and hence $T^{-1} : T(X) \to X$ exists, and we know that T^{-1} is a linear operator. Now for all $y \in T(X)$, we have

$$\begin{aligned} &\left\|T\left(T^{-1}y\right)\right\| \geq c\left\|T^{-1}y\right\| \\ &\text{i.e., } \|y\| \geq c\left\|T^{-1}y\right\| \\ &\text{i.e., } \left\|T^{-1}y\right\| \leq \frac{1}{c}\|y\|. \end{aligned}$$

Thus T^{-1} is a bounded linear operator on $T(X)$. Hence T^{-1} is continuous. □

Theorem 16.2.6. Let X and Y are any two normed linear spaces over the same scalar field. If X has finite dimension then every linear operator $T : X \to Y$ is continuous (equivalently bounded).

Proof. Proof. If $\dim X = 0$, then $X = \{0\}$. So the only linear operator $T : X = \{0\} \to Y$ is given by $T0 = 0$, which is trivially continuous.

Let now $\dim X = k \geq 1$. Then X has a basis of k vectors, $\{e_1, e_2, \ldots, e_k\}$, say. Then each $x \in X$ can be represented uniquely as a linear combination

$$x = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_k e_k.$$

$$Tx = T\left(\alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_k e_k\right) = \alpha_1 Te_1 + \alpha_2 Te_2 + \ldots + \alpha_k Te_k$$

So $\|Tx\| = \|\alpha_1 Te_1 + \alpha_2 Te_2 + \ldots + \alpha_k Te_k\|$

$$\begin{aligned} &\leq |\alpha_1| \cdot \|Te_1\| + |\alpha_2| \cdot \|Te_2\| + \ldots + |\ \alpha_k \\ &\leq M\left(|\alpha_1| + |\alpha_2| + \ldots + |\alpha_k|\right). \\ &M = \max\left\{\|Te_1\|, \|Te_2\|, \ldots, \|Te_k\|\right\}. \end{aligned}$$

a constant $\lambda > 0$ such that $\lambda\left(|\alpha_1| + |\alpha_2| + \ldots + |\alpha_k|\right) \leq \|\alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_k e_k\|$ for every set of scalars $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$. So in this case we get $\|Tx\| \leq M \cdot \frac{1}{\lambda}\|\alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_k e_k\| = \frac{M}{\lambda}\|x\|$. Thus $\|Tx\| \leq \frac{M}{\lambda}\|x\|$ for all $x \in X$. Hence T is a bounded linear operator and hence continuous. □

Lemma 16.2.7. Let X and Y are any two normed linear spaces over the same scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). Let $B(X, Y)$ denotes the set of all bounded linear operators $T : X \to Y$. Then $B(X, Y)$ is a linear space over Φ under addition and scalar multiplication defined point wise.

Proof. Let $Ox = 0 \in Y$ for all $x \in X$. Then $O \in B(X, Y)$. i.e., O is a continuous linear operator from $X \to Y$ Let now $T_1, T_2 \in B(X, Y)$ and $\alpha \in \Phi$. We define $T_1 + T_2 : X \to Y$ and $\alpha T_1 : X \to Y$ by the rule $(T_1 + T_2)(x) = T_1 x + T_2 x$ and $(\alpha T_1)(x) = \alpha T_1 x$ for all $x \in X$. Then $(T_1 + T_2)(\alpha x + \beta y) = T_1(\alpha x + \beta y) + T_2(\alpha x + \beta y) = \alpha T_1 x + \beta T_1 y + \alpha T_2 x + \beta T_2 y$

$$\begin{aligned} &= \alpha\left(T_1 x + T_2 x\right) + \beta\left(T_1 y + T_2 y\right) \\ &= \alpha\left(T_1 + T_2\right)(x) + \beta\left(T_1 + T_2\right)(y). \end{aligned}$$

Hence $T_1 + T_2$ is a linear operator. Also since T_1 and T_2 are bounded linear operators, so there are constants $c_1, c_2 > 0$, such that

$$||T_1 x|| \leq c_1||x|| \quad \text{and} \quad ||T_2 x|| \leq c_2||x|| \ \text{ for all } x \in X.$$

So,

$$||(T_1 + T_2)(x)|| = ||T_1 x + T_2 x|| \leq ||T_1 x|| + ||T_2 x|| \leq (c_1 + c_2)||x||.$$

Thus $T_1 + T_2$ is abounded linear operator. Thus $T_1 + T_2 \in B(X, Y)$.

Again,

$$\begin{aligned} (\alpha T_1)(\beta x + \gamma y) &= \alpha T_1(\beta x + \gamma y) \\ &= \alpha(\beta T_1 x + \gamma T_1 y) \\ &= \beta(\alpha T_1 x) + \gamma(\alpha T_1 y) \\ &= \beta(\alpha T_1)(x) + \gamma(\alpha T_1)(y). \end{aligned}$$

Thus, αT_1 is a linear operator.

Also, for all $x \in X$, we have

$$||(\alpha T_1)(x)|| = ||\alpha T_1 x|| = |\alpha|.||T_1(x)|| \leq |\alpha|.c_1||x||.$$

Hence αT_1 is a bounded linear operator on X into Y. So, $\alpha T_1 \in B(X, Y)$.

It is now clear that $B(X, Y)$ is a linear space over Φ under these operations of addition and scalar multiplication. □

Theorem 16.2.8. Let X and Y are any two normed linear spaces over the same scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). Then $B(X, Y)$ is a normed linear space over Φ under the norm defined for all $T \in B(X, Y)$ by

$$||T|| = \sup_{||x|| \leq 1;\ x \in X} ||Tx|| \quad \text{and} \quad ||Tx|| \leq ||T||||x|| \ \text{ for all } \ x \in X.$$

Proof. We know that $B(X,Y)$ is a linear space over Φ under addition and scalar multiplication defined point wise.

Now let $T \in B(X,Y)$. Since T is a bounded linear operator on X, so there is a real number $c > 0$ such that $||Tx|| \leq c||x||$ for all $x \in X$.

In particular, therefore, $||Tx|| \leq c.1$ for all $x \in X$ with $||x|| \leq 1$. Hence it follows that

$$0 \leq ||T|| = \sup\{||Tx|| : \ x \in X, \ ||x|| \leq 1\} \leq c < +\infty.$$

Also we have

$$||T0|| = ||0|| = 0 = ||T||.0 = ||T||.||0||,$$

and for all $x \in X$, $x \neq 0$, we have $\left|\left|\frac{1}{||x||}.x\right|\right| = 1$ and so by definition,

$$\left|\left|T\left(\frac{1}{||x||}.x\right)\right|\right| \leq ||T||$$

$$\text{i.e.,} \quad \frac{1}{||x||}||Tx|| \leq ||T||$$

$$\text{i.e.,} \quad ||Tx|| \leq ||T||||x||.$$

Since this is true even if $x = 0$, we have $||Tx|| \leq ||T||||x||$ for all $x \in X$.

Now for the zero operator $0 \in B(X,Y)$ be such that $||0|| = 0$, since $0x = 0$ for all $x \in X$.

Conversely, if $||T|| = 0$, then by the above $||Tx|| \leq ||T||||x|| = 0.||x|| = 0$ for all $x \in X$. So then $||Tx|| = 0$ for all $x \in X$. Therefore, T is the zero operator, 0.

Next for every $\alpha \in \Phi$, we have

$$||\alpha T|| = \sup_{||x||\leq 1; x\in X} ||(\alpha T)(x)|| = \sup_{||x||\leq 1; x\in X} ||\alpha Tx|| = \sup_{||x||\leq 1; x\in X} |\alpha|||Tx|| = |\alpha|. \sup_{||x||\leq 1; x\in X} ||Tx|| = |\alpha|||T||.$$

Finally, for $T_1, T_2 \in B(X,Y)$ and for all $x \in X$ with $\|x\| \leq 1$, we have

$$\|(T_1 + T_2)(x)\| = \|T_1x + T_2x\| \leq \|T_1x\| + \|T_2x\| \leq \|T_1\|\|x\| + \|T_2\|\|x\| \leq (\|T_1\| + \|T_2\|)\|x\|$$
$$\leq \|T_1\| + \|T_2\| \quad (\because \|x\| \leq 1).$$

Hence taking supremum value for such x it follows from definition that $\|T_1 + T_2\| \leq \|T_1\| + \|T_2\|$. Thus we conclude that $\|T\|$ defines a norm on $B(X,Y)$. So $B(X,Y)$ is a normed linear space under this norm. □

Theorem 16.2.9. If Y is a Banach Space, then so is also $B(X,Y)$.

Proof. We know that $B(X,Y)$ is a linear space over the same scalar field as that of X and Y, under norm defined for all $T \in B(X,Y)$ by

$$\|T\| = \sup_{\|x\|\leq 1; x\in X} \|Tx\|.$$

Now let $\{T_n\}$ be any Cauchy Sequence in $B(X,Y)$. Then for each $x \in X$, we have

$$\|T_mx - T_nx\| = \|(T_m - T_n)(x)\| \leq \|T_m - T_n\|\|x\| \to 0 \text{ as } m, n \to \infty.$$

Therefore, $\{T_n(x)\}_{n=1}^{\infty}$ is a Cauchy Sequence in the Banach Space Y. Hence the sequence converges to a unique element in Y, which we denote by Tx. Thus

$$\lim_{n\to\infty} T_nx = Tx \text{ for all } x \in X.$$

Thus we get a uniquely defined mapping $T : X \to Y,\ x \to Tx$. Now we have $T(\alpha x + \beta y) = \lim_{n\to\infty} T_n(\alpha x + \beta y)$

$$= \lim_{n\to\infty} (\alpha T_n x + \beta T_n y)$$
$$= \alpha Tx + \beta Ty.$$

Thus T is a linear operator. Further, since $\{T_n\}$ is a Cauchy Sequence in $B(X, Y)$, so it is bounded. Thus there is a real number $M > 0$ such that $\|T_n\| \leq M$ for $n = 1, 2, \ldots$ Then for each $x \in X$ and for each $n = 1, 2, \ldots$, we have

$$\|T_n x\| \leq \|T_n\| . \|x\| \leq M\|x\|. \tag{16.2.2}$$

Since $T_n x \to Tx$ in Y and the norm function is continuous (Theorem), so $\|T_n x\| \to \|Tx\|$. Hence letting $n \to \infty$ in (16.2.2), we get $\|Tx\| \leq M\|x\|$ for all $x \in X$. Thus T is a bounded linear operator on X to Y. Thus $T \in B(X, Y)$.

Finally, since $\{T_n\}$ is a Cauchy Sequence, for every $\epsilon > 0$ there is a positive integer N such that $\|T_m - T_n\| < \epsilon$ for $m, n \geq N$.

Thus, for all $x \in X$ with $\|x\| \leq 1$, we have $||(T_m - T_n)(x)|| \leq ||T_m - T_n||||x|| < \epsilon.1$ for all $m, n \geq N$. Since $\lim_{m\to\infty}(T_m - T_n)x = \lim_{m\to\infty} T_m x - T_n x = Tx - T_n x = (T - T_n)x$, so by the continuity of the norm in Y, it follows that $\lim_{m\to\infty} \|T_m x - T_n x\| = \|(T - T_n)x\|$. Hence it follows that $\|(T - T_n)x\| \leq \epsilon$ for all $n \geq N$ and for all $x \in X$ with $\|x\| \leq 1$. So, by definition, $\|T - T_n\| \leq \epsilon$ for all $n \geq N$. Thus, $T_n \to T$ in $B(X, Y)$.

Hence the normed linear space $B(X, Y)$ is complete, i.e., $B(X, Y)$ is a Banach Space. □

Theorem 16.2.10. For each T $\in B(X, Y)$ where $X \neq \{0\}$, we have

$$\|T\| = \sup_{\|x\|\leq 1} \|Tx\| = \sup_{\|x\|=1} \|Tx\| = \sup_{x\neq 0} \frac{\|Tx\|}{\|x\|} = \sup_{0<\|x\|\leq 1} \|Tx\|.$$

Proof. We write

$$\|T\| = \sup_{\|x\|\leq 1} \|Tx\|$$
$$\|T\|_1 = \sup_{\|x\|=1} \|Tx\|$$
$$\|T\|_2 = \sup_{x\neq 0} \frac{\|Tx\|}{\|x\|}$$
$$\|T\|_3 = \sup_{0<\|x\|\leq 1} \|Tx\|.$$

By definition $\|T\| = \sup_{\|x\|\leq 1} \|Tx\|$. Since here $X \neq \{0\}$ there is $x \in X,\ x \neq 0$. If $\|x\| \leq 1$ and $x \neq 0$, then $\left\|\frac{1}{\|x\|} \cdot x\right\| = 1$. So $\left\|T\left(\frac{x}{\|x\|}\right)\right\| \leq \|T\|_1$ i.e., $\frac{1}{||x||}\|Tx\| \leq \|T\|_1$, i.e., $\|Tx\| \leq \|T\|_1\|x\| \leq \|T\|_1.1 = \|T\|_1$. If x $= 0$ then $\|Tx\| = \|0\| = 0 \leq \|T\|_1$. Thus $\|Tx\| \leq \|T\|_1$ for all x with $\|x\| \leq 1$. Therefore $\|T\| \leq \|T\|_1$. But obviously $\|T\|_1 \leq \|T\|$. Hence $\|T\| = \|T\|_1$. Again for any $x \neq 0$, we have $\left\|\frac{1}{\|x\|} \cdot x\right\| = 1$. And so $\left\|T\left(\frac{x}{\|x\|}\right)\right\| \leq \|T\|_1$, i.e., $\frac{1}{\|x\|}\|Tx\| \leq \|T\|_1$.

Therefore, $\|T\|_2 \leq \|T\|_1$. But obviously, $\|T\|_1 \leq \|T\|_2$. Hence, $\|T\|_2 = \|T\|_1$.

We know that $\|Tx\| = 0$ if $x = 0$, but $\|Tx\| > 0$ if $x \neq 0$. Hence $\sup_{0\leq\|x\|\leq 1} \|Tx\| = \sup_{0<\|x\|\leq 1} \|Tx\|$, i.e., $\|T\| = \|T\|_3$. Thus we have $\|T\| = \|T\|_1 = \|T\|_2 = \|T\|_3$. □

Note 16.2.11. $\|T\| = \inf\{M \geq 0 : \|Tx\| \leq M.\|x\|,$ for all $x \in X\}$. Check yourself

Theorem 16.2.12. Let T_1, T_2 be any two bounded linear operators on a normed linear space X into itself, i.e., $T_1, T_2 \in B(X, X)$ and the product of composition $T_1T_2 : X \to X$ be defined by $(T_1T_2)(x) = T_1(T_2(x))$. Then T_1T_2 is a bounded linear operator on X into X and further $\|T_1T_2\| \leq \|T_1\| \|T_2\|$.

Proof. For all $x, y \in X$ and all scalars α, β we have

$$\begin{aligned}(T_1T_2)(\alpha x + \beta y) &= T_1(T_2(\alpha x + \beta y)) \\ &= T_1(\alpha T_2 x + \beta T_2 y) \\ &= \alpha T_1(T_2 x) + \beta T_1(T_2 y) \\ &= \alpha (T_1T_2)(x) + \beta (T_1T_2) y.\end{aligned}$$

Therefore T_1T_2 is a linear operator on X to X.
Next for all $x \in X$, we have

$$\|(T_1T_2)(x)\| = \|T_1(T_2(x))\| \leq \|T_1\| \|T_2(x)\| \leq \|T_1\| \|T_2\| \|x\|.$$

But $\|T\| = \inf\{M \geq 0 : \|Tx\| \leq M.\ \|x\|,\ \text{for all } x \in X\}$. Hence it follows that

$$\|T_1T_2\| \leq \|T_1\| \|T_2\|.$$

□

Example 16.2.13. We consider an example to show that in a normed linear space $T_1T_2 \neq T_2T_1$. Do it yourself.

Theorem 16.2.14. Let $T \in B(X, X)$ and $\|T\| < 1$, where X is a Banach Space. Let I denotes the identity mapping on X. Then prove that the range of $I - T$ is X, $(I - T)^{-1}$ exists and it belongs to $B(X, X)$ with $(I - T)^{-1} = I + T + T^2 + T^3 + \ldots$. Also show that $\left\|(I - T)^{-1}\right\| \leq \dfrac{1}{1 - \|T\|}$.

Proof. Since X is a Banach space, so $B(X, X)$ is also a Banach space.
Now $\{T^n\}_{n=1}^{\infty}$ is a sequence in the Banach space $B(X, X)$ with $\|T^n\| \leq \|T\|^n$ (by ...) Since $0 \leq \|T\| < 1$, so $\sum_{n=1}^{\infty} \|T^n\| \leq \sum_{n=1}^{\infty} ||T||^n = \dfrac{||T||}{1 - ||T||} < +\infty$. Hence, $\sum_{n=1}^{\infty} T^n$ converges in $B(X, X)$ (by ...)
We put $T_0 = I + \sum_{n=1}^{\infty} T^n = I + T + T^2 + \cdots$ Then $T_0 \in B(X, X)$. Also we have

$$TT_0 = T\left(I + T + T^2 + \cdots\right) = TI + T^2 + T^3 + \cdots = T + T^2 + T^3 + \cdots = T_0 - I$$

and

$$T_0T = \left(I + T + T^2 + \cdots\right)T = IT + T^2 + T^3 + \cdots = T + T^2 + T^3 + \cdots = T_0 - I$$

Therefore

$$(I - T)T_0 = IT_0 - TT_0 = T_0 - (T_0 - I) = I$$

and

$$T_0(I - T) = T_0I - T_0T = T_0 - (T_0 - I) = I$$

Hence it follows that $(I - T)^{-1}$ exists and is given by

$$(I - T)^{-1} = I + T + T^2 + T^3 + \ldots$$

We also note that for each $x \in X$ we have

$$((I-T)T_0)(x) = (I-T)(T_0x)$$
$$\text{i. e.,} \quad x = (I-T)(T_0x) \text{ for all } x \in X$$

Thus the range of $(I-T)$ is X. Finally,

$$\left\|(I-T)^{-1}\right\| = \left\|I+T+T^2+T^3+\ldots\right\|$$
$$\leq \|I\| + \|T\| + \left\|T^2\right\| + \cdots \leq 1 + \|T\| + \left\|T^2\right\| + \cdots = \frac{1}{1-\|T\|}.$$

□

Definition 16.2.15. Let $\mathrm{T} : \mathrm{X} \to \mathrm{Y}$ be an operator and let $\mathrm{B} \subset \mathrm{X}$. Then the restriction of T to B written as

$$\mathrm{T}|_{\mathrm{B}} : \mathrm{B} \to \mathrm{Y}$$

is defined as $\mathrm{T}|_{\mathrm{B}}(\mathrm{x}) = \mathrm{T}(\mathrm{x})$ for $\mathrm{x} \in \mathrm{B}$. An extension of T to $Z \supset X$ is an operator

$$\widetilde{\mathrm{T}} : \mathrm{Z} \to \mathrm{Y}$$

such that $\widetilde{\mathrm{T}} \mid \mathrm{x} = \mathrm{T}$, i.e., $\widetilde{\mathrm{T}}(\mathrm{x}) = \mathrm{T}(\mathrm{x})$ for all $\mathrm{x} \in \mathrm{X}$

If X is a proper subset of Z, then T can have many extensions. But we are more interested to study those operators which preserve certain basic properties, for example, linearity or boundedness. The following theorem is an important tool to extend an operator T to the closure of X such that the extended operator is again bounded and linear, and even has the same norm.

Theorem 16.2.16. Let $T : X \to Y$ be a bounded linear operator, where X lies in a normed space Z and Y is a Banach space. Then T has an extension

$$\tilde{T} : \overline{X} \to Y$$

where is a bounded linear operator of norm $\|\tilde{T}\| = \|T\|$.

Proof. We consider any $x \in \overline{X}$. Then there is a sequence $\{x_n\}$ in X such that $x_n \to x$. Since T is linear and bounded, we have

$$||T(x_n) - T(x_m)|| = ||T(x_n - x_m)|| \leq ||T||||x_n - x_m||.$$

Since $\{x_n\}$ is convergent, so $\{T(x_n)\}$ is Cauchy. Since Y is complete, so $\{T(x_n)\}$ converges to some $y \in Y$, say. We define $\tilde{T}$ as

$$\tilde{T}x = y.$$

We show that this definition is independent of the particular choice of a sequence in X converging to x. Suppose that $x_n \to x$ and $z_n \to x$. Then $v_m \to x$ where, $\{v_m\}$ is the sequence $\{x_1, z_1, x_2, z_2, \ldots\}$. Hence $\{T(v_m)\}$ converges and the two subsequences $\{T(x_n)\}$ and $\{T(z_n)\}$ must have the same limit. This proves that $\tilde{T}$ is uniquely defined at every point $x \in X$.

Clearly $\tilde{T}$ is linear and $\tilde{T}(x) = T(x)$ for every $x \in X$, so that $\tilde{T}$ is an extension of T. We now use

$$||T(x_n)|| \leq ||T||||x_n||$$

and let $n \to \infty$. Then $T(x_n) \to y = \tilde{T}(x)$. Since norm is a continuous function, we thus obtain

$$||\tilde{T}(x)|| \leq ||T||||x||.$$

Hence $\tilde{T}$ is bounded and $\|\tilde{T}|| \leq ||T||$. Obviously $\|\tilde{T}|| \geq ||T||$ because the norm, being defined by supremum, cannot decrease in an extension. Combining, we have

$$||\tilde{T}|| = ||T||.$$

Hence proved. □

Exercise 16.2.17.

1. Let $T : X \to Y$ be a linear operator, where X and Y are normed spaces. Show that T is bounded if and only if T maps bounded sets in X into bounded sets in Y
2. If $T = 0$ is a bounded linear operator, show that for any $x \in X$ such that $||x|| < 1,\ ||T(x)|| < \|T\|$.
3. Show that the null space of a bounded linear operator is closed.
4. Let M be a closed linear subspace of a normed linear space X and X/M be the quotient space. Let T be a mapping $T : X \to X/M$ given by $Tx = x + M$ for all $x \in X$. Show that T is a bounded linear operator with $\|T\| \leq 1$.

Unit 17

Course Structure

- Introduction
- Objectives
- Linear functionals
- Hahn-Banach theorem, simple applications.
- Normed conjugate space and separability of the space
- Uniform boundedness principle, simple application.

17.1 Introduction

Let X be any linear space over the scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). A linear operator $f : X \to \Phi$ i.e., a linear operator on X with values in the associated scalar field Φ of X is called a linear functional. We know that Φ is a Banach space over itself under the absolute value norm.

Keeping this in mind all the results proved for general linear operators hold true for all linear functionals. We note that for $f : X \to \Phi$ and $x \in X$, $\|f(x)\| = |f(x)|$.

Let X be a linear space over the scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). Then the space $B(X, \Phi)$ of all bounded linear functionals on X is called the conjugate space of X or the dual space of X; and is denoted by X^*.

Since Φ is a Banach space under the absolute value norm, so by a known result $X^* = B(X, \Phi)$ is also a Banach space.

17.2 Hahn-Banach Theorem

The Hahn-Banach theorem is an extension theorem for linear functionals. This theorem guarantees that a normed linear space is richly supplied with bounded linear functionals and makes possible an adequate theory of dual spaces, which is an essential part of the general theory of normed linear spaces. The theorem was given H. Hahn (1927), rediscovered in its present more general form by S. Banach (1929) and generalized to complex vector spaces by H. F. Bohnenblust and A. Sobczyk (1938). Before starting with the theorem, let us first see a few preliminary results which are required to prove it.

Theorem 17.2.1. Let $M \neq \emptyset$ be a partially ordered set. Suppose that every chain $C \subset M$ has an upper bound. Then, M has at least one maximal element.

In the Hahn-Banach theorem, the object to be extended is a linear functional f which is defined on a subspace Z of a vector space X and has a certain boundedness property which will be formulated in terms of a sublinear functional. By definition, this is a real-valued functional p on a vector space X which is subadditive, i.e.,

$$p(x+y) \leq p(x) + p(y), \ \ \forall x, y \in X,$$

and positive-homogeneous, i.e,

$$p(\alpha x) = \alpha p(x), \ \ \forall \alpha \geq 0 \ \text{ in } \ \mathbb{R} \ \text{ and } \ x \in X.$$

Theorem 17.2.2. (Hanh-Banach Theorem) Let X be a real vector space and p a sub linear functional on X. Furthermore, let $f$be a linear functional which is defined on a subspace Z of X and satisfies

$$f(x) \leq p(x) \, \forall x \in Z.$$

Then f has a linear extension $\tilde{f}$ from Z to X satisfying

$$\tilde{f}(x) \leq p(x) \, \forall x \in X,$$

i.e., $\tilde{f}$ is a linear functional on X, satisfies the above theorem and

$$\tilde{f}(x) = f(x), \ \forall x \in Z.$$

Proof. Let E be the set of all linear extensions g of f satisfying

$$g(x) \leq p(x)$$

on their respective domains. Clearly, $E \neq \emptyset$ since $f \in E$. On E, we can define a partial ordering by $g \leq h \Rightarrow h$ is an extension of g, i.e, by definition, $D(h) \supset D(g)$, where, $D(g)$ and $D(h)$ denote the domains of g and h respectively, and

$$h(x) = g(x) \text{ for all } x \in D(g).$$

For any chain $C \subset E$, we define $\hat{g}$ as

$$\hat{g}(x) = g(x) \text{ if } x \in D(g)$$

for $g \in C, \hat{g}$ is a linear functional, the domain being

$$D(\hat{g}) = \bigcup_{g \in C} D(g),$$

which is a vector space since C is a chain. The definition of $\hat{g}$ is unambiguous. Since for $x \in D(g_1) \cap D(g_2)$ with $g_1, g_2 \in C$, we have $g_1(x) = g_2(x)$ since C is a chain, so that $g_1 \leq g_2$ or $g_2 \leq g_1$.

Clearly, $g \leq \hat{g}$ for all $g \in C$. Hence $\hat{g}$ is an upper bound of C. Since $C \subset E$ was arbitrary, Zom's lemma thus implies that E has a maximal element $\tilde{f}$. By the definition of E, this is a linear extension of f which satisfies

$$\tilde{f}(x) \leq p(x), \quad x \in D(\tilde{f}).$$

This representation is unique. In fact, $y + \alpha y_1 = \tilde{y} + \beta y_1$ with $\tilde{y} \in D(\tilde{f})$ implies $y - \tilde{f} = (\beta - \alpha)y_1$, where, $y - \tilde{y} \in D(\tilde{f})$ whereas, $y_1 / \in D(\tilde{f})$, so that, the only solution is $y - \tilde{y} = 0$ and $\beta - \alpha = 0$. A functional g_1 on Y_1 is defined by

$$g_1(y + \alpha y_1) = \tilde{f}(y) + \alpha c \tag{17.2.1}$$

where c is any real constant. g_1 is linear. Moreover, for $\alpha = 0$ we have, $g_1(y) = \tilde{f}(y)$. Hence g_1 is a proper extension of $\tilde{f}$, that is, an extension such that $D(\tilde{f})$ is a proper subset of $D(g_1)$. So, we can prove that, $g_1 \in E$ by showing that

$$g_1(x) \leq p(x),\ x \in D(g_1), \tag{17.2.2}$$

this will contradict the maximality of $\tilde{f}$, so that $D(\tilde{f}) \neq X$ is false and so, $D(\tilde{f}) = X$.

We will now show that g_1, with a suitable c, satisfies equation (17.2.2). Let us consider $y, z \in D(\tilde{f})$. By the properties of $\tilde{f}$ and p, we get

$$\begin{aligned} \tilde{f}(y) - \tilde{f}(z) &= \tilde{f}(y - z) \\ &\leq p(y - z) \\ &= p\left(y + y_1 - y_1 - z\right) \\ &\leq p\left(y + y_1\right) + p\left(-y_1 - z\right). \end{aligned}$$

Rearranging, we get

$$-p\left(-y_1 - z\right) - \tilde{f}(z) \leq p\left(y + y_1\right) - \tilde{f}(y)$$

where y_1 is fixed. Since y does not appear on the left and z not on the right, the inequality continues to hold if we take the supremum over $z \in D(\tilde{f})$ on the left (call it m_0) and the infimum over $y \in D(\tilde{f})$ on the right, and call it m_1. Then, $m_0 \leq m_1$ and for a c with $m_0 \leq c \leq m_1$ we have, from the above equation

$$-p(-y_1 - z) - \tilde{f}(z) \leq c,\ \forall z \in D(\tilde{f}) \tag{17.2.3}$$

and

$$c < p\left(y + y_1\right) - \tilde{f}(y),\ \forall y \in D(\tilde{f}). \tag{17.2.4}$$

For negative α in equation (17.2.1), we use equation (17.2.3) with $\alpha^{-1}y$ in place of z, and get

$$-p\left(-y_1 - \frac{1}{\alpha}y\right) - \tilde{f}\left(\frac{1}{\alpha}y\right) \leq c,\ \ \forall y \in D(\tilde{f}).$$

Multiplying by $-\alpha > 0$ gives

$$\alpha p\left(-y_1 - \frac{1}{\alpha}y\right) - \tilde{f}(y) \leq -\alpha c,\ \ \forall y \in D(\tilde{f}).$$

From this and (17.2.1), using $y + \alpha y_1 = x$, we obtain

$$\begin{aligned} g_1(x) &= \tilde{f}(y) + \alpha c \\ &\leq -\alpha p\left(-y_1 - \frac{1}{\alpha}y\right) \\ &= p(\alpha y_1 + y) \\ &= p(x). \end{aligned}$$

For $\alpha = 0$, we have $x \in D(\tilde{f})$ and we have nothing to prove. For $\alpha > 0$ we use equation (17.2.4) with y replaced by $\alpha^{-1}y$ to get

$$c \leq p\left(y_1 + \frac{1}{\alpha}y\right) - \tilde{f}\left(\frac{1}{\alpha}\right).$$

Multiplying by $\alpha > 0$ gives

$$\alpha c \leq \alpha p\left(y_1 + \frac{1}{\alpha}y\right) - \alpha\tilde{f}\left(\frac{1}{\alpha}\right) = p(x) - \tilde{f}(y).$$

From this and equation (17.2.1), we get

$$g_1(x) = \tilde{f}(y) + \alpha c \leq p(x).$$

Hence the theorem. □

This was Hanh-Banach theorem in a real vector space. We will now state the theorem for any NLS.

17.2.1 Hahn Banach Theorem

Let X_0 be a linear subspace of a normed linear space X over the same scalar field $\Phi(\mathbb{R}$ or $\mathbb{C})$ and let $f_0 : X_0 \to \Phi$ be a bounded linear functional on X_0. Then f_0 on X_0 can be extended to a bounded linear functional $f : X \to \Phi$ on X such that $\|f\| = \|f_0\|$ and $f(x_0) = f_0(x_0)$ for all $x_0 \in X_0$.

Theorem 17.2.3. Let x_0 be a nonzero vector in a normed linear space X. Then there exists $f \in X^*$ such that $\|f\| = 1$ and $f(x_0) = \|x_0\|$.

Proof. Let X_0 be the linear subspace of X generated by the singleton $\{x_0\}$. Then each $x \in X_0$ has a unique representation $x = \alpha x_0$ for a suitable scalar α. We define a function f_0 on X_0 by $f_0(x) = \alpha||x_0||$, where $x = \alpha x_0$. If $x_1 = \alpha_1 x_0,\ x_2 = \alpha_2 x_0 \in X_0$, then for any two scalars β_1, β_2 we have

$$\begin{aligned} f_0(\beta_1 x_1 + \beta_2 x_2) &= f_0(\beta_1\alpha_1 \mathrm{x}_0 + \beta_2\alpha_2 \mathrm{x}_0) \\ &= f_0(\beta_1\alpha_1 + \beta_2\alpha_2)\,\mathrm{x}_0 \\ &= (\beta_1\alpha_1 + \beta_2\alpha_2)\,\|\mathrm{x}_0\| \\ &= \beta_1\alpha_1\,\|\mathrm{x}_0\| + \beta_2\alpha_2\,\|\mathrm{x}_0\| \\ &= \beta_1 f_0(x_1) + \beta_2 f_0(x_2). \end{aligned}$$

Thus f_0 is a linear functional on X_0. Now for all $\mathrm{x} = \alpha\mathrm{x}_0 \in \mathrm{X}_0$, we have

$$\|f_0(x)|| = |f_0(x)| = |\alpha|\|\,\mathrm{x}_0|| = ||\alpha\mathrm{x}_0|| = 1||x||.$$

Hence f_0 is a bounded linear functional. Also

$$\|f_0\| = \sup_{\|x\|=1;\ x\in X_0} |f_0(x)| = \sup_{\|\alpha x_0\|=1} |\alpha|\|x_0\| = \sup_{\|\alpha x_0\|=1} \|\alpha x_0\| = 1.$$

Finally, $f_0(x_0) = f_0(1.x_0) = 1.||x_0|| = \|x_0||$. Hence by Hahn Banach Theorem, f_0 on X_0 can be extended to a bounded linear functional $f \in X^*$ such that $\|f\| = \|f_0\| = 1$ and $f(x_0) = f_0(x_0) = \|x_0\|$.

This completes the proof. □

Exercise 17.2.4. Prove that

1. If $x_0 \in X$ is such that $f(x_0) = 0$ for all $f \in X^*$ then $x_0 = 0$.
2. If $x, y \in X$ be such that $x \neq y$ then there is an $f \in X^*$ such that $f(x) \neq f(y)$.
3. If $x, y \in X$ be such that $f(x) = f(y)$ for all $f \in X^*$ then $x = y$.

Theorem 17.2.5. For every non zero vector x in a normed linear space X,

$$\|x\| = \sup_{f\in X^*; f\neq 0} \frac{|f(x)|}{\|f\|}.$$

Proof. Since $x \neq 0$, there is $f_1 \in X^*$ such that $\|f_1\| = 1$ and $f_1(x) = \|x\| > 0$. Therefore,

$$\sup_{f\in X^*; f\neq 0} \frac{|f(x)|}{\|f\|} \geq \frac{|f_1(x)|}{||f_1||} = \frac{||x||}{1} = \|x\|.$$

On the other hand, we have

$$\|f(x)\| = |f(x)| \leq \|f\| \cdot \|x\| \text{ for every } f \in X^*.$$
$$\sup_{f\in X^*; f\neq 0} \frac{|f(x)|}{\|f\|} \leq \|x\|.$$

Combining both the cases we have the required result. □

Theorem 17.2.6. Let M be a proper closed subspace of a normed linear space X and let $x_0 \in X \setminus M$. If $\delta = \inf_{x\in M} \|x_0 - x\|$, then there exists $f \in X^*$ such that $\|f\| = 1$, $f(x_0) = \delta$ and $f(x) = 0$ for all $x \in M$.

Proof. Since M is closed in X and $x_0 \in X \setminus M$, then $\delta = \inf_{x\in M} \|x_0 - x\| = dist(x_0, M) > 0$. Let Y denote the linear subspace of X generated by $M \cup \{x_0\}$. Since M is a subspace of X and $x_0 \notin M$, so each $y \in Y$ has a unique representation $y = x + \alpha x_0$, where $x \in M$ and α is a scalar.

We define the scalar valued function f_0 on Y by setting $f_0(y) = \alpha\delta$, where $y = x+\alpha x_0$. For any two points $y_1 = x_1 + \alpha_1 x_0$ and $y_2 = x_2 + \alpha_2 x_0$ of Y and for any two scalars β_1 and β_2, we have $f_0(\beta_1 y_1 + \beta_2 y_2) = f_0(x + (\beta_1\alpha_1 + \beta_2\alpha_2) x_0)$, where $x = \beta_1 x_1 + \beta_2 x_2 \in M$. Thus $f_0(\beta_1 y_1 + \beta_2 y_2) = (\beta_1\alpha_1 + \beta_2\alpha_2)\delta = \beta_1\alpha_1\delta + \beta_2\alpha_2\delta = \beta_1 f_0(y_1) + \beta_2 f_0(y_2)$. So, f_0 is a linear functional on Y.

Consider now any $y = x + \alpha x_0 \in Y$. If $\alpha \neq 0$, then $\frac{1}{a}y = \frac{1}{a}x + x_0 = x_0 - \left(-\frac{1}{\alpha}x\right)$. Since $-\frac{1}{\alpha}x \in M$, we have

$$\left\|\frac{1}{\alpha}y\right\| = \left\|x_0 - \left(-\frac{1}{\alpha}x\right)\right\| \geq \delta.$$

So, $\|y\| \geq |\alpha|\delta = |\alpha\delta|$, which is also true for $\alpha = 0$. Thus $|f_0(y)| = |\alpha\delta| \leq 1||y||$ for all $y \in Y$. Hence f_0 is a bounded linear functional on Y with $\|f_0|| \leq 1$. Also, by definition of δ, there is a sequence $\{x_n\}$ in M such that $\|x_0 - x_n\| \to \delta$. Let us write $y_n = x_n - x_0 = x_n + (-1)x_0$. Then $y_n \in Y$ and by definition $f_0(y_n) = (-1)\delta$ for all n. So,

$$\|f_0\| = \sup_{\substack{y\in Y \\ y\neq 0}} \frac{|f_0(y)|}{\|y\|} \geq \frac{|f_0(y_n)|}{\|y_n\|} \text{ for all } n.$$

Since we cannot have $y_n = 0$ as $x_0 \notin M$. Thus $\|f_0\| \geq \frac{|f_0(y_n)|}{||y_n||} = \frac{|-\delta|}{||x_0 - x_n||} \to \frac{\delta}{\delta} = 1$. So, $\|f_0\| \geq 1$ provides $\|f_0\| = 1$.

Hence by Hahn-Banach theorem, f_0 can be extended to an $f \in X^*$ such that $\|f\| = \|f_0\| = 1$ and $f(y) = f_0(y)$ for all $y \in Y$.

Now for $x \in M$ the representation is $x = x + 0.x_0 \in Y$. So, $f(x) = f_0(x) = 0.\delta = 0$, for $x \in M$. Also since $x_0 = 0 + 1.x_0$, $0 \in M$, so

$$f(x_0) = f_0(x_0) = 1.\delta = \delta.$$

This completes the proof of the theorem. □

Theorem 17.2.7. If the dual space X^* of a normed linear space X is separable, then X is also separable.

Proof. We assume $X \neq \{0\}$, because otherwise there is nothing to prove. Then $X \neq \{0\}$ and so the unit sphere $S = \{f \in X^* : \|f\| = 1\}$ in X^* is non-empty. Since X^* is separable and every non-empty subset of a separable metric space is separable, so there is a countable subset $\{f_n\} \subset S$ which is dense in S.

Now, since $f_n \in S$, so $\|f_n\| = \sup_{\|x\|=1} |f_n(x)| = 1$. Then for each n there is a point $x_n \in X$ such that $\|x_n\| = 1$ and $|f_n(x_n)| > \frac{1}{2}$.

Let Y denote the linear sub space of X generated by the set of vectors $V = \{x_1, x_2, \ldots\}$. We assert that $\overline{Y} = X$. If not, then $\overline{Y}$ is a proper closed subspace of X. So by the previous theorem, there is an $f \in X^*$ such that $\|f\| = 1$ and $f(x) = 0$ for all $x \in \overline{Y}$. Then $f \in S$ and $f(x_n) = 0$ for all n.

Now, since $\{f_n\}$ is dense in S, and $f \in S$, we must have $\|f_n - f\| < \frac{1}{4}$, for at least one n. Then we have

$$\frac{1}{2} < |f_n(x_n)| = |(f_n - f)(x_n) + f(x_n)| = |(f_n - f)(x_n)| \leq ||f_n - f|| \cdot ||x_n|| < \frac{1}{4}.1.$$

This is a contradiction, which proves that $\overline{Y} = X$.

Now, the scalar field $\Phi(\mathbb{R}$ or $\mathbb{C})$ of X is separable, so there is a countable dense subset Ψ of Φ. Let Y_0 denotes the set of all finite linear combinations of the elements of V with coefficients from Ψ.

Since both V and Ψ are countable so Y_0 is also countable. Clearly, then Y_0 is a countable dense subset of Y. i.e., $Y \subset \overline{Y}_0$. Hence $X = \overline{Y} \subset \overline{Y}_0 = \overline{Y}_0$. Hence Y_0 is a countable dense subset of X. Thus the space X is separable. □

Theorem 17.2.8. Banach - Steinhaus Theorem (or Uniform Boundedness Principle). Let $\{T_\lambda\}_{\lambda\in\Lambda}$ be an arbitrary family of bounded linear operators $T_\lambda : X \to Y$, where X is a Banach space and Y is a normed linear space over the same scalar field. If for each $x \in X$ the set $\{T_\lambda(x)\}_{\lambda\in\Lambda}$ is bounded in Y, then $\{||T_\lambda||\}_{\lambda\in\Lambda}$ is bounded in $\mathbb{R}$.

Note 17.2.9. The conclusion of the theorem is that

$$M = \sup_{\lambda\in\Lambda} \|T_\lambda\| < +\infty$$

This means that $\|T_\lambda(x)\| \leq \|T_\lambda\| . \|x\| \leq M.\|x\|$ for all $x \in X$ and for all $\lambda \in \Lambda$. This is the meaning of uniform boundedness of the family $\{T_\lambda\}_{\lambda\in\Lambda}$.

Proof. For each positive integer n, we define

$$X_n = \{x \in X : ||T_\lambda(x)|| \leq n \text{ for all } \lambda \in \Lambda\}.$$

Then $\bigcup_{n=1}^{\infty} X_n \subset X$. On the other hand given any $x \in X$, since the set $\{T_\lambda(x)\}$ is bounded, so there is a positive integer $n = n(x)$ such that $||T_\lambda(x)|| \leq n$ for all $\lambda \in \Lambda$; and so then $x \in X_n$. Thus

$$X = \bigcup_{n=1}^{\infty} X_n. \tag{17.2.5}$$

Now, since the Banach space X is a non-empty complete metric space, so by Baire category theorem, it follows from (17.2.5) that some X_k is dense in some open ball B in X.

Fix any $y \in B$. Since X_k is dense in the open ball B, so there is a sequence $\{x_n\}_{n=1}^{\infty}$ in X_k such that $x_n \to y$. By definition of the set X_k, we have

$$||T_\lambda(x_n)|| \leq k \text{ for all } \lambda \in \Lambda \text{ and } n = 1, 2, \ldots \tag{17.2.6}$$

Since the operators T_λ are continuous and since the norm function is continuous, so considering $n \to \infty$ in (17.2.6), we get

$$\|T_\lambda(y)\| \leq k \text{ for all } \lambda \in \Lambda \text{ and for all } y \in B. \tag{17.2.7}$$

Now, let y_0 be the centre and $r > 0$ be the radius of the open ball B. Fix any $x \in X$ and $x \neq 0$.

Put $y_1 = \dfrac{r}{2||x||}x + y_0$. Then

$$\|y_1 - y_0\| = \frac{r}{2||x||}x = \frac{r}{2}\frac{\|x\|}{\|x\|} = \frac{r}{2} < r.$$

So, $y_1 \in B$. Since T_λ is linear, we have

$$\begin{aligned}
\|T_\lambda(x)\| &= \frac{2\|x\|}{r}\left\|T_\lambda\left(\frac{r}{2||x||}x\right)\right\| \\
&= \frac{2\|x\|}{r}\|T_\lambda(y_1 - y_0)\| \\
&\leq \frac{2\|x\|}{r}(\|T_\lambda(y_1)\| + \|T_\lambda(y_0)\| \\
&\leq \frac{2\|x\|}{r}(k + k) = \frac{4k}{r}\|x\|.
\end{aligned}$$

Since $\|T_\lambda(0)\| = \|0\| = 0$, it follows that $\|T_\lambda(x)\| \leq \dfrac{4k}{r}\|x\|$ for all $x \in X$ and for all $\lambda \in \Lambda$. This completes the proof of the theorem. □

Lemma 17.2.10. Let X be a normed linear space and $x_0 \in X$. Define F_{x_0} on X^* by $F_{x_0}(f) = f(x_0)$ for all $f \in X^*$. Then F_{x_0} is a bounded linear functional on X^*, i.e., $F_{x_0} \in (X^*)^* = X^{**}$. Also $\|F_{x_0}\| = \|x_0\|$.

Proof. By definition, for all $f \in X^*$, we have $F_{x_0}(f) = f(x_0)$. So, F_{x_0} is a scalar valued function on X^*. Now, for $f, g \in X^*$ and for any two scalars α, β we have

$$\begin{aligned}
F_{x_0}(\alpha f + \beta g) &= (\alpha f + \beta g)x_0 \\
&= (\alpha f)(x_0) + (\beta g)(x_0) \\
&= \alpha f(x_0) + \beta g(x_0) \\
&= \alpha F_{x_0}(f) + \beta F_{x_0}(g).
\end{aligned}$$

Thus F_{x_0} is a linear functional on X^*. Further we have

$$|F_{x_0}(f)| = |f(x_0)| \leq ||f|| \cdot ||x_0|| = ||x_0|| \cdot ||f||, \text{ for all } f \in X^*.$$

So, F_{x_0} is a bounded linear functional on X^* with $\|F_{x_0}|| \leq ||x_0||$. If $x_0 = 0$ then $\|x_0|| = 0$, and so $0 \leq ||F_{x_0}|| \leq 0$. Hence $\|F_{x_0}|| = 0 = ||x_0||$. If $x_0 \neq 0$ then by a known theorem we can find a bounded linear functional $f \in X^*$ such that $\|f\| = 1$ and $f(x_0) = \|x_0\|$. Then

$$|f(x_0)| = |F_{x_0}(f)| \leq ||F_{x_0}|| \cdot ||f||. \text{ i.e., } ||x_0|| \leq ||F_{x_0}|| \cdot 1 = ||F_{x_0}||.$$

Since also $\|F_{x_0}\| \leq ||x_0||$, we get $\|F_{x_0}|| = ||x_0||$. This completes the proof of the theorem. □

Theorem 17.2.11. Let E be a subset of a normed linear space X such that for every $f \in X^*$ the set of scalars $f(E)$ is bounded. Then E is bounded in X.

Proof. We know that X^* is a Banach space. Now, for each $x \in E$, let F_x denote the scalar valued function defined on X^* by

$$F_x(f) = f(x), f \in X^*.$$

We know that F_x is a bounded linear functional with $\|F_x\| = \|x\|$. Now, by hypothesis, for each $f \in X^*$, the set

$$\{F_x(f)\}_{x\in E} = \{f(x)\}_{x\in E} = f(E)$$

is bounded. Therefore, by the uniform boundedness principle, the set of norms

$$\{\|F_x\|\}_{x\in E} = \{\|x\|\}_{x\in E} = f(E)$$

is bounded. Hence the set E is bounded in X. □

Exercise 17.2.12.

1. Show that a norm on a vector space X is a sub linear functional on X.
2. Show that a sublinear functional p satisfies $p(0) = 0$ and $p(-x) \geq -p(x)$.
3. If p is a sub linear functional on a vector space X, show that $M = \{x|\ p(x) \leq \gamma,\ \gamma > 0 \text{ fixed}\}$ is a convex set.
4. If p and q are sublinear functionals on a vector space X and a and b are positive constants, then show that $ap + bq$ is sublinear in X.
5. Find all the nowhere (nw) dense sets in a discrete metric space.
6. Show that a subset M of a metric space X is nw dense X if and only if $X \setminus \overline{M}$ is dense in X.
7. Let X be a Banach space and Y a normed space and T_n are bounded linear operators from X to Y. Also, if $\{T_n(x)\}$ is Cauchy in Y for every $x \in X$, then show that $\{||T_n||\}$ is bounded.

Unit 18

Course Structure

- Introduction
- Objectives
- Inner product space.
- Some preliminary results
- Bessel's inequality and its generalisation

18.1 Introduction

In this unit we shall learn about inner product spaces or pre Hilbert spaces and several properties of inner products.

Objectives

After reading this unit, you will be able to

- define an inner product space
- learn certain examples of inner product spaces
- learn certain preliminary definitions and inequalities concerning inner product spaces
- apply them in various appropriate situations
- deduce important formulae and generalisations
- learn Bessel's inequality and its generalisations

18.2 Inner Product Spaces

Definition 18.2.1. Let X be a vector space over the scalar field Φ ($\mathbb{R}$ or $\mathbb{C}$). A mapping $(,) : X \times X \to \Phi$ which assigns to each ordered pair of elements $x, y \in X$, a unique scalar $(x, y) \in \Phi$ is called an inner product or scalar product on Xif the following properties hold for all $x, y, z \in X$ and $\alpha \in \Phi$:

1. $(x + y, z) = (x, z) + (y, z)$
2. $(\alpha x, y) = \alpha(x, y)$
3. $(y, x) = \overline{(x, y)}$ the bar denotes complex conjugate.
4. $(x, x) \geq 0$ and $(x, x) = 0$ if and only if $x = 0$.

The vector space X equipped with the inner product defined on it is called an inner product space or a pre-Hilbert space.

Lemma 18.2.2. An inner product space X has the following properties for all $x, y, z \in X$ and $\alpha, \beta \in \Phi$:

1. $(0, x) = (x, 0) = 0$
2. $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$
3. $(x, \alpha y) = \overline{\alpha}(x, y)$
4. $(x, \alpha y + \beta z) = \overline{\alpha}(x, y) + \overline{\beta}(x, z)$

Proof. 1. We have

$$(0, x) = (0 + 0, x) = (0, x) + (0, x) = 0.$$

So, $(0, x) = 0$. Further $(x, 0) = \overline{(0, x)} = \overline{0} = 0$.

2. $(\alpha x + \beta y, z) = (\alpha x, z) + (\beta y, z) = \alpha(x, z) + \beta(y, z)$.
3. $(x, \alpha y) = \overline{(\alpha y, x)} = \overline{\alpha}\overline{(y, x)} = \overline{\alpha}(x, y)$.
4. $(x, \alpha y + \beta z) = \overline{(\alpha y + \beta z, x)} = \overline{\alpha}\overline{(y, x)} + \overline{\beta}\overline{(z, x)} = \overline{\alpha}(x, y) + \overline{\beta}(x, z)$.

□

Theorem 18.2.3. (Cauchy-Schwarz's Inequality) For any two vectors x, y in an inner product space X we have

$$|(x, y)| \leq \sqrt{(x, x)}.\sqrt{(y, y)}.$$

Equality holds if and only if x and y are linearly dependent.

Proof. If $y = 0$, then x and y are linearly dependent and we have

$$(x, y) = (x, 0) = 0, \quad (y, y) = (0, 0) = 0.$$

So, if $y = 0$, $|(x, y)| = 0 = \sqrt{(x, x)}.\sqrt{(y, y)}$. Now, if $y \neq 0$, but x and y are linearly dependent, then we must have $x = cy$ for some scalar c.

So, $(x, y) = (cy, y) = c(y, y)$ and $(x, x) = (cy, cy) = c\bar{c}(y, y)$. Therefore $|(x, y)| = |c|(y, y)$ and

$$\sqrt{(x, x)} \cdot \sqrt{(y, y)} = |c|\sqrt{(y, y)} \cdot \sqrt{(y, y)} = |c|(y, y).$$

So, $|(x, y)| = \sqrt{(x, x)} \cdot \sqrt{(y, y)}$.

Now, let $y \neq 0$, and x and y are linearly independent. Then $x - \dfrac{(x,y)}{(y,y)}y \neq 0$, and so

$$\begin{aligned} 0 < \left(x - \frac{(x,y)}{(y,y)}y, x - \frac{(x,y)}{(y,y)}y\right) &= (x,x) - \frac{(x,y)}{(y,y)}(x,y) - \frac{(x,y)}{(y,y)}(y,x) + \frac{(x,y)}{(y,y)} \cdot \frac{(x,y)}{(y,y)}(y,y) \\ &= (x,x) - \frac{(x,y)}{(y,y)}(x,y) = (x,x) - \frac{|(x,y)|^2}{(y,y)}. \\ \text{i.e.,} \quad |(x,y)|^2 &< (x,x).(y,y). \end{aligned}$$

Therefore $|(x,y)| < \sqrt{(x,x)} \cdot \sqrt{(y,y)}$.

Thus the require inequality is proved. □

Theorem 18.2.4. An inner product space X is a normed linear space under the norm defined by

$$||x|| = +\sqrt{(x,x)} \text{ for all } x \in X.$$

Proof. By definition of inner product, we have
$0 \leq (x,x) < +\infty$ for all $x \in X$ and $(x,x) = 0$ if and only if $x = 0$.
$\therefore 0 \leq ||x|| < +\infty$ for all $x \in X$ and $||x|| = 0$ if and only if $x = 0$.

Also, for any scalar α, we have

$$||\alpha x|| = \sqrt{(\alpha x, \alpha x)} = \sqrt{\alpha\bar{\alpha}(x,x)} = \sqrt{|\alpha|^2 \cdot \|x\|^2} = |\alpha| \cdot ||x||.$$

Finally, for $x, y \in X$ we have

$$\begin{aligned} \|x+y\| = \sqrt{(x+y, x+y)} &= \sqrt{(x, x+y) + (y, x+y)} \\ &= \sqrt{(x,x) + (x,y) + (y,x) + (y,y)} \\ &= \sqrt{\|x\|^2 + (x,y) + \overline{(y,x)} + \|y\|^2}. \end{aligned}$$

Now, $(x,y) + \overline{(x,y)} = 2Rl(x,y) \leq 2|(x,y)| \leq 2\sqrt{(x,x)} \cdot \sqrt{(y,y)} = 2\|x\| \cdot \|y\|$. Therefore we get

$$\|x+y\| \leq \sqrt{\|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2} = \|x\| + \|y\|.$$

Hence, $\|\cdot\|$ a norm on X. Hence X is a normed linear space under this norm. □

Note 18.2.5. The metric induced by the inner product is given by

$$d(x,y) = \|x - y\| = \sqrt{(x-y, x-y)}.$$

Example 18.2.6.

1. The space $\mathbb{R}^n$ is an inner product with the inner product defined as

$$(x,y) = x_1y_1 + x_2y_2 + \ldots + x_ny_n$$

where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are two elements of $\mathbb{R}^n$.

2. The unitary space C^n is an inner product space where the inner product is defined as

$$(x,y) = x_1\overline{y}_1 + x_2\overline{y}_2 + \ldots + x_n\overline{y}_n$$

where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ are two elements of C^n.

3. The l_2 space forms an inner product space with the inner product defined as

$$(x, y) = \sum_{i=1}^{\infty} x_i \overline{y}_i.$$

The convergence of the series follows from the Cauchy-Schwarz inequality.

Theorem 18.2.7. Polarization Identity For all x, y in a real inner product space X we have

$$(x, y) = \frac{1}{4}\left[\|x + y\|^2 - \|x - y\|^2\right].$$

For all x, y in a complex inner product space X we have

$$\begin{aligned} Rl(x, y) &= \frac{1}{4}\left[\|x + y\|^2 - \|x - y\|^2\right] \\ Im(x, y) &= \frac{1}{4}\left[\|x + iy\|^2 - \|x - iy\|^2\right]. \end{aligned}$$

Proof. In any inner product space, we have

$$\begin{aligned} \|x + y\|^2 - \|x - y\|^2 &= (x + y, x + y) - (x - y, x - y) \\ &= (x, x) + (x, y) + (y, x) + (y, y) - (x, x) + (x, y) + (y, x) - (y, y) \\ &= = 2[(x, y) + \overline{(x, y)}] = 4Rl(x, y). \end{aligned} \tag{18.2.1}$$

Now, if X is a real inner product space, then $Rl(x, y) = (x, y)$, and so by (18.2.1)

$$\|x + y\|^2 - \|x - y\|^2 = 4(x, y),$$

which proves the first result.

If X is a complex inner product space, then changing y to iy in (18.2.1), we get

$$\begin{aligned} &\|x + iy\|^2 - \|x - iy\|^2 = 4Rl(x, iy) = 2[(x, iy) + \overline{(x, iy)}] = 2[\bar{i}(x, y) + i\overline{(x, y)}] \\ = \;& 2[-i(x, y) + i\overline{(x, y)}] = 2i[-(x, y) + \overline{(x, y)}] = 2i(-2i)Im(x, y) = 4Im(x, y). \end{aligned}$$

This together with (18.2.1) proves the second part. □

Theorem 18.2.8. (Parallelogram Law) For all x, y in an inner product space X we have

$$||x + y||^2 + ||x - y||^2 = 2\left[||x||^2 + ||y||^2\right].$$

Proof. Do it yourself. □

Note 18.2.9. A normed linear space in which the parallelogram law holds is necessarily an inner product space, i.e., the norm is induced by some inner product.

Theorem 18.2.10. The inner product in an inner product space X is continuous on $X \times X$.

Proof. Do it yourself. □

18.3 Orthogonality

Let X be an inner product space.

1. Two vectors x, y in X are said to be orthogonal, denoted by $x \perp y$, if $(x, y) = 0$, equivalently, $(y, x) = 0$, since $(y, x) = \overline{(x, y)}$.
2. A vector x in X is said to be orthogonal to a subset $M \subset X$, denoted by $x \perp M$, if $x \perp y$ for all $y \in M$.
3. A subset $M \subset X$ is said to be orthogonal to a subset $N \subset X$, denoted by $M \perp N$, if $x \perp y$ for all $x \in M$ and $y \in N$.
4. A subset $M \subset X$ is said to be an orthogonal set if $x \perp y$ for all $x, y \in M$ with $x \neq y$.
5. If $M \subset X$ is an orthogonal set of vectors such that $\|x\| = 1$ for all $x \in M$, then M is called orthonormal.
6. vi) A sequence $\{e_n\}_{n=1}^{\infty}$ in X is said to be orthogonal if $e_i \perp e_j$ for $i \neq j$. If further $\|e_n\| = 1$, then the sequence $\{e_n\}_{n=1}^{\infty}$ is called orthonormal.

Theorem 18.3.1. (Pythagoras Theorem.) If x and y are orthogonal vectors in an inner product space, then

$$||x + y||^2 = ||x||^2 + ||y||^2.$$

Proof. Do it yourself. □

Theorem 18.3.2. Every orthogonal set of non-zero vectors in an inner product space is linearly independent.

Proof. Do it yourself. □

Theorem 18.3.3. Let M be an orthonormal set of vectors in an inner product space X. Then

$$||x + y|| = ||x - y|| = \sqrt{2}$$

For all $x, y \in M$ with $x \neq y$. If X is separable then M is countable.

Proof. Do the first part yourself.

Now, let us assume that the space X is separable. Since every subspace of a separable metric space is separable, so M is separable. Then M has a countable dense subset E, say. Then for each $x \in M$, there exists $y \in E$ such that $||x - y|| < \sqrt{2}$.

Since, $x, y \in M$, the preceding result implies $x = y$. Thus $x = y \in E$ for all $x \in M$. Hence $M = E$, so that M is countable. □

Theorem 18.3.4. If $(x, u) = (x, v)$ for all x in an inner product space, then $u = v$.

Proof. Do it yourself. □

Theorem 18.3.5. Let $\{e_1, e_2, \ldots e_n\}$ be an orthonormal set in an inner product space X. Then

$$\sum_{i=1}^{k} |(x, e_j)|^2 \leq \|x\|^2$$

for all $x \in X$.

If $\{e_n\}$ is an orthonormal sequence in X then

$$\sum_{n=1}^{\infty} |(x, e_n)|^2 \leq \|x\|^2$$

for all $x \in X$.

Note 18.3.6. The scalars (x, e_n) 's are called the Fourier expansion of x with respect to $\{e_1, e_2, \ldots\ldots\}$.

Proof. We have

$$\begin{aligned} 0 &\leq \left(x - \sum_{j=1}^{k} (x, e_j)\, e_j, x - \sum_{j=1}^{k} (x, e_j)\, e_j \right) \\ &= (x, x) - \sum_{j=1}^{k} \overline{(x, e_j)}\, (x, e_j) - \sum_{j=1}^{k} (x, e_j)\, (e_j, x) + \sum_{i=1}^{k} \sum_{j=1}^{k} (x, e_i)\, \overline{(x, e_j)}\, (e_i, e_j) \\ &= \|x\|^2 - \sum_{j=1}^{k} |x, e_j|^2 - \sum_{j=1}^{k} (x, e_j)\, \overline{(x, e_j)} + \sum_{j=1}^{k} (x, e_j)\, \overline{(x, e_j)} \cdot 1 \end{aligned}$$

[Since $(e_i, e_j) = 0$ for $i \neq j$ and $\|e_i\| = 1$]. Thus $\sum_{j=1}^{k} |(x, e_j)|^2 \leq \|x\|^2$. Now, if $\{e_n\}$ is an orthonormal sequence in X then by the above

$$\sum_{j=1}^{\infty} |(x, e_j)|^2 = \lim_{k \to \infty} \sum_{j=1}^{k} |(x, e_j)|^2 \leq \| x|^2 .$$

□

Theorem 18.3.7. (Generalized Bessel's Inequality.) Let $\{e_1, e_2, \ldots e_n\}$ be any orthonormal set in an inner product space X. Then for all $x, y \in X$ we have

$$\sum_{n} \left| (x, e_n) \overline{(y, e_n)} \right| \leq \|x\| \cdot \|y\|.$$

Proof. Do it yourself. □

Exercise 18.3.8.

1. Show that every inner product space is a NLS.
2. If an inner product space X is real, show that the condition $\|x\| = \|y\|$ implies that $(x + y, x - y) = 0$.
3. Show that in an inner product space X, $x \perp y$ if and only if $\|x + \alpha y\| \geq \|x\|$ for all scalars α.
4. that in an inner product space X, $x \perp y$ if and only if $\|x + \alpha y\| = \|x - \alpha y\|$ for all scalars α.

Unit 19

Course Structure

- Introduction
- Objectives
- Hilbert's space
- Orthogonal complement
- Projection theorem

19.1 Introduction

In this unit, we will study the Hilbert spaces, which is nothing but the analogous of Banach spaces in an inner product space. We will study related concepts.

Objectives

After reading this unit, you will be able to

- define Hilbert spaces
- give examples of Hilbert spaces
- apply the related concepts in various cases
- define orthogonal complement
- learn the projection theorem

19.2 Hilbert Spaces

Definition 19.2.1. A Hilbert space is an inner product space H which is a Banach space under the norm induced by the inner product, i.e., which is complete with respect to the metric

$$d(x, y) = ||x - y|| = \sqrt{(x - y, x - y)}.$$

Example 19.2.2. Let $\Phi = \mathbb{R}$ or $\mathbb{C}$. The Euclidean space Φ^k is a Hilbert space under the inner product of $x = (\xi_1, \xi_2, \ldots, \xi_k)$ and $y = (\eta_1, \eta_2, \ldots, \eta_k)$ defined by

$$(x, y) = \xi_1\overline{\eta}_1 + \xi_2\overline{\eta}_2 + \ldots + \xi_k\overline{\eta}_k.$$

It is easy to verify that this is indeed an inner product on Φ^k. The norm induced by this inner product is given by $\|x\| = \sqrt{(x, x)} = \sqrt{\xi_1\overline{\xi}_1 + \xi_2\overline{\xi}_2 + \ldots + \xi_k\overline{\xi}_k} = \sqrt{|\xi_1|^2 + |\xi_2|^2 + \ldots + |\xi_k|^2}$ which defines a Euclidean norm.

We know that Φ^k is a Banach space under this norm. Hence Φ^k is a Hilbert space.

Example 19.2.3. The sequence space ℓ_2 is a is a separable Hilbert space of infinite dimension under the inner product of $x = \{\xi_j\}$ and $y = \{\eta_j\} \in \ell_2$ defined by

$$(x, y) = \sum_{j=1}^{\infty} \xi_j\overline{\eta}_j.$$

We first observe that for $x = \{\xi_j\}$ and $y = \{\eta_j\} \in \ell_2$, by definition, $\sum_{j=1}^{\infty} |\xi_j|^2 < +\infty$ and $\sum_{j=1}^{\infty} |\eta_j|^2 < +\infty$. Therefore, by Holder's inequality,

$$\sum_{j=1}^{\infty} \xi_j\overline{\eta}_j \leq \left(\sum_{j=1}^{\infty} |\xi_j|^2\right)^{1/2} \left(\sum_{j=1}^{\infty} |\eta_j|^2\right)^{1/2} < +\infty.$$

Thus the inner product $(x, y) = \sum_{j=1}^{\infty} \xi_j\overline{\eta_j}$ is a well defined scalar. It is easy to verify that the above definition defines an inner product on ℓ_2. The norm induced by this inner product is given by

$$\|x\| = \sqrt{(x, x)} = \sqrt{\sum_{j=1}^{\infty} \xi_j\overline{\xi_j}} = \sqrt{\sum_{j=1}^{\infty} |\xi_j|^2}.$$

But we know that ℓ_2 is a Banach space under this norm Hence ℓ_2 is a Hilbert space. Also, we know that ℓ_2 is separable and infinite dimensional.

Example 19.2.4. Not every Banach space is a Hilbert space.

1. For $p \neq 2$, ℓ_p is a Banach space but not a Hilbert space.

   We know that ℓ_p is a Banach space under the norm defined by $||x|| = \left(\sum_{j=1}^{\infty} \left|x^{(j)}\right|^p\right)^{1/p}$ for all $x = \left\{x^{(j)}\right\} \in \ell_p$

   Now, let $x = (1, 1, 0, 0, \ldots)$ and $y = (1, -1, 0, 0, \ldots)$. Then $||x|| = 2^{1/p}$ and $||y|| = 2^{1/p}$.

   Also, $\|x + y|| = 2$ and $\|x - y|| = 2$. So, $\|x + y\|^2 + \|x - y\|^2 = 2^2 + 2^2 = 8$. Whereas, $2\left(\|x\|^2 + \|y\|^2\right) = 2\left(2^{2/p} + 2^{2/p}\right) = 4 \cdot 2^{2/p} \neq 8$ for $p \neq 2$. Thus the parallelogram law fails in ℓ_p for $p \neq 2$. Hence the space ℓ_p is not a Hilbert space for $p \neq 2$.

2. The Banach space $C[a,b]$ is not a Hilbert space.

   We know that $C[a,b]$ is a Banach space under the norm defined by $\|x\| = \sup_{a\le t\le b}|x(t)|$. Let us take $x(t)=1$ and $y(t)=\dfrac{t-a}{b-a}\in C[a,b]$. Then $\|x\|=1$ and $\|y\|=1$. Also

$$||x+y|| = \sup_{a\le t\le b}\left|1+\frac{t-a}{b-a}\right| = 2 \quad \text{and} \quad ||x-y|| = \sup_{a\le t\le b}\left|1-\frac{t-a}{b-a}\right| = 1.$$

   So, $\|x+y\|^2+\|x-y\|^2 = 2^2+1^2 = 5$. Whereas, $2\left(||x||^2+||y||^2\right) = 2\left(1^2+1^2\right) = 4 \neq 5$. Thus the parallelogram law fails in $C[a,b]$. Hence the space $C[a,b]$ is not a Hilbert space.

Theorem 19.2.5. Let M be a non empty closed and convex subset of a Hilbert space H. Then for each $x_0 \in H$ there is a unique $y_0 \in M$ such that $\|x_0 - y_0\| = \text{dist}\,(x_0, M)$. In particular, M contains a unique vector of smallest norm.

Proof. Let $\delta = \text{dist}\,(x_0,M) = \inf_{y\in M}||x_0-y|| \ge 0$. If $u,v\in M$, then $\dfrac{1}{2}(u+v)\in M$, since M is convex. So, by the parallelogram law, we have

$$\begin{aligned}
\|u-v\|^2 &= \|(u-x_0)-(v-x_0)\|^2 \\
&= 2\left(\|u-x_0\|^2-\|v-x_0\|^2\right) - \|(u-x_0)+(v-x_0)\|^2 \\
&= 2\left(\|u-x_0\|^2-\|v-x_0\|^2\right) - 4||\frac{1}{2}(u+v)-x_0||^2 \\
&\le 2\left(\|u-x_0\|^2-\|v-x_0\|^2\right) - 4\delta^2 \qquad (19.2.1)
\end{aligned}$$

Now, there is a sequence $\{y_n\}$ in M such that $\|x_0-y_n\|\to\delta$. Then by (19.2.1), we have

$$\|y_m-y_n\|^2 \le 2\left(\|y_m-x_0\|^2-\|y_n-x_0\|^2\right) - 4\delta^2 \to 2\left(\delta^2+\delta^2\right)-4\delta^2 = 0$$

So, $\{y_n\}$ is a Cauchy sequence in $M\subset H$. Since, the Hilbert space H is complete, so, there is a point $y_0\in H$ such that $y_n\to y_0$. Since M is closed, $y_0\in M$.

Also, since $y_n-x_0\to y_0-x_0$, $\|x_0-y_n\|\to\|x_0-y_0\|$. Therefore, $\|x_0-y_0\| = \delta = \text{dist}(x_0,M)$. To prove the uniqueness of y_0, assume that

$$\|x_0-y\| = \delta \text{ for some } y\in M.$$

Then by (19.2.1)

$$0\le\|y-y_0\|^2 \le 2\left(\|y-x_0\|^2-\|y_0-x_0\|^2\right)-4\delta^2 = 2(\delta^2+\delta^2)-4\delta^2 = 0.$$

So, $\|y-y_0\|=0$, and hence $y=y_0$.

Finally, taking $x_0=0$, we see that $y_0\in M$ is the unique point such that

$$\begin{aligned}
&\|x_0-y_0\| = \|0-y_0\| = \|y_0\| = \text{dist}(0,M) \\
\Rightarrow\quad &\|y_0\| = \inf_{y\in M}\|0-y_0\| = \inf_{y\in M}\|y_0\|.
\end{aligned}$$

So, $y_0\in M$ has the smallest norm for all $y\in M$. This completes the proof of the theorem. □

Theorem 19.2.6. Let M be a closed linear sub space of a Hilbert space H. Then for each $x_0\in H$ there is a unique $y_0\in M$ such that $\|x_0-y_0\| = \inf_{y\in M}\|x_0-y\| = \text{dist}\,(x_0,M)$ and $(x_0-y_0)\perp M$.

Proof. The linear sub space M of the Hilbert space H is non-empty and convex. Since further M is closed, then by the previous theorem for each $x_0 \in H$ there is a unique $y_0 \in M$ such that $\|x_0 - y_0\| = \text{dist}\,(x_0, M) = \inf_{y\in M} ||x_0 - y|| = \delta(\text{say}) \geq 0$. We now show that $x_0 - y_0$ is orthogonal to M, i.e., $(x_0 - y_0, y) = 0$ for all $y \in M$. If $y = 0$, then $(x_0 - y_0, y) = (x_0 - y_0, 0) = 0$. Now, suppose, $y \neq 0$, then $(y, y) > 0$. Since $y_0, y \in M$ and M is a linear sub space of H. For any scalar α, we have $y_0 + \alpha y \in M$. Therefore,

$$\begin{aligned}\delta^2 \leq \|x_0 - (y_0 + \alpha y)\|^2 &= ((x_0 - y_0) - \alpha y, (x_0 - y_0) - \alpha y) \\ &= (x_0 - y_0, x_0 - y_0) - \bar{\alpha}\,(x_0 - y_0, y) - \alpha\,(y, x_0 - y_0) + \alpha\bar{\alpha}(y, y) \\ &= \delta^2 + \bar{\alpha}(y, y)\left(\alpha - \frac{(x_0 - y_0, y)}{(y, y)}\right) - \alpha\overline{(x_0 - y_0, y)}.\end{aligned}$$

Now, taking $\alpha = \dfrac{(x_0 - y_0, y)}{(y, y)}$, we get

$$\delta^2 \leq \delta^2 + 0 - \frac{(x_0 - y_0, y)}{(y, y)}\overline{(x_0 - y_0, y)}, \text{ i.e., } 0 \leq |(x_0 - y_0, y)|^2 \leq 0$$

Thus $(x_0 - y_0, y) = 0$. This completes the proof of the theorem. □

19.2.1 Orthogonal Complements and Direct Sums

Definition 19.2.7. A linear space X is said to be the direct sum of two subsets Y and Z, written as $X = Y \oplus Z$, if each $x \in X$ has a unique representation $x = y + z$ with $y \in Y$ and $z \in Z$.

Definition 19.2.8. The orthogonal complement of a subset Y of an inner product space X is defined as

$$Y^{\perp} = \{x \in X : x \perp Y\}.$$

Exercise 19.2.9. Show that $Y^{\perp}$ is a closed linear subspace of X even if Y is not.

Note 19.2.10. Since $(0, Y) = 0$, for all $y \in Y$, so $0 \in Y^{\perp}$. Now, if $x, y \in Y^{\perp}$ and α, β are any two scalars, then for all $z \in Y$, we have

$$(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z) = \alpha.0 + \beta.0 = 0.$$

So, $\alpha x + \beta y \in Y^{\perp}$. Thus $Y^{\perp}$ is a linear subspace of X even if Y is not.

Note 19.2.11. If $\{x_n\}$ is a sequence in $Y^{\perp}$ and if $x_n \to x \in X$. Then by the continuity of the inner product we have $(x_n, y) \to (x, y)$ for all $y \in Y$. So for $y \in Y$, $(x, y) = \lim_{n\to\infty}(x_n, y) = \lim_{n\to\infty} 0 = 0$. Therefore, $x \in Y^{\perp}$ and so $Y^{\perp}$ is closed.

Theorem 19.2.12. (Projection Theorem). For any closed linear subspace M of a Hilbert space H, we have $H = M \oplus M^{\perp}$.

Proof. Since M is a closed linear sub space of the vector space H, then by the previous theorem, for each $x \in H$ there is a unique $y \in M$ such that $(x - y) \perp M$.

Let us put $z = x - y$. Then $x = y + z$, where $y \in M$ and $z \in M^{\perp}$. To prove the uniqueness of this representation of the elements $x \in H$ assume that $x = y_1 + z_1$, where $y_1 \in M$ and $z_1 \in M^{\perp}$. Then

$$y + z = y_1 + z_1, \text{ i.e., } y - y_1 = z_1 - z.$$

Since both M and $M^{\perp}$ are linear subspaces of H, so $y - y_1 \in M$ and $z_1 - z \in M^{\perp}$, Hence $(y - y_1, z_1 - z) = 0$, i,e., $(y - y_1, y - y_1) = 0$. So, by definition of inner product, $y - y_1 = 0$ and hence $z_1 - z = 0$. Thus $y_1 = y$ and $z_1 = z$. Hence the representation of x as above is unique □

Note 19.2.13. In the above representation $x = y + z$, the unique vector $y \in M$ is called the orthogonal projection of the vector $x \in H$ on the closed subspace M.

Theorem 19.2.14. A linear subspace M of a Hilbert space H is closed in H if and only if $M = M^{\perp\perp}$.

Proof. We know that $M^{\perp} = \left(M^{\perp}\right)^{\perp}$ is a closed linear sub space of H. Hence if $M = M^{\perp\perp}$ then M is closed.

Conversely, let M be closed in H. If $x \in M$, then by definition of $M^{\perp}$, we have $x \perp M^{\perp}$, and so, $x \in M^{\perp\perp}$. Thus, $M \subset M^{\perp\perp}$. Next, let $y \in M^{\perp\perp}$. Since M is a closed linear subspace of H, so by projection theorem we can write

$$y = u + v, \text{ where } u \in M \subset M^{\perp\perp} \text{ and } v \in M^{\perp}.$$

Since $y, u \in M^{\perp\perp}$, so $y - u \in M^{\perp\perp}$. Thus $v \in M^{\perp}$ and also $v \in M^{\perp\perp}$. So $(v, v) = 0$ and hence $v = 0$. So, $y = u \in M$. Therefore, $M^{\perp\perp} \subset M$ which together with $M \subset M^{\perp\perp}$ provides $M = M^{\perp\perp}$.

This completes the proof of the theorem. □

Exercise 19.2.15.

1. Show that $y \perp x_n$ and $x_n \to x$ together imply $x \perp y$.
2. Show that for a sequence $\{x_n\}$ in an inner product space the conditions $\|x_n\| \to \|x\|$ and $\langle x_n, x\rangle \to \langle x, x\rangle$ imply $x_n \to x$
3. Show that for any set $M = \emptyset$, the set $M^{\perp}$ is a vector space.
4. Consider a subset M of $\mathbb{R}^2$. Find $M^{\perp}$ if M is
   (a) $\{x\}$ where $x = (\xi_1, \xi_2) = 0$,
   (b) a linearly independent set $\{x_1, x_2\} \subset \mathbb{R}^2$.
5. Let A and $B \supset A$ be non-empty subsets of an inner product space X. Then show that
   (a) $A \subset A^{\perp}$,
   (b) $B^{\perp} \subset A^{\perp}$
   (c) $A^{\perp\perp\perp} = A^{\perp}$.
6. Let $M = \emptyset$ is any subset of a Hilbert space H, show that $M^{\perp\perp}$ is the smallest closed subspace of H which contains M.
7. Let $\{x_n\}$ be a sequence in a Hilbert space H and $x \in H$ is such that $\lim_{n\to\infty} ||x_n|| = ||x||$, and $\lim_{n\to\infty} ||(x_n, x)|| = ||(x, x)||$, show that $\lim_{n\to\infty} x_n = x$.

Unit 20

Course Structure

- Introduction
- Objectives
- Riesz Representation theorem
- Convergence of series corresponding to orthogonal sequence
- Fourier coefficient

20.1 Introduction

Orthogonality of elements as defined in the preceding units plays a basic role in inner product and Hilbert spaces. Of particular interest are sets whose elements are orthogonal in pairs. In this unit, we will mainly focus on the concepts of orthonormal sets and sequences and their explicit applications.

Objectives

After reading this unit, you will be able to

- learn the Riesz Representation theorem
- apply the Riesz Representation theorem
- learn the concepts of orthogonal sets and sequences
- learn their properties
- give examples of each
- learn about the series related to the orthonormal sequences and sets
- apply them in different circumstances

20.2 Riesz Representation Theorem

Theorem 20.2.1. (Riesz Representation Theorem) For each bounded linear functional f on a Hilbert space H there is a unique vector $z \in H$ such that $f(x) = (x, z)$ for all $x \in H$; and further $\|f\| = \|z\|$.

Proof. **Existence of z.**

If $f = 0$, then $f(x) = 0 = (x, 0)$ for all $x \in H$. So, $z = 0$ is the solution in this case. Now, assume $f \neq 0$. Then the null space of $f, \mathcal{N}(f) = \{x \in H : \ f(x) = 0\}$ is a proper closed linear subspace of the Hilbert space H. By the projection theorem, we have $H = \mathcal{N}(f) \oplus \mathcal{N}(f)^{\perp}$. Since $\mathcal{N}(f) \neq H$, we have $\mathcal{N}(f)^{\perp} \neq \{0\}$. We select $z_0 \in \mathbb{N}(f)^{\perp}, z_0 \neq 0$, so that $(z_0, z_0) > 0$. Since f is linear, so for all $x \in H$, we have

$$f\left(f(x)z_0 - f\left(z_0\right)x\right) = f(x)f\left(z_0\right) - f\left(z_0\right)f(x) = 0.$$

Therefore,

$$f(x)z_0 - f\left(z_0\right)x \in \mathcal{N}(f).$$

Since, $z_0 \in \mathcal{N}(f)^{\perp}$, so,

$$0 = \left(f(x)z_0 - f\left(z_0\right)x, z_0\right) = f(x)\left(z_0, z_0\right) - f\left(z_0\right)\left(x, z_0\right).$$

So, $f(x) = \dfrac{f\left(z_0\right)\left(x, z_0\right)}{\left(z_0, z_0\right)} = (x, z)$, where $z = \dfrac{\overline{f\left(z_0\right)}}{\left(z_0, z_0\right)}z_0$. Hence $z = \dfrac{\overline{f\left(z_0\right)}}{\left(z_0, z_0\right)}z_0$ is the required in this case.

z is unique and $\|f\| = \|z\|$

Suppose that $f(x) = (x, z_1)$ for all $x \in H$. Then $f(x) = (x, z) = (x, z_1)$. Hence $z_1 = z$. This proves the uniqueness. Next, by Cauchy Schwarz's inequality, for all $x \in H$, we have

$$|f(x)| = |(x, z)| \leq ||x||||z|| = ||z||||x||.$$

So, $\|f\| \leq \|z\|$. On the other hand, we have

$$||z||^2 = (z, z) = f(z) \leq ||f||||z||.$$

So, if $\|z\| \neq 0$, then $\|z\| \leq \|f\|$. If $\|z\| = 0$, then $\|f\| \leq \|z\|$ provides that $\|f\| = 0 = \|z\|$. Hence, in all cases $\|f\| = \|z\|$. □

Exercise 20.2.2. Let X be an inner product space and let $y \in X$ be fixed. Define $f(x) = (x, y)$ for all $x \in X$. Show that f is a bounded linear functional on X with $\|f\| = \|y\|$.

20.2.1 Convergence of series corresponding to orthogonal sequence

Theorem 20.2.3. Let $\{e_n\}$ be an orthonormal sequence in a Hilbert space H and let $\{\alpha_n\}$ be any sequence of scalars. Then the series $\displaystyle\sum_{n=1}^{\infty} \alpha_n e_n$ converges in H if and only if the series $\displaystyle\sum_{n=1}^{\infty} |\alpha_n|^2$ converges in $\mathbb{R}$.

Also, if $\displaystyle\sum_{n=1}^{\infty} \alpha_n e_n = x \in H$, then $(x, e_n) = \alpha_n$ for all n and $\|x\|^2 = \displaystyle\sum_{n=1}^{\infty} |\alpha_n|^2$.

Proof. Let $s_n = \sum_{n=1}^{j} \alpha_j e_j$ and $\sigma_n = \sum_{n=1}^{j} |\alpha_j|^2$. Since $\{e_n\}$ is an orthonormal sequence, obviously, $\{\alpha_n e_n\}$ is an orthogonal sequence. Hence for all $m > n$, we have by Pythagoras theorem,

$$\begin{aligned} \|s_m - s_n\|^2 &= \|\alpha_{n+1}e_{n+1} + \alpha_{n+2}e_{n+2} + \cdots + \alpha_m e_m\|^2 \\ &= \|\alpha_{n+1}e_{n+1}\|^2 + \|\alpha_{n+2}e_{n+2}\|^2 + \cdots + \|\alpha_m e_m\|^2 \\ &= |\alpha_{n+1}|^2 1^2 + |\alpha_{n+2}|^2 1^2 + \cdots + |\alpha_m|^2 1^2 \\ &= |\alpha_{n+1}|^2 + |\alpha_{n+2}|^2 + \cdots + |\alpha_m|^2 = \sigma_m - \sigma_n. \end{aligned}$$

Hence it follows that $\{s_n\}$ is a Cauchy sequence in H if and only if $\{\sigma_n\}$ is a Cauchy sequence in $\mathbb{R}$. Since both H and $\mathbb{R}$ are complete, it follows that the series $\sum_{n=1}^{\infty} \alpha_n e_n$ converges in H if and only if the series $\sum_{n=1}^{\infty} |\alpha_n|^2$ converges in $\mathbb{R}$.

Now, suppose that the series $\sum_{n=1}^{\infty} \alpha_n e_n$ converges to some $x \in H$. Then $s_n \to x$. Given any n for all $m > n$, we have

$$\begin{aligned} (s_m, e_n) &= (\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_m e_m, e_n) \\ &= \alpha_1 (e_1, e_n) + \alpha_2 (e_2, e_n) + \cdots + \alpha_m (e_m, e_n) \\ &= \alpha_n . 1 = \alpha_n. \end{aligned}$$

Since inner product is continuous, so taking $m \to \infty$, we get $(x, e_n) = \alpha_n$ for all n. Moreover, for all n as above, we have

$$\begin{aligned} \|s_n\|^2 &= ||\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n\|^2 \\ &= |\alpha_1|^2 + |\alpha_2|^2 + \cdots + |\alpha_n|^2 . \end{aligned}$$

Now, considering $n \to \infty$, we get

$$||x||^2 = \sum_{n=1}^{\infty} |\alpha_n|^2 .$$

This completes the proof. □

Theorem 20.2.4. Let $\{e_n\}$ be an orthonormal sequence in a Hilbert space H Then for each $x \in H$, the series $\sum_{n=1}^{\infty} (x, e_n) e_n$ is convergent in H and further $\left(x - \sum_{n=1}^{\infty} (x, e_n) e_n \right) \perp e_k$ for $k = 1, 2, \ldots$.

Proof. Since $\{e_n\}$ is an orthonormal sequence in the Hilbert space H, by Bessel's inequality, we have

$$\sum_{n=1}^{\infty} |(x, e_n)|^2 \leq \|x\|^2 < \infty.$$

Thus the series $\sum_{n=1}^{\infty} |(x, e_n)|^2$ is convergent in $\mathbb{R}$ and hence by the previous theorem, the series $\sum_{n=1}^{\infty} (x, e_n) e_n$ is convergent in H. Now, let $x_m = x - \sum_{j=1}^{m} (x, e_j) e_j$. Then $x_m \to x - \sum_{n=1}^{\infty} (x, e_n) e_n$ as $m \to \infty$. Now,

given any positive integer k, for all $m > k$, we have

$$(x_m, e_k) = (x, e_k) - \sum_{j=1}^{m} (x, e_j)\,(e_j, e_k) = (x, e_k) - (x, e_k) = 0.$$

Since inner product is continuous, considering $m \to \infty$, we have

$$\left(x - \sum_{n=1}^{\infty} (x, e_n)\, e_n, e_k \right) = 0.$$

Hence, $\left(x - \sum_{n=1}^{\infty} (x, e_n)\, e_n \right) \perp e_k$ for $k = 1, 2, \ldots$ □

Definition 20.2.5. An orthonormal sequence $\{e_n\}$ in an inner product space X is said to be complete if the set $\{e_1, e_2, e_3, \ldots\}$ is not a proper subset of any orthonormal set in X.

Theorem 20.2.6. Let $\{e_n\}$ be an orthonormal sequence in a Hilbert space H. Then the following conditions are equivalent:

1. The orthonormal sequence $\{e_n\}$ is complete.
2. If $x \in H$ is such that $x \perp \{e_n\}$, then $x = 0$.
3. Each $x \in H$ has a Fourier expansion $x = \sum_{n=1}^{\infty} (x, e_n)\, e_n$.
4. Each $x \in H$ satisfies the Perseval identity $\|x\|^2 = \sum_{n=1}^{\infty} |(x, e_n)|^2$.

*Proof*1 $\Rightarrow$ 2: If $x \in H$ is such that $x \perp \{e_n\}$.

If possible, let $x \neq 0$. Then $e = \frac{1}{\|x\|} x \in H$ and $\|e\| = 1$. Further $(e, e_n) = \left(\frac{1}{\|x\|} x, e_n \right) = \frac{1}{\|x\|} \cdot 0 = 0$ for all n. Also $e \neq e_n$ for any n, for if $e = e_n$, $(e, e_n) = (e_n, e_n) = \|e_n\|^2 = 1 \neq 0$. Therefore, $\{e, e_1, e_2, e_3, \ldots\}$ is an orthonormal set of which $\{e_1, e_2, e_3, \ldots\}$ is a proper subset. This contradicts 1. Hence we must have $x = 0$. Thus $1 \Rightarrow 2$.

$2 \Rightarrow 3$: Since $\{e_n\}$ be an orthonormal sequence in a Hilbert space H, then by the previous theorem, for each $x \in H$, the series $\sum_{n=1}^{\infty} (x, e_n)\, e_n$ is convergent in H and further $\left(x - \sum_{n=1}^{\infty} (x, e_n)\, e_n \right) \perp e_k$ for $k = 1, 2, \ldots$ Then by 2, we have $x - \sum_{n=1}^{\infty} (x, e_n)\, e_n = 0$, i.e., $x = \sum_{n=1}^{\infty} (x, e_n)\, e_n$. Thus $2 \Rightarrow 3$.

$3 \Rightarrow 4$: By 3, for each $x \in H$, we have Fourier expansion $x = \sum_{n=1}^{\infty} (x, e_n)\, e_n$. Hence by a known result,

$$\|x\|^2 = \sum_{n=1}^{\infty} |\,(x, e_n)\,|^2.$$

Thus $3 \Rightarrow 4$.

$4 \Rightarrow 1$: By 4, we have

$$1 = \|e\|^2 = \sum_{n=1}^{\infty} |\,(e, e_n)\,|^2.$$

So, we must have $(e, e_n) \neq 0$ for at least one n. This means that $\{e, e_1, e_2, e_3, \ldots\}$ is not an orthonormal set. So, by definition, $\{e_1, e_2, e_3, \ldots\}$ is a complete orthonormal set, i.e., $\{e_n\}$ is a complete orthonormal sequence. Thus $4 \Rightarrow 1$.

This completes the proof of the theorem. □

Exercise 20.2.7.

1. Let $\{e_1, e_2, e_3, \ldots, e_n\}$ be an orthonormal set in a Hilbert space H where n is fixed. If $x \in H$ is a fixed member, show that for some scalars $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$, the value of $\left\| x - \sum_{n=1}^{\infty} \alpha_n e_n \right\|$ is minimum when $\alpha_i = (x, e_i)\,, i = 1, 2, \ldots, n$.

2. Let $\{e_n\}$ be an orthonormal sequence in an inner product space X. If $x \in H$ is such that $x = \sum_{n=1}^{\infty} (x, e_n)\, e_n$, then show that $(x, y) = \sum_{n=1}^{\infty} (x, e_n)\, \overline{(y, e_n)}$.

3. In the Hilbert space l^2, show that $\{e_n\}$ is a complete orthonormal sequence if

$$\begin{aligned} e_1 &= (1, 0, 0, 0, \ldots, 0) \\ e_2 &= (0, 1, 0, 0, \ldots, 0) \\ e_3 &= (0, 0, 1, 0, \ldots, 0) \\ &\vdots \end{aligned}$$

References

1. P.K. Jain and V.P. Gupta: Lebesgue Measure and Integration.
2. F. Burk: Lebesgue Measure and Integration An Introduction.
3. G. Barra: Measure Theory and Integration.
4. H.L. Royden and P.M. Fitzpatrick: Real Analysis.
5. M.E. Taylor: Measure Theory and Integration.
6. S. Ponnusamy: Foundations of Complex Analysis.
7. S. Ponnusamy H. Silverman: Complex Variables with Applications.
8. A.I. Markushevich: Theory of Functions of a Complex Variable, Volume I, II and III.
9. R.V. Churchill and J.W. Brown: Complex Variables and Applications.
10. E. Kreyszig:Introductory Functional Analysis with Applications.
11. G. Bachman and L. Narici:Functional Analysis.
12. W. Rudin:Functional Analysis.
13. N. Dunford and L. Schwartz :Linear Operators (Part I).
14. A. E. Taylor :Introduction to Functional Analysis.
15. B. V. Limaye:Functional Analysis.
16. K. Yoshida :Functional Analysis.
17. B. K. Lahiri:Elements of Functional Analysis.

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER II

SELF LEARNING MATERIAL

PAPER : COR 2.2
(Pure & Applied Streams)

Classical Mechanics

Abstract Algebra II

Operations Research II

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

Content Writers

Block - I : Classical Mechanics	Dr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani
Block - II : Abstract Algebra II	Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani
Block - III : Operations Research - II	Dr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani

July, 2022

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 Message

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, 2020 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.) Manas Kumar Sanyal, 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 well 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 coordinated 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 written and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Board of Studies Members of Department of Mathematics, Directorate of Open and Distance Learning (DODL), University of Kalyani

Sl No.	Name & Designation	Role
1	Dr. Animesh Biswas, Professor & Head, Dept. of Mathematics, KU	Chairperson
2	Dr. Pulak Sahoo, Professor, Dept. of Mathematics, KU	Member
3	Dr. Sahidul Islam, Assistant Professor, Dept. of Mathematics, KU	Member
4	Dr. Sushanta Kumar Mohanta, Professor, Dept. of Mathematics, West Bengal State University	External Nominated Member
5	Dr. Biswajit Mallick, Assistant Professor (Cont.), Department of Mathematics, DODL, KU	Member
6	Ms. Audrija Choudhury, Assistant Professor (Cont), Department of Mathematics, DODL, KU	Member
7	Director, DODL, KU	Convener

Core Paper

PURE & APPLIED STREAMS

COR 2.2
Marks : 100 (SEE : 80; IA : 20); Credit : 6

Classical Mechanics (Marks : 32 (SEE: 25; IA: 07))
Abstract Algebra II (Marks : 32 (SEE: 25; IA: 07))
Operations Research II (Marks : 36 (SEE: 30; IA: 06))

Syllabus

Block I

- **Unit 1:** Lagrangian Formulation: Generalised coordinates. Holonomic and nonholonomic systems. Scleronomic and rheonomic systems. D'Alembert's principle. Lagrange's equations. Energy equation for conservative fields. Cyclic (ignorable) coordinates. Generalised potential.
- **Unit 2:** Moving Coordinate System: Coordinate systems with relative translational motions, Rotating coordinate systems, The Coriolis force, Motion on the earth, Effect of Coriolis force on a freely falling particle.
- **Unit 3:** Euler's theorem, Euler's equations of motion for a rigid body, Eulerian angles.
- **Unit 4:** Variational Principle : Calculus of variations and its applications in shortest distance, minimum surface of revolution, Brachistochrone problem.
- **Unit 5:** Geodesic. Hamilton's principle. Lagrange's undetermined multipliers. Hamilton's equations of motion.
- **Unit 6:** Canonical Transformations: Canonical coordinates and canonical transformations. Poincare theorem.
- **Unit 7:** Lagrange's and Poisson's brackets and their variance under canonical transformations, Hamilton's equations of motion in Poisson's bracket. Jacobi's identity. Hamilton-Jacobi equation.
- **Unit 8:** Small Oscillations : General case of coupled oscillations. Eigen vectors and Eigen frequencies. Orthogonality of Eigen vectors. Normal coordinates. Two-body problem.

Block II

- **Unit 9:** Preliminaries: Review of earlier related concepts-Rings, integral domains, fields and their simple properties.
- **Unit 10:** Ideals in rings: Definitions, classifications with related theorems, examples and counter examples
- **Unit 11:** Detailed discussion on rings: Classification of rings, their definitions and characterization theorem with examples and counter examples.
- **Unit 12:** Polynomial rings, division algorithm.
- **Unit 13:** Domains in rings: Classification, definitions and related theories with example and counter examples, irreducible polynomials, Eisenstein's criterion for irreducibility.
- **Unit 14:** Field extensions: Definition and simple properties.

Block III

- **Unit 15:** Sensitivity Analysis: Changes in price vector of objective function, changes in resource requirement vector, addition of decision variable, addition of a constraint.
- **Unit 16:** Parametric Programming : Variation in price vector, Variation in requirement vector
- **Unit 17:** Replacement and Maintenance Models: Failure mechanism of items, General replacement policies for gradual failure of items with constant money value and change of money value at a constant rate over the time period, Selection of best item.
- **Unit 18:** Dynamic Programming (DP): Basic features of DP problems, Bellman's principle of optimality, Multistage decision process with Forward and Backward recursive relations, DP approach to stage-coach problems.
- **Unit 19:** Non-Linear Programming (NLP): Lagrange Function and Multipliers, Lagrange Multipliers methods for nonlinear programs with equality and inequality constraints.
- **Unit 20:** Separable programming, Piecewise linear approximation solution approach, Linear fractional programming.

Contents

Unit 1

Course Structure

- Lagrangian Formulation : Generalised coordinates, Holonomic and nonholonomic systems, Scleronomic and rheonomic systems, D'Alembert's principle, Lagrange's equations, Generalised potential.

1.1 Introduction

Lagrangian mechanics is a reformulation of classical mechanics, introduced by the Italian-French mathematician and astronomer *Joseph-Louis Lagrange*. In Lagrangian mechanics, the trajectory of a system of particles is derived by solving Lagrange equation which treat constraints explicitly as extra equations or directly by judicious choice of generalised coordinates. In this case, a mathematical function called the *Lagrangian* is a function of the generalised coordinates, their first derivatives, and time, and contains the information about the dynamics of the system. Although Lagrangian formulation reduce to Newton's laws, they are characterized not only by the relative ease with which many problems can be formulated and solved but by their relationship in both theory and application to such advanced fields as quantum mechanics, statistical mechanics and electrodynamics.

1.2 Basic Concepts

In this section, some basic concepts regarding the motion of particles are given below.

1.2.1 Coordinate Systems

The fundamental concept involved in the motion of a particle (or system) is the position coordinate and how it is changing with time. The position of a particle is represented by choosing a coordinate system. In the Cartesian or rectangular coordinate system, the position vector $\vec{r}$ of a particle is defined in terms $x,\ y$ and z coordinates. In a two dimensional motion, rectangular coordinates (x, y) or polar coordinates (r, θ) can represent the position of the particle [Fig. 1.2.1]. They are related as

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta; \quad r = \sqrt{x^2 + y^2} \quad \text{and} \quad \theta = \tan^{-1}\frac{y}{x}.$$

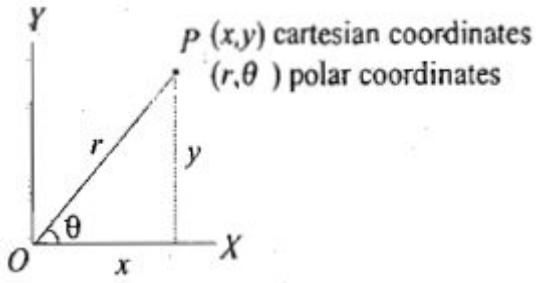


Figure 1.2.1: Rectangular and polar coordinates

In three dimensions, the cylindrical coordinates (ρ, θ, z) and the spherical coordinates (r, θ, ϕ) of the position of a particle are related to the Cartesian coordinates (x, y, z) as follows:

For cylindrical and Cartesian coordinates [Fig. 1.2.2]:

$$x = \rho\cos\theta, \quad y = \rho\sin\theta, \quad z = z; \quad \rho = \sqrt{x^2 + y^2}, \quad A = \tan^{-1}\frac{y}{x} = \sin^{-1}\frac{y}{\rho}$$

For spherical and cartesian coordinates [Fig. 1.2.2]:

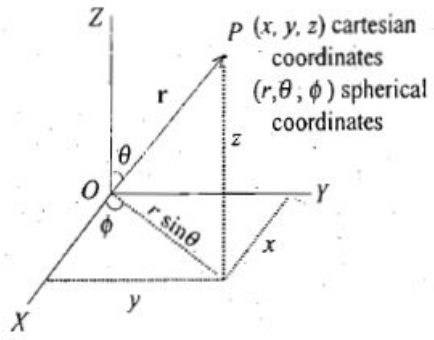


Figure 1.2.2: Cartesian and spherical coordinates

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta;$$

$$r = \left(x^2 + y^2 + z^2\right)^{1/2}, \theta = \tan^{-1}\frac{\sqrt{x^2 + y^2}}{z}, \phi = \tan^{-1}\frac{y}{x}$$

We may represent, for example, the relationships for spherical and cartesian coordinates as follows :

$$x = x(r, \theta, \phi), y = y(r, \theta, \phi), z = z(r, \theta)$$
$$\text{or, } \vec{r} = \vec{r}(r, \theta, \phi)$$

If we include the time variable also, then

$$\vec{r} = \vec{r}(r, \theta, \phi, t),$$

In general , we may represent the coordinates by q_1, q_2, q_3, having the relationships with the cartesian coordinates as

$$x = x\left(q_1, q_2, q_3, t\right), \quad y = y\left(q_1, q_2, q_3, t\right), \quad z = z\left(q_1, q_2, q_3, t\right)$$
$$\text{or, } \vec{r} = \vec{r}\left(q_1, q_2, q_3, t\right)$$

In fact, these are the transformation equations from a general system to the cartesian coordinate system.

1.2.2 Degrees of Freedom - Configuration Space

The minimum number of independent variables or coordinates required to specify the position of a dynamical system, consisting of one or more particles, is called the number of degrees of freedom of the system. For example, the motion of a particle, moving freely in space, can be described by a set of three coordinates e.g.s (x, y, z) and hence the number of degrees of freedom, possessed by the particle, is three. A system of two particles, moving freely in space, requires two sets of three coordinates [e.g., (x_1, y_1, z_1) and (x_2, y_2, z_2)] i.e., six coordinates to specify its position. Thus the system has six degrees of freedom. If a system consists of N particles, moving freely in space, we need $3N$ independent coordinates to descrie its position. Hence the number of degrees of freedom of the system is $3N$.

The configuration of the system of N particles, moving freely in space, may be represented by the position of a single point in $3N$ dimensional space, which is called configuration space of the system. The configuration space for a system of one freely moving particle is 3 -dimensional and for a system of two freely moving particles, it is six dimensional. In the later case, the configuration of the system of the two particles can be represented by the position of a single point with six coordinates in six dimensional space. This system has six degrees of freedom and its configuration space is six dimensional.

The number of coordinates, needed to specify a dynamical system, becomes smaller, when the constraints (which we describe below) are present in the system. Hence *the degrees of freedom of a dynamical system is defined as the minimum number of independent coordinates (or variables) required to specify the system compatible with the constraints*. If there are n independent variables, say $q_1, q_2, \ldots, q_n$ and n constants $C_1 \cdot C_2, \ldots, C_n$ such that

$$\sum_{i=1}^{n} C_i dq_i = 0 \tag{1.2.1}$$

at any position of the system, then we must have

$$C_1 = C_2 = \cdots = C_n = 0.$$

1.2.3 Constraints

Often the motion of a particle or system of particles is restricted by one or more conditions. *The limitations on the motion of a system are called constraints and the motion is said to be constrained motion.*

Holonomic constraints

Suppose the constraints are present in the system of N particles. If the constraints are expressed in the form of equations of the form

$$f\left(\vec{r}_1, \vec{r}_2, \ldots, t\right) = 0 \tag{1.2.2}$$

then they are called *holonomic constraints*. Let there be m number of such equations to describe the constraints in the N particle system. Now, we may use these equations to eliminate m of the $3N$ coordinates and we need only n independent coordinates to describe the motion, given by

$$n = 3N - m$$

The system is said to have n or $3N - m$ degrees of freedom. The elimination of the dependent coordinates can be expressed by introducing $n = 3N - m$ independent variables $q_1, q_2, \ldots, q_n$. These are referred as *generalized coordinates.*

Nonholonomic constraints

The constraints which are not expressible in the form of Eq. (1.2.2) are called *nonholonomic*. For example, the motion of a particle, placed on the surface of a sphere of radius a, will be described by

$$|\vec{r}| \geq a \quad \text{or} \quad r - a \geq 0$$

in a gravitational field, where $\vec{r}$ is the position vector of the particle relative to the centre of the sphere. The particle will first slide down the surface and then fall off. The gas molecules in a container are constrained to move inside it and the related constraint is another example of nonholonomic constraints. If the gas container is in spherical shape with radius a and $\vec{r}$ is the position vector of a molecule, then the condition of constraint for the motion of molecules can be expressed as

$$|\vec{r}| \leq a \quad \text{or} \quad r - a \leq 0.$$

It is to be mentioned that in holonomic constraints, each coordinate can vary independently of the other. In a nonholonomic system, all the coordinates cannot vary independently and hence the number of degrees of freedom of the system is less than the minimum number of coordinates need to specify the configuration of the system. We shall in general consider the holonomic systems.

Constraints are further described as (i) *rheonomous* and (ii) *scleronomous*. In the former, the equations of constraint contain the time as an explicit variable, while in the later they are not explicitly dependent on time. Constraints may also be classified as (i) *conservative* and (ii) *dissipative*. In case of conservative constraints, total mechanical energy of the system is conserved during the constrained motion and the constraint forces do not do any work. In dissipative constraints, the constraint forces do work and the total mechanical energy is not conserved. Time-dependent or rheonomic constraints are generally dissipative.

1.3 Generalized Coordinates

The name generalized coordiniates is given to a set of independent coordinates sufficient in number to describe completely the state of configuration of a dynamical system. These coordinates are denoted as

$$q_1, q_2, q_3, \ldots, q_k, \ldots q_n \tag{1.3.1}$$

where n is the total number of generalized coordinates. In fact, these are the minimum number of coordinates, needed to describe the motion of the system. For example, for a particle constrained to move on the circumference of a circle only one generalized coordinate $q_1 = \theta$ is sufficient and two generalized coordinates $q_1 = \theta$, and $q_2 = \phi$ for a particle moving on the surface of a sphere. The generalized coordinates for a system of N particles, constrained by m equations, are $n = 3N - m$. It is not necessary that these coordinates should be rectangular, spherical or cylindrical. In fact, the quantities like length, (length) ^{2}, angle, energy or a dimensionless quantity may be used as generalized coordinates but they should completely describe the state of the system. Further these n generalized coordinates are not restricted by any constraint.

For a system of N particles, if (x_i, y_i, z_i) are the cartesian coordinates of the i-th particle, then these coordinates in terms of the generalized coordinates q_k can be expressed as

$$\begin{aligned} x_i &= x_i\left(q_1, q_2, \ldots, q_k, \ldots, q_n, t\right) \\ y_i &= y_i\left(q_1, q_2, \ldots, q_k, \ldots, q_n, t\right) \\ z_i &= z_i\left(q_1, q_2, \ldots, q_k, \ldots, q_n, t\right) \end{aligned} \tag{1.3.2}$$

or, in general the position vector $\vec{r}_i\,(x_i, y_i, z_i)$ of the i-th particle is

$$\vec{r}_i = \vec{r}_i\,(q_1, q_2, \ldots, q_k, \ldots, q_n, t) \tag{1.3.3}$$

Eqs. (1.3.2) or (1.3.3) give the transformation equations. It may be mentioned that the generalized coordinates may be the cartesian coordinates.

One should note that the system is said to be rheonomic, when there is an explicit time dependence in some or all of the functions defined by Eq. (1.3.2) or (1.3.3). If there is not the explicit time dependence, the system is called scleronomic and t is not written in the functional dependence, i.e.,

$$\vec{r}_i = \vec{r}_i\,(q_1, q_2, \ldots, q_k, \ldots, q_n) \tag{1.3.4}$$

1.4 Principle of virtual work

In order to investigate the properties of a system, we can imagine arbitrary instantaneous change in the position vectors of the particles of the system e.g., virtual displacements. An infinitesimal virtual displacement of i-th particle of a system of N particles is denoted by δr_i. This is the displacement of position coordinates only and does not involve variation of time, i.e.,

$$\delta\vec{r}_i = \delta\vec{r}_i\,(q_1, q_2, \ldots, q_n)$$

Suppose the system is in equilibrium, then the total force on any particle is zero i.e.,

$$\vec{F}_i = 0, \quad i = 1, 2, \ldots, N$$

The virtual work of the force $\vec{F}_i$ in the virtual displacement $\delta\vec{r}_i$ will also be zero, i.e.,

$$\delta W_i = \vec{F}_i \cdot \delta\vec{r}_i = 0$$

Similarly, the sum of virtual work for all the particles must vanish, i.e.,

$$\delta W = \sum_{i=1}^{N} \vec{F}_i \cdot \delta\vec{r}_i = 0 \tag{1.4.1}$$

This result represents the principle of virtual work which states that the work done is zero in the case of an arbitrary virtual displacement of a system from a position of equilibrium .

The total force $\vec{F}_i$ on the i-th particle can be expressed as

$$\vec{F}_i = \vec{F}_i^a + \vec{f}_i.$$

where $\vec{F}_i^a$ is the applied force and $\vec{f}_i$ is the force of constraint. Hence, Eq. (1.4.1) assumes the form

$$\sum_{i=1}^{N} \vec{F}_i^a \cdot \delta\vec{r}_i + \sum_{i=1}^{N} \vec{f}_i \cdot \delta\vec{r}_i = 0$$

We restrict ourselves to the systems where the virtual work of the forces of constraints is zero, e.g., in case of a rigid body. Then

$$\sum_{i=1}^{N} \vec{F}_i^a \cdot \delta\vec{r}_i = 0$$

i.e., for equilibrium of a system, the virtual work of applied forces is zero. We see that the principle of virtual work deals with the statics of a system of particles. However, we want a principle to deal with the general motion of the system and such a principle was developed by D'Alembert.

1.5 D'Alembert's Principle

According to Newton's second law of motion, the force acting on the i-th particle is given by

$$\vec{F}_i = \frac{d\vec{p}_i}{dt} = \dot{\vec{p}}_i$$

This can be written as

$$\vec{F}_i - \dot{\vec{p}}_i = 0, \quad i = 1, 2, \ldots, N$$

These equations mean that any particle in the system is in equilibrium under a force, which is equal to the actual force $\vec{F}_i$ plus a reversed effective force $\dot{\vec{p}}_i$. Therefore, for virtual displacements $\delta\vec{r}_i$,

$$\sum_{i=1}^{N} \left(\vec{F}_i - \dot{\vec{p}}_i \right) \cdot \delta\vec{r}_i = 0$$

But $\vec{F}_i = \vec{F}_i^a + \vec{f}_i$, then

$$\sum_{i=1}^{N} \left(\vec{F}_i^a - \dot{\vec{p}}_i \right) \cdot \delta\vec{r}_i + \sum_{i=1}^{N} \vec{f}_i \cdot \delta\vec{r}_i = 0$$

Again, we restrict ourselves to the systems for which the virtual work of the constraints is zero, i.e., $\sum_i \vec{f}_i \cdot \delta\vec{r}_i = 0$. Then

$$\sum_{i=1}^{N} \left(\vec{F}_i^a - \dot{\vec{p}}_i \right) \cdot \delta\vec{r}_i = 0 \tag{1.5.1}$$

This is known as *D'Alembert's principle*. Since the forces of constraints do not appear in the equation and hence now we can drop the superscript. Therefore, the D'Alembert's principle may be written as

$$\sum_{i=1}^{N} \left(\vec{F}_i - \dot{\vec{p}}_i \right) \cdot \delta\vec{r}_i = 0 \tag{1.5.2}$$

1.6 Lagrange's equations from D'Alembert's Principle

Consider a system of N particles. The transformation equations for the position vectors of the particles

$$\vec{r}_i = \vec{r}_i \left(q_1, q_2, \ldots, q_k, \ldots, q_n, t \right) \tag{1.6.1}$$

where t is the time and $q_k (k = 1, 2, \ldots, n)$ are the generalized coordinates. Differentiating Eq. (1.6.1) with respect to t, we obtain the velocity of the i-th particle, i.e.,

$$\frac{d\vec{r}_i}{dt} = \frac{\partial\vec{r}_i}{\partial q_1}\frac{dq_1}{dt} + \frac{\partial\vec{r}_i}{\partial q_2}\frac{dq_2}{dt} + \cdots + \frac{\partial\vec{r}_i}{\partial q_k}\frac{dq_k}{dt} + \cdots + \frac{\partial\vec{r}_i}{\partial q_n}\frac{dq_n}{dt} + \frac{\partial\vec{r}_i}{\partial t}$$

$$\text{or,} \quad \vec{v}_i = \dot{\vec{r}}_i = \sum_{k=1}^{n} \frac{\partial\vec{r}_i}{\partial q_k}\dot{q}_k + \frac{\partial\vec{r}_i}{\partial t} \tag{1.6.2}$$

where $\dot{q}_k$ are the *generalized velocities*. The virtual displacement is given by

$$\delta\vec{r}_i = \frac{\partial\vec{r}_i}{\partial q_1}\delta q_1 + \frac{\partial\vec{r}_i}{\partial q_2}\delta q_2 + \cdots + \frac{\partial\vec{r}_i}{\partial q_k}\delta q_k + \cdots + \frac{\partial\vec{r}_i}{\partial q_n}\delta q_n$$

$$\text{or,} \quad \delta\vec{r}_i = \sum_{k=1}^{n} \frac{\partial\vec{r}_i}{\partial q_k}\delta q_k \tag{1.6.3}$$

since by definition the virtual displacements do not depend on time. According to D'Alembert's principle,

$$\sum_{i=1}^{N}\left(\vec{F}_i-\dot{\vec{p}}_i\right)\cdot\delta\vec{r}_i=0 \tag{1.6.4}$$

Here

$$\sum_{i=1}^{N}\vec{F}_i\cdot\delta\vec{r}_i=\sum_{i=1}^{N}\vec{F}_i\cdot\sum_{k=1}^{n}\frac{\partial\vec{r}_i}{\partial q_k}\delta q_k=\sum_{k=1}^{n}\sum_{i=1}^{N}\left[\vec{F}_i\cdot\frac{\partial\vec{r}_i}{\partial q_k}\right]\delta q_k=\sum_{k=1}^{n}G_k\delta q_k \tag{1.6.5}$$

where

$$G_k=\sum_{i=1}^{N}\vec{F}_i\cdot\frac{\partial\vec{r}_i}{\partial q_k}=\sum_{i=1}^{N}\left[F_{x_i}\frac{\partial x_i}{\partial q_k}+F_{y_i}\frac{\partial y_i}{\partial q_k}+F_{z_i}\frac{\partial z_i}{\partial q_k}\right] \tag{1.6.6}$$

are called the components of *generalized force* associated with the generalized coordinates q_k. This may be mentioned that as the dimensions of the generalized coordinates need not be those of length, similarly the generalized force components G_k may have dimensions different than those of force. However, the dimensions of $G_k\delta q_k$ are those of work. For example, if δq_k has the dimensions of length, G_k will have the dimensions of force; if δq_k has the dimensions of angle (θ), G_k will have the dimensions of torque (τ).

Further

$$\sum_{i=1}^{N}\dot{\vec{p}}_i\cdot\delta\vec{r}_i=\sum_{i=1}^{N}m_i\ddot{\vec{r}}_i\cdot\sum_{k=1}^{n}\frac{\partial\vec{r}_i}{\partial q_k}\delta q_k=\sum_{k=1}^{n}\left[\sum_{i=1}^{N}m_i\ddot{\vec{r}}_i\cdot\frac{\partial\vec{r}_i}{\partial q_k}\right]\delta q_k \tag{1.6.7}$$

Now

$$\sum_{i=1}^{N}m_i\ddot{\vec{r}}_i\cdot\frac{\partial\vec{r}_i}{\partial q_k}=\sum_{i=1}^{N}\left[\frac{d}{dt}\left(m_i\dot{\vec{r}}_i\cdot\frac{\partial\vec{r}_i}{\partial q_k}\right)-m_i\dot{\vec{r}}_i\cdot\frac{d}{dt}\left(\frac{\partial\vec{r}_i}{\partial q_k}\right)\right] \tag{1.6.8}$$

It is easy to prove that

$$\frac{d}{dt}\left(\frac{\partial\vec{r}_i}{\partial q_k}\right)=\frac{\partial}{\partial q_k}\left(\frac{d\vec{r}_i}{dt}\right)=\frac{\partial\vec{v}_i}{\partial q_k} \tag{1.6.9}$$

and

$$\frac{\partial\vec{r}_i}{\partial q_k}=\frac{\partial\vec{v}_i}{\partial\dot{q}_k} \tag{1.6.10}$$

Therefore, Eq. (1.6.8) becomes

$$\sum_{i=1}^{N}m_i\ddot{\vec{r}}_i\cdot\frac{\partial\vec{r}_i}{\partial q_k}=\sum_{i=1}^{N}\left[\frac{d}{dt}\left[m_i\vec{v}_i\cdot\frac{\partial\vec{v}_i}{\partial\dot{q}_k}\right]-m_i\vec{v}_i\cdot\frac{\partial\vec{v}_i}{\partial q_k}\right] \tag{1.6.11}$$

Substituting in (1.6.7), we get

$$\begin{aligned}\sum_{i=1}^{N}\dot{\vec{p}}_i\cdot\delta\vec{r}_i &= \sum_{k=1}^{n}\sum_{i=1}^{N}\left[\frac{d}{dt}\left(m_i\vec{v}_i\cdot\frac{\partial\vec{v}_i}{\partial\dot{q}_k}\right)-m_i\vec{v}_i\cdot\frac{\partial\vec{v}_i}{\partial q_k}\right]\delta q_k\\ &= \sum_{k=1}^{n}\left[\frac{d}{dt}\left\{\frac{\partial}{\partial\dot{q}_k}\left(\sum_{i=1}^{N}\frac{1}{2}m_i\left(\vec{v}_i\cdot\vec{v}_i\right)\right)\right\}-\frac{\partial}{\partial q_k}\left\{\sum_{i=1}^{N}\frac{1}{2}m_i\left(\vec{v}_i\cdot\vec{v}_i\right)\right\}\right]\delta q_k\\ &= \sum_{k=1}^{n}\left[\frac{d}{dt}\left(\frac{\partial T}{\partial\dot{q}_k}\right)-\frac{\partial T}{\partial q_k}\right]\delta q_k\end{aligned} \tag{1.6.12}$$

where $\sum_i \frac{1}{2} m_i (\vec{v}_i \cdot \vec{v}_i) = \sum_i \frac{1}{2} m_i v_i^2 = T$ is the *kinetic energy* of the system.

Substituting for $\sum_i \vec{F}_i \cdot \delta\vec{r}_i$ from (1.6.5) and $\sum_i \dot{\vec{p}}_i \cdot \delta\vec{r}_i$ from (1.6.12) in Eq. (1.6.4), the D'Alembert's principle becomes

$$\sum_{k=1}^{n} \left[\left\{ \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} \right\} - G_k \right] \delta q_k = 0 \tag{1.6.13}$$

As the constraints are holonomic, it means that any virtual displacement δq_k is independent of δq_j. Therefore, the coefficient in the square bracket for each δq_k must be zero, i.e.,

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} - G_k = 0 \quad \text{or} \quad \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} = G_k \tag{1.6.14}$$

This represents the *general form of Lagrange's equations*.

For a *conservative system*, the force is derivable from a scalar potential V :

$$\vec{F}_i = -\nabla_i V = -\hat{i}\frac{\partial V}{\partial x_i} - \hat{j}\frac{\partial V}{\partial y_i} - \hat{k}\frac{\partial V}{\partial z_i} \tag{1.6.15}$$

Hence from Eq. (1.6.6), the generalized force components are

$$G_k = -\sum_{i=1}^{N} \left[\frac{\partial V}{\partial x_i}\frac{\partial x_i}{\partial q_k} + \frac{\partial V}{\partial y_i}\frac{\partial y_i}{\partial q_k} + \frac{\partial V}{\partial z_i}\frac{\partial z_i}{\partial q_k} \right] \tag{1.6.16}$$

Clearly the right hand side of the above equation is the partial derivative of $-V$ with respect to q_k, i.e.,

$$G_k = -\frac{\partial V}{\partial q_k} \tag{1.6.17}$$

Thus Eq. (1.6.14) assumes the form

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} = -\frac{\partial V}{\partial q_k} \tag{1.6.18}$$

$$\text{or,} \quad \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial (T-V)}{\partial q_k} = 0 \tag{1.6.19}$$

Since the scalar potential V is the function of generalized coordinates q_k only not depending on generalized velocities, we can write Eq. (1.6.19) as

$$\frac{d}{dt}\left[\frac{\partial (T-V)}{\partial \dot{q}_k}\right] - \frac{\partial (T-V)}{\partial q_k} = 0 \tag{1.6.20}$$

We define a new function given by $L = T - V$ which is called the *Lagrangian* of the system. Thus, Eq. (1.6.20) takes the form

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0 \tag{1.6.21}$$

for $k = 1, 2, \ldots, n$.

These equations are inown as *Lagrange's equations* for conservative system. They are n in number and there is one equation for each generalized coordinate. In order to solve these equations, we must know the Lagrangian function $L = T - V$ in the appropriate generalized coordinates.

1.7 Procedure for formation of Lagrange's equations

The Lagrangian function L of a system is given by

$$L = T - V \tag{1.7.1}$$

In order to form L, kinetic energy T and potential energy V are to be written in generalized coordinates. This is then substituted in the Lagrangian equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0 \tag{1.7.2}$$

to obtain the equations of motion of the system. This involves first to find the partial derivatives of L, i.e:, $\partial L/\partial q_k$ and $\partial L/\partial \dot{q}_k$ and then to put their values in Eq. (1.7.2).

Kinetic Energy in Generalized Coordinates: The transformation equations (1.6.1) and (1.6.2) are used to transform T from cartesian coordinates to generalized coordinates. Therefore

$$T = \sum_i \frac{1}{2} m_i v_i^2 = \sum_i \frac{1}{2} m_i \dot{\vec{r}}_i^2 = \sum_i \frac{1}{2}\left(\sum_{k=1}^{n} \frac{\partial \vec{r}_i}{\partial q_k}\dot{q}_k + \frac{\partial \vec{r}_i}{\partial t}\right)^2$$

$$\text{or, } T = M_0 + \sum_k M_k \dot{q}_k + \frac{1}{2}\sum_{k,l} M_{kl}\dot{q}_k\dot{q}_l \tag{1.7.3}$$

where $M_0 = \sum_k \frac{1}{2} m_i \left(\frac{\partial \vec{r}_i}{\partial t}\right)^2$, $M_k = \sum_i m_i \frac{\partial \vec{r}_i}{\partial t} \cdot \frac{\partial \vec{r}_i}{\partial q_k}$ and $M_{kl} = \sum_i m_i \frac{\partial \vec{r}_i}{\partial q_k} \cdot \frac{\partial \vec{r}_i}{\partial q_l}$

Thus we see from (1.7.3) that in the expression for kinetic energy, first term is independent of generalized velocities, while second and third terms are linear and quadratic in generalized velocities respectively.

For scleronomic systems, the transformation equations do not contain time explicitly, so that

$$\vec{v}_i = \dot{\vec{r}}_i = \sum_k \frac{\partial \vec{r}_i}{\partial q_k}\dot{q}_k$$

Therefore,

$$T = \sum_i \frac{1}{2} m v_i^2 = \frac{1}{2}\sum_{k,l} M_{kl}\dot{q}_k\dot{q}_l \tag{1.7.4}$$

In such a case, the expression for T is a homogeneous quadratic form in generalized velocities.

Example 1.7.1. Obtain the equation of motion of a simple pendulum by using Lagrangian method and hence deduce the formula for its time period for small amplitude oscillations.

Solution: Let θ be the angular displacement of the simple pendulum from the equilibrium position. If l be the effective length of the pendulum and m be the mass of the bob, then the displacement along arc $OA = s$ is given by

$$s = l\theta \quad \left[\text{because } \theta = \frac{\text{Arc}}{\text{Radius}} = \frac{s}{l}\right]$$

Kinetic energy $T = \frac{1}{2}mv^2 = \frac{1}{2}ml^2\dot{\theta}^2 \quad \left[\because v = \frac{ds}{dt} = \frac{d(l\theta)}{dt} = \frac{l\, d\theta}{dt} = \dot{\theta}\right]$

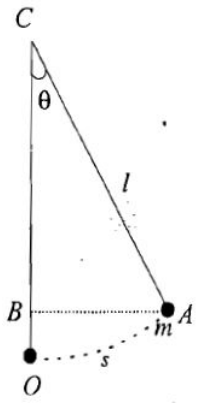


Figure 1.7.1: Simple pendulum

If the potential energy of the system, when the bob is at O, is zero, then the potential energy, when the bob is at A, is given by

$$V = mg\ (OB) = mg\ \ (OC - BC) = mg(l - l\cos\theta) = mgl(1 - \cos\theta)$$

Hence

$$L = T - V, \quad \text{or} \quad L = \frac{1}{2}ml^2\dot{\theta}^2 - mgl(1 - \cos\theta)$$

Now,

$$\frac{\partial L}{\partial \theta} = -mgl\sin\theta \quad \text{and} \quad \frac{\partial L}{\partial \dot{\theta}} = ml^2\dot{\theta}$$

Substituting these values in the Lagrange's equation (here there is only one generalized coordinate $q_1 = \theta$)

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}}\right) - \frac{\partial L}{\partial \theta} = 0$$

we get

$$\frac{d}{dt}\left[ml^2\dot{\theta}\right] + mgl\sin\theta = 0 \Rightarrow ml^2\ddot{\theta} + mgl\sin\theta = 0 \Rightarrow \ddot{\theta} + \frac{g}{l}\sin\theta = 0$$

This represents the equation of motion of a simple pendulum. For small amplitude oscillations, $\sin\theta \cong \theta$, and therefore the equation of motion of a simple pendulum is

$$\ddot{\theta} + \frac{g}{l}\theta = 0$$

This represents a *simple harmonic motion* of period, given by

$$T = 2\pi\sqrt{\frac{l}{g}}$$

Example 1.7.2. Use Lagrange's equations to find the equation of motion of a compound pendulum in a vertical plane about a fixed horizontal axis. Hence find the period of small amplitude oscillations of the compound pendulum.

Solution. Let the compound pendulum be suspended from S with C as centre of mass. It is oscillating in the vertical plane which is the plane of the paper.

Moment of inertia of the pendulum about the axis of rotation through S is given by

$$I = I_c + Ml^2 = M(K^2 + l^2)$$

where M is the mass of the pendulum, $I_c = MK^2$ (K = radius of gyration) about a parallel axis through C and l the distance between centre of suspension and centre of mass. If θ is the instantaneous angle which SC

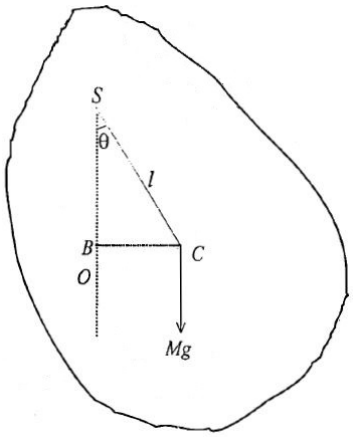


Figure 1.7.2: Compound pendulum

makes with the vertical axis through O, then the kinetic energy of the oscillating system is

$$T = \frac{1}{2}I\dot{\theta}^2 = \frac{1}{2}M\left(K^2 + l^2\right)\dot{\theta}^2$$

Potential energy with respect to horizontal plane through S is $V = -Mgl\cos\theta$ and Lagrangian $L = T - V = \frac{1}{2}M\left(K^2 + l^2\right)\dot{\theta}^2 + Mgl\cos\theta$.

Now, $\dfrac{\partial L}{\partial \theta} = -Mgl\sin\theta$ and $\frac{\partial L}{\partial \dot{\theta}} = M\left(K^2 + l^2\right)\dot{\theta}$

Lagrange's equation in θ coordinate is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}}\right) - \frac{\partial L}{\partial \theta} = 0$$

Therefore,

$$\frac{1}{2}M\left(K^2 + l^2\right)\ddot{\theta} + Mgl\sin\theta = 0 \Rightarrow \ddot{\theta} + \frac{gl}{K^2 + l^2}\sin\theta = 0$$

This is the equation of motion of the compound pendulum. If θ is small, $\sin\theta \cong \theta$

$$\ddot{\theta} + \frac{gl}{K^2 + l^2}\theta = 0$$

This equation represents a simple harmonic motion whose period is given by

$$T = 2\pi\sqrt{\frac{K^2 + l^2}{lg}} = 2\pi\sqrt{\frac{K^2}{l} + l}.$$

1.8 Lagrange's equations in presence of non-conservative forces

When the forces acting on the system consist of non-conservative forces $(\vec{f}_i)$ in addition to the conservative forces $(\vec{F}_i)$, then the components of generalized force can be written as [using Eq. (1.6.6)]

$$G_k = \sum_{i=1}^{N} \left[\vec{F}_i + \vec{f}_i\right] \cdot \frac{\partial \vec{r}_i}{\partial q_k} = \sum_i \vec{F}_i \cdot \frac{\partial \vec{r}_i}{\partial q_k} + \sum_i \vec{f}_i \cdot \frac{\partial \vec{r}_i}{\partial q_k} \Rightarrow G_k = -\frac{\partial V}{\partial q_k} + G'_k \tag{1.8.1}$$

where $G'_k = \sum \vec{f}_i \cdot \frac{\partial \vec{r}_i}{\partial q_k}$ are the components of generalized non-potential force resulting from non-conservative forces and $\sum \vec{f}_i \cdot \frac{\partial \vec{r}_i}{\partial q_k} = -\frac{\partial V}{\partial q_k}$ for conservative part [Eq. (1.6.17)].

Here V is the scalar potential for conservative forces. In such a case, Eq. (1.6.21) assumes the form

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \left(\frac{\partial L}{\partial q_k}\right) = G'_k \tag{1.8.2}$$

where $L = T - V$. Eqs. (1.8.2) represent the Lagrange's equations in the presence of non-conservative forces.

1.9 Generalized Potential

In general, the Lagrange's equations can be written as

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_k}\right) - \frac{\partial T}{\partial q_k} = G_k \tag{1.9.1}$$

For a conservative system, $G_k = -\frac{\partial V}{\partial q_k}$ and then the Lagrange's equations in the usual form are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0 \quad \text{with} \quad L = T - V \tag{1.9.2}$$

However, Lagrange's equations can be put in the form (1.9.2), provided the generalized forces are obtained from a function $U(q_k, \dot{q}_k)$, given by

$$G_k = -\frac{\partial U}{\partial q_k} + \frac{d}{dt}\left(\frac{\partial U}{\partial \dot{q}_k}\right) \tag{1.9.3}$$

In such a case,

$$L = T - U \tag{1.9.4}$$

where $U(q_k, \dot{q}_k)$ is called *velocity dependent potential or generalized potential*. This type of case occurs in case of a charge moving in an electromagnetic field.

Exercise 1.9.1.

1. What are generalized coordinates? Describe the advantage of their use in the solutions of mechanical problems.
2. What are constraints? Classify the constraints with some examples.
3. What is D'Alembert's principle? Derive Lagrange's equations of motion from it for conservative system. How will the result be modified for non-conservative system?

4. Use D'Alembert's principle to determine the equation of motion of a simple pendulum.

5. Prove that a system describe the motion of harmonic oscillator whose Lagrangian is given by

$$L = \frac{1}{2}\dot{q}^2 + q\dot{q} - \frac{1}{2}q^2.$$

6. The Lagrangian of a system is given by $L = \frac{1}{2}ml^2(\dot{\theta} + \sin\theta\dot{\phi}^2) - mgl\cos\theta$, where m, l and g are constants. Prove that the quantity $\dot{\phi}\sin^2\theta$ is conserved.

7. The Lagrangian of a particle of mass m moving in one dimensional is given by $L = \frac{1}{2}m\dot{x}^2 - bx$, where b is a positive constant. Show that the coordinate of the particle $x(t)$ at time t is given by $-\frac{b}{2m}t^2 + c_1t + c_2$, where c_1 and c_2 are constants.

8. Prove that the dynamics of a particle is governed by the Lagrangian $L = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2 - kx\dot{x}t$ describes a free particle.

9. Consider a particle of mass m attached to two identical springs each of length l and spring constant k. The equilibrium configuration is the one where the springs are unstretched. There are no other external forces on the system. If the particle is given a small displacement along the X-axis, then show that the equation of motion for small oscillations is governed by the equation

$$m\ddot{x} + \frac{kx^3}{l^2} = 0.$$

10. The parabolic coordinate (ξ, η) are related to the Cartesian coordinates (x, y) by $x = \xi\eta$ and $y = \frac{1}{2}(\xi^2 - \eta^2)$. Show that the Lagrangian of a two-dimensional simple harmonic oscillator of mass m and angular frequency ω is

$$\frac{1}{2}m(\xi^2 + \eta^2)\left[\left(\dot{\xi}^2 + \dot{\eta}^2\right) - \frac{1}{4}\omega^2\left(\xi^2 - \eta^2\right)\right].$$

Unit 2

Course Structure

- Moving Coordinate System : Coordinate systems with relative translational motions.
- Rotating coordinate systems, The Coriolis force.
- Motion on the earth. Effect of Coriolis force on a freely falling particle.

2.1 Introduction

We have seen earlier that Newton's laws of motion are valid in inertial frame of reference and these inertial frames are unaccelerated. The accelerated frames are called as non-inertial frames because in such a frame, a force-free particle will seem to have an acceleration. If we do not consider the acceleration of the frame but apply Newton's laws to the motion of the force free-particle, then it will appear that a force is acting on it. This means that in the accelerated frames, Newton's law of inertia is not valid. Thus a non-inertial frame of reference is defined as a frame of reference in which Newton's first law does not hold true. An observer of a rotating frame will also see a force on a force-free particle and hence all rotating frames are also non-inertial.

2.2 Fictitious or Pseudo Force

Suppose that S is an inertial frame and another frame S' is moving with an acceleration a relative to S. The acceleration of a particle P, on which no external force is acting, will be zero in the frame S; but in frame S' the observer will find that an acceleration $-a_0$ is acting on it. Thus, in frame S' the observed force on the particle is $-m\vec{a}_0$, where m is the mass of the particle. Such a force, which does not really act on the particle but appears due to the acceleration of the frame, is called a *fictitious* or *pseudo force*. Hence fictitious force on the particle P is

$$\vec{F}_0 = -m\vec{a}_0 \tag{2.2.1}$$

Here, the accelerated frame S' is non-inertial. Now, if a force $\vec{F}_i$ is applied on the particle and $\vec{a}_i$ is the observed acceleration in the inertial frame (S), then according to Newton's law

$$\vec{F}_i = m\,\vec{a}_i \tag{2.2.2}$$

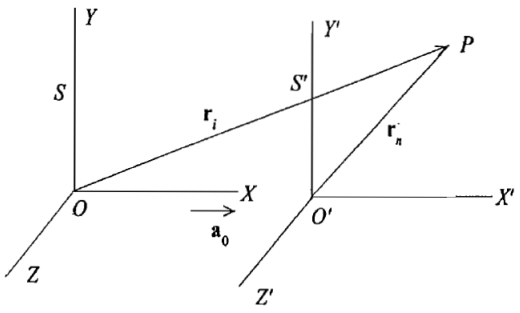


Figure 2.2.1: Non-inertial (accelerated) frames

Suppose frame S' coincides at $t = 0$ with the initial frame S. Then at any time t, the position vectors of a particle $\vec{r}_i$ and $\vec{r}_n$ in the inertial and non-inertial frames respectively are connected as

$$\vec{r}_i = \vec{r}_n + \frac{1}{2}\vec{a}_0 t^2$$

where a_0 is the acceleration of the frame S' with respect to S. Double differentiation with respect to time t gives

$$\frac{d^2\vec{r}_i}{dt^2} = \frac{d^2\vec{r}_n}{dt^2} + \vec{a}_0 \tag{2.2.3}$$

As $\frac{d^2\vec{r}_i}{dt^2} = \vec{a}_i$ is the acceleration in the inertial frame, and $\frac{d^2\vec{r}_n}{dt^2} = \vec{a}_n$, the acceleration observed in the non-inertial frame, we can write Eq. (2.2.3) as

$$\vec{a}_i = \vec{a}_n + \vec{a}_0 \Rightarrow \vec{a}_i - \vec{a}_0 = \vec{a}_n \Rightarrow m\vec{a}_i - m\vec{a}_0 = m\vec{a}_n \tag{2.2.4}$$

If we define the force on the particle in the accelerated system according to Newton's second law, i.e., $m\vec{a}_n = \vec{F}_n$, then using Eqs. (2.2.1) and (2.2.2), we get

$$\vec{F}_n = \vec{F}_i + \vec{F}_0 \tag{2.2.5}$$

where $\vec{F}_i\,(= m\vec{a}_i)$ is the real force acting on the particle and $\vec{F}_0\,(= -m\vec{a}_0)$ is the fictitious force. Thus, the observer in the accelerated frame will measure the resultant (total) force which is the sum of real and fictitious forces on the particle i.e.,

Total force = True force + Fictitious force

2.3 Centrifugal Force

Let us consider a mass m, moving on the circumference of a circle of radius $\vec{r}$ with an angular velocity $\vec{\omega}$. For example, consider a stone attached at the end of a string. In an inertial frame, the centripetal force acting on the mass m is given by

$$\vec{F}_i = -m\omega^2\vec{r} \tag{2.3.1}$$

where $\vec{r}$ is directed outward from the centre O. In case of rotating string with stone, this centripetal force is provided by the tension $\vec{T}$ of the string. So that

$$\vec{F}_i = \vec{T} = -m\omega^2\vec{r} \tag{2.3.2}$$

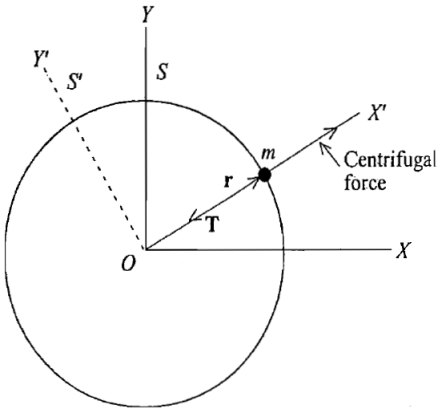


Figure 2.3.1: Centrifugal Force

Now suppose that a frame is rotating with an angular velocity ω relative to the inertial frame so that in the rotating frame the mass m is at rest. In this non-inertial (rotating) frame, the observed acceleration $(\vec{a}_n)$ of the mass m is zero, i.e., $\vec{a}_n = 0$ and consequently the total force $\left(\vec{F}_n\right)$ is given by

$$\vec{F}_n = m\vec{a}_n = 0. \tag{2.3.3}$$

Now

$$\vec{F}_i + \vec{F}_0 = \vec{F}_n \quad \text{i.e.,} \quad -m\omega^2\vec{r} + \vec{F}_0 = 0 \tag{2.3.4}$$

Thus

$$\vec{F}_0 = m\omega^2\vec{r} \tag{2.3.5}$$

This fictitious force $\left(\vec{F}_0\right)$ is directed away from the centre (along $\vec{r}$) and is called the *centrifugal force*.

2.4 Uniformly Rotating Frames

We know that the earth itself rotates about its axis in 24 hours. Therefore, any frame fixed with the earth will also rotate with it and so it will be a non-inertial frame.

Suppose that a frame $S'(X_r, Y_r, Z_r)$ is rotating with an angular velocity ω relative to an inertial frame $S(X_i, Y_i, Z_i)$. For simplicity, we assume that both of the frames have common origin O and common Z-axis. In case of the earth, the common origin O may be considered as the centre of the earth, Z-axes as coinciding with its rotational axis and the frame S' as rotating with earth relative to the non-rotating frame S.

The position vector of a particle P in both frames will be the same, i.e., $\vec{R}_i = \vec{R}_r = \vec{R}$, because the origins are coincident. Now, if the particle P is stationary in the frame S, the observer: in the rotating frame S' will see that the particle is moving oppositely with linear velocity $-\omega \times R$. Thus, if the velocity of the particle in the frame S is $\left(\frac{d\vec{R}}{dt}\right)_i$, then its velocity $\left(\frac{d\vec{R}}{dt}\right)_r$ in the rotating frame will be given by

$$\left(\frac{d\vec{R}}{dt}\right)_r = \left(\frac{d\vec{R}}{dt}\right)_i - \vec{\omega} \times \vec{R} \Rightarrow \left(\frac{d\vec{R}}{dt}\right)_i = \left(\frac{d\vec{R}}{dt}\right)_r + \vec{\omega} \times \vec{R} \tag{2.4.1}$$

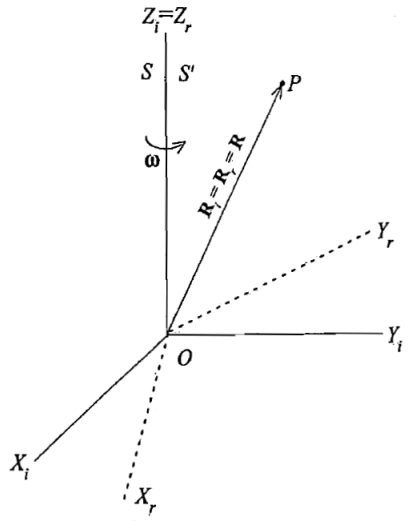


Figure 2.4.1: Uniformly rotating frame

In fact this equation holds for all vectors and relates the time derivatives of a vector in the frames S and S'. Therefore, relation (2.4.1) may be written in the form of operator equation:

$$\left(\frac{d}{dt}\right)_i = \left(\frac{d}{dt}\right)_r + \vec{\omega} \times \tag{2.4.2}$$

Writing $d\vec{R}/dt = v$ for the velocity of the particle, we have

$$\vec{v}_i = \vec{v}_r + \vec{\omega} \times \vec{R} \tag{2.4.3}$$

Now, if we operate Eq. (2.4.2) on velocity vector v_i, we have

$$\left(\frac{d\vec{v}_i}{dt}\right)_i = \left(\frac{d\vec{v}_i}{dt}\right)_r + \vec{\omega} \times \vec{v}_i$$

Substituting the value of $\vec{v}_i$ in the right hand side of this relation from Eq. (2.4.3), we obtain

$$\begin{aligned}\left(\frac{d\vec{v}_i}{dt}\right)_i &= \left[\frac{d}{dt}\left(\vec{v}_r + \vec{\omega} \times \vec{R}\right)\right]_r + \vec{\omega} \times \left(\vec{v}_r + \vec{\omega} \times \vec{R}\right) \\ &= \left(\frac{d\vec{v}_r}{dt}\right)_r + \frac{d\vec{\omega}}{dt} \times \vec{R} + \vec{\omega} \times \left(\frac{d\vec{R}}{dt}\right)_r + \omega \times \vec{v}_r + \vec{\omega} \times (\vec{\omega} \times \vec{R})\end{aligned}$$

If we write the acceleration $\dfrac{d\vec{v}}{dt} = \vec{a}$ and $\left(\dfrac{d\vec{R}}{dt}\right)_r = \vec{v}_r$, then

$$\begin{aligned}\vec{a}_i &= \vec{a}_r + 2\vec{\omega} \times \vec{v}_r + \vec{\omega} \times (\vec{\omega} \times \vec{R}) + \frac{d\vec{\omega}}{dt} \times \vec{R} \\ \text{or,}\quad \vec{a}_r &= \vec{a}_i - 2\vec{\omega} \times \vec{v}_r - \vec{\omega} \times (\vec{\omega} \times \vec{R}) - \vec{\omega} \times \vec{R}\end{aligned} \tag{2.4.4}$$

which relates the acceleration, $(\vec{a}_r)$ in the rotating frame to that $(\vec{a}_i)$ in the non-rotating (inertial) frame. If m is the mass of the particle, then force in the rotating frame is

$$m\vec{a}_r = m\vec{a}_i - 2m\vec{\omega} \times \vec{v}_r - m\vec{\omega} \times (\vec{\omega} \times \vec{R}) - m(\vec{\omega} \times \vec{R})$$

But $m\vec{a}_r = \vec{F}_i + \vec{F}_0$, therefore fictitious force $\vec{F}_0$ is given by

$$\vec{F}_0 = -2m(\vec{\omega} \times \vec{v}_r) - \vec{\omega} \times (\vec{\omega} \times \vec{R}) - m(\dot{\vec{\omega}} \times \vec{R})$$

where $-2m(\vec{\omega} \times \vec{v}_r)$ is the *Coriolis force*, $-m\vec{\omega} \times (\vec{\omega} \times \vec{R})$ the *centrifugal force* and $-m(\dot{\vec{\omega}} \times \vec{R})$ the *Euler force*. These forces appear to exist only in the rotating frame of reference.

For earth, $\vec{\omega}$ is constant, hence Euler force is zero, i.e.,

$$\vec{F}_0 = -2m\vec{\omega} \times \vec{v}_r - m\vec{\omega} \times (\vec{\omega} \times \vec{R})$$

This means that in case of a frame rotating with the earth, only Coriolis force $[-2m(\vec{\omega} \times \vec{v}_r)]$ and centrifugal force $[-m\vec{\omega} \times (\vec{\omega} \times \vec{R})]$ are the fictitious forces, acting on a moving particle. These forces are not due to any specific action applied to the particle. If the particle is at rest (i.e., $\vec{v}_r = \vec{0}$) in the rotating frame, then centrifugal force is only the fictitious force acting on the particle. On the other hand, if the particle is moving with velocity $\vec{v}_r \neq \vec{0}$ in the rotating frame, then in addition to centrifugal force, Coriolis force acts on the particle.

2.5 Motion Relative to the Earth

In the rotating frame with the earth, the acceleration of a particle is obtained from Eq. (2.4.4) with $\dot{\omega} = 0$ i.e.;

$$\vec{a}_r = \vec{a}_i - \vec{\omega} \times (\vec{\omega} \times \vec{R}) - 2\vec{\omega} \times \vec{v}_r \tag{2.5.1}$$

where $\vec{a}_i$ is the acceleration in the non-rotating or inertial frame, $-\vec{\omega} \times (\vec{\omega} \times \vec{R})$ is the *centrifugal acceleration* and $-2\vec{\omega} \times \vec{v}_r$ is the *Coriolis acceleration*.

An interesting application of Eq. (2.5.1) is the study of the motion of a body relative to a frame (S') rotating

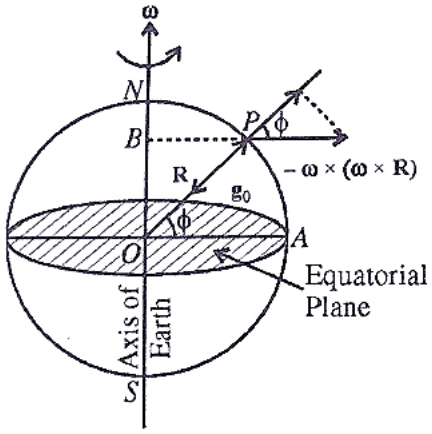


Figure 2.5.1: Motion relative to the earth

with the earth. The angular velocity of the earth is $\vec{\omega}$ along its axis of rotation. Now consider a particle P close to the earth's surface [Fig. 2.5.1]. If we call $\vec{g}_0$ as the acceleration due to gravity with respect to an inertial or non-rotating frame, then $\vec{a}_i = \vec{g}_0$ and hence the acceleration as measured by an observer rotating with the earth is [from Eq. (2.5.1)]

$$\vec{a}_r = \vec{g}_0 - \vec{\omega} \times (\vec{\omega} \times \vec{R}) - 2\vec{\omega} \times \vec{v}_r \tag{2.5.2}$$

Hence the acceleration of a body relative to the rotating earth depends on its velocity relative to the earth and on the position vector $\vec{R}$ of the body.

Now we shall discuss the effect of centrifugal and Coriolis accelerations separately.

1. **Effect of centrifugal force:** If the particle P is at rest on the earth's surface, then Coriolis term is zero and the acceleration $\vec{a}_r$ measured relative to the earth is called *effective acceleration due to gravity* $\vec{g}$. Thus

$$\vec{g} = \vec{g}_0 - \vec{\omega} \times (\vec{\omega} \times \vec{R}) \tag{2.5.3}$$

Obviously the centrifugal acceleration $-\vec{\omega} \times (\vec{\omega} \times \vec{R})$ acts in the outward direction along BP and its magnitude is $\omega^2 R \cos\phi$, where ϕ is the latitude at P.

Assuming earth to be spherical, we may consider g_0 to be pointing towards the centre of the earth along the radial direction. Due to the second term in (2.5.3), the direction of $\vec{g}$, called the vertical, deviates slightly from the radial direction and is determined by a plumb line. For practical purposes, the vertical may be assumed to coincide with the radial direction. The magnitude of $\vec{g}$ is slightly less than the value of $\vec{g}_0$ and can be expressed as

$$\vec{g} = \vec{g}_0 - \omega^2 R \cos^2\phi \tag{2.5.4}$$

The second term in Eq. (2.5.4) is very small (about 0.3%) compared to g_0 and accounts for most of the observed variations of acceleration due to gravity with latitude at earth's surface.

2. **Effect of Coriolis force:** Coriolis force $(-2m\vec{\omega} \times \vec{v}_r)$ is a fictitious force which acts on a particle only if it is in motion with respect to the rotating frame. Hence, in the rotating frame if a particle moves with velocity $\vec{v}_r$, then it always experiences a force $(-2m\vec{\omega} \times \vec{v}_r)$ perpendicular to its path opposite to the direction of vector product $\vec{\omega} \times \vec{v}_r$. The effect of Coriolis force is not considerable at small (particle) speeds. Let us discuss the effect of Coriolis force, when the particle is moving in a horizontal plane or falling freely on the earth.

   (i) **Particle moving in a horizontal plane:** The Coriolis force causes a moving particle in a horizontal plane in the northern hemisphere to deflect towards the right of its path. In the southern hemisphere, the deflection is towards the left of the path. Let a particle of mass m be projected with velocity $\mathbf{v}$ in a horizontal plane at a point P on earth's surface with latitude ϕ. Consider a frame XYZ fixed on the earth at P so that X-axis is vertical and YZ is horizontal plane [Fig. 2.5.2]. Now, $\vec{v} = v_y\hat{j} + v_z\hat{k}$ and $\vec{\omega} = \omega \sin\phi\hat{i} + \vec{\omega}\cos\phi\hat{k}$. The Coriolis force acting on the particle

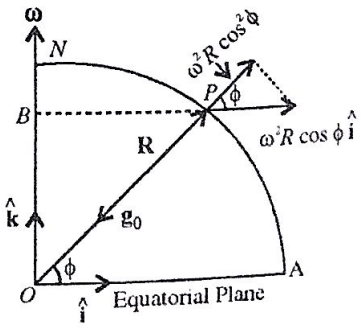


Figure 2.5.2: Particle moving in a horizontal plane

is

$$\begin{aligned}\vec{F} &= -2m(\vec{\omega}\times\vec{v}) \\ &= -2m(\omega\sin\phi\hat{i}+\omega\cos\phi\hat{k})\times\left(v_y\hat{j}+v_z\hat{k}\right) \\ &= 2m\omega v_y\cos\phi\hat{i}-2m\omega v_z\sin\phi\hat{j}+2m\omega v_y\sin\phi\hat{k} \qquad (2.5.5)\end{aligned}$$

If $\vec{F}_H$ and $\vec{F}_V$ are the horizontal and vertical components of the Coriolis force, then

$$\vec{F}_H = 2m\omega v_z\sin\phi\hat{j}-2m\omega v_y\sin\phi\hat{k}$$

Its magnitude $F_H = 2m\omega v\sin\phi$ Also $\vec{F}_V = 2m\omega v_y\cos\phi\hat{i}$. Its magnitude $F_V = 2m\omega v_y\cos\phi$. Obviously the horizontal component of the Coriolis force tends to deviate the path of the particle towards the right in the northern hemisphere. For example, if the particle is projected towards north, due to the Coriolis force it is deflected towards east.

The magnitude of horizontal Coriolis force is $2m\omega v\sin\phi$ and is zero at $\phi = 0$, i.e., at equator. The effect of Coriolis force is appreciable due to its horizontal component because in the vertical direction, its effect is masked by the large gravitational force.

(ii) **Free Fall of a Body on Earth's Surface:** One of the important effect of Coriolis force is that a freely falling body deviates from its true vertical path. This deviation is always towards east direction in either of the hemisphere of the earth. Due to to the horizontal component of earth's angular velocity, Coriolis force stars to act on freely falling boy in the horizontal direction and deviates it from the true vertical direction. We deduce below an expression fr this deviation.

Let us consider the free fall of a body from a height h on the surface of the earth at a latitude ϕ. The earth is rotating about its axis with an angular velocity $\vec{\omega}$. At the point P of the earth, take X-axis vertically, Y-axis along east and Z-axis along north [Fig. 2.6.1]. If $\hat{i},\hat{j},\hat{k}$, are unit vectors

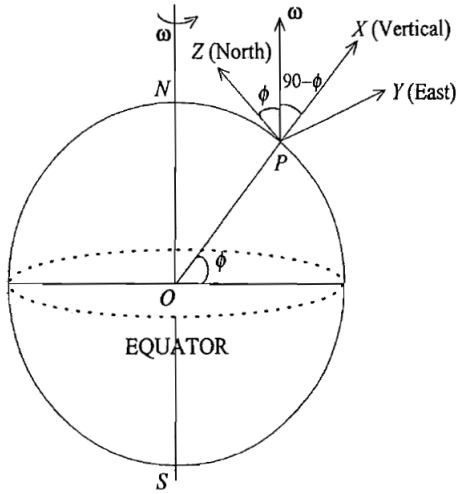


Figure 2.5.3: Free Fall of a Body on Earth's Surface

along these axes, then the angular velocity $\vec{\omega}$ can be represented as

$$\vec{\omega} = \omega\cos\left[\frac{\pi}{2}-\phi\right]\hat{i}+\omega\cos\phi\hat{k} = \omega(\sin\phi\hat{i}+\cos\phi\hat{k}) \qquad (2.5.6)$$

As the effective value of the acceleration due to gravity $\vec{g}$ is the combined effect of the centripetal acceleration and acceleration in the inertial frame, then substituting $\vec{a}_i - \vec{\omega} \times (\vec{\omega} \times \vec{R}) = -\hat{i}g$ in the equation $\vec{a}_i = \vec{a}_r + 2\vec{\omega} \times \vec{v}_r + \vec{\omega} \times (\vec{\omega} \times \vec{R})$, we get

$$-\hat{i}g = \vec{a}_r + 2\vec{\omega} \times \vec{v}_r \tag{2.5.7}$$

Here the velocity of the body $\vec{v}_r$ is almost along X-axis with negligible Y and Z components and hence we can have

$$\vec{v}_r = \hat{i}\frac{dx}{dt} \tag{2.5.8}$$

Writing $\vec{a}_r$ in component form, we get from Eq. (2.5.7)

$$\begin{aligned} -\hat{i}g &= \hat{i}\frac{d^2x}{dt^2} + \hat{j}\frac{d^2y}{dt^2} + \hat{k}\frac{d^2z}{dt^2} + 2\omega(\sin\phi\hat{i} + \cos\phi\hat{k}) \times \hat{i}\frac{dx}{dt} \\ \Rightarrow -\hat{i}g &= \hat{i}\frac{d^2x}{dt^2} + \hat{j}\left[\frac{d^2y}{dt^2} + 2\omega\frac{dx}{dt}\cos\phi\right] + \hat{k}\frac{d^2z}{dt^2} \end{aligned}$$

Now, comparing coefficients of $\hat{i}, \hat{j}$ and $\hat{k}$ of both sides of the above equation, we get the three component equations as

$$\frac{d^2x}{dt^2} = -g \tag{2.5.9}$$

$$\frac{d^2y}{dt^2} = -2\omega\frac{dx}{dt}\cos\phi \tag{2.5.10}$$

$$\frac{d^2z}{dt^2} = 0 \tag{2.5.11}$$

Integrating Eq. (2.5.9), we get

$$\frac{dx}{dt} = -gt + C$$

But initially at $t = 0$, $\dfrac{dx}{dt} = 0$, therefore $C = 0$. Thus

$$\frac{dx}{dt} = -gt. \tag{2.5.12}$$

Integrating it further, we get

$$x = -\frac{1}{2}gt^2 + C'$$

Initially at $t = 0$, the distance of the stone from the earth $x = h$. This gives $h = C'$, and hence

$$x = -\frac{1}{2}gt^2 + h$$

Finally, when at $t = T$, the stone touches the ground $x = 0$. Therefore

$$h - \frac{1}{2}gT^2 = 0 \quad \Rightarrow \quad T = \sqrt{2h/g} \tag{2.5.13}$$

Now,

$$\frac{d^2y}{dt^2} = -2\omega\frac{dx}{dt}\cos\phi \quad \text{or} \quad \frac{d^2y}{dt^2} = 2\omega gt\cos\phi$$

Integrating it, we get

$$\frac{dy}{dt} = \omega g t^2 \cos\phi \quad \text{and hence} \quad y = \frac{\omega g t^3}{3} \cos\phi \tag{2.5.14}$$

The constants of integration in (2.5.14) have been taken zero, because initially the stone has no displacement and velocity in the Y-direction is zero. Finally, at $t = T$, the maximum horizontal displacement will be given by

$$Y = \frac{\omega g T^3}{3} \cos\phi \tag{2.5.15}$$

Substituting the value of T from Eq. (2.5.13), we get

$$Y = \frac{\omega g}{3}\left[\frac{2h}{g}\right]^{3/2} \cos\phi \quad \text{or} \quad Y = \left[\frac{8}{9g}\right]^{1/2} h^{3/2}\omega\cos\phi \tag{2.5.16}$$

Thus a freely falling body at latitude ϕ is displaced horizontally due east by Coriolis force by an amount given by Eq. (2.5.16). At the equator ($\phi = 0$) the easterly deflection is obtained to be

$$Y = \left[\frac{8}{9g}\right]^{1/2} h^{3/2}\omega.$$

2.6 Some other effects of Coriolis force

1. **Formation of cyclones:** When a low pressure zone is created at a place on earth, the wind will flow radially from high pressure regions to the central low pressure region. However the Coriolis force deviates the air molecules toward the right of their path in the northern hemisphere, resulting in anticlockwise motion of the air particles and formation of the cyclone. The cyclone also moves along with the centre of low pressure region. In the southern hemisphere, the formation of cyclone is clockwise.

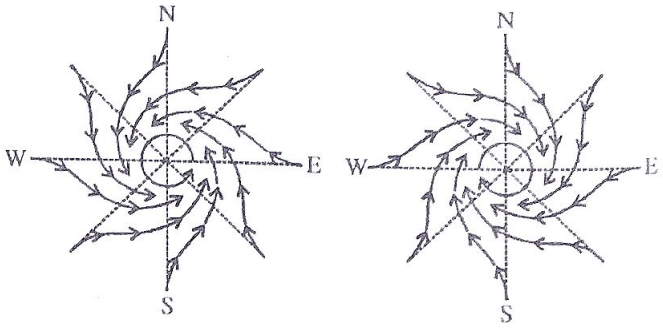


Figure 2.6.1: Formation of cyclones at Northern hemisphere (left side figure) and Southern hemisphere (right side figure)

2. In the northern hemisphere, flowing water in a river is acted by Coriolis force towards right of its path so that the right bank of the river is eroded more in comparison to its left bank. In the southern hemisphere, the effect is more on the left bank.

3. **Rotation of plane of oscillation of Foucault's pendulum:** Foucault's pendulum is similar to a simple pendulum, with a heavy bob and long strong suspension wire. If the pendulum is oscillated, Coriolis

force acts toward the right of its path which results in the rotation of the plane of oscillation slowly in the clockwise direction in the northern hemisphere. This experimental fact demonstrates that the earth rotates about its axis.

Exercise 2.6.1. 1. What are non-inertial frames and fictitious forces? Is the centrifugal force fictitious one?

2. Obtain equation of motion for a particle is rotating coordinate system.

3. What are Coriolis force? Show that the total Coriolis force acting on a body of mass m in a rotating frame is $-2m\vec{\omega} \times \vec{v}$, where $\vec{\omega}$ is the angular velocity of rotating frame and $\vec{v}$ is the velocity of the body in rotating frame.

4. A reference frame A rotates with respect to another reference frame B with uniform angular velocity $\vec{\omega}$. If the position, velocity and acceleration of a particle in frame A are represented by $\vec{R}$, $\vec{v}_a$ and $\vec{f}_a$ respectively, show that the acceleration of that particle in frame B is given by

$$\vec{f}_b = \vec{f}_a + 2\vec{\omega} \times \vec{v}_a + \vec{\omega} \times (\vec{\omega} \times \vec{R}).$$

Interpret this equation with reference to the motion of bodies on earth's surface.

5. A stone is allowed to fall under gravity from the top of a h meter high tower at the equator. Show that the horizontal displacement of the stone due to the earth's rotation is given by

$$y = \left(\frac{8}{9g}\right)^{1/2} h^{3/2}\omega.$$

6. A body is thrown vertically upward with a velocity u. Prove that it will fall back on a point displaced to the way by a distance equal to $\frac{4}{3}\left(\frac{8h^3}{g}\right)^{1/2} \omega \cos\phi$, where ϕ is the latitude and $h = u^2/2g$.

Unit 3

Course Structure

- Euler's theorem. Euler's equations of motion for a rigid body. Euler's angles.

3.1 Generalized coordinates of a rigid body

A rigid body is defined as a system of particles in which the distance between any two particles remains fixed throughout the motion. Thus a system of N particles is said to be a rigid body if it is subjected to holonomic constraints of the form

$$r_{ij} = C_{ij} \tag{3.1.1}$$

where r_{ij} is the distance between i-th and j-th particles and C_{ij} is the constant. In a rigid body motion, the deformations, occurring in actual bodies, are neglected and a rigid body maintains its shape during its motion.

We cannot obtain the actual number of degrees of freedom just by subtracting the number of constraint equations from $3N$ because there are $\frac{1}{2}N(N-1)$ possible constraint equations of the form (3.1.1). Obviously for large value of N, these constraint equations are more in number than $3N$. In fact, the equations represented by (3.1.1) are not all independent.

We can show in the following two ways that *the number of degrees of freedom for the general motion of a rigid body is six*, i.e., six independent coordinates are needed to specify the motion.

Let us consider three non-collinear particles P_1, P_2 and P_3 in a rigid body (Fig. 3.1.1). As each particle has three degrees of freedom, nine degrees of freedom in total is required. From (3.1.1), the three equations of constraints, expressed in terms of coordinates of the points relative to an arbitrary origin fixed in the body, are

$$\begin{aligned} r_{12} &= \left\{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2\right\}^{1/2} = C_{12} \quad \text{(constant length)} \\ r_{23} &= \left\{(x_2-x_3)^2+(y_2-y_3)^2+(z_2-z_3)^2\right\}^{1/2} = C_{23} \\ r_{13} &= \left[(x_1-x_3)^2+(y_1-y_3)^2+(z_1-z_3)^2\right]^{1/2} = C_{13} \end{aligned} \tag{3.1.2}$$

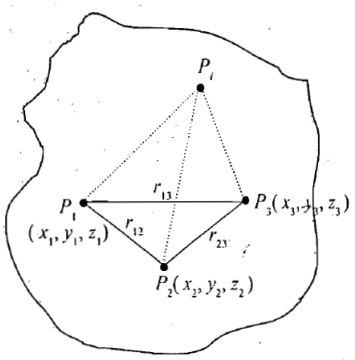


Figure 3.1.1: Degrees of freedom for the motion of a rigid boy - 3 reference points P_1, P_2, P_3 of the rigid body with three equations of rigid constraints allow the body to have six degrees of freedom

Hence the number of degrees of fredom of this three particle system is reduced to $9 - 3 = 6$. The position of any other particle in the body, say P_i, needs three coordinates and obviously there are three equations of constraints, because the distances of P_i from P_1, P_2 and P_3 are fixed. Hence any other particle will not add any new degree of freedom to six degrees of freedom of the three-particle system. Thus the motion of a rigid body be specified by six degrees of freedom. In other words, we need six independent or generalized coordinates to specify the motion of a rigid body.

To look the situation in other way, the position of the particle P_1 needs three coordinates. Relative to P_1, the position of P_2 can be specified by only two coordinates because of one constraint equation $r_{12} = C_{12}$. The third particle P_3 relative to P_1 and P_2 has only one degree of freedom because of two constraints $r_{13} = C_{13}$ and $r_{23} = C_{23}$. Thus the three particles (non-collinear) of the rigid body have $3 + 2 + 1 = 6$ degrees of freedom. It is to be noted that (3.1.1) particle P_2 relative to P_1 is constrained to move on the surface of a sphere and its position can be specified by two angles, and (3.1.2) particle P_3 relative to P_1 and P_2 can only rotate about the axis joining P_1 and P_2 which can be specified by third angle. Intuitively, one can think that rigid body should possess three translational and three rotational degrees of freedom. Therefore, in order to describe the motion of a rigid body, we usually choose three of these coordinates to be the coordinates of a point in the-body (generally the centre of mass) and the remaining three to be the three angles (usually three *Eulerian angles*) which describe the rotation of the body about the point.

In addition to the constraints of rigidity, if the body has additional constraints, this will further reduce the number of degrees of freedom and hence the number of independent generalized coordinates.

3.1.1 Body and space reference system

We may describe the motion of a rigid body by using two coordinate systems

1. **Body coordinate system:** A coordinate system, fixed in the rigid boy, is called a body coordinate system and its axes are called body set of axes.

2. **Space coordinate system:** The axes of such a coordinate system are fixed in the space set of axes.

3.2 Euler's equations of motion for a rigid body

3.2.1 Newtonian method

If a rigid body is rotating under the action of a torque τ with one point fixed, then the torque is expressed as

$$\tau = \left[\frac{dJ}{dt}\right]_s \tag{3.2.1}$$

where J is the angular momentum and its time derivative refers to the space set of axes, represented by the subscript s, because the equation holds in an inertial frame.

The body coordinate system is rotating with an instantaneous angular velocity ω. The time derivatives of angular momentum J in the body coordinate and space coordinate systems are related as

$$\left[\frac{dJ}{dt}\right]_s = \left[\frac{dJ}{dt}\right]_b + \omega \times J \tag{3.2.2}$$

Thus

$$\vec{\tau} = \frac{d\vec{J}}{dt} + \vec{\omega} \times \vec{J} \tag{3.2.3}$$

where we have dropped the body subscript because we shall represent the physical quantities of right hand side in the body coordinate system.

We choose principal axes for body set of axes. If I_1, I_2 and I_3 are the principal moments of inertia, then

$$\vec{J} = I_1\omega_1\hat{i} + I_2\omega_2\hat{j} + I_3\omega_3\hat{k} \tag{3.2.4}$$

where $\vec{\omega} = \omega_1\hat{i} + \omega_2\hat{j} + \omega_3\hat{k}$ is the angular velocity with components ω_1, ω_2 and ω_3 along the principal axes.

As the principal moments of inertia and body base vectors $\hat{i}, \hat{j}$ and $\hat{k}$ are constants in time with respect to the body coordinate system, we find that in the body coordinate system, using (3.2.4) the time derivative of $\vec{J}$ is

$$\frac{d\vec{J}}{dt} = I_1\dot{\omega}_1\hat{i} + I_2\dot{\omega}_2\hat{j} + I_3\dot{\omega}_3\hat{k} \tag{3.2.5}$$

Substituting in (3.2.3), we obtain

$$\vec{\tau} = I_1\dot{\omega}_1\hat{i} + I_2\dot{\omega}_2\hat{j} + I_3\dot{\omega}_3\hat{k} + \left(\omega_1\hat{i} + \omega_2\hat{j} + \omega_3\hat{k}\right) \times \left(I_1\omega_1\hat{i} + I_2\omega_2\hat{j} + I_3\omega_3\hat{k}\right) \tag{3.2.6}$$

Writing $\vec{\tau} = \tau_1\hat{i} + \tau_2\hat{j} + \tau_3\hat{k}$, we can obtain the x, y, z components of the torque $\vec{\tau}$ as

$$\tau_1 = I_1\dot{\omega}_1 + (I_3 - I_2)\,\omega_2\omega_3 \tag{3.2.7}$$

$$\tau_2 = I_2\dot{\omega}_2 + (I_1 - I_3)\,\omega_3\omega_1 \tag{3.2.8}$$

$$\tau_3 = I_3\dot{\omega}_3 + (I_2 - I_1)\,\omega_1\omega_2 \tag{3.2.9}$$

Eqs. (3.2.7), (3.2.8), (3.2.9) are known as *Euler's equations* for the motion of a rigid body with one point fixed under the action of a torque. These equations can also be derived from Lagrange's equations, when the generalized forces G_k are the torques and Euler's angles (ϕ, θ, ψ) are the generalized coordinates.

3.2.2 Lagrange's method

When a rigid body is rotating with one point fixed, Euler's angles completely describe the orientation of the rigid body. In case of the rotating rigid body, we take the Euler's angles ϕ, θ, ψ as the generalized coordinates and components of the applied torque as the generalized forces corresponding to these angles. For conservative system, Lagrangian for the system is

$$L = T(\dot{\phi}, \dot{\theta}, \dot{\psi}, \phi, \theta, \psi) - V(\phi, \theta, \psi) \tag{3.2.10}$$

where T is the rotational kinetic energy and is given by

$$T = \frac{1}{2}\left(I_1\omega_1^2 + I_2\omega_2^2 + I_3\omega_3^2\right) \tag{3.2.11}$$

where the body axes are taken as principal axes. In view of Euler's geometrical equations, the angular velocity components ω_1, ω_2 and ω_3 along the principal axes can be written as

$$\begin{aligned} \omega_1 &= \dot{\phi}\sin\theta\sin\psi + \dot{\theta}\cos\psi \\ \omega_2 &= \dot{\phi}\sin\theta\cos\psi - \dot{\theta}\sin\psi \\ \omega_3 &= \dot{\phi}\cos\theta + \dot{\psi} \end{aligned} \tag{3.2.12}$$

The Lagrange's equation for ψ coordinate is

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{\psi}}\right] - \frac{\partial L}{\partial \psi} = 0$$

But for $L = T - V$, given by (3.2.10),

$$\frac{d}{dt}\left[\frac{\partial T}{\partial \dot{\psi}}\right] - \frac{\partial T}{\partial \psi} = -\frac{\partial V}{\partial \psi} \tag{3.2.13}$$

because $\partial V/\partial\dot{\psi} = 0$. However, the angle ψ is the angle of rotation about the principal Z-axis and is one of the generalized coordinates in the present problem. The generalized force $[G_\psi = -\partial V/\partial\psi]$ corresponding to the generalized coordinate ψ is obviously the Z-component of the impressed torque i.e.,

$$\tau_3 = G_\psi = -\frac{\partial V}{\partial \psi} \tag{3.2.14}$$

Thus Eq. (3.2.13) assumes the form

$$\begin{aligned} \tau_3 &= \frac{d}{dt}\left[\frac{\partial T}{\partial \dot{\psi}}\right] - \frac{\partial T}{\partial \psi} \\ \text{or} \quad \tau_3 &= \frac{d}{dt}\left[\sum_i \frac{\partial T}{\partial \omega_i}\frac{\partial \omega_i}{\partial \dot{\psi}}\right] - \frac{\partial}{\partial \psi}\left[\sum_i \frac{\partial T}{\partial \omega_i}\frac{\partial \omega_i}{\partial \psi}\right] \end{aligned} \tag{3.2.15}$$

But from (3.2.11), we get

$$T = \frac{1}{2}\sum_i I_i\omega_i^2$$

Therefore, $\dfrac{\partial T}{\partial \omega_i} = I_i\omega_i$. From (3.2.12), we obtain

$$\frac{\partial \omega_1}{\partial \dot{\psi}} = \frac{\partial \omega_2}{\partial \dot{\psi}} = 0 \quad \text{and} \quad \frac{\partial \omega_3}{\partial \dot{\psi}} = 1$$

So that

$$\sum_i \frac{\partial T}{\partial \omega_i}\frac{\partial \omega_i}{\partial \dot{\psi}} = I_3\omega_3 \tag{3.2.16}$$

Also from (3.2.12), we get

$$\frac{\partial \omega_1}{\partial \psi} = -\dot{\phi}\sin\theta\cos\psi - \dot{\theta}\sin\psi = \omega_2,$$
$$\frac{\partial \omega_2}{\partial \psi} = -\dot{\phi}\sin\theta\sin\psi - \dot{\theta}\cos\psi = -\omega_1, \quad \frac{\partial \omega_3}{\partial \psi} = 0$$

Hence

$$\begin{aligned}\sum_i \frac{\partial T}{\partial \omega_i}\frac{\partial \omega_i}{\partial \psi} &= \frac{\partial T}{\partial \omega_1}\frac{\partial \omega_1}{\partial \psi} + \frac{\partial T}{\partial \omega_2}\frac{\partial \omega_2}{\partial \psi} + \frac{\partial T}{\partial \omega_3}\frac{\partial \omega_3}{\partial \psi} \\ &= I_1\omega_1\omega_2 + I_2\omega_2(-\omega_1) = -(I_2 - I_1)\omega_1\omega_2\end{aligned} \tag{3.2.17}$$

Substituting the values from (3.2.7), (3.2.8), (3.2.9) and (3.2.17) in (3.2.15), we get

$$\begin{aligned}\tau_3 &= \frac{d}{dt}(I_3\omega_3) + (I_2 - I_1)\omega_1\omega_2 \\ \text{or,} \quad \tau_3 &= I_3\dot{\omega}_3 + (I_2 - I_1)\omega_1\omega_2\end{aligned} \tag{3.2.18}$$

which is the third Euler's equation obtained earlier. One may obtain the other two Euler's equations by simply cyclic permutation. Note that these two equations do not correspond to θ and ϕ coordinates.

In case a rigid body is rotating about a fixed axis, say principal Z-axis, then

$$\omega_1 = \omega_2 = 0 \quad \text{and} \quad \omega_3 = \omega$$

Therefore, from Eqs. (3.2.18) we have the equations of motion as

$$\tau_1 = \tau_2 = 0$$

and

$$\tau_3 = I_3\dot{\omega} \quad \text{or} \quad \tau = I\dot{\omega} \tag{3.2.19}$$

where we have put $\tau_3 = \tau$ and $I_3 = I$ corresponding to Z-axis.

Instantaneous angular momentum about Z-axis is

$$J_3 = I_3\omega_3 \quad \text{or} \quad J = I\omega \tag{3.2.20}$$

and instantaneous rotational kinetic energy is

$$T = \frac{1}{2}\vec{\omega}\cdot\vec{J} = \frac{1}{2}I\omega^2 \tag{3.2.21}$$

3.3 Torque free motion of a rigid body

Equations of motion: When a rigid body is not subjected to any net torque, the Euler's equations of motion of the body with one point fixed reduce to

$$I_1\dot{\omega}_1 = (I_2 - I_3)\,\omega_2\omega_3 \tag{3.3.1}$$
$$I_2\dot{\omega}_2 = (I_3 - I_1)\,\omega_3\omega_1 \tag{3.3.2}$$
$$I_3\dot{\omega}_3 = (I_1 - I_2)\,\omega_1\omega_2 \tag{3.3.3}$$

In case the body is not subjected to any net forces or torques, its centre of mass is either at rest or moves with uniform velocity. Obviously we may discuss the rotational motion of the rigid body in a reference system in which the centre of mass is stationary and choose the centre of mass as fixed point and origin for the principal axes in the body. In such a case, we obtain from (3.3.1), (3.3.2) and (3.3.3) two integrals of motion, describing the kinetic energy and angular momentum as constant in time.

If we multiply Eqs. (3.3.1), (3.3.2) and (3.3.3) by $\omega_1, \omega_2, \omega_3$ respectively and then add, we obtain

$$\begin{aligned} &I_1\omega_1\dot{\omega}_1 + I_2\omega_2\dot{\omega}_2 + I_3\omega_3\dot{\omega}_3 = (I_2 - I_3 + I_3 - I_1 + I_1 - I_2)\,\omega_1\omega_2\omega_3 = 0 \\ &\Rightarrow \frac{d}{dt}\left(\frac{1}{2}I_1\omega_1^2 + \frac{1}{2}I_2\omega_2^2 + \frac{1}{2}I_3\omega_3^2\right) = 0 \\ &\Rightarrow \frac{1}{2}I_1\omega_1^2 + \frac{1}{2}I_2\omega_2^2 + \frac{1}{2}I_3\omega_3^2 = \frac{1}{2}\vec{\omega}\cdot\vec{J} = \text{ constant} \end{aligned} \tag{3.3.4}$$

which is the *principle of conservation of total rotational kinetic energy* in absence of external torque. As

$$\tau = \frac{d\vec{J}}{dt} = 0\vec{J} = I_1\omega_1\hat{i} + I_2\omega_2\hat{j} + I_3\omega_3\hat{k} = \text{ constant}$$

describes another constant of motion, representing *the principle of conservation of angular momentum.*

3.4 Euler's Angles

We are interested in knowing three independent parameters to specify the orientation of body set of axes relative to the space set of axes. For this purpose, we use three, angles. These angles may be chosen in various ways, but the most commonly used set of three angles are the Euler's angles, represented by ϕ, θ and ψ.

We can reach an arbitrary orientation of the body set of axes $X'Y'Z'$ from space set of axes (XYZ) by making three successive rotations performed in a specific order.

1. **First rotation** (ϕ)**:** First the space set of axes is rotated through an angle ϕ counter-clockwise about the Z-axis so that YZ plane takes the new position Y_1Z_1 and this new plane Y_1Z_1 contains the Z'-axis of the body coordinate system.

   Now the new position of the coordinate system is $X_1Y_1Z_1$ (with $Z = Z_1$). [Fig. 3.4.1]. If $\hat{i}', \hat{j}', \hat{k}'$ are the unit vectors along X, Y, Z axes and $\hat{i}_1, \hat{j}_1, \hat{k}_1$ along X_1, Y_1, Z_1 axes respectively, then the transformation to this new set of axes from space set of axes is represented by the equations

$$\begin{aligned} \hat{i}_1 &= \cos\phi\hat{i} + \sin\phi\hat{j} \\ \hat{j}_1 &= -\sin\phi\hat{i} + \cos\phi\hat{j} \\ \hat{k}_1 &= \hat{k} \end{aligned} \tag{3.4.1}$$

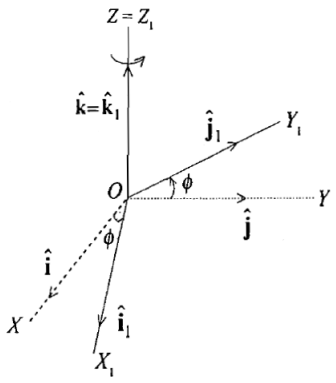


Figure 3.4.1: Euler's angles - First rotation ϕ, defining precession angle

$$\text{or,} \quad \begin{bmatrix} \hat{i}_1 \\ \hat{j}_1 \\ \hat{k}_1 \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{i} \\ \hat{j} \\ \hat{k} \end{bmatrix} \tag{3.4.2}$$

Thus XYZ axes are transformed to $X_1Y_1Z_1$ by the matrix of transformation

$$\begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.4.3}$$

The angle ϕ is called the *precession angle*.

2. **Second rotation (θ):** Next intermediate axes $X_1Y_1Z_1$ are rotated about X_1 axis counter-clockwise through an angle θ to the position X_2. Y_2Z_2 so that Y_1, Z_1 axes acquire the positions Y_2, Z_2 with $Z_2 = Z'$ [Fig. 3.4.2]. This also results the plane $X_2.Y_2$ in plane $X'Y'$. If $\hat{i}_2, \hat{j}_2, \hat{k}_2$ are unit vectors along X_2, Y_2, Z_2 axes respectively, then

$$\begin{aligned} \hat{i}_2 &= \hat{i}_1 \\ \hat{j}_2 &= \cos\theta\hat{j}_1 + \sin\theta\hat{k}_1 \\ \hat{k}_2 &= -\sin\theta\hat{j}_1 + \cos\theta\hat{k}_1 \end{aligned}$$

or

$$\begin{bmatrix} \hat{i}_2 \\ \hat{j}_2 \\ \hat{k}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \hat{i}_1 \\ \hat{j}_1 \\ \hat{k}_1 \end{bmatrix} \tag{3.4.4}$$

In this case the matrix of transformation is

$$C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \tag{3.4.5}$$

The angle θ is called the *nutation angle*. Then $X_2 = X_1$ axis is at the intersection of the XY and X_2Y_2 planes and is called the *line of nodes*.

3. **Third rotation (ψ):** Finally the third rotation is performed about $Z_2 = Z'$ axis through an angle ψ counter-clockwise so that X_2, Y_2 axes coincide $X_3 = X', Y_3 = Y'$ [Fig 3.4.3].

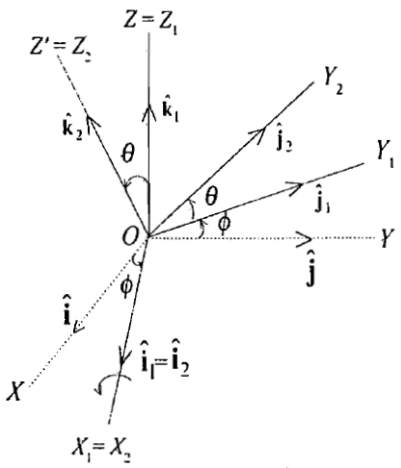


Figure 3.4.2: Euler's angles - Second rotation θ, defining nutation angle

Thus these three rotations ϕ, θ and ψ bring the space set of axes to coincide with body set of axes. The ϕ, θ and ψ are the Euler's angles and completely specify the orientation of the $X'Y'Z'$ system relative to the XYZ system. These ϕ, θ and ψ angles can be taken as three generalized coordinates. Now

$$\begin{aligned}
\hat{i}_3 &= \hat{i}' = \hat{i}_2 \cos\psi + \hat{j}_2 \sin\psi \\
\hat{j}_3 &= \hat{j}' = -\hat{i}_2 \sin\psi + \hat{j}_2 \cos\psi \\
\hat{k}_3 &= \hat{k}' = \hat{k}_2
\end{aligned}$$

$$\text{or} \quad \begin{bmatrix} \hat{i}' \\ \hat{j}' \\ \hat{k}' \end{bmatrix} = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{i}_2 \\ \hat{j}_2 \\ \hat{k}_2 \end{bmatrix} \tag{3.4.6}$$

So that the transformation matrix is

$$B = \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{3.4.7}$$

The angle ψ is called the *body angle*.

In this way we have reached at the body set of axes after three successive rotations of space set of axes. We may write the complete matrix of transformations A as

$$\begin{bmatrix} \hat{i}' \\ \hat{j}' \\ \hat{k}' \end{bmatrix} = A \begin{bmatrix} \hat{i} \\ \hat{j} \\ \hat{k} \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = A \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{3.4.8}$$

But using Eqs. (3.4.2), (3.4.3), (3.4.4), (3.4.5), (3.4.6) and (3.4.7)

$$\begin{bmatrix} \hat{i}' \\ \hat{j}' \\ \hat{k}' \end{bmatrix} = \begin{bmatrix} \hat{i}_3 \\ \hat{j}_3 \\ \hat{k}_3 \end{bmatrix} = B \begin{bmatrix} \hat{i}_2 \\ \hat{j}_2 \\ \hat{k}_2 \end{bmatrix} = BC \begin{bmatrix} \hat{i}_1 \\ \hat{j}_1 \\ \hat{k}_1 \end{bmatrix} = BCD \begin{bmatrix} \hat{i} \\ \hat{j} \\ \hat{k} \end{bmatrix} \tag{3.4.9}$$

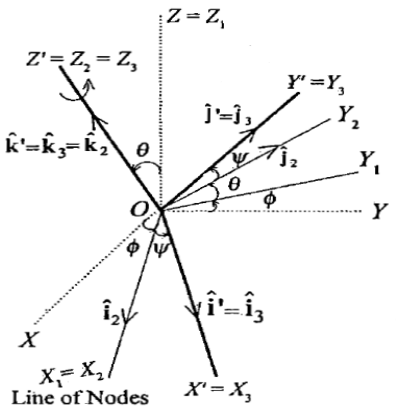


Figure 3.4.3: Euler's angles - Third rotation Ψ, defining body angle

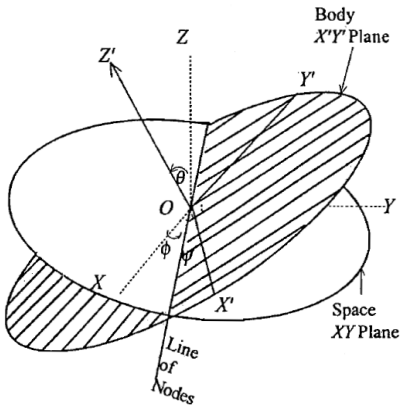


Figure 3.4.4: The three Eulerian angle ϕ, θ and Ψ in different planes

From (3.4.8) and (3.4.9) we see that the complete matrix of transformation from space set of axes to body set of axes is

$$A = BCD \tag{3.4.10}$$

The inverse transformation from body set of axes to space set of axes will be given by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = A^{-1} \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}$$

Now

$$\begin{aligned} A = BCD &= \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\phi & \sin\phi & 0 \\ -\cos\theta\sin\phi & \cos\theta\cos\phi & \sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{pmatrix} \end{aligned}$$

The inverse transformation matrix from body set of axes to space set of axes is given by $A^{-1} = A_T$ because A represents a proper orthogonal matrix. Thus

$$\mathbf{A}^{-1} = \begin{pmatrix} \cos\psi\cos\phi & -\sin\psi\cos\phi & \sin\theta\sin\phi \\ -\cos\theta\sin\phi\sin\psi & -\cos\psi\cos\theta\sin\phi & \\ \cos\psi\sin\phi & -\sin\psi\sin\phi & -\sin\theta\cos\phi \\ +\sin\psi\cos\theta\cos\phi & +\cos\psi\cos\theta\cos\phi & \\ \sin\psi\sin\theta & \cos\psi\sin\theta & \cos\theta \end{pmatrix} \tag{3.4.11}$$

Exercise 3.4.1. 1. How will you assign the generalized coordinates for the motion of a rigid body? For a rigid body consisting of N particles, how many generalized coordinates will have to be specified?

2. Define Euler's angle for the orientation of a rigid body.

3. Define Euler's angle and obtain an expression for the complete transformation matrix.

4. Derive Euler's equations of motion for a rigid body.

5. If T be the kinetic energy, $\vec{G}$ be the external torque about the instantaneous axis of rotation and $\vec{\omega}$ the angular velocity, then prove that
$$\frac{dT}{dt} = \vec{G} \cdot \vec{\omega}$$

6. From Euler's equations of motion for a rigid body, having no external torque about a fixed point, prove that
$$T = \frac{1}{2} I_1 \omega_1^2 + \frac{1}{2} I_2 \omega_2^2 + \frac{1}{2} I_3 \omega_3^2 = \text{constant, and } \vec{J} = I_1 \omega_1 \hat{i} + I_2 \omega_2 \hat{j} + I_3 \omega_3 \hat{k} = \text{constant}$$

Unit 4

Course Structure

- Variational Principle
- Calculus of variations and its applications in shortest distance
- Minimum surface of revolution
- Brachistochrone problem.
- Geodesics

4.1 Introduction

The Euler-Lagrangian approach to classical mechanics stems from a deep philosophical belief that the laws of nature are based on a principle of economy.That is, the physical universe follows paths through space and time that are based on extrema principles. The standard Lagrangian L is defined as the difference between the kinetic and potential energy, that is $L - U$. The laws of classical mechanics can be expressed in terms of Hamilton's variational principle which states that the motion of the system between the initial time t_1 and final time t_2 follows a path that minimizes the scalar action integral S defined as the time integral of the Lagrangian.

$$S = \int_{t_1}^{t_2} L \, dt$$

4.2 The Calculus of Variations and Euler-Lagrange Equation

Let us have a function $f\,(y, y', x)$ defined on a curve given by

$$y = y(x) \tag{4.2.1}$$

between two points $A\,(x_1, y_1)$ and $B\,(x_2, y_2)$. Here, $y' = dy/dx$. We are interested in finding a particular curve $y(x)$ for which the line integral I of the function f between the two points

$$I = \int_{x_1}^{x_2} f(y, y', x) \, dx \tag{4.2.2}$$

has a stationary value. Suppose that APB be the curve for which I is stationary. Now, consider a neighbouring curve $AP'B$ with the same end points A and B. The point $P(x, y)$ of the curve APB corresponds to the point $P'(x, y + \delta y)$ of the curve $AP'B$, keeping x coordinate of the points fixed. This defines a δ-variation of the curve. The variation is arbitrary but small and may be expressed as

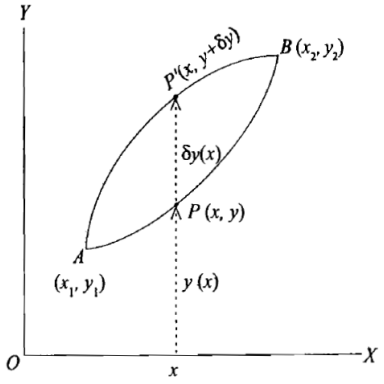


Figure 4.2.1: δ - variation

$$\delta y = \frac{\partial y}{\partial \alpha}\delta\alpha = \eta(x)\delta\alpha \tag{4.2.3}$$

where α is a parameter (independent of x) common to all points of the path and $\eta(x)$ is a function of x with the condition that

$$\delta y_1 = \delta y_2 = \eta\left(x_1\right) = \eta\left(x_2\right) = 0 \tag{4.2.4}$$

By choosing different $\eta(x)$, we may construct different varied paths. The corresponding variation in y' is

$$\delta y' = \eta'(x)\ \delta\alpha \tag{4.2.5}$$

Now, the integral on the varied path is

$$I' = \int_{x_1}^{x_2} f\left(y + \delta y, y' + \delta y', x\right) dx$$

$$\text{or,}\quad I' = \int_{x_1}^{x_2} f\left(y + \eta\delta\alpha, y' + \eta'\delta\alpha, x\right) dx \tag{4.2.6}$$

Since the variation is small, the integral I' may be obtained by considering only first order terms in the Taylor expansion of the function f i.e.,

$$I' = \int_{x_1}^{x_2} \left[f\left(y, y', x\right) + \frac{\partial f}{\partial y}\eta\delta\alpha + \frac{\partial f}{\partial y'}\eta'\delta\alpha\right] dx \tag{4.2.7}$$

Hence

$$\delta I = I' - I = \delta\alpha \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\eta + \frac{\partial f}{\partial y'}\eta'\right) dx \tag{4.2.8}$$

But

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\eta' dx = \frac{\partial f}{\partial y'}\eta\Big|_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\eta\, dx = -\int_{x_1}^{x_2} \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\eta\, dx \quad [\because \eta(x_1) = \eta(x_2) = 0]$$

Therefore,

$$\delta I = \delta\alpha \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\right]\eta\, dx \tag{4.2.9}$$

The condition that the integral I is stationary means that $\delta I = 0$, i.e.,

$$\int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\right]\eta\, dx = 0 \tag{4.2.10}$$

As η is arbitrary, the integrand of (4.2.10) must be zero, i.e.,

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0 \tag{4.2.11}$$

which is known as *Euler-Lagrange equation*.

The result can easily be generalized to the case where f is a function of many independent variables y_k and their derivatives y_k'. However, y_k and y_k' are function of x. Then

$$\delta I = \delta \int_{x_1}^{x_2} f\left(y_1, y_2, \ldots, y_k, \ldots, y_n, y_1', y_2', \ldots, y_k', \ldots y_n', x\right)\, dx = 0 \tag{4.2.12}$$

leads to the Euler-Lagrange equations

$$\frac{\partial f}{\partial y_k} - \frac{d}{dx}\left(\frac{\partial f}{\partial y_k'}\right) = 0 \tag{4.2.13}$$

where, $k = 1, 2, \ldots, n$. It is to be pointed out that in most of the problems the stationary value of the integral is seen to be a minimum but occasionally maximum.

4.3 Application of Variational Principle to Shortest Distance

Example 4.3.1. Show that the shortest distance between two points in a plane is a straight line.

Solution. Suppose $A\,(x_1, y_1)$ and $B\,(x_2, y_2)$ are two points in XY plane. An element of length ds of any curve, say $AP'B$, passing through A and B points is given by

$$ds^2 = dx^2 + dy^2 \Rightarrow ds = \sqrt{1 + y'^2}dx$$

Total length of the curve from point A to the point B is given by

$$S = \int_A^B \sqrt{1 + y'^2}dx = \int_A^B f\, dx$$

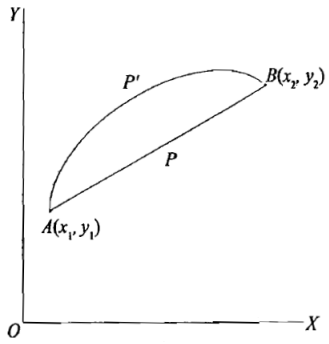


Figure 4.3.1: Shortest distance between two points in a plane

where $f = \sqrt{1 + y'^2}$. The length of the curve s will be minimum, when $\delta s = 0$. This means that f should satisfy the Euler Lagrange's equation, i.e.,

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0$$

Here, $\dfrac{\partial f}{\partial y} = 0$ and $\dfrac{\partial f}{\partial y'} = \dfrac{y'}{\sqrt{i + y'^2}}$.
Therefore,

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{1 + y'^2}}\right) = 0 \quad \text{or} \quad \frac{y'}{\sqrt{1 + y'^2}} = C, \text{ a constant}$$

$$\Rightarrow y'^2 = C^2\left(1 + y'^2\right) \text{ or } y'^2\left(1 - C^2\right) = C^2 \text{ or } y' = \frac{C}{\sqrt{1 - C^2}} = a \text{ (constant } \Big)$$

$$\Rightarrow \frac{dy}{dx} = a$$

Integrating it, we get

$$y = ax + b$$

where b is a constant of integration. This represents a straight line. Therefore the shortest distance between any two points in a plane is a straight line. The constants of integration a and b can be determined by the condition that the straight line passes through $A\,(x_1, y_1)$ and $B\,(x_2, y_2)$.

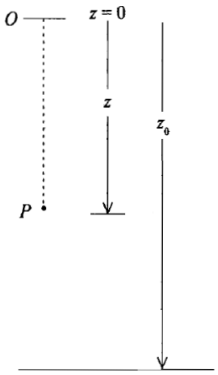
Example 4.3.2. A particle of mass m falls a given distance z_0 in time $t_0 = \sqrt{2z_0/g}$ and the distance travelled in time t is given by $z = at + bt^2$, where constants a and b are such that the time t_0 is always the same. Show that the integration $\int_0^{t_0} L dt$ is an extremum for real values of the coefficients only when $a = 0$ and $b = \dot{g}/2$.

Solution. Solution : Let the particle fall from $O(z = 0)$ to $P(OP = z)$ in time t. Kinetic energy of the particle at P,

$$T = \frac{1}{2}m\dot{z}^2.$$

Potential energy of the particle at P, $V = -mgz$ Hence

$$L = T - V = \frac{1}{2}m\dot{z}^2 + mgz \tag{4.3.1}$$



According to the Hamilton's principle

$$\delta \int_0^{t_0} L dt = 0 \Rightarrow \int_0^{t_0} L dt = \text{ extremum, for which}$$

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{z}}\right] - \frac{\partial L}{\partial z} = 0 \tag{4.3.2}$$

is to be satisfied. Here, $\frac{\partial L}{\partial \dot{z}} = m\dot{z}$ and $\frac{\partial L}{\partial z} = mg$. Hence the Euler-Lagrange's equation becomes

$$\frac{d}{dt}(m\dot{z}) - mg = 0 \Rightarrow \ddot{z} = g \tag{4.3.3}$$

$$\text{But } z = at + bt^2 \text{ and therefore } \dot{z} = a + 2bt \text{ and } \ddot{z} = 2b \tag{4.3.4}$$

From (4.3.3) and (4.3.4), we get

$$2b = g \Rightarrow b = g/2 \tag{4.3.5}$$

Also at $t = t_0, z = z_0$, we have

$$z_0 = at_0 + bt_0^2 \tag{4.3.6}$$

But

$$t_0 = \sqrt{\frac{2z_0}{g}} \Rightarrow z_0 = \frac{1}{2}gt_0^2 \tag{4.3.7}$$

Comparing (4.3.6) and (4.3.7) and putting $b = g/2$, we get

$$at_0 + \frac{g}{2}t_0^2 = \frac{1}{2}gt_0^2 \Rightarrow at_0 = 0$$

Since $t_0 \neq 0$, therefore, $a = 0$. Thus we find that $\int_0^{t_0} L dt$ is extremum, when $a = 0, b = \mathrm{g}/2$.

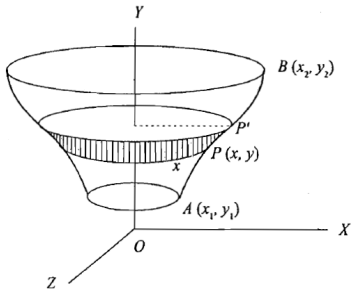


Figure 4.4.1: Minimum surface area of revolution

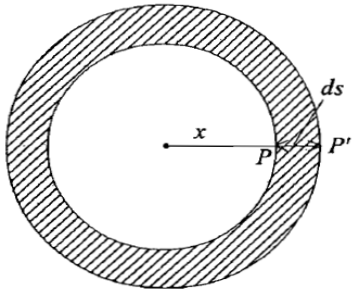


Figure 4.4.2: Circular strip of area $2\pi x\ ds$

4.4 Application of Variational Principle to Minimum Surface of Revolution

Example 4.4.1. We take a curve passing through the fixed points (x_1, y_1) and (x_2, y_2) and revolve it about Y-axis to form a surface of revolution. Find the equation of the curve for which the surface area is minimum.

Solution. Let AB be the curve which passes through the fixed points $A\,(x_1, y_1)$ and $B\,(x_2, y_2)$. The curve AB has been revolved about Y-axis to generate a surface. Consider a strip of the surface whose radius is x and breadth is $PP' = ds$, given by

$$ds^2 = dx^2 + dy^2 \text{ or } ds = \sqrt{1 + y'^2}dx$$

Area of the strip $dS = 2\pi x\ ds == 2\pi x\sqrt{1 + y'^2}dx$ (Fig. 4.4.2).
Total area of revolution

$$S = 2\pi \int_A^B x\sqrt{1 + y'^2}\ dx \tag{4.4.1}$$

This area will be minimum, strictly speaking extremum, if $\delta S = 0$, for which Euler-Lagrange equation is to be satisfied, i.e.,

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0 \tag{4.4.2}$$

where $f = x\sqrt{1 + y'^2}$, when compared to Eq. (4.2.2). Here

$$\frac{\partial f}{\partial y} = 0, \quad \frac{\partial f}{\partial y'} = \frac{xy'}{\sqrt{1 + y'^2}}$$

Substituting in Eq. (4.4.2), we have

$$\frac{d}{dx}\frac{xy'}{\sqrt{1 + y'^2}} = 0 \Rightarrow \frac{xy'}{\sqrt{1 + y'^2}} = a \tag{4.4.3}$$

where a is constant of integration. Squaring (4.4.3), we get

$$x^2y'^2 = a^2 + a^2y'^2 \Rightarrow y' = \frac{dy}{dx} = \frac{a}{\sqrt{x^2 - a^2}}$$

Therefore,

$$y = \int \frac{a}{\sqrt{x^2 - a^2}}\ dx = a\cosh^{-1}\frac{x}{a} + b \tag{4.4.4}$$

where b is another constant of integration. From (4.4.4) we have

$$\cosh^{-1}\frac{x}{a}=\frac{y-b}{a}\Rightarrow x=a\cosh\frac{y-b}{a} \tag{4.4.5}$$

which is the equation of a *catenary*. This is the equation of the curve for which the surface of revolution is minimum The two constants a and b can be determined by the condition that the curve (4.4.5) passes through (x_1,y_1) and (x_2,y_2) points.

4.5 Brachistochrone Problem

Example 4.5.1. A particle slides from rest at one point on a frictionless wire in a vertical plane to another point under the influence of the earth's gravitational field. If the particle travels in the shortest time, show that the path followed by it is a cycloid.

Solution. Let the shape of wire be in the form of a curve OA. The particle starts to travel from $O(0;0)$ from rest and moves to $A\,(x_1,y_1)$ under the influence of gravity on the frictionless wire.

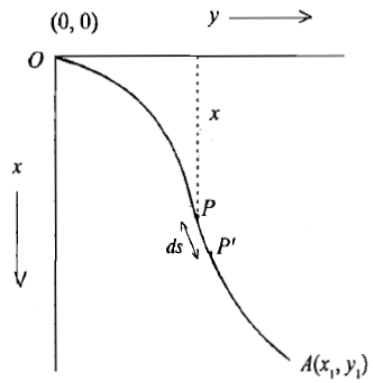


Figure 4.5.1: The Brachistochrone Problem

Let v be the speed at P. Then in moving $PP'=ds$ element, the time taken will be ds/v. Therefore, total time taken by the particle in moving from the higher point O to the lower point A is

$$t=\int_0^A\frac{ds}{v} \tag{4.5.1}$$

If the vertical distance of fall from O to P be x, then from the principle of conservation of energy $\frac{1}{2}mv^2=mgx\Rightarrow v=\sqrt{2gx}$. Therefore,

$$t=\int_0^A\frac{\sqrt{1+y'^2}dx}{\sqrt{2gx}}\qquad\left[ds=\sqrt{dx^2+dy^2}=dx\sqrt{1+y'^2}\right] \tag{4.5.2}$$

So that $f = \sqrt{\dfrac{1+y'^2}{2gx}}$ and for t to be minimum,

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0 \tag{4.5.3}$$

Here, $\dfrac{\partial f}{\partial y} = 0$ and $\dfrac{\partial f}{\partial y'} = \dfrac{y'}{\sqrt{2gx}\sqrt{1+y'^2}}$. Substituting in (4.5.3), we get

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{2gx}\sqrt{1+y'^2}}\right) = 0 \Rightarrow \frac{y'}{\sqrt{x}\sqrt{1+y'^2}} = C, \quad \text{constant}$$

$$\Rightarrow \frac{y'^2}{C^2} = x\left(1+y'^2\right) \Rightarrow y'^2\left(\frac{1}{C^2} - x\right) = x \Rightarrow y'^2 = \frac{x}{b-x} \quad (\text{where } b = 1/C^2, \text{ a constant}).$$

$$\Rightarrow \frac{dy}{dx} = \sqrt{\frac{x}{b-x}} \Rightarrow y = \int \sqrt{\frac{x}{b-x}}dx + C', \quad \text{another constant} \tag{4.5.4}$$

Let $x = b\sin^2\theta$, then $dx = 2b\sin\theta\ \cos\theta\ d\theta$. Therefore,

$$\begin{aligned} y &= \int \frac{\sin\theta}{\cos\theta} 2b\sin\theta\cos\theta\ d\theta + C' \\ &= b\int 2\sin^2\theta d\theta + C' = b\int (1-\cos 2\theta)d\theta + C' \\ &= b\left[\theta - \frac{\sin 2\theta}{2}\right] + C' = \frac{b}{2}[2\theta - \sin 2\theta] + C' \end{aligned}$$

Thus the parametric equations of the curve are

$$x = b\sin^2\theta = \frac{b}{2}(1-\cos 2\theta) \quad \text{and} \quad y = \frac{b}{2}(2\theta - \sin 2\theta) + C' \tag{4.5.5}$$

Since the curve passes through $(0,0)$ we have $C = 0$. Therefore,

$$x = \frac{b}{2}(1-\cos 2\theta) \text{ and } y = \frac{b}{2}(2\theta - \sin 2\theta) \tag{4.5.6}$$

Let $2\theta = \phi$ and $b/2 = a$. Then the parametric equations of the curve are

$$x = a(1-\cos\phi) \text{ and } y = a(\phi - \sin\phi) \tag{4.5.7}$$

This represents a cycloid [Fig. 4.5.2]. The constant a can be determined because the curve passes through the point $A\,(x_1, y_1)$.

Example 4.5.2. Apply variational principle to find the equation of one dimensional harmonic oscillator.

Solution. The Lagrangian L for one dimensional harmonic oscillator is

$$L = T - V = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2 \quad \text{or} \quad L = f(x,\dot{x},t) = \frac{1}{2}m\dot{x}^2 - \frac{1}{2}kx^2$$

According to variational principle $\int L\ dt$ or $\int f(x,\dot{x},t)\ dt$ is extremum. Euler-Lagrange's equation is

$$\frac{d}{dt}\left(\frac{\partial f}{\partial \dot{x}}\right) - \frac{\partial f}{\partial x} = 0$$

. Here, $\dfrac{\partial f}{\partial x} = -kx$, $\dfrac{\partial f}{\partial \dot{x}} = m\dot{x}$. Therefore, $m\ddot{x} + kx = 0$ which is the equation of motion for one-dimensional harmonic oscillator.

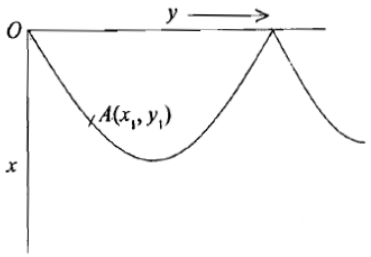


Figure 4.5.2: Cycloid

Example 4.5.3. Show that for a spherical surface, the geodesics are the great circles. (For a non-flat surface, the curves of extremal lengths are called geodesics.)

Solution. $ds^2 = a^2\left(d\theta^2 + \sin^2\theta d\phi^2\right)$ or $ds = ad\theta\sqrt{1+\sin^2\theta\phi'^2}$.
According to the variational principle,

$$\delta s = \delta\int ds = \delta\int a\, d\theta\sqrt{1+\sin^2\theta\phi'^2} = 0 \quad \text{or} \quad \delta\int_{\theta_1,\phi_1}^{\theta_2,\phi_2} d\theta\,\sqrt{1+\sin^2\theta\phi'^2} = 0$$

Here, $f = \sqrt{1+\sin^2\theta\phi'^2}$; $\therefore \dfrac{\partial f}{\partial\phi} = 0$ and $\dfrac{\partial f}{\partial\phi'} = \dfrac{\phi'\sin^2\theta}{\sqrt{1+\sin^2\theta\phi'^2}}$ Now,

$$\frac{\partial f}{\partial\phi} - \frac{d}{d\theta}\left(\frac{\partial f}{\partial\phi'}\right) = 0 \Rightarrow \frac{\phi'\sin^2\theta}{\sqrt{1+\sin^2\theta\phi'^2}} = C$$

$$\Rightarrow \phi' = \frac{C\,^2\theta}{\left(1-C^2-C^2\cot^2\phi\right)^{1/2}} = \frac{d\phi}{d\theta}, \quad \therefore \phi = \alpha - \sin^{-1}\left(C'\cot\theta\right)$$

where α and C' are constants and these may be fixed by limits θ_1, ϕ_1 and θ_2, ϕ_2

$$C'\cot\theta = \sin(\alpha-\phi) \Rightarrow C'r\cos\theta = r\sin(\alpha-\phi)\sin\theta$$
$$\Rightarrow C'r\cos\theta = \sin\alpha r\cos\phi\sin\theta - \cos\alpha r\sin\phi\sin\theta$$
$$\Rightarrow C'z = x\sin\alpha - y\cos\alpha$$

where we have transformed from spherical coordinates to Cartesian coordinates.

The above equation represents a plane passing through the origin $(0,0,0)$. This plane will cut the surface of the sphere in a great circle (whose centre is at the origin). This indicates that the shortest or longest distance between two points on the surface of the sphere is an arc of the circle with its centre at the origin.

4.6 Geodesics

A line is the shortest path between two points in a plane. We also wish to find shortest paths between pairs of points on other, more general, surfaces. To find these geodesics, we must minimize arc length.

The simplest case arises when the surface is a level set for one of the coordinates in a system of orthogonal curvilinear coordinates. The arc length can then be written using the scale factors of the coordinate system.

Consider, for example, two points, A and B, on a sphere of radius R centered at the origin. We wish to join A and B by the shortest, continuously differentiable curve lying on the sphere. We start by specifying position,

$$\vec{r}(x, y, z) = x\hat{i} + y\hat{j} + z\hat{k},$$

using the Cartesian coordinates x, y, and z and Cartesian basis vectors $\hat{i}, \hat{j}$, and $\hat{k}$. For points on the surface of a sphere, we now switch to the spherical coordinates r, θ, and ϕ (see Figure 4.6.1). Since

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta,$$

the position vector $\vec{r}$ now takes the form

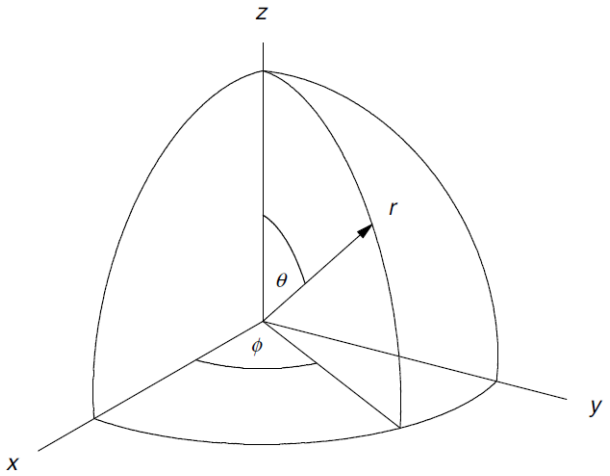


Figure 4.6.1: Spherical coordinates

$$\vec{r}(r, \theta, \phi) = r\sin\theta\cos\phi\hat{i} + r\sin\theta\sin\phi\hat{j} + r\cos\theta\hat{k}.$$

Since this position vector depends on r, θ, and ϕ,

$$d\vec{r} = \frac{\partial\vec{r}}{\partial r}dr + \frac{\partial\vec{r}}{\partial\theta}d\theta + \frac{\partial\vec{r}}{\partial\phi}d\phi.$$

The three partial derivatives on the right-hand side of this equation are vectors tangent to motions in the r, θ, and ϕ directions. Thus

$$d\vec{r} = h_r\, dr\, \hat{e}_r + h_\theta\, d\theta\, \hat{e}_\theta + h_\phi\, d\phi\, \hat{e}_\phi,$$

where $\hat{e}_r, \hat{e}_\theta$, and $\hat{e}_\phi$ are unit vectors in the r, θ, and ϕ directions and

$$h_r = \left\|\frac{\partial\vec{r}}{\partial r}\right\| = 1, \quad h_\theta = \left\|\frac{\partial\vec{r}}{\partial\theta}\right\| = r, \quad h_\phi = \left\|\frac{\partial\vec{r}}{\partial\phi}\right\| = r\sin\theta$$

are the scale factors for spherical coordinates.

The element of arc length in spherical coordinates is given by

$$\begin{aligned} ds &= \sqrt{d\vec{r}\cdot d\vec{r}} = \sqrt{h_r^2 dr^2 + h_\theta^2 d\theta^2 + h_\phi^2 d\phi^2} \\ &= \sqrt{dr^2 + r^2 d\theta^2 + r^2\sin^2\theta d\phi^2}. \end{aligned}$$

For a sphere of radius $r = R$, this element reduces to

$$ds = R\sqrt{d\theta^2 + \sin^2\theta d\phi^2}.$$

If we assume that $\phi = \phi(\theta)$, finding the curve that minimizes the arc length between the points $A = (\theta_a, \phi_a)$ and $B = (\theta_b, \phi_b)$ simplifies to finding the function $\phi(\theta)$ that minimizes the integral

$$s = \int_A^B ds = R\int_{\theta_A}^{\theta_B} \sqrt{1 + \sin^2\theta(d\phi/d\theta)^2}d\theta$$

subject to the boundary conditions

$$\phi(\theta_a) = \phi_a, \quad \phi(\theta_b) = \phi_b.$$

Unfortunately, we cannot expect every interesting surface to be the level set for some common coordinate. We may, however, hope to represent our surface parametrically. We may prescribe the x, y, and z coordinates of points on the surface using the parameters u and v and write our surface in the vector form

$$\vec{r}(u, v) = x(u, v)\hat{i} + y(u, v)\hat{j} + z(u, v)\hat{k}.$$

We can now specify a curve on this surface by prescribing u and v in terms of a single parameter - call it t - so that

$$u = u(t), \quad v = v(t).$$

The vector

$$\dot{\vec{r}} \equiv \frac{d\vec{r}}{dt} = \frac{\partial\vec{r}}{\partial u}\dot{u} + \frac{\partial\vec{r}}{\partial v}\dot{v}$$

is tangent to both the curve and the surface. We find the square of the distance between two points on a curve by integrating

$$ds^2 = d\vec{r}\cdot d\vec{r} = \left(\frac{\partial\vec{r}}{\partial u}du + \frac{\partial\vec{r}}{\partial v}dv\right)\cdot\left(\frac{\partial\vec{r}}{\partial u}du + \frac{\partial\vec{r}}{\partial v}dv\right) \tag{4.6.1}$$

along the curve. Equation (4.6.1) is often written

$$ds^2 = E\,du^2 + 2\,F\,du\,dv + G\,dv^2, \tag{4.6.2}$$

where

$$E = \frac{\partial\vec{r}}{\partial u}\cdot\frac{\partial\vec{r}}{\partial u}, \quad F = \frac{\partial\vec{r}}{\partial u}\cdot\frac{\partial\vec{r}}{\partial v}, \quad G = \frac{\partial\vec{r}}{\partial v}\cdot\frac{\partial\vec{r}}{\partial v}$$

The right-hand side of equation (4.6.2) is called the first fundamental form of the surface. The coefficients $E(u, v), F(u, v)$, and $G(u, v)$ have many names. They are sometimes called first-order fundamental magnitudes or quantities. Other times, they are simply called the coefficients of the first fundamental form.

The distance between the two points $A = (u_a, v_a)$ and $B = (u_b, v_b)$ on the curve $u = u(t), v = v(t)$ may now be written

$$s = \int_{t_a}^{t_b} \sqrt{E\left(\frac{du}{dt}\right)^2 + 2F\frac{du}{dt}\frac{dv}{dt} + G\left(\frac{dv}{dt}\right)^2}dt,$$

with

$$u(t_a) = u_a, \quad v(t_a) = v_a, \quad u(t_b) = u_b, v(t_b) = v_b.$$

In this formulation, we have two dependent variables, $u(t)$ and $v(t)$, and one independent variable, t. If v can be written as a function of u, $v = v(u)$, we can instead rewrite our integral as

$$s = \int_{u_a}^{u_b} \sqrt{E + 2F\left(\frac{dv}{du}\right) + G\left(\frac{dv}{du}\right)^2}\, du$$

with

$$v\left(u_a\right) = v_a, \quad v\left(u_b\right) = v_b.$$

This is now a problem with one dependent variable and one independent variable.

Illustration: To make all this concrete, let us take, as an example, the pseudosphere (see Figure 4.6.2), half of the surface of revolution generated by rotating a tractrix about its asymptote. If the asymptote is the z-axis,

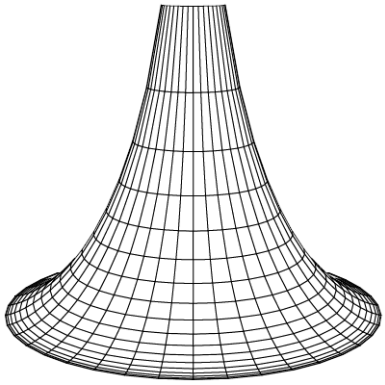


Figure 4.6.2: Pseudosphere

we can write the equation for a pseudosphere, parametrically, as

$$\vec{r}(u, v) = a \sin u \cos v\hat{i} + a \sin u \sin v + a\left(\cos u + \ln \tan \frac{u}{2}\right)\hat{k}.$$

Since

$$\vec{r}_u = \frac{\partial \vec{r}}{\partial u} = (a \cos u \cos v,\ a \cos u \sin v,\ -a \sin u + a \csc u)$$

and

$$\vec{r}_v = \frac{\partial \vec{r}}{\partial v} = (-a \sin u \sin v,\ a \sin u \cos v, 0),$$

the first-order fundamental quantities reduce to

$$\begin{aligned} E &= \vec{r}_u \cdot \vec{r}_u = a^2 \cot^2 u \\ F &= \vec{r}_u \cdot \vec{r}_v = 0 \\ G &= \vec{r}_v \cdot \vec{r}_v = a^2 \sin^2 u \end{aligned}$$

To determine a geodesic on the pseudosphere, we must thus find a curve, $u = u(t)$ and $v = v(t)$, that minimizes the arc-length integral

$$s = a \int_{t_a}^{t_b} \sqrt{\cot^2 u \dot{u}^2 + \sin^2 u \dot{v}^2}\, dt$$

subject to the boundary conditions

$$u(t_a) = u_a, \quad v(t_a) = v_a, \quad u(t_b) = u_b, \quad v(t_b) = v_b.$$

Alternatively, we may look for a curve, $v = v(u)$, that minimizes the integral

$$s = a \int_{u_a}^{u_b} \sqrt{\cot^2 u + \sin^2 u \left(\frac{dv}{du}\right)^2}\, du$$

subject to the boundary conditions

$$v(u_a) = v_a, \quad v(u_b) = v_b.$$

John Bernoulli (1697) posed the problem of finding geodesics on convex surfaces. In 1698, he remarked, in a letter to Leibniz, that geodesics always have osculating planes that cut the surface at right angles. (An osculating plane is the plane that passes through three nearby points on a curve as two of these points approach the third point.) This geometric property is frequently used as the definition of a geodesic curve, irrespective of whether the curve actually minimizes arc length. Later, Euler (1732) derived differential equations for geodesics on surfaces using the calculus of variations. This was Euler's earliest known use of the calculus of variations.

Exercise 4.6.1. 1. Show that for a function $f = f(y_1, y_2, \ldots, y_n, y_1', y_2', \ldots, y_n', x)$, the integral

$$I = \int_{x_1}^{x_2} f\, dx$$

will be extremum, if

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y_k'}\right) - \frac{\partial f}{\partial y_k} = 0$$

.

2. What do you mean by variational principle? Prove that the equation of curve for which surface area of revolution is minimum, is a catenary $x = a\cosh(y-b)/a$ where a and b are constants.

3. Use the variational principle to show that the shortest distance between two points in space is a straight line joining them.

4. Apply the vaiational principle to deduce the equation for stable equilibrium configuration of a uniform heavy flexible string fixed between two points $A(x_1, y_1)$ and $B(x_2, y_2$ in the constant gravity field of the earth.

5. A curve AB, having end points $A(x_1, y_1)$ and $B(x_2, y_2)$, is revolved about X- axis so that the area of the surface of revolution is a minimum. Show that $S = 2\pi \int_{x_1}^{x_2} y\sqrt{1+y'^2}\, dx$. Obtain the differential equation of the curve and prove that the curve represents a catenary.

6. Determine the first fundamental form for the following surfaces.
(i) the helicoid $x = u\cos v,\ y = u\sin v,\ z = av$;
(ii) the catenoid $x = a\cosh\dfrac{u}{a}\cos v,\ y = a\cosh\dfrac{u}{a}\sin v,\ z = u$;
(iii) the hyperbolic paraboloid $x = a(u+v),\ y = b(u-v),\ z = uv$.

Unit 5

Course Structure

- Hamilton's principle.
- Lagrange's undetermined multipliers.
- Hamilton's equations of motion.

5.1 Hamilton's Principle

This principle states that for a conservative holonomic system, its motion from time t_1 to time t_2 is such that the line integral (known as *action* or *action integral*)

$$S = \int_{t_1}^{t_2} L \, dt \tag{5.1.1}$$

with $L = T - V$ has stationary (extremum) value for the correct path of the motion.

The quantity S is called as *Hamilton's principal function*. The principle may be expressed as

$$\delta \int_{l_1}^{t_2} L \, dt = 0 \tag{5.1.2}$$

where δ is the variation symbol.

5.1.1 Lagrange's equation from Hamilton's principle

The Lagrangian L is a function of generalized coordinates q_k's and generalized velocities $\dot{q}_k$'s and time t, i.e.,

$$L = L\left(q_1, tq_2, \ldots, q_k, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_k, \ldots, \dot{q}_n, t\right)$$

If the Lagrangian does not depend on time t explicitly, then the variation δL can be written as

$$\delta L = \sum_{k=1}^{n} \frac{\partial L}{\partial q_k} \delta q_k + \sum_{k=1}^{n} \frac{\partial L}{\partial \dot{q}_k} \delta \dot{q}_k \tag{5.1.3}$$

Integrating both sides from $t = t_1$ to $t = t_2$, we get

$$\int_{t_1}^{t_2} \delta L \, dt = \int_{t_1}^{t_2} \sum_k \frac{\partial L}{\partial q_k} \, \delta q_k \, dt + \int_{t_1}^{t_2} \sum_k \frac{\partial L}{\partial \dot{q}_k} \delta \dot{q}_k \, dt$$

But in view of the Hamilton's principle

$$\delta \int_{l_1}^{t_2} L \, dt = 0$$

Therefore,

$$\int_{t_1}^{t_2} \sum_k \frac{\partial L}{\partial q_k} \delta q_k \, dt + \int_{t_1}^{t_2} \sum_k \frac{\partial L}{\partial \dot{q}_k} \delta \dot{q}_k \, dt = 0 \tag{5.1.4}$$

where $\delta \dot{q}_k = \dfrac{d}{dt} (\delta q_k)$. Integrating by parts the second term on the left hand side of Eq. (5.1.4), we get

$$\int_{t_1}^{t_2} \sum_k \frac{\partial L}{\partial \dot{q}_k} \delta \dot{q}_k \, dt = \sum_k \left[\frac{\partial L}{\partial \dot{q}_k} \delta q_k \right]_{t_1}^{t_2} - \int_{t_1}^{t_2} \sum_k \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_k} \right) \delta q_k \, dt \tag{5.1.5}$$

At the end points of the path at the times t_1 and t_2, the coordinates must have definite values $q_k (t_1)$ and

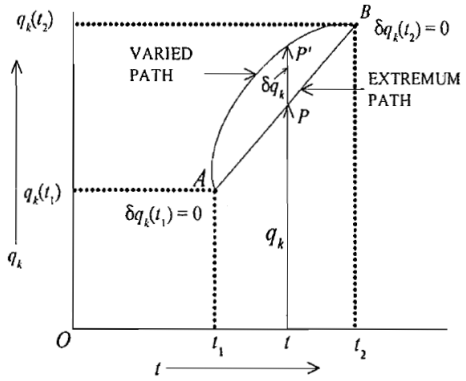


Figure 5.1.1: δ - variation - extremum path

$q_k (t_2)$ respectively, i.e., $\delta q_k (t_1) = \delta q_k (t_2) = 0$ (Fig. 5.1.1) and hence

$$\sum_k \left[\frac{\partial L}{\partial q_k} \delta q_k \right]_{t_1}^{t_2} = 0$$

Therefore, Eq. (5.1.4) takes the form

$$\int_{t_1}^{t_2} \sum_k \frac{\partial L}{\partial q_k} \delta q_k \, dt - \int_{t_1}^{t_2} \sum_k \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_k} \right) \delta q_k \, dt = 0$$

$$\Rightarrow \sum_k \int_{t_1}^{t_2} \left[\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_k} \right) - \frac{\partial L}{\partial q_k} \right] \delta q_k \, dt = 0 \tag{5.1.6}$$

For holonomic system, the generalized cooordinates δq_k are independent of each other. Therfore, the coefficien of each δq_k must vanish, i.e.,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0 \tag{5.1.7}$$

where $k = 1, 2, \ldots, n$ are the generalized coordinates. Eq. (5.1.7) are the *Lagrange's equations of motion*.

5.2 Lagrange's Equations of Motion for Non-holonomic systems

(Lagrange's method of undetermined multipliers)

In the derivation of Lagrange's equations from D'Alembert's principle or Hamilton's principle, we need the requirement of holonomic constraints at the final step, when the variations δq_k are considered to be independent of each other. In case of non-holonomic systems, the generalized coordinates are not independent of each other. However, we can treat certain types of non-holonomic systems for which the equations of constraint can be put in the form

$$\sum_k a_{lk}\, dq_k + a_{lt}\, dt = 0 \tag{5.2.1}$$

These equations of constraints connect the differentials dq_k's by linear relations. For each l, there is one equation and we assume that there are m such equations for $l = 1, 2, \ldots, m$.

In case of δ-variation, the virtual displacements δq_k are taken at constant times and hence the m equations of constraints, consistent for virtual displacements, are

$$\sum_k a_{lk}\delta q_k = 0 \tag{5.2.2}$$

Eq. (5.2.2) now can be used to reduce the number of virtual displacements to independent ones. The procedure applied for this purpose is called *Lagrange's method of undetermined multipliers*.

If Eq. (5.2.2) is valid, then the multiplication of this equation by λ_l, an undetermined quantity, yields

$$\lambda_l \sum_k a_{lk}\, \delta q_k = 0 \quad \text{or} \quad \sum_k \lambda_l a_{lk}\, \delta q_k = 0 \tag{5.2.3}$$

where λ_l $(1, 2, \ldots, m)$ are undetermined quantities and they are functions in general of the coordinates and time. Summing Eq. (5.2.3) over l and then integrating the sum with respect to time from $t = t_1$ to $t = t_2$, we get

$$\int_{t_1}^{t_2} \sum_{k,l} \lambda_l\, a_{lk}\, \delta q_k\, dt = 0 \tag{5.2.4}$$

We assume the Hamilton's principle $\delta \int_{t_1}^{t_2} L\, dt = 0$ to hold for the non-holonomic system. This implies that

$$\int_{t_1}^{t_2} \sum_k \left[\frac{\partial L}{\partial q_k} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right)\right] \delta q_k\, dt = 0 \tag{5.2.5}$$

Adding (5.2.4) and (5.2.5), we obtain

$$\int_{t_1}^{t_2} \sum_{k=1}^{n} \left[\frac{\partial L}{\partial q_k} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) + \sum_l \lambda_l a_{lk} \right] \delta q_k \, dt = 0 \tag{5.2.6}$$

Still all δq_k's $(k = 1, 2, \ldots, n)$ are not independent of each other. First $n - m$ of these δq_k's may be chosen independently and the last m of these δq_k's are then fixed by the Eq. (5.2.2).

Till now the values of λ_l, have not been specified. We choose the λ_l's such that

$$\frac{\partial L}{\partial q_k} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) + \sum_{l=1}^{m} \lambda_k a_{lk} = 0 \tag{5.2.7}$$

which are $n - m$ equations for $k = 1, 2, \ldots, n - m$. Adding Eqs. (5.2.7) (45) and (47), we get the complete set of the Lagrange's equations for the non-holonomic system, i.e.,

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}_k}\right] - \frac{\partial L}{\partial q_k} = \sum_{l=1}^{m} \lambda_l a_{lk} \tag{5.2.8}$$

where $k = 1, 2, \ldots n$. Eq. (5.2.8) gives us n equations, but there are $n + m$ unknowns, n coordinates q_k and m Lagrange's multipliers. The remaining m unknown q_k's are determined from m equations of constraints (5.2.1), written in the following form of m first-order differential equations:

$$\sum_k a_{lk}\dot{q}_k + a_{lt} = 0 \tag{5.2.9}$$

5.3 Physical Significance of Lagrange's Multipliers λ_i

Suppose we remove the constraints on the system, but apply external forces G_k so that the motion of the system remains unchanged. Now, the equations of motion are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = G_k \tag{5.3.1}$$

Since the applied force are equal to the forces of constraints, Eqs. (5.2.8) and (5.3.1) must be identical, resulting

$$G_k = \sum_{l=1}^{n} \lambda_l a_{lk} \tag{5.3.2}$$

Thus the generalized forces of constraints G_k have been identified to $\sum \lambda_l a_{lk}$ We observe that in such problems, we need not to eliminate the forces of the constraints but the procedure itself determines these forces by Eq. (5.3.2).

Eq. (5.2.1) does not represent the most general type of non-holonomic constraints because it does not include equations of constraints in the form of inequalities. However, it includes holonomic constraints. Equation representing holonomic constraints is given by

$$f\left(q_\ell, q_2, \ldots, q_n, t\right) = 0 \tag{5.3.3}$$

so that

$$\sum_{k=1}^{n} \frac{\partial f}{\partial q_k}\, dq_k + \frac{\partial f}{\partial t}\, dt = 0 \tag{5.3.4}$$

This is identical in form to Eq. (5.2.1) with the coefficients a_{lk} and a_{lt}, given by

$$a_{lk} = \frac{\partial f}{\partial q_k} \quad \text{and} \quad a_{lt} = \frac{\partial f}{\partial t} \tag{5.3.5}$$

Thus one can use Lagrange's method of undetermined multipliers for holonomic constraints when it is not easy to reduce all the q_k's to independent coordinates or we may be interested in knowing the force of constraints.

Example 5.3.1. *Simple pendulum:* Find the equation of motion and force of constraint in case of simple pendulum by using Lagrange's method of undetermined multipliers.

Solution. Referring Fig. 5.3.1, the Lagrangian L is given by

$$= \frac{1}{2}mr^2\dot{\theta}^2 + mgr\cos\theta \tag{5.3.6}$$

where $V = -mgr\cos\theta$ with respect to position S. The equation of constraint is

$$r = l \quad \text{or} \quad dr = 0 \tag{5.3.7}$$

Here there is only one constraint equation, hence only one Lagrange's multiplier λ will be needed. There are

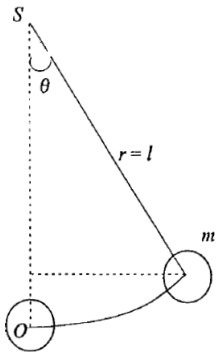


Figure 5.3.1: Simple pendulum

two coordinates r and θ and the general constraint equation will be

$$a_r\, dr + a_\theta\, d\theta = 0 \tag{5.3.8}$$

Therefore $a_r = 1$ and $a_\theta = 0$.
Equation of motion are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}}\right) - \frac{\partial L}{\partial r} = \lambda a_r \tag{5.3.9}$$

$$\text{and} \quad \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}}\right) - \frac{\partial L}{\partial \theta} = \lambda a_\theta \tag{5.3.10}$$

Here, $\frac{\partial L}{\partial \dot{r}} = 0$, $\frac{\partial L}{\partial r} = mr\dot{\theta}^2 + mg\cos\theta$, $\frac{\partial L}{\partial \dot{\theta}} = mr^2\dot{\theta}$, $\frac{\partial L}{\partial \theta} = -mgr\sin\theta$. Thus,

$$-mr\theta^2 - mg\cos\theta = \lambda \tag{5.3.11}$$

$$mr^2\ddot{\theta} + mgr\sin\theta = 0 \tag{5.3.12}$$

where $\dot{r} = 0$ from (5.3.7). Using $r = l$, (the constraint equation), equation of motion of simple pendulum is given by Eq. (5.3.12), i.e.,

$$l\ddot{\theta} + g\sin\theta = 0$$

For small θ, $\sin\theta = \theta$ and hence

$$\ddot{\theta} + \frac{g}{l}\theta = 0 \tag{5.3.13}$$

The force of constraint is

$$\lambda = -ml\dot{\theta}^2 - mg\cos\theta \tag{5.3.14}$$

which gives the force of constraint, i,e. tension $F = ml\dot{\theta}^2 + mg\cos\theta$ in magnitude.

5.4 Hamiltonian Dynamics

In the Lagrangian formulation, the equations of motion are in the form of a set of second order differential equations. An alternative formulation, given by Hamilton and known as the Hamiltonian dynamics, makes use of the generalized momenta $p_1, p_2, \ldots, p_n$ in place of the generalized velocities $\dot{q}_1, \dot{q}_2, \ldots, \dot{q}_n$ used in the Lagrangian formulation. In the Hamiltonian formulation, two sets of first order differential equations are used instead of a set of second order differential equations. Both the formulations are equivalent, but the Hamiltonian formulation is more fundamental to the foundations of statistical and quantum mechanics.

5.4.1 Generalized Momentum and Cyclic Coordinates

In order to define the generalized momentum, we take a simple example of a single particle, moving with velocity $\dot{x}$ along X-axis. The kinetic energy of the particle is

$$T = \frac{1}{2}m\dot{x}^2 \tag{5.4.1}$$

The derivative of T with respect to $\dot{x}$ i.e., $\dfrac{\partial T}{\partial \dot{x}} = m\dot{x} = p$ defines the momentum. If V is not a function of the velocity $\dot{x}$, i.e., $V = V(x)$ and $\dfrac{\partial V}{\partial \dot{x}} = 0$, then the momentum p can be written also as

$$p = \frac{\partial}{\partial \dot{x}}(T - V) \quad \text{or} \quad p = \frac{\partial L}{\partial \dot{x}} \tag{5.4.2}$$

Similarly for a system described by a set of generalized coordinates q_k's and generalized velocities $\dot{q}_k$'s, we define the generalized momentum corresponding to the generalized coordinate q_k as

$$p_k = \frac{\partial L}{\partial \dot{q}_k} \tag{5.4.3}$$

This is also called *conjugate momentum* (conjugate to the coordinate q_k) or *canonical momentum*. For a conservative system, the Lagrange's equations are given by

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}_k}\right] - \frac{\partial L}{\partial q_k} = 0 \tag{5.4.4}$$

Substituting for $\dfrac{\partial L}{\partial \dot{q}_k} = p_k$, we get

$$\frac{dp_k}{dt} - \frac{\partial L}{\partial q_k} = 0 \quad \text{or} \quad \dot{p}_k = \frac{\partial L}{\partial q_k} \tag{5.4.5}$$

Now, suppose in the expression for Lagrangian L of a system, a certain coordinate q_k does not appear explicitly. Then

$$\frac{\partial L}{\partial q_k} = 0 \tag{5.4.6}$$

This means from Eq. (5.4.5) that

$$\dot{p}_k = \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}_k}\right] = 0 \tag{5.4.7}$$

and hence on integration, we get

$$p_k = \frac{\partial L}{\partial \dot{q}_k} = \text{ a constant} \tag{5.4.8}$$

Thus whenever the Lagrangian function does not contain a coordinate q_k explicitly, the generalized momentum p_k is a constant of motion. The coordinate q_k is called *cyclic* or *ignorable*. In other words, the generalized momentum associated with an ignorable coordinate is a constant of motion for the system.

5.4.2 Hamiltonian Function H and Conservation of Energy

In the Lagrangian formulation one may expect the deduction of the theorem of conservation of the total energy for a system where the potential energy is a function of position only. In fact we shall see, as discussed below, the theorem of conservation of total energy is a special case of a more general conservation theorem.

Consider a general Lagrangian L of a system given by

$$L = L\left(q_1, q_2, \ldots, q_k, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_k, \ldots, \dot{q}_n, t\right)$$

We denote it for our convenience by

$$L = L\left(q_k, \dot{q}_k, t\right)$$

The total time derivative of L is

$$\frac{dL}{dt} = \sum_k \frac{\partial L}{\partial q_k}\frac{dq_k}{dt} + \sum_k \frac{\partial L}{\partial \dot{q}_k}\frac{d\dot{q}_k}{dt} + \frac{\partial L}{\partial t} \tag{5.4.9}$$

From Lagrangian equations, we have

$$\frac{\partial L}{\partial q_k} = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right)$$

Substituting for $\partial L/\partial q_k$ in Eq. (5.4.9), we get

$$\begin{aligned}
\frac{dL}{dt} &= \sum_k \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right)\dot{q}_k + \sum_k \frac{\partial L}{\partial \dot{q}_k}\frac{d\dot{q}_k}{dt} + \frac{\partial L}{\partial t} \\
\Rightarrow \frac{dL}{dt} &= \sum_k \frac{d}{dt}\left(\dot{q}_k \frac{\partial L}{\partial \dot{q}_k}\right) + \frac{\partial L}{\partial t} \\
\Rightarrow \frac{d}{dt}&\left(\sum_k \dot{q}_k \frac{\partial L}{\partial \dot{q}_k} - L\right) = -\frac{\partial L}{\partial t}
\end{aligned} \tag{5.4.10}$$

The quantity in the bracket is sometimes called the *energy function* and is denoted by h:

$$h\left(q_1, q_2, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \ldots, \dot{q}_n, t\right) = \sum_k \dot{q}_k \frac{\partial L}{\partial \dot{q}_k} - L \tag{5.4.11}$$

Thus from Eq. (5.4.10) the total time derivative of h is

$$\frac{dh}{dt} = -\frac{\partial L}{\partial t} \tag{5.4.12}$$

If the Lagrangian L does not depend on time t explicitly, then $\partial L/\partial t = 0$, so that

$$\frac{dh}{dt} = 0 \text{ i.e., } h = \text{ Constant} \tag{5.4.13}$$

Thus when the lagrangian is not explicit function of time, the energy function is the constant of motion. It is one of the first integrals of the motion and is called *Jacobi's integral*. But from Eq. (5.4.8) $\partial L/\partial \dot{q}_k = p_k$, hence Eq. (5.4.10) can be written as

$$\frac{d}{dt}\left(\sum_k p_k \dot{q}_k - L\right) = -\frac{\partial L}{\partial t} \tag{5.4.14}$$

The quantity in the bracket is called the *Hamiltonian function* H, i.e.,

$$H = \sum_k p_k \dot{q}_k - L \tag{5.4.15}$$

In general, the Hamiltonian function H is the function of generalized momenta p_k, generalized coordinates q_k and time t i.e.,

$$H = H\left(p_1, p_2, \ldots, p_k, .., p_n, q_1, q_2, \ldots, q_k, \ldots, q_n, t\right) \tag{5.4.16}$$

$$\text{or } H = H\left(p_k, q_k, t\right) \tag{5.4.17}$$

It is to be seen that the energy function h is identical in value with the Hamiltonian H. It is given a different name and symbol because h is a function of $q_k, \dot{q}_k$ and t, while H that of q_k, p_k and t.

If t does not appear in the Lagrangian L explicitly, then $\partial L/\partial t = 0$ and Eqs. (5.4.14) and (5.4.15) give

$$\frac{dH}{dt} = 0 \quad \text{or} \quad H = \sum_k p_k \dot{q}_k - L = \text{ constant} \tag{5.4.18}$$

Thus, if the time t does not appear in the Lagrangian L explicitly, we see that the Hamiltonian H is constant in time, i.e., conserved. This is the conservation theorem for the Hamiltonian of the system. Under special circumstances, the Hamiltonian H is equal to the total energy E of the system. In fact, this is the case in most of the physical problems.

Conservation of Energy-Physical Significance

The Hamiltonian takes a special form, if the system is conservative i.e., the potential energy V is independent of velocity coordinates $\dot{q}_k$ and the transformation equations for coordinates do not contain time explicitly i.e.,

$$r_i = r_i\left(q_1, q_2, \ldots, q_k, \ldots, q_n\right).$$

For a conservative system $\partial V \partial \dot{q}_k = 0$. From Eq. (5.4.8), we have

$$p_k = \frac{\partial L}{\partial \dot{q}_k} = \frac{\partial}{\partial \dot{q}_k}(T - V) = \frac{\partial T}{\partial \dot{q}_k}$$

So that Eq. (5.4.18) is

$$H = \sum_k p_k \dot{q}_k - L = \sum_k \frac{\partial T}{\partial \dot{q}_k} \dot{q}_k - L \tag{5.4.19}$$

If r_i does not depend on time t explicitly, then the kinetic energy T is a homogeneous quadratic function. It is easy to show that

$$\sum_k \frac{\partial T}{\partial \dot{q}_k} \dot{q}_k = 2T \tag{5.4.20}$$

In fact, for a natural conservative system neither T nor V contains any explicit time dependence (i.e., the Lagrangian does not depend on time explicitly) and T is a homogeneous quadratic function of the time derivatives $\dot{q}_k$. Hence from Eq. (5.4.19) and Eq. (5.4.20),

$$H = 2T - L = 2T - (T - V) \Rightarrow H = T + V = E, \quad \text{constant} \tag{5.4.21}$$

Thus the Hamiltonian H represents the total energy of the system E and is conserved, provided the system is conservative and T is a homogeneous quadratic function.

5.4.3 Hamilton's Equation

The Hamiltonian, in general, is a function of generalized coordinates q_k, generalized momenta p_k and time t, i.e.,

$$H = H\left(q_1, q_2, \ldots, q_k, \ldots, q_n, p_1, p_2, \ldots, p_k, \ldots, p_n t\right)$$

We may write the differential dH as

$$dH = \sum_k \frac{\partial H}{\partial q_k} dq_k + \sum_k \frac{\partial H}{\partial p_k} dp_k + \frac{\partial H}{\partial t} dt \tag{5.4.22}$$

But as defined in Eq. (5.4.15), $H = \sum_k p_k \dot{q}_k - L$ and hence

$$dH = \sum_k \dot{q}_k dp_k + \sum_k p_k \, d\dot{q}_k - dL \tag{5.4.23}$$

Also, $L = L\left(q_1, q_2, \ldots, q_k, \ldots, q_n, \dot{q}_1, \dot{q}_2, \ldots, \dot{q}_k, \ldots, \dot{q}_n, t\right)$. Therefore,

$$dL = \sum_k \frac{\partial L}{\partial q_k} dq_k + \sum_k \frac{\partial L}{\partial \dot{q}_k} d\dot{q}_k + \frac{\partial L}{\partial t} dt$$

But $\dot{p}_k = \dfrac{\partial L}{\partial q_k}$ [Eq. (5.4.5)] and $p_k = \dfrac{\partial L}{\partial \dot{q}_k}$ [Eq. (5.4.3)]. Therefore,

$$dL = \sum_k \dot{p}_k \, dq_k + \sum_k p_k \, d\dot{q}_k + \frac{\partial L}{\partial t} dt \tag{5.4.24}$$

Substituting for dL from Eq. (5.4.24) in Eq. (5.4.23), we get

$$dH = \sum_k \dot{q}_k \, dp_k - \sum_k \dot{p}_k dq_k - \frac{\partial L}{\partial t} dt \tag{5.4.25}$$

Comparing the coefficients of dp_k, dq_k and dt in Eqs. (5.4.22) and (5.4.25), we obtain

$$\dot{q}_k = \frac{\partial H}{\partial p_k} \tag{5.4.26}$$

$$-\dot{p}_k = \frac{\partial H}{\partial q_k} \tag{5.4.27}$$

$$-\frac{\partial L}{\partial t} = \frac{\partial H}{\partial t} \tag{5.4.28}$$

Eqs. (5.4.26) and (5.4.27) are known as *Hamilton's equations* or *Hamilton's canonical equations of motion*.

It is clear from Eq. (5.4.27) that if any coordinate q_k is cyclic, i.e., not contained in H, then

$$\frac{\partial H}{\partial q_k} = 0 \quad \text{or} \quad \dot{p}_k = 0 \quad \text{or} \quad p_k = \text{ constant in time} \tag{5.4.29}$$

Thus for any cyclic coordinate, corresponding conjugate momentum is a constant of motion. Further from Eq. (5.4.22), we have

$$\frac{dH}{dt} = \sum_k \frac{\partial H}{\partial q_k}\dot{q}_k + \sum_k \frac{\partial H}{\partial p_k}\dot{p}_k + \frac{\partial H}{\partial t} \tag{5.4.30}$$

Substituting for $\dot{q}_k$ and $\dot{p}_k$ from Eqs. (5.4.26) and (5.4.27) in Eq. (5.4.30), we get

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} = -\frac{\partial L}{\partial t} \tag{5.4.31}$$

If the Lagrangian L and hence H does not depend on time t explicitly; then $\partial L/\partial t = -\partial H/\partial t = 0$ and hence

$$\frac{dH}{dt} = 0 \quad \text{or} \quad H = \text{ constant .} \tag{5.4.32}$$

We are mainly interested in the conservative systems for which $H = T + V = E$ is a constant of motion, as discussed earlier.

Example 5.4.1. Write the Hamiltonian for a simple pendulum and deduce its equation of motion.

Solution. We know for a simple pendulum the kinetic energy $T = \frac{1}{2}ml^2\dot{\theta}^2$, Potential energy $V = mgl(1 - \cos\theta)$. Therefore,

$$\text{Lagrangian} \quad L = T - V = \frac{1}{2}ml^2\dot{\theta}^2 - mgl(1 - \cos\theta)$$

Hence, $p_\theta = \dfrac{\partial L}{\partial \dot{\theta}} = ml^2\dot{\theta}$, Now, Hamiltonian

$$\begin{aligned} H &= \sum_k p_k\dot{q}_k - L = p_\theta\dot{\theta} - \left[\frac{1}{2}ml^2\dot{\theta}^2 - mgl(1 - \cos\theta)\right] \\ &= ml^2\dot{\theta}^2 - \frac{1}{2}ml^2\dot{\theta}^2 - mgl(1 - \cos\theta) \\ &= \frac{1}{2}ml^2\dot{\theta}^2 + mgl(1 - \cos\theta) = T + V = \text{ Total energy} \\ &= \frac{1}{2}ml^2\left[\frac{p_\theta}{ml^2}\right]^2 + mgl(1 - \cos\theta) = \frac{p_\theta^2}{2ml^2} + mgl(1 - \cos\theta) \end{aligned}$$

Hence Hamilton's equations are

$$\dot{\theta} = \frac{\partial H}{\partial p_\theta} = \frac{p_\theta}{ml^2} \quad \text{and} \quad -\dot{p}_\theta = \frac{\partial H}{\partial \theta} = mgl\sin\theta$$

Thus

$$\dot{p}_\theta = ml^2\ddot{\theta} = -mgl\sin\theta \Rightarrow l\ddot{\theta} + g\sin\theta = 0 \Rightarrow \ddot{\theta} + \frac{g}{l}\theta = 0 \quad \text{for small} \quad \theta \ (\because \sin\theta \approx \theta)$$

This is the equation of motion of simple pendulum.

Example 5.4.2. Find the Hamiltonian corresponding to Lagrangian $L = a\dot{x} + b\dot{y} - kxy$.

Solution. We know $H = \sum_k p_k\dot{q}_k - L,\ p_k = \frac{\partial L}{\partial \dot{q}_k}$. Here $L = a\dot{x} + b\dot{y} - kxy$.

Now $p_x = \frac{\partial L}{\partial \dot{x}} = 2a\dot{x},\ p_y = \frac{\partial L}{\partial \dot{y}} = 2b\dot{y}$.

Therefore,

$$H = p_x\dot{x} + p_y\dot{y} - L = 2a\dot{x}^2 + 2b\dot{y}^2 - a\dot{x}^2 - b\dot{y}^2 + kxy = a\dot{x}^2 + b\dot{y}^2 + kxy$$

As $\dot{x} = \frac{p_x}{2a}$ and $\dot{x} = \frac{p_y}{2b}$, we obtain

$$H = a\frac{p_x^2}{4a^2} + b\frac{p_y^2}{4b^2} + kxy = \frac{p_x^2}{4a} + \frac{p_y^2}{4b} + kxy$$

Example 5.4.3. Find the Lagrangian for the case when the Hamiltonian is $H(p,r) = \frac{p^2}{2m} - (\vec{a}\cdot\vec{p})$, $\vec{a} = a_x\hat{i} + a_y\hat{j} + a_z\hat{k}$ being a constant vector.

Solution. Given $H(p,r) = \frac{p^2}{2m} - (\vec{a}\cdot\vec{p})$. Now,

$$H(p,x,y,z) = \frac{p_x^2}{2m} + \frac{p_y^2}{2m} + \frac{p_z^2}{2m} - (a_xp_x + a_yp_y + a_zp_z) \tag{5.4.33}$$

Hamilton's equations are $\dot{q} = \frac{\partial H}{\partial p_k},\ -\dot{p}_k = \frac{\partial H}{\partial q_k}$.

In this problem, $\dot{x} = \frac{\partial H}{\partial p_k},\ \dot{y} = \frac{\partial H}{\partial p_y},\ \dot{z} = \frac{\partial H}{\partial p_z}$.

Using (5.4.33)

$$\dot{x} = \frac{p_x}{m} - a_x,\ \dot{y} = \frac{p_y}{m} - a_y,\ \dot{z} = \frac{p_z}{m} - a_z \tag{5.4.34}$$

But $= \sum p_k\dot{q}_k - L$. Hence

$$\frac{p_x^2}{2m} + \frac{p_y^2}{2m} + \frac{p_z^2}{2m} - (a_xp_x + a_yp_y + a_zp_z = p_x\dot{x} + p_y\dot{y} + p_z\dot{z} - L$$

$$\Rightarrow \frac{p_x^2}{2m} + \frac{p_y^2}{2m} + \frac{p_z^2}{2m} - (a_xp_x + a_yp_y + a_zp_z = p_x\left(\frac{p_x}{m} - a_x\right) + p_y\left(\frac{p_y}{m} - a_y\right) + p_z\left(\frac{p_z}{m} - a_z\right) - L$$

$$\Rightarrow \frac{p_x^2}{2m} + \frac{p_y^2}{2m} + \frac{p_z^2}{2m} = \frac{p_x^2}{2m} + \frac{p_y^2}{2m} + \frac{p_z^2}{2m} - L$$

Therefore, $L = \frac{p_x^2}{2m} + \frac{p_y^2}{2m} + \frac{p_z^2}{2m}$.
But form (5.4.34), $p_x = m(\dot{x} + a_x),\ p_y = m(\dot{y} + a_y),\ p_z = m(\dot{z} + a_z)$.
Hence

$$\begin{aligned} L &= \frac{1}{2}m\left[(\dot{x}+a_x)^2 + (\dot{y}+a_y)^2 + (\dot{z}+a_z)^2\right] \\ &= \frac{1}{2}m\left[(v_x+a_x)^2 + (v_y+a_y)^2 + (v_z+a_z)^2\right] \\ &= \frac{1}{2}m[\vec{v}+\vec{a}]^2 \end{aligned}$$

Example 5.4.4. Using Hamilton's equation of motion, show that the Hamiltonian $H = \frac{p^2}{2m}e^{-m} + \frac{1}{2}m\omega^2x^2e^{rt}$ leads to the equation of motion of a damped harmonic oscillator $\ddot{x} + r\dot{x} + \omega^2 x = 0$.

Solution. Equations of motion are $\dot{q} = \frac{\partial H}{\partial p}$ and $\dot{p} = -\frac{\partial H}{\partial q}$.

For $q = x$, $\dot{x} = \frac{\partial H}{\partial p}$ and $\dot{p} = -\frac{\partial H}{\partial x}$.

Here

$$\dot{x} = \frac{p}{m}e^{-rt} \quad \text{and} \quad \dot{p} = -m\omega^2 x e^{rt} \tag{5.4.35}$$

whence $p = m\dot{x}e^{rt}$ and $\dot{p} = m\ddot{x}e^{rt} + mr\dot{x}e^{rt}$.

Substituting for $\dot{p}$ in (5.4.35), we get

$$m\ddot{x}e^{rt} + mr\dot{x}e^{rt} = -m\omega^2 x e^{rt} \Rightarrow \ddot{x} + r\dot{x} + \omega^2 x = 0$$

which is the desired equation of damped harmonic oscillator.

Exercise 5.4.5.

1. Deduce the Hamiltonian function and equation of motion for a compound pendulum.
2. The Lagrangian for anharmonic oscillator is given by $L(x, \dot{x}) = \frac{1}{2}\dot{x} - \frac{1}{2}\omega^2x^2 - \alpha x^3$. Find the Hamiltonian.
3. The Hamiltonian of a system with generalized coordinate and momentum (q, p) is $H = p^2q^2$. Show that the solution of the Hamiltonian equation of motion is $p = Be^{-2At},\ q = \frac{A}{B}e^{2At}$, where A and B are constants.
4. A system is governed by the Hamiltonian $H = \frac{1}{2}(p_x - ay)^2 + \frac{1}{2}(p_y - bx)^2$ where a and b are constants and $p_x,\ p_y$ are momenta conjugate to x and y respectively. For what values of a and b will the quantities $(p_x - 3y)$ and $(p_y + 2x)$ be conserved?
5. A particle in two dimension is in a potential $V(x, y) = x + 2y$. Show that $p_y - 2p_x$ is a constant of motion.
6. The Hamiltonian for a system described by the generalized coordinate x and generalized momentum p is $H = ax^2p + \frac{p^2}{2(1+2\beta x)} + \frac{1}{2}\omega^2x^2$ where α, β and ω are constants. Find the corresponding Lagrangian.

Unit 6

Course Structure

- Canonical Transformations:
- Canonical coordinates and canonical transformations
- Poincare theorem.

6.1 Canonical Transformations

In several problems, we nay need to change one set of position and momentum coordinates into another set of position and momentum coordinates. Suppose that q_k and p_k are the old position and momentum coordinates and Q_k and P_k are the new ones. Let these coordinates be related by the following transformations:

$$\begin{aligned} P_k &= P_k\left(p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_n, t\right) \\ \text{and } \ Q_k &= Q_k\left(p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_n, t\right) \end{aligned} \tag{6.1.1}$$

Now, if there exists a Hamiltonian H' in the new coordinates such that

$$\dot{P}_k = -\frac{\partial H'}{\partial Q_k} \text{ and } \dot{Q}_k = \frac{\partial H'}{\partial P_k} \tag{6.1.2}$$

where $H' = \sum_{k=1}^{n} P_k \dot{Q}_k - L'$ and L' substituted in the Hamilton's principle $\delta \int L' dt = 0$ gives the correct equations of motion in terms of the new coordinates P_k and Q_k, then the transformations (6.1.1) are known as *canonical (or contact) transformations.*

6.2 Legendre Transformations

This is a mathematical technique used to change the basis from one set of coordinates to another. If $f(x, y)$ is a function of two variables x and y, then the differential of this function can be written as

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad \text{or} \quad df = u dx + v dy \tag{6.2.1}$$

where $u = \partial f/\partial x$ and $v = \partial f/\partial y$ Now, we want to change the basis from (x, y) to (u, v) so that u is now an independent variable and x is a dependent one. Let f' be a function of u and y such that

$$f' = f - ux. \tag{6.2.2}$$

Then $df' = df - u\,dx - x\,du$. Substituting for df from (6.2.1), we get

$$\begin{aligned} df' &= u\,dx + v\,dy - u\,dx - x\,du \\ \Rightarrow df' &= v\,dy - x\,du \end{aligned} \tag{6.2.3}$$

But f' is a function of u and y, therefore

$$df' = \frac{\partial f'}{\partial u}du + \frac{\partial f'}{\partial y}dy \tag{6.2.4}$$

Comparing Eqs. (6.2.3) and (6.2.4), we get

$$x = -\frac{\partial f'}{\partial u} \quad \text{and} \quad v = \frac{\partial f'}{\partial y} \tag{6.2.5}$$

These are the necessary relations for Legendre transformations.

6.3 Generating Functions

For canonical transformations, the Lagrangian L in p_k, q_k coordinates and L' in P_k, Q_k coordinates must satisfy the Hamilton's principle, i.e.,

$$\delta \int_{t_1}^{t_2} L\,dt = 0 \quad \text{and} \quad \delta \int_{t_1}^{t_2} L'\,dt = 0 \tag{6.3.1}$$

But $L = \sum_{k=1}^{n} p_k\dot{q}_k - H$ and $L' = \sum_{k=1}^{n} P_k\dot{Q}_k - H'$, therefore,

$$\delta \int_{t_1}^{t_2} \left[\sum_k p_k\dot{q}_k - H\right] dt = 0 \tag{6.3.2}$$

$$\text{and} \quad \delta \int_{t_1}^{t_2} \left[\sum_k P_k\dot{Q}_k - H'\right] dt = 0 \tag{6.3.3}$$

Subtracting Eq. (6.3.3) from Eq. (6.3.2), we get

$$\delta \int_{t_1}^{t_2} \left[\left(\sum_k p_k\dot{q}_k - H\right) - \left(\sum_k P_k\dot{Q}_k - H'\right)\right] dt = 0 \tag{6.3.4}$$

In δ-variation process, the condition $\delta \int f dt = 0$ is to be satisfied, in general, by $f = dF/dt$, where F is an arbitrary function. Therefore,

$$\delta \int_{t_1}^{t_2} \frac{dF}{dt}dt = 0 \tag{6.3.5}$$

where

$$\frac{dF}{dt} = L - L'$$
$$\frac{dF}{dt} = \left(\sum_k p_k \dot{q}_k - H\right) - \left(P_k \dot{Q}_k - H'\right) \tag{6.3.6}$$

The function F is known as the generating function. The first bracket in (6.3.6) is a function of p_k, q_k and t and the second as a function of $P_{k'}, Q_k$ and t. F is therefore, in general, a function of $(4n+1)$ variables p_k, q_k, P_k, Q_k and t. It is to be remembered that the variables are subjected to the transformation equations (6.1.1) and therefore F may be regarded as the function of $(2n+1)$ variables, comprising t and any $2n$ of the p_k, q_k, P_k, Q_k. Thus we see that F can be written as a function of $(2n+1)$ independent variables in the following four forms:

$$\text{(i) } F_1(q_k, Q_k, t), \quad \text{(ii) } F_2(q_k, P_k, t), \quad \text{(iii) } F_3(p_k, Q_k, t), \quad \text{and} \quad \text{(iv) } F_4(p_k, P_k, t) \tag{6.3.7}$$

The choice of the functional form of the generating function F depends on the problem under consideration.

Case I : If we choose the form (i), i.e.,

$$F_1 = F_1(q_1, q_2, \ldots, q_k, \ldots, q_n, Q_1, Q_2, \ldots, Q_{k'}, \ldots, Q_{n'} t) \tag{6.3.8}$$

$$\text{then,} \quad \frac{dF_1}{dt} = \sum_k \frac{\partial F_1}{\partial q_k}\dot{q}_k + \sum_k \frac{\partial F_1}{\partial Q_k}\dot{Q}_k + \frac{\partial F_1}{\partial t} \tag{6.3.9}$$

Subtracting (6.3.9) from (6.3.6), we can write

$$\sum_k \left(p_k - \frac{\partial F_1}{\partial q_k}\right)\dot{q}_k - \sum_k \left(P_k + \frac{\partial F_1}{\partial Q_k}\right)\dot{Q}_k + H' - H - \frac{\partial F_1}{\partial t} = 0$$
$$\text{or,} \quad \sum_k \left(p_k - \frac{\partial F_1}{\partial q_k}\right) dq_k - \sum_k \left(P_k + \frac{\partial F_1}{\partial Q_k}\right) dQ_k + \left[H' - H - \frac{\partial F_1}{\partial t}\right] dt = 0 \tag{6.3.10}$$

As q_k, Q_k and t may be regarded as independent variables,

$$p_k = \frac{\partial}{\partial q_k} F_1(q_k, Q_k, t), \; P_k = -\frac{\partial}{\partial Q_k} F_1(q_k, Q_k, t), \quad \text{and} \quad H' - H = \frac{\partial}{\partial t} F_1(q_k, Q_k, t) \tag{6.3.11}$$

In principle, first equation of (6.3.11) may be solved to give

$$Q_k = Q_k(q_k, p_k, t) \tag{6.3.12}$$

Substituting this in the second equation of (6.3.11), one gets

$$P_k = P_k(q_k, p_k, t) \tag{6.3.13}$$

In fact, these are the transformation equations (6.1.1). Thus we find that transformation equations can be derived from a knowledge of the function F. This is why F is known as the *generating function of the transformation*.

Case II : If the generating function is of the type $F_2(q_k, P_k, t)$, then it can be dealt with by affecting Legendre transformation of $F_1(q_k, Q_k, t)$. In case of Legandre transformation (6.2.2) :

$$f' = f - ux, \quad \text{where} \quad u = \partial f / \partial x$$

Here, since $P_k = -\partial F_1/\partial Q_k$, we have $u = -P_k, x = Q_k, f' = F_2$ and $f = F_1$. Therefore

$$F_2\left(q_k, P_k, t\right) = F_1\left(q_k, Q_k, t\right) + \sum_k P_k Q_k \tag{6.3.14}$$

Evidently, F_2 is independent of Q_k variables, because

$$\frac{\partial F_2}{\partial Q_k} = \frac{\partial F_1}{\partial Q_k} + P_k = -P_k + P_k = 0 \quad \text{as} \quad \frac{\partial F_1}{\partial Q_k} = -P_k \text{ in (6.3.11).}$$

Using Eq. (6.3.6)

$$\left(\sum_k p_k \dot{q}_k - H\right) - \left(\sum_k P_k \dot{Q}_k - H'\right) = \frac{dF_1}{dt} = \frac{d}{dt}\left[F_2 - \sum_k P_k Q_k\right]$$

$$\text{or,} \quad \frac{dF_2}{dt} = \sum_k p_k \dot{q}_k + \sum_k Q_k \dot{P}_k + H' - H \tag{6.3.15}$$

Total time derivative of $F_2\left(q_k, P_k, t\right)$ is

$$\frac{dF_2}{dt} = \sum_k \frac{\partial F_2}{\partial q_k}\dot{q}_k + \sum_k \frac{\partial F_2}{\partial P_k}\dot{P}_k + \frac{\partial F_2}{\partial t} \tag{6.3.16}$$

From (6.3.15) and (6.3.16), we get

$$p_k = \frac{\partial F_2}{\partial q_k}, \quad Q_k = \frac{\partial F_2}{\partial P_k} \quad \text{and} \quad H' - H = \frac{\partial F_2}{\partial t} \tag{6.3.17}$$

If we look (6.3.11) and (6.3.17), we find $\dfrac{\partial F_1}{\partial t} = \dfrac{\partial F_2}{\partial t}$. Further as $\dfrac{\partial F_1}{\partial q_k} = \dfrac{\partial F_2}{\partial q_k}$, first equation of (6.3.11) and that of (6.3.17) are identical. Second equation of (6.3.17) appears to be different from the second equation of (6.3.11), but in fact it is a rearrangement of it.

Case III : We can again relate the third type of generating function $F_3\left(p_k, Q_k, t\right)$ to F_1 by a Legendre transformation in view of the relation $p_k = \partial F_1/\partial q_k$. Here $u = p_k, x = q_k, f' = F_3$ and $f = F_1$. Therefore,

$$F_3\left(p_k, Q_k, t\right) = F_1\left(q_k, Q_k, t\right) - \sum_k p_k q_k$$

$$F_1\left(q_k, Q_k, t\right) = F_3\left(p_k, Q_k, t\right) + \sum_k p_k q_k \tag{6.3.18}$$

Using Eq. (6.3.6), we have

$$\left(\sum_k p_k \dot{q}_k - H\right) - \left(\sum_k P_k \dot{Q}_k - H'\right) = \frac{dF_1}{dt} = \frac{d}{dt}\left(F_3 + \sum p_k q_k\right)$$

$$\text{or,} \quad \frac{dF_3}{dt} = -\sum_k \dot{p}_k q_k - \sum_k P_k \dot{Q}_k + H' - H$$

$$\text{Also,} \quad \frac{dF_3}{dt} = \sum_k \frac{\partial F_3}{\partial p_k}\dot{p}_k + \sum_k \frac{\partial F_3}{\partial Q_k}\dot{Q}_k + \frac{\partial F_3}{\partial t}$$

Therefore, the new transformation equations are

$$q_k = -\frac{\partial F_3}{\partial p_k}, P_k = -\frac{\partial F_3}{\partial Q_k} \quad \text{and} \quad H' - H = \frac{\partial F_3}{\partial t} \tag{6.3.19}$$

Case IV : Using Legendre transformations, the generating function $F_4\left(p_k, P_k, t\right)$ can be connected to $F_1\left(q_k, Q_k, t\right)$ as

$$F_4\left(p_k, P_k, t\right) = F_1\left(q_k, Q_k, t\right) + \sum_k P_k Q_k - \sum_k p_k q_k \tag{6.3.20}$$

Using Eq. (6.3.6), we have

$$\left(\sum_k p_k \dot{q}_k - H\right) - \left(\sum_k P_k \dot{Q}_k - H'\right) = \frac{d}{dt}\left(F_4 - \sum_k P_k Q_k + \sum_k p_k q_k\right)$$

$$\text{or,} \quad \frac{dF_4}{dt} = -\sum_k q_k \dot{p}_k + \sum_k Q_k \dot{P}_k + H' - H$$

But

$$\frac{dF_4}{dt} = \sum_k \frac{\partial F_4}{\partial p_k}\dot{p}_k + \sum_k \frac{\partial F_4}{\partial P_k}\dot{P}_k + \frac{\partial F_4}{\partial t}$$

A comparison of the above two equations gives the fourth set of transformation equations:

$$q_k = -\frac{\partial F_4}{\partial p_k}, \quad Q_k = \frac{\partial F_4}{\partial P_k}, \quad H' - H = \frac{\partial F_4}{\partial t} \tag{6.3.21}$$

6.4 Procedure for Application of Canonical Transformations

We note that the relation between,H and H' in all the cases has the same form i.e., $H' = H + \partial F/\partial t$. Now, if F has no explicit time dependence, then $\partial F/\partial t = 0$ and hence

$$H' = H \tag{6.4.1}$$

Thus, when the generating function has no explicit time dependence, the new Hamiltonian H' is obtained from the old Hamiltonian H by substituting for p_k, q_k in terms of the new variables P_k, Q_k. Further we note that the time t has been treated as an invariant parameter of the motion and we have not made any provision for a transformation of the time coordinate alongwith the other coordinates.

If in the new set of coordinates (P_k, Q_k, t) all coordinates Q_k are cyclic, then

$$\dot{P}_k = -\frac{\partial H'}{\partial Q_k} = 0 \quad \text{or} \quad P_k = \text{ Constant, say } \alpha_k \tag{6.4.2}$$

If the generating function F does not depend on time t explicitly and H is a constant of motion, not depending on time, then from (6.4.1) H' is also constant of motion. Thus H' will not involve Q_k and t (explicit time dependence). Therefore,

$$H\left(q_k, p_k\right) = H'\left(Q_k, P_k\right) = H'\left(P_k\right) = H'\left(\alpha_1, \alpha_2, \ldots, \alpha_n\right)$$

Hamilton's equations for Q_k are

$$\dot{Q}_k = \frac{\partial H'}{\partial P_k} = \frac{\partial H'}{\partial \alpha_k} = \omega_k \tag{6.4.3}$$

where ω_k's are functions of the α_k's only and are constant in time.

Eq. (6.4.3) has the solution

$$Q_k = \omega_k t + \beta_k \tag{6.4.4}$$

where β_k's are the constants of integration, determined by the initial conditions.

6.5 Condition for Canonical Transformations

Suppose $F = F(q_k, Q_k)$, then obviously $\partial F/\partial t = 0$ and $H = H'$ [from (6.3.11)]. Further from (6.3.11), we have

$$p_k = \frac{\partial F}{\partial q_k} \quad \text{and} \quad P_k = -\frac{\partial F}{\partial Q_k}$$

Also

$$dF = \sum_k \frac{\partial F}{\partial q_k} dq_k + \sum_k \frac{\partial F}{\partial Q_k} dQ_k$$

$$\text{or,} \quad dF = \sum_k p_k dq_k - \sum_k P_k dQ_k \tag{6.5.1}$$

The left hand side of Eq. (6.5.1) is an exact differential, hence for a given transformation to be canonical, the right hand side of Eq. (6.5.1), i.e., $\sum_k p_k dq_k - \sum_k P_k dQ_k$ must be an exact differential.

Example 6.5.1. Prove that the generating function $F = \sum_i q_i P_i$ generates the identity transformation.

Solution. Here, the generating function is $F_2 = \sum_i q_i P_i$ and hence applying Eq. (6.3.17), we get

$$p_i = \partial F_2/\partial q_i = P_i, \quad Q_i = \partial F_2/\partial P_i = q_i$$
$$H' = H \ (\because F_2 \text{ is not } t \text{ dependent })$$

Thus the new and old variables are separately equal and hence F generates an identity transformation.

Example 6.5.2. Show that for the function $F = \sum_k q_k Q_k$, the transformations are $p_k = Q_k, P_k = -q_k$ and $H' = H$.

Solution. Here $F = \sum_k q_k Q_k$ is F_1 and hence applying Eqs. (6.3.11), we get

$$p_k = \frac{\partial F_1}{\partial q_k} = Q_k, P_k = -\frac{\partial F_1}{\partial Q_k} = -q_k \quad \text{and} \quad H' = H.$$

Example 6.5.3. Show that the transformation

$$P = \frac{1}{2}\left(p^2 + q^2\right), Q = \tan^{-1}\frac{q}{p}$$

is canonical.

Solution. The transformation will be canonical, if $p\,dq - P\,dQ$ is an exact differential. Here

$$dQ = (p\,dq - q\,dp)/p^2 + q^2$$

Therefore,

$$p\,dq - P\,dQ = p\,dq - \frac{1}{2}\left(p^2 + q^2\right)\frac{p\,dq - q\,dp}{p^2 + q^2} = \frac{1}{2}(p\,dq + q\,dp) = d\left(\frac{1}{2}pq\right) = \text{ an exact differential}$$

This means that the given transformation is canonical.

Example 6.5.4. The transformation equations between two sets of coordinates are

$$P = 2\left(1 + q^{1/2}\cos p\right) q^{1/2}\sin p \quad \text{and} \quad Q = \log\left(1 + q^{1/2}\cos p\right)$$

Show that the transformation is canonical and the generating function of this transformation is

$$F_3 = -\left(e^Q - 1\right)^2 \tan p.$$

Solution. Here,

$$\begin{aligned}
(p\,dq - P\,dQ) &= p\,dq - 2\left[1 + q^{1/2}\cos p\right] q^{1/2}\cos p \times \frac{\left(-q^{1/2}\sin p\,dp + \frac{1}{2}\cos p\,dq/q^{-1/2}\right)}{\left(1 + q^{1/2}\cos p\right)} \\
&= p\,dq + 2q\,\sin^2 p\,dp - \sin p\cos p\,dq \\
&= \left(p - \frac{1}{2}\sin 2p\right)\,dq + q(1 - \cos 2p)\,dp \\
&= d\left[q\left(p - \frac{1}{2}\sin 2p\right)\right]
\end{aligned}$$

which is an exact differential and hence the transformation is canonical. Further

$$\begin{aligned}
Q = \log_e\left(1 + q^{1/2}\cos p\right) &\Rightarrow e^Q = 1 + q^{1/2}\cos p \\
\Rightarrow q^{1/2}\cos p = e^Q - 1 &\Rightarrow q = \left(e^Q - 1\right)^2/\cos^2 p
\end{aligned}$$

For this transformation, we take $F = F_3(p, Q)$, so that

$$q = -\frac{\partial F_3}{\partial p} \quad \text{and} \quad P = -\frac{\partial F_3}{\partial Q}$$

Thus

$$\begin{aligned}
-\frac{\partial F_3}{\partial p} = \left(e^Q - 1\right)^2 \frac{1}{\cos^2 p} \quad &\text{or} \quad F_3 = -\int \frac{\left(e^Q - 1\right)^2}{\cos^2 p}\,dp \\
\text{or,} \quad F_3 = -\left(e^Q - 1\right)^2 \tan p + &\text{ constant}
\end{aligned}$$

If the constant of integration is zero,

$$F_3 = -\left(e^Q - 1\right)^2 \tan p.$$

6.6 Bilinear Invariant Condition

According to this condition, if a transformation (q_k, p_k) coordinates to (Q_k, P_k) coordinates is canonical, then bilinear form

$$\sum_k \left(\delta p_k dq_k - \delta q_k dp_k\right) \tag{6.6.1}$$

remains invariant. This statement means that

$$\sum_k \left(\delta p_k dq_k - \delta q_k dp_k\right) = \sum_k \left(\delta P_k dQ_k - \delta Q_k dP_k\right) \tag{6.6.2}$$

Proof. From Hamilton's canonical equations, we have

$$\dot{q}_k = \frac{\partial H}{\partial p_k} \quad \text{or} \quad dq_k = \frac{\partial H}{\partial p_k} dt \tag{6.6.3}$$

$$\text{and} \quad \dot{p}_k = -\frac{\partial H}{\partial q_k} \quad \text{or} \quad dp_k = -\frac{\partial H}{\partial q_k} dt \tag{6.6.4}$$

Similarly,

$$dQ_k = \frac{\partial H}{\partial P_k} dt \quad \text{and} \quad dP_k = -\frac{\partial H}{\partial Q_k} dt \tag{6.6.5}$$

Since δp_k and δq_k are arbitrary,

$$\sum_k \delta p_k \left(dq_k - \frac{\partial H}{\partial p_k} dt \right) - \sum_k \delta q_k \left(dp_k + \frac{\partial H}{\partial q_k} dt \right) = 0 \tag{6.6.6}$$

Obviously in order to satisfy this equation, the coefficients of δp_k and δq_k must be zero and this gives Eqs. (**??**). Therefore, Eq. (6.6.6) is correct and it can be written as

$$\sum_k (\delta p_k dq_k - \delta q_k dp_k) - \sum_k \left(\frac{\partial H}{\partial p_k} \delta p_k + \frac{\partial H}{\partial q_k} \delta q_k \right) dt = 0$$

$$\sum_k (\delta p_k dq_k - \delta q_k dp_k) - \delta H \; dt = 0 \tag{6.6.7}$$

Similarly, for $H' = H$, when F does not depend on time,

$$\sum_k (\delta P_k \; dQ_k - \delta Q_k \; dP_k) - \delta H \; dt = 0 \tag{6.6.8}$$

Eliminating $\delta H \; dt$ from Eqs. (6.6.7) and (6.6.8), we obtain

$$\sum_k (\delta p_k \; dq_k - \delta q_k \; dp_k) = \sum_k (\delta P_k \; dQ_k - \delta Q_k \; dP_k) \tag{6.6.9}$$

which proves the statement

Example 6.6.1. Show that the transformation $Q = \dfrac{1}{p}$ and $P = qp^2$ is canonical.

Solution. Since $Q = \dfrac{1}{p}$, therefore

$$dQ = \frac{\partial Q}{\partial p} dp + \frac{\partial Q}{\partial q} dq$$

$$\text{or,} \quad dQ = \frac{\partial}{\partial p} \left(\frac{1}{p} \right) dp + \frac{\partial}{\partial q} \left(\frac{1}{p} \right) dq = -\frac{1}{p^2} dp$$

$$\text{Also,} \quad \delta Q = \frac{\partial Q}{\partial p} \delta p + \frac{\partial Q}{\partial q} \delta q = -\frac{1}{p^2} \delta p$$

Similarly, $dP = p^2 \; dq + 2qp \; dp \quad \left(\because P = qp^2 \right)$ and $\delta P = p^2 \delta q + 2qp \; \delta p$.
Therefore,

$$\begin{aligned} \delta P \; dQ - \delta Q \; dP &= \left(p^2 \; \delta q + 2qp \; \delta p \right) \left(-\frac{1}{p^2} \; dp \right) - \left(-\frac{1}{p^2} \; \delta p \right) \left(p^2 dq + 2qp \; dp \right) \\ &= -\delta q \; dp - \frac{2q}{p} \delta p \; dp + \delta p \; dq + \frac{2q}{p} \; \delta p \; dp \\ &= \delta p \; dq - \delta q \; dp \end{aligned}$$

Therefore, the bilinear form is invariant and hence the transformation is canonical.

6.7 Integral Invariance of Poincare

Phase space is defined as a $2n$ dimensional space formed by the $2n$ coordinates $q_1, q_2, \ldots, q_n, p_1, p_2, \ldots, p_n$. In this space a complete dynamical specification of a mechanical system is given by a point.

According to *Poincare's theorem*, the integral

$$I = \iint_S \sum_k dq_k \, dp_k \tag{6.7.1}$$

taken over an arbitrary two dimensional surface S of $2n$ dimensional phase space is invariant under canonical transformation, i.e.,

$$\iint_S \sum_k dq_k \, dp_k = \iint_S \sum_k dQ_k dP_k \tag{6.7.2}$$

If S is a 4-dimensional surface in $2n$-dimensional phase space, then according to Poincare's theorem,

$$\iint_S \sum_k \sum_l dq_k dq_l dp_k dp_l = \iint_S \sum_k \sum_l dQ_k dQ_l dP_k dP_l$$

In general, if the surface is $2n$-dimensional in $2n$-dimensional phase space, then the integral invariance of Poincare means

$$\iint \cdots \int dq_1 \, dq_2 \, \ldots \, dq_n \, dp_1 \cdots dp_n = \iint \cdots \int dQ_1 \, dQ_2 \ldots dQ_n \, dP_1 \ldots dP_n \tag{6.7.3}$$

which shows that the volume in phase space is invariant under canonical transformation. In the advanced calculus, we have the relation

$$\int \cdots \int dQ_1 \, dQ_2 \ldots dQ_n \, dP_1 \ldots dP_n = \int \cdots \int D \, dq_1 \, dq_2 \, \ldots \, dq_n \, dp_1 \ldots dp_n \tag{6.7.4}$$

where D is known as the Jacobian of the transformation, given by

$$D = \frac{\partial \left(Q_1, Q_2, \ldots, Q_n, P_1, P_2, \ldots, P_n\right)}{\partial \left(q_1, q_2, \ldots, q_n, p_1, p_2, \ldots, p_n\right)} \tag{6.7.5}$$

This means that in order to prove the integral invariance (6.7.3), we have to show $D = 1$. By using the properties of the Jacobian, it can be written as

$$D = \frac{\frac{\partial(Q_1, Q_2, \ldots, Q_n, P_1, P_2, \ldots, P_n)}{\partial(q_1, q_2, \ldots, q_n, P_1, P_2, \ldots, P_n)}}{\frac{\partial(q_1, q_2, \ldots, q_n, p_1, p_2, \ldots, p_n)}{\partial(q_1, q_2, \ldots, q_n, P_1, P_2, \ldots, P_n)}} \tag{6.7.6}$$

In the calculus, we know that if the same variables are present in both the partial differentials, the Jacobian is reduced to fewer variables in which the repeated variables are treated as constants in carrying out the differentiation. Thus

$$D = \frac{\left[\frac{\partial(Q_1, Q_2, \ldots, Q_n)}{\partial(q_1, q_2, \ldots, q_n)}\right]_{P_1, P_2, \ldots, P_n \text{ as constants}}}{\left[\frac{\partial(p_1, p_2, \ldots, p_n)}{\partial(P_1, P_2, \ldots, P_n)}\right]_{q_1, q_2, \ldots, q_n \text{ as constants}}} \tag{6.7.7}$$

The numerator is a determinant of order n whose element in the i-th row and k-th column is $\partial Q_k/\partial q_{p'}$, i.e.,

$$\frac{\partial(Q_1, Q_2, \ldots, Q_n)}{\partial(q_1, q_2, \ldots, q_n)} = \begin{vmatrix} \frac{\partial Q_1}{\partial q_1} & \frac{\partial Q_2}{\partial q_1} & \cdots & \frac{\partial Q_n}{\partial q_1} \\ \frac{\partial Q_1}{\partial q_2} & \frac{\partial Q_2}{\partial q_2} & \cdots & \frac{\partial Q_n}{\partial q_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial Q_1}{\partial q_i} & \frac{\partial Q_2}{\partial q_i} & \cdots & \frac{\partial Q_n}{\partial q_i} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial Q_1}{\partial q_n} & \frac{\partial Q_2}{\partial q_n} & \cdots & \frac{\partial Q_n}{\partial q_n} \end{vmatrix} \tag{6.7.8}$$

Similarly, the denominator is a determinant of the same order n whose element in the i-th row and k-th column is $\dfrac{\partial p_k}{\partial P_i}$. If the generating function of the above canonical transformation is written as $F_2\,(q_k, P_k)$, then from Eq. (6.3.17), we obtain

$$Q_k = \frac{\partial F_2}{\partial P_k} \quad \text{and} \quad p_k = \frac{\partial F_2}{\partial q_k}$$

and hence

$$\frac{\partial Q_k}{\partial q_i} = \frac{\partial^2 F_2}{\partial q_i \partial P_k} \quad and \quad \frac{\partial p_k}{\partial P_i} = \frac{\partial^2 F_2}{\partial P_i \partial q_k} \tag{6.7.9}$$

Thus, we see that the ik-element of the numerator is the same as the ki-element of the denominator. Since in a determinant, rows and columns can be interchanged and hence the determinant of the numerator is equal to the determinant of the denominator. Therefore, from (6.7.7), we get

$$D = 1 \tag{6.7.10}$$

Thus we see that Eq. (6.7.3) is true, i.e., the volume in phase space is invariant under canonical transformation.

Also, if we take a two dimensional surface S of $2n$-dimensional phase space, then the invariance of Poincare's integral under canonical transformation means that

$$\iint_S \sum_k dq_k dp_k = \iint_S \sum_k dQ_k dP_k.$$

6.7.1 Infinitesimal Contact Transformations

Those transformations in which the new set of coordinates (Q_k, P_k) differ from the old set (q_k, p_k) by infinitesimals i.e., $Q_k = q_k + \delta q_k$ and $P_k = p_k + \delta p_k$, are called *infinitesimal contact transformations*.

It is known that the generating function $F_2 = \sum_k q_k P_k$ generates the identity transformation i.e., $Q_k = q_k$ and $P_k = p_k$. The generating function, giving an infinitesimal change in the variables, can be readily written as

$$F_2 = \sum_k q_k P_k + \varepsilon G\,(q_k, P_k) \tag{6.7.11}$$

where ε is an infinitesimal parameter of the transformation and $G\,(q_k, P_k)$, is arbitrary. Substitution of (6.7.11) in Eqs. (6.3.17) gives

$$p_k = \frac{\partial F_2}{\partial q_k} = P_k + \varepsilon \frac{\partial G}{\partial q_k}, \quad Q_k = \frac{\partial F_2}{\partial P_k} = q_k + \varepsilon \frac{\partial G}{\partial P_k}, \quad H' = H \tag{6.7.12}$$

Therefore,

$$Q_k - q_k = \delta q_k = \varepsilon \frac{\partial G}{\partial P_k} \quad \text{and} \quad P_k - p_k = \delta p_k = -\varepsilon \frac{\partial G}{\partial q_k} \tag{6.7.13}$$

Since the difference $(P_k - p_k)$ is infinitesimal. We can replace P_k by p_k in the derivative and also $G(q_k, P_k)$ by $G(q_k, p_k)$. So that Eqs. (6.7.13) are

$$\delta q_k = \varepsilon \frac{\partial G}{\partial p_k} \quad \text{and} \quad \delta p_k = -\varepsilon \frac{\partial G}{\partial q_k} \tag{6.7.14}$$

In case of infinitesimal contact transformations, the description is transferred to the function G instead of the original generating function F. Thus G is the new generating function which generates the infinitesimal contact transformation.

Let us consider a special case in which $\epsilon = dt$ and $G = H$. Eqs. (6.7.14) can be written by using Hamilton's equations of motion as

$$\delta q_k = dt \frac{\partial H}{\partial p_k} = dt\, \dot{q}_k = dq_k \quad \text{and} \quad \delta p_k = -dt \frac{\partial H}{\partial q_k} = dt\, \dot{p}_k = dp_k \tag{6.7.15}$$

These changes in the conjugate variables represent an infinitesimal change in coordinates in time dt. Eqs. (6.7.15) givë thus a transformation from the variables q_k, p_k at time t to $q_k + dq_k, p_k + dp_k$ at time $t + dt$. Hence *the motion of the system in a small time dt can be described by an infinitesimal canonical transformation generated by the Hamiltonian H of the system.* Evidently the motion of the system in a finite interval of time is described by a succession of infinitesimal canonical transformations generated by the same Hamiltonian. In other words, the motion of a system corresponds to the continuous evolution of canonical transformation. Thus we can say that the Hamiltonian of the system is the generator of the motion of the system in phase space with time.

Exercise 6.7.1. 1. Let F be a generating function depend only on $Q_\alpha,\ P_\alpha,\ t$. Prove that

$$P_\alpha = -\frac{\partial F}{\partial Q_\alpha},\ q_\alpha = -\frac{\partial F}{\partial p_\alpha},\ H' = \frac{\partial F}{\partial t} + H.$$

2. Determine the values of α and β so that the equations $Q = q^\alpha \cos \beta p$ and $P = q^2 \sin \beta p$ is canonical transformation. Also find the generating function F_3 for this case.

3. Given that the linear transformation of a generalized coordinate q and the corresponding momentum p is canonical. Find the value of constant α.

4. Show that the transformation $Q = \sqrt{2q}e^\alpha \cos p$ and $P = \sqrt{2q}e^{-\alpha} \sin p$, with α being constant, is canonical.

5. Prove that the transformation $P = q \cot p$ and $Q = \log \dfrac{\sin p}{q}$ is canonical. Show that the generating function is $F(q, Q) = e^{-Q}\left(1 - q^2 e^{2Q}\right)^{1/2} + q \sin^{-1}\left(qe^Q\right)$.

6. Show that the transformation $Q = p + iaq,\ P = \dfrac{p - iaq}{2ia}$ is canonical and find a generating function.

7. Find the canonical transformation defined by the generating function $F_1(q, Q) = qQ - \dfrac{1}{2}m\omega q^2 - Q^2/4m\omega$.

8. A canonical transformation $(q, p) \to (Q, P)$ is made through the generating function $F(q, p) = q^2 P$ on the Hamiltonian $H(p, q) = \dfrac{p^2}{2\alpha q^2} + \dfrac{\beta}{4}q^4$ where α and β are constants. Find the equations of motion for (Q, P).

Unit 7

Course Structure

- Lagrange's and Poisson's brackets and their variance under canonical transformations,
- Hamilton's equations of motion in Poisson's bracket.
- Jacobi's identity.
- Hamilton-Jacobi equation.

7.1 Introduction

In the previous unit, we have shown that in the case of infinitesimal contact transformations, the changes in the conjugate variables p_k and q_k are given by

$$\delta q_k = \varepsilon \frac{\partial G}{\partial p_k} \quad \text{and} \quad \delta p_k = -\varepsilon \frac{\partial G}{\partial q_k} \tag{7.1.1}$$

where ε is an infinitesimal parameter and the generating function $G(q_k, p_k)$ is arbitrary. Now let us consider some function $F(q_k, p_k)$. The change in the value of $F(q_k, p_k)$ with the changes δq_k and δp_k in the coordinates q_k and p_k respectively can be expressed as

$$\delta F = \sum_k \left(\frac{\partial F}{\partial q_k} \delta q_k + \frac{\partial F}{\partial p_k} \delta p_k \right) \tag{7.1.2}$$

If the transformation (7.1.1), generated by the function G, is applied, we get

$$\delta F = \sum_k \left[\frac{\partial F}{\partial q_k} \left(\varepsilon \frac{\partial G}{\partial p_k} \right) + \frac{\partial F}{\partial p_k} \left(-\varepsilon \frac{\partial G}{\partial q_k} \right) \right]$$

Since the parameter ε is independent of q_k and p_k, we have

$$\delta F = \varepsilon \left[\sum_k \left(\frac{\partial F}{\partial q_k} \frac{\partial G}{\partial p_k} - \frac{\partial F}{\partial p_k} \frac{\partial G}{\partial q_k} \right) \right] \tag{7.1.3}$$

The quantity in the big bracket in (7.1.3) is called the *Poisson bracket* of two functions or dynamical variables $F(q_k, p_k)$ and $G(q_k, p_k)$ and is denoted by $[F, G]$. This definition of Poisson bracket is true for F and G, being functions of time. Thus

$$\delta F = \varepsilon [F, G] \tag{7.1.4}$$

7.2 Poisson's Brackets

If the functions F and G depend upon the position coordinates q_k, momentum coordinates p_k and time t, the Poisson bracket of F and G is defined as

$$[F, G]_{q,p} = \sum_{k=1}^{n} \left(\frac{\partial F}{\partial q_k} \frac{\partial G}{\partial p_k} - \frac{\partial F}{\partial p_k} \frac{\partial G}{\partial q_k} \right) \tag{7.2.1}$$

For brevity, we may drop the subscripts $q,\ p$ and write the Poisson bracket as $[F, G]$. The total time derivative of the function F can be written as

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + \sum_{k=1}^{n} \left(\frac{\partial F}{\partial q_k} \dot{q}_k + \frac{\partial F}{\partial p_k} \dot{p}_k \right) \tag{7.2.2}$$

Using, Hamilton's equations $\dot{q}_k = \dfrac{\partial H}{\partial p_k}$ and $-\dot{p}_k = \dfrac{\partial H}{\partial q_k}$, Eq. (7.2.2) is obtained to be

$$\frac{dF}{dt} = \dot{F} = \frac{\partial F}{\partial t} + \sum_{k=1}^{n} \left(\frac{\partial F}{\partial q_k} \frac{\partial H}{\partial p_k} - \frac{\partial F}{\partial p_k} \frac{\partial H}{\partial q_k} \right) \tag{7.2.3}$$

In view of the definition of Poisson's bracket given by Eq. (7.2.1), we obtain

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + [F, H] \tag{7.2.4}$$

From this equation we see that the function F is a constant of motion, if

$$\frac{dF}{dt} = 0 \quad \text{or} \quad \frac{\partial F}{\partial t} + [F, H] = 0 \tag{7.2.5}$$

Now, if the function F does not depend on time explicitly, $\dfrac{\partial F}{\partial t} = 0$ and then the condition for F to be constant of motion is obtained to be

$$[F, H] = 0 \tag{7.2.6}$$

Thus *if a function F does not depend on time explicitly and is a constant of motion, its Poisson bracket with the Hamiltonian vanishes. In other words, a function whose Poisson bracket with Hamiltonian vanishes is a constant of motion.* This result does not depend whether H itself is constant of motion.

Equations of motion in Poisson bracket form : Special cases of (7.2.4) are

$$F = q_k, \quad \dot{q}_k = [q_k, H] \tag{7.2.7}$$

$$F = p_k, \quad \dot{p}_k = [p_k, H] \tag{7.2.8}$$

$$F = H, \quad \dot{H} = \frac{\partial H}{\partial t} \tag{7.2.9}$$

These equations (7.2.7), (7.2.8), (7.2.9) are identical to Hamilton's equations and referred as *equations of motion in Poisson bracket form.*

Properties of Poisson brackets and Fundamental Poisson brackets : The Poisson bracket has the property of anti-symmetry, given by

$$[F, G] = -[G, F], \tag{7.2.10}$$

because

$$[F,G]=\sum_k\left[\frac{\partial F}{\partial q_k}\frac{\partial G}{\partial p_k}-\frac{\partial F}{\partial p_k}\frac{\partial G}{\partial q_k}\right]=-\sum_k\left[\frac{\partial G}{\partial q_k}\frac{\partial F}{\partial p_k}-\frac{\partial G}{\partial p_k}\frac{\partial F}{\partial q_k}\right]=-[G,F].$$

Thus Poisson bracket does not obey the commutative law of algebra. As an application of the Poisson brackets, we are giving below some of the special cases :

1. When $G=q_l$,

$$[F,q_l]=\sum_k\left[\frac{\partial F}{\partial q_k}\frac{\partial q_l}{\partial p_k}-\frac{\partial F}{\partial p_k}\frac{\partial q_l}{\partial q_k}\right]=-\sum_k\frac{\partial F}{\partial p_k}\delta_{lk}=-\frac{\partial F}{p_l}$$

   Also if $F=q_k,\quad [q_k,q_l]=-\dfrac{\partial q_k}{\partial p_l}=0$ and if $F=p_k,\quad [p_k,q_l]=-\dfrac{\partial p_k}{\partial p_l}=-\delta_{kl}$

2. When $G=p_t,[F,p_l]=\sum_k\dfrac{\partial F}{\partial q_k}\delta_{kl}=[F,p_l]=\dfrac{\partial F}{\partial q_l}.$

   For $F=p_k,\quad [p_k,p_l]=\dfrac{\partial p_k}{\partial q_l}=0$ and for $F=q_k,\quad [q_k,p_l]=\dfrac{\partial q_k}{\partial q_l}=\delta_{kl}.$
   The above results can be summarized as follows:

$$[q_k,q_l]=[p_k,p_l]=0\quad\text{and}\quad[q_k,p_l]=\delta_{kl}\tag{7.2.11}$$

   where δ_{kl} is the kronecker delta symbol with the property

$$\delta_{kl}=0\quad\text{for}\quad k\neq l\quad\text{and}\quad\delta_{kl}=1\quad\text{for}\quad k=l.$$

   Equations (7.2.11) are called the *fundamental Poisson's brackets*.

Further from the definition of Poisson bracket of any two dynamical variables or functions, one can obtain the following identities:

(i) $[F,F]=0$ (ii) $[F,C]=0,C=$ constant (iii) $[CF,G]=C[F,G]$

(iv) $[F_1+F_2,G]=[F_1,G]+[F_2,G]$ (v)] $[F,G_1G_2]=G_1\,[F,G_2]+[F,G_1]\,G_2$

(vi) $\dfrac{\partial}{\partial t}[F,G]=\left[\dfrac{\partial F}{\partial t},G\right]+\left[F,\dfrac{\partial G}{\partial t}\right]$ (vii) $[F,[G,K]]+[G,[K,F]]+[K,[F,G]]=0$ (Jacobi's identity)

7.3 Lagrange Brackets

The Lagrange bracket of two dynamical variables $F\,(q_k,p_k)$ and $G\,(q_k,p_k)$ is defined as

$$\{F,G\}=\sum_k\left[\frac{\partial q_k}{\partial F}\frac{\partial p_k}{\partial G}-\frac{\partial p_k}{\partial F}\frac{\partial q_k}{\partial G}\right]\tag{7.3.1}$$

The Langrange's bracket does not obey the commutative law of algebra i.e., for Lagrangian bracket

$$\{F,G\}=-\{G,F\}\tag{7.3.2}$$

because

$$\{F,G\}=-\sum_k\left[\frac{\partial q_k}{\partial G}\frac{\partial p_k}{\partial F}-\frac{\partial p_k}{\partial G}\frac{\partial q_k}{\partial F}\right]=-\{G,F\}$$

Further

$$\{q_i, q_j\} = \sum_k \left[\frac{\partial q_k}{\partial q_i}\frac{\partial p_k}{\partial q_j} - \frac{\partial p_k}{\partial q_i}\frac{\partial q_k}{\partial q_j}\right] = 0 \tag{7.3.3}$$

because

$$\frac{\partial p_k}{\partial q_j} = \frac{\partial p_k}{\partial q_i} = 0.$$

Similarly, one can prove that for Lagrange brackets

$$\{p_i, p_j\} = 0; \{q_i, p_j\} = \delta_{ij} \tag{7.3.4}$$

7.4 Relation between Lagrange and Poisson Brackets

If $F_k,\ k = 1, 2, \ldots, 2n$, are $2n$ independent functions such that each F_k is a function of $2n$ coordinates q_1, $q_2, \ldots, q_n; p_1, p_2, \ldots, p_n$, then

$$\sum_{k=1}^{2n} \{F_k, F_i\}\,[F_k, F_j] = \delta_{ij} \tag{7.4.1}$$

In order to prove the relation (7.4.1), we take the left hand side of this equation and use the definitions of Poisson and Lagrange brackets :

$$\begin{aligned}\sum_{k=1}^{2n} \{F_k, F_i\}\,[F_k, F_j] &= \sum_{k=1}^{2n}\left[\sum_{l=1}^{n}\sum_{m=1}^{n}\left(\frac{\partial q_l}{\partial F_k}\frac{\partial p_l}{\partial F_i} - \frac{\partial p_l}{\partial F_k}\frac{\partial q_l}{\partial F_i}\right)\left(\frac{\partial F_k}{\partial q_m}\frac{\partial F_j}{\partial p_m} - \frac{\partial F_k}{\partial p_m}\frac{\partial F_j}{\partial q_m}\right)\right] \\ &= \sum_{l=1}^{n}\left(\frac{\partial F_j}{\partial p_l}\frac{\partial p_l}{\partial F_i} + \frac{\partial F_j}{\partial q_l}\frac{\partial q_l}{\partial F_i}\right) = \frac{\partial F_j}{\partial F_i} = \delta_{ij}\end{aligned}$$

In general, Poisson bracket is relatively much more useful than Lagrange bracket.

7.5 Invariance of Poisson Bracket with respect to Canonical Transformations

Poisson brackets are invariant under canonical transformations. First we shall prove this statement for fundamental Poisson brackets and then in general.

7.5.1 Fundamental Poisson brackets under canonical transformation

The fundamental Poisson brackets are invariant under canonical transformation means that if

$$[q_k, q_l] = [p_k, p_l] = 0,\ [q_k, p_l] = \delta_{kl} \tag{7.5.1}$$

and the transformation $(q_k, p_k) \to (Q_k, P_k)$ is canonical, then

$$[Q_k, Q_l] = [P_k, P_l] = 0, [Q_k, P_l] = \delta_{kl} \tag{7.5.2}$$

According to the definition of Poisson bracket [Eq. (7.2.1)], we have

$$[F, G]_{q,p} = \sum_{i=1}^{n}\left(\frac{\partial F}{\partial q_i}\frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial G}{\partial q_i}\right) \tag{7.5.3}$$

Therefore,

$$[Q_k, Q_l]_{q,p} = \sum_i \left[\frac{\partial Q_k}{\partial q_i} \frac{\partial Q_l}{\partial p_i} - \frac{\partial Q_k}{\partial p_i} \frac{\partial Q_l}{\partial q_i} \right] \tag{7.5.4}$$

From Eq. (6.3.11) of Unit 6, we get

$$\frac{\partial p_k}{\partial Q_l} = \frac{\partial}{\partial Q_l} \frac{\partial F_1}{\partial q_k} = \frac{\partial}{\partial q_k} \frac{\partial F_1}{\partial Q_l} = -\frac{\partial P_l}{\partial q_k} \tag{7.5.5}$$

Similarly Eqs. (6.3.17), (6.3.19) and (6.3.21) of Unit 6 yield

$$\frac{\partial p_k}{\partial P_l} = \frac{\partial}{\partial P_l} \frac{\partial F_2}{\partial q_k} = \frac{\partial}{\partial q_k} \frac{\partial F_2}{\partial P_l} = \frac{\partial Q_l}{\partial q_k} \tag{7.5.6}$$

$$\frac{\partial q_k}{\partial Q_l} = -\frac{\partial}{\partial Q_l} \frac{\partial F_3}{\partial p_k} = -\frac{\partial}{\partial p_k} \frac{\partial F_3}{\partial Q_l} = \frac{\partial P_l}{\partial p_k} \tag{7.5.7}$$

$$\frac{\partial q_k}{\partial P_l} = -\frac{\partial}{\partial P_l} \frac{\partial F_4}{\partial p_k} = -\frac{\partial}{\partial p_k} \frac{\partial F_4}{\partial P_l} = -\frac{\partial Q_l}{\partial p_k} \tag{7.5.8}$$

Hence Eq. (7.5.4) is [using (7.5.5) and (7.5.7)]

$$[Q_k, Q_l]_{q,p} = \sum_i \left(-\frac{\partial Q_k}{\partial q_i} \frac{\partial q_i}{\partial P_l} - \frac{\partial Q_k}{\partial p_i} \frac{\partial p_i}{\partial P_l} \right) = -\frac{\partial Q_k}{\partial P_l} = 0 \tag{7.5.9}$$

because Q_k and P_k are independent variables. Also we note that

$$[Q_k, Q_l]_{Q,P} = \sum_i \left(-\frac{\partial Q_k}{\partial Q_i} \frac{\partial Q_l}{\partial P_i} - \frac{\partial Q_k}{\partial P_i} \frac{\partial Q_l}{\partial Q_i} \right) = 0$$

Therefore,

$$[Q_k, Q_l]_{q,p} = [Q_k, Q_l]_{Q,P} = 0 \tag{7.5.10}$$

Similarly we can prove

$$[P_k, P_l]_{q,p} = [P_k, P_l]_{Q,P} = 0 \tag{7.5.11}$$

Now,

$$[Q_k, P_l]_{q,p} = \sum_i \left(\frac{\partial Q_k}{\partial q_i} \frac{\partial P_l}{\partial p_i} - \frac{\partial Q_k}{\partial p_i} \frac{\partial P_l}{\partial q_i} \right)$$

Using Eqs. (7.5.5) and (7.5.7), we obtain

$$[Q_k, P_l]_{q,p} = \sum_i \left(\frac{\partial Q_k}{\partial q_i} \frac{\partial q_i}{\partial Q_l} + \frac{\partial Q_k}{\partial p_i} \frac{\partial p_i}{\partial Q_l} \right) = \frac{\partial Q_k}{\partial Q_l} = \delta_{kl} \tag{7.5.12}$$

By definition:

$$[Q_k, P_l]_{Q,P} = \delta_{kl} \tag{7.5.13}$$

Thus

$$[Q_k, P_l]_{q,p} = [Q_k, P_l]_{Q,P} = \delta_{kl}. \tag{7.5.14}$$

Eqs. (7.5.10), (7.5.11) and (7.5.14) show the invariance of fundamental Poisson brackets with respect to canonical transformation.

7.5.2 Poisson brackets under canonical transformation

In general, if Poisson bracket is invariant under canonical transformation (q, p) to (Q, P), we mean that

$$[F, G]_{q,p} = [F, G]_{Q,P} \tag{7.5.15}$$

In order to prove this, let us start from the definition of Poisson bracket, i.e.,

$$[F, G]_{q,p} = \sum_k \left(\frac{\partial F}{\partial q_k} \frac{\partial G}{\partial p_k} - \frac{\partial F}{\partial p_k} \frac{\partial G}{\partial q_k} \right) \tag{7.5.16}$$

As $p_k = P_k\left(Q_1, Q_2, \ldots, Q_k, \ldots, P_1, P_2, \ldots, P_k, \ldots,\right)$ and $q_k = q_k\left(Q_1, Q_2, \ldots, Q_k, \ldots, P_1, P_2, \ldots, P_k, \ldots,\right)$ we can write

$$[F, G]_{q,p} = \sum_k \sum_l \left[\frac{\partial F}{\partial q_k} \left(\frac{\partial G}{\partial Q_l} \frac{\partial Q_l}{\partial p_k} + \frac{\partial G}{\partial P_l} \frac{\partial P_l}{\partial p_k} \right) - \frac{\partial F}{\partial p_k} \left(\frac{\partial G}{\partial Q_l} \frac{\partial Q_l}{\partial q_k} + \frac{\partial G}{\partial P_l} \frac{\partial P_l}{\partial p_k} \right) \right]$$

$$\Rightarrow [F, G]_{q,p} = \sum_l \left(\frac{\partial G}{\partial Q_l} [F, Q_l]_{q,p} + \frac{\partial G}{\partial P_l} [F, P_l]_{q,p} \right) \tag{7.5.17}$$

In Eq. (7.5.17), substituting $F = Q_i$ and $G = F$, we get

$$[Q_i, F]_{q,p} = \sum_l \left(\frac{\partial F}{\partial Q_l} [Q_i, Q_l]_{q,p} + \frac{\partial F}{\partial P_l} [Q_i, P_l]_{q,p} \right) = \frac{\partial F}{\partial P_i}. \tag{7.5.18}$$

Similarly, substituting $F = P_i$ and $G = F$ in (7.5.17), we have

$$\{P_i, F]_{q,p} = \sum_l \left(\frac{\partial F}{\partial Q_l} [P_i, Q_l]_{q,p} + \frac{\partial F}{\partial P_l} [P_i, P_l]_{q,p} \right) = -\frac{\partial F}{\partial Q_i} \tag{7.5.19}$$

Substituting (7.5.18) and (7.5.19) in (7.5.17), we obtain

$$[F, G]_{q,p} = \sum_l \left(-\frac{\partial G}{\partial Q_l} \frac{\partial F}{\partial P_l} + \frac{\partial G}{\partial P_l} \frac{\partial F}{\partial Q_l} \right) = [F, G]_{Q,P} \tag{7.5.20}$$

This proves the statement (7.5.15). Thus for the canonical variables, we can drop the subscripts of Poisson brackets.

7.6 Invariance of Lagrange's Bracket with respect to Canonical Transformations

According to the Poincare theorem, the integral

$$I_1 = \iint_S \sum_k dq_k \, dp_k \tag{7.6.1}$$

taken over an arbitrary two dimensional surface S of $2n$ dimensional phase space (q_k, p_k) is invariant under canonical transformation, i.e.,

$$\iint_S \sum_k dq_k \, dp_k = \iint_S \sum_k dQ_k \, dP_k \tag{7.6.2}$$

The position of a point on any two dimensional surface can be completely specified by two parameters, say u and v, so that

$$q_k = q_k(u, v) \quad \text{and} \quad p_k = p_k(u, v) \tag{7.6.3}$$

Transforming the area element in terms of new variables (u, v) by means of Jacobian, we have

$$dq_k\, dp_k = \frac{\partial\,(q_k, p_k)}{\partial(u, v)} du\, dv \tag{7.6.4}$$

with

$$\frac{\partial\,(q_k, p_k)}{\partial(u, v)} = \begin{vmatrix} \dfrac{\partial q_k}{\partial u} & \dfrac{\partial p_k}{\partial u} \\ \dfrac{\partial q_k}{\partial v} & \dfrac{\partial p_k}{\partial v} \end{vmatrix} \tag{7.6.5}$$

as the Jacobian. Eq. (7.6.2) in view of Eq. (7.6.4) is obtained to be

$$\iint_S \sum_k \frac{\partial\,(q_k, p_k)}{\partial(u, v)}\, du\, dv = \iint_S \sum_k \frac{\partial\,(Q_k, P_k)}{\partial(u, v)}\, du\, dv \tag{7.6.6}$$

As the surface S is arbitrary, area $du\, dv$ is arbitrary and therefore the expressions on both sides of Eq. (7.6.6) will be equal in the condition that the integrals are equal, i.e.,

$$\sum_k \frac{\partial(q_k, p_k)}{\partial(u, v)} = \sum_k \frac{\partial(Q_k, P_k)}{\partial(u, v)} \Rightarrow \sum_k \begin{vmatrix} \dfrac{\partial q_k}{\partial u} & \dfrac{\partial p_k}{\partial u} \\ \dfrac{\partial q_k}{\partial v} & \dfrac{\partial p_k}{\partial v} \end{vmatrix} = \sum_k \begin{vmatrix} \dfrac{\partial Q_k}{\partial u} & \dfrac{\partial P_k}{\partial u} \\ \dfrac{\partial Q_k}{\partial v} & \dfrac{\partial P_k}{\partial v} \end{vmatrix}$$

$$\Rightarrow \sum_k \left(\frac{\partial q_k}{\partial u}\frac{\partial p_k}{\partial v} - \frac{\partial q_k}{\partial v}\frac{\partial p_k}{\partial u} \right) = \sum_k \left(\frac{\partial Q_k}{\partial u}\frac{\partial P_k}{\partial v} - \frac{\partial Q_k}{\partial v}\frac{\partial P_k}{\partial u} \right) \Rightarrow \{u, v\}_{q,p} = \{u, v\}_{Q,P}$$

Thus, Lagrange's bracket is invariant under canonical transformation. Therefore it is immaterial which set of canonical coordinates is to be used i.e., that subscripts q, p can be dropped in writing Lagrange's brackets.

Example 7.6.1. Show that transformation defined by $q = \sqrt{2P}\sin Q, \quad p = \sqrt{2P}\cos Q$ is canonical by using Poisson bracket.

Solution. The transformation is

$$q = \sqrt{2P}\sin Q, \quad p = \sqrt{2P}\cos Q$$

From these equations, we can write the transformation as

$$\tan Q = \frac{q}{p} \quad \text{and} \quad P = \frac{1}{2}\left(q^2 + p^2\right) \tag{7.6.7}$$

In order to show that the given transformation is canonical, the Poisson bracket conditions are

$$[Q, Q] = [P, P] = 0 \quad \text{and} \quad [Q, P] = 1 \tag{7.6.8}$$

Here, $[Q, Q] = \dfrac{\partial Q}{\partial q}\dfrac{\partial Q}{\partial p} - \dfrac{\partial Q}{\partial p}\dfrac{\partial Q}{\partial q} = 0$. Similarly, $[P, P] = 0$. Also

$$[Q, P] = \frac{\partial Q}{\partial q}\frac{\partial P}{\partial p} - \frac{\partial Q}{\partial p}\frac{\partial P}{\partial q} \tag{7.6.9}$$

But from (7.6.7),

$$\sec^2 Q\frac{\partial Q}{\partial q}=\frac{1}{p},\quad \frac{\partial P}{\partial p}=p,\quad \sec^2 Q\frac{\partial Q}{\partial p}=-\frac{q}{p^2},\quad \frac{\partial P}{\partial q}=q$$

Substituting these values in (7.6.9), we get

$$\begin{aligned}[Q,P] &= \frac{\cos^2 Q}{p}p+\frac{q\cos^2 Q}{p^2}q=\cos^2 Q+\frac{q^2}{p^2}\cos^2 Q\\ &= \cos^2 Q\left[1+\frac{q^2}{p^2}\right]=\cos^2 Q[1+\tan^2 Q]=\cos^2 Q\sec^2 Q=1\end{aligned}$$

Thus we prove the condition (7.6.8) which means that the given transformation is canonical.

7.7 Jacobi's identity

For any three functions F, G and K of p_k and q_k, the following relation holds true:

$$[F,[G,K]]+[G,[K,F]]+[K,[F,G]]=0$$

This relation is known as Jacobi's identity.

Proof. Let us consider the expression for the following:

$$\begin{aligned}&[F,[G,K]]-[G,[F,K]]\\ &=\left[F,\sum_k\left(\frac{\partial G}{\partial q_k}\frac{\partial K}{\partial p_k}-\frac{\partial G}{\partial p_k}\frac{\partial K}{\partial q_k}\right)\right]-\left[G,\sum_k\left(\frac{\partial F}{\partial q_k}\frac{\partial K}{\partial p_k}-\frac{\partial F}{\partial p_k}\frac{\partial K}{\partial q_k}\right)\right]\\ &=\left[F,\sum_k\left(\frac{\partial G}{\partial q_k}\frac{\partial K}{\partial p_k}\right)\right]-\left[F,\sum_k\left(\frac{\partial G}{\partial p_k}\frac{\partial K}{\partial q_k}\right)\right]-\left[G,\sum_k\left(\frac{\partial F}{\partial q_k}\frac{\partial K}{\partial p_k}\right)\right]+\left[G,\sum_k\left(\frac{\partial F}{\partial p_k}\frac{\partial K}{\partial q_k}\right)\right]\end{aligned}$$

Now, using the property $[F,GK]=[F,G]K+[F,K]G$, we have

$$\begin{aligned}[F,[G,K]]-[G,[F,K]] &= \left[F,\sum_k\frac{\partial G}{\partial q_k}\right]\sum_k\frac{\partial K}{\partial p_k}+\left[F,\sum_k\frac{\partial K}{\partial p_k}\right]\sum_k\frac{\partial G}{\partial q_k}-\left[F,\sum_k\frac{\partial G}{\partial p_k}\right]\sum_k\frac{\partial K}{\partial q_k}\\ &\quad-\left[F,\sum_k\frac{\partial K}{\partial q_k}\right]\sum_k\frac{\partial G}{\partial p_k}-\left[G,\sum_k\frac{\partial F}{\partial q_k}\right]\sum_k\frac{\partial K}{\partial p_k}-\left[G,\sum_k\frac{\partial K}{\partial p_k}\right]\sum_k\frac{\partial F}{\partial q_k}\\ &\quad+\left[G,\sum_k\frac{\partial F}{\partial p_k}\right]\sum_k\frac{\partial K}{\partial q_k}+\left[G,\sum_k\frac{\partial K}{\partial q_k}\right]\sum_k\frac{\partial F}{\partial p_k}\\ &= \sum_k\left\{-\frac{\partial K}{\partial q_k}\left(\left[\frac{\partial F}{\partial p_k},G\right]+\left[F,\frac{\partial G}{\partial p_k}\right]\right)+\frac{\partial K}{\partial p_k}\left(\left[\frac{\partial F}{\partial q_k},G\right]+\left[F,\frac{\partial G}{\partial q_k}\right]\right)\right\}\\ &\quad+\sum_k\left\{\frac{\partial G}{\partial q_k}\left[F,\frac{\partial K}{\partial p_k}\right]-\frac{\partial G}{\partial p_k}\left[F,\frac{\partial K}{\partial q_k}\right]-\frac{\partial F}{\partial q_k}\left[G,\frac{\partial K}{\partial p_k}\right]+\frac{\partial F}{\partial p_k}\left[G,\frac{\partial K}{\partial q_k}\right]\right\}\end{aligned}$$

Using the identity $\frac{\partial}{\partial x}[F,G]=\left[\frac{\partial F}{\partial x},G\right]+\left[F,\frac{\partial G}{\partial x}\right]$, we obtain

$$[F,[G,K]]-[G,[F,K]]=\sum_k\left[-\frac{\partial K}{\partial q_k}\frac{\partial}{\partial p_k}[F,G]+\frac{\partial K}{\partial p_k}\frac{\partial}{\partial q_k}[F,G]\right]+0=-[K,[F,G]]$$

Thus, $[F,[G,K]]+[G,[K,F]]+[K,[F,G]]=0$, which proves the Jacobi's identity. □

Example 7.7.1. Show that the Poisson bracket of two constants of motion is itself a constant of motion.

Solution: In Jacobi's identity, we put $K = H$, then

$$[F,[G,H]]+[G,[H,F]]+[H,[F,G]]=0$$

Now, if F and G are constants of motion, then $[F,H]=0$ and $[G,H]=0$ Therefore, $[H,[F,G]]=0$ which means that the dynamic variable $[F,G]$ is constant of motion. Thus the Poisson bracket of two constants of motion is itself a constant of motion.

7.8 Hamilton-Jacobi Equation

If we make a canonical transformation from the old set of variables (q_k, p_k) to a new set of variables (Q_k, P_k), then the new equations of motion are,

$$\dot{P}_k=-\frac{\partial H'}{\partial Q_k} \quad \text{and} \quad \dot{Q}_k=\frac{\partial H'}{\partial P_k} \tag{7.8.1}$$

Now, if we require that the transformed Hamiltonian H' is identically zero i.e., $H'=0$, then equations of motion (7.8.1) assume the form

$$\dot{P}_k=0 \text{ and } \dot{Q}_k=0 \Rightarrow P_k= \text{ constant and } Q_k= \text{ constant} \tag{7.8.2}$$

Thus the new coordinates and momenta are constants in time and they are cyclic. The new Hamiltonian H' is related to the old Hamiltonian H by the relation

$$H'=H+\frac{\partial F}{\partial t}$$

which will be zero only when F satisfies the relation

$$H\left(q_k,p_k,t\right)+\frac{\partial F}{\partial t}=0 \tag{7.8.3}$$

where $H\left(q_k,p_k,t\right)$ is written for $H\left(\dot{q}_1,q_2,\ldots,q_n,p_1,p_2,\ldots,p_n,t\right)$. For convenience, we take the generating function F as a function of the old coordinates q_k, the new constant momenta P_k and time t i.e.; $F_2\left(q_k,P_k,t\right)$. Then $p_k=\dfrac{\partial F_2}{\partial q_k}$. Therefore,

$$H\left(q_k,\frac{\partial F_2}{\partial q_k},t\right)+\frac{\partial F_2}{\partial t}=0 \tag{7.8.4}$$

Let us see what is the physical meaning of the generating function $F_2\left(q_k,P_k,t\right)$. The total time derivative of F_2 is

$$\frac{\partial F_2}{\partial t}=\sum_{k=1}^{n}\frac{\partial F_2}{\partial q_k}\dot{q}_k+\sum_{k=1}^{n}\frac{\partial F_2}{\partial P_k}\dot{P}_k+\frac{\partial F_2}{\partial t}$$

Here, $\dot{P}_k=0$, $\dfrac{\partial F_2}{\partial t}=-H$ and $\dfrac{\partial F_2}{\partial q_k}=p_k$. Therefore,

$$\frac{\partial F_2}{\partial t}=\sum_{k=1}^{n}p_k\dot{q}_k-H=L\Rightarrow F_2=\int L\,dt=S \tag{7.8.5}$$

where S is the familiar *action* of the system, known as the *Hamilton's principal function* in relation to the variational principle. Writing $F_2 = S$ in Eq. (7.8.4), we get

$$H\left(q_k, \frac{\partial S}{\partial q_k}, t\right) + \frac{\partial S}{\partial t} = 0 \tag{7.8.6}$$

This is known as the *Hamilton-Jacobi equation* which is a partial differential equation of first order in $(n+1)$ variables $q_1, q_2, \ldots, q_n, t$.

Let the complete solution of equation (7.8.6) be of the form

$$S = S\left(q_1, q_2, \ldots, q_n, \alpha_1, \alpha_2, \ldots, \alpha_n, t\right) \tag{7.8.7}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are n independent constants of integration. Here, we have omitted one arbitrary additive constant which has no importance in a generating function because only partial derivatives of the generating function appear in the transformation equations.

In Eq. (7.8.7), the solution S is a function of n coordinates q_k, time t and n independent constants. We can take these n constants of integration as the new constant momenta i.e.,

$$P_k = \alpha_k \tag{7.8.8}$$

Now, the n transformation equations [Eqs. (6.3.17) of Unit 6] are

$$p_k = \frac{\partial S\left(q_1, \ldots, q_n, \alpha_1, \ldots, \alpha_n, t\right)}{\partial q_k} \tag{7.8.9}$$

These are n equations, which at $t = t_0$ (initially) give the n values of α_k in terms of the initial values of q_k and p_k. The other n transformation equations are

$$Q_k = \frac{\partial S}{\partial P_k} = \text{ constant, say } \beta_k$$

or

$$\beta_k = \frac{\partial S\left(q_1, \ldots, q_n, \alpha_1, \ldots, \alpha_n, t\right)}{\partial \alpha_k} \tag{7.8.10}$$

Similarly, one can calculate the constants β_k by using initial conditions i.e., at $t = t_0$, the known initial values of q_k, in Eq. (7.8.10). Thus α_k and β_k constants are known and Eq. (7.8.10) will give q_k in terms of α_k, β_k and t i.e.,

$$q_k = q_k\left(\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n, t\right) \tag{7.8.11}$$

After performing the differentiation in Eq. (7.8.9), Eq. (7.8.11) may be substituted for q_k to obtain momenta p_k. Thus p_k will be obtained as functions of constants α_k, β_k and time t i.e.,

$$p_k = p_k\left(\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_n, t\right) \tag{7.8.12}$$

In this way we obtain the desired complete solution of the mechanical problem. Thus we see that the Hamilton's principal function S is the generator of a canonical transformation to constant coordinates (β_k) and momenta (α_k). Also in solving the Hamilton-Jacobi equation, we obtain simultaneously a solution to the mechanical problem.

7.9 Solution of Harmonic Oscillator Problem by Hamilton-Jacobi Method

Let us consider a one-dimensional harmonic oscillator. The force acting on the oscillator at a displacement q is $F = -kq$, where k is force constant. Potential energy, $V = \int_0^q kq\, dq = \frac{1}{2}kq^2$ and Kinetic energy, $T = \frac{1}{2}mv^2 = \frac{p^2}{2m}$. Therefore, Hamiltonian, $H = T + V = \frac{p^2}{2m} + \frac{1}{2}kq^2$. But $p = \frac{\partial S}{\partial q}$, therefore,

$$H = \frac{1}{2}\left[\frac{\partial S}{\partial q}\right]^2 + \frac{1}{2}kq^2 \tag{7.9.1}$$

Hence the Hamilton-Jacobi equation corresponding to this Hamiltonian is

$$\frac{1}{2m}\left[\frac{\partial S}{\partial q}\right]^2 + \frac{1}{2}kq^2 + \frac{\partial S}{\partial t} = 0 \tag{7.9.2}$$

As the explicit dependence of S on t is involved only in the last term of left hand side of Eq. (7.9.2), a solution to this equation can be assumed in the form

$$S_i = S_1(q) + S_2(t) \tag{7.9.3}$$

Thus

$$\frac{1}{2m}\left[\frac{\partial S_1}{\partial q}\right]^2 + \frac{1}{2}kq^2 = -\frac{\partial S_2}{\partial t} \tag{7.9.4}$$

Setting each side of Eq. (7.9.4) equal to a constant, say α, we get

$$\frac{1}{2m}\left[\frac{\partial S_1}{\partial q}\right]^2 + \frac{1}{2}kq^2 = \alpha \quad \text{and} \quad -\frac{\partial S_2}{\partial t} = \alpha$$

$$\text{So that} \quad \frac{\partial S_1}{\partial q} = \sqrt{2m\left(\alpha - \frac{1}{2}kq^2\right)} \quad \text{and} \quad -\frac{\partial S_2}{\partial t} = \alpha$$

Integrating, we get

$$S_1 = \int \sqrt{2m\left(\alpha - \frac{1}{2}kq^2\right)}dq + C_1 \quad \text{and} \quad S_2 = -\alpha t + C_2$$

Therefore,

$$S = \int \sqrt{2m\left(\alpha - \frac{1}{2}kq^2\right)}dq - \alpha t + C$$

where $C = (C_1 + C_2)$ the constant of integration. It is to be noted that C is an additive constant and will not affect the transformation, because to obtain the new position coordinate ($Q = \partial S/\partial P$ or $\beta = \partial S/\partial \alpha$) only partial derivative of S with respect to $\alpha (= P$, new momentum) is required. This is why this additive constant C has no effect on transformation and is dropped. Thus

$$S = \int \sqrt{2m\left(\alpha - \frac{1}{2}kq^2\right)}dq - \alpha t \tag{7.9.5}$$

In this expression for the Hamilton's principal function S, first part is the function of α and q and is denoted as $W(q, \alpha)$. This is called *Hamilton's characteristic function*. Thus $S = W(q, \alpha) - \alpha t$.

We designate the constant α as the new momentum P. The new constant coordinate $(Q = \beta)$ is obtained by the transformation

$$\beta = \frac{\partial S}{\partial \alpha} = \frac{\sqrt{2m}}{2} \int \frac{dq}{\sqrt{\alpha - \frac{1}{2}kq^2}} - t = \sqrt{\frac{m}{2\alpha}} \int \frac{dq}{\sqrt{1 - \frac{kq^2}{2\alpha}}} - t$$

or

$$\beta = \sqrt{\frac{m}{k}} \sin^{-1} q\sqrt{\frac{k}{2\alpha}} - t$$

Therefore, $\sqrt{\frac{m}{k}} \sin^{-1} q\sqrt{\frac{k}{2\alpha}} = t + \beta$ or $\sin^{-1} q\sqrt{\frac{k}{2\alpha}} = \sqrt{\frac{k}{m}}(t + \beta)$. Writing $\omega = \sqrt{k/m}$, we obtain

$$q = \sqrt{\frac{2\alpha}{m\omega^2}} \sin \omega(t + \beta) \tag{7.9.6}$$

which is the *familiar solution of the harmonic oscillator*. Now,

$$p = \frac{\partial S}{\partial q} = \sqrt{2m\left(\alpha - \frac{1}{2}kq^2\right)} = \sqrt{2m\alpha - m^2\omega^2 q^2} \tag{7.9.7}$$

Putting the value of q from (7.9.6), we get

$$p = \sqrt{2m\alpha\left(1 - \sin^2 \omega(t + \beta)\right)} \quad \text{or} \quad p = \sqrt{2m\alpha} \cos \omega(t + \beta) \tag{7.9.8}$$

The constants α and β are to be known from initial conditions. Suppose at $t = 0$, the particle is at rest, i.e., $p_0 = 0$ and it is at the displacement $q = q_0$, from the equilibrium position. Then from Eq. (7.9.7)

$$p_0 = 0 = \sqrt{2m\alpha - m^2\omega^2 q_0^2} \quad \text{or} \quad \alpha = \frac{1}{2}m\omega^2 q_0^2 = \frac{1}{2}kq_0^2 \tag{7.9.9}$$

Also $H' = H + \partial S/\partial t = H - \alpha = 0$ [$\because \partial S/\partial t = -\alpha$ from (7.9.4)]. This gives $H = \alpha$. But the system is conservative and hence $H = E$. Thus the *new canonical momentum* $(P = \alpha)$ *is identified as the total energy of the oscillator*.

Also from (7.9.9), $q_0 = \sqrt{2\alpha/m\omega^2}$ and hence the solution (7.9.6) takes the more familiar form

$$q = q_0 \sin \omega(t + \beta) \tag{7.9.10}$$

Also from (7.9.8) and (7.9.10) at $t = 0, \cos \omega\beta = 0$ and $\sin \omega\beta = 1$. Therefore, $\omega\beta = \pi/2$ or $\beta = \pi/2\omega$. Thus the *new constant canonical coordinate*, measures the initial phase angle and in the present initial conditions the initial phase $\omega\beta = \pi/2$. Therefore, Eq. (7.9.10) is

$$q = q_0 \cos \omega t \tag{7.9.11}$$

In view of Eq. (7.9.7), and then (7.9.8), Hamilton's principal function S from (7.9.5) is obtained to be

$$S = \int p\, dq - \alpha t = \int \sqrt{2m\alpha}\omega \cos \omega(t + \beta) q_0 \cos \omega(t + \beta) dt - \alpha t$$
$$= 2\alpha \int \cos^2 \omega(t + \beta) dt - \alpha t = 2\alpha \int \left[\cos^2 \omega(t + \beta) - \frac{1}{2}\right] dt.$$

The Lagrangian L is given by

$$\begin{aligned} L &= \frac{p^2}{2m} - \frac{1}{2}kq^2 = \alpha\cos^2\omega(t+\beta) - \frac{1}{2}kq_0^2\sin^2\omega(t+\beta) \\ &= \alpha\left[\cos^2\omega(t+\beta) - \sin^2\omega(t+\beta)\right] \quad \left(\text{as } \alpha = \frac{1}{2}kq_0^2\right) \\ &= 2\alpha\left\{\cos^2\omega(t+\beta) - \frac{1}{2}\right] \end{aligned}$$

Therefore,

$$S = \int L dt$$

Thus for harmonic oscillator we prove that the Hamilton's principal function is the time integral of Lagrangian.

Exercise 7.9.1.

1. The Lagrangian for a simple pendulum is given by $L = \frac{1}{2}ml^2\dot{\theta} - mgl(1-\cos\theta)$. Find the Poisson bracket between θ and $\dot{\theta}$.
2. Let q and p be the canonical coordinate and momentum of a dynamical system. Use the concept of Poisson bracket to show that the transformation $Q = \frac{1}{\sqrt{2}}(p+q)$ and $P = \frac{1}{\sqrt{2}}(p-q)$ is canonical.
3. If $[\alpha, \beta]$ is the Poisson bracket, prove that $\frac{\partial}{\partial t}\left[\alpha, \beta\right] = \left[\frac{\partial\alpha}{\partial t}, \beta\right] + \left[\alpha, \frac{\partial\beta}{\partial t}\right]$.
4. For a simple harmonic oscillator, the Lagrangian is given by $L = \frac{1}{2}\dot{q}^2 - \frac{1}{2}q^2$. If $A(p,q) = \frac{p+iq}{\sqrt{2}}$ and $H(p,q)$ is the Hamiltonian of the system, then show that the Poisson bracket $\{A(p,q),\ H(p,q)\}$ is given by $iA(p,q)$.
5. The coordinates and momenta $x_i,\ P_i\ (i = 1,2,3)$ of a particle satisfy the canonical Poisson bracket relations $\{x_i, p_j\} = \delta_{ij}$. If $C_1 = x_2p_3 + x_3p_2$ and $C_2 = x_1p_2 - x_2p_1$ are constants of motion, and if $C_3\{C_1, C_2\} = x_1p_3 + x_3p_1$, then show that $\{C_2, C_3\} = C_1$ and $\{C_3, C_1\} = -C_2$.

Unit 8

Course Structure

- Small Oscillations: General case of coupled oscillations.
- Eigen vectors and Eigen frequencies.
- Orthogonality of Eigen vectors.
- Normal coordinates. Two-body problem.

8.1 Introduction

The theory of small oscillations about the equilibrium position is of importance in molecular spectra, acoustics, vibrations of atoms in solids, vibrations of coupled mechanical systems and coupled electrical circuits. If the displacement from the stable equilibrium conditions are small, the motion can be described as that of a system of coupled linear harmonic oscillators with each generalized co-ordinate expressed as a function of the different frequencies of vibrations of the system. The problem can be simplified further by a transformation of the generalized co-ordinates to another set of co-ordinates, each of which undergoes periodic changes with a well-defined single frequency. In this unit we develop a theory of small oscillations based on Lagrangian formulation.

8.2 General Theory of Small Oscillations

The potential energy of a conservative system, specified by n generalized coordinates $q_1, q_2, \ldots, q_n$, is represented as

$$V = V(q_1, q_2, \ldots, q_n) \tag{8.2.1}$$

We are interested in the motion of the system, when the displacements of the particles are small from the position of stable equilibrium. We denote the displacements of the generalized coordinates from equilibrium position by u_i, i.e.,

$$q_i = q_i^0 + u_i \tag{8.2.2}$$

Since q_i^0 is fixed, u_i may be taken as new generalized coordinates of the motion. Expanding the potential energy about the position of equilibrium, we obtain

$$V(q_1,\ldots,q_n) = V\left(q_1^0, q_2^0, \ldots, q_n^0\right) + \sum_{i=1}^{n}\left[\frac{\partial V}{\partial q_i}\right]_0 \left(q_i - q_i^0\right) + \frac{1}{2!}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{\partial^2 V}{\partial q_i \partial q_j}\right]_0 \left(q_i - q_i^0\right)\left(q_j - q_j^0\right) + \ldots \tag{8.2.3}$$

In consequence of equilibrium, $(\partial V/\partial q_i)_0 = 0$. First term in the expansion represents the potential energy in the equilibrium position and is constant for the system. Assuming the potential energy in the equilibrium to be zero and writing $u_i = q_i - q_i^0$ and $u_j = q_j - q_j^0$, we get

$$V = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} V_{ij} u_i u_j \tag{8.2.4}$$

where $V_{ij} = \left[\dfrac{\partial^2 V}{\partial q_i \partial q_j}\right]_0 = \left[\dfrac{\partial^2 V}{\partial u_i \partial u_j}\right]_0$ = constant which is to be evaluated at $q_i = q_i^0$ and $q_j = q_j^0$.

The constant $V_{ij} = V_{ji}$ form a symmetric matrix V. In Eq. (8.2.4), we retain the terms quadratic in the coordinates. The kinetic energy of the system is given by

$$T = \sum_i \sum_j \frac{1}{2} m_{ij} \dot{q}_i \dot{q}_j = \sum_i \sum_j \frac{1}{2} m_{ij} \dot{u}_i \dot{u}_j \tag{8.2.5}$$

because the generalized coordinates do not involve time explicitly and therefore the kinetic energy is a homogeneous quadratic function of generalized velocities. The coefficients are, in general, functions of generalized coordinates and therefore expanding m_{ij} in Taylor's series, we get

$$m_{ij}(q_1,\ldots,q_n) = m_{ij}\left(q_0^1,\ldots,q_0^n\right) + \sum_{k=1}^{n}\left[\frac{\partial m_{ij}}{\partial q_k}\right]_0 u_k + \ldots \tag{8.2.6}$$

In Eq. (8.2.5), the term is already quadratic in the u_i 's, we obtain the lowest non-vanishing approximation to T in quadratic form only by retaining the first term in the expansion. If the constant values of the function m_{ij} are denoted by T_{ij}, then the kinetic energy is

$$T = \frac{1}{2}\sum_i \sum_j T_{ij} \dot{u}_i \dot{u}_j \tag{8.2.7}$$

Obviously the constants T_{ij} are elements of symmetric matrix T. Now, the Lagrangian $L(= T - V)$ can be written as

$$T = \frac{1}{2}\sum_i \sum_j \left[T_{ij}\dot{u}_i \dot{u}_j - V_{ij} u_i u_j\right] \tag{8.2.8}$$

Using u_i 's as generalized coordinates, the Lagrange's equations $\dfrac{d}{dt}\left[\dfrac{\partial L}{\partial \dot{q}_i}\right] - \dfrac{\partial L}{\partial q_i} = 0$ take the form

$$\sum_{j=1}^{n}\left[T_{ij}\ddot{u}_j + V_{ij} u_j\right] = 0 \tag{8.2.9}$$

For $i = 1, 2, \ldots, n$, Eqs. (8.2.9) represent n equations which are to be solved to obtain the motion near the position of equilibrium.

8.2.1 Secular Equation and Eigen value Equation

We try an oscillatory solution of Eq. (8.2.9) in the form

$$u_i = Ca_i e^{i\omega t} \tag{8.2.10}$$

where Ca_i is the complex amplitude of the oscillation for each coordinate u_i the factor C being used for convenience as a scale factor, the same for all the coordinates.

Substituting for u_j from Eq. (8.2.10) into Eq. (8.2.9), we obtain

$$\sum_{j=1}^{n} \left[V_{ij}a_j e^{i\omega t} - \omega^2 T_{ij}a_j e^{i\omega t}\right] = 0 \quad \text{or} \quad e^{i\omega t}\sum_{j=1}^{n}\left[V_{ij}a_j - \omega^2 T_{ij}a_j\right] = 0$$

In general, $e^{i\omega t}$ is not zero, hence

$$\sum_{j=1}^{n}\left[V_{ij}a_j - \omega^2 T_{ij}a_j\right] = 0 \tag{8.2.11}$$

or in matrix form

$$Va - \omega^2 Ta = 0 \tag{8.2.12}$$

where the matrix V, T and a are

$$V = \begin{bmatrix} V_{11} & V_{12} & \cdots & V_{1n} \\ V_{21} & V_{22} & \cdots & V_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ V_{n1} & V_{n2} & \cdots & V_{nn} \end{bmatrix} \quad T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1n} \\ T_{21} & T_{22} & \cdots & T_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ T_{n1} & T_{n2} & \cdots & T_{nn} \end{bmatrix} \quad \text{and} \quad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Eqs. (8.2.11)/(8.2.12) represent n tinear, homogeneous, algebraic equations in a_j and ω, i.e.,

$$\begin{aligned} &\left[V_{11} - \omega^2 T_{11}\right] a_1 + \left[V_{12} - \omega^2 T_{12}\right] a_2 + \cdots + \left[V_{1n} - \omega^2 T_{1n}\right] a_n = 0 \\ &\vdots \quad \vdots \\ &\left[V_{n1} - \omega^2 T_{n1}\right] a_1 + \left[V_{n2} - \omega^2 T_{n2}\right] a_2 + \cdots + \left[V_{nn} - \omega^2 T_{nn}\right] a_n = 0 \end{aligned} \tag{8.2.13}$$

Let us assume that inverse of T matrix exists. Multiplying Eq. (8.2.12) by T^{-1}, we get

$$T^{-1}Va - \omega^2 T^{-1}Ta = 0$$

Since $T^{-1}T = I$, unit matrix and $T^{-1}V = P$ (say), then

$$Pa - \omega^2 Ia = 0 \quad \text{or} \quad \left(P - \omega^2 I\right) a = 0 \tag{8.2.14}$$

Eq. (8.2.14) is the *eigen value equation*. Here ω^2 are the eigenvalues of P and a is the eigenvector with n components.

8.2.2 Solution of the Eigenvalue Equation

The eigenvalues are obtained by solving the determinant

$$|P - \omega^2 I| = 0 \quad \text{or} \quad |V - \omega^2 T| = 0 \tag{8.2.15}$$

or

$$\begin{vmatrix} V_{11}-\omega^2 T_{11} & V_{12}-\omega^2 T_{12} & \cdots & V_{1n}-\omega^2 T_{1n} \\ V_{21}-\omega^2 T_{21} & V_{22}-\omega^2 T_{22} & \cdots & V_{2n}-\omega^2 T_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ V_{n1}-\omega^2 T_{n1} & V_{n2}-\omega^2 T_{n2} & \cdots & V_{nn}-\omega^2 T_{nn} \end{vmatrix} = 0 \tag{8.2.16}$$

Eq. (8.2.15)- (8.2.16) is called *secular equation*. This determinantal condition is in effect an algebraic equation of n-th degree for ω^2 and the roots of the determinant provide n frequencies $\left(\omega_1^2, \omega_2^2, \ldots, \omega_n^2\right)$. These values are the *normal mode frequencies*.

For each of the value of ω^2, say $\omega_k^2 (k = 1, 2, \ldots, n)$ in the k-th mode of vibration, Eqs. (8.2.13) may be solved for amplitudes a_i. Corresponding to this ω_k^2, we denote the amplitudes by a_{ik} $(i = 1, 2, \ldots, n)$. Thus a_{ik} is the amplitude in the k-th mode of the i-th coordinate. Only for $\omega_k^2 > 0$, the motion is oscillatory about the position of stable equilibrium.

In order to find the amplitudes a_{ik}, we use eqs. (8.2.14) for a particular value of ω, say ω_1 and then we know $a_{11}, a_{21}, \ldots, a_{n1}$. Similarly for $\omega_2, a_{12}, a_{22}, \ldots, a_{n2}$ and for $\omega_n, a_{1n}, a_{2n}, \ldots, a_{nn}$ are known. More correctly speaking, we may find $n-1$ amplitudes for a particular frequency. For example, for frequency ω_k, we can determine all the amplitudes except one, say $a_{2k}, a_{3k}, \ldots, a_{nk}$ except a_{1k}. In other words, we may determine the coefficients a_{ik} in terms of a_{1k} in the form of ratios :

$$\frac{a_{2k}}{a_{1k}}, \frac{a_{3k}}{a_{1k}}, \cdots, \frac{a_{nk}}{a_{1k}} \tag{8.2.17}$$

A general solution of equation of motion (8.2.11)/(8.2.12)/(8.2.13) involves a superposition of oscillations with all the permitted frequencies. Thus if the system is displaced slightly from the equilibrium position and then released, it performs small amplitude oscillations about the equilibrium position with frequencies $\omega_1, \omega_2, \ldots, \omega_n$. The solutions of the secular equation (8.2.15)/(8.2.16) are therefore often called as the *frequencies of free vibrations* or as the *resonant frequencies* of the system.

The general solution may now be written as

$$u_i = \sum_{k=1}^{n} C_k a_{ik} e^{i\omega_k t} \tag{8.2.18}$$

where we have used index k for summation for displacements due to all the allowed frequencies. Corresponding to the normal frequency ω_k (k-th mode of vibration), the eigenvector is a_k with n components given by the matrix

$$a_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{bmatrix} \tag{8.2.19}$$

For the oscillating system, there are n eigen vectors $a_1, a_2, \ldots, a_k, \ldots, a_n$, where a_k is given by (8.2.19). Thus in all there are $n \times n$ eigenvector components for the system, which may be represented by the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \tag{8.2.20}$$

Obviously one may say that for each solution ω_k^2 of the secular equation (8.2.15), there are two resonant frequencies $+\omega_k$ and $-\omega_k$. The eigenvector a_k is the same for the two frequencies, but the scaling factors C_k^+ and C_k^- may be much different. Thus the general solution should be

$$u_i = \sum_{k=1}^{n} a_{ik} \left[C_k^+ e^{+i\omega_k t} + C_k^- e^{-i\omega_k t} \right] \tag{8.2.21}$$

The actual motion is the real part of the complex solution (8.2.21) which can be expressed as

$$u_i = \sum_{k=1}^{n} f_k a_{ik} \cos\left(\omega_k t + \phi_k\right) \tag{8.2.22}$$

where f_k and ϕ_k are determined from initial conditions.

8.2.3 Small Oscillations in Normal Coordinates

Let us define

$$u_i = \sum_{k=1}^{n} a_{ik} Q_k \tag{8.2.23}$$

In terms of single column matrices

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{bmatrix}$$

we have,

$$u = AQ \tag{8.2.24}$$

The potential energy V can be written as

$$V = \frac{1}{2} \sum_i \sum_j V_{ij} u_i u_j = \frac{1}{2} \sum_i \sum_j u_i V_{ij} u_j \quad \text{or} \quad V = \frac{1}{2} u^T V u \tag{8.2.25}$$

where u^T is the transpose of u or single row matrix. From Eq. (8.2.24)

$$u = (AQ)^T = Q^T A^T$$

Therefore

$$V = \frac{1}{2} Q^T A^T V A Q \tag{8.2.26}$$

The kinetic energy K similarly is

$$T = \sum_i \sum_j \dot{u}_i T_{ij} \dot{u}_j = \frac{1}{2} \dot{Q} A^T T A \dot{Q} \tag{8.2.27}$$

From Eq. (8.2.11), writing $\omega_k^2 = \lambda_k$,

$$\sum_{j=1}^{n} \left[V_{ij} a_{jk} - \lambda_k T_{ij} a_{jk} \right] = 0 \tag{8.2.28}$$

The complex conjugate of this equation is

$$\sum_{i=1}^{n} [V_{ij} a_{il}^* - \lambda_i^* T_{ij} a_{il}^*] = 0 \tag{8.2.29}$$

As a_{ij} are real, we eliminate V_{ij} from (8.2.28) and (8.2.29) by multiplying the former by a_{il} and summing over i and the latter by a_{jk} and summing over j. Thus

$$(\lambda_k - \lambda_l^*) \sum_i \sum_j a_{jk} T_{ij} a_{il} = 0 \tag{8.2.30}$$

If all λ_k are distinct, i.e., $(\lambda_k - \lambda_l^*)$ is not zero, then

$$\sum_i \sum_j a_{jk} T_{ij} a_{il} = 0 \tag{8.2.31}$$

The coefficients a_{jk} in eq. (8.2.28) cannot be completely determined, because this is a set of linear equations. This indeterminacy can be removed by requiring that

$$\sum_i \sum_j a_{jk} T_{ij} a_{ik} = 1 \tag{8.2.32}$$

The two equations (8.2.31) and (8.2.32) can be combined into one by means of Kronecker delta symbol δ_{kl}, i.e.,

$$\sum_i \sum_j a_{jk} T_{ij} a_{il} = \delta_{kl} \tag{8.2.33}$$

Eqs. (8.2.31) and (8.2.32) can be written as

$$A^T T A = I \tag{8.2.34}$$

Writing $\lambda_l = \lambda_k \delta_{lk}$, we obtain from eq. (8.2.28)

$$\sum_{j=1}^{n} V_{ij} a_{jk} = \sum_{j=1}^{n} T_{ij} a_{jk} \lambda_k \delta_{lk} \tag{8.2.35}$$

which is in matrix notation

$$VA = TA\lambda \tag{8.2.36}$$

Multiplying by A^T from left, we get

$$A^T V A = A^T T A \lambda \tag{8.2.37}$$

But $A^T T A = I$ [eq. (8.2.34)],

$$A^T V A = \lambda \tag{8.2.38}$$

In view of eq. (8.2.38), eq. is obtained to be

$$\begin{aligned} V &= \frac{1}{2} Q^T \lambda Q = \frac{1}{2} (Q_1, Q_2 \ldots, Q_n) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} \cdot \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_n \end{pmatrix} \\ &= \frac{1}{2} \left(\lambda_1 Q_1^2 + \lambda_2 Q_2^2 + \ldots + \lambda_n Q_n^2 \right) \\ &= \frac{1}{2} \sum_{k=1}^{n} \lambda_k Q_k^2 = \frac{1}{2} \sum_{k=1}^{n} \omega_k^2 Q_k^2 \end{aligned} \tag{8.2.39}$$

Similarly from eqs. (8.2.27) and (8.2.34), we have

$$T = \frac{1}{2}\dot{Q}^T I \dot{Q} = \frac{1}{2}\sum_{k=1}^{n} \dot{Q}_k^2 \tag{8.2.40}$$

We see from eqs. (8.2.39) and (8.2.40) that in the new coordinates, both the potential and kinetic energies are the sums of squares only without any cross terms.

Now, the Lagrangian $L = T - V$ is

$$L = \frac{1}{2}\sum_{k=1}^{n} \dot{Q}_k^2 - \frac{1}{2}\sum_{k=1}^{n} \omega_k^2 Q_k^2 \tag{8.2.41}$$

Hence the Lagrangian equations

$$\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{Q}_k}\right] - \frac{\partial L}{\partial Q_k} = 0$$

for the new coordinates are

$$\ddot{Q}_k + \omega_k^2 Q_k = 0 \tag{8.2.42}$$

which are n equations for $k = 1, 2, \ldots, n$.

Thus each new coordinate executes simple harmonic motion with a single frequency and therefore, Q_1, $Q_2, \ldots, Q_n$ are called *normal coordinates*. The frequencies $\omega_1, \omega_2, \ldots, \omega_n$ are referred as *normal frequencies*. The solution of eq. (8.2.42) is

$$Q_k = f_k \cos(\omega_k t + \phi_k) \tag{8.2.43}$$

From eqs. (8.2.43) and (8.2.23), we see,

$$u_i = \sum_{k=1}^{n} a_{lk} Q_k = \sum_{k=1}^{n} f_k a_{ik} \cos(\omega_k t + \phi_k) \tag{8.2.44}$$

Thus (8.2.43) could have been obtained directly from (8.2.22) and (8.2.23).

It may be reminded again that each normal coordinate corresponds to a vibration of the system with only one frequency and these component oscillations are called as the *normal modes of vibration*. *In each mode all the particles vibrate with the same frequency and with the same phase (the particle may be out of phase, if the a's have opposite sign)*, the relative amplitudes being determined by the matrix elements a_{ik}. The complete motion is then composed of sum of the normal modes weighted with proper amplitude and phase factors contained in the scaling factor C_k's.

8.3 Two body problems

Here we discuss important examples of two coupled oscillators.

8.3.1 Two coupled pendulums

Consider two identical pendulums as shown in Fig 8.3.1. Each pendulum has a bob of mass m with an effective length l. The two bobs of the pendulums are connected by a light spring of force constant k. The relaxed length of the spring is equal to the distance between the two bobs at equilibrium. We shall consider small

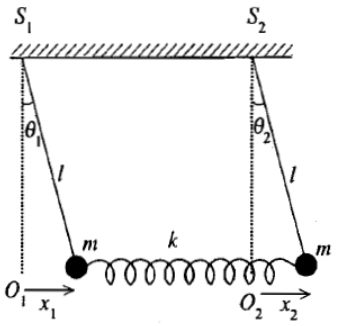


Figure 8.3.1: Two coupled pendulums

amplitude oscillations, restricted to the plane in equilibrium configuration. Thus the system of two coupled pendulums, under consideration, has two degrees of freedom.

Let the system of two coupled pendulums be allowed to oscillate so that x_1 and x_2 represent displacements from the equilibrium positions O_1 and O_2 respectively. If θ_1 and θ_2 be the angular displacements at any instant t, then the potential energy of the system is given by

$$V = mgl\,(1-\cos\theta_1) + mgl\,(1-\cos\theta_2) + \frac{1}{2}k\,(x_1-x_2)^2$$

where the potential energy in the equilibrium configuration is assumed to be zero. For small amplitude oscillations.

$$1-\cos\theta_1 = 1-\left(1-\theta_1^2/2\right) = \theta_1^2/2 = x_1^2/2l^2$$

and similarly $1-\cos\theta_2 = x_2^2/2l^2$, where $\theta_1 = x_1/l$ and $\theta_2 = x_2/l$. Thus

$$V = \frac{1}{2}\frac{mg}{l}x_1^2 + \frac{1}{2}\frac{mg}{l}x_2^2 + \frac{1}{2}k\,(x_1-x_2)^2 \tag{8.3.1}$$

The kinetic energy of the system is

$$T = \frac{1}{2}m\dot{x}_1^2 + \frac{1}{2}m\dot{x}_2^2 \tag{8.3.2}$$

The V and T matrices for the system are

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}.$$

Here

$$V_{11} = \left[\frac{\partial^2 V}{\partial x_1^2}\right]_{x_1=0,x_2=0} = k + \frac{mg}{l}, V_{12} = \left[\frac{\partial^2 V}{\partial x_1 \partial x_2}\right]_{x_1=0,x_2=0} = -k,$$

$$V_{21} = \left[\frac{\partial^2 V}{\partial x_2 \partial x_1}\right]_{x_1=0,x_2=0} = -k, V_{22} = \left[\frac{\partial^2 V}{\partial x_2^2}\right]_{x_1=0,x_2=0} = k + \frac{mg}{l}$$

Since

$$T = \frac{1}{2}\left[T_{11}\dot{x}_1^2 + T_{12}\dot{x}_1\dot{x}_2 + T_{21}\dot{x}_1\dot{x}_2 + T_{22}\dot{x}_2^2\right]$$

$$T_{11} = m = T_{22} \text{ and } T_{12} = T_{21} = 0$$

Thus

$$V = \begin{bmatrix} k+\frac{mg}{l} & -k \\ -k & k+\frac{mg}{l} \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix} \tag{8.3.3}$$

The normal frequencies are determined from the equation

$$\begin{aligned} & V - \omega^2 T = 0 \\ \Rightarrow & \begin{vmatrix} k + \frac{mg}{l} - m\omega^2 & -k \\ -k & k + \frac{mg}{l} - m\omega^2 \end{vmatrix} = 0 \\ \text{or,} \quad & \left[k + \frac{mg}{l} - m\omega^2\right]^2 - k^2 = 0 \quad \text{or} \quad \left[\frac{mg}{l} - m\omega^2\right]\left[2k + \frac{mg}{l} - m\omega^2\right] = 0 \end{aligned} \tag{8.3.4}$$

which gives

$$\omega^2 = \omega_1^2 = \frac{g}{l} \text{ and } \omega^2 = \omega_2^2 = \frac{g}{l} + \frac{2k}{m} \quad \text{or,} \quad \omega_1 = \pm\sqrt{\frac{g}{l}} \text{ and } \omega_2 = \pm\sqrt{\frac{g}{l} + \frac{2k}{m}}$$

Thus the normal frequencies of the system are

$$\omega_1 = \sqrt{\frac{g}{l}} \text{ and } \omega_2 = \sqrt{\frac{g}{l} + \frac{2k}{m}} \tag{8.3.5}$$

To determine the eigenvectors, we use the equation

$$\begin{aligned} & \left[V - \omega_k^2 T\right] a_k = 0 \\ \text{or,} \quad & \begin{pmatrix} k + \frac{mg}{l} - m\omega_k^2 & -k \\ -k & k + \frac{mg}{l} - m\omega_k^2 \end{pmatrix} \begin{pmatrix} a_{1k} \\ a_{2k} \end{pmatrix} = 0. \end{aligned}$$

For

$$\begin{aligned} & \omega^2 = \omega_1^2 = g/l, \text{ we have} \\ & \begin{pmatrix} k & -k \\ -k & k \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = 0 \quad \text{or,} \quad \frac{a_{21}}{a_{11}} = 1 \end{aligned}$$

If $a_{11} = \alpha$, then $a_{21} = \alpha$. For $\omega^2 = \omega_2^2 = \dfrac{g}{l} + \dfrac{2k}{m}$, we have

$$\begin{pmatrix} -k & -k \\ -k & -k \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = 0 \quad \text{or} \quad \frac{a_{22}}{a_{12}} = -1$$

If $a_{12} = \beta$, then $a_{22} = -\beta$. Thus the eigenvectors are

$$a_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha \end{pmatrix} \quad \text{and} \quad a_2 = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = \begin{pmatrix} \beta \\ -\beta \end{pmatrix} \tag{8.3.6}$$

Now, the matrix A is

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \alpha & -\beta \end{pmatrix}$$

Therefore, the transpose of A matrix i.e., A^T is

$$A^T = \begin{bmatrix} \alpha & \alpha \\ \beta & -\beta \end{bmatrix}$$

We impose the condition $A^T T A = I$ i.e.,

$$\begin{pmatrix} \alpha & \alpha \\ \beta & -\beta \end{pmatrix} \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \alpha & -\beta \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

or

$$\begin{pmatrix} 2m\alpha^2 & 0 \\ 0 & 2m\beta^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

whence $\alpha = \beta = 1/\sqrt{2m}$.
Thus the eigenvecfors are

$$a_1 = \frac{1}{\sqrt{2m}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad a_2 = \frac{1}{\sqrt{2m}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{8.3.7}$$

The generalized coordinates x_1 and x_2 are related to normal coordinates Q_1 and Q_2 by using the relation :

$$u_i = \sum_{k=1}^{2} a_{ik} Q_k$$

where for $i = 1, 2, u_1 = x_1$ and $u_2 = x_2$. Therefore,

$$\begin{aligned} & x_1 = a_{11}Q_1 + a_{12}Q_2 \quad \text{and} \quad x_2 = a_{21}Q_1 + a_{22}Q_2 \\ \text{or,} \quad & x_1 = \frac{1}{\sqrt{2m}}Q_1 + \frac{1}{\sqrt{2m}}Q_2 \quad \text{and} \quad x_2 = \frac{1}{\sqrt{2m}}Q_1 - \frac{1}{\sqrt{2m}}Q_2 \end{aligned} \tag{8.3.8}$$

Hence the normal coordinates Q_1 and Q_2 are

$$Q_1 = \sqrt{2m}\,(x_1 + x_2) \quad \text{and} \quad Q_2 = \sqrt{2m}\,(x_1 - x_2) \tag{8.3.9}$$

Further the normal coordinates Q_1 oscillates with frequency ω_1 and Q_2 with ω_2. So that

$$Q_1 = f_1 \cos(\omega_1 t + \phi_1) \text{ and } Q_2 = f_2 \cos(\omega_2 t + \phi_2) \tag{8.3.10}$$

Thus

$$x_1 = \frac{f_1}{\sqrt{2m}} \cos(\omega_1 t + \phi_1) + \frac{f_2}{\sqrt{2m}} \cos(\omega_2 t + \phi_2) \tag{8.3.11}$$

and

$$x_2 = \frac{f_1}{\sqrt{2m}} \cos(\omega_1 t + \phi_1) - \frac{f_2}{\sqrt{2m}} \cos(\omega_2 t + \phi_2) \tag{8.3.12}$$

Putting $f_1/\sqrt{2m} = A_1$ and $f_2/\sqrt{2m} = A_2$, we get

$$x_1 = A_1 \cos(\omega_1 t + \phi_1) + A_2 \cos(\omega_2 t + \phi_2) \tag{8.3.13}$$

$$x_2 = A_1 \cos(\omega_1 t + \phi_1) - A_2 \cos(\omega_2 t + \phi_2) \tag{8.3.14}$$

Thus the displacement of a pendulum is obtained by the superposition of harmonic oscillations of ω_1 and ω_2 frequencies.

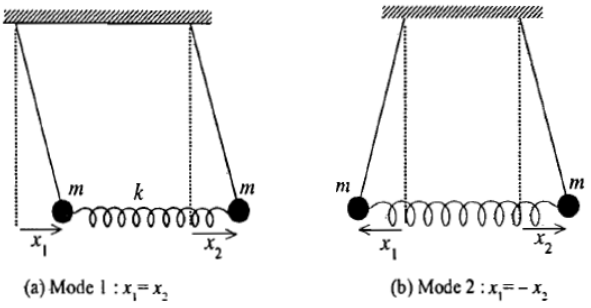
Eqs. (8.3.5), (8.3.7), (8.3.9) and (8.3.10) completely describe the motion. Eqs. (8.3.11), (8.3.12) [or (8.3.13), (8.3.14)] are the result of eqs. (8.3.9) and (8.3.10). If we put $Q_2 = 0$ (or f_2 or A_2), then from eq. (8.3.9)

$$x_1 = x_2$$

which means that in Mode 1 (Q_1), the two pendula oscillate with the same frequency $\omega_1 = \sqrt{g/l}$ in the same phase. If we put $Q_1 = 0$ (or f_1 or A_1), then from eq. (8.3.9), we have

$$x_1 = -x_2$$

This means that in Mode 2 (Q_2), the two pendula oscillate exactly out of phase (with a phase difference of π) with the same frequency $\omega_2 = \sqrt{\frac{g}{l} + \frac{2k}{m}}$. It is to be noted that in Mode 1 $(x_1 = x_2)$, there is no stretching or compression of the spring so that ω_1 does not depend on spring constant k, while in Mode 2, due to compression or stretching of the spring the force constant contributes in ω_2.



8.3.2 Double Pendulum

A double pendulum consists of a pendulum of mass m_1 and length l_1 to which a second pendulum of mass m_2 and length l_2 is suspended [Fig. 8.3.2]. The motion is considered in a plane so that the system has two degrees of freedom. If (x_1, y_1) and (x_2, y_2) be the coordinates of the masses m_1 and m_2 respectively, then from the figure, we have

$$x_1 = l_1 \sin\theta_1, \quad x_2 = l_1 \sin\theta_1 + l_2 \sin\theta_2$$
$$\text{and } y_1 = l_1 \cos\theta_1, \; y_2 = l_1 \cos\theta_1 + l_2 \cos\theta_2$$

where θ_1 and θ_2 are the angles, made by the lengths of the pendulums with the vertical. These are taken as generalized coordinates.

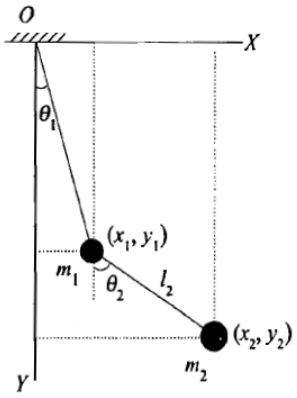


Figure 8.3.2: Double Pendulums

Thus the potential energy of the system is

$$V = -m_1 g y_1 - m_2 g y_2 = -m_1 g l_2 \cos\theta_1 - m_2 g(l_1 \cos\theta_1 + l_2 \cos\theta_2)$$

where the potential energy is considered to be zero at O. For small θ_1, $\cos\theta_1 = 1 - \theta_1^2/2, \;\; \cos\theta_2 = 1 - \theta_2^2/2$. Therefore,

$$V = -m_1 g l_1 - m_2 g(l_1 + l_2) + \frac{1}{2} m_1 g l_1 \theta_1^2 + \frac{1}{2} m_2 g l_2 \theta_2^2.$$

The V matrix is

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

Here

$$V_{11} = \left[\frac{\partial^2 V}{\partial \theta_1^2}\right]_{\theta_1=0,\theta_2=0} = (m_1+m_2)gl_1,\ V_{12} = V_{21} = 0,\ \text{and } V_{22} = \left[\frac{\partial^2 V}{\partial \theta_2^2}\right]_{\theta_1=0,\theta_2=0} = m_2 g l_2$$

$$\text{Therefore,}\quad V = \begin{bmatrix} (m_1+m_2)gl_1 & 0 \\ 0 & m_2 g l_2 \end{bmatrix} \tag{8.3.15}$$

The kinetic energy of the system is

$$T = \frac{1}{2}m_1\left(\dot{x}_1^2 + \dot{y}_1^2\right) + \frac{1}{2}m_2\left(\dot{x}_2^2 + \dot{y}_2^2\right)$$
$$T = \frac{1}{2}m_1 l_1^2\dot{\theta}_1^2 + \frac{1}{2}m_2\left[l_1^2\dot{\theta}_1^2 + l_2^2\dot{\theta}_2^2 + 2l_1 l_2\dot{\theta}_1\dot{\theta}_2\cos\left(\theta_1-\theta_2\right)\right]$$

As θ_1 and θ_2 are small $\cos\left(\theta_1-\theta_2\right) \simeq 1$,

$$T = \frac{1}{2}\left(m_1+m_2\right)l_1^2\dot{\theta}_1^2 + m_2 l_1 l_2\dot{\theta}_1\dot{\theta}_2 + \frac{1}{2}m_2 l_2^2\dot{\theta}_2^2$$

Since for two degrees of freedom

$$T = \frac{1}{2}T_{11}\dot{\theta}_1^2 + \frac{1}{2}T_{12}\dot{\theta}_1\dot{\theta}_2 + \frac{1}{2}T_{21}\dot{\theta}_2\dot{\theta}_1 + \frac{1}{2}T_{22}\dot{\theta}_2^2,$$

therefore, $T_{11} = \left[m_1+m_2\right]l_1^2, T_{12} = T_{21} = m_2 l_1 l_2$ and $T_2 = m_2 l_2^2$ Thus

$$T = \begin{pmatrix} \left(m_1+m_2\right)l_1^2 & m_2 l_1 l_2 \\ m_2 l_1 l_2 & m_2 l_2^2 \end{pmatrix} \tag{8.3.16}$$

The normal mode frequencies are determined from the equation

$$\left|V-\omega^2 T\right| = 0 \text{ or } \begin{bmatrix} \left(m_1+m_2\right)g l_1 - \omega^2\left(m_1+m_2\right)l_1^2 & -\omega^2 m_2 l_1 l_2 \\ -\omega^2 m_2 l_1 l_2 & m_2 g l_2 - \omega^2 m_2 l_2^2 \end{bmatrix} = 0$$

Dividing by l_1 in the first row and $m_2 l_2$ in the second row, we get

$$\left[\left(m_1+m_2\right)g - \omega^2\left(m_1+m_2\right)l_1\right]\left(g-\omega^2 l_2\right) - \omega^4 m_2 l_1 l_2 = 0$$
$$\text{or, } \omega^4 m_1 l_1 l_2 - \omega^2 g\left(m_1+m_2\right)\left(l_1+l_2\right) + \left(m_1+m_2\right)g^2 = 0$$
$$\text{or, } \omega^2 = \frac{g\left(m_1+m_2\right)\left(l_1+l_2\right) \pm \sqrt{\left[g\left(m_1+m_2\right)\left(l_1+l_2\right)\right]^2 - 4m_1\left(m_1+m_2\right)l_1 l_2 g^2}}{2m_1 l_1 l_2} \tag{8.3.17}$$

which gives two normal mode frequencies. Now, we consider following three special cases for the determination of frequencies.

Case I : When $m_1 \gg m_2$, then $m_1+m_2 \simeq m_1$ Now from eq. (8.3.17), we have

$$\omega^2 = \frac{gm_1\left(l_1+l_2\right) \pm gm_1\sqrt{\left(l_1+l_2\right)^2 - 4l_1 l_2}}{2m_1 l_1 l_2}$$

whence the two normal frequencies are

$$\omega_1^2 = g/l_2 \text{ and } \omega_2^2 = g/l_1$$

Case II : When $m_1 \ll m_2$, then $m_1 + m_2 \simeq m_2$ Now from eq. (8.3.17), we obtain

$$\omega^2 = \frac{gm_2(l_1+l_2) \pm g\sqrt{[m_2(l_1+l_2)]^2 - 4m_1m_2l_1l_2}}{2m_1l_1l_2} = \frac{gm_2(l_1+l_2)\left[1 \pm \left(1 - \frac{4m_1l_1l_2}{m_2(l_1+l_2)^2}\right)^{\frac{1}{2}}\right]}{2m_1l_1l_2}$$
$$= \frac{gm_2(l_1+l_2)\left[1 \pm \left(1 - \frac{2m_1l_1l_2}{m_2(l_1+l_2)^2}\right)\right]}{2m_1l_1l_2}$$

Hence

$$\omega_1^2 = \frac{gm_2}{m_1}\left[\frac{1}{l_1} + \frac{1}{l_2}\right] \quad \text{and} \quad \omega_2^2 = \frac{g}{l_1+l_2} \tag{8.3.18}$$

Case III : When $m_1 = m_2 = m$ and $l_1 = l_2 = l$,

$$\omega^2 = \frac{4gml \pm \sqrt{(4gml)^2 - 8m^2l^2g^2}}{2ml^2} = \frac{2g \pm g\sqrt{2}}{l} = \frac{g}{l}(2 \pm \sqrt{2})$$

Thus the two normal frequencies are

$$\omega_1^2 = \frac{g}{l}[2+\sqrt{2}] \quad \text{and} \quad \omega_2^2 = \frac{g}{l}[2-\sqrt{2}] \tag{8.3.19}$$

Example 8.3.1. Determine the eigen frequencies and normal coordinates of a system with two degrees of freedom whose Lagrangian is

$$L = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) - \frac{1}{2k}(x^2+y^2) + \alpha xy, \quad \alpha > 0$$

Solution.

$$L = \frac{m}{2}\left(\dot{x}^2 + \dot{y}^2\right) - \frac{1}{2k}\left(x^2+y^2\right) + \alpha xy \tag{8.3.20}$$

As $L = T - V$, hence

$$T = \frac{m}{2}\left(\dot{x}^2 + \dot{y}^2\right) \tag{8.3.21}$$

$$\text{and } V = \frac{1}{2k}\left(x^2+y^2\right) - \alpha xy \tag{8.3.22}$$

The V and T matrices are

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \tag{8.3.23}$$

Kinetic energy

$$T = \frac{1}{2}\sum_i \sum_j T_{ij}\dot{u}_i \cdot \dot{u}_j \text{ with } i,j = 1,2$$

or

$$T = \frac{1}{2}\left[T_{11}\dot{u}_1^2 + T_{12}\dot{u}_1\dot{u}_2 + T_{21}\dot{u}_2\dot{u}_1 + T_{22}\dot{u}_2^2\right] \tag{8.3.24}$$

Therefore, from (8.3.21) and (8.3.24)

$$T_{11} = m, T_{22} = m, T_{12} = T_{21} = 0$$

Also,

$$V_{11} = \left[\frac{\partial^2 V}{\partial x^2}\right]_0 = \frac{1}{k}, V_{12} = \left[\frac{\partial^2 V}{\partial x \partial y}\right]_0 = -\alpha, V_{21} = -\alpha \text{ and } V_{22} = \left[\frac{\partial^2 V}{\partial y^2}\right]_0 = \frac{1}{k}$$

Hence,

$$T = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} \text{ and } V = \begin{pmatrix} 1/k & -\alpha \\ -\alpha & \frac{1}{k} \end{pmatrix} \tag{8.3.25}$$

Eigen frequencies are determined from

$$\left|V - \omega^2 T\right| = 0$$

Thus

$$\begin{vmatrix} \frac{1}{k} - m\omega^2 & -\alpha \\ -\alpha & \frac{1}{k} - m\omega^2 \end{vmatrix} = 0$$

$$\text{or, } \left(\frac{1}{k} - m\omega^2\right)^2 - \alpha^2 = 0 \quad \text{or, } \left(\frac{1}{k} - m\omega^2 - \alpha\right)\left(\frac{1}{k} - m\omega^2 + \alpha\right) = 0$$

$$\text{or, } \frac{1}{k} - m\omega_1^2 = \alpha \text{ and } \frac{1}{k} - m\omega_2^2 = -\alpha$$

Hence,

$$\omega_1^2 = \frac{1}{mk} - \frac{\alpha}{m} \text{ and } \omega_2^2 = \frac{1}{mk} + \frac{\alpha}{m}$$

i.e., *eigen frequencies*, $\omega_1 = \sqrt{\frac{1}{mk} - \frac{\alpha}{m}}$ and $\omega_2 = \sqrt{\frac{1}{mk} + \frac{\alpha}{m}}$.

To determine the eigen vectors, we use the equation

$$\left[V - \omega_k^2 T\right] a_k = 0$$

$$\begin{pmatrix} \frac{1}{k} - m\omega_k^2 & -\alpha \\ -\alpha & \frac{1}{k} - m\omega_k^2 \end{pmatrix} \begin{pmatrix} a_{1k} \\ a_{2k} \end{pmatrix} = 0$$

For $k = 1, \omega_1^2 = \frac{1}{mk} - \frac{\alpha}{m}$ and substituting in above, we have

$$\begin{bmatrix} \frac{1}{k} - \frac{1}{k} + \alpha & -\alpha \\ -\alpha & \frac{1}{k} - \frac{1}{k} + \alpha \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} = 0 \quad \text{or} \quad \begin{bmatrix} \alpha & -\alpha \\ -\alpha & \alpha \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} = 0 \quad \text{whence } \frac{a_{21}}{a_{11}} = 1 \tag{8.3.26}$$

If $a_{11} = p$, then $a_{21} = p$. For $k = 2, \quad \omega_2^2 = \frac{1}{mk} + \frac{\alpha}{m}$, we have

$$\begin{pmatrix} \frac{1}{k} - \frac{1}{k} - \alpha & -\alpha \\ \alpha & \frac{1}{k} - \frac{1}{k} - \alpha \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = 0 \text{ or } \begin{pmatrix} -\alpha & -\alpha \\ -\alpha & -\alpha \end{pmatrix} \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = 0 \text{ or } \frac{a_{22}}{a_{12}} = -1$$

If $a_{12} = q$, then $a_{22} = -q$. Thus the eigen vectors are $a_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} p \\ p \end{pmatrix}$ and $a_2 = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = \begin{pmatrix} q \\ -q \end{pmatrix}$. Now, the matrix A is

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} p & q \\ p & -q \end{pmatrix}$$

Transpose of A i.e.,

$$A^T = \begin{pmatrix} p & p \\ q & -q \end{pmatrix}$$

Using the condition $A^T T A = I$ i.e.,

$$\begin{pmatrix} p & p \\ q & -q \end{pmatrix} \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} \begin{pmatrix} p & q \\ p & -q \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ or } \begin{pmatrix} 2mp^2 & 0 \\ 0 & 2mq^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where $p = \dfrac{1}{\sqrt{2m}} = q$. Thus the eigen vectors are

$$a_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} p \\ p \end{pmatrix} = \frac{1}{\sqrt{2m}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and

$$a_2 = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = \begin{pmatrix} q \\ -q \end{pmatrix} = \frac{1}{\sqrt{2m}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The generalized coordinates x and y related to the normal coordinates Q_1 and Q_2 as

$$u_i = \sum_{k=1}^{2} a_{ik} Q_k$$

where for $i = 1, u_1 = x$ and for $i = 2, u_2 = y$. Therefore,

$$u_1 = a_{11}Q_1 + a_{12}Q_2 \text{ and } u_2 = a_{21}Q_1 + a_{22}Q_2$$

or

$$x = \frac{1}{\sqrt{2m}}Q_1 + \frac{1}{\sqrt{2m}}Q_2 \quad \text{and} \quad y = \frac{1}{\sqrt{2m}}Q_1 - \frac{1}{\sqrt{2m}}Q_2$$

Hence normal coordinates are

$$Q_1 = \frac{\sqrt{2m}}{2}(x + y) \quad \text{and} \quad Q_2 = \frac{\sqrt{2m}}{2}(x - y)$$

Further normal coordinates Q_1 and Q_2 oscillates with frequencies ω_1 and ω_2 respectively. Therefore.

$$Q_1 = f_1 \cos(\omega_1 t + \phi_1) \quad \text{and} \quad Q_2 = f_2 \cos(\omega_1 t + \phi_2)$$

$$\text{with } x = \frac{f_1}{\sqrt{2m}} \cos(\omega_1 t + \phi_1) + \frac{f_2}{\sqrt{2m}} \cos(\omega_2 t + \phi_2)$$

$$\text{and } y = \frac{f_1}{\sqrt{2m}} \cos(\omega_1 t + \phi_1) - \frac{f_2}{\sqrt{2m}} \cos(\omega_2 t + \phi_2)$$

Exercise 8.3.2.

1. Establish the Lagrangian and deduce the Lagrangian's equations of motion for small oscillations of a system in the neighbourhood of stable equilibrium.

2. Obtain the equation of motion for small oscillations of a system around a position of a stable equilibrium.

3. Deduce the eigen value equation for small oscillations. How will you obtain the eigen values and eigenvectors from this equation?

4. What do you understand by normal modes of vibration? Explain the meaning of normal coordinates and normal frequencies. Show that when the kinetic and potential energies are expressed in terms of normal coordinates, both kinetic and potential energies are homogeneous quadratic functions.

5. Consider the case of two coupled pendulums. Determine T and V matrices, the normal frequencies, the normal coordinates, the equation of motion, the eigenvectors and the general solution.

6. Two equal masses (m) are connected to each other with the help of a spring of force constant k and then upper mass is connected to a rigid support by an identical spring. The system is allowed to oscillate in the vertical direction. Show that the frequencies of two normal modes are $\omega^2 = (3 \pm \sqrt{5})k/2m$ and the ratios of the amplitudes of two masses in the two modes are $\frac{1}{2}(\sqrt{5} \pm +1)$.

7. Determine the normal mode frequency of the Lagrangian, given by

$$L = \frac{1}{2}(\dot{x}^2 + \dot{y}^2) - \frac{1}{2}(\omega_1^2 x^2 + \omega_2^2 y^2) + \alpha xy$$

8. A particle of mass m is in a potential given by

$$V(r) = -\frac{a}{r} + \frac{ar_0^2}{3r^3},$$

where a and r_0 are positive constants. When disturbed slightly from its stable equilibrium position, it undergoes a simple harmonic oscillation. Show that time period of oscillation is $2\pi\sqrt{\frac{mr_0^3}{2a}}$.

9. A particle of mass m is moving in the the potential $V(x) = -\frac{1}{2}ax^2 + \frac{1}{4}bx^4$ where $a,\ b$ are positive constants. Show that the frequency of small oscillations about a point of stable equilibrium is $\sqrt{2a/m}$.

10. A particle of mass m is moving in the one-dimensional potential $V(x) = \frac{\alpha}{3}x^3 + \frac{\beta}{4}x^4$ where $\alpha,\ \beta < 0$. One of the equilibrium points is $x = 0$. Show that the angular frequency of small oscillations about the other equilibrium point is $\frac{\alpha}{\sqrt{m\beta}}$.

Unit 9

Course Structure

- Rings and properties
- Integral Domains and properties
- Fields and properties

9.1 Introduction

This unit is a recollection of the concept of a ring, which is a generalization of the addition and multiplication operations of standard numbers. The term “ring" was first coined by David Hilbert in 1892, although he only referred to a particular type of ring. It wasn’t until 1920 that Emmy Noether gave an abstract definition of a ring, which would apply to the “hyper-complex” number systems, developed earlier by William Hamilton and Hermann Grassmann. The groups, for example was an abstract algebraic structure together with a single binary operation that satisfies certain axioms. There are however certain groups in which we could define two binary operations. An easy example is the set of integers. It forms a group with respect to addition. We could also define multiplication on the elements of $\mathbb{Z}$, which however does not form a group (you know why!). This extra operation gives $\mathbb{Z}$ a much richer structure than standard groups. Such algebraic structure that has two binary operations satisfying certain axioms is called a ring. Certain specific kinds of rings such as integral domains, fields, etc. will also be explored in this unit.

Objectives

After studying this unit you will be able to

- recollect the idea of rings and its properties
- explore certain particular types of rings such as integral domains and fields

9.2 Rings

We earlier read that a non-empty set R with two binary operations '$+$' and '$\circ$' is said to be a **ring** if it satisfies the following:

1. R is an Abelian group with respect to '$+$'
2. R is a semi-group with respect to '$\circ$'
3. Multiplication distributes over addition, i.e.,
   (a) $a \circ (b + c) = a \circ b + a \circ c, a, b, c \in R$
   (b) $(b + c) \circ a = b \circ a + c \circ a,\ a, b, c \in R$

The triple $(R, +, \circ)$, or simply R is called a ring. Since $(R, \circ)$ is not a group, unit element may or may not exist in R with respect to multiplication. If the unit element exists, that is, there exists an element $e \in R$ such that $a \circ e = a = e \circ a, \forall a \in R$, then R is said to be a ring with unity. R is called commutative only if it is commutative with respect to '$\circ$'. For the sake of simplicity, we will thereafter write $a \circ b$ as ab.

Example 9.2.1.

1. $R = \mathbb{Z}$, is the ring of integers for usual addition and multiplication. It has 1 as unit element and is commutative.
2. $R = n\mathbb{Z}$, $n \in \mathbb{N}$ is also a ring for usual addition and multiplication. But it has no unit, and is commutative.
3. $R = \mathbb{Z}_n$, $n \in \mathbb{N}$, the additive Abelian group of residue classes modulo n. It is a ring, known to be *ring of residue classes modulo* n. It has $\bar{1}$ as unit element and it is commutative.
4. Let R be the set of all 2×2 matrices with real entries. R forms a ring with respect to usual matrix addition and multiplication. It is not commutative.
5. Let F be the set of all continuous functions $f : \mathbb{R} \to \mathbb{R}$. Then F forms a ring under addition and multiplication defined by
   $$(f + g)(x) = f(x) + g(x), \quad \text{and} \quad (fg)(x) = f(x)g(x)\ \ \forall x \in \mathbb{R}$$
   for any $f, g \in F$. The zero of this ring is the mapping $O : \mathbb{R} \to \mathbb{R}$ such that $O(x) = 0$ for all $x \in \mathbb{R}$. The additive inverse of any $f \in F$ is the function $(-f) : \mathbb{R} \to \mathbb{R}$ such that $(-f)(x) = -f(x)$. In fact F contains a unit elements too. The function $i : \mathbb{R} \to \mathbb{R}$ defined by $i(x) = 1$ for all $x \in \mathbb{R}$ is the unit element of F.
6. Let $\mathbb{Z}[i] = \{a + ib |\ a, b \in \mathbb{Z}\}$ forms a ring under usual addition and multiplication of complex numbers. $a + ib$, where $a, b \in \mathbb{Z}[i]$ is called a Gaussian integers and $\mathbb{Z}[i]$ is called the ring of Gaussian integers.
7. We can similarly define the ring of Gaussian integers modulo n, denoted by $\mathbb{Z}_n[i] = \{a + ib |\ a, b \in \mathbb{Z}_n\}$.
8. Let X be a non empty set. Then $\mathcal{P}(X)$, the power set of X forms a ring under '$+$' and '$\cdot$' which are respectively defined as
   $$A + B = (A \cup B) \setminus (A \cap B), \quad A \cdot B = (A \cap B).$$
   In fact, it is a commutative ring with unity.

9. Let M be the set of all 2×2 matrices over members from the ring of integers modulo 2. Then it forms a finite non commutative ring.

10. Let R be a ring. A polynomial with coefficients in R is of the form

$$a_0 + a_1x + \ldots + a_mx^m,$$

$a_i \in R$. The set of polynomials over R is denoted by $R[x]$ and it forms a ring with respect to the operations $+$ and $\cdot$ defined as follows.

Let $f, g \in R[x]$ be such that

$$\begin{aligned} f(x) &= a_0 + a_1x + \ldots + a_mx^m, \\ g(x) &= b_0 + b_1x + \ldots + b_nx^n. \end{aligned}$$

Then

$$\begin{aligned} f(x) + g(x) &= (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \ldots \\ f(x)g(x) &= c_0 + c_1x + c_2x^2 + \ldots + c_{m+n}x^{m+n} \end{aligned}$$

where

$$c_i = \sum_{j+k=i} a_jb_k, \;\; 0 \leq i \leq m+n.$$

$R[x]$ is then called the polynomial ring over R. (detailed discussion in later units)

11. Let $(R_1, +_1, \cdot_1), (R_2, +_2, \cdot_2), \ldots, (R_n, +_n, \cdot_n)$ be rings. Similar to the idea of external direct product of groups, we can construct a new ring as follows. Let

$$R = R_1 \oplus R_2 \oplus \ldots \oplus R_n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in R_i\}$$

and perform componentwise addition and multiplication as follows.

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 +_1 b_1, a_2 +_2 b_2, \ldots a_n +_n b_n)$$

and

$$(a_1, a_2, \ldots, a_n) \cdot (b_1, b_2, \ldots, b_n) = (a_1 \cdot_1 b_1, a_2 \cdot_2 b_2, \ldots a_n \cdot_n b_n).$$

Then $(R, +, \cdot)$ forms a ring called the *direct sum of* $R_1, R_2, \ldots, R_n$.

Theorem 9.2.2. Let R be a ring. Then

1. $a0 = 0 = 0a$, $a \in R$
2. $a(-b) = (-a)b = -(ab)$, $a, b \in R$.
3. $(-a)(-b) = ab$, $a, b \in R$.
4. $a(b - c) = ab - ac$ and $(b - c)a = ba - ca$.

If R has unity 1, then $(-1)a = -a$ and $(-1)(-1) = 1$. Further, if m, n are integers, then for any $a, b \in R$, we have

1. $n(a + b) = na + nb$; $(m + n)a = na + ma$; and $(mn)a = m(na)$;
2. $a^m a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$.

Example 9.2.3. Let $(R, +, \cdot)$ be a ring where $(R, +)$ is cyclic. Show that R is a commutative ring.

Solution. Let $(R, +)$ be generated by a. Let $x, y \in R$ be any two elements. Then $x = ma$ and $y = na$ for some integers m and n. Now,

$$\begin{aligned} xy &= (ma)(na) \\ &= (\underbrace{a + a + \ldots + a}_{m \text{ times}})(\underbrace{a + a + \ldots + a}_{n \text{ times}}) \\ &= (mn)a^2 = (nm)a^2 = (na)(ma) = yx. \end{aligned}$$

Hence, R is commutative. ■

The proof of this theorem follows directly from the definition of rings.

9.2.1 Subrings

Definition 9.2.4. Let $(R, +, \cdot)$ be a ring and let S be a nonempty subset of R. Then S is called a subring if $(S, +, \cdot)$ is itself a ring.

We note that addition and multiplication of elements of S are to coincide with addition and multiplication of these elements considered as elements of the larger ring R. Every ring R has two trivial subrings, $\{0\}$ or simply 0 and R. Let R be a ring (with or without unity). A subring of R may or may not have a unity; also if it has a unity, it can be different from the unity of R.

Example 9.2.5. If $R = \mathbb{Z}$ and $S = 2\mathbb{Z} = \{2x |\ x \in \mathbb{Z}\}$, then S does not have unity even though R has one. Further, if $R = \mathbb{Z}_{10}$ and $S = \{\bar{0}, \bar{2}, \bar{4}, \bar{6}, \bar{8}\}$, then R has unity $\bar{1}$ whereas S has unity $\bar{6}$. (verify!)

The following result is frequently useful:

Theorem 9.2.6. A non empty subset S of a ring R is a subring if and only if for all $a, b \in S$, we have $a - b \in S$ and $ab \in S$.

The proof is elementary and is left as exercise.

Definition 9.2.7. Let R be a ring. Then the set $Z(R) = \{a \in R |\ xa = ax \text{ for all } x \in R\}$ is called the center of the ring R.

Theorem 9.2.8. The center of a ring is a subring.

Definition 9.2.9. Let S be a subset of a ring R. Then the smallest subring of R containing S is called the subring generated by S.

The intersection of subrings is a subring. Hence it follows that the subring generated by a subset S of R is the intersection of all subrings of R containing S. The subring generated by the empty set is clearly $\{0\}$, and the subring generated by a single element a in R consists of all elements of the form $n_1 a + n_2 a^2 + \ldots + n_k a^k$, $n_i \mathbb{Z}$, and k is a positive integer.

Exercise 9.2.10.

1. If $C[0, 1]$ is the set of all real-valed continuous functions on $[0, 1]$ then show that for $a \in [0, 1]$, $T = \{f \in C[0, 1] \mid f(a) = 0\}$ is a subring of $C[0, 1]$.
2. Give an example of a finite non-commutative ring. Also, give an example of an infinite non-commutative ring that does not have the unity.

3. Let R be a ring such that $x^3 = x$ for all $x \in R$. Check if R is commutative.
4. Give an example of a subset of a ring that is a subgroup under addition but not a subring.
5. Describe all the subrings of the ring of integers.
6. Prove that the intersection of any collection of subrings of a ring R is a subring of R.

9.2.2 Integral Domains

We know that in a ring R, a is called left zero divisor of b if $ab = 0$. And b is called a right zero divisor, where $a, b \in R$. A commutative ring which has no zero divisors is called an *integral domain*.

Example 9.2.11.

1. The rings $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ of integers, rational, real or complex numbers are all integral domains.
2. $\mathbb{Z}_5$ is an integral domain.
3. The ring $C[0,1]$ of real valued continuous functions on [0,1] is not an integral domain because if

$$\begin{aligned} f(t) &= 0, \ 0 \le t \le 1/2 \\ &= t - \frac{1}{2}, \ 1/2 \le t \le 1 \end{aligned}$$

and

$$\begin{aligned} g(t) &= \frac{1}{2} - t, \ 0 \le t \le 1/2 \\ &= 0, \ 1/2 \le t \le 1 \end{aligned}$$

then both $f(t)$ and $g(t)$ belongs to $C[0,1]$, and both of them are not identically equal to 0. But,

$$fg \equiv 0$$

which shows that $C[0,1]$ has zero divisors.

4. $\mathbb{Z}_6$ is not an integral domain since $\overline{2}$ and $\overline{3}$ are both non-zero elements; although, we have,

$$\overline{2}.\overline{3} = \overline{6} = \overline{0}$$

showing that $\mathbb{Z}_6$ has zero divisors.

5. Consider the ring of all 2×2 matrices over the set of integers. Let

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

Both A and B are non-zero matrices but their product is equal to the zero matrix. Hence the ring of all 2×2 matrices over the set of integers has zero divisors.

6. $\mathbb{Z} \oplus \mathbb{Z}$ is not an integral domain (verify!).

Observe that, in the definition of ring, it is only a semi-group with respect to '$\circ$'.

Theorem 9.2.12. A commutative ring R is an integral domain if and only if for all $a, b, c \in R$, with $a \neq 0$, $ab = ac \Rightarrow b = c$.

Proof. Let R be an integral domain. Let $ab = ac$, $a \neq 0$. Then

$$\begin{aligned} & ab - ac = 0 \\ \Rightarrow \quad & a(b-c) = 0 \\ \Rightarrow \quad & a = 0 \text{ or } b - c = 0. \end{aligned}$$

Since $a \neq 0$, so we get $b = c$.

Conversely suppose the given condition hold. Let $a, b \in R$ be any two elements with $a \neq 0$. Suppose $ab = 0$. Then

$$ab = a \cdot 0 \Rightarrow b = 0$$

using the given condition. Hence, $ab = 0 \Rightarrow b = 0$ whenever $a \neq 0$. Hence, R is an integral domain. □

A ring R is said to satisfy left cancellation law if for all $a, b.c \in R$, $a \neq 0$, $ab = ac \Rightarrow b = c$. Similarly we can talk of right cancellation law. Thus, we can say that a commutative ring is an integral domain if and only if it satisfies the cancellation laws.

Exercise 9.2.13.

1. Check whether $\mathbb{Z} \oplus \mathbb{Z}$ is an integral domain.
2. Show that a commutative ring with the cancellation property (under multiplication) has no zero-divisors.
3. Give an example of a commutative ring without divisor of zero that is not an integral domain.
4. A ring element is called idempotent if $a^2 = a$. Prove that the only idempotent elements in an integral domain as 0 and 1.
5. Let R be the ring of real valued continuous functions on $[0, 1]$. Show that R has zero divisors.

9.3 Fields

We know that in a ring $(R, +, \cdot)$, $(R, \cdot)$ is a subring. That means, the elements of R need to have inverse elements with respect to '$\cdot$' in R. However, if some element has an inverse in R, then it is called invertible.

Definition 9.3.1. An element a in a ring R with unity 1 is called invertible (or a unit) with respect to the multiplication operation in R, if there exists some $b \in R$ such that $ab = 1 = ba$.

It is to be noted that unity and unit are two different concepts in a ring. A ring R in which every non-zero element is a unit is called a division ring. The definition can also be given as follows.

Definition 9.3.2. A ring R whose non-zero elements form a group with respect to multiplication is called a division ring or a skew field.

Example 9.3.3.

1. In a field or a skew field, every non-zero element is a unit.
2. In the ring of integers $\mathbb{Z}$, the only units are ± 1.
3. In the ring $\mathbb{Z}_n$ of integers modulo n, the units are the prime residue classes modulo n.

Further, a commutative division ring is a field. The definition of a field can also be given as follows.

Definition 9.3.4. A commutative ring R with unit element $1 \neq 0$ in which every non-zero element has an inverse with respect to multiplication, is called a field.

In this case, $(R, \circ)$ forms a commutative group which was not the case for a general ring.

Example 9.3.5. 1. The ring of rational, real or complex numbers are all fields.

2. The ring $\mathbb{Z}_p$, where p is a prime, forms a field with respect to addition and multiplication modulo p. We have seen that $\mathbb{Z}_p$ is a commutative ring with unit element $\overline{1}$. Now, for any non-zero $\overline{a} \in \mathbb{Z}_p$, a is prime to p. So, there exist some λ, μ in $\mathbb{Z}$ such that

$$\lambda a + \mu p = 1$$

Consequently,

$$\overline{\lambda}.\overline{a} = \overline{1}$$

Hence, $\overline{a}$ has an inverse in $\mathbb{Z}_p$ with respect to $\circ$. Hence, the non-zero elements of $\mathbb{Z}_p$ forms a commutative group with respect to $\circ$. It is a finite field.

3. Consider the ring of Gaussian integers modulo 3, that is,

$$\mathbb{Z}_3[i] = \{a + bi \mid a, b \in \mathbb{Z}_3\}.$$

It contains 9 elements, viz., $\overline{0}, \overline{1}, \overline{2}, i, \overline{1}+i, \overline{2}+i, \overline{2}i, \overline{1}+\overline{2}i, \overline{2}+\overline{2}i$. Elements are added and multiplied as in the complex numbers, except that the coefficients are reduced to modulo 3. In particular, $\overline{-1} = \overline{2}$. The multiplication table of the non-zero elements of $\mathbb{Z}_3[i]$ is given below.

	$\overline{1}$	$\overline{2}$	i	$\overline{1}+i$	$\overline{2}+i$	$\overline{2}i$	$\overline{1}+\overline{2}i$	$\overline{2}+\overline{2}i$
$\overline{1}$	$\overline{1}$	$\overline{2}$	i	$\overline{1}+i$	$\overline{2}+i$	$\overline{2}i$	$\overline{1}+\overline{2}i$	$\overline{2}+\overline{2}i$
$\overline{2}$	$\overline{2}$	$\overline{1}$	$\overline{2}i$	$\overline{2}+\overline{2}i$	$\overline{1}+\overline{2}i$	i	$\overline{2}+i$	$\overline{1}+i$
i	i	$\overline{2}i$	$\overline{2}$	$\overline{2}+i$	$\overline{2}+\overline{2}i$	$\overline{1}$	$\overline{1}+i$	$\overline{1}+\overline{2}i$
$\overline{1}+i$	$\overline{1}+i$	$\overline{2}+\overline{2}i$	$\overline{2}+i$	$\overline{2}i$	$\overline{1}$	$\overline{1}+\overline{2}i$	$\overline{2}$	i
$\overline{2}+i$	$\overline{2}+i$	$\overline{1}+\overline{2}i$	$\overline{2}+\overline{2}i$	$\overline{1}$	i	$\overline{1}+i$	$\overline{2}i$	$\overline{2}$
$\overline{2}i$	$\overline{2}i$	i	$\overline{1}$	$\overline{1}+\overline{2}i$	$\overline{1}+i$	$\overline{2}$	$\overline{2}+\overline{2}i$	$\overline{2}+i$
$\overline{1}+\overline{2}i$	$\overline{1}+\overline{2}i$	$\overline{2}+i$	$\overline{1}+i$	$\overline{2}$	$\overline{2}i$	$\overline{2}+\overline{2}i$	i	$\overline{1}$
$\overline{2}+\overline{2}i$	$\overline{2}+\overline{2}i$	$\overline{1}+i$	$\overline{1}+\overline{2}i$	i	$\overline{2}$	$\overline{2}+i$	$\overline{1}$	$\overline{2}i$

Thus, $\mathbb{Z}_3[i]$ is a field with 9 elements.

4. Let $\mathbb{Z}_5[i] = \left\{a + bi \mid a, b \in \mathbb{Z}_5, i^2 = -1\right\}$. This ring has 25 elements but is not an integral domain because $(\overline{1}+\overline{2}i)(\overline{1}-\overline{2}i) = \overline{1} - \overline{4}i^2 = \overline{0}$.

From the definition it is clear that every field is a skew field. But the converse may not be true as can be seen from the following example.

Example 9.3.6. Let H be the set of 2×2 complex matrices of the form

$$\begin{bmatrix} a & b \\ -\bar{b} & \bar{a} \end{bmatrix}$$

where $a, b \in \mathbb{C}$ and their bars denote their conjugates. Then H is a non-commutative division ring. For if

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

is a nonzero matrix, then its determinant $d = a\bar{a} + b\bar{b} \neq 0$, and, hence

$$A^{-1} = \begin{bmatrix} \frac{\bar{a}}{d} & -\frac{b}{d} \\ \frac{\bar{b}}{d} & \frac{a}{d} \end{bmatrix}.$$

Theorem 9.3.7. Every field is an integral domain.

Proof. Let R be a field. Suppose $a \in R$, $b \in R$, $a \neq 0$, $b \neq 0$ and $ab = 0$. Since R is a field, and $b \neq 0$, b has an inverse $b' \in R$ such that $b.b' = b'.b = 1$. Now, $(ab)b' = 0.b' = 0$, but, $a(bb') = a.1 = a$. Hence, $a = 0$, which is a contradiction. Thus $ab \neq 0$ showing that R is not an integral domain. □

Corollary 9.3.8. If R is a ring with unit element, and $b \in R$ be a unit, then b is not a zero divisor.

Corollary 9.3.9. $\mathbb{Z}_n$ is not a field, when n is not a prime.

Proof. If n is not a prime, $n = m_1.m_2$, where m_1, m_2 are proper divisors. Hence,

$$\overline{m}_1.\overline{m}_2 = \overline{m_1 m_2} = \overline{n} = \overline{0},$$

where, $\overline{m_1} \neq 0$, $\overline{m_2} \neq 0$. Hence, $\mathbb{Z}_n$ is not an integral domain. Hence, it is not a field. □

Thus, a skew field is also an integral domain. But the converse is not true. For example, $\mathbb{Z}$ is an integral domain without being a skew field. Also the same example shows that an integral domain may not be a field as well. Thus, the converse of the above theorem may not be true in general. However, we see that the set of integers is infinite. What happens if the integral domain is finite? The next result gives an answer to this question.

Theorem 9.3.10. Let R be a finite integral domain. Then R is a field.

Proof. We shall first show that R has a unit element. Let $R = \{a_1, a_2, \cdots, a_n\}$ be the distinct elements of R and let a be any non-zero element of R. Then the elements $aa_1, aa_2, aa_3, \cdots, aa_n$ are all distinct points of R, because if $aa_i = aa_j$, then $a(a_i - a_j) = 0$ and since R is an integral domain and $a \neq 0$, so $a_i = a_j$. Hence, the set $\{aa_1, aa_2, aa_3, \cdots, aa_n\}$ coincide with R. In particular, $aa_k = a$ for some k. We shall show that a_k is the unit element of R. Let a_j be any element of R. Then $a_j = aa_i$ for some i. Now,

$$\begin{aligned} a_j a_k &= a_k a_j \\ &= a_k(aa_i) \\ &= (a_k a)a_i \\ &= (aa_k)a_i \\ &= aa_i \\ &= a_j \end{aligned}$$

Thus, a_k is the unit element of R. The unit element is unique and we shall denote it by 1. Let $a \in R$, $a \neq 0$. Since $1 \in R$, $aa_l = a_l a = 1$, for some a_l. Thus a has an inverse with respect to multiplication showing that R is a field. □

Corollary 9.3.11. For any prime p, $\mathbb{Z}_p$ is a field.

Proof. $\mathbb{Z}_p$ is an integral domain, because if $\overline{a}.\overline{b} = 0$, that is, $\overline{ab} = \overline{0}$, then p about ab. Since p is a prime, p divides either a or b, that is, $\overline{a} = \overline{0}$ or, $\overline{b} = \overline{0}$. Hence, $\mathbb{Z}_p$ is an integral domain, and since it is a finite ring, so it is a field. □

Another term that needs to be introduced is the nilpotent element.

Definition 9.3.12. Let $(R, +, \cdot)$ be a commutative ring with additive identity 0. An element $a \in R$ is said to be Nilpotent if there exists an $n \in \mathbb{N}$ such that $a^n = 0$.

Example 9.3.13. 1. By definition, the additive identity 0 is always a nilpotent element in a ring $(R, +, \cdot)$.

2. In the commutative ring $(\mathbb{Z}_4, +, \cdot)$. Then the element $2 \in \mathbb{Z}_4$ is nilpotent since $2^2 = 4 \equiv 0 \pmod 4$.

Theorem 9.3.14. Let $(R, +, \cdot)$ be a commutative ring and let $a, b \in R$ be nilpotent. Then $a + b$ is nilpotent.

Proof. Let $a, b \in R$ be nilpotent. Then there exists $n, m \in \mathbb{N}$ such that $a^n = 0$ and $b^m = 0$. Let $t = nm + 1$. Then t is a positive integer and by the binomial theorem:

$$(a+b)^t = \binom{t}{0} a^t b^0 + \binom{t}{1} a^{t-1} b^1 + ... + \binom{t}{t} a^0 b^t$$

Each term in the expression is of the form $\binom{t}{k} a^{t-k} b^k$. Since $t = nm + 1$, we must have at least $t - k > n$ or $k > m$ (since if $t - k < n$ and $k < m$ then $t = (t - k) + k < n + m \leq nm < t$). So each term in the expression above is equal to zero. So $(a + b)$ is nilpotent. □

Exercise 9.3.15. 1. Show that a non-zero element a in $\mathbb{Z}_n$ is a unit if and only if a and n are relatively prime. Also show that if a is not a unit, then it is a zero divisor.

2. Show that $\mathbb{Z}_p$ is a field if and only if p is a prime.

3. Let R be a commutative ring with unity. Show that

   (a) a is a unit if and only if a^{-1} is a unit.

   (b) a, b are units if and only if ab us a unit.

4. Show that the set of all units in a commutative ring with unity forms an Abelian group.

5. If $(R, +, \cdot)$ is a commutative ring and $a \in R$ is nilpotent, then show that for all $r \in R$, $r \cdot a$ and $a \cdot r$ are nilpotent.

6. Let $(R, +, \cdot)$ be a commutative ring and let $u, a \in R$. If u is a unit and a is nilpotent, show that $u - a$ is a unit.

9.4 Characteristic of a Ring

Definition 9.4.1. Let R be a ring. The *characteristic* of R is the smallest positive integer n, if it exists, such that $na = 0$ for all $a \in R$. If no such n exists, the characteristic of R is said to be zero. We denote the characteristic of R as char$R$.

Example 9.4.2. 1. If $R = \mathbb{Z}$, then char$R = 0$, because there exists no such positive integer m such that, $m\mathbb{Z} = 0$.

2. char$\mathbb{Z}_n = n$ since $n\overline{a} = 0$ for all $\overline{a} \in \mathbb{Z}_n$.

Theorem 9.4.3. Let R be a ring with unit element e. Then

1. char$R = n$, if n is the smallest positive integer such that $ne = 0$, and

2. char$R = 0$ if no such n exists.

Proof. 1. Suppose n is the smallest positive integer such that $ne = 0$. Then, for any $a \in R$,

$$\begin{aligned} na &= a + a + \cdots + a(n \text{ times}) \\ &= ae + ae + \cdots + ae(n \text{ times}) \\ &= a(e + e + \cdots + e) \\ &= a(ne) \\ &= a0 \\ &= 0 \end{aligned}$$

Also, for $0 < m < n$, $me \neq 0$. Hence, char$R = n$.

2. Suppose no such n exists. If possible, let there exists some $m > 0$, with $ma = 0$ for all $a \in R$. In particular, $me = 0$, $m > 0$ which contradicts the given hypothesis. Hence, our assumption is wrong and char$R$ has to be 0.

□

Theorem 9.4.4. Let R be an integral domain. Then char$R$ is either 0 or a prime.

Proof. Suppose char$R = n \neq 0$ and suppose n is not prime. Then $n = m_1.m_2$, where m_1 and m_2 are proper divisors of n. For any $a \in R$, $a \neq 0$, we have,

$$0 = na^2 = (m_1 m_2)a^2 = (m_1 a)(m_2 a).$$

Since R is an integral domain, $m_1 a = 0$ or $m_2 a = 0$. Suppose $m_1 a = 0$. Then we show that $m_1 x = 0$ for any $x \in R$. Now, $m_1(xa) = (m_1 x)a = x(m_1 a) = x0 = 0$. Since $a \neq 0$, and R is an integral domain, $m_1 x = 0$. Thus, $m_1 x = 0$ for all x, $m_1 < n$. This contradicts the assumption that char$R = n$. Hence proved. □

Corollary 9.4.5. The characteristic of a field is either 0 or a prime.

Exercise 9.4.6.

1. Suppose that R is a commutative ring without zero-divisors. Show that the characteristic of R is 0 or prime.
2. Let R be a ring with m elements. Show that the characteristic of R divides m.
3. Let F be a field of order 2^n. Show that char $F = 2$.
4. Find the characteristic of $\mathbb{Z}_4 \oplus 4\mathbb{Z}$.
5. If R is a ring of characteristic $m > 0$ and S is a subring of R, what can you say about the characteristic of S?

Sample Questions

1. In a ring $(R, +, \circ)$ if $(R, +)$ is cyclic, show that R is a commutative ring.
2. Show that the center of a ring is its subring.
3. Show that a finite integral domain is a division ring.
4. Show that a field is an integral domain. Is the converse true? Justify your answer.
5. Let F be a field of characteristic 2 with more than two elements. Show that $(x + y)^3 = x^3 + y^3$ for some x and y in F.

Unit 10

Course Structure

- Definition of Ideals
- Classification of ideals
- Factor Rings

10.1 Introduction

We have studied the concept of subrings of a given ring. Recollect that in the theory of groups, the normal subgroups played a special role. They helped in the construction of quotient groups. In a similar manner, we will study a special kind of subring of a ring which will enable us to define the concept of quotient rings. Such subrings will be called ideals. Factor rings will be constructed in the same fashion as factor groups. The role of an ideal in a "homomorphism between rings" is similar to the role of a normal subgroup in a "homomorphism between groups.

Objectives

After reading this unit, you will be able to

- define left and right ideals and use them to define ideal
- visualise the connection between ideals and normal subgroups
- define certain important kinds of ideals and deal with their properties
- define factor rings with the help of an ideal and characterize them in accordance with the type of ideal

10.2 Ideals in rings

Let R be a ring and $I \subset R$. I is called a left(right) ideal of R if

1. $a - b \in I$ where $a, b \in I$.

2. For each $a \in I$ and $x \in R$, $xa \in I (ax \in I)$.

Clearly, a right or left ideal is a subring of the ring. An ideal is both a left and right ideal, and in a commutative ring, any left or right ideal is an ideal. The following result gives a way to check ideals of R

Theorem 10.2.1. A nonempty subset A of a ring R is an ideal of R if

1. $a - b \in A$ whenever $a, b \in A$.

2. ra and ar are in A whenever $a \in A$ and $r \in R$.

Example 10.2.2.

1. For any ring R, $\{0\}$ and R are ideals of R. These are called the *trivial ideals* of R.

2. For any positive integer n, the set $n\mathbb{Z} = \{nz : z \in \mathbb{Z}\}$ is an ideal of $\mathbb{Z}$.

3. Let R be a commutative ring with unit element an let $a_1, a_2, \ldots, a_n \in R$. Then $I = \langle a_1, a_2, \ldots, a_n \rangle = \{r_1a_1 + r_2a_2 + \cdots r_na_n : r_i \in R\}$ is an ideal of R called the $ideal$ generated by $a_1, a_2, \ldots, a_n$. This is the smallest ideal containing $a_1, a_2, \ldots, a_n$. (discussed shortly)

4. Let R be the ring of 2×2 upper triangular matrices over a field F. Then the subset

$$I = \left\{ \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \mid a \in F \right\}$$

   is an ideal in R.

5. Let R be the ring of all functions from the closed interval $[0, 1]$ to the field of real numbers. Let $c \in [0, 1]$ and $I = \{f \in R \mid f(c) = 0\}$. Then I is an ideal of R.

6. Let $\mathbb{R}[x]$ denote the set of all polynomials with real coefficients and let A denote the subset of all polynomials with constant term 0. Then A is an ideal of $\mathbb{R}[x]$ and $A = \langle x \rangle$.

7. Let $Z[x]$ denote the ring of all polynomials with integer coefficients and let I be the subset of $Z[x]$ of all polynomials with even constant terms. Then I is an ideal of $Z[x]$ and $I = \langle x, 2 \rangle$.

8. Let R be the ring of all real-valued functions of a real variable. The subset S of all differentiable functions is a subring of R but not an ideal of R.

Also, if I is an ideal of R, then I forms a normal subgroup of $(R, \circ)$(check it).

Theorem 10.2.3. Let $\{A_i\}_{i \in \Lambda}$ be a family of right (left) ideals in a ring R. Then $\bigcap_{i \in \Lambda} A_i$ is also a right (left) ideal of R.

Proof. Let $A = \bigcap_{i \in \Lambda} A_i$ and $a, b \in A$ and $r \in R$. Then for all $i \in \Lambda$, $a, b \in A_i$ and ar $(ra) \in A_i$ since A_i's are right (left) ideals. Hence the result. □

Next, let S be a subset of a ring R. Let $\mathscr{A} = (A \mid A$ is a right ideal of R containing S). Then $\mathscr{A} \neq \varnothing$ because $R \in \mathscr{A}$. Let $I = \bigcap_{A \in \mathcal{A}} A$. Then I is the smallest right ideal of R containing S and is denoted by $\langle S \rangle_{\mathrm{r}}$. The smallest right ideal of R containing a subset S is called a right ideal generated by S. If $S = \{a_1, \ldots, a_m\}$ is a finite set, then $\langle S \rangle_{\mathrm{r}}$ is also written $\langle a_1, \ldots, a_m \rangle_{\mathrm{r}}$. Similarly, we define the left ideal and the ideal generated by a subset S, denoted, respectively, by $\langle S \rangle_l$ and $\langle S \rangle$.

Definition 10.2.4. A right ideal I of a ring R is called finitely generated if $I = \langle a_1, \ldots, a_m \rangle_r$ for some $a_i \in R, 1 \leq i \leq m$.

Definition 10.2.5. A right ideal I of a ring R is called principal if $I = \langle a \rangle_r$ for some $a \in R$.

In a similar manner we define a finitely generated left ideal, a finitely generated ideal, a principal left ideal $\langle a \rangle_l$, and a principal ideal $\langle a \rangle$.

Exercise 10.2.6.

1. Let A and B be ideals in a ring. Show that $AB \subseteq A \cap B$.
2. Prove that $I = \{f(x) \in \mathbb{Z}[x] \mid f(1) \text{ is even}\}$ is an ideal of $\mathbb{Z}[x]$.
3. Let $S = \{a + bi \mid a, b \in \mathbb{Z},\ b \text{ is even}\}$. Show that S is a subring of $\mathbb{Z}[i]$. Is it an ideal of $\mathbb{Z}[i]$? Justify your answer.
4. If A and B are ideals of a ring, show that the sum of A and B, that is, $A + B = \{a + b \mid a \in A,\ b \in B\}$ is also an ideal of the ring.
5. Show that

$$\begin{aligned} \langle a \rangle &= \left\{ \sum_{i=1}^{k} r_i a s_i + ra + as + na \mid r, s, r_i, s_i \in R\, \forall i,\ n, k \in \mathbb{Z} \right\} \\ \langle a \rangle_r &= \{ar + na \mid r \in R,\ n \in \mathbb{Z}\} \\ \langle a \rangle_l &= \{ra + na \mid r \in R,\ n \in \mathbb{Z}\}. \end{aligned}$$

6. If R contains unity, then show that

$$\begin{aligned} \langle a \rangle &= \left\{ \sum_{i=1}^{k} r_i a s_i \mid r_i, s_i \in R\, \forall i,\ k \in \mathbb{Z} \right\} \\ \langle a \rangle_r &= \{ar \mid r \in R,\ n \in \mathbb{Z}\} \\ \langle a \rangle_l &= \{ra \mid r \in R,\ n \in \mathbb{Z}\}. \end{aligned}$$

(In this case, $\langle a \rangle_r$ and $\langle a \rangle_l$ are also denoted by aR and Ra respectively.

10.3 Factor Rings (or Quotient Rings)

Let R be a ring and let A be an ideal of R. Since R is a group under addition and A is a normal subgroup of R, we may form the factor group $R/A = \{r + A \mid r \in R\}$. The natural question at this point is: How may we form a ring of this group of cosets? The addition is already taken care of, and, by analogy with groups of cosets, we define the product of two cosets $s + A$ and $t + A$ of R as $st + A$. The next theorem shows that this definition works as long as A is an ideal of R, and not just a subring of R.

Theorem 10.3.1. (Existence of factor rings) Let R be a ring and A be a subring of R. The set of cosets $\{r + A \mid r \in R\}$ is a ring under the operations

1. $(s + A) + (t + A) = s + t + A$,
2. $(s + A)(t + A) = st + A$

$s, t \in R$, if and only if A is an ideal of R.

Proof. We know that the set of cosets forms a group under addition. Once we know that multiplication is indeed a binary operation on the cosets, it is trivial to check that the multiplication is associative and that multiplication is distributive over addition. Hence, the proof reduces to showing that multiplication is well-defined if and only if A is an ideal of R. To do this, let us suppose that A is an ideal and let $s + A = s' + A$ and $t + A = t' + A$. Then we must show that $st + A = s't' + A$. Well, by definition, $s = s' + a$ and $t = t' + b$, where a and b belong to A. Then

$$st = \left(s' + a\right)\left(t' + b\right) = s't' + at' + s'b + ab,$$

and so

$$st + A = s't' + at' + s'b + ab + A = s't' + A,$$

since A absorbs $at' + s'b + ab$. Thus, multiplication is well-defined when A is an ideal.

On the other hand, suppose that A is a subring of R that is not an ideal of R. Then there exist elements $a \in A$ and $r \in R$ such that $ar \notin A$ or $ra \notin A$. For convenience, say $ar \notin A$. Consider the elements $a + A = 0 + A$ and $r + A$. Clearly, $(a + A)(r + A) = ar+$ but $(0 + A)(r + A) = 0 \cdot r + A = A$. Since $ar + A \neq A$, the multiplication is not well-defined and the set of cosets is not a ring. □

Thus, the ring $(R/A, +, \cdot)$ is called the factor ring.

Example 10.3.2.

1. $\mathbb{Z}/4\mathbb{Z}$=$\{0 + 4\mathbb{Z}, 1 + 4\mathbb{Z}, 2 + 4\mathbb{Z}, 3 + 4\mathbb{Z}\}$ is a factor ring under the two operations modulo 4.

2. $2\mathbb{Z}/6\mathbb{Z}$ is a commutative example of an ideal and factor ring.

3. Consider the factor ring $R = \mathbb{Z}_3[x]/\left\langle x^2 + 1\right\rangle$. To simplify the notation let $I = \left\langle x^2 + 1\right\rangle$ and let us write the elements of $\mathbb{Z}_3$ as simply numbers without using bars. By definition, the elements of R have the form $f(x) + I$ where $f(x)$ is a polynomial with coefficients from $\mathbb{Z}_3$. But what are the distinct elements of R? The fact that $x^2 + 1 + I = 0 + I$ means that when dealing with coset representatives we may treat x^2+1 as equivalent to 0 and therefore $x^2 = -1$. For example, the coset $2x^2+x+1+I = -2+x+1+I = x - 1 + I$. Moreover, when dealing with coset representatives we have, $x^2 = -1$, which implies that $x^3 = -x$ and $x^4 = 1$. So, $x^4+2x^3+x^2+2+I = 1-2x-1+2+I = -2x+2+I = x+2+I$ (because $-2 = 1$ in $\mathbb{Z}_3$). In the same way, we have $x^5 = x$, $x^6 = x^2x^4 = -1$ and so on. Thus we see that $R = \{ax + b + I \mid a, b \in \mathbb{Z}_3\}$. This means that R has order 9 . We can make one more simplification by suppressing the use of I and just refer to the coset $ax + b + I$ as $ax + b$. All we need to keep in mind is that when we preform the product $(ax + b)(cx + d) = acx^2 + (ad + bc)x + bd$ we replace acx^2 with $-ac$. We can now ask if any particular non-zero element of R is a unit or a zero-divisor. Consider $x + 1$. Note that $(x + 1)^2 = x^2 + 2x + 1 = 2x$. Then $(x + 1)^4 = (2x)^2 = 4x^2 = -4 = -1$. So, $x + 1$ is a unit and $|x + 1| = 8$. This means the eight non-zero elements of R form a cyclic group and R is a field of order 9.

4. Consider the factor ring $R = \mathbb{Z}_5[x]/\langle x^2 + 1\rangle$. This time, $|R| = 25$ and $(x + 1)^4 = (2x)^2 = 4x^2 = -4 = 1$. So, $x + 1$ is a unit in R and $|x + 1| = 4$. Further, $(x + 2)(x + 3) = x^2 + 1 = 0$ and so $x + 2$ and $x + 3$ are zero divisors and hence R is not a field.

10.4 Types of Ideals

We will come across three types of ideals. First we have already seen the principal ideals. In this section, we will come across two more ideals, viz., Prime and Maximal ideals.

Definition 10.4.1. A prime ideal A of a commutative ring R is a proper ideal of R such that $a, b \in R$ and $ab \in A$ imply $a \in A$ or $b \in A$.

The motivation for the definition of prime ideals comes from the integers.

Definition 10.4.2. A maximal ideal of a commutative ring R is a proper ideal of R such that, whenever B is an ideal of R and $A \subseteq B \subseteq R$, then $B = A$ or $B = R$.

So, the only ideal that properly contains a maximal ideal is the entire ring.

Example 10.4.3.
1. Let n be an integer greater than 1. Then, in the ring of integers, the ideal $n\mathbb{Z}$ is prime if and only if n is prime.
2. $\langle 2\rangle$ and $\langle 3\rangle$ are maximal ideals of $\mathbb{Z}_{36}$.
3. $\langle x^2 + 1\rangle$ is maximal in $\mathbb{R}[x]$.

Theorem 10.4.4. In a commutative ring R, let A be an ideal. If a unit element is in A, then $A = R$.

Proof. Let $u \in A$ be an unit element of R. Let $v \in R$ be such that $uv = 1$. Now, since A is an ideal of R, so for $u \in A$ and $v \in R$, we must have $uv \in A$, i.e., $1 \in A$. Now, for every $r \in R$, we must have $r1 \in A$, i.e., $r \in A\ \forall r \in R$. Thus, $A = R$. □

Theorem 10.4.5. Let R be a commutative ring with unity and let A be an ideal of R. Then, R/A is an integral domain if and only if A is prime.

Proof. Let R/A is an integral domain and $ab \in A$. Then, $(a + A)(b + A) = ab + A = A$, the zero element of the ring R/A. So, either $a + A = A$ or $b + A = A$, that is, either $a \in A$ or $b \in A$. Hence, A is prime.

We observe that R/A is a commutative ring with unity for any proper ideal A. Our task is only to show that R/A has no zero divisors. So, let A be prime ideal and $(a + A)(b + A) = 0 + A = A$. Then, $ab \in A$ and hence, $a \in A$ or $b \in A$. Thus, one of $a + A$ or $b + A$ is the zero coset in R/A. □

Theorem 10.4.6. Let R be a commutative ring with unity and let A be an ideal of R. Then, R/A is a field if and only if A is maximal.

Proof. Let R/A be a field and B is an ideal of R that properly contains A. Let $b \in B$ but $b \notin A$. Then $b + A$ is a nonzero element of R/A and hence there exists an element $c + A$ such that $(b + A).(c + A) = 1 + A$, the multiplicative identity of R/A. Since $b \in B$, we have $bc \in B$. Since

$$1 + A = (b + A)(c + A) = bc + A,$$

we have $1 - bc \in A \subset B$. So, $1 = (1 - bc) + bc \in B$. So, $B = R$. Thus, A is maximal.

Now let A is maximal and $b \in R$ but $b \notin A$. It suffices to show that $b + A$ has a multiplicative inverse. Consider $B = \{br + a | r \in R, a \in A\}$. This is an ideal in R that properly contains A. Since A is maximal, we must have $B = R$. Thus, $1 \in B$, and $1 = bca'$(say), where $a' \in A$. Then

$$1 + A = bc + a' + A = bc + A = (b + A)(c + A).$$

□

Since $\mathbb{Z}$ is an integral domain but not a field, and since

$$\mathbb{Z}[x]/ < x > \simeq \mathbb{Z}$$

(follows from the next unit) we must have from the previous theorems, ideal $\langle x\rangle$ is prime but not maximal.

Theorem 10.4.7. In a commutative ring with unity, every maximal ideal is a prime ideal.

Proof. Left as an exercise. □

But the converse of the theorem is not true in general. For example, $\langle 0 \rangle$ is a prime ideal in $\mathbb{Z}$ but not a maximal ideal.

Exercise 10.4.8.

1. Find all the maximal ideals of $\mathbb{Z}_8,\ \mathbb{Z}_{10},\ \mathbb{Z}_{12},\ \mathbb{Z}_{30}$.
2. Let $n = st$, where s and t are divisors of n greater than 1. Prove that $\langle s \rangle$ is a maximal ideal in $\mathbb{Z}_n$ if and only if s is prime.
3. In $\mathbb{Z}[x]$, prove that $\langle 2x, 3 \rangle = \langle x, 3 \rangle$.
4. If n is an integer greater than 1, show that $\langle n \rangle = n\mathbb{Z}$ is a prime ideal of $\mathbb{Z}$ if and only if n is a prime number.
5. Give an example of a commutative ring that has a maximal ideal that is not a prime ideal.
6. Let R be a commutative ring with unity. Suppose that the only ideals of R are $\{0\}$and R. Show that R is a field.

10.5 Factorization

We assume throughout this section that R is an integral domain with unit element.

Definition 10.5.1. Let $a \in R$ and $b \in R$, $a \neq 0$. a is said to divide b if there exists $c \in R$ such that $b = ac$.

Example 10.5.2.

1. In $R = \mathbb{Z}$, 3 divides 15.
2. In $R = \mathbb{Z} + i\mathbb{Z} = \{a + ib | a, b \in \mathbb{Z}\}$, $(1 + 3i)$ divides 10 as $10 = (1 + 3i)(1 - 3i)$.

Definition 10.5.3. Two non-zero elements a and b are said to be $associates$ if $a|b$ and $b|a$.

That is, a and b are associates if and only if $b = au$, for some unit $u \in R$.

Definition 10.5.4. Any element a in ring R is called $irreducible$ if

1. a in not an unit,
2. the only divisors of a are units and associates of a

Example 10.5.5.

1. Let $R = \mathbb{Z}$ and $n > 1$. Then n is irreducible if and only if the only divisors of n are 1,-1,n,-n. Thus n is a prime integer.
2. In $R = \mathbb{Z}$, 5 and -5 are associates as $-5 = (-1).5$.
3. Let $R = \{a + b\sqrt{-5} | a, b \in \mathbb{Z}\}$. Then, $1 + 2\sqrt{-5}$ is an irreducible element of R.

Definition 10.5.6. Let $p \in R$ which is not a unit, p is called a $prime$ element if whenever $p|ab$, then either $p|a$ or $p|b$.

Theorem 10.5.7. Every prime element is irreducible.

Proof. Let p is prime, and let a be any divisor of p so that $p = ab$. Since p is a prime $p|a$ or $p|b$. If $p|a$, since we have $a|p$ we have a is an associate of p and b is a unit. Similarly, if $p|b$, then b is an associate of p and a is a unit. Thus the only divisors of p are units and associates of p. Hence, p is irreducible. □

In general rings, the converse of the theorem is not true. But we know, in $R = \mathbb{Z}$, every irreducible element is prime.

Proposition 10.5.8. An element p$\in R$ is prime if and only if the ideal $\langle p \rangle$ is prime ideal.

Proof. Let p is a prime element and let $ab \in < p >$. Then $ab = cp$ for some $c \in R$. Since, p is a prime, $p|a$ or $p|b$, i.e., $a \in P$ or $b \in P$. Hence, P is a prime ideal. Conversely, let P be a prime ideal and let $p|ab$. Then $ab = cp \in P$ and since P is a prime ideal, $a \in P$ or $b \in P$, i.e., $p|a$ or $p|b$. Hence, p is a prime element. □

Definition 10.5.9. Let $a \in R$ and $b \in R$. An element $d \in R$ is called a greatest common divisor(gcd) of a and b if

1. $d|a$ and $d|b$,
2. whenever $d'|a$ and $d'|b$, then $d'|d$.

Example 10.5.10. In $R = \mathbb{Z}$, if $a = 9$, $b = -48$ then $d = 3$ is a gcd of a, b.

Example 10.5.11. Let R be an integral domain and let a and b are two elements in R.

1. b divides a if and only if $Ra \subset Rb$ and vice versa. The proof is obvious since if $b|a$ there exists some c in R, such that $a = bc$. Thus, for any x in R, we have

$$\begin{aligned} x &= ya \\ &= ybc \\ &= ycb \end{aligned}$$

   Hence, $x \in Rb$. Thus, $Ra \subset Rb$. the converse part is left as an exercise.

2. a and b are associates if and only if $Ra = Rb$. Since a and b are associates, $a|b$ which yields $Rb \subset Ra$; and $b|a$ which gives $Ra \subset Rb$. Combining, we get the desired result. The converse part is left as exercise.

Example 10.5.12. Let R be an integral domain. If $a \in R$ such that Ra is a maximal ideal, then a is irreducible. Since Ra is maximal ideal, a is not a unit. If possible, let a be reducible. Then there exists b and c in R such that $a = bc$. Then b and c are proper divisors of a. Hence, by the previous example, $Ra \subset Rb$ and $Ra \subset Rc$. Which is impossible since Ra is a maximal ideal in R.

Exercise 10.5.13.

1. Show that the union of a chain $I_1 \subset I_2 \subset \ldots$ of ideals of a ring R is an ideal of R
2. Show that the product of an irreducible and a unit is irreducible in an integral domain. Is the result true in any arbitrary ring?
3. Let D be an integral domain. Define $a\rho b$ if a and b are associates. Show that 'ρ' is an equivalence relation on R.

Sample Questions

1. Show that the set of all upper triangular matrices form an ideal of the ring of all 2×2 matrices over some field F.
2. Show that a R/A forms a factor ring if and only if A is an ideal of R.
3. With proper justifications, list the elements of $\mathbb{Z}_5[x]/\langle x^2 + 1\rangle$.
4. If R is a commutative ring with unity, then show that
   (a) A is a prime ideal of R if and only if R/A is an integral domain;
   (b) A is a maximal ideal of R if and only if R/A is a field.
5. Show that an ideal A of R containing a unit is equal to R.
6. Show that in a commutative ring with unity, every maximal ideal is a prime ideal.
7. Show that every prime element in an integral domain with unity is irreducible.
8. Let R be an integral domain with unity. Show that a and b are associates if and only if $Ra = Rb$.

Unit 11

Course Structure

- Ring homomorphisms: basic definitions and properties
- Classification of rings, their definitions and characterization theorem with examples and counter examples.

11.1 Introduction

We have constantly tried to establish the theory of rings taking motivation from the theory of groups. We had studied the idea of group homomorphisms which preserve the algebraic structure of the groups. Also, depending upon the algebraic structure, the groups were classified using the idea of homomorphisms. We will do somewhat similar in case of rings as well. We note that there are two binary operations in rings. A structure preserving map between rings will have to preserve both the operations. Such maps are called ring homomorphisms. In an similar way, we will use the ring homomorphism to characterise rings depending upon their algebraic structures.

Objectives

After reading this unit, you will be able to

- define ring homomorphism and learn its basic properties
- define various types of homomorphisms and look into their examples
- characterise rings on the basis of ther algebraic structures

11.2 Ring Homomorphisms

A *ring homomorphism* ϕ from a ring $(R, +_R, \cdot_R)$ to $(S, +_S, \cdot_S)$ is a mapping from R to S that preserves the two ring operations, that is, for all $a, b \in R$,

1. $\phi(a +_R b) = \phi(a) +_S \phi(b)$

2. $\phi(a \cdot_R b) = \phi(a) \cdot_S \phi(b)$.

For the sake of simplicity, we will use just the symbols '$+$' and '$\cdot$' instead of $+_R$ (or $+_S$) or $\cdot_R$ (or $\cdot_S$).

If f is $1-1$, then f is called a monomorphism from R into S. In this case f is also called an embedding of the ring R into the ring S (or R is embeddable in S); we also say that S contains a copy of R, and R may be identified with a subring of S.

If a homomorphism f from a ring R into a ring S is both 1-1 and onto, then there exists a homomorphism g from S into R that is also 1-1 and onto. In this case we say that the two rings R and S are isomorphic, and, abstractly speaking, these rings can be regarded as the same (algebraically). We write $R \simeq S$ whenever there is a $1-1$ homomorphism (isomorphism) of R onto S. As stated above $R \simeq S$ implies $S \simeq R$. Also, the identity mapping gives $R \simeq R$ for any ring R. It is easy to verify that if $f : R \to S$ and $g : S \to T$ are isomorphisms of R onto S and S onto T, respectively, then gf is also an isomorphism of R onto T. Hence, $R \simeq S$ and $S \simeq T$ imply $R \simeq T$. Thus, we have shown that

Isomorphism is an equivalence relation in the class of rings.

Example 11.2.1.

1. For any two rings R and S, the maps $f : R \to S$ and $g : R \to S$ defined as $f(r) = 0$ and $g(r) = 1$, for all $r \in R$, where 0 and 1 are the additive identity and unity in S, then f and g are ring homomorphisms called the trivial homomorphisms.

2. For any positive integer n, the mapping $k \to k$ mod n is a ring homomorphism from $\mathbb{Z}$ onto $\mathbb{Z}_n$. This mapping is called the natural homomorphism from $\mathbb{Z}$ onto $\mathbb{Z}_n$.

3. The mapping $a + bi \to a - bi$ is a ring isomorphism from the complex numbers onto the complex numbers.

4. Let $\phi : \mathbb{Z}_2 \to \mathbb{Z}_2$ be defined by $\phi(x) = x^2$. Then,

$$\phi(x+y) = (x+y)^2 = x^2 + 2xy + y^2 = x^2 + y^2 = \phi(x) + \phi(y)$$

since $2xy = 0$ since the characteristic of $\mathbb{Z}_2$ is 2. Next,

$$\phi(xy) = (xy)^2 = x^2y^2 = \phi(x)\phi(y).$$

The second equality follows from the fact that $\mathbb{Z}_2$ is commutative. Also, note that $\phi(1) = 1^2 = 1$. Thus, ϕ is a ring homomorphism.

5. Let R be a commutative ring of characteristic 2. Then the mapping $a \to a^2$ is a ring homomorphism from R to R.

6. The function $\phi : \mathbb{Z} \to \mathbb{Z}$ defined by $\phi(x) = 2x$ is not a ring homomorphism. Indeed, for any $x, y \in \mathbb{Z}$, we have

$$\phi(x+y) = 2(x+y) = 2x + 2y = \phi(x) + \phi(y).$$

Thus, ϕ is additive. But,

$$\phi(1.3) = \phi(3) = 2.3 = 6, \text{ while } \phi(1)\phi(3) = (2.1)(2.3) = 12.$$

Thus, $\phi(1.3) \neq \phi(1)\phi(3)$.

Exercise 11.2.2.

1. Show that $\phi : \mathbb{Z} \to \mathbb{Z}$ defined by $\phi(x) = 2x + 5$ is not a ring homomorphism.

2. Check whether $\phi : \mathbb{Z} \to \mathbb{Z}$ defined by $\phi(x) = 3x$ is a ring homomorphism.

Theorem 11.2.3. Let R, S, T be rings and let $f : R \to S$ and $g : S \to T$ be ring homomorphisms. Then the composite map $g \circ f : R \to T$ is also so.

Proof. Let $x, y \in R$ and $h = g \circ f$. Then

$$h(x+y) = g(f(x+y)) = g(f(x)+f(y)) = g(f(x)) + g(f(y)) = h(x) + h(y).$$
$$h(xy) = g(f(xy)) = g(f(x)f(y)) = g(f(x))g(f(y)) = h(x)h(y).$$

Hence the result. □

Let us now list a few elementary but fundamental properties of homomorphisms.

Theorem 11.2.4. (Properties of Ring Homomorphism) Let ϕ be a ring homomorphism from a ring R to a ring S. Let A be a subring of R.

1. For any $r \in R$ and any positive integer n, $\phi(nr) = n\phi(r)$ and $\phi(r^n) = (\phi(r))^n$.
2. $\phi(A) = \{\phi(a) | a \in A\}$ is a subring of S.
3. If A is an ideal and ϕ is onto S, then $\phi(A)$ is an ideal.
4. $\phi^{-1}(B) = \{r \in R \mid \phi(r) \in B\}$ is an ideal of R.
5. If R is commutative, then $\phi(R)$ is commutative.
6. If R has a unity 1, $S \neq \{0\}$, and ϕ is onto, then $\phi(1)$ is the unity of S and units in R map to units in S.
7. ϕ is an isomorphism if and only if ϕ is onto and the kernel of ϕ, $ker\phi = \{r \in R | \phi(r) = 0\} = \{0\}$.
8. ϕ is an isomorphism from R onto S, then ϕ^{-1} is an isomorphism from S onto R.

Proof. Proofs are similar to those of groups. Left as exercise. □

Example 11.2.5. Show by an example that we can have a homomorphism $\phi : R \to S$, such that $\phi(1)$ is not unity in S, where 1 is the unity of R.

Solution. Consider the map $\phi : \mathbb{Z} \to \mathbb{Z}$ such that $\phi(x) = 0$ for all $x \in \mathbb{Z}$. Then ϕ is a homomorphism (verify!). Again, $\phi(1) = 0$, but 0 is not unity in $\mathbb{Z}$.

We could also take the same mapping from $\mathbb{Z}$ onto $2\mathbb{Z}$ to solve the given problem. ■

Example 11.2.6. Find all the ring homomorphisms from $\mathbb{Z}_{20} \to \mathbb{Z}_{30}$.

Solution. Let $f : \mathbb{Z}_{20} \to \mathbb{Z}_{30}$ be any ring homomorphism and $f(1) = a$. Then $f(x) = f(x.1) = xa$ by the additive property of f. Also since f satisfies the definition of group homomorphism, so $o(a)|o(\mathbb{Z}_{30}) = 30$ and $o(a)|o(\mathbb{Z}_{20}) = 20$. Thus, the possible values of $o(a)$ are $1, 2, 5, 10$ and so the possible values of a will be $0, 3, 6, 9, 12, 15, 18, 21, 24, 27$ which gives us ten group homomorphisms. Now, f is a ring homomorphism and in $\mathbb{Z}_{20}$, $1.1 = 1$, we find $f(1.1) = f(1) \Rightarrow f(1)f(1) = f(1) \Rightarrow a^2 = a$ in $\mathbb{Z}_{30}$. This is satisfied by $0, 6, 15, 21$. Hence, there are four ring homomorphisms from $\mathbb{Z}_{20} \to \mathbb{Z}_{30}$. ■

Example 11.2.7. Show that $2\mathbb{Z}$ is not isomorphic to $3\mathbb{Z}$ as rings. What can be said about the isomorphism between $m\mathbb{Z}$ and $n\mathbb{Z}$, where m, n are positive integers?

Solution. Suppose $2\mathbb{Z} \simeq 2\mathbb{Z}$ and let $f : 2\mathbb{Z} \to 3\mathbb{Z}$ be the isomorphism. As $2 \in 2\mathbb{Z}$, $f(2) = 3n$ for some $n \in \mathbb{Z} \setminus \{0\}$. Now,

$$f(4) = f(2+2) = f(2) + f(2) = 6n.$$
$$f(4) = f(2.2) = f(2).f(2) = (3n)^2.$$

Thus, $6n = 9n^2 \Rightarrow 2 = 3n$. But this is not possible for any $n \in \mathbb{Z}$. Hence, f is not an isomorphism.

Suppose now $f : m\mathbb{Z} \to n\mathbb{Z}$ is any ring isomorphism. Then

$$\begin{aligned} & f(m + m + \ldots + m) = f(m) + f(m) + \ldots + f(m) \\ & \qquad\qquad m \text{ times} \\ \Rightarrow \quad & f(mm) = mf(m) \\ \Rightarrow \quad & f(m)f(m) = mf(m) \\ \Rightarrow \quad & f(m) = m. \end{aligned} \tag{11.2.1}$$

Again as f is onto and $n \in n\mathbb{Z}$, there exists $mr \in m\mathbb{Z}$ such that

$$f(mr) = n \Rightarrow rf(m) = n \Rightarrow f(m)|n.$$

Again, as $m \in m\mathbb{Z}$, $f(m) \in n\mathbb{Z} \Rightarrow f(m) = nk$ for some k. Thus, $n|f(m)$ and hence $f(m) = n$. This and equation (11.2.1) together implies that $m = n$.

So, if $m\mathbb{Z} \simeq n\mathbb{Z}$ then $m = n$. The converse of course is obviously true.

Hence we conclude that $m\mathbb{Z} \simeq n\mathbb{Z}$ if and only if $m = n$. ■

Example 11.2.8. Show that the only homomorphism from $\mathbb{Z}$ to itself is the identity or the zero mappings.

Solution. Let $f : \mathbb{Z} \to \mathbb{Z}$ be a homomorphism. Since

$$(f(1))^2 = f(1)f(1) = f(1.1) = f(1),$$

so $f(1)[f(1) - 1] = 0 \Rightarrow f(1) = 0,$ or $f(1) = 1$.

If $f(1) = 0$ then $f(x) = 0$ for all $x \in \mathbb{Z}$. Thus in this case f is the zero homomorphism.

If $f(1) = 1$, then for all $x \in \mathbb{Z}$,

$$\begin{aligned} & f(x) = f(1 + 1 + \ldots + 1) = xf(1) = x \ (x > 0) \\ & f(x) = f(-y) = -f(y) = -[f(1 + 1 + \ldots + 1)] = -yf(1) = xf(1) = x \ (x < 0, \ y = -x) \\ & f(0) = 0. \end{aligned}$$

So in this case f is identity map , which proves the result. ■

Example 11.2.9. Let R and S be two commutative rings with unity and let $f : R \to S$ be an onto homomorphism. If char$R \neq 0$, show that char$S$ divides char$R$.

Solution. Suppose char$R = n$. Then $n.1 = 0$, n being the smallest such integer and 0 being the additive identity of R. Thus, $1 + 1 + \ldots + 1 = 0$ and so additive order of 1 is n. Again as f is onto, $f(1)$ is unity of S and so char$S$ is additive order of $f(1)$. As $o(f(1))|o(1)$, we find the desired result. ■

Exercise 11.2.10.

1. Find all the ring homomorphisms from $\mathbb{Z}_{12}$ to $\mathbb{Z}_{30}$.
2. Determine all ring homomorphisms from $\mathbb{Z}_{25}$ to $\mathbb{Z}_{20}$.
3. Determine all ring homomorphisms from $\mathbb{Z}_n$ to itself.

4. Let
$$R = \left\{ \begin{bmatrix} a & b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{Z} \right\}$$
and let ϕ be the mapping that takes $\left(\begin{smallmatrix} a & b \\ b & a \end{smallmatrix}\right)$ to $a - b$.
   (a) Show that ϕ is a homomorphism.
   (b) Determine the kernel of ϕ
   (c) Show that $R/Ker\phi$ is isomorphic to $\mathbb{Z}$
   (d) Is $Ker\phi$ a prime ideal? Is it maximal?

Theorem 11.2.11. Let ϕ be a ring homomorphism from a ring R to a ring S. Then $ker\phi = \{r \in R \mid \phi(r) = 0\}$ is an ideal of R.

Proof. Let $x, y \in ker\phi$. Then $\phi(x) = \phi(y) = 0$. Then $\phi(x - y) = \phi(x) - \phi(y) = 0$. Thus, $x - y \in ker\phi$. Also, for $r \in R$, we have $\phi(rx) = \phi(r)\phi(x) = \phi(r).0 = 0$. Hence, $rx \in Ker\phi$. Thus, $Ker\phi$ is an ideal of R. □

From the above theorem, we can have an ideal from a homomorphism. The converse is also true. Suppose I is an ideal of a ring R. Then it induces a homomorphism from R onto the factor ring R/I, which is called the *canonical homomorphism* for I.

Theorem 11.2.12. Let I be an ideal of a ring R. Then the function $\phi : R \to R/I$ defined by $\phi(r) = r + I$ is a ring homomorphism and its kernel is I.

Proof. For all $r, s \in R$,
$$\phi(r + s) = (r + s) + I = (r + I) + (s + I) = \phi(r) + \phi(s)$$
and
$$\phi(rs) = rs + I = (r + I)(s + I) = \phi(r)\phi(s).$$
So, ϕ is a ring homomorphism. Further,
$$\begin{aligned} Ker\phi &= \{r \in R \mid \phi(r) = 0 + I = I\} \\ &= \{r \in R \mid r + I = I\} \\ &= \{r \in R \mid r \in I\} \\ &= I. \end{aligned}$$
□

Theorem 11.2.13. (First Isomorphism Theorem) Let ϕ be a ring homomorphism from R to S. Then $R/Ker\phi \simeq \phi(R)$.

Proof. Put $I = Ker\phi(R)$. Define $f : R/I \to \phi(R)$ by $f(r + I) = \phi(r)$. Then, for $r, s \in R/I$, we have
$$f((r + I) + (s + I)) = f(r + s + I) = \phi(r + s) = \phi(r) + \phi(s) = f(r + I) + f(s + I)$$
$$f((r + I)(s + I)) = f(rs + I) = \phi(rs) = \phi(r)\phi(s) = f(r + I)f(s + I).$$
Hence, f is a ring homomorphism. Further, for $f(r + I) = f(s + I)$ we have, $\phi(r) = \phi(s) \Rightarrow \phi(r - s) = 0 \Rightarrow r - s \in I \Rightarrow r + I = s + I$. Hence, f is injective. For surjectivity, let $y \in \phi(R)$. Then, there exists $r \in R$ such that $\phi(r) = y$. Now, $r + I \in R/I$ such that $f(r + I) = \phi(r) = y$. Hence f is surjective. Thus the result. □

Corollary 11.2.14. If $\phi : R \to S$ be a surjective homomorphism. Then, $S \simeq R/Ker\phi$.

Theorem 11.2.15. Let R be a ring with unity 1. The mapping $\phi : \mathbb{Z} \to R$ given by $n \to n.1$ is a ring homomorphism.

Proof. Since the multiplicative group property $a^{m+n} = a^m a^n$ translates to $(m+n)a = ma + na$ when the operation is addition, we have

$$\phi(m+n) = (m+n).1 = m.1 + n.1$$

So, ϕ preserves addition.

ϕ also follows multiplication. We know that $(m.a)(n.b) = (mn)(ab)$ for all integers. Thus,

$$\phi(mn) = (mn).1 = (mn).((1)(1)) = (m.1)(n.1) = \phi(m)\phi(n)$$

So, ϕ preserves multiplication as well. □

Corollary 11.2.16. If R is a ring with unity and characteristic of R is $n > 0$, then R contains a subring isomorphic to $\mathbb{Z}_n$. If the characteristic of R is 0, then R contains a subring isomorphic to $\mathbb{Z}$.

Proof. Let 1 be the unity of ring R and let $S = \{k.1 | k \in \mathbb{Z}\}$. The previous theorem shows that ϕ from $\mathbb{Z}$ to S given by $\phi(k) = k.1$ is a homomorphism, and by the First Isomorphism theorem for rings, we have

$$\mathbb{Z}/Ker\phi \simeq S.$$

Clearly, $Ker\phi = \langle n \rangle$, where n is the additive order of 1 and n is also the characteristic of R. So, R has characteristic n,

$$S \simeq \mathbb{Z}/\langle n \rangle \simeq \mathbb{Z}_n.$$

When R has characteristic 0, then

$$S \simeq \mathbb{Z}/\langle 0 \rangle \simeq \mathbb{Z}.$$

□

Corollary 11.2.17. For any positive integer m, the mapping $\phi : \mathbb{Z} \longrightarrow \mathbb{Z}_m$ given by $x \longrightarrow x \bmod m$ is a ring homomorphism.

Proof. Left as exercise. □

Corollary 11.2.18. If F is a field of characteristic p, then F contains a subfield isomorphic to $\mathbb{Z}_p$. If F is a field of characteristic 0, then F contains a subfield isomorphic to $\mathbb{Q}$.

Theorem 11.2.19. (Second Isomorphism Theorem) Let $B \subseteq A$ be two ideals of a ring R. Then

$$R/A \simeq (R/B)/(A/B).$$

Proof. Define a mapping $f : R/B \to R/A$ such that $f(r+B) = r + A$. Then it is easy to check that f is an onto homomorphism (verify!). By the first isomorphism theorem, $(R/B)/Kerf \simeq R/A$. We will be done if we find the kernel of f. We have

$$\begin{aligned} Kerf &= \{r + B \in R/B \mid f(r+B) = A\} \\ &= \{r + B \in R/B \mid r + A = A\} \\ &= \{r + B \in R/B \mid r \in A\} \\ &= A/B. \end{aligned}$$

Hence the result. □

Theorem 11.2.20. (Third Isomorphism Theorem) Let A, B be two ideals of a ring R. Then

$$(A+B)/A \simeq B/(A \cap B).$$

Proof. Define a mapping $f : B \to (A+B)/A$ such that $f(b) = b + A$ for all $b \in B$. Then f is a well-defined homomorphism (check!). Again, if $x + A \in (A + B)/A$ be any element then

$$x \in A + B \Rightarrow x = a + b,\ a \in A,\ b \in B.$$

So,

$$x + A = (a + b) + A = (b + a) + A = b + (a + A) = b + A.$$

Thus, $x + A = b + A = f(b)$, that is, b is the pre-image of $x + A$ under f or that f is onto. By the first isomorphism theorem, $(A + B)/A \simeq B/Kerf$.

Now,

$$\begin{aligned} x \in Kerf &\Leftrightarrow f(x) = A \\ &\Leftrightarrow x + A = A \\ &\Leftrightarrow x \in A \\ &\Leftrightarrow x \in A \cap B \ \ (x \in Kerf \subseteq B). \end{aligned}$$

Hence, $Kerf = A \cap B$. Hence the result. □

Example 11.2.21. Show that $\mathbb{Z}/\langle 2\rangle \simeq 5\mathbb{Z}/10\mathbb{Z}$.

Solution. Take $A = \langle 2\rangle = 2\mathbb{Z}$, $B = \langle 5\rangle = 5\mathbb{Z}$, the ideals of $\mathbb{Z}$. Then $A + B = \langle d\rangle$, where $d = gcd(2, 5) = 1$. $A \cap B = \langle l\rangle$, where $l = lcm(2, 5) = 10$. So, $A + B = \mathbb{Z}$ and $A \cap B = 10\mathbb{Z}$. Hence using the third isomorphism theorem, we get the desired result. ■

Exercise 11.2.22.

1. Prove that the ring $\mathbb{Z}_3/\langle x^2 + 1\rangle$ is isomorphic to the field $\mathbb{Z}_3[i]$.
2. For any integer $n > 1$, prove that $\mathbb{Z}_n[x]/\langle x\rangle$ is isomorphic to $\mathbb{Z}_n$.
3. Is there a ring homomorphism from the reals to some ring whose kernel is the integers?

Sample Questions

1. Show that the homomorphic image of a commuattive ring is commutative. Is converse true? Justify your answer.
2. Show that the homomorphic image of a ring with unity is an ring with unity. Is the converse true? Justify your answer.
3. State and prove the first isomorphism theorem.
4. State and prove the second isomorphism theorem.
5. State and prove the third isomorphism theorem.
6. Show that if m and n are distinct positive integers, then $m\mathbb{Z}$ is not ring-isomorphic to $n\mathbb{Z}$.
7. Determine all ring homomorphisms from $\mathbb{Z}$ to $\mathbb{Z}$.

Unit 12

Course Structure

- Polynomial rings: definition and properties
- Division algorithm and its applications

12.1 Introduction

We have primarily come across integer or real polynomials in one variable, mostly x, (having coefficients in $\mathbb{Z}$ and $\mathbb{R}$ respectively) previously. But we can also form polynomials using elements from any arbitrary ring R. Such polynomials along with addition and multiplication, as defined in unit 9, also forms a ring, called the polynomial rings. The study of polynomial rings are closely related to the idea of field extensions, vector spaces, etc. It is needless to say that the idea of polynomial rings in more than one variable can be drawn from polynomial rings in one variable.

Objectives

After reading this unit, you will be able to

- define polynomial rings and come across several examples and properties of polynomial rings
- come across the idea of division algorithm and its applications

12.2 Polynomial rings

We are all familiar with expressions like, $x^2 + 4x + 1$. This is a polynomial with integer coefficients. We can factorise them, find their roots, etc. Likewise, here we will look into those polynomials with coefficients from commutative rings only.

Definition 12.2.1. Let R be a ring. A polynomial in x with coefficients from R is an expression of the type $f(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_i \in R$, $i = 0, 1, \ldots, n$.

Example 12.2.2. 1. $f(x) = x^4 + 4x^3 + 1$ is a polynomial with coefficients in $\mathbb{Z}$.

2. $f(x) = x^3 + (4+i)x^2 + 5 + 6i$ is a polynomial with coefficients from the ring of Gaussian integers, that is, $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$.

3. If $R = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$, then

$$f(x) = (3 + 6\sqrt{-5})x^7 + 5x^3 + (8 + 7\sqrt{-5})x + (1 + \sqrt{-5})$$

is a polynomial with coefficients from R.

We now recapitulate the definitions of addition and multiplication between two polynomials from a set of polynomials with coefficients from R.

Definition 12.2.3. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$, where $a_n,\ b_m \neq 0$ be two polynomials over R. Then their sum and multiplication are defined as,

$$(f + g)(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots$$

and,

$$(fg)(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{m+n}x^{m+n}$$

where, c_i is defined as

$$c_i = \sum_{k=0}^{i} a_k b_{i-k}$$

for $i = 0, 1, \ldots, m + n$.

Then it is clear that addition is associative since addition in R is so. Also, the zero element is given by the polynomial

$$0(x) = 0 + 0x + 0x^2 + \cdots$$

The inverse element of any polynomial $f(x) = a_0 + a_1x + \cdots + a_nx^n$ is given by $-f(x) = -a_0 - a_1x - \cdots - a_nx^n$. by virtue of all these properties, we have the following theorem:

Theorem 12.2.4. The set of polynomials over the ring R forms a ring under addition and multiplication. This ring of polynomials is denoted by $R[x]$.

Example 12.2.5. The ring of polynomials with coefficients from $\mathbb{R}$, $\mathbb{C}$ are denoted by $\mathbb{R}[x]$, $\mathbb{C}[x]$ respectively.

Exercise 12.2.6. 1. Compute (i) $(x^2 + 2x + 2) + (x^2 + 3)$ (ii) $(x^2 + 2x + 2)(x^2 + 3)$ in $\mathbb{Z}_5[x]$

2. Compute (i) $(2x^2 + 1) + (4x^2 + 5)$ (ii) $(3x + 2)(2x + 3)$ in $\mathbb{Z}_6[x]$.

$R[x]$ may or may not have unit element. For example, consider the ring $2\mathbb{Z}$. It has no unit element.The polynomial ring over $2\mathbb{Z}$ also does not have unit element. Also see that $\mathbb{R}$, $\mathbb{C}$ both have unit element 1 and also their corresponding polynomials have so. So we see that the existence of unit element in R can be a possible requisite for $R[x]$ to contain so. We can see from the following theorem:

Theorem 12.2.7. Let R be a commutative ring with unit element 1. Then, $R[x]$ is also a commutative ring with unit element.

Proof. Since R is commutative, for any $a,\ b \in R$, $a.b = b.a$. So, for $f(x) = a_0 + a_1x + \cdots + a_nx^n$, and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ in $R[x]$, we have

$$(fg)(x) = c_0 + c_1x + c_2x^2 + \cdots + c_{m+n}x^{m+n}$$

where, c_i is given as

$$\begin{aligned} c_i &= \sum_{k=0}^{i} a_k b_{i-k} \\ &= \sum_{k=0}^{i} b_{i-k} a_k \end{aligned}$$

since R is commutative. This shows that, $fg = gf$, for any polynomials $f(x)$ and $g(x)$ in $R[x]$. Also, if we define the polynomial $1(x) = 1 + 0x + 0x^2 + \cdots$, then we can easily check that, for any polynomial $f(x) \in R[x]$, we have $f.1 = f$. So, the above defined polynomial is the unit element in $R[x]$. □

All the examples of polynomial rings given before are polynomial rings with unity. The next theorem that directly follows from the above is

Theorem 12.2.8. If R is an integral domain, then $R[x]$ is so.

Proof. Since, R is an integral domain, it is a commutative ring with unity. By the previous theorem, we can say that $R[x]$ is also so. We are left only to show that $R[x]$ does not contain any divisors of zero. For this, let us assume that $f(x) = a_0 + a_1x + \cdots + a_nx^n$, and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ are two polynomials in $R[x]$, with $a_n, b_m \neq 0$. So, we have,

$$(fg)(x) = a_0b_0 + (a_1b_0 + b_1a_0)x + \cdots + a_nb_mx^{m+n}$$

Since R does not contain any zero divisor, and $a_n, b_m \neq 0$, so $a_nb_m \neq 0$. Hence, $fg \neq 0$. Thus, $R[x]$ does not contain any zero divisor. Hence the result. □

Corollary 12.2.9. If R is a field, then $R[x]$ is an integral domain.

But can we say that $R[x]$ is a field if R is so? It is evident that the constant polynomials of $R[x]$ are units and have an inverse in $R[x]$. Indeed, we have the following result.

Theorem 12.2.10. Let R be an integral domain with unity 1. Then the units of $R[x]$ are the same as those of R.

To prove the above theorem, we will need a few more results.

Definition 12.2.11. Let $f(x)$ be a non-zero polynomial in $R[x]$. Then the largest n for which the coefficient of x^n is non-zero, is called the *degree* of f. It is generally denoted by $\deg f$.

Theorem 12.2.12. Let R be any commutative ring, $f(x), g(x) \in R[x]$. Then $\deg(fg) \leq \deg(f) + \deg(g)$ and equality holds when R is an integral domain.

Proof. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_n \neq 0$, so that $\deg f = n$ and let $g(x) = b_0 + b_1x + \cdots + b_mx^m$, $b_m \neq 0$ with $\deg g = m$. Then, $f(x)g(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + \cdots + a_nb_mx^{m+n}$. Thus, $\deg(fg) \leq m+n = \deg f + \deg g$. If R is an integral domain, then $a_nb_m \neq 0$. So, $\deg(fg) = m+n = \deg f + \deg g$. □

Corollary 12.2.13. If F is a field, $\deg(fg) = \deg f + \deg g$ and in particular $\deg(fg) \geq \deg f$, as $\deg g \geq 0$.

Definition 12.2.14. If $f(x) \in R[x]$ is such that $\deg f = 0$, it is called a constant polynomial.

If $\deg f = 0$, then $f(x) = a_0 \in R$ and conversely, if $a \in R$, then a can be written as $a + 0x + 0x^2 + ...$ of degree 0. Thus the constant polynomials can be identified with elements of R and under this identification, R is a subring of $R[x]$.

Theorem 12.2.15. Let R be an integral domain. Then If $f, g \in R[x]$, then

$$\deg(f + g) \leq \max\{\deg f, \deg g\}.$$

Proof. Let $f(x) = a_0 + a_1x + \ldots + a_nx^n \in R[x]$, $a_n \neq 0$ and $g(x) = b_0 + b_1x + \ldots + b_mx^m \in R[x]$, $b_m \neq 0$. Then,

$$\begin{aligned}(f+g)(x) = f(x) + g(x) &= (a_0 + a_1x + \ldots + a_nx^n) + (b_0 + b_1x + \ldots + b_mx^m) \\ &= (a_0 + b_0) + (a_1 + b_1)x + \ldots + a_nx^n + \ldots + b_mx^m \text{ if } m > n \\ &= (a_0 + b_0) + (a_1 + b_1)x + \ldots + b_mx^m + \ldots + a_nx^n \text{ if } n > m \\ &= (a_0 + b_0) + (a_1 + b_1)x + \ldots + (a_n + b_m)x^n \text{ if } n = m \text{ and } a_n \neq -b_m.\end{aligned}$$

Hence the result. □

Note that the result may also be true for arbitrary rings instead of integral domains. Also, there may be cases when we can get strict inequalities.

Now we prove theorem 12.2.10.

Proof. Let $f(x) \in R[x]$ be a unit so that there exists some $g(x) \in R[x]$ such that $f(x)g(x) = g(x)f(x) = 1$. Now, $0 = \deg 1 = \deg(fg) = \deg f + \deg g$. Thus, $\deg f = \deg g = 0$, that is, f and g are constant polynomials. The relation $fg = gf = 1$ implies that f and g are units in R. Thus, the units of $R[x]$ are units of R. Conversely, the units of R is also a unit of $R[x]$. □

Exercise 12.2.16.

1. What are the units of the ring $\mathbb{Z}_7[x]$?
2. Let R be a commutative ring with 1 and $p(x) = \sum_{j=0}^{n} a_jx^j \in R[x]$. Prove that $p(x)$ is a unit if and only if a_0 is a unit and a_j are nilpotent for all $j \geq 1$.
3. Give examples of polynomials $f, g \in R[x]$ such that $\deg(f + g) < \max\{\deg f, \deg g\}$.
4. Give examples of polynomials $f, g \in R[x]$ such that $\deg(fg) < \deg f + \deg g$.

12.2.1 Division Algorithm

Theorem 12.2.17. Let F be a field and let $f(x), g(x) \in F[x]$ with $g(x) \neq 0$. Then there exists unique polynomials $q(x)$ and $r(x)$ in $F[x]$ such that

$$f(x) = g(x)q(x) + r(x)$$

and either $r(x) = 0$ or $\deg r(x) < \deg g(x)$.

Proof. If $f(x) = 0$ or $\deg f(x) < \deg g(x)$, we simply set $q(x) = 0$ and $r(x) = f(x)$. So, we assume that $n = \deg f(x) \geq \deg g(x) = m$. Let

$$\begin{aligned} f(x) &= a_n x^n + \cdots + a_0, \text{ and} \\ g(x) &= b_m x^m + \cdots + b_0 \end{aligned}$$

By, doing long division method, we let

$$f_1(x) = f(x) - a_n b_m x^{n-m} g(x)$$

Then, $f_1(x) = 0$ or $\deg f_1(x) < \deg f(x)$. By induction hypothesis, there exists $q_1(x)$ and $r_1(x)$ in $F[x]$ such that

$$f_1(x) = g(x)q_1(x) + r_1(x)$$

where, $r_1(x) = 0$ or $\deg r_1(x) < \deg g(x)$. Thus,

$$\begin{aligned} f(x) &= a_n b_m^{-1} x^{n-m} g(x) + f_1(x) \\ &= a_n b_m^{-1} x^{n-m} g(x) + g(x)q_1(x) + r_1(x) \\ &= [a_n b_m^{-1} x^{n-m} + q_1(x)]g(x) + r_1(x) \end{aligned}$$

So the polynomials

$$\begin{aligned} q(x) &= a_n b_m^{-1} x^{n-m} + q_1(x), \ \textit{and} \\ r(x) &= r_1(x) \end{aligned}$$

have the desired properties.
To prove the uniqueness, let

$$\begin{aligned} f(x) &= g(x)q(x) + r(x) \ \textit{and} \\ f(x) &= g(x)\bar{q}(x) + \bar{r}(x) \end{aligned}$$

where, and either $r(x) = 0$ or $\deg r(x) < \deg g(x)$ and and either $\bar{r}(x) = 0$ or $\deg \bar{r}(x) < \deg g(x)$. Subtracting these two equations, we obtain

$$\begin{aligned} 0 &= g(x)[q(x) - \bar{q}(x)] + [r(x) - \bar{r}(x)] \\ \bar{r}(x) - r(x) &= g(x)[q(x) - \bar{q}(x)] \end{aligned}$$

Thus, $\bar{r}(x) - r(x) = 0$, or the degree of $\bar{r}(x) - r(x)$ is at least that of $g(x)$. Since the latter is impossible, we have $\bar{r}(x) = r(x)$ and $q(x) = \bar{q}(x)$ as well. □

The polynomials $q(x)$ and $r(x)$ in the division algorithm are called *quotient* and *remainder* in the division.

12.2.2 Remainder Theorem

Corollary 12.2.18. Let F be a field, $a \in F$, and $f(x) \in F[x]$. Then $f(a)$ is the remainder in the division of $f(x)$ by $x - a$.

Proof. Left as exercise. □

Example 12.2.19. Divide $3x^4 + 2x^3 + x + 2$ by $x^2 + 4$ in $\mathbb{Z}_5[x]$. As we follow the division, we note that $-4 = 1, -3 = 2$, and $-2 = 3-$ we are doing arithmetic mod 5.

$$\begin{array}{r|l}
 & 3x^2 + 2x + 3 \\
\hline
x^2 + 4 & 3x^4 + 2x^3 + x + 2 \\
 & 3x^4 \quad + 2x^2 \\
 & - \qquad - \\
\hline
 & 2x^3 + 3x^2 + x \\
 & 2x^3 \quad + 3x \\
 & - \qquad - \\
\hline
 & 3x^2 + 3x + 2 \\
 & 3x^2 \quad + 2 \\
 & - \qquad - \\
\hline
 & 3x
\end{array}$$

The quotient is $3x^2 + 2x + 3$ and remainder is $3x$. Thus,

$$3x^4 + 2x^3 + x + 2 = (3x^2 + 2x + 3)(x^2 + 4) + 3x.$$

12.2.3 Factor Theorem

Corollary 12.2.20. Let F be a field, $a \in F$, and $f(x) \in F[x]$. Then a is the zero of $f(x)$ if and only if $x - a$ is a factor of $f(x)$.

Proof. Left as exercise. □

Corollary 12.2.21. A polynomial of degree n over a field has at most n zeros, counting multiplicities.

Proof. We use induction over n. Clearly, a polynomial of degree 0 over a field has no zeros. Now, let $f(x)$ is a polynomial of degree n over a field and a is a zero of $f(x)$ of multiplicity k. Then

$$\begin{aligned} f(x) &= (x-a)^k q(x) \ \textit{and} \\ q(a) &\neq 0 \end{aligned}$$

and since

$$\begin{aligned} n &= \deg f(x) \\ &= (x-a)^k q(x) \\ &= k + \deg q(x) \end{aligned}$$

we have $k \leq n$. If $f(x)$ has no zeros other than a, we are done. On the other hand, if $b \neq a$ and b is a zero of $f(x)$, then

$$0 = f(b) = (b-a)^k q(b)$$

so that b is also a zero of $q(x)$ with the same multiplicity as it has for $f(x)$. By the Second Principle of Induction, we know that $q(x)$ has at most $\deg q(x) = n - k$ zeros, counting multiplicity. Thus $f(x)$ has at most $k + n - k = n$ zeros, counting multiplicity. □

Exercise 12.2.22.

1. If R and R' are two isomorphic rings, show that $R[x]$ and $R'[x]$ are isomorphic.
2. Show that for any R, $R[x]$ can never be a field.

Sample Questions

1. If R is an integral domain, show that $R[x]$ is also so.
2. For a commutative ring R, show that $\deg(fg) \leq \deg f + \deg g$, for $f(x), g(x) \in R[x]$. Can we write equality in the given inequation? Justify your answer.
3. In an integral domain R, show that $\deg(f + g) \leq \max\{\deg f, \deg g\}$, for $f(x), g(x) \in R[x]$.
4. Show that the units in an integral domain are precisely those in $R[x]$.
5. State and prove the division algorithm in $F[x]$, for a field F.

Unit 13

Course Structure

- Domains in rings
- Classification of domains
- Irreducible Polynomials and Eisenstein's Criterion for irreducibility

---

13.1 Introduction

This unit is dedicated to the study of different kinds of rings. We have come across integral domains and fields in our preceding units. A field is an integral domain (inclusion is shown in figure 13.1.1). The converse is not true as $\mathbb{Z}$ is an integral domain without being a field. There are however certain types of domain that

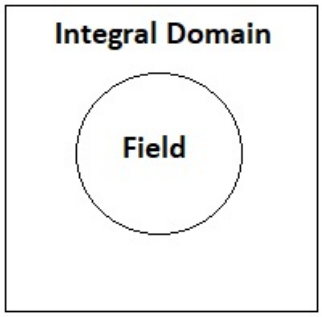


Figure 13.1.1: Basic Inclusion in domains

comes in between these two extremes. They are the Euclidean domains (ED), Principal Ideal Domains (PID) and Unique Factorisation Domains (UFD). We will explore the inclusions of these domains and their basic properties. Having equipped with the basic idea on domains, we will discuss the irreducibility of polynomials under certain conditions.

Objectives

After reading this unit, you will be able to

- define various types of domains and explore their characteristics
- define irreducibility of polynomials and learn certain criteria to check for the irreducibility of polynomials in a given polynomial ring

13.2 Euclidean Domain

Definition 13.2.1. A *Euclidean domain* is an integral domain R in which there exists an integer valued function d on the non-zero elements of R, satisfying the following conditions:

1. $d(a) \geq 0$ for all non-zero $a \in R$.
2. $d(ab) \geq d(a)$, $a, b \in R$
3. For $a, b \in R$, $b \neq 0$ there exists $q, r \in R$ such that $a = bq + r$ with either $r = 0$ or $d(r) < d(b)$.

Example 13.2.2.
1. If $R = \mathbb{Z}$, then d can be defined on R as $d(a) = |a|$. Then check that d satisfies all the conditions of an Euclidean domain.
2. Let $R = \{a+ib : a, b \in \mathbb{Z}\}$ be the ring of Gaussian Integers. The if d is defined as $d(z) = |z|^2 = a^2+b^2$, where, $z = a + ib$, then R forms a Euclidean Domain.
3. Let R be a field. Then, d can be defined as, $d(a) = a.a^{-1}$. This definition is well defined since each element a has an inverse. d satisfies all the properties of an integral domain.

Theorem 13.2.3. Let R be an Euclidean Domain. Then, every ideal I of R is of the form $I = Ra$ for some $a \in R$.

Proof. If $I = 0$, then $a = 0$ and we are done. So, we take $I \neq 0$. Choose some $a \neq 0$ such that $d(a)$ is the least in R. Such an a exists by well-ordeing principle of real numbers. We claim that $I = Ra$. Since $a \in R$, so $Ra \subset I$. Let $b \in I$. By the third condition of the definition of Euclidean domain, there exists q and r in R such that $b = aq + r$, where, either $r = 0$ or $d(r) < d(a)$. Now, $r = b - aq$ is in R. If $d(r) < d(a)$, then this contradicts the fact that a is the element with the least value of d. Then obviously $r = 0$. Thus, we get $b = aq$. Hence, $I \subset Ra$. Combining. we get $I = Ra$. □

Theorem 13.2.4. Let R be a Euclidean Domain. Then any two elements a, b in R, have a gcd.

Proof. Let $I = Ra$ and $J = Rb$. Then $I + J$ is also an ideal of R. By the previous theorem, we have $I + J = Rd$ for some d in R. We claim that d is the gcd of a and b. We can see that $I = Ra \subset Rd$ and $J = Rb \subset Rd$. Hence, $d|a$ and $d|b$. Let d' be another element in R such that $d'|a$ and $d'|b$. So, $Ra \subset Rd'$ and $Rb \subset Rd'$. This implies that $Rd = I + J \subset Rd'$. Hence, $d'|d$. Hence, by definition, d is the required gcd. □

Theorem 13.2.5. Let R be a Euclidean domain. Then, $a \in R$ is a unit if and only if $d(a) = d(1)$.

Proof. Let a be a unit in R. Then, $a.a^{-1} = 1$. By the second condition of d, $d(1) = d(a.a^{-1}) \geq d(a)$. Also, $d(a) = d(a.1) \geq d(1)$. So, combining, we get $d(a) = d(1)$. Conversely, let $d(a) = d(1)$. By the third condition of d, we have $1 = qa + r$, with either $r = 0$ or $d(r) < d(a) = d(1)$. But, $d(r) < d(1)$ is impossible since $d(r) = d(r.1) \geq d(1)$. Hence, $r = 0$. Thus, $aq = 1$. Hence, a is a unit in R. □

Theorem 13.2.6. Let R be a Euclidean Domain and a, b be two elements in R. If a is a proper divisor of b, then $d(a) < d(b)$.

Proof. Since $a|b$, there exists some $c \in R$ such that $b = ac$. So we have, $d(b) \geq d(a)$. We will show that $d(b) \neq d(a)$. If possible, $d(a) = d(b)$. Then, for any $x \in Ra$ we have $d(x) \geq d(a) = d(b)$ which means that $d(b)$ is least in Ra. This implies that, $Ra = Rb$, which implies that a and b are associates which is a contradiction. Hence the result. □

Theorem 13.2.7. Let R be an Euclidean Domain. Then any $a \in R$ which is not a unit can be expressed as a product of irreducible elements.

Proof. If a is irreducible, then there is nothing to prove. Otherwise, it has a proper divisor b, that is, $a = bc$. Then, $d(b) < d(a)$ and $d(c) < d(a)$. If b and c are irreducible, then we are done. If $b(or\ c)$ is irreducible, then we can write $b = ef$, where $d(e) < d(b)$ and $d(f) < d(b)$. If this process is continued, then after a finite number of stages all the factors will be irreducible since the values of d is strictly reducing at each stage. Thus, after a finite number of steps, all the factors will be irreducible and a will be expressible as the product of irreducible elements. □

Example 13.2.8. Let R be an Euclidean Domain and let

$$I = \{a : d(a) > d(1)\} \cup \{0\}$$

Then I does not form an ideal in R. In fact, if we consider the ring $\mathbb{Z}$, the only units in $\mathbb{Z}$ are ± 1. Now, 2, 3 belong to I but $2 - 3$ does not belong to I. Hence, I is not an ideal in this case.

Example 13.2.9. Let R be a Euclidean Domain. If d' be another function defined as $d'(a) = d(a) + m$, for some positive integer m, with $d(1) + m \geq 0$. Check that R is a Euclidean Domain with respect to the function d'.

Example 13.2.10. If R is a Euclidean Domain and for a, b in R, if $a|b$ and $d(a) = d(b)$, then, a and b are associates.

Exercise 13.2.11.

1. Show that the units in the ring of Gaussian integers, that is, $R = \{a + ib \mid a, b \in \mathbb{Z}\}$ are ± 1, $\pm i$.
2. Show that the ring of even integers is not an Euclidean domain.
3. Show that every field in an ED.
4. If a is an irreducible element in an ED R, show that Ra is a maximal ideal.

13.3 Principal Ideal Domain

Definition 13.3.1. An integral domain R is called a *Principal Ideal Domain* (PID) if every ideal of R is a principal ideal, that is, for any ideal I in R, there exists an element a in R such that $I = \langle a \rangle = Ra$.

It is clear from theorem 13.2.7, we can say that every Euclidean Domain is a Principal Ideal Domain. But the converse is not necessarily true. For example, the ring $R = \left\{ a + \frac{b}{2}(1 + \sqrt{-19}) : a, b \in \mathbb{Z} \right\}$ is a Principal ideal domain, but not a Euclidean domain.

Theorem 13.3.2. Any two elements in a PID R have gcd.

Proof. Similar as before. □

Theorem 13.3.3. Let R be a PID. Then every $a \in R$ can be expressed as a product of irreducible elements.

Proof. If $a \in R$ is irreducible, then there is nothing to prove. Otherwise, let $a = bc$, where b, c are proper divisors of a. If both b, c are irreducible, then we are done. Suppose, b is irreducible . Then we have

$$b = ef$$

where e, f are proper divisors of b. If we continue in this way, after a finite number of steps, all the factors will be irreducible for otherwise, there will be an infinite sequence of elements,

$$a_0 = a, a_1 = b, a_2 = e, \cdots, a_n, \ldots$$

such that a_{n+1} is a proper divisor of a_n. We will show that this is impossible.

Let such a sequence exists. Let $I_n = Ra_n$, so that we have an increasing sequence of ideals

$$I_0 \subset I_1 \subset I_2 \cdots I_n ....$$

Since a_{n+1} is a proper divisor of a_n, $I_n \neq I_{n+1}$. Let

$$I = \bigcup_{k=0}^{\infty} I_k.$$

Then I is an ideal of R, because if $a, b \in I$, then $a \in I$ and $b \in I$, where either $I_r \subset I_s$ or $I_s \subset I_r$, so that $a - b \in I_r \cup I_s \subset I$, and $xa \in I_r \subset I$ for all $x \in R$. Since R is a principal ideal domain, $I = \langle d \rangle$ for some $d \in R$. Now, $d \in I_m$ for some m, so that $I = \langle d \rangle \subset I_m \subset I_{m+1} \subset \cdots \subset I$, i.e., $I_m = I_{m+1} = \cdots = I$, a contradiction. Hence proved. □

Example 13.3.4. The ring $R = \{m/n : m, n \in \mathbb{Z}, n \text{ odd}\}$ is a principal ideal domain.

Exercise 13.3.5.

1. Show that a subring of a PID need not be a PID.
2. If R is a PID, and $p \in R$, $p \neq 0$. Show that the following conditions are equivalent:
   a) p is a prime
   b) p is an irreducible element
   c) Rp is a prime ideal
   d) Rp is a maximal ideal
3. Show that in a PID, every non-zero prime ideal is maximal.
4. Show that every ideal in a PID is contained in a maximal ideal.

13.4 Unique Factorisation Domain

Definition 13.4.1. An integral domain R is called a *Factorisation Domain* (FD) if every element $a \in R$, which is not a unit can be expressed as a product of irreducible elements.

Thus, the Euclidean domain and principal ideal domains are factorisation domains.

Definition 13.4.2. $a \in R$ is said to be expressible uniquely as a product of irreducible elements if whenever $a = p_1 p_2 \ldots p_m = q_1 q_2 \ldots q_n$, where p_i, q_j are irreducible, then $m = n$ and each $p_i = u_i q_i$, where u_i is a unit in some order.

If $a \in R$ be expressed as a product of irreducible elements, the expression need not be unique.

Example 13.4.3. Let $R = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$. Then the only units in R are ± 1. Also, $1 + 2\sqrt{-5}$ is an irreducible element. Similarly, we can show that 3 and 7 are irreducible elements in R.Then

$$21 = 3.7 = (1 + 2\sqrt{-5})(1 - 2\sqrt{-5}).$$

Hence there are two distinct factorisations of 21 into irreducible elements of R since $1 \pm 2\sqrt{-5}$ are not associates of 3 or 7.

Definition 13.4.4. An integral domain R is said to be a *Unique Factorisation Domain* (UFD) if every $a \in R$ which is not a unit can be expressed uniquely as a product of irreducible elements.

Theorem 13.4.5. Let R be an integral domain in which

1. Every $a \in R$ which is non-unit can be expressed as a product of irreducible elements.

2. Every irreducible element is prime.

Then R is a ufd.

Proof. It is sufficient to show that factorisation is unique. Let $a = p_1 p_2 \dots p_m = q_1 q_2 \dots q_n$, where p_i, q_j are irreducible, and hence prime. Since $p_1 | a$, we have, $p_1 | q_1 q_2 \dots q_n$. So, $p_1 | q_j$ for some j. Without any loss of generality, let us assume that $p_1 | q_1$. Since q_1 is irreducible,and p_1 is not a unit, p_1 is an associate of q_1, that is, $q_1 = u_1 p_1$, where u_1 is a unit. Thus, $p_1 p_2 \dots p_m = (u_1 p_1) q_2 \dots q_n$. Since R is an integral domain, we have, $p_2 p_3 \dots p_m = u_1 q_2 q_3 \dots q_n$. Repeating the same process for p_2, we have, $q_2 = u_2 p_2$, where u_2 is a unit. Continuing this process we must have neither p_t not q_t left after a finite number of steps, for otherwise, in either case, a unit will be expressible as a product of irreducible elements, which is impossible. Thus, $m = n$, and each $p_t = u_i q_i$, where u_i is a unit. □

Corollary 13.4.6. Let R be a Euclidean domain or a PID. Then, R is a ufd.

Proof. It is sufficient to show that every irreducible element is prime. Let p be an irreducible element and $p | ab$. Consider the $\gcd(p, a)$ of p and a. It is either p or 1, and if $\gcd(p, a) = p$, then $p | a$. If $\gcd(p, a) = 1$, then $\lambda p + \mu a = 1$ for some $\lambda, \mu \in R$. Multiplying both sides by b, we have, $\lambda pb + \mu ab = b$. Since $p | ab$, it follows that $p | b$. Hence p is a prime. □

There exists ufd's which are not PID's (or ED's).

Theorem 13.4.7. If R is a ufd, then any two elements of R has a gcd.

Exercise 13.4.8.

1. Show that in a UFD, $a | c,\ b | c$ and $\gcd(a, b) = 1$ implies $ab | c$.

2. Show that in a UFD, $\gcd(a, c) = 1$, $\gcd(b, c) = 1$ implies $\gcd(ab, c) = 1$.

3. In a FD R, if any two elements have a gcd, show that R is a UFD.

13.5 Irreducible Polynomials

Let us use the idea of domains in the context of polynomial rings and try to establish new results in this direction.

Theorem 13.5.1. If F is a field, then $F[x]$ is a Euclidean domain.

Proof. By discussions in previous units, $F[x]$ is an integral domain. Define $d : F[x] \to \mathbb{Z}$ as $d(f(x)) = \deg f(x)$, for all $f(x) \in F[x]$. It is easy to check that d satisfies all the properties of definition 13.2.1 and hence is an ED (verify!). □

Theorem 13.5.2. If F is a field, then $F[x]$ is a principal ideal domain.

Proof. We know that $F[x]$ is an integral domain. Let I be an ideal in $F[x]$. If $I \neq \{0\}$, then among all the elements of I, let $g(x)$ be one of minimum degree. We will show that $I = \langle g(x)\rangle$. Since $g(x) \in I$, we have $\langle g(x)\rangle \subseteq I$. Now, let $f(x) \in I$. Then by division algorithm, we may write,

$$f(x) = g(x)q(x) + r(x)$$

and either $r(x) = 0$ or $\deg r(x) < \deg g(x)$. Since $r(x0 = f(x) - g(x)q(x) \in I$, the minimality of $\deg g(x)$ implies that the latter condition cannot hold. So, $r(x) = 0$ and hence $f(x) \in \langle g(x)\rangle$. This shows that $I \subseteq \langle g(x)\rangle$. □

The converse of the above theorem is not true in general. For example, $\mathbb{Z}[x]$ is a PID (how?) but $\mathbb{Z}$ is not a field.

Theorem 13.5.3. Let F be a field, I a non-zero ideal in $F[x]$, and $g(x)$ an element of $F[x]$. Then $I = \langle g(x)\rangle$ if and only if $g(x)$ is a nonzero polynomial of minimum degree in I.

Recollect all that we have learnt in the preceding units concerning polynomial rings. We will be defining an irreducible element in $R[x]$ for an integral domain R.

Definition 13.5.4. Let R be an integral domain. A polynomial $f(x) \in R[x]$ which is neither the zero polynomial nor a unit in $R[x]$ is said to be irreducible over R if, whenever $f(x)$ is expressed as $f(x) = g(x)h(x)$, where $h(x), g(x) \in R[x]$, then $g(x)$ and $h(x)$ are units in $R[x]$. A nonzero, nonunit element in $R[x]$ that is not irreducible over R is said to be reducible over R.

When R is a field, then it can be more conveniently said that a polynomial $f(x)$ is irreducible if it can't be expressed as the product of two polynomials of lower degree.

Example 13.5.5. The polynomial $f(x) = 2x^2 + 4$ is irreducible over $\mathbb{Q}$ but reducible over $\mathbb{Z}$, since $2x^2 + 4 = 2(x^2 + 2)$ and neither 2 nor $x^2 + 2$ is a unit in $\mathbb{Z}$.

In fact, we will soon see that any polynomial reducible over $\mathbb{Q}$ is reducible over $\mathbb{Z}$ but the converse is not true as we have seen in the above example.

Example 13.5.6. The polynomial $f(x) = 2x^2 + 4$ is irreducible over $\mathbb{R}$but reducible over $\mathbb{C}$.

Example 13.5.7. The polynomial $f(x) = x^2 - 2$ is irreducible over $\mathbb{Q}$ but reducible over $\mathbb{R}$ since $x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2})$.

In general it is not a very easy job to decide whether a given polynomial is irreducible over some given integral domain. So, we have certain theorems to help us decide that. Our first such theorem is

Theorem 13.5.8. Let F be a field. If $f(x) \in F[x]$ and deg$f(x) = 2$, or 3, then $f(x)$ is reducible over F if and only if $f(x)$ has a zero in F.

Proof. Let $f(x) = g(x)h(x)$, where $h(x), g(x) \in F[x]$ and have degrees less than that of $f(x)$. Since deg$f(x)$ =deg$g(x)$+deg$h(x)$, and deg$f(x)$ is 2 or 3, at least one of $g(x)$ and $h(x)$ has degree 1. Say $g(x) = ax + b$. Then of course $-a^{-1}b$ is a root of $g(x)$ and hence a zero of $f(x)$ too.

Conversely, suppose that $f(a) = 0$, where $a \in F$. Then, by the Factor Theorem, we know that $x - a$ is a factor of $f(x)$ and therefore $f(x)$ is reducible over F. □

One may think whether the theorem is true for polynomials over degree 3. For this, let us consider the following example:

Example 13.5.9. Consider $f(x) = x^4 + 2x^2 + 1$ over $\mathbb{Q}[x]$. Then clearly, $x^4 + 2x^2 + 1 = (x^2 + 1)(x^2 + 1)$. Hence $f(x)$ is reducible without having zeros in $\mathbb{Q}$.

Of course if $f(x)$ has a root in $R[x]$ then it is reducible. But the converse is true only if the conditions in the above theorem hold.

We have another powerful theorem fo checking irreducibility of polynomials in a UFD called the Eisenstein's Criterion. Lets check that out.

13.5.1 Eisenstein's criterion for irreducibility

Definition 13.5.10. The content of a nonzero polynomial $a_n x^n + \cdots + a_0$, where the a's are the integers, is the greatest common divisor of the integers $a_n, \cdots, a_0$.

A primitive polynomial is an element of $\mathbb{Z}[x]$ with content 1.

The product of two primitive polynomials is primitive.

Theorem 13.5.11. Let f(x)$\in \mathbb{Z}[x]$. If f(x) is reducible over $\mathbb{Q}$, then it is reducible over $\mathbb{Z}$

Proof. Let $f(x) = g(x)h(x)$, where $g(x)$ and $h(x) \in \mathbb{Q}[x]$. Clearly, we assume that $f(x)$ is a primitive because we can divide both $f(x)$ and $g(x)$ by the content of $f(x)$. Let a be the least common multiple of the denominators of the coefficients of $g(x)$, and b be the least common multiple of the denominators of the coefficients of $h(x)$. Then

$$abf(x) = ag(x).bh(x),$$

where $ag(x)$ and $bh(x)$ in $\mathbb{Z}[x]$. Let c_1 be the content of $ag(x)$ and c_2 be the content of $bh(x)$. Then

$$\begin{aligned} ag(x) &= c_1 g_1(x) \ and \\ bh(x) &= c_2 h_1(x) \end{aligned}$$

where both $g_1(x)$ and $h_1(x)$ are primitive, and

$$abf(x) = c_1 c_2 g_1(x) h_1(x)$$

Since, $f(x)$ is primitive, the content of $abf(x)$ is ab. Also, since the product of two primitives is primitive, it follows that the content of $c_1 c_2 g_1(x) h_1(x)$ is $c_1 c_2$. Thus, $ab - c_1 c_2$ and $f(x) = g_1(x) h_1(x)$, where $g_1(x)$ and $h_1(x) \in \mathbb{Z}[x]$ and $\deg g_1(x) = \deg g(x)$ and $\deg h_1(x) = \deg h(x)$. □

Theorem 13.5.12. Let

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x]$$

If there is a prime p such that p does not divide a_n, $p|a_{n-1}, \cdots, p|a_0$ and p^2 does not divide a_0, then f(x) is irreducible over $\mathbb{Q}$.

This is Eisenstein's Criterion. This theorem can also be stated for any arbitrary UFD.

Proof. If $f(x)$ is reducible over $\mathbb{Q}$ then there exists $g(x)$ and $h(x)$ in $\mathbb{Z}[x]$ such that

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

$1 \leq \deg g(x)$ and $1 \leq \deg h(x) < n$. Say

$$\begin{aligned} g(x) &= b_r x^r + \cdots + b_0 \ \ and \\ h(x) &= c_s x^s + \cdots + c_0. \end{aligned}$$

Then, since $p|a_0$, and p^2 does not divide a_0, and $a_0 = b_0 c_0$, it follows that p divides one of b_0 and c_0, but not the other. Let $p|b_0$ and not c_0. Also, since p does not divides $a_n = b_r c_s$, we know that p does not divide b_r. So, there is a least integer t such that p does not divide b_t. Now, consider

$$a_t = b_t c_0 + b_{t-1} c_1 + \cdots + b_0 c_t.$$

By assumption, p divides a_t and by choice of t, every summand on the right after the first one is divisible by p. Clearly, this forces p to divide $b_t c_0$ as well. This is impossible, however, since p is prime and p divides neither b_t nor c_0. □

Corollary 13.5.13. For any prime p, the p-th cyclotomic polynomial

$$\phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible over $\mathbb{Q}$.

Example 13.5.14. The polynomial $3x^5 + 15x^4 - 20x^3 + 10x + 20$ is irreducible over $\mathbb{Q}$ because 5 does not divide 3 and 25 does not divide 20 but 5 divides 15, -20, 10 and 20.

Theorem 13.5.15. Let F be a field and p(x)$\in F[x]$. Then $\langle p(x) \rangle$ is maximal ideal in $F[x]$ if and only if p(x) is irreducible over F.

Proof. Let $\langle p(x) \rangle$ be maximal ideal in $F[x]$. Clearly, $p(x)$ is neither zero nor unit in $F[x]$, because neither $\{0\}$ nor $F[x]$ is a maximal ideal in $F[x]$. If $p(x) = g(x)h(x)$ is a factorization of $p(x)$ over F, then $\langle p(x) \rangle \subseteq \langle g(x) \rangle \subseteq F[x]$. Thus, $\langle p(x) \rangle = \langle g(x) \rangle$ or $F[x] = \langle g(x) \rangle$. In the first case, we must have $\deg p(x) = \deg g(x)$. In the second case, it follows that $\deg g(x) = 0$ and consequently, $\deg h(x) = \deg p(x)$. Thus $p(x)$ cannot be written as product of two polynomials in $F[x]$ of lower degree.
Now, let $p(x)$ is irreducible over F. Let I be any ideal of $F[x]$ such that $\langle p(x) \rangle \subseteq I \subseteq F[x]$. Since $F[x]$ is a principal ideal domain, we know that $I = \langle g(x) \rangle$ for some $g(x) \in F[x]$. So, $p(x) \in \langle g(x) \rangle$ and thus $p(x) = g(x)h(x)$, where $h(x) \in F[x]$. Since, $p(x)$ is irreducible over F, it follows that either $g(x)$ is a constant or $h(x)$ is a constant. In the first case, we have $I = F[x]$; in the second case, we have $\langle p(x) \rangle = \langle g(x) \rangle = I$. So, $\langle p(x) \rangle$ is maximal in $F[x]$. □

Corollary 13.5.16. Let F be a field and $p(x)$ be an irreducible polynomial over F. Then, $F[x]/\langle p(x) \rangle$ is a field.

Proof. This easily follows. □

Corollary 13.5.17. Let F be a field and let $p(x), a(x), b(x) \in F[x]$. If p(x) is irreducible over F and $p(x)|a(x)b(x)$, then $p(x)|a(x)$ or $p(x)|b(x)$.

Proof. Since, $p(x)$ is irreducible, $F[x]/\langle p(x)\rangle$ is a field and, therefore, an integral domain. From one of the previous theorems, we know that $\langle p(x)\rangle$ is a prime ideal, and since $p(x)$ divides $a(x)bx)$, we have $a(x)b(x) \in \langle p(x)\rangle$. Thus, $a(x) \in \langle p(x)\rangle$ or $b(x) \in \langle p(x)\rangle$. This means that $p(x)|a(x)$ or $p(x)|b(x)$. □

Exercise 13.5.18. 1. Show that every polynomial with real coefficients of odd degree is reducible.

2. Find all roots of $f(x) = x^3 + x^2 + x + 1$ in $\mathbb{Z}_5$.
3. Let $f(x) = x^2 + 8x - 2$. Show that it is irreducible over $\mathbb{Q}$. Is is irreducible over $\mathbb{R}$?
4. Test for irreducibility of the following polynomial over $\mathbb{Q}$:
   a) $8x^3 + 6x^2 - 9x + 24$
   b) $x^4 + 9x + 3$
   c) $x^5 + 9x^4 + 12x^2 + 6$
5. Show that $x^2 + x + 4$ is irreducible over $\mathbb{Z}_{11}$.

Sample Questions

1. Show that every ED is a PID. Is the converse true? Justify your answer.
2. Show that every PID is a UFD.
3. Show that any two elements in an ED have a gcd.
4. If F is a field, show that $F[x]$ is a PID.
5. Let $f(x) \in F[x]$, where $\deg f(x) = 2$ or 3 and F is a field. Show that then $f(x)$ is reducible over F if and only if $f(x)$ has a zero in F.
6. State the Eisenstein's criteion. Hence check for the irreducibility of the $x^7 + 48x - 24$ over $\mathbb{Q}$.

Unit 14

Course Structure

- Extension of fields: Simple extension, Algebraic and transcendental extensions
- Splitting fields, normal extensions
- Separable extensions.

14.1 Introduction

Consider the expression $x^2 + 1$. It is a polynomial with real coefficients. Hence, if we name it as $f(x)$, then $f(x) \in \mathbb{R}[x]$. But, this polynomial clearly has no root in the underlying field $\mathbb{R}$. However, if we adjoin the number $i = \sqrt{-1}$ to the field $\mathbb{R}$ and consider a new field containing $\mathbb{R}$ and i, then that new field can be identified with the complex field $\mathbb{C}$. $\mathbb{C}$ can be thought of as some sort of "extension" to the field of real numbers which contains both the roots of $f(x)$, that is $\pm i$. Thus, in order to find the roots of a general polynomial over a certain field F, we may be required to go to some higher field, say K containing F. This K gives a clearer picture of the roots of the polynomial in question. The nature of the roots, such as whether they are simple or multiple; the nature of the polynomial, for example $f(x)$ as defined above, is irreducible in $\mathbb{R}$, whereas, it is reducible in $\mathbb{C}$; etc. can be studied having a systematic introduction to the idea of field extensions.

Objectives

After reading this unit, you will be able to

- learn the basic idea of extension of fields and related terminologies
- get an idea of the roots of a given polynomial by the application of appropriate extension to a given field
- find a certain extension that contains all the roots of a given polynomial

14.2 Field extensions

Definition 14.2.1. Let F be a field and K be another field containing F. Then K is called an extension of F and is denoted by the symbol K/F.

Example 14.2.2.

1. Let $F = \mathbb{Q}$, $K = \mathbb{R}$ and $L = \mathbb{C}$, the fields of rational, real, and complex numbers. Then, K and L are extensions of F.
2. Let K be a field of Char $p \geq 0$. If $p = 0$, K contains a field F isomorphic to $\mathbb{Q}$, that is, K is an extension of F. If $p > 0$, K contains a field F isomorphic to $\mathbb{Z}_p$, that is, K is an extension of F.

Definition 14.2.3. Let K/F be an extension, $a_1, a_2, \cdots, a_n \in K$. The smallest subfield of K containing F and $a_1, a_2, \cdots, a_n$ is called the field generated by $a_1, a_2, \cdots, a_n$ over F and is denoted by $F(a_1, a_2, \cdots, a_n)$.

Theorem 14.2.4.

$$F(a_1, a_2, \cdots, a_n) = \left\{ \frac{f(a_1, a_2, \cdots, a_n)}{g(a_1, a_2, \cdots, a_n)} : f, g \in F[x_1, x_2, \cdots, x_n], g(a_1, a_2, \cdots, a_n) \neq 0 \right\}$$

Proof. Let

$$L = \left\{ \frac{f(a_1, a_2, \cdots, a_n)}{g(a_1, a_2, \cdots, a_n)} : f, g \in F[x_1, x_2, \cdots, x_n], g(a_1, a_2, \cdots, a_n) \neq 0 \right\}$$

Then L is evidently a field for addition and multiplication induced from K. Also, $L \supset F$ as $a \in F$ can be represented as $a = a/1 \in L$. Let K be some other field containing F and $a_1, a_2, \cdots, a_n$, then K contains $f(a_1, a_2, \cdots, a_n)$ and also $\dfrac{f(a_1, a_2, \cdots, a_n)}{g(a_1, a_2, \cdots, a_n)}$, if $g(a_1, a_2, \cdots, a_n) \neq 0$. Hence, $L \subset K$. Hence the result. □

Definition 14.2.5. K/F is called a simple extension if $K = F(a)$.

Example 14.2.6. Let $F = Q$ and

$$K = \left\{ \frac{a + b\sqrt{2}}{c + d\sqrt{2}} \mid a, b, c, d \in Q,\ c \text{ or } d \neq 0 \right\}.$$

Then K is a field and $K = Q(\sqrt{2})$ is a simple extension of Q.

Theorem 14.2.7. Let K/F be a simple extension with $K = F(a)$. The, either

1. there does not exist any non-zero polynomial $g(x) \in F[x]$ with $g(a) = 0$, or
2. there exists a unique monic polynomial $f(x)$ of least degree with $f(a) = 0$.

Proof. Let $\phi : F[x] \to F[a]$ be defined by $\phi(f(x)) = f(a)$, $f(x) \in F[x]$. Then ϕ is a ring homomorphism which is onto. Then we have two cases:

1. If $Ker\ \phi = \{0\}$, there does not exist any non-zero polynomial $g(x) \in F[x]$ such that $g(a) = 0$.
2. If $Ker\ \phi \neq \{0\}$, then since $F[x]$ is a PID, so there exists a polynomial $h(x) \in F[x]$ such that $Ker\ \phi = \langle h(x) \rangle$, $h(x)$ is unique if chosen to be monic. Now, since $h(x) \in Ker\ \phi$, so $h(a) = 0$. Let $f(x) \in F[x]$ with $f(a) = 0$. Then $f(x) \in Ker\ \phi$ and $f(x)$ is a multiple of $h(x)$, that is, degree of $f \geq \deg h$. Hence the second part is proved.

□

We further see that $Ker\ \phi$ is a non-zero prime ideal, since $F[a]$ is a PID, hence $Ker\ \phi$ is a maximal ideal, that is, $F[x]/(Ker\ \phi)$ is isomorphic to $F[a]$, a field. This shows that $F[a] = F(a)$.

Definition 14.2.8. Let K/F be an extension of $a \in K$. Then a is called *transcendental* over F if there does not exist any non-zero $f(x) \in F[x]$ with $f(a) = 0$. Otherwise, a is called *algebraic* over F and the unique monic polynomial $f(x) \in F[x]$ of least degree with $f(a) = 0$ is called the *minimum polynomial* of a. And, if every element of K is algebraic, then K/F is called an algebraic extension.

Example 14.2.9.
1. Let $F = \mathbb{R}$. Then $a = i = \sqrt{-1}$ is algebraic over F with minimal polynomial $x^2 + 1$.
2. $F = \mathbb{Q}$. Then $a = e$ or π. Then a is transcendental over F.
3. $F = \mathbb{Q}$ and $a = \sqrt{3}$. Then a is algebraic over F since the minimum polynomial is $x^2 - 3 \in \mathbb{Q}[x]$.

Theorem 14.2.10. Let $a \in K$ be algebraic over F and $f(x)$ be the minimum polynomial of a with degree n. Then $F(a)$ forms a vector space over F with dimension n.

Proof. We have,

$$F(a) = \frac{F[x]}{\langle f(x) \rangle} = F[a]$$

where

$$f(x) = x^n + c_1 x^{n-1} + \cdots + c_n,$$

where, $c_i \in F$. We claim that $1, a, a^2, \ldots, a^{n-1}$ is a basis of $F(a)$. They are linearly independent since otherwise a will satisfy a polynomial of degree less than n over F contradicting that $f(x)$ is the minimum polynomial of a. Now,

$$a^n = -(c_1 a^n + c_2 a^{n-1} + \cdots + c_n a).$$

This is an F-linear combination of $1, a, \cdots, a^{n-1}$ by substituting for a^n. Similarly, any power of a can be expressed as a linear combination of $1, a, a^2, \ldots, a^{n-1}$. Thus, $1, a, a^2, \ldots, a^{n-1}$ generates $F[a] = F(a)$ over F. Hence, the result. □

Definition 14.2.11. Let K/F be an extension. It is called a *finite* extension if dimension of K over F, denoted as $\dim_F K < \infty$, or $[K : F] < \infty$.

Example 14.2.12. If a is algebraic over F with minimum polynomial of degree n, then $[F(a) : F] = n$.

Theorem 14.2.13. Let K/F be a finite extension. Then K/F is an algebraic extension.

Proof. Let $[K : F] = n$. Then, by the previous theorem, $1, a, a^2, \ldots, a^n$ are linearly dependent over F. In particular, there exists $c_0, c_1, \ldots, c_n$, not all zero, such that,

$$c_0 + c_1 a + c_2 a^2 + \cdots + c_n a^n = 0$$

Hence, if we define $f(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n \not\equiv 0$, then $f(a) = 0$. Hence, a is algebraic over F. Since every $a \in K$ is algebraic over F, K/F is an algebraic extension. □

Theorem 14.2.14. Let $F \subset K \subset L$ be extensions such that K/F and L/K are finite. Then L/F is finite and

$$[L : F] = [L : K][K : F].$$

Proof. Let $\{e_1, e_2, \cdots, e_n\}$ be a basis of K/F and $\{f_1, f_2, \cdots, f_m\}$ be a basis of L/K. We claim that $\{e_i f_j\}$, $1 \leq i \leq n$ and $1 \leq j \leq m$ is a basis of L/F. Let $a \in L$. Then,

$$a = \sum_{j=1}^{m} b_j f_j$$

where, $b_j \in K$. Now, for $b_j \in K$, we have

$$b_j = \sum_{i=1}^{n} c_{ij} e_i$$

where, $c_{ij} \in F$. Then,

$$a = \sum_{j=1}^{m} \sum_{i=1}^{n} c_{ij} e_i f_j$$

Thus showing that $\{e_i f_j\}$ generates L over F. In order to prove the linear independence, let

$$\sum_{i,j} \mu_{ij} e_i f_j = 0,$$

where, $\mu_{ij} \in F$. Then,

$$\sum_{j} \left(\sum_{i} \mu_{ij} e_i \right) f_j = 0.$$

Owing to the linear independence of $\{f_j\}$ over K,

$$\sum_{i} \mu_{ij} e_i = 0$$

for all j. Since $\{e_i\}$ are linearly independent over F, so $\mu_{ij} = 0$ for all i, j. Hence, $\{e_i f_j\}$ is a basis of L over F. Hence the result. □

Corollary 14.2.15. Let K/F be a finite extension and $a \in K$ with minimum polynomial of degree n. Then n divides $[K : F]$.

Proof. Follows directly from the above theorem. □

Corollary 14.2.16. Let K/F be an extension, $a_1, a_2, \ldots, a_n \in K$ are algebraic over F. Then $F(a_1, a_2, \ldots, a_n)/F$ is a finite extension.

Proof. The proof is by induction on n. For $n = 1$, the result follows from theorem 14.2.10. By induction hypothesis, $F' = F(a_1, a_2, \ldots, a_{n-1})/F$ is a finite extension. Since a_n is algebraic over F, it is also so over F'. Hence, $F'(a_n)/F'$ is a finite extension. By the previous theorem, $F'(a_n)/F$ is a finite extension, that is, $F(a_1, a_2, \ldots, a_n)/F$ is a finite extension. □

Corollary 14.2.17. Let K/F be an extension and $a, b \in K$ are algebraic over F. Then $a \pm b$, ab, a/b, $b \neq 0$ are all algebraic over F.

Proof. Follows from the previous corollary. □

Theorem 14.2.18. Let K/F and L/K are algebraic extensions. Then L/F is an algebraic extension.

Proof. Since L/K is algebraic, every $a \in L$ satisfies the relation $a^n + c_1 a^{n-1} + \cdots + c_n = 0$, $a_i \in K$. Then a is also algebraic over $F' = F(c_1, c_2, \ldots, c_n)$ as the above relation is also a relation over the field F'. Since $c_i \in K$, they are all algebraic over F. Hence, F/F' is a finite extension . Now, $F'(a)/F'$ is a finite extension and hence $F'(a)/F$ is finite. In particular, a is algebraic over F. Hence, L/F is an algebraic extension. □

Definition 14.2.19. A field F is called *algebraically closed* if it has no proper algebraic extension, that is, if K/F is an algebraic extension of F, then $K = F$.

The complex field $\mathbb{C}$ is algebraically closed.

Exercise 14.2.20. 1. Find the degree of the following field extensions

i) $\mathbb{Q}(\sqrt[3]{2}, \sqrt{3})/\mathbb{Q}$

ii) $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})/\mathbb{Q}$

2. Show that a finite extension of prime degree is a simple extension.

3. Let K/F be an extension and $a, b \in K$ are algebraic over F with degrees m and n respectively. Show that if $\gcd(m, n) = 1$, then $[F(a, b) : F] = mn$.

14.3 Normal Extensions

We have seen that a polynomial $f(x) \in F[x]$ may not always have a root in F, but we can obtain a root of $f(x)$ in an extension field K/F. We see the following theorem:

Theorem 14.3.1. Let $f(x) \in F[x]$ be an irreducible polynomial. Then there exists an extension K/F which contains a root of $f(x)$.

Proof. Since $f(x)$ is irreducible, the ideal $I = \langle f(x) \rangle$ in $F[x]$ is a maximal ideal and hence $K = F[x]/I$ is a field. The map $F \to K$ given by $a \mapsto a + I$ is an isomorphism of F onto its image $F' \subset K$. Identifying F with $F' \subset K$, we can consider K as an extension of F. Also, $\alpha = x + I \in K$ is a root of $f(x)$ since $f(\alpha) = f(x + I) = f(x) + I = \overline{0}$ as $f(x) \in I$. Hence the result. □

Corollary 14.3.2. Let $g(x) \in F[x]$ be any non-constant polynomial. Then there exists an extension K/F which contains a root of $g(x)$.

Proof. Consider an irreducible factor $f(x)$ of $g(x)$ and take the extension K/F which contains a root α of $f(x)$. Then α is a root of $g(x)$. □

Now, if $g(x) \in F[x]$ is of degree n, then we know that $g(x)$ can have at most n roots in any extension K/F, counting a root of multiplicity k, as k roots. We now show that there exists a field K/F which contains all the n roots of $f(x)$ and in fact we can obtain a smallest extension which contains all the n roots.

Definition 14.3.3. Let $f(x) \in F[x]$ be of degree n. An extension K/F is called a *splitting field* of $f(x)$ if

1. K contains all the n roots $a_1, a_2, \ldots, a_n$ of $f(x)$, and

2. $K = F(a_1, a_2, \ldots, a_n)$.

Theorem 14.3.4. Let $f(x) \in F[x]$ be of degree n. Then $f(x)$ has a splitting field.

Proof. The proof is by induction on n. If $n = 1$, $f(x) = cx + d$, where $c \neq 0$. Then $a_1 = -d/c \in F$ is clearly a root of $f(x)$ and $K = F$. Let $n > 1$. Let $f_1(x)$ be an irreducible factor of $f(x)$ and let $L = F(a_1)$ be an extension such that $f_1(a_1) = 0$. Then $f(a_1) = 0$ and hence $f(x) = (x - a_1)g(x)$, $g(x) \in L[x]$. Since $\deg g(x) = n - 1$, by induction, there exists an extension K/L which contains $n - 1$ roots $a_2, a_3, \ldots, a_n$ of $g(x)$ such that $K = L(a_2, a_3, \ldots, a_n)$. Then $K = F(a_1, a_2, \ldots, a_n)$ is the required splitting field of $f(x)$. □

Let $f(x) \in F[x]$ and K/F be a splitting field of $f(x)$ with roots $a_1, a_2, \ldots, a_n$. Then $f(x) = (x-a_1)(x-a_2)\ldots(x-a_n)$ over $K[x]$. This means that a polynomial completely splits over its splitting field.

Example 14.3.5. 1. The splitting filed of $x^2 - 3$ is $\mathbb{Q}(\sqrt{3})$.

2. The splitting filed of $x^3 - 2$ is $\mathbb{Q}(\omega, \sqrt[3]{2})$, where ω is a cube root of unity.

Definition 14.3.6. Let K/F and K'/F be extension fields. An isomorphism $\sigma : K \to K'$ of K onto K' is called an F-isomorphism if $\sigma|_F = Id$, where, Id is the identity mapping.

Theorem 14.3.7. Let $\sigma : F \to F'$ be an isomorphism of F onto F', $f(x) \in F[x]$ an irreducible polynomial, $f(x) = \sum a_i x^i$, $\overline{f}(x) = \sum \overline{a}_i x^i$, where $\overline{a}_i = \sigma(a_i)$, the image polynomial over $F'[x]$. Let a, a' be respectively roots of $f(x)$ and $\overline{f}(x)$. Then σ can be extended to an isomorphism $\overline{\sigma} : F(a) \to F'(a')$ such that $\overline{\sigma}|_F = \sigma$.

Proof. Since σ is an isomorphism and $f(x) \in F[x]$ is irreducible, $\overline{f}(x) \in F'[x]$ is also irreducible, σ can be extended to an isomorphism $\sigma_1 : F[x] \to F'[x]$, where $\sigma_1\left(\sum b_i x^i\right) = \sum \overline{b}_i x^i$. Then $\sigma_1(f(x)) = \overline{f}(x)$ and this induces an isomorphism of the quotient rings

$$\overline{\sigma}_1 : \frac{F[x]}{\langle f(x)\rangle} \to \frac{F'[x]}{\langle \overline{f}(x)\rangle}$$

with $\overline{\sigma}_1(x + \langle f(x)\rangle) = x + \langle \overline{f'}(x)\rangle$. Since for any root α of $f(x)$, $F(\alpha) \simeq \dfrac{F[x]}{\langle f(x)\rangle}$, we have an isomorphism $\overline{\sigma} : F(\alpha) \to F'(\alpha')$. Clearly $\overline{\sigma}(\alpha) = \alpha'$ and $\overline{\sigma}|_F = \sigma$. □

Corollary 14.3.8. Let $f(x) \in F[x]$ be an irreducible polynomial and α, α' be roots of $f(x)$ in some extension. Then there exists an F-isomorphism $\overline{\sigma} : F(\alpha) \to F(\alpha')$ such that $\overline{\sigma}(\alpha) = \alpha'$.

Proof. Take $F = F'$ and $\sigma = Id$. □

Theorem 14.3.9. Let $\sigma : F \to F'$ be an isomorphism from F onto F', $f(x) \in F[x]$, $f(x) = \sum a_i x^i$, $\overline{f}(x) = \sigma(f(x)) = \sum \overline{a}_i x^i$, where $\overline{a}_i = \sigma(a_i)$.

Proof. The proof is by induction on $n = \deg f$. If $n = 0$, there is nothing to prove. Assume $n \geq 1$. Let $f_1(x)$ be an irreducible factor of $f(x)$, α a root of $f_1(x)$ and α' a root of $\sigma(f_1)$. By the previous theorem, σ can be extended to an isomorphism $\sigma_1 : F(\alpha) \to F'(\alpha')$ with $\sigma_1|F = \sigma$.

Let $f(x) = (x - \alpha)g(x)$, $g(x) \in F(\alpha)[x]$, so that $\sigma(f(x)) = (x - \alpha')\sigma_1(g)$. Then K(respectively K') is a splitting field of $g(x)$ over $F(x)$(respectively $\sigma_1(g)$ over $F'(\alpha')$). By induction, σ_1 can be extended to an isomorphism $\overline{\sigma} : K \to K'$ of K onto K'. Clearly, $\overline{\sigma}|F = \sigma_1|F = \sigma$. □

Corollary 14.3.10. Any two splitting fields of $f(x) \in F[x]$ are isomorphic.

Proof. Take $F = F'$ and $\sigma = Id$. □

We have seen earlier that an extension K/F of F may contain some roots of $f(x)$ but not all. We will now deal with those extensions which have that property.

Definition 14.3.11. An extension K/F is called a normal extension if it is algebraic and for any irreducible polynomial $f(x) \in F[x]$ which has a root in K, $f(x)$ has all its roots in K.

Example 14.3.12. Let K/F be an extension of degree 2. Then it is normal. Let $\alpha \in K$ with minimum polynomial $f(x)$. Then deg$f(x) \leq$ 2and if deg$f(x) = 1$, $\alpha \in F$. If deg$f(x) = 2$, then $f(x) = x^2 + ax + b,\ a, b \in F$. Then the other root is $\alpha - a$ as the sum of the roots is $-a$. Hence $\alpha - a \in K$, thus K/F is normal.

Definition 14.3.13. Two algebraic elements α and α' are said to be conjugate over F if there exists an F-isomorphism $\sigma : F(\alpha) \to F(\alpha')$ with $\sigma(\alpha) = \alpha'$.

Theorem 14.3.14. Two elements α and α' are conjugate over F iff they have the same minimum polynomial over F.

Proof. Let α and α' are conjugate over F and let $\sigma : F(\alpha) \to F(\alpha')$ with $\sigma(\alpha) = \alpha'$. If $f(x)$ is the minimum polynomial of α, then $\sigma(f(x)) = f(x)$ is the minimum polynomial of $\sigma(\alpha) = \alpha'$.

Conversely, assume that α and α' have the same minimum polynomial. Then using a previous theorem, taking $F = F'$ and $\sigma = Id$, we have an F-isomorphism $\overline{\sigma} : F(\alpha) \to F(\alpha')$ with $\overline{\sigma}(\alpha) = \alpha'$. Hence α and α'. □

Exercise 14.3.15.

1. Show that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$.
2. Find the degree of the splitting field of $x^4 + 1 \in \mathbb{Q}[x]$ over $\mathbb{Q}$.
3. Fine the splitting field of $x^4 - x^2 - 2$ over $\mathbb{Q}$.
4. Find all the conjugates of the following elements over $\mathbb{Q}$
   i) $\sqrt{1 + \sqrt{2}}$
   ii) $\sqrt{2} + i$
5. Show that if n is not an integer which is not a perfect square and $\alpha = a + b\sqrt{n}, a\ b \in \mathbb{Q}$ is a root of a polynomial $f(x) \in \mathbb{Q}[x]$, then $a + b\sqrt{n}$ is also a root of $f(x)$.

14.4 Separable Extensions

Definition 14.4.1. Let $f(x) \in F[x]$ be an irreducible polynomial. $f(x)$ is called a separable polynomial if all its roots in its splitting field are simple. A polynomial $g(x) \in F[x]$ is called separable if all its irreducible factors are separable. A polynomial which is not separable is called an inseparable polynomial.

Example 14.4.2. $x^3 - 2 \in \mathbb{Q}[x]$ is a separable polynomial as its roots $\sqrt[3]{2}, \omega\sqrt[3]{2}, \omega^2\sqrt[3]{2}$ are distinct, $\omega = e^{2\pi i/3}$.

Theorem 14.4.3. Let $f(x) \in F[x]$ be a non-constant polynomial. A root a of $f(x)$ in some extension field is a multiple root if and only if $f'(a) = 0$.

Proof. If $f(x) = \sum_i a_i x^i,\ f'(x) = \sum_i i a_i x^{i-1}$. Since $a \in K/F$ is a root of $f(x)$, $f(x) = (x - a)g(x)$, $g(x) \in K[x]$. a is a multiple root if and only if it is a root of $g(x)$. Now, $f'(x) = g(x) + (x - a)g'(x)$. Thus, a is a root of $g(x)$ if and only if it is a root of $f'(x)$, that is, $f'(a) = 0$. Hence the result. □

Corollary 14.4.4. Let $f(x) \in F[x]$ be a monic irreducible polynomial. Then f has a multiple root if and only if $f' \equiv 0$.

Proof. If a is a multiple root of $f(x)$ in some extension, then $f'(a) = 0$. Since $f(x)$ is the minimum polynomial of a, $f(x)$ divides $f'(x)$. But $\deg f'(x) < \deg f(x)$ so that $f'(x) \equiv 0$. □

Corollary 14.4.5. Let Char$F = 0$ and $f(x) \in F[x]$ be a monic irreducible polynomial. Then $f(x)$ is separable.

Proof. Let $f(x) = \sum_i a_i x^i$ be inseparable. Then $f'(x) = \sum_i i a_i x^{i-1} \equiv 0$, by the previous corollary. In particular, $ia_i = 0$ for all $i \geq 1$, that is, $f(x) = a_0$, a constant polynomial, which is a contradiction since $f(x)$ is irreducible. Hence, $f(x)$ is separable. □

Corollary 14.4.6. Let Char$F = p > 0$ and $f(x) \in F[x]$ be a monic irreducible polynomial. Then $f(x)$ is inseparable if and only if $f(x)$ is a polynomial in x^p.

Proof. Let $f(x) = \sum_i a_i x^i$. Then $f(x)$ is inseparable if and only if $f'(x) \equiv 0$, that is, $ia_i = 0,\ i \geq 1$. Thus implies $a_i = 0$ if $\gcd(i, p) = 1$, that is, $f(x)$ is a polynomial in x^p and conversely. □

Definition 14.4.7. Let K/F be an extension. $a \in K$ is called separable over F if it is algebraic and its minimum polynomial over F is a separable polynomial. Otherwise, a is called inseparable.

Definition 14.4.8. An extension K/F is called separable if it is algebraic and every $a \in K$ is separable over F.

Example 14.4.9.

1. If Char$F = 0$, any algebraic extension is separable.
2. Let Char$F = p > 0$ and $f(x) = x^p - a \in F[x]$ with no root in F. Then $f(x)$ is an inseparable polynomial for let a_1, a_2 be two roots of $f(x)$ in some extension field. Then $a_1^p = a = a_2^p$. Now, $(a_1 - a_2)^p = a_1^p - a_2^p$, as all the other terms in the binomial expansion are zero, the coefficients being divisible by p. Hence, $(a_1 - a_2)^p = a_1^p - a_2^p = a - a = 0$, that is, $a_1 = a_2$. Thus, all the roots of $f(x)$ are the same and there is only one root of f, say α of multiplicity p. We claim that $f(x)$ is irreducible over F. Let $g(x)$ be an irreducible factor of $f(s)$ so that $g(\alpha) = 0$, that is, $g(x)$ is the minimum polynomial of α. Thus, the only irreducible factor of $f(x)$ is the minimum polynomial $g(x)$ of α, that is, $f = g^m$. In particular, $\deg f = p = m \deg g$, that is, $\deg g$ divides p. But $\deg g > 1$, as $f(x)$ has no root in F. Hence $m = 1$, that is, $f = g$ is irreducible.

Certain types of fields do not admit any inseparable extensions. These are called perfect fields.

Definition 14.4.10. F is called a perfect field if every algebraic extension K/F is separable.

By the previous example, every field F having characteristic 0 is perfect.

Exercise 14.4.11.

1. Examine whether $x^4 + x + 1 \in \mathbb{Q}[x]$ is a separable polynomial.
2. Let Char$F = p > 0$ and K/F be a finite extension. If it is inseparable, show that p divides $[K : F]$.

Sample Questions

1. Define algebraic extension. If $a \in K$ is algebraic over F with minimum polynomial of degree n, show that $F(a)$ forms a vector space over F with dimension n.
2. Show that every finite extension is algebraic.
3. If $F \subset K \subset L$ be fields such that K/F and L/K are finite. Show that $[K : F]$ divides $[L : F]$.
4. For every irreducible polynomial $f(x) \in F[x]$, show that there exists an extension which contains a root of $f(x)$.
5. Define splitting field. Every polynomial has a splitting field. Comment with justifications.
6. For every non-constant polynomial $f(x)$ show that a is a multiple root if and only if $f'(a) = 0$.
7. If F has characteristic 0, show that any algebraic extension of F is separable.

Unit 15

Course Structure

- Sensitivity Analysis: Changes in price vector of objective function, changes in resource requirement vector, addition of decision variable, addition of a constraint.

15.1 Introduction

Once the optimal solution to a linear programming problem has been attained, it may be desirable to study how the current solution changes when the parameters of the problem get changed. The study of the effect of *discrete* changes in the values of the parameters on the optimal solution is called *sensitivity analysis* or *post-optimality analysis*. The objective is to determine how *sensitive* is the optimal solution is to the changes in the values of these parameters.

In general, once the optimal solution to a linear programming problem has been attained, two situations may arise which require additional computations :

1. During the formulation it is assumed that the parameters such as market demand, equipment capacity, resource consumption, resource availability, the relevant costs or profits are all known with certainty and do not change over time. In actual practice the markets fluctuate, material and labour costs go up or down, production times change and equipment availability varies from time to time. It is, therefore, desirable to study how the *current* optimal solution changes when the parameters of the problem get changed. In these problems this information may be more important than the single result provided by the optimal solution. Such an analysis converts the static linear programming solution into a dynamic tool to study the effect of changing conditions such as in business and industry.

2. The second situation is rather unpleasant, yet one may be encountered with it quite often. After attaining the optimal solution, one may discover that a wrong value of a cost coefficient was used or a particular variable or constraint was omitted or one or more of right-hand constants used were wrong.

The changes in parameters of the problem may be discrete or continuous. The study of the effect of discrete changes in parameters on the optimal solution is called the *sensitivity analysis* or the *post optimality analysis*, while that of continuous changes in parameters is called *parametric programming*. One way to determine the effects of parameter changes is to solve the problem anew, which may be computationally inefficient. Alternatively, the *current* optimal solution may be investigated, making use of the properties of the simplex criterion.

The second method reduces additional computations considerably and hence forms the subject of the present discussion.

The changes in the parameters of a linear programming problem include:

1. Changes in the cost/profit coefficients or cost/profit contribution per unit of decision variables (c_j).
2. Changes in the right-hand side of the constraints or availability of resources (b_i).
3. Addition of new variables.
4. Changes in the coefficients of constraints or consumption of resources per unit of decision variables (a_{ij}).
5. Addition of new constraints.

Generally, these parameter changes result in one of the following three cases :

1. The optimal solution remains unchanged i.e., the basic variables and their values remain unchanged.
2. The basic variables remain unchanged but their values change.
3. The basic variables as well as their values are changed.

While dealing with these changes, one important objective is to find the maximum extent to which a parameter or a set of parameters can be changed so that the current optimal solution remains optimal. In other words, the objective is to determine how *sensitive* is the optimal solution to the changes in those parameters. Such an analysis is called *sensitivity analysis*.

In this topic we shall see how to minimize the additional computations necessary to study the changes in various parameters. In many cases it may not be necessary to solve the problem all over again. A small amount of computational work applied to the optimal solution will suffice. However, when large modifications in parameters are made, the post-optimal computations may become so tedious that there is no alternative but to go back to the beginning and resolve the problem.

15.2 Changes in the Cost/Profit Coefficient c_j

Changes in the coefficients of the objective function may take place due to a change in cost or profit of either basic variables or non-basic variables. Each of these two cases will first be considered separately. The discussion, will then, be followed by a combined case. All the three cases will be studied by considering a few examples.

Example 15.2.1. A company wants to produce three products A, B and C. The unit profits on these products are Rs. 4, Rs. 6 and Rs. 2 respectively. These products require two types of resources: man-power and material. The following L.P. model is formulated for determining the optimal product mix:

$$\begin{array}{lll} \text{maximize} & Z = & 4x_1 + 6x_2 + 2x_3, \\ \text{subject to} & & x_1 + x_2 + x_3 \leq 3, \text{ (manpower)} \\ & & x_1 + 4x_2 + 7x_3 \leq 9, \text{ (material)} \\ & & x_1, x_2, x_3 \geq 0, \end{array}$$

where x_1, x_2, x_3 are the number of products A, B and C produced.

(a) Find the optimal product mix and the corresponding profit to the company.

(b) (i) Find the range on the values of non-basic variable coefficient c_3 such that the current optimal product mix remains optimal.

(ii) What happens if c_3 is increased to Rs. 12? What is the new optimal product mix in this case?

(c) (i) Find the range on basic variable coefficient c_1 such that the current optimal product mix remains optimal.

(ii) Find the effect when $c_1 =$ Rs. 8 on the optimal product mix.

(d) Find the effect of changing the objective function to $Z = 2x_1 + 8x_2 + 4x_3$ on the current optimal product mix.

Solution. The standard form of the problem is

$$\begin{array}{lll} \text{maximize} & Z = & 4x_1 + 6x_2 + 2x_3 + 0x_4 + 0x_5, \\ \text{subject to} & & x_1 + x_2 + x_3 + x_4 = 3, \\ & & x_1 + 4x_2 + 7x_3 + x_5 = 9, \\ & & x_1, x_2, x_3, x_4, x_5 \geq 0, \end{array}$$

Putting $x_1 = x_2 = x_3 = 0$ in the constraint equations, we get $x_4 = 3$ and $x_5 = 9$ as the initial basic feasible solution which can be expressed in the form of a simple matrix or table as shown below. Performing iterations we get the remaining tables.

Table 1

	c_j	4	6	2	0	0			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ	
0	x_4	1	1	1	1	0	3	3	
0	x_5	1	(4)	7	0	1	9	9/4	←
$Z_j = \Sigma c_B a_{ij}$		0	0	0	0	0			
$\overline{c}_j = c_j - Z_j$		4	6	2	0	0			
			↑						*Initial feasible solution*

Table 2

	c_j	4	6	2	0	0			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ	
0	x_4	$(3/4)$	0	$-3/4$	1	$-1/4$	$3/4$	1	←
6	x_2	/	1	$7/4$	0	$1/4$	$9/4$	9	
$Z_j = \Sigma c_B a_{ij}$		$3/2$	6	$21/2$	0	$3/2$			
$\overline{c}_j = c_j - Z_j$		$5/2$	0	$-17/2$	0	$-3/2$			
		↑							*Second feasible solution*

Therefore, the optimal solution is $x_1 = 1,\ x_2 = 2,\ x_3 = 0$ and $Z_{\max} =$ Rs. $(4\times1+6\times2+2\times0) =$ Rs. 16.

Table 3

	c_j	4	6	2	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
4	x_1	1	0	–1	$4/3$	$-1/3$	1
6	x_2	0	1	2	$-1/3$	$1/3$	2
$Z_j = \Sigma c_B a_{ij}$		4	6	8	$10/3$	$2/3$	
$\bar{c}_j = c_j - Z_j$		0	0	–6	$-10/3$	$-2/3$	

Optimal feasible solution

Effect of changing the objective function coefficient of a non-basic variable

(b) (i) The coefficient c_3 corresponds to the non-basic variable x_3 for product C. In the optimal product mix shown in Table 3, product C is not produced because of the low associated profit of Rs. 2 per unit (c_3). Clearly, if c_3 further decreases, it will have no effect on the current optimal product mix. However, if c_3 is increased beyond a certain value, it may become profitable to produce the product C.

Table 4

	c_j	4	6	12	0	0			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ	
4	x_1	1	0	–1	$4/3$	$-1/3$	1	–1	
6	x_2	0	1	(2)	$-1/3$	$1/3$	2	1	←
$Z_j = \Sigma c_B a_{ij}$		4	6	8	$10/3$	$2/3$			
$\bar{c}_j = c_j - Z_j$		0	0	4	$-10/3$	$-2/3$			
				↑					

Replace x_2 by x_3.

Table 5

	c_j	4	6	12	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
4	x_1	1	$1/2$	0	$7/6$	$-1/6$	2
12	x_3	0	$1/2$	1	$-1/6$	$1/6$	1
$Z_j = \Sigma c_B a_{ij}$		4	8	12	$8/3$	$4/3$	
$\bar{c}_j = c_j - Z_j$		0	–2	0	$-8/3$	$-4/3$	

Optimal feasible solution

As a rule, the sensitivity of the current optimal solution is determined by studying how the current optimal solution given in Table 3 changes as a result of changes in the input data. When value of c_3 changes, the value

of net evaluation (relative profit coefficient) of the non-basic variable x_3 i.e., $\bar{c}_3$ in Table 3 also changes. The table will remain optimal, as long as $\bar{c}_3$ remains non-positive.

Therefore, for Table 3 to remain optimal,

$$\overline{c_3} \leq 0 \Rightarrow c_3 - (4,6)\begin{bmatrix} -1 \\ 2 \end{bmatrix} \leq 0 \Rightarrow c_3 - (-4+12) \leq 0 \Rightarrow c_3 \leq 8.$$

This means that as long as the unit profit of product C is less than Rs. 8, it is not profitable to produce it. The current optimal solution remains optimal.

(ii) If $c_3 = 12$, then

$$\bar{c}_3 = c_3 - (4,6)\begin{bmatrix} -1 \\ 2 \end{bmatrix} = 12 - (-4+12) = 12 - 8 = 4.$$

As $\bar{c}_3$ becomes positive, the current product mix given by Table 3 does not remain optimal. The optimal profit can be increased further by producing product C. Non-basic variable x_3 can enter the solution to increase Z. This is shown in Table 4 and Table 5.

Therefore, new optimal product mix is $x_1 = 2, x_2 = 0, x_3 = 1$ and $Z_{\max}$ = Rs. $(4 \times 2 + 6 \times 0 + 12 \times 1)$ = Rs. 20.

Effect of changing the objective function coefficient of a basic variable

(c) (i) Clearly, when c_1 decreases below a certain level, it may no longer remain profitable to produce product A. On the other hand, if c_1 increases beyond a certain value, it may becomes so profitable that it is most paying to produce only product A. In either case the optimal product mix will change and hence there is lower as well as upper limit on c_1 within which the optimal product mix will not be affected.

Referring again to Table 3, it can be seen that any variation in c_1 (and/or in c_2 also) will not change $\bar{c}_1$ and $\bar{c}_2$ (i.e., they remain zero), while $\bar{c}_3, \bar{c}_4, \bar{c}_5$ will change. However, as long as $\bar{c}_j (j = 3, 4, 5)$ remain non-positive, Table 3 will remain optimal. $\bar{c}_3, \bar{c}_4$ and $\bar{c}_5$ can be expressed as functions of c_1 as follows :

$$\bar{c}_3 = 2 - (c_1, 6)\begin{bmatrix} -1 \\ 2 \end{bmatrix} = 2 - (-c_1 + 12) = c_1 - 10$$

$$\bar{c}_4 = 0 - (c_1, 6)\begin{bmatrix} \frac{4}{3} \\ -\frac{1}{3} \end{bmatrix} = 0 - \left(\frac{4}{3}c_1 - 2\right) = -\frac{4}{3}c_1 + 2$$

$$\bar{c}_5 = 0 - (c_1, 6)\begin{bmatrix} -\frac{1}{3} \\ \frac{1}{3} \end{bmatrix} = 0 - \left(-\frac{1}{3}c_1 + 2\right) = \frac{1}{3}c_1 - 2$$

$$\begin{aligned} &\text{For } \bar{c}_3 \text{ to be } \leq 0, \ c_1 - 10 \leq 0 \Rightarrow c_1 \leq 10, \\ &\text{for } \bar{c}_4 \text{ to be } \leq 0, -\frac{4}{3}c_1 + 2 \leq 0 \ \Rightarrow c_1 \geq \frac{3}{2}, \\ &\text{for } \bar{c}_5 \text{ to be } \leq 0, \ \frac{1}{3}c_1 - 2 \leq 0 \Rightarrow c_1 \leq 6. \end{aligned}$$

Therefore, range on c_1 for the optimal product mix to remain optimal is $\frac{3}{2} \leq c_1 \leq 6$. Thus so long as c_1 lies within these limits, the optimal solution in Table 3 viz., $x_1 = 1, x_2 = 2, x_3 = 0$ remains optimal.

However, within this range, as the value of c_1 is changed, $Z_{\max}$ undergoes a change. For example, when $c_1 = 3, Z_{\max} = \text{Rs. } (3 \times 1 + 6 \times 2) = \text{Rs. } 15$.

(ii) When $c_1 = 8$,

$$\bar{c}_3 = c_1 - 10 = 8 - 10 = -2, \quad \bar{c}_4 = -\frac{4}{3}c_1 + 2 = -\frac{4}{3} \times 8 + 2 = -\frac{26}{3},$$
$$\bar{c}_5 = \frac{1}{3}c_1 - 2 = \frac{8}{3} - 2 = +\frac{2}{3}, \quad \bar{c}_1 = \bar{c}_2 = 0.$$

As $\bar{c}_5$ becomes positive, the solution given in Table 3 no longer remains optimal. Slack variable x_5 enters the solution. This shown in Table 6 and Table 7.

Table 6

	c_j	8	6	2	0	0		
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ
8	x_1	1	0	–1	$4/3$	$-1/3$	1	–3
6	x_2	0	1	2	$-1/3$	$(1/3)$	2	6 ←
$Z_j = \Sigma c_B a_{ij}$		8	6	4	$26/3$	$-2/3$		
$\bar{c}_j = c_j - Z_j$		0	0	–2	$-26/3$	$2/3$ ↑		

Replacer x_2 by x_5.

Table 7

	c_j	8	6	2	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
8	x_1	1	1	1	1	0	3
0	x_5	0	3	6	–1	1	6
$Z_j = \Sigma c_B a_{ij}$		8	8	8	8	0	
$\bar{c}_j = c_j - Z_j$		0	–2	–6	–8	0	

Optimal feasible solution

Thus the optimal product mix changes to $x_1 = 3,\ x_2 = 0$ and $x_3 = 0$ units with $Z_{\max} = \text{Rs. } 24$.

Effect of changing the objective function coefficients of basic as well as non-basic variables

(d) The effect on the optimal product mix can be determined by checking whether the $\bar{c}_j$ row in Table 3 remains non-positive.

$$\bar{c}_1 = 0, \quad \bar{c}_2 = 0,$$
$$\bar{c}_3 = 4 - (2, 8)\begin{bmatrix} -1 \\ 2 \end{bmatrix} = 4 - (-2 + 16) = -10 \leq 0,$$
$$\bar{c}_4 = 0 - (2, 8)\begin{bmatrix} 4/3 \\ -1/3 \end{bmatrix} = 0 - \left(\frac{8}{3} - \frac{8}{3}\right) = 0,$$
$$\bar{c}_5 = 0 - (2, 8)\begin{bmatrix} -1/3 \\ 1/3 \end{bmatrix} = 0 - (-2/3 + 8/3) = -2 < 0.$$

Hence the optimal solution does not change. The optimal product mix remains $x_1 = 1, x_2 = 2, x_3 = 0$ and $Z_{\max} = (1 \times 2 + 2 \times 8 + 0 \times 4) = 18$. There is indication of an alternate optimal solution since $\bar{c}_4 = 0$.

15.3 Changes in the Right-Hand Side of the Constraints b_i

Suppose that an optimal solution to a linear programming problem has already been found and it is desired to find the effect of increasing or decreasing some resource. Clearly, this will affect not only the objective function but also the solution. Large changes in the limiting resources may even change the variables in the solution since one or more current basic variables becomes negative. Dual simplex method is used to remove infeasibility and to get a feasible optimal solution.

Example 15.3.1. (a) Solve the problem

$$\begin{array}{lll} \text{maximize} & Z = & 5x_1 + 12x_2 + 4x_3, \\ \text{subject to} & & x_1 + 2x_2 + x_3 \leq 5, \\ & & 2x_1 - x_2 + 3x_3 = 2, \\ & & x_1, x_2, x_3 \geq 0, \end{array}$$

(b) Discuss the effect of changing the requirement vector from $\begin{bmatrix} 5 \\ 2 \end{bmatrix}$ to $\begin{bmatrix} 7 \\ 2 \end{bmatrix}$ on the optimum solution.

(c) Discuss the effect of changing the requirement vector from $\begin{bmatrix} 5 \\ 2 \end{bmatrix}$ to $\begin{bmatrix} 3 \\ 9 \end{bmatrix}$ on the optimum solution.

(d) Which resource should be increased and how much to achieve the best marginal increase in the value of the objective function?

Solution. (a) The standard form of this problem is

$$\begin{array}{lll} \text{maximize} & Z = & 5x_1 + 12x_2 + 4x_3 + 0x_4 - Mx_5, \\ \text{subject to} & & x_1 + 2x_2 + x_3 + x_4 = 5, \\ & & 2x_1 - x_2 + 3x_3 + x_5 = 2, \\ & & x_1, x_2, x_3, x_4, x_5 \geq 0, \end{array}$$

Putting $x_1 = x_2 = x_3 = 0$ in the constraint equations, we get $x_4 = 5$ and $x_5 = 2$ as the initial basic solution which can be expressed in the form of a simple matrix or table. Performing iterations yields the table given below.

Table 1

	c_j	5	12	4	0	–M			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ	
0	x_4	1	2	1	1	0	5	5	
–M	x_5	2	–1	(3)	0	1	2	2/3	← *key row*
$Z_j = \Sigma c_B\, a_{ij}$		–2M	M	–3M	0	–M			
$\bar{c}_j = c_j - Z_j$		5+2M	12–M	4+3M	0	0			
				↑K					*Initial feasible solution*

Table 2

	c_j	5	12	4	0	–M			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ	
0	x_4	$1/3$	$(7/3)$	0	1	$-1/3$	$13/3$	$13/7$	← Key row
4	x_3	$2/3$	$-1/3$	1	0	$1/3$	$2/3$	–2	
$Z_j = \Sigma c_B a_{ij}$		$8/3$	$-4/3$	4	0	$4/3$			
$\bar{c}_j = c_j - Z_j$		$7/3$	$40/3$	0	0	$-M-4/3$			
			↑K						*Second feasible solution*

Table 3

	c_j	5	12	4	0	– M			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	θ	
12	x_2	$1/7$	1	0	$3/7$	$-1/7$	$13/7$	13	
4	x_3	$(5/7)$	0	1	$1/7$	$2/7$	$9/7$	$9/5$	←*Key row*
$Z_j = \Sigma c_B\, a_{ij}$		$32/7$	12	4	$40/7$	$-4/7$			
$\bar{c}_j = c_j - Z_j$		$3/7$	0	0	$-40/7$	$-M+4/7$			
		↑K							*Third feasible solution*

Table 4

	c_j	5	12	4	0	– M	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
12	x_2	0	1	$-1/5$	$2/5$	$-1/5$	$8/5$
5	x_1	1	0	$7/5$	$1/5$	$2/5$	$9/5$
$Z_j = \Sigma c_B\, a_{ij}$		5	12	$23/5$	$29/5$	$-2/5$	
$\bar{c}_j = c_j - Z_j$		0	0	$-3/5$	$-29/5$	$-M+2/5$	
							Optimal feasible solution

Thus the optimal solution is $x_1 = 9/5,\ x_2 = 8/5, x_3 = 0$, and $Z_{\max} = 5 \times 9/5 + 12 \times 8/5 + 0 = 141/5$.

(b) New values of the current basic variables are given by

$$\begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = B^{-1}b = \begin{bmatrix} 2/5 & -1/5 \\ 1/5 & 2/5 \end{bmatrix} \begin{bmatrix} 7 \\ 2 \end{bmatrix} = \begin{bmatrix} 14/5 - 2/5 \\ 7/5 + 4/5 \end{bmatrix} = \begin{bmatrix} 12/5 \\ 11/5 \end{bmatrix}$$

Since both x_1 and x_2 are non-negative, the current basic solution consisting of x_1 and x_2 remains feasible and optimal at the new values $x_1 = 11/5,\ x_2 = 12/5$ and $x_3 = 0$. The new optimum value of Z is

$5 \times 11/5 + 12 \times 12/5 + 4 \times 0 = 199/5.$

(c) New values of the current basic variables are

$$\begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = B^{-1}b = \begin{bmatrix} \frac{2}{5} & -\frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} \end{bmatrix} \begin{bmatrix} 3 \\ 9 \end{bmatrix} = \begin{bmatrix} \frac{6}{5} - \frac{9}{5} \\ \frac{3}{5} + \frac{18}{5} \end{bmatrix} = \begin{bmatrix} -\frac{3}{5} \\ \frac{21}{5} \end{bmatrix}$$

Since x_2 become $-$ve, the *current optimal solution becomes infeasible.* Dual simplex method may be used to clear infeasibility of the problem. Table 4 is modified and written in Table 5.

Table 5

	c_j	5	12	4	0	–M		
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b	
12	x_2	0	1	$(-1/5)$	$2/5$	$-1/5$	$-3/5$	← *Key row*
5	x_1	1	0	$7/5$	$1/5$	$2/5$	$21/5$	
$Z_j = \Sigma c_B\, a_{ij}$		5	12	$23/5$	$29/5$	$-2/5$		
$\bar{c}_j = c_j - Z_j$		0	0	$-3/5$	$-29/5$	–M+$2/5$		
				↑K				

As $b_1 = -3/5$, the first row is the key row and x_2 is the outgoing variable. Find the ratios of non-basic elements of $\bar{c}_j$ row to the elements of key row. Neglect the ratios corresponding to positive or zero elements of key row and choose the lowest ratio. The desired ratio is $\dfrac{-3/5}{-1/5} = 3$. Hence 'x_3'-column is the key column, x_3 is the incoming variable and $-1/5$ is the key element. Replace x_2 by x_3. This is shown in Table 6.

Table 6

	c_j	5	12	4	0	–M	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
4	x_3	0	–5	1	–2	1	3
5	x_1	1	7	0	3	–1	0
$Z_j = \Sigma c_B a_{ij}$		5	15	4	7	–1	
$\bar{c}_j = c_j - Z_j$		0	–3	0	–7	–M+1	

As all elements in $\bar{c}_j$-row are negative or zero and all b_i are positive, the solution given by Table 6 is optimal. The optimal solution is

$$x_1 = 0, x_2 = 0, x_3 = 3,$$
$$Z_{\max} = 5(0) + 12(0) + 4 \times 3 = 12.$$

(d) In order to find the resource that should be increased (or decreased), we shall write the dual objective function, which is

$$G = 5y_1 + 2y_2,$$

where $y_1 = 29/5$ and $y_2 = 2/5$ are the optimal dual variables. Thus the first resource should be increased as each additional unit of the first resource increases the objective function by $29/5$. Next we are to find how

much the first resource should be increased so that each additional unit continues to increase the objective function by $29/5$. This requirement will be met so long as the primal problem remains feasible. If Δ be increase in the first resource, it can be determined from the condition

$$\begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = B^{-1}b = \begin{bmatrix} 2/5 & -1/5 \\ 1/5 & 2/5 \end{bmatrix} \begin{bmatrix} 5+\Delta \\ 2 \end{bmatrix} = \begin{bmatrix} 10/5 + 2\Delta/5 - 2/5 \\ 5/5 + \Delta/5 + 4/5 \end{bmatrix} = \begin{bmatrix} \frac{8+2\Delta}{5} \\ \frac{9+\Delta}{5} \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

As x_1 and x_2 remain feasible (≥ 0) for all values of $\Delta \geq 0$, the first resource can be increased indefinitely while maintaining the condition that each additional unit will increase the objective function by $29/5$.

The second resource should be decreased as each additional unit of the second resource decreases the objective function by $2/5$. Let Δ be the decrease in the second resource. To find its extent, we make use of the condition that the current solution remains feasible so long as

$$\begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = B^{-1}b = \begin{bmatrix} 2/5 & -1/5 \\ 1/5 & 2/5 \end{bmatrix} \begin{bmatrix} 5 \\ 2-\Delta \end{bmatrix} = \begin{bmatrix} 10/5 - 2/5 + \Delta/5 \\ 5/5 + 4/5 - 2\Delta/5 \end{bmatrix} = \begin{bmatrix} \frac{8+\Delta}{5} \\ \frac{9-2\Delta}{5} \end{bmatrix} \geq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Evidently x_1 remains positive only so long as $\dfrac{9-2\Delta}{5} \geq 0$ or $\Delta \leq 9/2$. If $\Delta > 9/2, x_1$ becomes negative and must leave the solution.

15.4 Addition of a New Variable

Addition of a new variable in physical sense means introduction of a new product to the current product mix. Intuitively, it is desirable only if it is profitable i.e., if it improves the optimal value of the objective function.

Example 15.4.1. Consider the L.P. problem

$$\begin{array}{lll} \text{maximize} & Z = & 45x_1 + 100x_2 + 30x_3 + 50x_4, \\ \text{subject to} & & 7x_1 + 10x_2 + 4x_3 + 9x_4 \leq 1200, \\ & & 3x_1 + 40x_2 + x_3 + x_4 \leq 800, \\ & & x_1, x_2, x_3, x_4 \geq 0, \end{array}$$

The optimal table is given below.

Table 1

	c_j	45	100	30	50	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	b
30	x_3	$5/3$	0	1	$7/3$	$4/15$	$-1/15$	$400/3$
100	x_2	$1/30$	1	0	$-1/30$	$-1/150$	$2/75$	$40/3$
$\bar{c}_j = c_j - Z_j$		$-25/3$	0	0	$-50/3$	$-22/3$	$-2/3$	

If a new variable x_7 is added to this problem with a column $\begin{bmatrix} 10 \\ 10 \end{bmatrix}$ and $c_7 = 120$, find the change in the optimal solution.

Solution. From the Revised Simplex method, we know that

$$\bar{c}_7 = c_7 - c_B\overline{P}_7 = c_7 - c_B \cdot B^{-1} \cdot P_7 = c_7 - \pi P_7,$$

where $c_7 = 120, P_7 = \begin{bmatrix} 10 \\ 10 \end{bmatrix}$ and π, the simplex multiplier corresponding to the original optimal solution in Table 1 is given by

$$\pi = (\pi_1, \pi_2) = c_B B^{-1} = (30, 100)\begin{bmatrix} 4/15 & -1/15 \\ -1/150 & 2/75 \end{bmatrix} = \left(\frac{22}{3}, \frac{2}{3}\right).$$

$$\therefore \quad \bar{c}_7 = c_7 - \pi P_7 = 120 - \left(\frac{22}{3}, \frac{2}{3}\right)\begin{bmatrix} 10 \\ 10 \end{bmatrix} = 120 - \left(\frac{220}{3} + \frac{20}{3}\right) = 40.$$

Since $\bar{c}_7$ is positive, the existing optimal solution can be improved. Now

$$\overline{P}_7 = B^{-1}P_7 = \begin{bmatrix} 4/15 & -1/15 \\ -1/150 & 2/75 \end{bmatrix}\begin{bmatrix} 10 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/5 \end{bmatrix}.$$

Now we start with the original optimal (Table 1) and add entries corresponding to variable x_7 as follows :

Table 2

	c_j	45	100	30	50	0	0	120			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	x_7	b	θ	
30	x_3	$\frac{5}{3}$	0	1	$\frac{7}{3}$	$\frac{4}{15}$	$-\frac{1}{15}$	2	$\frac{800}{3}$	$\frac{400}{3}$	
100	x_2	$\frac{1}{30}$	1	0	$-\frac{1}{30}$	$-\frac{1}{150}$	$\frac{2}{75}$	$\left(\frac{1}{5}\right)$	$\frac{40}{3}$	$\frac{200}{3}$	←Key row
$\bar{c}_j = c_j - Z_j$		$-\frac{25}{3}$	0	0	$-\frac{50}{3}$	$-\frac{22}{3}$	$-\frac{2}{3}$	+40			
								↑K			

Replace x_2 by x_7.

Table 3

	c_j	45	100	30	50	0	0	120	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	x_7	b
30	x_3	$\frac{4}{3}$	–10	1	$\frac{8}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$	0	$\frac{400}{3}$
120	x_7	$\frac{1}{6}$	5	0	/	$-\frac{1}{30}$	$\frac{2}{15}$	1	$\frac{200}{3}$
$Z_j = \Sigma c_B a_{ij}$		60	300	30	60	6	6	120	
$\bar{c}_j = c_j - Z_j$		–15	–200	0	–10	–6	–6	0	
									Optimal feasible solution

Since $\bar{c}_j$ is negative Table 3 gives the optimal solution with $x_3 = 400/3$, $x_7 = 200/3$ (basic variable), $x_1 = x_2 = x_4 = x_5 = x_6 = 0$ (non-basic variables) and $Z_{\max} = 30 \times 400/3 + 120 \times 200/3 = 4000 + 8000 = 12000$.

15.5 Changes in the Coefficients of the Constraints (Resource requirement vector) a_{ij}

When changes take place in the constraint coefficients of a *non-basic variable* in a current optimal solution, feasibility of the solution is not affected. The only effect, if any, may be on the optimality of the solution. This effect can be studied by following the steps given in §15.4.

However, if the constraint coefficients of a *basic variable* get changed, things become more complicated since the feasibility of the current optimal solution may also be affected (lost). The basic matrix is affected, which, in turn, may affect all the quantities given in the current optimal table. Under such circumstances, it may be better to solve the problem all over again.

Example 15.5.1. Find the effect of the following changes in the original optimal Table 1 of Example 15.4.1.

(a) 'x_1'-column in the problem changes from $\begin{bmatrix} 7 \\ 3 \end{bmatrix}$ to $\begin{bmatrix} 7 \\ 5 \end{bmatrix}$.

(b) 'x_1'-column changes from $\begin{bmatrix} 7 \\ 3 \end{bmatrix}$ to $\begin{bmatrix} 5 \\ 8 \end{bmatrix}$.

Solution. (a) x_1 is a non-basic variable in the optimal solution.

$$\begin{aligned} \overline{c_1} &= c_1 - c_B\overline{P}_1 = c_1 - c_B B^{-1}P_1 \\ &= c_1 - \pi P_1, \text{ where } c_1 = 45, P_1 = \begin{bmatrix} 7 \\ 5 \end{bmatrix}, \end{aligned}$$

and $\pi = c_B B^{-1} = (30, 100)\begin{bmatrix} \frac{4}{15} & -\frac{1}{15} \\ -\frac{1}{150} & \frac{2}{75} \end{bmatrix} = \left(\frac{22}{3}, \frac{2}{3}\right)$.

$$\therefore \bar{c}_1 = 45 - \left(\frac{22}{3}, \frac{2}{3}\right)\begin{bmatrix} 7 \\ 5 \end{bmatrix} = 45 - \left(\frac{154}{3}, \frac{10}{3}\right) = 45 - \frac{164}{3} = -\frac{29}{3}.$$

Since $\bar{c}_1$ remains non-positive, the original optimum solution remains optimum for the new problem also.

(b) $\bar{c}_1 = c_1 - c_B\overline{P}_1 = c_1 - c_B B^{-1}P_1 = c_1 - \pi P_1 = 45 - \left(\frac{22}{3}, \frac{2}{3}\right)\begin{bmatrix} 5 \\ 8 \end{bmatrix} = 45 - \left(\frac{110}{3} + \frac{16}{3}\right) = +3.$

Table 1

	c_j	45	100	30	50	0	0			
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	b	θ	
30	x_3	$4/5$	0	1	$7/3$	$4/15$	$-1/15$	$800/3$	$1,000/3$	
100	x_2	$(27/150)$	1	0	$-1/30$	$-1/150$	$2/75$	$40/3$	$2,000/27$	← key row
$\bar{c}_j = c_j - Z_j$		+3	0	0	$-50/3$	$-22/3$	$-2/3$			
		↑K								

Replace x_2 by x_1.

As $\bar{c}_1$ is positive, the existing optimum solution can be improved. Now

$$\overline{P}_1 = B^{-1}P_1 = \begin{bmatrix} \frac{4}{15} & -\frac{1}{5} \\ -\frac{1}{150} & \frac{2}{75} \end{bmatrix} \begin{bmatrix} 5 \\ 8 \end{bmatrix} = \begin{bmatrix} \frac{4}{5} \\ \frac{27}{150} \end{bmatrix}.$$

Now we start with the original optimal table and incorporate the changes due to variable x_1.

Table 2

	c_j	45	100	30	50	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	b
30	x_3	0	$-40/9$	1	$67/27$	$8/27$	$-5/27$	$5,600/27$
45	x_1	1	$50/9$	0	$-5/27$	$-1/27$	$4/27$	$2,000/27$
$Z_j = \Sigma c_B a_{ij}$		45	$350/3$	30	$595/9$	$65/9$	$10/9$	
$\bar{c}_j = c_j - Z_j$		0	$-50/3$	0	$-145/9$	$-65/9$	$-10/9$	

Optimal feasible solution

Since $\bar{c}_j$ is non-positive, Table 2 gives the optimal solution with

$$x_1 = \frac{2000}{27}, \quad x_3 = \frac{5600}{27} \quad \text{(basic variables)}$$
$$x_2 = x_4 = x_5 = x_6 = 0 \quad \text{(non-basic variables)},$$
$$Z_{\max} = \frac{2000}{27} \times 45 + \frac{5600}{27} \times 30 = \frac{10000}{3} + \frac{56000}{9} = \frac{86000}{9}.$$

15.6 Addition of a New Constraint

Addition of a new constraint may or may not affect the feasibility of the current optimal solution. For this, it is sufficient to check whether new constraint is satisfied by the current optimal solution or not. If it is satisfied, the inclusion of the constraint has no effect on the current optimal solution i.e., it remains feasible as well as optimal. If, however, the constraint is not satisfied, the current optimal solution becomes infeasible. Dual simplex method is then used to find the new optimal solution.

Example 15.6.1. In problem 15.4.1 an administrative constraint is added. Products A, B and C require 2, 3 and 2 hours of administrative services, while the total available administrative hours are 10. How does the optimal solution given by Table 3 of Example 15.2.1 change?

If the total available administrative time is 4 hours, find the new optimal solution.

Solution. The optimal feasible solution given by Table 3 of Example 15.2.1 is $x_1 = 1, x_2 = 2, x_3 = 0$; while the additional constraint is $2x_1 + 3x_2 + 2x_3 \leq 10$. As this constraint is satisfied by the optimal solution, the solution remains feasible and optimal for the modified problem.

As the additional constraint $2x_1 + 3x_2 + 2x_3 \leq 4$ is not satisfied by the current optimal solution, Table 3 of Example 15.2.1 is no longer optimal for the modified problem. In order to find the new optimal solution, we add the new constraint as the third row in Table 1 below. Using s_3 as the slack variable for this constraint, the (modified) optimal table may be written as

Table 1

c_B	c_j Basis	4 x_1	6 x_2	2 x_3	0 x_4	0 x_5	0 x_6	b
4	x_1	1	0	–1	$4/3$	$-1/3$	0	1
6	x_2	0	1	2	$-1/3$	$1/3$	0	2
0	x_6	2	3	2	0	0	1	4
$Z_j = \Sigma c_B a_{ij}$		4	6	8	$10/3$	$2/3$	0	
$\bar{c}_j = c_j - Z_j$		0	0	–6	$-10/3$	$-2/3$	0	

Since x_1 and x_2 are in the basic solution, their corresponding coefficients in the basic constraint must be zero. To eliminate the coefficients of x_1 and x_2, we multiply the first row by -2, the second row by -3 and add them to the third row. Table 2 below represents the new table after the row operations. Note that $\bar{c}_j$ row is not affected since the new basic variable x_6 is the slack variable.

Table 2

c_B	c_j Basis	4 x_1	6 x_2	2 x_3	0 x_4	0 x_5	0 x_6	b	
4	x_1	1	0	–1	$4/3$	$-1/3$	0	1	
6	x_2	0	1	2	$-1/3$	$1/3$	0	2	
0	x_6	0	0	(–6)	$-5/3$	$-1/3$	1	–4	← *Key row*
$Z_j = \Sigma c_B a_{ij}$		4	6	8	$10/3$	$2/3$	0		
$\bar{c}_j = c_j - Z_j$		0	0	–6	$-10/3$	$-2/3$	0		
				↑K					

In Table 2, $\bar{c}_j$ row is optimal, but since b_3 is negative, the current basic solution is infeasible. In other words, Table 2 is dual feasible and, therefore, dual simplex method is applied to find the new optimal solution.

Evidently x_6 is the variable that leaves the basis. The ratios for the non-basis. The ratios for the non-basic variables are $1,\ 2,\ 2$ respectively. The variable x_3 which corresponds to the minimum ratio is the entering variable. The key element, -6 has been shown bracketed. Regular simplex method is used to find the optimal solution.

Table 3 is optimal and the optimal product mix is to produce $5/3$ units of product A, $2/3$ units of product B and $2/3$ units of product C with the new maximum profit = Rs. $(4 \times 5/3 + 6 \times 2/3 + 2 \times 2/3)$ = Rs. 12. Thus the addition of a new constraint decreases the optimum profit from Rs. 16 (Table 3 of Example 15.2.1) to Rs. 12. This is true of every linear programming problem. In general, whenever a new constraint is added to a linear programming problem, the old optimal value will always be better or at least equal to the new optimal value. In other words, addition of a new constraint cannot improve the optimal value of any linear programming problem.

Table 3

	c_j	4	6	2	0	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	b
4	x_1	1	0	0	$29/18$	$-5/18$	$-1/6$	$5/3$
6	x_2	0	1	0	$-8/9$	$2/9$	$1/3$	$2/3$
2	x_3	0	0	1	$5/18$	$1/18$	$-1/6$	$2/3$
$Z_j = \Sigma c_B a_{ij}$		4	6	2	$5/3$	$1/3$	1	
$\bar{c}_j = c_j - Z_j$		0	0	0	$-5/3$	$-1/3$	-1	

Optimal feasible solution

Note 15.6.2. The idea of adding new constraints can sometimes be used to reduce the computational time and hence cost of solving a linear programming problem. As the computational effort in solving a linear programming problem increases with the number of constraints, it will be advantageous to identify and delete the constraints that are not binding. Such constraints are called inactive or secondary constraints. These may pertain to resources which can be obtained easily or can be directly controlled. The new problem with fewer number of constraints is then solved. After the optimal solution is obtained, the secondary constraints are added to verify whether the optimal solution satisfies them or not. If not, the dual simplex method is applied to get the new optimal solution. No doubt, the overall saving in computational time and cost will depend on how accurately the initial judgements were made while identifying the secondary constraints.

Exercise 15.6.3.

1. What do you understand by sensitivity analysis? Explain how it is carried out.
2. What is sensitivity analysis? Discuss the effect of (i) variation of b_i, (ii) variation of c_j.
3. Discuss sensitivity analysis with respect to (i) change in constraint matrix, (ii) Addition of a new constraint.
4. Explain the basic concepts of sensitivity analysis. What are the different factors affecting the given solutions and how do we resolve them? Give a brief comment on each of them.
5. Referring to the Example 15.2.1, let us suppose that Research and Development department of the company has proposed a fourth product D which requires 1 unit of manpower and 1 unit of material and earns a unit profit of Rs. 3 when sold in the market. It is desired to find whether it is profitable to produce product D.
6. Consider the L.P. problem

$$\begin{array}{lll} \text{maximize} & Z = & 3x_1 + 5x_2 + 4x_3, \\ \text{subject to} & & 2x_1 + 3x_2 \leq 8, \\ & & 2x_2 + 5x_3 \leq 10, \\ & & 3x_1 + 2x_2 + 4x_3 \leq 15, \\ & & x_1, x_2, x_3 \geq 0, \end{array}$$

The optimal table is given below.

(a) How much c_3 and c_4 can be increased till the optimal solution given by Table 1 ceases to be optimal? Also find the new value of the objective function if possible.

	c_j	3	5	4	0	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	b
5	x_2	0	1	0	$15/41$	$8/41$	$-10/41$	$50/41$
4	x_3	0	0	1	$-6/41$	$5/41$	$/41$	$62/41$
3	x_1	1	0	0	$-2/41$	$-12/41$	$15/41$	$89/41$
	Z_j	3	5	4	$45/41$	$24/41$	$11/41$	$765/41$
	$\overline{c}_j$	0	0	0	$-45/41$	$-24/41$	$-11/41$	

(b) Find the range over which b_2 can be changed maintaining the feasibility of the solution.

7. Consider the L.P. problem

$$\begin{array}{lll} \text{maximize} & Z = & -x_1 + 2x_2 - x_3, \\ \text{subject to} & & 3x_1 + x_2 - x_3 \leq 10, \\ & & -x_1 + 4x_2 + x_3 \geq 6, \\ & & x_2 + x_3 \leq 4, \\ & & x_1, x_2, x_3 \geq 0, \end{array}$$

and the optimal solution given by table below, find the separate ranges of $b_1,\ b_2$ and b_3 consistent with the optimal solution.

	c_j	–1	2	–1	0	0	0	–M	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	A_2	b
0	x_4	3	0	–2	1	0	–1	0	6
2	x_2	0	1	1	0	0	1	0	4
0	x_5	1	0	3	0	1	4	–1	10
	Z_j	0	2	2	0	0	2	0	
	$\overline{c}_j$	–1	0	–3	0	0	–2	–M	

8 Consider the following table which represents an optimal solution to some L.P.P.: If the additional

	c_j	2	4	1	3	2	0	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	b
2	x_1	1	0	0	–1	0	$1/2$	$1/5$	–1	3
4	x_2	0	1	0	2	1	–1	0	$1/2$	1
1	x_3	0	0	1	–1	–2	5	$-3/10$	2	7
	$\overline{c}_j$	0	0	0	–2	0	–2	$-1/10$	–2	17

constraint $2x_1 + 3x_2 - x_3 + 2x_4 + 4x_5 \leq 5$ is annexed to the system, will there be any change in the optimal solution? Justify your answer.

Unit 16

Course Structure

- Parametric Programming : Variation in price vector, Variation in requirement vector.

16.1 Introduction

The study of the effect of *continuous* changes in the values of the parameters on the optimal solution to a linear programming problem is called *parametric programming*. It is an extension of sensitivity analysis and aims at finding the various basic solutions that become optimal, one after the other, as the parameters of the problem change continuously their values.

In general, parametric linear programming investigates the effect of *predetermined continuous* variations of the input coefficients on the optimal solution. It is simply an extension of sensitivity analysis and aims at finding the various basic solutions that become optimal, one after the other, as the coefficients of the problem change continuously. The coefficients change as a linear function of a single parameter, hence the name parametric linear programming for this computational technique. As in sensitivity analysis, the purpose of this technique is to reduce the additional computations required to obtain the changes in the optimal solution. The various types of parametric problem that one may come across are

1. Parametric cost problem, in which the cost coefficients c_j vary linearly as a function of parameter λ.

2. Parametric right-hand side problem, in which the resources availability coefficients b_i vary linearly as a function of parameter λ.

3. Parametric problem involving linear variations in the non-basic vector P_j of A.

4. Parametric problem involving simultaneous linear variations in $c_j,\ b_i$ and P_j.

In this unit, we will cover type 1 and type 2 parametric problems in details.

16.2 Parametric Cost Problem

Let the linear programming problem before parametrization be

$$\begin{aligned} \text{minimize} \quad & Z = CX, \\ \text{subject to} \quad & AX = b, \\ & X \geq 0, \end{aligned}$$

where C is the given cost vector. Let this cost vector change to $C + \lambda C'$ so that the parametric cost problem becomes

$$\begin{aligned} \text{minimize} \quad & Z = (C + \lambda C')X, \\ \text{subject to} \quad & AX = b, \\ & X \geq 0, \end{aligned}$$

where C' is the given predetermined cost variation vector and λ is an unknown (positive or negative) parameter. As λ changes, the cost coefficients of all variables also change. We wish to determine the family of optimal solutions as λ changes from $-\infty$ to $+\infty$.

This problem is solved by using the simplex method and sensitivity analysis. When $\lambda = 0$, the parametric cost problem reduces to the original L.P. problem; simplex method is used to find its optimal solution. Let B and X_B represent the optimal basis matrix and the optimal basic feasible solution respectively for $\lambda = 0$. The net evaluations or relative cost coefficients are all non-negative (minimization problem) and are given by

$$\bar{c}_j = c_j - Z_j = c_j - \sum c_B a_{ij} = c_j - c_B \overline{P}_j,$$

where c_B is the cost vector of the basic variables and $\overline{P}_j$ is the j-th column (corresponding to the variable x_j) in the optimal table.

As λ changes from zero to a positive or negative value, the feasible region and values of the basic variables X_B remain unaltered, but the relative cost coefficients change. For any variable x_j, the relative cost coefficient is given by

$$\begin{aligned} \bar{c}_j(\lambda) &= \left(c_j + \lambda c'_j\right) - \left(c_B + \lambda c'_B\right) \overline{P}_j \\ &= \left(c_j - c_B \overline{P}_j\right) + \lambda \left(c'_j - c'_B \overline{P}_j\right) = \bar{c}_j + \lambda \bar{c}'_j. \end{aligned}$$

Since vectors C and C' are known, $\bar{c}_j$ and $\bar{c}'_j$ can be determined. For the current minimization problem, $\bar{c}_j(\lambda)$ must be non-negative for the solution to be optimal [$\bar{c}_j(\lambda)$ must be non-positive for a maximization problem]. Thus

$$\bar{c}_j(\lambda) \geq 0, \quad \bar{c}_j + \lambda \bar{c}'_j \geq 0$$

In other words, for a given solution we can determine the range for λ within which the solution remains optimal.

Example 16.2.1. Consider the linear programming problem

$$\begin{aligned} \text{maximize} \quad & Z = 4x_1 + 6x_2 + 2x_3, \\ \text{subject to} \quad & x_1 + x_2 + x_3 \leq 3, \\ & x_1 + 4x_2 + 7x_3 \leq 9, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

The optimal solution to this problem is given by the following table: Solve this problem if the variation cost

Table 1

	c_j	4	6	2	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
4	x_1	1	0	–1	$4/3$	$-1/3$	1
6	x_2	0	1	2	$-1/3$	$1/3$	2
$Z_j = \sum c_B a_{ij}$		4	6	8	$10/3$	$2/3$	
$\bar{c}_j = c_j - Z_j$		0	0	–6	$-10/3$	$-2/3$	

vector $C' = (2, -2, 2, 0, 0)$. Identify all critical values of the parameter λ.

Solution. The given parametric cost problem is

$$\begin{aligned} \text{maximize} \quad Z =& (4+2\lambda)x_1 + (6-2\lambda)x_2 + (2+2\lambda)x_3 + 0x_4 + 0x_5, \\ \text{subject to} \quad & x_1 + x_2 + x_3 + x_4 = 3, \\ & x_1 + 4x_2 + 7x_3 + x_5 = 9, \\ & x_1, x_2, x_3, x_4, x_5 \geq 0. \end{aligned}$$

When $\lambda = 0$, the problem reduces to the L.P. problem, whose optimal solution is given by Table 1. The relative profit coefficients in this optimal table are all non-positive. For values of λ other than zero, the relative profit coefficients become linear functions of λ. To compute them, we, first, add a new relative profit row called $\bar{c}'_j$ row to Table 1. This is shown in Table 2.

Table 2

		c'_j	2	– 2	2	0	0	
		c_j	4	6	2	0	0	
c'_B	c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
2	4	x_1	1	0	–1	$4/3$	$-1/3$	1
– 2	6	x_2	0	1	2	$-1/3$	$1/3$	2
	$\bar{c}_j$		0	0	–6	$-10/3$	$-2/3$	$Z = 16$
	$\bar{c}'_j$		0	0	8	$-10/3$	$4/3$	$Z' = -2$

In Table 2, $\bar{c}'_j$ is calculated just as $\bar{c}_j$ row except that vector C is replaced by C'. For example,

$$\bar{c}_2 = c_2 - Z_2 = c_2 \sum c_B a_{i2} = c_2 - c_B \overline{P}_2 = 6 - (4, 6) \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 6 - 6 = 0.$$

Therefore,

$$\bar{c}_1' = c_1' - c_B'\overline{P}_1 = 2 - (2,-2)\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0,$$
$$\bar{c}_2' = -2 - (2,-2)\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0,$$
$$\bar{c}_3' = 2 - (2,-2)\begin{bmatrix} -1 \\ 2 \end{bmatrix} = 2 - (-2-4) = 8,$$
$$\bar{c}_4' = 0 - (2,-2)\begin{bmatrix} 4/3 \\ -1/3 \end{bmatrix} = -\left(\frac{8}{3} + \frac{2}{3}\right) = -\frac{10}{3},$$
$$\bar{c}_5' = 0 - (2,-2)\begin{bmatrix} -1/3 \\ 1/3 \end{bmatrix} = -\left(-\frac{2}{3} - \frac{2}{3}\right) = \frac{4}{3},$$
$$Z' = (1 \times 2) - (2 \times 2) = -2.$$

Table 2 represents a basic feasible solution for the given parametric cost problem. It is given by

$$x_1 = 1, x_2 = 2, x_3 = x_4 = x_5 = 0.$$

Value of the objective function, $Z(\lambda) = Z + \lambda Z' = 16 - 2\lambda$.

The relative profit coefficients, which are linear functions of λ, are given by

$$\bar{c}_j(\lambda) = \bar{c}_j + \lambda \bar{c}_j', \quad j = 1, 2, 3, 4, 5.$$

Table 2 will be optimal if $\bar{c}_j(\lambda) \leq 0$ for $j = 3, 4, 5$. Thus we can determine the range of λ for which Table 2 remains optimal as follows:

$$\bar{c}_3(\lambda) = \bar{c}_3 + \lambda \bar{c}_3' = -6 + 8\lambda \leq 0 \Rightarrow \lambda \leq 3/4$$
$$\bar{c}_4(\lambda) = \bar{c}_4 + \lambda \bar{c}_4' = -\frac{10}{3} - \frac{10}{3}\lambda \leq 0 \Rightarrow \lambda \geq -1$$
$$\bar{c}_5(\lambda) = \bar{c}_5 + \lambda \bar{c}_5' = -\frac{2}{3} + \frac{4}{3}\lambda \leq 0 \Rightarrow \lambda \leq \frac{1}{2}$$

Thus $x_1 = 1, x_2 = 2, x_3 = x_4 = x_5 = 0$ is an optimal solution for the given parametric problem for all values of λ between -1 and $1/2$ and $Z_{\max} = 16 - 2\lambda$.

For $\lambda > 1/2$, the relative profit coefficient of the non-basic variable x_5, namely $\bar{c}_5(\lambda)$ becomes positive and Table 2 no longer remains optimal. Regular simplex method is used to iterate towards optimality. x_5 is the entering variable and computation indicates x_2 to be the variable that leaves the basis matrix so that the key element is $\frac{1}{3}$. The key element is made unity and x_2 is replaced by x_5 in Table 3.

Table 3 will be optimal if $\bar{c}_j(\lambda) \leq 0$, for $j = 2, 3, 4$. Now

$$\bar{c}_2(\lambda) = \bar{c}_2 + \lambda \bar{c}_2' = 2 - 4\lambda \leq 0 \quad \therefore \lambda \geq \frac{1}{2}$$
$$\bar{c}_3(\lambda) = \bar{c}_3 + \lambda \bar{c}_3' = -2 \leq 0, \text{ which is true}$$
$$\bar{c}_4(\lambda) = \bar{c}_4 + \lambda \bar{c}_4' = -4 - 2\lambda \leq 0 \quad \therefore \lambda \geq -2$$

$\therefore$ For all $\lambda \geq \dfrac{1}{2}$, the optimal solution is given by

$$x_1 = 3, x_2 = x_3 = x_4 = 0, x_5 = 6 \text{ and } \mathrm{Z}_{\max} = 12 + 6\lambda.$$

Table 3

		c'_j	2	–2	2	0	0	
		c_j	4	6	2	0	0	
c'_B	c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
2	4	x_1	1	1	1	1	0	3
0	0	x_5	0	3	6	–1	1	6
		$\bar{c}_j$	0	2	–2	–4	0	Z = 12
		$\bar{c}'_j$	0	–4	0	–2	0	Z′ = 6

For $\lambda < -1$, the relative profit coefficient of the non-basic variable x_4, namely $\bar{c}_4(\lambda)$ becomes positive and again Table 2 no longer remains optimal. x_4 becomes the entering variable and x_1 the leaving variable. Key element is $4/3$. This element is made unity and x_1 is replaced by x_4 in Table 4.

Table 4

		c'_j	2	–2	2	0	0	
		c_j	4	6	2	0	0	
c'_B	c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
0	0	x_4	$\frac{3}{4}$	0	$-\frac{3}{4}$	1	$-\frac{1}{4}$	$\frac{3}{4}$
– 2	6	x_2	$\frac{1}{4}$	1	$\frac{7}{4}$	0	$\frac{1}{4}$	$\frac{9}{4}$
		$\bar{c}_j$	$\frac{5}{2}$	0	$-\frac{17}{2}$	0	$-\frac{3}{2}$	Z = $\frac{27}{2}$
		$\bar{c}'_j$	$\frac{5}{2}$	0	$\frac{11}{2}$	0	$\frac{1}{2}$	Z′ = $-\frac{9}{2}$

Table 4 will be optimal if $\bar{c}'_1(\lambda) \leq 0$ for $j = 1, 3, 5$. Now

$$\bar{c}_1(\lambda) = \bar{c}_1 + \lambda \bar{c}'_1 = \frac{5}{2} + \frac{5}{2}\lambda \leq 0 \qquad \therefore \lambda \leq -1,$$
$$\bar{c}_3(\lambda) = \bar{c}_3 + \lambda \bar{c}'_3 = -\frac{17}{2} + \frac{11}{2}\lambda \leq 0 \quad \therefore \lambda \leq \frac{17}{11},$$
$$\bar{c}_5(\lambda) = \bar{c}_5 + \lambda \bar{c}'_5 = -\frac{3}{2} + \frac{3}{2}\lambda \leq 0 \qquad \therefore \lambda \leq 3.$$

$\therefore$ For all $\lambda \leq -1$, the optimal solution is given by

$$x_1 = 0, x_2 = \frac{9}{4}, x_3 = 0, x_4 = \frac{3}{4}, x_5 = 0 \text{ and } Z_{\max} = \frac{27}{2} - \frac{9}{2}\lambda.$$

Thus Tables 2, 3 and 4 give families of optimal solutions for $-1 \leq \lambda \leq \frac{1}{2}, \lambda \geq \frac{1}{2}$ and $\lambda \leq -1$ respectively.

16.3 Parametric Right-Hand Side Problem

The right-hand side constants in a linear programming problem represent the limits in the resources and the outputs. In some practical problems all the resources are not independent of one another. A shortage of one resource may cause shortage of other resources at varying levels. Same is true for outputs also. For example, consider a firm manufacturing electrical appliances. A shortage in electric power will decrease the demand of all the electric items produced, in varying degrees depending upon the electric energy consumed by them.

In all such problems, we are to consider simultaneous changes in the right-hand side constants, which are functions of one parameter and study how the optimal solution is affected by these changes.

Let the linear programming problem before parameterization be

$$\begin{aligned} &\text{maximize} Z = cX, \\ &\text{subject to } AX = b, \\ &\qquad X \geq 0, \end{aligned}$$

where b is the known requirement (right-hand side) vector. Let this requirement vector b change to $b + \lambda b'$ so that parametric right-hand side problem becomes

$$\begin{aligned} &\text{maximize} Z = cX, \\ &\text{subject to } AX = b + \lambda b', \\ &\qquad X \geq 0, \end{aligned}$$

where b' is the given and predetermined variation vector and λ is an unknown parameter. As λ changes, the right-hand constants also change. We wish to determine the family of optimal solutions as λ changes from $-\infty$ to $+\infty$.

When $\lambda = 0$, the parametric problem reduces to the original L.P. problem; simplex method is used to find its optimal solution.

Let B and X_B represent the optimal basis matrix and the optimal basic feasible solution respectively for $\lambda = 0$. Then $X_B = B^{-1}b$. As λ changes from zero to a positive or negative value, the values of the basic variables change and the new values are given by

$$X_B = B^{-1}\left(b + \lambda b'\right) = B^{-1}b + \lambda B^{-1}b' = \overline{b} + \lambda \overline{b}'.$$

A change in λ has no effect on the values of relative profit coefficients $\bar{c}_j$ i.e., $\bar{c}_j$ values remain non-positive (maximization problem). For a given basis matrix B, values of $\overline{b}$ and $\overline{b}'$ can be calculated. The solution $X_B = \overline{b} + \lambda \overline{b}'$ is feasible and optimal as long as $\overline{b} + \lambda \overline{b}' \geq 0$. In other words, for a given solution we can determine the range for λ within which the solution remains optimal.

Example 16.3.1. Consider the linear programming problem

$$\begin{array}{lll} \text{maximize} & Z = & 4x_1 + 6x_2 + 2x_3, \\ \text{subject to} & & x_1 + x_2 + x_3 \leq 3, \\ & & x_1 + 4x_2 + 7x_3 \leq 9, \\ & & x_1, x_2, x_3 \geq 0. \end{array}$$

The optimal solution to this problem is given by

Solve the problem if the variation right-hand side vector $b' = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$. Perform complete parametric analysis and identify all critical values of parameter λ.

Solution. The given parametric right-hand side problem is

$$\begin{array}{lll} \text{maximize} & \text{Z} = & 4x_1 + 6x_2 + 2x_3 + 0x_4 + 0x_5, \\ \text{subject to} & & x_1 + x_2 + x_3 + x_4 = 3 + 3\lambda, \\ & & x_1 + 4x_2 + 7x_3 + x_5 = 9 - 3\lambda, \\ & & x_1, x_2, x_3, x_4, x_5 \geq 0. \end{array}$$

Table 1

	c_j	4	6	2	0	0	
c_B	Basis	x_1	x_2	x_3	x_4	x_5	b
4	x_1	1	0	–1	$4/3$	$-1/3$	1
6	x_2	0	1	2	$-1/3$	$1/3$	2
$Z_j = \sum c_B a_{ij}$		4	6	8	$10/3$	$2/3$	
$\bar{c}_j = c_j - Z_j$		0	0	–6	$-10/3$	$-2/3$	

When $\lambda = 0$, the problem reduces to the L.P. problem whose optimal solution is given by Table 1. For values of λ other than zero, the values of right-hand constants change because of the variation vector $\mathbf{b}'$. This is shown in the expanded Table 2.

Table 2

	c_j	4	6	2	0	0		
c_B	Basis	x_1	x_2	x_3	x_4	x_5	$\bar{\mathbf{b}}$	$\bar{\mathbf{b}}'$
4	x_1	1	0	–1	$4/3$	$-1/3$	1	5
6	x_2	0	1	2	$(-1/3)$	$1/3$	2	–2 ←*Key row*
	$\bar{c}_j$	0	0	–6	$-10/3$	$-2/3$	Z = 16	Z′ = 8
					↑K			

The vectors $\bar{b}$ and $\bar{b}'$ are computed as follows :

$$\bar{b} = B^{-1}b = \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 3 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

$$\bar{b}' = B^{-1}b' = \begin{bmatrix} 4/3 & -1/3 \\ -1/3 & 1/3 \end{bmatrix} \begin{bmatrix} 3 \\ -3 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \end{bmatrix}.$$

For a fixed λ, the value of basic variables in Table 2 are given by

$$x_1 = \bar{b}_1 + \lambda \bar{b}'_1 = 1 + 5\lambda, \quad x_2 = \bar{b}_2 + \lambda \bar{b}'_2 = 2 - 2\lambda.$$

$\bar{c}_j$ values are not affected as long as the basis consists of variables x_1 and x_2. As λ changes, values of basic variables x_1 and x_2 change and Table 2 remains optimal as long as the basis (x_1, x_2) remains feasible. In other words, Table 2 remains optimal as long as

$$x_1 = 1 + 5\lambda \geq 0 \Rightarrow \lambda \geq -\frac{1}{5},$$
$$x_2 = 2 - 2\lambda \geq 0 \Rightarrow \lambda \leq 1.$$

Therefore, Table 2 remains optimal as λ varies from $-1/5$ to 1 . Thus for all $-1/5 \leq \lambda \leq 1$, the optimal solution is given by

$$x_1 = 1 + 5\lambda, x_2 = 2 - 2\lambda, x_3 = x_4 = x_5 = 0, Z_{\max} = 16 + 8\lambda.$$

For $\lambda > 1$, the basic variable x_2 becomes negative. Although this makes Table 2 infeasible for the primal, it remains feasible for the dual since all $\bar{c}_j$ coefficients are non-positive. Dual simplex method can, therefore, be applied to find the new optimal solution for $\lambda > 1$. Evidently x_2 is the variable that leaves the basis. The ratios of the non-basic variables are $-3,\ 10,\ -2$. Thus variable x_4 is the entering variable. The key element $-1/3$ has been shown bracketed. Regular simplex method is now used to find the new optimal solution. In Table 3, the key element has been made unity and x_2 is replaced by x_4.

Table 3

	c_j	4	6	2	0	0		
c_B	Basis	x_1	x_2	x_3	x_4	x_5	$\bar{\mathbf{b}}$	$\bar{\mathbf{b}}'$
4	x_1	1	4	7	0	1	9	–3
0	x_4	0	–3	–6	1	–1	–6	6
$Z_j = \sum c_B a_{ij}$		4	16	28	0	4		
$\bar{c}_j = c_j - Z_j$		0	–10	–26	0	–4		

The basic solution given by Table 3 is

$$x_1 = 9 - 3\lambda,\ x_2 = 0,\ x_3 = 0,\ x_4 = -6 + 6\lambda,\ x_5 = 0,\ Z_{\max} = 36 - 12\lambda.$$

This solution is optimal as long as the basic variables x_1 and x_4 remain non-negative i.e., as long as

$$x_1 = 9 - 3\lambda \geq 0 \Rightarrow \lambda \leq 3,$$
$$x_4 = -6 + 6\lambda \geq 0 \Rightarrow \lambda \geq 1.$$

Thus the above solution is optimal for all $1 \leq \lambda \leq 3$.

For $\lambda > 3$, the basic variable x_1 becomes negative. As there is no negative coefficient in the first row, the primal solution is infeasible. Hence there exists no optimal solution to the problem for all $\lambda > 3$.

For $\lambda \leq -1/5$, the basic variable x_1 in Table 2 becomes negative. Although this makes Table 2 infeasible for the primal, it remains feasible for the dual, since all $\bar{c}_j$ coefficients are non-positive. Dual simplex method can, therefore, be applied to find the new optimal solution for $\lambda \leq -1/5$. Evidently x_1 is the variable that leaves the basis. The ratios of non-basic variables are $6,\ -5/2,\ 2$. Thus variable x_5 is the entering variable and $-1/3$ is the key element. This element is made unity in Table 4. Also x_1 is replaced by x_5. The basic

Table 4

	c_j	4	6	2	0	0		
c_B	Basis	x_1	x_2	x_3	x_4	x_5	$\bar{\mathbf{b}}$	$\bar{\mathbf{b}}'$
0	x_5	–3	0	3	–4	1	–3	–15
6	x_2	1	1	1	1	0	3	3
$Z_j = \sum c_B a_{ij}$		6	6	6	6	0		
$\bar{c}_j = c_j - Z_j$		–2	0	–4	–6	0		

solution given by Table 4 is and

$$x_1 = 0, x_2 = 3 + 3\lambda, x_3 = 0, x_4 = 0, x_5 = -3 - 15\lambda \text{ and } Z_{\max} = 18 + 18\lambda.$$

This solution is optimal so long as

$$x_2 = 3 + 3\lambda \geq 0 \Rightarrow \lambda \geq -1,$$
$$x_5 = -3 - 15\lambda \geq 0 \Rightarrow \lambda \leq -1/5$$

Thus the above solution is optimal for all $-1 \leq \lambda \leq -1/5$.

For $\lambda < -1$, the basic variable x_2 in Table 4 becomes negative. As there is no negative coefficient in the second row, the primal solution is infeasible. Hence there exists no optimal solution to the problem for all $\lambda < -1$. Thus Tables 2, 3 and 4 give families of optimal solutions for $-\frac{1}{5} \leq \lambda \leq 1, 1 \leq \lambda \leq 3$ and $-1 \leq \lambda \leq -\frac{1}{5}$ respectively.

Exercise 16.3.2. 1 Explain parametric linear programming. How does it differ from sensitivity analysis?

2 What are different types of parametric linear programming problems? Explain their solution procedures.

3 Consider the parametric problem

$$\begin{aligned} \text{maximize} \quad & Z = (\theta - 1)x_1 + x_2, \\ \text{subject to} \quad & x_1 + 2x_2 \leq 10, \\ & 2x_1 + x_2 \leq 11, \\ & x_1 - 2x_2 \leq 3, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Perform a complete parametric analysis. Identify all the critical values of the parameter θ and the optimal basic solutions.

4 Consider the parametric problem

$$\begin{aligned} \text{maximize} \quad & Z = (3 - 6\lambda)x_1 + (2 - 2\lambda)x_2 + (5 + 5\lambda)x_3, \\ \text{subject to} \quad & x_1 + 2x_2 + x_3 \leq 430, \\ & 3x_1 + 2x_3 \leq 460, \\ & x_1 + 4x_2 \leq 420, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Perform a complete parametric analysis and identify all the critical values of the parameter λ.

5 Consider the parametric problem

$$\begin{aligned} \text{maximize} \quad & Z = 3x_1 + 2x_2 + 5x_3, \\ \text{subject to} \quad & x_1 + 2x_2 + x_3 \leq 430 + \theta, \\ & 3x_1 + 2x_3 \leq 460 - 4\theta, \\ & x_1 + 4x_2 \leq 420 - 4\theta, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Determine the critical values (range) of θ for which the solution remains optimal basic feasible.

Unit 17

Course Structure

- Replacement and Maintenance Models: Failure mechanism of items, General replacement policies for gradual failure of items with constant money value and change of money value at a constant rate over the time period, Selection of best item.

17.1 Introduction

Replacement models find applications in the following situations:

1. All industrial and military equipment gets worn with time and usage and it functions with decreasing efficiency. For example, a machine requires higher operating cost, a transport vehicle such as a car or airplane requires more and more maintenance cost, a railway timetable becomes more and more out of date with the passage of time. The ever increasing repair, maintenance and operating cost necessitates the replacement of the equipment. However, there is no sharp, clearly defined time which indicates the need for this replacement. The replacement policy, in this case, consists of calculating the increased operating cost, maintenance cost, forced idle time cost together with cost of the new equipment and scrap value of the old.

2. A separate but similar problem involves the replacement of items such as electric bulbs, radio tubes, television parts, etc. which do not deteriorate with time but suddenly fail. The problem, in this case, is of finding which items to replace and whether or not to replace them in a group and, if so, when. The objective is to minimize the sum of the cost of the item, cost of replacing the item and the cost associated with failure of item.

3. Another situation in which replacement becomes necessary is obsolescence due to new discoveries and better design of the equipment. The equipment needs replacement not because it no longer performs to the designed standards, but because more modern equipment performs higher standards. For example, an equipment may have an economic life of 20 years, yet it may become obsolete after 10 years because of better technical developments.

4. Still another situation involving replacement is the staff in an organisation that gradually decreases due to death, retrenchment and other reasons.

Thus in these situations there is need to formulate a replacement policy to determine the time or age at which the replacement of the given equipment is most economical, taking into consideration all the alternatives.

17.2 Types of Failures

There are two types of failures: 1. Gradual failure and 2. Sudden failure.

1. **Gradual Failure:** Gradual failure is progressive in nature. As the life of the equipment increases, its operational efficiency decreases. This results in

   (i) increased running (repair, maintenance and operating) costs.

   (ii) decreased productivity.

   (iii) decreased resale or scrap value.

   Machines, vehicles, tyres, tubes, pistons, piston rings, bearings, etc. fall in this category.

2. **Sudden Failure:** Some items do not deteriorate with time. They give the desired level of service for some period, after which they fail. The period of desired service is not constant but follows some frequency distribution which may be progressive, retrogressive or random in nature.

   (i) *Progressive failure:* If the probability of failure of an item increases with increase in its life, then such a failure is called a progressive failure [Fig. 17.2.1 (a)]. Electric bulbs and tubes fall under this category of failure.

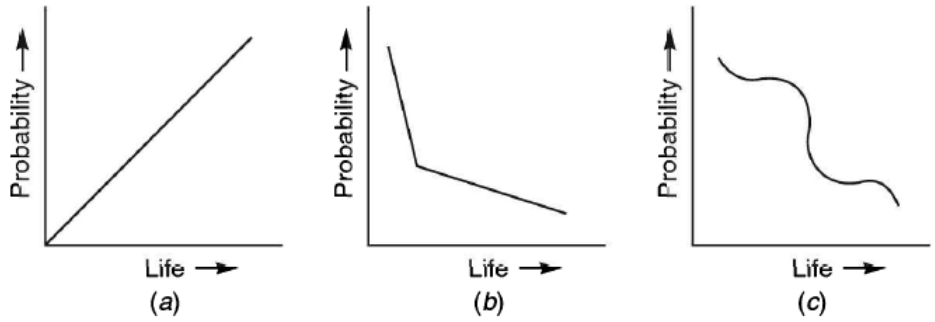


Figure 17.2.1: Sudden Failure

   (ii) *Retrogressive failure:* If the probability of failure of an item is more in the beginning but decreases with the life of an item, then such a failure is called a retrogressive failure [Fig. 17.2.1 (b]). Automobile engines fall under this category.

   (iii) *Random failure:* If the probability of failure of the item is due to random causes such as physical shock, irrespective of its age, then such a failure is called a random failure [Fig. 17.2.1 (c)]. Failure of vacuum tubes and electronic items is generally random in nature.

17.3 Replacement of items that deteriorate

(Maintenance costs increase with time

Quite often the repair, maintenance and operating costs of items increase with time and a stage may come when these costs become so high that it is more economical to replace the item by a new one. Since these costs tend to increase with time, they are grouped while analysing a problem. If these costs decrease or remain constant with time, the best policy is never to replace the item. However, this condition is hardly met with in practice. If these costs fluctuate with time, the item should be replaced only when they are increasing, of

course, the analysis becomes more involved.

Generally, all costs that depend upon the choice or age of the equipment must be taken into account while analysing the decision of its replacement. However, in special situations, certain costs may not be considered. For example, costs (such as labour cost, electric cost, etc.) that do not change with the age of the equipment may not be included in calculations. Now we shall consider a few cases of items that deteriorate with time and it will be assumed that suitable expressions for maintenance costs are available.

17.3.1 Replacement of items whose maintenance and repair costs increase with time, ignoring changes in the value of money during the period

Let us first consider a simple situation which consists of minimizing the average annual cost of an equipment whose maintenance cost is a function increasing with time and whose scrap value is constant. As the time value of money is not to be considered, the interest rate is zero and the calculations can be based on average annual cost.

Case 1. When time 't' is a continuous variable

Let

$C =$ Capital cost of the item,
$S =$ Scrap value of the item,
$T_{\text{ave}} =$ Average annual total cost of the item,
$n =$ Number of years the item is to be in use,
$f(t) =$ Operating and maintenance cost of the item at time t.

Annual cost of the item at any time $t =$ capital cost $-$ scrap value $+$ maintenance cost at time t.

Now total maintenance cost incurred during n years $= \int_0^n f(t)\, dt$.

$\therefore$ Total cost incurred during n years, $TC = C - S + \int_0^n f(t)\, dt$.

$\therefore$ Average annual cost incurred on the item,

$$\text{ATC}_n = \frac{1}{n}\left[C - S + \int_0^n f(t)\, dt\right].$$

It is desired to find the value of n for which ATC_n is minimum. Differentiating ATC_n w.r.t. n we get

$$\frac{d}{dn}(\text{ATC}_n) = -\frac{1}{n^2}(C - S) - \frac{1}{n^2}\int_0^n f(t)\, dt + \frac{1}{n}f(n). \tag{17.3.1}$$

For $\frac{d}{dn}(\text{ATC}_n) - 0$, we have

$$f(n) = \frac{1}{n}\left[C - S + \int_0^n f(t)\, dt\right] = \text{ATC}_n \tag{17.3.2}$$

Thus the item should be replaced when the average annual cost to date becomes equal to the current maintenance cost. Using this result we can decide when to replace an item provided an explicit expression is given for the maintenance and repair costs.

Case 2: When time ' t ' is a discrete variable

In this case, the total cost incurred during n years,

$$\text{TC} = C - S + \sum_{t=0}^{n} f(t). \tag{17.3.3}$$

$\therefore$ Average annual cost incurred on the item,

$$\text{ATC}_n = \frac{1}{n}\left[C - S + \sum_{t=0}^{n} f(t)\right]. \tag{17.3.4}$$

We want to find the value of n for which ATC_n is minimum.

Thus we have the inequalities $\text{ATC}_{n-1} > \text{ATC}_n < \text{ATC}_{n+1}$, which gives

$$\text{ATC}_{n-1} - \text{ATC}_n > 0 \quad \text{and} \quad \text{ATC}_{n-1} - \text{ATC}_n > 0.$$

Rewriting Eq. (17.3.4) for period $n+1$, we get

$$\begin{aligned}
\text{ATC}_{n+1} &= \frac{1}{n+1}\left[C - S + \sum_{t=1}^{n+1} f(t)\right] = \frac{1}{n+1}\left[C - S + \sum_{t=1}^{n} f(t) + f(n+1)\right] \\
&= \frac{n}{n+1}\left[\frac{1}{n}\left\{c - s + \sum_{t=1}^{n} f(t)\right\}\right] + \frac{f(n+1)}{n+1} = \frac{n}{n+1}\cdot \text{ATC}_n + \frac{f(n+1)}{n+1}.
\end{aligned}$$

$$\begin{aligned}
\therefore \quad \text{ATC}_{n+1} - \text{ATC}_n &= \frac{n}{n+1}\cdot \text{ATC}_n + \frac{f(n+1)}{n+1} - \text{ATC}_n \\
&= \frac{f(n+1)}{n+1} + \text{ATC}_n\left(\frac{n}{n+1} - 1\right) = \frac{f(n+1)}{n+1} - \frac{\text{ATC}_n}{n+1}.
\end{aligned}$$

Since $\text{ATC}_{n+1} - \text{ATC}_n > 0$, we get

$$\frac{f(n+1)}{n+1} - \frac{\text{ATC}_n}{n+1} > 0 \quad \text{or} \quad f(n+1) - \text{ATC}_n > 0 \quad \text{or} \quad f(n+1) > \text{ATC}_n.$$

Similarly, $\text{ATC}_{n-1} - \text{ATC}_n > 0$ yields $f(n) < \text{ATC}_{n-1}$. These results provide the following replacement policy:

(i) If the running cost (operating and maintenance cost) for the next year, $f(n+1)$ is more than the average annual cost of nth year, ATC_n then replace at the end of n years. That is

$$f(n+1) > \frac{1}{n}\left[C - S + \sum_{t=0}^{n} f(t)\right].$$

(ii) If the running cost of the present year is less than the previous year's average annual cost, ATC_{n-1}, then do not replace. That is

$$f(n) < \frac{1}{n-1}\left[C - S + \sum_{t=0}^{n-1} f(t)\right].$$

The above policy implies that n is optimal at the minimum average annual cost. Tabular method is used in this case. It has the advantage of being a simpler method. The examples that follow explain this method.

Example 17.3.1. The cost of a machine is Rs. 6100 and its scrap value is Rs. 100. The maintenance costs found from experience are as follows:

$$\begin{bmatrix} \text{Year} & : & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \text{Maintenance Cost} & : & 100 & 250 & 400 & 600 & 900 & 1200 & 1600 & 2000 \end{bmatrix}$$

When should the machine be replaced?

Solution. Let is be profitable to replace the machine after n years. Then n is determined by the minimum value of T_{ave}. Values of T_{ave} for various years are computed in Table 17.1.

Table 17.1

(1) *Years of service* (n)	(2) *Purchase price − scrap value* $(C - S)$	(3) *Annual maintenance cost* $f(t)$	(4) *Summation of maintenance cost* $\sum_{t=0}^{n} f(t)$	(5) *Total cost* $C - S + \sum_{t=0}^{n} f(t)$	(6) *Average annual cost* $-\left[C - S + \sum f(t)\right]$
1	6,000	100	100	6,100	6,100
2	6,000	250	350	6,350	3,175
3	6,000	400	750	6,750	2,250
4	6,000	600	1,350	7,350	1,837.50
5	6,000	900	2,250	8,250	1,650
6	6,000	1,200	3,450	9,450	**1,575**
7	6,000	1,600	5,050	11,050	1,578
8	6,000	2,000	7,050	13,050	1,631

Table 17.1 shows that the average annual cost is minimum (Rs. 1575) during sixth year then rises. Hence the machine should be replaced after 6 years of its use.

Example 17.3.2. The maintenance cost and resale value per year of a machine whose purchase price is Rs. 7000 is given below.

$$\begin{bmatrix} \text{Year} & : & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \text{Maintenance Cost} & : & 900 & 1200 & 1600 & 2100 & 2800 & 3700 & 4700 & 5900 \\ \text{Resale value} & : & 4000 & 2000 & 1200 & 600 & 500 & 400 & 400 & 400 \end{bmatrix}$$

When should the machine be replaced?

Solution. Capital cost $C =$ Rs. 7000. Let it be profitable to replace the machine after n years. Then n should be determined by the minimum value of T_{ave}. Values of T_{ave} for various years are computed in Table

Table 17.2

(1) *Years of service*	(2) *Resale value*	(3) *Purchase price–resale value (C – S)*	(4) *Annual maintenance cost* $f(t)$	(5) *Summation of maintenance cost* $\sum_{t=0}^{n} f(t)$	(6) *Total cost* $\left[C - S + \sum_{t=0}^{n} f(t)\right]$	(7) *Average annual cost* $\frac{1}{n}\left[C - S + \sum_{t=0}^{n} f(t)\right]$
1	4,000	3,000	900	900	3,900	3,900
2	2,000	5,000	1,200	2,100	7,100	3,550
3	1,200	5,800	1,600	3,700	9,500	3,166.67
4	600	6,400	2,100	5,800	12,200	3,050
5	500	6,500	2,800	8,600	15,100	**3,020**
6	400	6,600	3,700	12,300	18,900	3,150
7	400	6,600	4,700	17,000	23,600	3,371.43
8	400	6,600	5,900	22,900	29,500	3,687.50

17.2.

We observe form the table that average annual cost is minimum (Rs. 3020) in the 5th year. Hence the machine should be replaced at the end of 5 years of service.

Example 17.3.3. (a) Machine A costs Rs. $9,000$. Annual operating costs are Rs. 200 for the first year, and then increase by Rs. $2,000$ every year. Determine the best age at which to replace the machine. If the optimum replacement policy is followed, what will be the average yearly cost of owning and operating the machine? Assume that the machine has no resale value when replaced and that future costs are not discounted.

(b) Machine B costs Rs. $10,000$. Annual operating costs are Rs. 400 for the first year and then increase by Rs. 800 every year. You have now a machine of type A which is one year old. Should you replace it with B, and if so, when?

(c) Suppose you are just ready to replace machine A with another machine of the same type, when you hear that machine B will become available in a year. What would you do?

Solution. (a) It is given that the machine A has no resale value when replaced. The average annual cost is computed in Table 17.3. From Table 17.3 we find that machine A should be replaced at the end of 3 years and the average yearly cost of owning and operating the machine at this time of replacement if Rs. 5200.

(b) The average annual cost for machine B is computed in Table 17.4

Table 17.4 indicates that machine B should be replaced at the end of 5 years. Moreover, since the lowest average cost of Rs. 4000 for machine B is less than the lowest average cost of Rs. 5200 for machine A, machine A should be replaced by machine B.

Now we have to determine as to when machine A should be replaced. Machine A should be replaced when the cost for next year of running this machine becomes more than the average yearly cost for machine B.

Table 17.3

(1) Years of service (n)	(2) Resale value (S)	(3) Purchase price – resale value (C – S)	(4) Annual maintenance cost f (t)	(5) Summation of maintenance cost $\sum_{t=0}^{n} f(t)$	(6) Total cost (3) + (5)	(7) Average annual cost (6)/(1)
1	Zero	9,000	200	200	9,200	9,200
2	Zero	9,000	2,200	2,400	11,400	5,700
3	Zero	9,000	4,200	6,600	15,600	**5,200**
4	Zero	9,000	6,200	12,800	21,800	5,450
5	Zero	9,000	8,200	21,000	30,000	6,000

Table 17.4

(1) Years of service (n)	(2) Resale value (S)	(3) Purchase price– resale value (C – S)	(4) Annual maintenance cost f (t)	(5) Summation of maintenance cost $\sum_{t=0}^{n} f(t)$	(6) Total cost (3)+(5)	(7) Average annual cost (6)/(1)
1	Zero	10,000	400	400	10,400	10,400
2	Zero	10,000	1,200	1,600	11,600	5,800
3	Zero	10,000	2,000	3,600	13,600	4,533.33
4	Zero	10,000	2,800	6,400	16,400	4,100
5	Zero	10,000	3,600	10,000	20,000	**4,000**
6	Zero	10,000	4,400	14,400	24,400	4,066.67

Now, Total cost of machine A in the first year Rs. $= 9,200$,
total cost of machine A in the second year Rs. $= 11,400 - 9,200$ = Rs. 2,200,
total cost of machine A in the third year Rs. $= 4,200$,
total cost of machine A in the fourth year Rs. $= 6,200$.

As the cost of running machine A in third year (Rs. $4,200$) is more than the average yearly cost for machine B (Rs. $4,000$); machine A should be replaced at the end of two years i.e., one year after it is one year old (one year hence).

(c) As seen from part (b), machine A should be replaced one year hence and machine B will also be available at that time. Therefore, machine A should be replaced by machine B after one year from now.

17.3.2 Replacement of items whose maintenance costs increase with time and value of money also changes with time

As the money value changes with time, we must calculate the present value or present worth of the money to be spent a few years hence. If it is the interest rate (i may also be considered as the rate of inflation or the sum of the rates of interest and inflation) per year, a rupee invested at present will be equivalent to $(1 + i)$ a

year hence, $(1+i)^2$ two years hence, and $(1+i)^n$ in n years time. In other words, making a payment of one rupee after n years is equivalent to paying $(1+i)^{-n}$ now. The quantity $(1+i)^{-n}$ is called the present worth or present value of one rupee spent n years from now.

Present value of a rupee spent n years hence $= (1+i)^{-n} = \nu^n$, where, $\nu = (1+i)^{-1} = \frac{1}{1+i}$ is called discount rate or discount factor or present worth factor (pwf) and is always less than unity.

In order to find the optimal policy of replacement i.e., when a manufacturer should replace a machine on which he is working, let us assume that the machine is replaced after n years. Let C be the purchase price of the machine and $R_1, R_2, \ldots, R_n$ be the running costs is 1st, 2nd, ..., n-th year respectively. Assuming that scrap value of the machine is zero and that all payments (cash outflows) are made at the beginning of each year, the present worth of expenditure in n years is

$$P_n = C + R_1 + \nu R_2 + \nu^2 R_3 + \cdots + \nu^{n-1} R_n \tag{17.3.5}$$

Thus P_n is the amount of money required now to pay all future costs of acquiring and operating the machine assuming that it is to be replaced after n years.

Now P_n increases as n increases which means that the present worth, if the machine is replaced after $n+1$ years is greater than if it is replaced after n years. Thus for any additional amount spent we get an extra year's service. We are, therefore, interested in finding some function of the replacement interval which allows for this.

In order to do so, let us assume that the manufacturer invests the amount P_n by borrowing money at the interest rate i and repays it off in fixed annual payments, each of value x, throughout the life of the machine. Thus after n years he will have paid off the total cost P_n of the machine.

The present worth of fixed annual payments, each of value x, for n years is

$$x + \nu x + \nu^2 x + \cdots + \nu^{n-1} x = \frac{1-\nu^n}{1-\nu} x.$$

Since this is equal to the sum P_n borrowed,

$$P_n = \frac{1-\nu^n}{1-\nu} x \quad \Rightarrow \quad x = \frac{1-\nu}{1-\nu^n} P_n. \tag{17.3.6}$$

Thus the best period to replace the machine is the period n which minimizes $x = \dfrac{1-\nu}{1-\nu^n} P_n$. However, since $(1-\nu)$ is a positive constant, the period at which to replace the machine is the period n which minimizes the function $F_n = \dfrac{P_n}{1-\nu^n}$.

Since n can have only discrete values, method of finite differences can be used to calculate its optimal value. By this method, n will be optimal i.e., F_n will be minimum if

$$\Delta F_{n-1} < 0 < \Delta F_n. \tag{17.3.7}$$

Now

$$\begin{aligned} \Delta F_n = F_{n+1} - F_n &= \frac{P_{n+1}}{1-\nu^{n+1}} - \frac{P_n}{1-\nu^n} = \frac{\left(1-\nu^n\right) P_{n+1} - \left(1-\nu^{n+1}\right) P_n}{\left(1-\nu^{n+1}\right)\left(1-\nu^n\right)} \\ &= \frac{1}{\left(1-\nu^{n+1}\right)\left(1-\nu^n\right)} \left[\left(P_{n+1} - P_n\right) + \left(\nu^{n+1} P_n - \nu^n P_{n+1}\right)\right]. \end{aligned} \tag{17.3.8}$$

Further,

$$P_{n+1} = \left(C + R_1 + \nu R_2 + \cdots + \nu^{n-1} R_n\right) + \nu^n R_{n+1} = P_n + \nu^n R_{n+1}.$$

From equation (17.3.8) we get

$$\begin{aligned}
\Delta F_n &= \frac{1}{(1-\nu^{n+1})(1-\nu^n)} \left[(\nu^n R_{n+1}) + \nu^{n+1} P_n - \nu^n \{P_n + \nu^n R_{n+1}\}\right] \\
&= \frac{1}{(1-\nu^{n+1})(1-\nu^n)} \left[\nu^n R_{n+1}(1-\nu^n) - \nu^n P_n (1-\nu)\right] \qquad (17.3.9) \\
&= \frac{\nu^n (1-\nu)}{(1-\nu^{n+1})(1-\nu^n)} \left[\frac{1-\nu^n}{1-\nu} R_{n+1} - P_n\right] \\
&= \text{a positive constant} \left[\frac{1-\nu^n}{1-\nu} R_{n+1} - P_n\right]
\end{aligned}$$

Therefore, F_n has always the same sign as the quantity in brackets. Hence, from inequation (17.3.7), n will be optimal if

$$\frac{1-\nu^{n-1}}{1-\nu} R_n - P_{n-1} < 0 < \frac{1-\nu^n}{1-\nu} R_{n+1} - P_n. \qquad (17.3.10)$$

From inequation (17.3.10) we have,

$$\begin{aligned}
&\frac{1-\nu^n}{1-\nu} R_{n+1} - P_n > 0 \quad \Rightarrow \quad R_{n+1} > P_n \cdot \frac{1-\nu}{1-\nu^n} \quad \Rightarrow \quad R_{n+1} > P_n \Big/ \frac{1-\nu^n}{1-\nu} \\
&\Rightarrow R_{n+1} > \frac{C + R_1 + \nu R_2 + \nu^2 R_3 + \cdots + \nu^{n-1} R_n}{1 + \nu + \nu^2 + \cdots + \nu^{n-1}} \qquad (17.3.11) \\
&\Rightarrow \text{Next periods cost} \quad > \quad \text{Weighted average of previous costs.}
\end{aligned}$$

Since the expression on the R.H.S. of inequation (17.3.11) is the weighted average of all costs upto and including period $n-1$. The weights $1, \nu, \nu^2, \ldots, \nu^{n-1}$ are the discount factors applied to the costs in each period.

The other part of inequation (17.3.10) can, similarly, be expressed as

$$R_n < \frac{C + R_1 + \nu R_2 + \nu^2 R_3 + \cdots + \nu^{n-2} R_{n-1}}{1 + \nu + \nu^2 + \cdots + \nu^{n-2}}. \qquad (17.3.12)$$

From expressions (17.3.11) and (17.3.12) we conclude that

(a) The machine should be replaced if the next period's cost is greater than the weighted average of previous costs.

(b) The machine should not be replaced if the next period's cost is less than the weighted average of previous costs.

The corresponding value of the minimum annual payment x is obtained from equation (17.3.6) as

$$x = \frac{1-v}{1-v^n} P_n.$$

Further, if x_1 and x_2 are the minimum annual payments for two machines A and B, A will be preferred if $x_1 < x_2$ and vice versa.

It may be noted that the replacement policy of §17.3.1 which money value is ignored is a special case of this section. As interest rate $i \to 0$, the discount rate $\nu \to 1$ and expression (17.3.11) reduces to

$$R_{n+1} > \frac{C + R_1 + R_2 + \cdots + R_n}{1 + 1 + 1 + \cdots + n \text{ times}} \quad \Rightarrow \quad R_{n+1} > \frac{P_n}{n}$$

which is identical to equation (17.3.2). In actual practice, this type of replacement problem may be further complicated by the prevailing tax laws. A discussion of tax laws is beyond the scope of this book, but in any real problem the effect of taxes has got to be taken into account.

Example 17.3.4. The yearly cost of two machines A and B, when money value is neglected is shown in table below. Find their cost patterns if money is 10% per year and hence find which machine is more economical.

$$\begin{bmatrix} \text{Year} & 1 & 2 & 3 \\ \text{Machine A} & 1800 & 1200 & 1400 \\ \text{Machine B} & 2800 & 200 & 1400 \end{bmatrix}$$

Solution. The total expenditure for each machine in three years when money value is not considered is Rs. 4400. Thus the two machines are equally good if the money has no value over time. When the value of money 10% per year, the discount rate

$$\nu = \frac{1}{1 + 0.10} = \frac{1}{1.1} = 0.9091$$

The discounted cost patterns for machines A and B are shown in table below.

Year	1	2	3	Total cost (Rs.)
Machine A (Discounted cost in Rupees)	1800	1200×0.9091 $= 1090.90$	1400×0.9091^2 1157.04	404794
Machine B (Discounted cost in Rupees)	2800	200×0.9091 $= 181.82$	1400×0.9091^2 $= 1157.04$	4138.86

As total cost for machine A is less than that for machine B, machine A is more economical.

Example 17.3.5. A machine costs Rs. 500. Operation and maintenance costs are zero for the first year and increase by Rs. 100 every year. If money is worth 5% every year, determine the best age at which the machine should be replaced. The resale value of the machine is negligibly small. What is the weighted average cost of owning and operating the machine?

Solution. Discount rate,

$$v = \frac{1}{1 + r} = \frac{1}{1 + 0.05} = 0.9524$$

To find the best replacement age, we enter the calculations in a table. Table 17.5 represents these calculations. From this table we find that

$$200 < 217.61 < 300,$$

where 200 is the running cost of 3rd year and 300 is that of 4th year. Therefore, the machine should be replaced after third year. The weighted average cost of owning and operating the machine is Rs. 217.61.

Table 17.5

(1) Years of service (r)	(2) Maintenance cost (R_r)	(3) Discount factor (v^{r-1})	(4) Discounted cost ($R_r v^{r-1}$)	(5) Discounted total cost $C+\sum_{r=1}^{n} R_r v^{r-1}$	(6) Cumulative discount factor $\sum_{r=1}^{n} v^{r-1}$	(7) Weighted average annual cost $\frac{C+\sum_{r=1}^{n} R_r v^{r-1}}{\sum_{r=1}^{n} v^{r-1}}$
1	0	1.0000	0.00	500.00	1.0000	500.00
2	100	0.9524	95.24	595.24	1.9524	304.88
3	200	0.9070	181.40	776.64	2.8594	**217.61 Replace**
4	300	0.8638	259.14	1,035.78	3.7232	278.20
5	400	0.8227	329.08	1,364.86	4.5459	300.25

Example 17.3.6. The cost of a new machine is Rs. 5000. The maintenance cost during the n-th year is given by $M_n =$ Rs. $500(n-1)$, where $n = 1, 2, 3, \ldots$. If the discount rate per year is 0.05, after how many years will it be economical to replace the machine by a new one?

Solution. Since the discount rate of money is 0.05 per year, the present worth of the money to be spent after a year is

$$\nu = \frac{1}{1+0.05} = 0.9523$$

From Table 17.6 it is clear that it will be economical to replace the machine at the end of 5th year.

Table 17.6

Year (r) (1)	Maintenance cost (R_r) (2)	Discount factor (v^{r-1}) (3)	Discounted maintenance cost ($R_r v^{r-1}$) (4)	Cumulative total discounted cost $C+\sum_{r=1}^{n} R_r v^{r-1}$ (5)	Cumulative discount factor $\sum_{r=1}^{n} v^{r-1}$ (6)	Weighted average annual cost (7) = (5)/(6) (7)
1	0	1.0000	0	5,000	1.0000	5,000
2	500	0.9523	476	5,476	1.9523	2,805
3	1,000	0.9070	907	6,383	2.8593	2,232
4	1,500	0.8638	1,296	7,679	3.7231	2,063
5	2,000	0.8227	1,645	9,324	4.5458	**2,051**
6	2,500	0.7835	1,959	11,283	5.3293	2,117

Exercise 17.3.7. 1. Explain how the theory of replacement is used in replacement of items whose maintenance cost varies with time.

2. For an equipment the maintenance cost is a function increasing with time and scrap value is constant. Ignoring time value of money and considering interest rate as zero, find at what time it is advisable to replace the equipment?

3. The cost of a new machine is Rs. 5000. The maintenance cost of n-th year is given by $R_n = 500(n-1);\ n = 1, 2, \ldots$. Assuming that the money value will not change with time, after how many years will it be economical to replace the machine by new one?

4. Madras Cola Inc. uses a bottling machine that costs Rs. 50000 when new. Table below gives the expected operating costs per year, the annual expected production per year and the salvage value of the machine. The wholesale price for a bottle of drink is Rs. 1.00.

Table 17.6
Data Associated with Age of Bottling Machine

Age	1	2	3	4	5
Operating costs :	7,000	8,000	10,000	14,000	20,000
Production (Bottles) :	2,08,000	2,08,000	2,00,000	1,90,000	1,75,000
Salvage value :	30,000	19,000	15,000	12,000	10,000

When should the machine be replaced?

5. Derive the expression for the condition to replace the equipment whose maintenance costs increase with time and the value of money also changes with time.

6. Derive the following rule for minimizing costs in case of replacement of item whose maintenance costs increase with time:

   (i) Replace if the next period's cost is greater than the weighted average of the previous costs.

   (ii) Do not replace if the next period's cost is less than the weighted average of the previous costs.

7. Purchase price of a machine is Rs. 3000 and its running cost is given in the table below. If the discount rate is 0.90, find at what age the machine should be replaced.

$$\begin{bmatrix} \text{Year} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \text{Running cost} & 500 & 600 & 800 & 1000 & 1300 & 1600 & 2000 \end{bmatrix}$$

8. A company has the option to buy one of the minicomputers: MINICOMP and CHIPCOMP. MINICOMP costs Rs. 5 lakhs, and running and maintenance costs are Rs. $60,000$ for each of the first five years, increasing by Rs. $20,000$ in the sixth and subsequent years. CHIPCOMP has the same capacity as MINICOMP but costs only Rs. $2,50,000$. However, its running and maintenance costs are Rs. $1,20,000$ per year in the first five years and increase by Rs. $20,000$ per year thereafter. If the money is worth 10% per year, which computer should be purchased? What are the optimal replacement periods for each computer? Assume that there is no salvage value for either computer. Explain your analysis.

Unit 18

Course Structure

- Dynamic Programming (DP): Basic features of DP problems, Bellman's principle of optimality, Multistage decision process with Forward and Backward recursive relations, DP approach to stage-coach problems.

18.1 Introduction

In optimization problems involving a large number of decision variables or the inequality constraints, it may not be possible to use the methods of calculus for obtaining a solution. Classical mathematics handles the problems in a way to find the optimal values for all the decision variables simultaneously which for large problems rapidly increases the computations that become uneconomical or difficult to handle even by the available computers. The obvious solution is to split up the original large problem into small subproblems involving a few variables and that is precisely what the dynamic programming does. It uses recursive equations to solve a large, complex problem, broken into a series of interrelated decision stages (subproblems) wherein the outcome of the decision at one stage affects the decisions at the remaining stages.

Dynamic programming is a mathematical technique dealing with the optimization of multistage decision problems. The technique was originated in 1952 by Richard Bellman and G.B. Dantzig, and was initially referred to as the stochastic linear programming. Today dynamic programming has been developed as a mathematical technique to solve a wide range of decision problems and it forms an important part of every operation researcher's tool kit.

18.2 Characteristics of dynamic programming

The important features of dynamic programming which distinguish it from other quantitative techniques of decision-making can be summarized as follows:

1. Dynamic programming splits the original large problem into smaller subproblems (also called stages) involving only a few variables, wherein the outcome of decision at one stage affects the decisions at the remaining stages.

2. It involves a multistage process of decision-making. The points at which decisions are called for are called stages. The stages may be certain time intervals or certain subdivisions of the problems, for which independent feasible decisions are possible. Each stage can be thought of having a beginning and an end. The stages come in a sequence, the end of a stage forming the beginning of the next stage.

3. In dynamic programming, the variable that links up two stages is called a state variable. At any stage, the status of the problem can be described by the values the state variable can take. These values are referred to as states. Each stage may have, associated with it, a certain number of states. It is not essential to know about the previous decisions and how the states arise. This enables us to consider decisions one at a time.

4. In dynamic programming the outcome of decisions depends upon a small number of variables; that is, at any stage only a few variables should define the problem. For example, in the production smoothening problem, all that one needs to know at any stage is the production capacity, cost of production in regular and overtime, storage costs and the time remaining to the last decision.

5. A stage decision does not alter the number of variables on which the outcome depends, but only changes the numerical value of these variables. For the production smoothening problem, the number of variables which describe the problem i.e., production capacity, production costs, storage costs and time to the last decision, remain the same at all stages. No variable is added or dropped. The effect to decision at any stage will be to alter the used production capacity, storage cost, production cost and time remaining to the last decision.

6. Principle of Optimality. Dynamic programming is based on Bellman's Principle of Optimality, which states, "An optimal policy (a sequence of decisions) has the property that whatever the initial state and decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision". This principle implies that a wrong decision taken at one stage does not prevent from taking of optimum decisions for the remaining stages. For example, in a production scheduling problem, wrong decisions made during first and second months do not prevent taking correct decisions during third, fourth month, etc. Using this principle of optimality, we find the best policy by solving one stage at a time, and then adding a series of one-stage-problems until the overall optimum of the original problem is attained.

7. Bellman's principle of optimality forms the basis of dynamic programming technique. With this principle in mind, recursive equations are developed to take optimal decision at each stage. A recursive equation expresses subsequent state conditions and it is based on the fact that a policy is 'optimal' if the decision made at each stage results in overall optimality over all the stages and not only for the current stage.

8. Dynamic programming provides a systematic procedure wherein starting with the last stage of the problem and working backwards one makes an optimal decision for each stage of the problem. The information for the last stage is the information derived from the previous stages. It may be noted that D.P. problems can also be solved by working forward i.e., starting with the first stage and then working forward upto the last stage.

18.3 Dynamic programming approach

Before discussing the solutions to numerical problems, it will be worthwhile to know a little more about some fundamental concepts of dynamic programming. The first concept is stage. As already discussed, the problem is broken down into sub-problems and each sub-problem is referred to as a stage. A stage signifies a portion

of decision problem for which a separate decision can be made. At each stage there are a number of alternatives and the decision-making process involves the selection of one feasible alternative which may be called as stage decision. The stage decision may not be optimal for the considered stage, but contributes to make an overall optimal decision for the entire problem.

The other important concept is state. A state represents the status of the problem at a particular stage. The variables which specify the condition of decision process and summarize the current 'status' of the system are called state variables. For example, in the capital budgeting problem, the capital is the state variable. The amount of capital allocated to the present stage and the preceding stages (or the capital remaining) defines the status of the problem. The number of state variables should be as small as possible. With the increase in number of state variables, increases the difficulty of problem solving.

The procedure adopted in the analysis of dynamic programming problems can be summarized as follows:

1. Define the problem variables, determine the objective function and specify the constraints.
2. Define the stages of the problem. Determine the state variables whose values constitute the state at each stage and the decision required at each stage. Specify the relationship by which the state at one stage can be expressed as a function of the state and decisions at the next stage.
3. Develop the recursive relationship for the optimal return function which permits computation of the optimal policy at any stage. Decide whether to follow the forward or the backward method to solve the problem. Specify the optimal return function at stage 1 , since it is generally a bit different from the general optimal return function for the other stages.
4. Make a tabular representation to show the required values and calculations for each stage.
5. Find the optimal decision at each stage and then the overall optimal policy. There may be more than one such optimal policy.

18.4 Formulation of dynamic programming problems

Consider a situation wherein a certain quantity 'R' of a resource (such as men, machines, money, materials, etc.) is to be distributed among 'n' number of different activities. The return 'P' depends upon the activities and the quantities of resource allotted to them and the objective is to maximize the total return.

If $p_i\left(R_i\right)$ denotes the return form the i-th activity with the resource R_i, then the total return may be expressed as

$$P\left(R_1, R_2, \ldots, R_n\right) = p_1\left(R_1\right) + p_2\left(R_2\right) + \cdots + p_n\left(R_n\right). \tag{18.4.1}$$

The quantity of resource R is limited, which gives rise to the constraint

$$R = R_1 + R_2 + \cdots + R_n, \quad R_i \geq 0, i = 1, 2, \ldots, n. \tag{18.4.2}$$

The problem is to maximize the total return given by equation (18.4.1) subject to constraint (18.4.2). If

$$f_n(R) = \max_{0 \leq R_i \leq R} \left[P\left(R_1, R_2, \cdots, R_n\right)\right] = \max_{0 \leq R_i \leq R} \left[p_1\left(R_1\right) + p_2\left(R_2\right) + \cdots + p_n\left(R_n\right)\right], \tag{18.4.3}$$

then $f_n(R)$ is the maximum return from the distribution of the resource R to the n activities. Let us now allocate the resource to the activities, one by one, starting from the last i.e., n-th activity. An expression

connecting $f_n(R)$ and $f_{n-1}(R)$ for arbitrary values of R and n may now be obtained with the help of principle of optimality. If R_n is the quantity of resource allocated to the n-th activity such that $0 \leq R_n \leq R$, then regardless of the values of R_n, a quantity $(R - R_n)$ of the resource will be distributed among the remaining $(n-1)$ activities. Let $f_{n-1}(R - R_n)$ denote the return from the $(n-1)$ activities, then the total return from all the n activities will be

$$p_n(R_n) + f_{n-1}(R - R_n).$$

An optimal choice of R_n will maximize the above function and thus the fundamental dynamic programming model may be expressed as

$$f_n(R) = \max_{0 \leq R_n \leq R} [p_n(R_n) + f_{n-1}(R - R_n)], \quad n = 2, 3, \ldots, \tag{18.4.4}$$

where $f_1(R)$, when $n = 1$ is obtained from equation (18.4.3) as

$$f_1(R) = p_1(R). \tag{18.4.5}$$

Equation (18.4.5) gives the return from the first activity when whole of the resource R is allotted to it. Once $f_1(R)$ is known, equation (18.4.4) provides a relation to evaluate $f_2(R), f_3(R), \ldots$. This recursive process ultimately leads to the value of $f_{n-1}(R)$ and finally $f_n(R)$ at which the process stops.

Example 18.4.1. A firm has divided its marketing area into three zones. The amount of sales depends upon the number of salesman in each zone. The firm has been collecting the data regarding sales and salesmen in each area over number of past years. The information is summarized in Table 18.1. For the next year firm has only 9 salesmen and the problem is to allocate these salesmen to three different zones so that the total sales are maximum. *Solution.* In this problem the three zones represent the three stages and the number of salesmen

Table 18.1

No. of salesmen	*Zone 1*	*Zone 2*	*Zone 3*
0	*30*	*35*	*42*
1	*45*	*45*	*54*
2	*60*	*52*	*60*
3	*70*	*64*	*70*
4	*79*	*72*	*82*
5	*90*	*82*	*95*
6	*98*	*93*	*102*
7	*105*	*98*	*110*
8	*100*	*100*	*110*
9	*90*	*100*	*110*

represent the state variables.

Stage 1: We start with zone 1. The amount of sales corresponding to different number of salesmen allocated to zone 1 are given in Table 18.1 and are reproduced in Table 18.2.

Table 18.2

Zone 1

No. of salesmen:	0	1	2	3	4	5	6	7	8	9
Profit:	30	45	60	70	79	90	98	105	100	90

Stage 2: Now consider the first two zones, zone 1 and 2. Nine salesmen can be divided among two zones in 10 different ways : as 9 in zone 1 and 0 in zone 2,8 in zone 1 and 1 in zone 2,7 in zone 1 and 2 in zone 2, etc. Each combination will have associated with it certain returns. The returns for all number of salesmen (total) $9, 8, 7, \ldots, 0$ are shown in Table 18.3. For a particular number of salesmen, the profits for all possible combinations can be read along the diagonal. Maximum profits are marked by *.

Table 18.3

Zone 1 x_1:	0	1	2	3	4	5	6	7	8	9
$f_1(x_1)$:	30	45	60	70	79	90	98	105	100	90
Zone 2 x_2 / $f_2(x_2)$										
0 / 35	65*	80*	95*	105*	114	125*	133	140	135	125
1 / 45	75	90	105*	115*	124	135*	143*	150	145	
2 / 52	82	97	112	122	131	142	150	157		
3 / 64	94	109	124	134	143*	154*	162			
4 / 72	102	117	132	142	151	162				
5 / 82	112	127	142	152	161					
6 / 93	123	138	153	163*						
7 / 98	128	143	158							
8 / 100	130	145								
9 / 100	130									

Stage 3: Now consider the distribution of 9 salesmen in three zones 1, 2 and 3. The decision at this stage will result in allocating certain number of salesmen to zone 3 and the remaining to zone 2 and 1 combined; and then by following the backward process, they will be distributed to zones 2 and 1. For total of 9 salesmen to be allocated to the three zones, the returns are shown in Table 18.4 below. From Table 18.4, the maximum

Table 18.4

No. of salesmen :	0	1	2	3	4	5	6	7	8	9
Total profit $f_2(x_2) + f_1(x_1)$:	65	80	95	105	115	125	135	143	154	163
Salesmen in zone 2 + *zone* 1:	0+0	0+1	0+2	0+3	1+3	0+5	1+5	3+4	3+5	6+3
$(x_2 + x_1)$				1+2				1+6		
No. of salesmen in Zone 3:	9	8	7	6	5	4	3	2	1	0
Profit $f_3(x_3)$:	110	110	110	102	95	82	70	60	54	42
Total profit $f_3(x_3) + f_2(x_2)$:	175	190	205	207	210*	207	205	203	208	205
$+ f_1(x_1)$										

profit for 9 salesmen is Rs. 210000 if 5 salesmen are allotted to zone 3 and from the remaining four, 1 is allotted to zone 2 and 3 to zone 1.

Example 18.4.2. An oil company has 8 units of money available for exploration of three sites. If oil is present at a site, the probability of finding it depends upon the amount allocated for exploiting the site, as given below.

Table 18.5
Units of money allocated

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Site 1	*0.0*	*0.0*	*0.1*	*0.2*	*0.3*	*0.5*	*0.7*	*0.9*	*1.0*
Site 2	*0.0*	*0.1*	*0.2*	*0.3*	*0.4*	*0.6*	*0.7*	*0.8*	*1.0*
Site 3	*0.0*	*0.1*	*0.1*	*0.2*	*0.3*	*0.5*	*0.8*	*0.9*	*1.0*

The probability that the oil exists at sites 1, 2 and 3 is 0.4, 0.3 and 0.2 respectively. Find the optimal allocation of money.

Solution. In this oil exploration problem, the objective is to maximize the probability of finding oil by allocating the available amount of money to the three potential oil sites. Let x_1, x_2 and x_3 be the units of money allocated to the sites 1,2 and 3 respectively, and $p_1(x_1), p_2(x_2)$ and $p_3(x_3)$ be the corresponding probabilities of finding oil, if it exists. Then actual probabilities of finding oil at the three sites are $p_1(x_1)\times 0.4, p_2(x_2)\times 0.3$ and $p_3(x_3) \times 0.2$.

Thus the objective function can be written as

$$\begin{aligned} &\text{maximize } Z = 0.4p_1(x_1) + 0.3p_2(x_2) + 0.2p_3(x_3), \\ &\text{subject to constraint } x_1 + x_2 + x_3 \leq 8, \\ &\text{where } x_1, x_2, x_3 \text{ are non-negative integers.} \end{aligned}$$

The probabilities of finding the oil, taking into consideration the availabilities of oil at different sites, in the percentage form can be expressed as below.

Table 18.6
Units of money allocated

	0	1	2	3	4	5	6	7	8
Site 1, $f_1(x_1)$:	0	0	4	8	12	20	28	36	40
Site 2, $f_2(x_2)$:	0	3	6	9	12	18	21	24	30
Site 3, $f_3(x_3)$:	0	2	2	4	6	10	16	18	20

Here the three sites are regarded as the three stages and the money allocated is the state variable.

Stage 1: We start with site 1. The actual probabilities of finding the oil when expressed as percentages are shown in Table 18.7.

Table 18.7
Site 1

Units of money allocated:	0	1	2	3	4	5	6	7	8
$f_1(x_1)$:	0	0	4	8	12	20	28	36	40

Stage 2: Now consider the first two sites 1 and 2. Eight units of money can be divided among the two sites in 9 different ways as shown in Table 18.8.

Table 18.8

	x_1	0	1	2	3	4	5	6	7	8
	$f_1(x_1)$	0	0	4	8	12	20	28	36	40
x_2	$f_2(x_2)$									
0	0	0*	0	4	8	12*	20*	28*	36*	40*
1	3	3*	3	7	11	15	23	31	39	
2	6	6*	6	10	14	18	26	34		
3	9	9*	9	13	17	21	29			
4	12	12*	12	16	20	24				
5	18	18	18	22	26					
6	21	21	21	25						
7	24	24	24							
8	30	30								

The optimal values of $f_2(x_2) + f_1(x_1)$ are given in Table 18.9.

Table 18.9

Units of money :	0	1	2	3	4	5	6	7	8
$f_2(x_2) + f_1(x_1)$:	0	3	6	9	12	20	28	36	40
$x_2 + x_1$:	0+0	1+0	2+0	3+0	4+0 0+4	0+5	0+6	0+7	0+8

State 3: Now consider the allocation of 8 units of money to the three sites. The corresponding probabilities expressed as percentages are shown in Table 18.9.

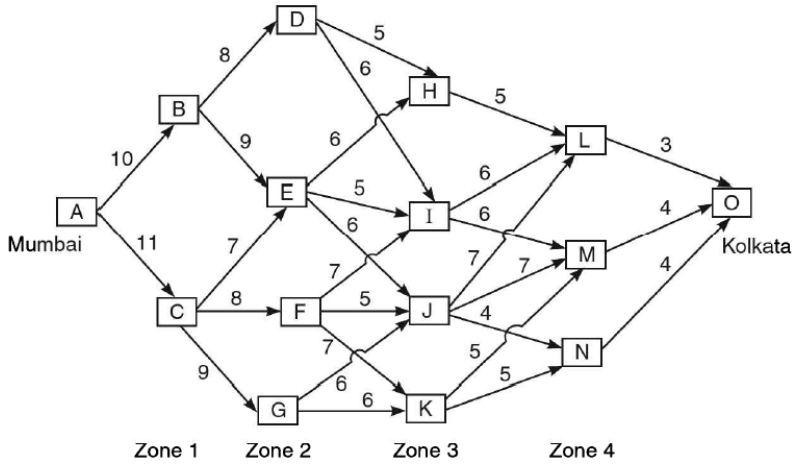
Table 18.9

Units of money :	0	1	2	3	4	5	6	7	8
$f_2(x_2) + f_1(x_1)$:	0	3	6	9	12	20	28	36	40
$x_2 + x_1$:	0+0	1+0	2+0	3+0	4 + 0 0 + 4	0+5	0+6	0+7	0+8
x_3 :	8	7	6	5	4	3	2	1	0
$f_3(x_3)$:	20	18	16	10	6	4	2	2	0
$f_3(x_3) + f_2(x_2) + f_1(x_1)$:	20	21	22	19	18	24	30	38	40

Thus the maximum probability is 40%, which is obtained if $x_3 = 0,\ x_2 = 0$, and $x_1 = 8$, i.e., if entire 8 units of money are allocated to site 1 only.

18.5 Dynamic programming approach to stage-coach problems

Example 18.5.1. A salesman is planning a business tour from Mumbai to Kolkata in the course of which he proposes to cover one city from each of the company's different marketing zones route. As has limited time at



his disposal, he has to complete his tour in the shortest possible time. The network in the above figure shows the number of days' time involved for covering any of the various intermediate cities (time includes travel as well as working time). Determine the optimum tour plan.

Solution. Starting from A, the cities of various marketing zones may be considered as distinct stages.

$$\begin{aligned} &\text{Stage1}: \quad \text{B or C ?} \\ &\text{Stage2}: \quad \text{D, E, F or G ?} \\ &\text{Stage3}: \quad \text{H, I, J or K ?} \\ &\text{Stage4}: \quad \text{L, M or N ?} \\ &\text{Stage5}: \quad \text{Best route to O.} \end{aligned}$$

Stage 1: At this stage it is not known whether B lies on the overall shortest route; but if it does, the shortest route from A to B is AB.

$$\left.\begin{aligned} &\text{A to B} = 6 \\ &\text{A to C} = 11 \end{aligned}\right\} \text{the only routes.}$$

Stage 2: It is not known whether D lies on the overall shortest route; but if it does, the only route from A is $\text{ABD} = 10 + 8 = 18$.

Similarly,

$$\left.\begin{aligned} &\text{ABE} = 10 + 9 = 19 \\ &\text{ACE} = 11 + 7 = 18 \\ &\text{ACF} = 11 + 8 = 19 \\ &\text{ACG} = 11 + 9 = 20. \end{aligned}\right\}$$

From the above, shortest routes are :

$$\left.\begin{aligned} &\text{A to } D = 18 \\ &A \text{ to } E = 18 \\ &A \text{ to } F = 19 \\ &A \text{ to } G = 20. \end{aligned}\right\}$$

Stage 3: It is not known whether H lies on the overall shortest route; but if it does, is it through D or E ?

Both D and E are reached in 18 days by the quickest route from A (from the optimal result from stage 2).

Therefore

$$\left.\begin{aligned} \text{ADH} &= 18 + 5 = 23 \\ \text{AEH} &= 18 + 6 = 24 \end{aligned}\right\}$$

Similarly,

$$\begin{aligned} \text{ADI} &= 18 + 6 = 24 \\ \text{AEI} &= 18 + 5 = 23 \\ \text{AFI} &= 19 + 7 = 26 \\ \text{AEJ} &= 18 + 6 = 24 \\ \text{AFJ} &= 19 + 5 = 24 \\ \text{AGJ} &= 20 + 6 = 26 \\ \text{AFK} &= 19 + 7 = 26 \\ \text{AGK} &= 20 + 6 = 26. \end{aligned}$$

From the above, shortest routes from A are

$$\left.\begin{aligned} \text{A to H} &= 23 \\ \text{A to I} &= 23 \\ \text{A to J} &= 24 \\ \text{A to K} &= 26. \end{aligned}\right\}$$

Stage 4: Proceeding in the same way as for stage 3 , we have

$$\begin{aligned} \text{AHL} &= 23 + 5 = 28 \\ \text{AIL} &= 23 + 6 = 29 \\ \text{AJL} &= 24 + 7 = 31 \\ \text{AIM} &= 23 + 6 = 29 \\ \text{AJM} &= 24 + 7 = 31 \\ \text{AKM} &= 26 + 5 = 31 \\ \text{AJN} &= 24 + 4 = 28 \\ \text{AKN} &= 26 + 5 = 31. \end{aligned}$$

$\therefore$ The shortest routes from A are

$$\left.\begin{aligned} A \text{ to } L &= 28 \\ \text{A to } M &= 29 \\ \text{and } A \text{ to } N &= 28. \end{aligned}\right\}$$

Final stage: There are three alternatives to reach O from the 4th stage viz. LO, MO and NO. Using the optimal times at 4th stage,

$$\left.\begin{aligned} \text{ALO} &= 28 + 3 = 31 \\ \text{AMO} &= 29 + 4 = 33 \\ \text{ANO} &= 28 + 4 = 32 \end{aligned}\right\}$$

Thus the shortest time from A to O $= 31$. Now we retrace the steps backwards along the network to identify the intermediate cities along the shortest route.

$$\begin{aligned} \text{A} - \text{O} - &\quad \text{Final} \\ \text{A} - \text{L} - \text{O} - &\quad \text{Stage } 4 \\ \text{A} - \text{H} - \text{L} - \text{O} - &\quad \text{Stage } 3 \\ \text{A} - \text{D} - \text{H} - \text{L} - \text{O} - &\quad \text{Stage } 2 \\ \text{A} - \text{B} - \text{D} - \text{H} - \text{L} - \text{O} - &\quad \text{Optimal route.} \end{aligned}$$

The problem of finding the shortest route is known as the *stage coach problem.*

18.6 Application of dynamic programming

We have discussed some over-simplified examples from the various fields of applications of dynamic programming. Many more applications are found for this decision-making technique. Whereas the linear programming has found its applications in large-scale complex situations, dynamic programming has more applications in smaller-scale systems. Following are a few of the large number of fields in which dynamic programming has ben successfully applied:

1. Production: In the production area, this technique has been employed for production, scheduling and employment smoothening, in the face of widely fluctuating demand requirements.
2. Inventory Control: This technique has been used to determine the optimum inventory level and for formulating the inventory reordering rules, indicating when to replenish an item and by what amount.
3. Allocation of Resources: It has been employed for allocating the scarce resources to different alternative uses, such as allocating salesmen to different sales zones and capital budgeting procedures.
4. Spare part level determination to guarantee high efficiency utilisation of expensive equipment.
5. Scheduling methods for routine and major overhauls on complex machinery.
6. Systematic plan or search to discover the whereabouts of a valuable resource.

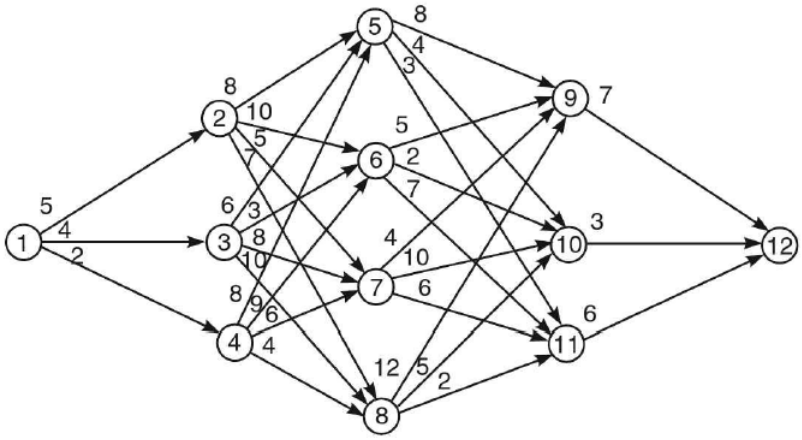
These are only a few of the wide range of situations to which dynamic programming has been successfully applied. Many real operating systems call for thousands of such decisions. The dynamic programming models make it possible to make all these decisions, of course with the help of computers. These decisions individually may not appear to be of much economic benefit, but in aggregate they exert a major influence on the economy of a firm.

Exercise 18.6.1. 1. What is dynamic programming and what sort of problems can be solved by it? State Bellman's principle of optimality and explain why it holds.

2. What is the need of dynamic programming and how is it differ from linear programming? Write some applications of dynamic programming.
3. Explain the following in the context of dynamic programming:

   (i) Principle of optimality, (ii) State, (iii) Stage

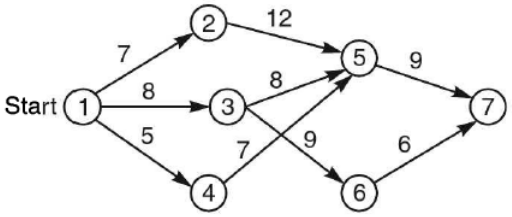
4. Write short note on characteristics of dynamic programming?



5. Find the shortest path from 1 to 12 through the network given in figure above. Also find the longest path connecting 1 and 12.

6. (a) What do you mean by forward and backward recursion in dynamic programming?

   (b) Suppose that a person wants to select the shortest highway route between two cities. The network shown below provides the possible routes between the starting city at node 1 and the destination city at node 7. The routes pass through intermediate cities designated by nodes 2 to 6. Solve the problem of finding the shortest route using dynamic programming.



Unit 19

Course Structure

- Non-Linear Programming (NLP): Lagrange Function and Multipliers, Lagrange Multipliers methods for nonlinear programs with equality and inequality constraints.

19.1 Constrained Extremal Problem for non-linear programming

The optimization problems having continuous objective function and equality or inequality type constraints are called constrained extremal problems. The solution of such problems, having differentiable objective function and equality type constraints can be obtained by a number of methods, but the most common is the Lagrange multipliers method.

19.1.1 Problem with one Equality Constraint

The use of Lagrange function can best be understood with the help of an example. Let us consider a simple two-variable problem having a single equality type constraint.

$$\begin{aligned} &\text{Maximize or minimize } Z = f(x_1, x_2), \\ &\text{subject to } g(x_1, x_2) = b, \\ &\qquad x_1, x_2 \geq 0, \end{aligned}$$

where the objective function as well as the constraint are differentiable w.r.t. x_1 and x_2 and $f(x_1, x_2)$ or $g(x_1, x_2)$ or both are non-linear. The constraint function can be replaced by another differentiable function $h(x_1, x_2)$ such that

$$h(x_1, x_2) = g(x_1, x_2) - b = 0.$$

The problem, then, reduces to

$$\begin{aligned} &\text{maximize or minimize} \quad z = f(x_1, \\ &\text{subject to} \quad h(x_1, x_2) = 0, \\ &\qquad x_1, x_2 \geq 0. \end{aligned} \tag{19.1.1}$$

The Lagrangian function can now be formulated as

$$L(x_1, x_2, \lambda) = f(x_1, x_2) - \lambda h(x_1, x_2)$$

where λ is the Lagrange multiplier. The necessary conditions for the maximum or minimum of $f(x_1, x_2)$, subject to the constraint $h(x_1, x_2) = 0$, can be obtained as

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial x_1} &= 0, \\ \frac{\partial \mathrm{L}}{\partial x_2} &= 0, \\ \text{and} \quad \frac{\partial \mathrm{L}}{\partial \lambda} &= 0, \end{aligned}$$

where $L = L(x_1, x_2, \lambda)$. If $f = f(x_1, x_2)$ and $h = h(x_1, x_2)$, the above three necessary conditions for optimization are given by

$$\begin{aligned} \frac{\partial \mathrm{L}}{\partial x_1} = \frac{\partial f}{\partial x_1} - \lambda \frac{\partial h}{\partial x_1} = 0 \quad \text{or} \quad \frac{\partial f}{\partial x_1} = \lambda \frac{\partial h}{\partial x_1}, \\ \frac{\partial \mathrm{L}}{\partial x_2} = \frac{\partial f}{\partial x_2} - \lambda \frac{\partial h}{\partial x_2} = 0 \quad \text{or} \quad \frac{\partial f}{\partial x_2} = \lambda \frac{\partial h}{\partial x_2}, \\ \text{and} \quad \frac{\partial \mathrm{L}}{\partial \lambda} = 0 - h = 0 \quad \text{or} \quad -h = 0. \end{aligned}$$

The necessary conditions for optima of $f(x_1, x_2)$, subject to $h(x_1, x_2) = 0$, are thus given by

$$\begin{aligned} f_1 &= \lambda h_1, \\ f_2 &= \lambda h_2, \\ \text{and} \quad -h &= 0. \end{aligned}$$

These necessary conditions are also the sufficient conditions for a maximum if the objective function is concave and for a minimum if the objective function is convex.

19.1.2 Necessary and Sufficient Conditions for a General NLPP

A general NLPP having n variables and m constraints $(n \geq m)$, can be expressed as

$$\begin{aligned} &\text{maximize or minimize } \; Z = f(X), \;\; X = (x_1, x_2, \ldots, x_n), \\ &\text{subject to } \; g^i(X) = b_i, \quad i = 1, 2, \ldots, m, \\ &\qquad\qquad X \geq 0. \end{aligned}$$

The constraint can also be written as

$$h^i(X) = g^i(X) - b_i = 0, i = 1, 2, \ldots, m.$$

By introducing the Lagrange multipliers, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$, the Lagrange function is formed as

$$L(X, \lambda) = f(X) - \sum_{i=1}^{m} \lambda_i h^i(X)$$

Assuming that all the functions L, f and h^i are differentiable partially w.r.t. $x_1, x_2, \ldots, x_n$ and $\lambda_1, \lambda_2, \ldots, \lambda_m$, the necessary conditions for the objective function to be a maximum or a minimum are

$$\begin{aligned} \frac{\partial L}{\partial x_j} = \frac{\partial f}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \frac{\partial h^i}{\partial x_j} = 0 \qquad \text{or} \qquad \frac{\partial f}{\partial x_j} = \sum_{i=1}^{m} \lambda_i \frac{\partial h^i}{\partial x_j} \\ \text{and} \quad \frac{\partial L}{\partial \lambda_i} = 0 - h^i = 0 \qquad \text{or} \qquad -h^i = 0, \;\; i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n. \end{aligned}$$

These $(m + n)$ necessary conditions also become the sufficient conditions for a maximum if the objective function is concave and for a minimum if the objective function is convex.

19.1.3 When concavity (convexity) is not known

As discussed in §19.1.2, for an n-variable non-linear programming problem having one equality type constraint, the necessary conditions for a stationary point to be a maximum or minimum are

$$\frac{\partial L}{\partial x_j} = \frac{\partial f}{\partial x_j} - \lambda \frac{\partial h}{\partial x_j} = 0, \quad j = 1, 2, \ldots, n,$$

and

$$\frac{\partial L}{\partial \lambda} = -h(X) = 0.$$

From the first condition, $\lambda = \dfrac{\partial f}{\partial x_j} \Big/ \dfrac{\partial h}{\partial x_j}$, for $j = 1, 2, \ldots, n$.

These necessary conditions provide an optimal solution to the problem. The sufficient conditions for determining whether the solution results in maximization or minimization of the objective function involve the solution of $(n-1)$ principal minors of the following determinant: If the signs of minors $\Delta_3, \Delta_4, \Delta_5$, etc. are

$$\Delta_{n+1} = \begin{vmatrix} 0 & \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} & \cdots & \frac{\partial h}{\partial x_n} \\ \frac{\partial h}{\partial x_1} & \frac{\partial^2 f}{\partial x_1^2} - \lambda \frac{\partial^2 h}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} - \lambda \frac{\partial^2 h}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} - \lambda \frac{\partial^2 h}{\partial x_1 \partial x_n} \\ \frac{\partial h}{\partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} - \lambda \frac{\partial^2 h}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} - \lambda \frac{\partial^2 h}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} - \lambda \frac{\partial^2 h}{\partial x_2 \partial x_n} \\ \vdots & & & & \\ \frac{\partial h}{\partial x_n} & \frac{\partial^2 f}{\partial x_n \partial x_1} - \lambda \frac{\partial^2 h}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} - \lambda \frac{\partial^2 h}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} - \lambda \frac{\partial^2 h}{\partial x_n^2} \end{vmatrix}$$

alternatively $+ve$ and $-ve$, the stationary point is a local maximum, and if all the minors are negative, the local stationary point is a minimum.

Example 19.1.1. Solve the NLPP:

$$\begin{aligned} \text{Maximize} \quad & Z = 4x_1 - x_1^2 + 8x_2 - x_2^2, \\ \text{subject to} \quad & x_1 + x_2 = 2, \\ & x_1, x_2 \geq 0. \end{aligned}$$

Solution. The objective function as well as the constraint are differentiable w.r.t. x_1 and x_2. The constraint can be replaced by another differentiable function such as

$$x_1 + x_2 - 2 = 0.$$

The Lagrangian function can be written as

$$L(x_1, x_2, \lambda) = 4x_1 - x_1^2 + 8x_2 - x_2^2 - \lambda(x_1 + x_2 - 2)$$

The necessary conditions for a maxima or minima of the objective function are and

$$\frac{\partial L}{\partial x_1} = 4 - 2x_1 - \lambda = 0 \tag{19.1.2}$$

$$\frac{\partial L}{\partial x_2} = 8 - 2x_2 - \lambda = 0 \tag{19.1.3}$$

$$\frac{\partial L}{\partial \lambda} = -(x_1 + x_2 - 2) = 0 \tag{19.1.4}$$

From (19.1.2) and (19.1.3), $4-2x_1 = 8-2x_2 \Rightarrow x_2 - x_1 = 2$, and from (19.1.4), $x_2 + x_1 = 2$, which give $x_1 = 0, x_2 = 2$ and $\lambda = 4$.

The sufficient conditions for determining whether the above solution results in maximization or minimization of the objective function involve the solution of $(n-1) = 2-1 = 1$ principal minor of the following determinant of order 3 :

$$D_3 = \begin{vmatrix} 0 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 0 & -2 \end{vmatrix} = -1(-2) + 1(2) = 2 + 2 = 4$$

Since D_3 is positive, the solution $x_1 = 0, x_2 = 2$ maximizes the objective function and

$$Z_{\max} = 0 - 0 + 16 - 4 = 12.$$

Example 19.1.2. Obtain the necessary and sufficient conditions for the optimal solution of the following problem. What is the optimal solution?

$$\begin{aligned} \text{Minimize} \quad & Z = 2e^{3x_1+1} + e^{2x_2+3}, \\ \text{subject to} \quad & x_1 + 2x_2 = 5, \\ & x_1, x_2 \geq 0. \end{aligned}$$

Solution. The objective function as well as the constraint are differentiable with respect to x_1 and x_2 and the Lagrangian function for the above problem can be formed as

$$L(X, \lambda) = 2e^{3x_1+1} + e^{2x_2+3} - \lambda\left(x_1 + x_2 - 5\right).$$

The necessary and sufficient conditions for maximization or minimization of $Z = f\left(x_1, x_2\right)$ can be obtained as

$$\frac{\partial \mathrm{L}}{\partial x_1} = 6e^{3x_1+1} - \lambda = 0 \tag{19.1.5}$$

$$\frac{\partial \mathrm{L}}{\partial x_2} = 2e^{2x_2+3} - \lambda = 0 \tag{19.1.6}$$

$$\frac{\partial \mathrm{L}}{\partial \lambda} = -\left(x_1 + x_2 - 5\right) = 0. \tag{19.1.7}$$

From (19.1.5) and (19.1.6), $6e^{3x_1+1} = 2e^{2x_2+3}$ and from (19.1.7) $x_1 + x_2 = 5$.

Therefore,

$$\begin{aligned} & 6e^{3x_1+1} = 2e^{2(5-x_1)+3} = 2e^{13-2x_1} \\ \Rightarrow\ & 3e^{3x_1+1} = e^{13-2x_1} \\ \Rightarrow\ & \log_e 3 + 3x_1 + 1 \qquad\qquad = 13\text{-}2\ x_1 \\ \Rightarrow\ & x_1 = \frac{1}{5}\left(12 - \log_e 3\right), \end{aligned}$$

and

$$x_2 = 5 - \frac{1}{5}\left(12 - \log_e 3\right) = \frac{1}{5}\left(13 + \log_e 3\right).$$

Now

$$\begin{aligned} D_3 &= \begin{vmatrix} 0 & 1 & 1 \\ 1 & 18e^{3x_1+1} & 0 \\ 1 & 0 & 4e^{2x_2+3} \end{vmatrix} \\ &= -1\left(4e^{2x_2+3}\right) + 1\left(-18e^{3x_1+1}\right) \\ &= -2\left(9e^{2x_1+3} + 2e^{2x_2+3}\right) \end{aligned}$$

Since the expression within parenthesis is positive for all values of x_1 and x_2, hence D_3 is negative. Thus, the above solution minimizes the objective function and

$$\begin{aligned} Z_{\min} &= 2e^{3\left\{\frac{1}{5}(12-\log_e 3)\right\}+1} + e^{2\left\{\frac{1}{5}(13-\log_e 3)\right\}+3} \\ &= 2e^{\frac{1}{5}(41-3\log_e 3)} + e^{\frac{1}{5}(41+2\log_e 3)} \end{aligned}$$

Example 19.1.3. Determine the optimal solution for the following NLPP and check whether it maximizes or minimizes the objective function:

$$\begin{aligned} \text{Optimize} \quad & Z = x_1^2 - 10x_1 + x_2^2 - 6x_2 + x_3^2 - 4x_3, \\ \text{subject to} \quad & x_1 + x_2 + x_3 = 7, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Solution. The objective function as well as the constraint are differentiable with respect to x_1, x_2 and x_3 and the Lagrangian function can be formed as

$$L(X, \lambda) = x_1^2 - 10x_1 + x_2^2 - 6x_2 + x_3^2 - 4x_3 - \lambda(x_1 + x_2 + x_3 - 7).$$

The necessary conditions for Z to be maximum or minimum are

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= 2x_1 - 10 - \lambda = 0, \\ \frac{\partial L}{\partial x_2} &= 2x_2 - 6 - \lambda = 0, \\ \frac{\partial L}{\partial x_3} &= 2x_3 - 4 - \lambda = 0, \\ \frac{\partial L}{\partial \lambda} &= -\left(x_1 + x_2 + x_3 - 7\right) = 0. \end{aligned}$$

The resulting solution is $x_1 = 4, x_2 = 2, x_3 = 1$ and $\lambda = -2$. To determine whether this solution results in maximization or minimization, $(n-1) = 3-1 = 2$ principal minors D_3 and D_4 of the determinants of order 3 and 4 are solved.

$$D_3 = \begin{pmatrix} 0 & \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} \\ \frac{\partial h}{\partial x_1} & \frac{\partial^2 f}{\partial x_1^2} - \lambda\frac{\partial^2 h}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} - \lambda\frac{\partial^2 h}{\partial x_1 \partial x_2} \\ \frac{\partial h}{\partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} - \lambda\frac{\partial^2 h}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} - \lambda\frac{\partial^2 h}{\partial x_2^2} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix} = -1(2)+1(-2) = -4,$$

$$D_4 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix} = -1\begin{vmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 2 \end{vmatrix} + 1\begin{vmatrix} 1 & 2 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 2 \end{vmatrix} - 1\begin{vmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \\ 1 & 0 & 0 \end{vmatrix}$$

$$= -1\{1(4)\} + 1\{1(0) - 2(2)\} - 1\{1(0) - 2(-2)\} = -4 - 4 - 4 = -12.$$

Since the principal minors D_3 and D_4 are negative, the solution $x_1 = 4, x_2 = 2, x_3 = 1$ minimizes the objective function, and

$$Z_{\min} = 16 - 40 + 4 - 12 + 1 - 4 = -35.$$

19.2 Constrained extremal problem with more than one equality constant

The non-linear programming problem having n variables and m equality constraints ($m < n$), can be expressed in the general form as

$$\begin{aligned} \text{maximize (or minimize)} \quad & Z = f(\mathrm{X}), \\ \text{subject to} \quad & h^i(X) = 0,\ i = 1, 2, \ldots, m, \\ & X \geq 0. \end{aligned}$$

The Lagrangian function can be formed as

$$L(X, \lambda) = f(X) - \sum_{i=1}^{m} \lambda_i h^i(X),$$

where $\lambda_i,\ (i = 1, 2, \ldots, m)$ are the Lagrangian multipliers. As in the previous cases, here again it is assumed that the functions $L(X, \lambda), f(X)$ and $h^i(X)$ are partially differentiable w.r.t. X and λ.

The necessary conditions for the optimum solution are

$$\frac{\partial L}{\partial x_j} = 0, j = 1, 2, \ldots, n$$
$$\frac{\partial L}{\partial \lambda_i} = 0, i = 1, 2, \ldots, m.$$

The sufficient conditions for the stationary point to be a maxima or minima are obtained by solving the principal minors of the *bordered Hessian* matrix,

$$H^B = \begin{pmatrix} O & P \\ P^T & Q \end{pmatrix}_{(m+n)\times(m+n)}$$

where O is an $m \times m$ null matrix,

$$P = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \frac{\partial h_1}{\partial x_3} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \frac{\partial h_2}{\partial x_3} & \cdots & \frac{\partial h_2}{\partial x_n} \\ \vdots & & & & \\ \frac{\partial h_m}{\partial x_1} & \frac{\partial h_m}{\partial x_2} & \frac{\partial h_m}{\partial x_3} & \cdots & \frac{\partial h_m}{\partial x_n} \end{pmatrix}_{(m\times n)} \qquad Q = \begin{pmatrix} \frac{\partial^2 L}{\partial x_1^2} & \frac{\partial^2 L}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 L}{\partial x_1 \partial x_n} \\ \frac{\partial^2 L}{\partial x_2 \partial x_1} & \frac{\partial^2 L}{\partial x_2^2} & \cdots & \frac{\partial^2 L}{\partial x_2 \partial x_n} \\ \vdots & & & \\ \frac{\partial^2 L}{\partial x_n \partial x_1} & \frac{\partial^2 L}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 L}{\partial x_n^2} \end{pmatrix}$$

If (X^*, λ^*) is the stationary point for the function $L(X, \lambda)$ and H^{B^*} is the corresponding bordered Hessian matrix, the sufficient but not necessary condition for the maxima and minima is determined by the signs of the last $(n - m)$ principal minors of H^{B^*}, starting with the principal minor of the order $(2m + 1)$.

X^* maximizes the function if the signs alternate, starting with $(-1)^{m+n}$ and X^* minimizes the function if all the signs are same and of the $(-1)^m$ type.

Example 19.2.1. Solve the non-linear programming problem given below:

$$\begin{array}{ll} \text{Optimize} & Z = x_1^2 + x_2^2 + x_3^2, \\ \text{subject to} & x_1 + x_2 + 3x_3 = 2, \\ & 5x_1 + 2x_2 + x_3 = 5, \\ & x_1, x_2, x_3 \geq 0. \end{array}$$

Solution. The objective function as well as constraints are differentiable with respect of x_1, x_2 and x_3 and the Lagrangian function is formed as

$$L(X, \lambda) = x_1^2 + x_2^2 + x_3^2 - \lambda_1 (x_1 + x_2 + 3x_3 - 2) - \lambda_2 (5x_1 + 2x_2 + x_3 - 5)$$

The necessary conditions for the maxima or minima of the objective function are obtained as

$$\frac{\partial L}{\partial x_1} = 2x_1 - \lambda_1 - 5\lambda_2 = 0 \tag{19.2.1}$$

$$\frac{\partial L}{\partial x_2} = 2x_2 - \lambda_1 - 2\lambda_2 = 0, \tag{19.2.2}$$

$$\frac{\partial L}{\partial x_3} = 2x_3 - 3\lambda_1 - \lambda_2 = 0, \tag{19.2.3}$$

$$\frac{\partial L}{\partial \lambda_1} = -(x_1 + x_2 + 3x_3 - 2) = 0, \tag{19.2.4}$$

$$\frac{\partial L}{\partial \lambda_2} = -(5x_1 + 2x_2 + x_3 - 5) = 0. \tag{19.2.5}$$

Substituting the values of x_1, x_2, x_3 from , and in and , we get

$$\frac{\lambda_1 + 5\lambda_2}{2} + \frac{\lambda_1 + 2\lambda_2}{2} + 3\left(\frac{3\lambda_1 + \lambda_2}{2}\right) - 2 = 0 \quad \text{or} \quad 11\lambda_1 + 10\lambda_2 = 4, \tag{19.2.6}$$

$$\text{and } \frac{5 \cdot (\lambda_1 + 5\lambda_2)}{2} + \frac{2(\lambda_1 + 2\lambda_2)}{2} + \frac{3\lambda_1 + \lambda_2}{2} - 5 = 0 \quad \text{or,} \quad \lambda_1 + 3\lambda_2 = 1. \tag{19.2.7}$$

Solving (19.2.6) and (19.2.7), we have $\lambda_1 = 0.087$ and $\lambda_2 = 0.304$. Equations (19.2.1), (19.2.1) and (19.2.1) yield $x_1 = 0.804, x_2 = 0.348$ and $x_3 = 0.283$ as the solution.

To determine whether this solution point is a maxima or minima, the following bordered Hessian matrix is constructed:

$$H^B = \begin{bmatrix} O & P \\ P^T & Q \end{bmatrix}_{(m+n)\times(m+n)}$$

where

$$O = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad P = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 1 \end{bmatrix}, \quad P^T = \begin{bmatrix} 1 & 5 \\ 1 & 2 \\ 3 & 1 \end{bmatrix}, \quad \text{and} \quad Q = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Therefore,

$$H^B = \left[\begin{array}{cc|ccc} 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 5 & 2 & 1 \\ \hline 1 & 5 & 2 & 0 & 0 \\ 1 & 2 & 0 & 2 & 0 \\ 3 & 1 & 0 & 0 & 2 \end{array}\right].$$

Since $n = 3, m = 2; n - m = 1$ and $2m + 1 = 5$. This means that only one principal minor of H^B of order 5 needs to be solved. For maximization, the sign should be $(-1)^{m+n} = (-1)^5 = -ve$ and for minimization, the sign should be $(-1)^m = (-1)^2 = +$ve. Now the determinant of H^B of order 5 is

$$\begin{vmatrix} 0 & 0 & 1 & 1 & 3 \\ 0 & 0 & 5 & 2 & 1 \\ 1 & 5 & 2 & 0 & 0 \\ 1 & 2 & 0 & 2 & 0 \\ 3 & 1 & 0 & 0 & 2 \end{vmatrix} = 1\begin{vmatrix} 0 & 0 & 2 & 1 \\ 1 & 5 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 3 & 1 & 0 & 2 \end{vmatrix} - 1\begin{vmatrix} 0 & 0 & 5 & 1 \\ 1 & 5 & 2 & 0 \\ 1 & 2 & 0 & 0 \\ 3 & 1 & 0 & 2 \end{vmatrix} + 3\begin{vmatrix} 0 & 0 & 5 & 2 \\ 1 & 5 & 2 & 0 \\ 1 & 2 & 0 & 2 \\ 3 & 1 & 0 & 0 \end{vmatrix}$$

$$= 1\left[2\begin{vmatrix} 1 & 5 & 0 \\ 1 & 2 & 0 \\ 3 & 1 & 2 \end{vmatrix} - 1\begin{vmatrix} 1 & 5 & 0 \\ 1 & 2 & 2 \\ 3 & 1 & 0 \end{vmatrix}\right] - 1\left[5\begin{vmatrix} 1 & 5 & 0 \\ 1 & 2 & 0 \\ 3 & 1 & 2 \end{vmatrix} - 1\begin{vmatrix} 1 & 5 & 2 \\ 1 & 2 & 0 \\ 3 & 1 & 0 \end{vmatrix}\right]$$
$$+3\left[5\begin{vmatrix} 1 & 5 & 0 \\ 1 & 2 & 2 \\ 3 & 1 & 0 \end{vmatrix} - 2\begin{vmatrix} 1 & 5 & 2 \\ 1 & 2 & 0 \\ 3 & 1 & 0 \end{vmatrix}\right]$$

$$\begin{aligned} &= 1[2\{1(4-0) - 5(2-0)\} - 1\{1(0-2) - 5(0-6)\}] \\ &\quad -1[5\{1(4-0) - 5(2-0)\} - 1\{1(0-0) - 5(0-0) + 2(1-6)\}] \\ &\quad +3[5\{1(0-2) - 5(0-6)\} - 2\{1(0-0) - 5(0-0) + 2(1-6)\}] \\ &= 1[2(4-10) - 1(-2+30)] - 1[5(4-10) - 1(0-0-10)] \\ &\quad +3[5(-2+30) - 2(0-0-10)] \\ &= 1[-12-28] - 1[-30+10] + 3[140+20] = -40 + 20 + 480 = 460 \end{aligned}$$

Since the value is $+ve$, the above solution minimizes the objective function and

$$Z_{\min} = (0.804)^2 + (0.348)^2 + (0.283)^2 = 0.847.$$

19.3 Non-linear programming problem with one inequality constraint

Consider a general non-linear programming problem having one inequality constraint of the type

$$\begin{aligned} \text{Maximize} \quad & \text{Z} = f(X), \\ \text{subject to} \quad & g(X) \leq b, \\ & X \geq 0, X = x_1, x_2, \ldots, x_n. \end{aligned}$$

Introducing a slack variable S in the form of S^2 so as to ensure that it is always non-negative, the constraint equation can be modified to $h(X) + S^2 = 0$, where $h(X) = g(X) - b \leq 0$.

The problem can now be expressed as

$$\begin{aligned} \text{Maximize} \quad & Z = f(X), \\ \text{subject to} \quad & h(X) + S^2 = 0, \\ & X \geq 0, \end{aligned}$$

which is an $(n+1)$ variable, single equality constraint problem of constrained optimization and can be solved by the method of Lagrange multipliers. The Lagrangian function can be constructed as

$$L(X, S, \lambda) = f(X) - \lambda \left[h(X) + S^2\right]$$

The necessary conditions for the stationary point are

$$\frac{\partial L}{\partial x_j} = \frac{\partial f}{\partial x_j} - \lambda \frac{\partial h}{\partial x_j} = 0, j = 1, 2, \ldots, n$$
$$\frac{\partial L}{\partial \lambda} = -\left[h(\mathrm{X}) + S^2\right] = 0,$$
$$\text{and } \frac{\partial L}{\partial S} = -2S\lambda = 0.$$

The condition $\dfrac{\partial L}{\partial S} = 0$ implies that either $S = 0$ or $\lambda = 0$. If $S = 0$, then condition $\frac{\partial L}{\partial \lambda} = 0$ gives $h(X) = 0$. Thus either λ or $h(X) = 0$, i.e., $\lambda \cdot h \cdot (\mathrm{X}) = 0$.

Since S^2 has been taken to be a non-negative slack variable, $h(X) \leq 0$. This implies that when $h(X) < 0, \lambda = 0$; and when $\lambda > 0, h(X) = 0$. The necessary conditions for maximization problem can thus be summarized as

$$\frac{\partial f}{\partial x_j} - \lambda \frac{\partial h}{\partial x_j} = 0,$$
$$\lambda h(X) = 0,$$
$$h(X) \leq 0,$$
$$\lambda \geq 0.$$

These necessary conditions are also called *Kuhn-Tucker conditions*. A similar argument holds for the minimization non-linear programming problem :

$$\begin{array}{ll} \text{Minimize} & Z = f(X) \\ \text{subject to} & g(X) \geq b \\ & X \geq 0 \end{array}$$

Introduction of $h(X) = g(X) - b$, reduces the constraint to $h(X) \geq 0$. The surplus variable S^2 can be introduced so that the constraint becomes $h(X) - S^2 = 0$. The appropriate Lagrangian functions is

$$L(X, S, \lambda) = f(X) - \lambda \left[h(X) - S^2\right].$$

Following an analysis similar to the one for maximization problem, the *Kuhn-Tucker conditions* for the minimization non-linear programming problem can be obtained as:

$$\frac{\partial f}{\partial x_j} - \lambda \frac{\partial h}{\partial x_j} = 0,$$
$$\lambda h(X) = 0,$$
$$h(X) \geq 0,$$
$$\lambda \leq 0.$$

For a single constraint non-linear programming problem, the necessary Kuhn-Tucker conditions are also the sufficient conditions for

1. the maximization problem, when $f(X)$ is concave and $h(X)$ is convex.

2. the minimization problem, when both $f(X)$ and $h(X)$ are convex.

Example 19.3.1. Solve the following non-linear programming problem:

$$\begin{aligned} \text{Maximize} \quad & Z = 4x_1 - x_1^3 + 2x_2, \\ \text{subject to} \quad & x_1 + x_2 \leq 1, \\ & x_1, x_2 \geq 0. \end{aligned}$$

Solution. The problem can be put as

$$\begin{aligned} f(X) &= 4x_1 - x_1^3 + 2x_2, \\ h(X) &= x_1 + x_2 - 1. \end{aligned}$$

The problem is of maximization, and Kuhn-Tucker conditions are

$$\begin{aligned} \frac{\partial f(X)}{\partial x_j} - \lambda \frac{\partial h(X)}{\partial x_j} &= 0, \\ \lambda h(X) &= \\ h(X) &\leq 0, \\ \lambda &\geq 0. \end{aligned}$$

Applying these conditions, we get

$$4 - 3x_1^2 - \lambda = 0, \tag{19.3.1}$$

$$2 - \lambda = 0, \tag{19.3.2}$$

$$\lambda(x_1 + x_2 - 1) = 0, \tag{19.3.3}$$

$$x_1 + x_2 - 1 \leq 0, \tag{19.3.4}$$

$$\lambda \geq 0. \tag{19.3.5}$$

From (19.3.2) $\lambda = 2$, therefore from (19.3.3) $x_1 + x_2 - 1 = 0$. These results satisfy the conditions (19.3.4) and (19.3.5). Solution of (19.3.1), (19.3.2) and (19.3.3) yields

$$x_1 = \sqrt{2/3} = 0.8165 \quad \text{and} \quad x_2 = 1 - \sqrt{2/3} = 0.1835$$

It can be easily observed that $f(X)$ is concave in X, while $h(X)$ is a convex function. Hence, the solution $X^* = (0.8165,\ 0.1835)$ maximizes the objective function which comes to $Z_{\max} = 3.0887$.

19.4 Non-linear programming problem with more than one inequality constraint

Let us consider a general non-linear programming problem of the maximization type.

$$\begin{aligned} \text{Maximize} \quad Z = \quad & f(X), \\ \text{subject to} \quad & g^i(X) \leq b_i, \\ & X \geq 0; i = 1, 2, \ldots, m. \end{aligned}$$

The constraint equation can be written in the form

$$h^i(X) = g^i(X) - b_i \leq 0,$$

which can be further modified to equality constraint by introducing slack variables.

$$\therefore \quad h^i(X) + S_i^2 = 0, \quad i = 1, 2, \ldots, m.$$

The Lagrangian function is constructed as

$$L(X, S, \lambda) = f(X) - \sum_{i=1}^{m} \lambda_i \left[h^i(X) + S_i^2\right].$$

The necessary conditions for maximization are

$$\frac{\partial L}{\partial x_j} = \frac{\partial f(X)}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \frac{\partial h^i X}{\partial x_j} = 0, \tag{19.4.1}$$

$$\frac{\partial L}{\partial \lambda_i} = -\left[h^i(X) + S_i^2\right] = 0, \tag{19.4.2}$$

$$\frac{\partial L}{\partial S_i} = -2S_i\lambda_i = 0, \tag{19.4.3}$$

$$i = 1, 2, \ldots, m,$$

$$j = 1, 2, \ldots, n.$$

The conditions (19.4.2) and (19.4.3) can be replaced by the following set of conditions, by carrying out analysis similar to the one done in case of single inequality constraint.

$$\lambda_i h^i(X) = 0 \tag{19.4.4}$$

$$h^i(X) \leq 0 \tag{19.4.5}$$

$$\lambda_i \geq 0. \tag{19.4.6}$$

Thus the Kuhn-Tucker conditions for a non-linear programming problem of maximizing $f(X)$ subject to the constraints $h^i(X) \leq 0$, can be summarized as

$$f_j(X) - \sum_{i=1}^{m} \lambda_i h_j^i(X) = 0$$

$$\lambda_i h^i(X) = 0,$$

$$h^i(X) \leq 0,$$

$$\lambda_i \geq 0,$$

$$i = 1, 2, \ldots, m,$$

$$j = 1, 2, \ldots, n.$$

It can be shown that the Kuhn-Tucker conditions for a minimization non-linear programming problem are

$$f_j(X) - \sum_{i=1}^{m} \lambda_i h_j^i(X) = 0$$

$$\lambda_i h^i(X) = 0$$

$$h^i(X) \geq 0$$

$$\lambda_i \geq 0$$

The Kuhn-Tucker conditions are also the sufficient conditions.

1. for a maximum, if $f(X)$ is concave and all $h^i(X)$ are convex in X.
2. for a minimum, if $f(X)$ is convex and all $h^i(X)$ are concave in X.

In both the maximization and minimization problems, the Lagrange's multipliers corresponding to the equality constraints must be unrestricted in sign. In maximization problems all constraints should be of $\leq$ type, while in minimization, the constraints should be of $\geq$ type. These conditions can be obtained by performing the necessary transformations as discussed in linear programming.

Example 19.4.1. Solve the following NLPP:

$$\begin{array}{ll} \text{Maximize} & Z = 7x_1^2 + 6x_1 + 5x_2^2 \\ \text{subject to} & x_1 + 2x_2 \leq 10 \\ & x_1 - 3x_2 \leq 9 \\ & x_1, x_2 \geq 0 \end{array}$$

Solution. We have

$$\begin{aligned} f(X) &= 7x_1^2 + 6x_1 + 5x_2^2 \\ h^1(X) &= x_1 + 2x_2 - 10 \\ h^2(X) &= x_1 - 3x_2 - 9 \end{aligned}$$

The Kuhn-Tucker conditions for a maximization problem are

$$\begin{aligned} f_j(X) - \sum_{i=1}^{m} \lambda_i h_j^i(X) &= 0 \\ \lambda_i h^i(X) &= 0 \\ h^i(X) &\leq 0 \end{aligned}$$

Applying these conditions, we get

$$14x_1 + 6 - \lambda_1 - \lambda_2 = 0, \tag{19.4.7}$$
$$10x_2 - 2\lambda_1 - 3\lambda_2 = 0, \tag{19.4.8}$$
$$\lambda_1(x_1 + 2x_2 - 10) = 0, \tag{19.4.9}$$
$$\lambda_2(x_1 - 3x_2 - 9) = 0, \tag{19.4.10}$$
$$x_1 + 2x_2 - 10 \leq 0,$$
$$x_1 - 3x_2 - 9 \leq 0,$$
$$\lambda_1,\ \lambda_2 \geq 0.$$

Here we have two Lagrange's multipliers λ_1 and λ_2 which can take zero or non-zero positive values. Thus four solutions corresponding to the following four combinations of $\lambda_i (i = 1, 2)$ values can be obtained:

(i) $\lambda_1 = 0, \quad \lambda_2 = 0;$

(ii) $\lambda_1 = 0, \quad \lambda_2 \neq 0;$

(iii) $\lambda_1 \neq 0, \quad \lambda_2 = 0;$

(iv) $\lambda_1 \neq 0, \quad \lambda_2 \neq 0;$

Solution 1: $\lambda_1 = 0$ and $\lambda_2 = 0$ result in $x_1 = -\frac{6}{14}$ and $x_2 = 0$, which is an infeasible solution.

Solution 2: $\lambda_1 = 0, \lambda_2 \neq 0$. Since $\lambda_2 \neq 0$, from (19.4.10) $x_1 - 3x_2 - 9 = 0$, from (19.4.7) and (19.4.8), $14x_1 + 6 - \lambda_2 = 0$, and $10x_2 + 3\lambda_2 = 0$. Solution of these equations yields

$$x_1 = \frac{19}{119}, x_2 = -\frac{1,052}{357}, \lambda_1 = 0, \lambda_2 = \frac{980}{119}.$$

This again is an infeasible solution.

Solution 3: $\lambda_1 \neq 0, \lambda_2 = 0$. From (19.4.7), (19.4.8) and (19.4.9) we have

$$\begin{aligned} 14x_1 + 6 - \lambda_1 &= 0, \\ 10x_2 - 2\lambda_1 &= 0 \\ x_1 + 2x_2 - 10 &= 0 \end{aligned}$$

The solution of these equations yields

$$x_1 = \frac{38}{33}, x_2 = \frac{146}{33}, \lambda_1 = \frac{730}{33}, \lambda_2 = 0.$$

This is a feasible solution giving $Z = 114.061$.

Solution 4: $\lambda_1 \neq 0, \lambda_2 \neq 0$ From (19.4.7), (19.4.8), (19.4.9) and (19.4.10), we have

$$\begin{aligned} 14x_1 + 6 - \lambda_1 - \lambda_2 &= 0, \\ 10x_2 - 2\lambda_1 + 3\lambda_2 &= 0, \\ x_1 + 2x_2 - 10 &= 0, \\ \text{and} \quad x_1 - 3x_2 - 9 &= 0. \end{aligned}$$

The solution of these four equations yields

$$x_1 = \frac{48}{5}, x_2 = \frac{1}{5}, \lambda_1 = \frac{2,116}{25}, \lambda_2 = \frac{1,394}{25}.$$

This also is a feasible solution giving $Z = 702.92$. Since the maximum value of Z is obtained for solution 4 , where $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, the optimal solution is

$$x_1{}^* = \frac{48}{5}, x_2{}^* = \frac{1}{5} \text{ and } Z_{\max} = 702.92.$$

Exercise 19.4.2. 1. Solve the NLPP:

$$\begin{aligned} \text{Maximize} \quad & Z = 4x_1 + 6x_2 - 2x_1^2 - 2x_1x_2 - 2x_2^2, \\ \text{subject to} \quad & x_1 + 2x_2 = 2, \\ & x_1, x_2 \geq 0. \end{aligned}$$

2. Solve the following NLPP by using Lagrange multipliers method:

$$\begin{aligned} \text{Maximize} \quad & Z = x_1^2 + x_2^2 + x_3^2, \\ \text{subject to} \quad & 4x_1 + x_2^2 + 2x_3 = 14, \\ & x_1, x_2 \geq 0. \end{aligned}$$

3. Use the method of Lagrangian multipliers to solve the following NLPP. Does the solution maximize or minimize the objective function?

$$\begin{aligned} \text{Optimize} \quad & Z = 2x_1^2 + x_2^2 + 3x_3^2 + 10x_1 + 8x_2 + 6x_3 - 100, \\ \text{subject to} \quad & x_1 + x_2^2 + x_3 = 20, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

4. Solve the following non-linear programming problem, using the Lagrange multipliers:

$$\begin{aligned} \text{Optimize} \quad & Z = 4x_1^2 + 2x_2^2 + x_3^2 - 4x_1x_2, \\ \text{subject to} \quad & x_1 + x_2^2 + x_3 = 15, \\ & 2x_1 - x_2 + 2x_3 = 20, \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

5. Solve the following non-linear programming problem:

$$\begin{aligned} \text{Optimize} \quad & Z = 4x_1 + 9x_2 - x_1^2 - x_2^2, \\ \text{subject to} \quad & 4x_1 + 3x_2 = 15, \\ & 3x_1 + 5x_2 = 14, \\ & x_1, x_2 \geq 0. \end{aligned}$$

6. Solve the following non-linear programming problem using the Kuhn-Tucker conditions:

$$\begin{aligned} \text{Maximize} \quad & Z = 10x_1 + 4x_2 - 2x_1^2 - x_2^2, \\ \text{subject to} \quad & 2x_1 + x_2 \leq 5, \\ & x_1, x_2 \geq 0. \end{aligned}$$

7. Solve the following NLPP using the Kuhn-Tucker conditions:

$$\begin{aligned} \text{Maximize} \quad & Z = 2x_1^2 - 7x_2^2 + 12x_1x_2, \\ \text{subject to} \quad & 2x_1 + 5x_2 \leq 98, \\ & x_1, x_2 \geq 0. \end{aligned}$$

8. Use the Kuhn-Tucker conditions to solve the following non-linear programming problem:

$$\begin{aligned} \text{Maximize} \quad & Z = 7x_1^2 - 6x_1 + 5x_2^2, \\ \text{subject to} \quad & x_1 + 2x_2 \leq 10, \\ & x_1 - 3x_2 \leq 9, \\ & x_1, x_2 \geq 0. \end{aligned}$$

9. Use the Kuhn-Tucker conditions to solve the following non-linear programming problem:

$$\begin{aligned} \text{Optimize} \quad & Z = 2x_1 + 3x_2 - (x_1^2 + x_2^2 + x_3^2), \\ \text{subject to} \quad & x_1 + x_2 \leq 1, \\ & 2x_1 + 3x_2 \leq 6, \\ & x_1, x_2 \geq 0. \end{aligned}$$

Unit 20

Course Structure

- Separable programming, Piecewise linear approximation solution approach, Linear fractional programming.

20.1 Introduction

Separable programming is one of the indirect methods used to solve a non-linear programming problem. Indirect methods solve an NLP problem by dealing with one or more linear problems that are extracted from the original problem.

Separable programming is useful in solving those NLP problems in which the objective function and constraints are separable. Sometimes, functions that are not separable can be made separable by using simplified approximation. Such approximation reduces the single variable non-linear function into piece-wise linear functions, as shown in Fig 20.1.1.

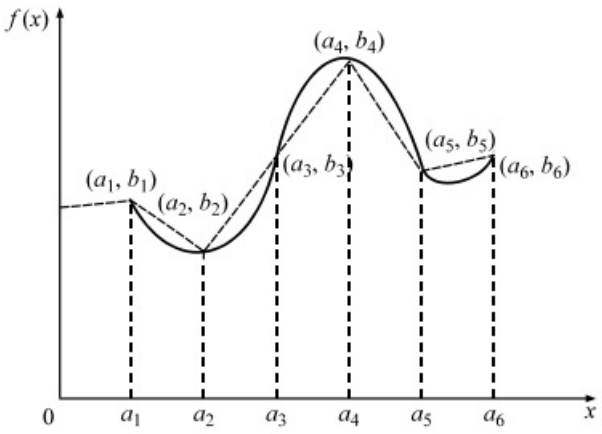


Figure 20.1.1: Linear approximation of a function

There is no particular method to determine the exact number of such piece-wise linear segments. Efforts should be made to have large number of linear functions (or segments) to reduce the chance of error in the approximation. However, such a number will increase the size of the problem and more computational time would obviously be required to obtain the optimal solution.

In this unit we shall discuss to obtain an approximate solution for any separable problem by linear approximation and the simplex method of linear programming.

20.2 Separable Functions

A function $f(x_1, x_2, \ldots, x_n)$ that can be expressed as the sum of n single-variable functions, $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ such that:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n)$$

is said to be a *separable function*. For example, the linear function

$$h(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \ldots + c_nx_n$$

(where c's are constants) is a separable function. But the function

$$h(x_1, x_2, \ldots, x_n) = x_1^2 + x_1 \cos(x_2 + x_3) + x_3 2^{x_2}$$

is not a separable function.

20.2.1 Reduction to separable form

A few non-linear functions are not directly separable, but can be separated by applying suitable substitutions. For example, in the function $y = x_1 \cdot x_2$, the non-separable term $x_1 \cdot x_2$ can be expressed in terms of two linear separable functions by taking log on both sides:

$$\log y = \log x_1 + \log x_2$$

The problem can then be stated as:

$$\begin{aligned} &\text{Maximize } Z = y \\ &\text{subject to the constraint} \\ &\quad \log y = \log x_1 + \log x_2 \end{aligned}$$

This problem is separable. Since logarithmic function is undefined for non-positive values, substitution assumes that both x_1 and x_2 are positive.

If x_1 and x_2 assume zero values (i.e. $x_1, x_2 \geq 0$), then two new variables u_1 and u_2 are defined as follows:

$$\begin{aligned} u_1 = x_1 + a_1 \quad &\text{and} \quad u_2 = x_2 + a_2 \\ \text{or } x_1 = u_1 - a_1 \quad &\text{and} \quad x_2 = u_2 - a_2 \end{aligned}$$

where a_1 and a_2 are positive constants. This implies that both u_1 and u_2 are strictly positive. Now:

$$x_1x_2 = (u_1 - a_1)(u_2 - a_2) = u_1u_2 - a_1u_2 - a_2u_1 + a_1a_2$$

Let $y = u_1 u_2$. The original problem is then stated as:

$$\text{Maximize } Z = y - a_1 u_2 - a_2 u_1 + a_1 a_2$$
$$\text{subject to the constraint}$$
$$\log y = \log u_1 + \log u_2; \;\; u_1 \geq a_1, \;\; u_2 \geq a_2$$

This problem is also separable. Few other functions that can also be expressed as separable functions using suitable substitution are:

$$e^{x_1+x_2}, \;\; x_1^{x_2}, \;\; (x_1)^{1/2}(x_2^2 + e^{x_2})^{-2} \;\; \text{etc.}$$

Definition 20.2.1. Separable programming problem: If the objective function of an NLP problem can be expressed as a linear combination of several different one-variable functions, of which some or all are non-linear, then such an NLP problem is called a *separable programming problem*.

Definition 20.2.2. Separable convex programming: It is the special case of separable programming in which separate functions are convex. Also, the non-linear function $f(x)$ is convex in case of minimization and concave in case of maximization.

For example, if $f(x)$ is the non-linear objective function, then for separable convex programming, it is expressed as:

$$f(x) = \sum_{j=1}^{n} f_j(x_j)$$

where all $f_j(x_j)$ are convex.

Illustration: Let $f_(x) = 9x_1^2 + 5x_2^2 - 5x_1 + 2x_2$. Then $f(x)$ is separated as:

$$f_1(x_1) = 9x_1^2 - 5x_1 \quad \text{and} \quad f_2(x_2) = 5x_2^2 + 2x_2$$

where both $f_1(x_1)$ and $f_2(x_2)$ are convex functions, such that $f(x) = f_1(x_1) + f_2(x_2)$.

20.3 Piece-Wise Linear Approximation of Non-linear Functions

In this section, we shall discuss piece-wise linear approximation method to reduce a separable convex (or concave) non-linear programming problem to a linear programming problem. Consider the following NLP problem:

$$\text{Optimize (Max or Min) } Z = \sum_{j=1}^{n} f_j(x_j)$$
$$\text{subject to the constraints}$$
$$\sum_{j=1}^{n} a_{ij} x_j = b_i; \;\; i = 1, 2, \ldots, m$$
$$\text{and } \; x_j \geq 0 \;\; \text{for all } \; j$$

where $f_j(x_j)$ is the j-th separable function to be approximated over a defined interval.

Define (a_k, b_k) for all $k = 1, 2, \ldots, K$ as the k-th breaking point joining a linear segment, which approximate the non-linear function $f(x)$, as shown in Fig. 20.1.1. Further define W_k as the non-negative weight

associated with the k-th breaking point such that $\sum_{k=1}^{K} W_K = 1$.

Let us impose an additional condition (if necessary) so that all W_k and $W_k + 1$ are equated with zero to determine the weighted average of breaking points. This means that W_k and $W_k + 1$ will represent the weighted average of breaking point (a_k, b_k) and (a_{k+1}, b_{k+1}) respectively. Thus $f(x)$ is approximated as follows:

$$f(x) = \sum_{k=1}^{K} b_k W_k; \quad x = \sum_{k=1}^{K} a_k W_k$$

This approximation is valid, provided the following conditions hold good:

$$\begin{aligned}
&0 \le W_1 \le y_1 \\
&0 \le W_2 \le y_1 + y_2 \\
&0 \le W3 \le y_2 + y_3 \\
&\vdots \\
&0 \le W_{k\check{}1} \le y_{k\check{}2} + y_{k\check{}1} \\
&0 \le W_k \le y_{k-1} \\
\text{and } &\sum_{k=1}^{K} W_k = 1; \quad \sum_{k=1}^{K-1} y_k = 1 \\
&y_k = 0 \text{ or } 1 \text{ for all } k.
\end{aligned}$$

The variables for approximation are now W_k and y_k. The last constraints implies that if $y_k = 1$, then all other $y_k = 0$. Consequently immediately preceding constraints ensure that $0 \le W_k \le y_k = 1$ and $0 \le W_{k+1} \le y_k = 1$. This means all other constraints should give $W_k \le 0$.

20.4 Mixed-Integer Approximation of Separable NLP Problem

The single-variable non-linear separable function $f(x)$, as defined earlier, can also be approximated by a piece-wise linear function, using mixed-integer programming. Let the number of breaking points for the j-th variable, x_j, be equal to K_j and g_{jk} be its k-th breaking value. Also, let w_{jk} be the weight associated with the k-th breaking point of j-th variable, x_j. Then the equivalent mixed integer programming problem is stated as:

$$\text{Optimize (Max or Min) } Z = \sum_{j=1}^{n} \sum_{k=1}^{K_j} f_j(a_{jk}) w_{jk}$$

subject to the constraints

$$\begin{aligned}
&\sum_{j=1}^{n} \sum_{k=1}^{K_j} g_{ij}(a_{jk}) w_{jk} \le b_i; \quad i = 1, 2, \ldots, m \\
&0 \le w_{j1} \le y_{j1} \\
&\vdots \\
&0 \le w_{jk} \le y_{jk-1} + y_{jk}; \quad k = 2, 3, \ldots, K_j - 1 \\
\text{and } &\sum_{k=1}^{K_j} w_{jk} = 1; \quad \sum_{k=1}^{K_j - 1} y_{jk} \\
&y_{jk} = 0 \text{ or } 1 \text{ for all } j \text{ and } k.
\end{aligned}$$

The approximation is valid only under following two conditions:

(i) For each j, no more than two w_{jk} should appear in the basis. That is, no more than two w_{jk} are positive for each j.

(ii) Two w_{jk} can be positive only if they are adjacent.

The simplex method can now be used to solve the above stated problem, along with additional constraints involving y_{jk} variables. Thus, the optimality criterion of the simplex method can be used to select the entering variable w_{jk} into the basis, only if it satisfies the above two conditions. Otherwise the variable w_{jk}, having the next best optimality indicator $(c_{jk} \breve{} z_{jk})$, is considered for entering the basis. The process is repeated until the optimality criterion is satisfied or until it is impossible to introduce a new w_{jk} without violating the restricted basis condition, whichever occurs first. The last simplex table provides the approximate optimal solution to the given problem.

Remark 20.4.1.
1. It is important to note that the restricted basis method yields only a local optimum, whereas mixed integer programming method guarantees a global optimum to the approximate problem.
2. The approximate solution obtained by using any of the two methods may not be feasible for the original NLP problem.
3. The solution space of approximate problem may have additional extreme points that do not exist in the solution space of the original problem. However, this depends on the degree of accuracy while obtaining linear approximation.

The Procedure

Step 1: Convert minimization objective function of the given NLP problem into that of maximization, with the usual method as discussed earlier.

Step 2: Examine whether the functions $f_j(x_j)$ and $g_{ij}(x_j)$ satisfy the concavity (convexity) conditions required for the maximization of NLP problem. If yes, then go to Step 3. Otherwise stop where $f(x)$ is to be approximated.

Step 3: Divide the interval $0 \leq x_j \leq t_j$ $(j = 1, 2, \ldots, n)$ into a number of breaking points a_{jk} $(k = 1, 2, \ldots, K_j)$ such that $a_{j1} = 0, a_{j1} < a_{j2} < \ldots < a_{jK_j}$.

Step 4: For each point a_{jk}, compute piece-wise linear approximation $f_j(x_j)$ and $g_{ij}(x_j)$, for all i and j.

Step 5: Write down piece-wise linear approximation of the given NLP problem obtained from Step 4.

Step 6: Solve the resulting LP problem using *two-phase simplex method* treating $w_{i1}(i = 1, 2, \ldots, m)$ as artificial variables. The coefficients associated with these variables are assumed to be zero. This assumption yields optimal simplex table of Phase I and hence would be considered as the initial simplex table for Phase II.

Step 7: Obtain optimum solution of the original NLP problem by using the relations:

$$x_j^* = \sum_{k=1}^{K_j} a_{jk} w_{jk}; \quad j = 1, 2, \ldots, n.$$

Example : Solve the following non-linear programming problem using separable programming algorithm.

$$\begin{aligned}&\text{Max} Z = 3x_1 + 2x_2\\&\text{subject to the constraints}\\&\quad g(x) = 4x_1^2 + x_2^2 \leq 16\\&\text{and}\quad x_1, x_2 \geq 0\end{aligned}$$

Solution : The objective function is already of maximization form. Consider the separable functions:

$$\begin{aligned}f_1(x_1) &= 3x_1; \quad & f_2(x_2) &= 2x_2\\g_{11}(x_1) &= 4x_1^2; \quad & g_{12}(x_2) &= x_2^2\end{aligned}$$

Since $f_1(x_1)$ and $f_2(x_2)$ are in linear form, can leave them in their present form. Further, it may be observed that these functions satisfy the concavity (convexity) conditions.

By inspection, constraints of the problem suggest the values of variables as: $x_1 \leq 2$ and $x_2 \leq 4$. Therefore, we take $t_1 = 4$ and $t_2 = 4$ as the upper limits for the variables x_1 and x_2 respectively. Thus, we divide the closed interval $[0, 4]$ into four subintervals of equal size for both x_1 and x_2. It is important to note that the number of subintervals for x_1 and x_2 should be the same, but they need not be equal in size.

To obtain the approximate LP problem for the given NLP problem, divide the interval $0 \leq x_j \leq 4$ into five breaking points a_{jk} $(j = 1, 2; k = 1, 2, 3, 4, 5)$ such that:

$$a_{j1} = 0,\ a_{j1} \leq a_{j2} \leq a_{j3} \leq a_{j4} \leq a_{j5} = 4$$

For each point a_{jk}, compute the piece-wise linear approximation for each of $f_j(x_j)$ and $g_{1j}(x_j);\ j = 1, 2$ as follows:

k	a_{jk}	$f_1(x_1 = a_{jk})$	$f_2(x_2 = a_{jk})$	$g_{11}(x_1 = a_{jk})$	$g_{12}(x_2 = a_{jk})$
1	0	0	0	0	0
2	1	3	2	4	1
3	2	6	4	16	4
4	3	9	6	36	9
5	4	12	8	64	16

Using this data, we have the following piece-wise linear approximation:

$$\begin{aligned}f_1(x_1) &= 0\ w_{11} + 3\ w_{12} + 6\ w_{13} + 9w_{14} + 12\ w_{15}\\f_2(x_2) &= 0\ w_{21} + 2\ w_{22} + 4\ w_{23} + 6w_{24} + 8\ w_{25}\\g_{11}(x_1) &= 0\ w_{11} + 4\ w_{12} + 16\ w_{13} + 36\ w_{14} + 64w_{15}\\g_{12}(x_2) &= 0\ w_{21} + 1\ w_{22} + 4\ w_{23} + 9\ w_{24} + 16w_{25}\end{aligned}$$

Using the data of Step 4, the approximating LP problem can now be stated as follows:

$$\text{Max}\, f(x) = (0\ w_{11} + 3\ w_{12} + 6\ w_{13} + 9\ w_{14} + 12\ w_{15}) + (0\ w_{21} + 2\ w_{22} + 4\ w_{23} + 6\ w_{24} + 8\ w_{25})$$

subject to the constraints

$$\begin{aligned}(0\,w_{11} + 4\,w_{12} + 16\,w_{13} + 36\,w_{14} + 64\,w_{15})\\ + (0\,w_{21} + w_{22} + 4\,w_{23} + 9\,w_{24} + 16\,w_{25}) \le 16\\ w_{11} + w_{12} + w_{13} + w_{14} + w_{15} = 1\\ w_{21} + w_{22} + w_{23} + w_{24} + w_{25} = 1\\ \text{and} \quad w_{jk} \ge 0 \quad \text{for } j \text{ and } k\end{aligned}$$

with the additional two restricted basis conditions:

(i) for each j no more than two w_{jk} are positive, and

(ii) if two w_{jk} are positive, they must correspond to adjacent points.

To solve the LP problem formulated in Step 5 by using simplex method, introduce the slack variables s_i (≥ 0) to convert inequality constraint into equation. For using Phase II of simplex method to solve the given LP problem, treat w_{11} and w_{21} as artificial variables whose coefficients in the objective function of reduced LP problem are zero. The initial simplex table for Phase II is shown in Table 20.1.

Table 20.1

		$c_j \rightarrow$	3	6	9	12	2	4	6	8	0	0	0
c_B	Basic Variables (B)	Solution Values $b(= x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1	w_{11}	w_{21}
0	s_1	16	4	16	36	64	1	4	9	16	1	0	0
0	w_{11}	1	1	1	1	1	0	0	0	0	0	1	0
0	w_{21}	1	0	0	0	0	1	1	1	(1)	0	0	1→
$f(x) = 0$		$c_j - z_j$	3	6	9	12	2	4	6	8 ↑	—	—	—

In Table 20.1, the entries in $c_j - z_j$ row indicate that variable w_{15} must enter into the new solution and variable s_1 should leave the current solution. But this does not satisfy the additional Conditions (i) and (ii). The next best variable to enter the basis is therefore w_{25} and variable w_{21} leaves the basis. The new solution, so obtained, is shown in Table 20.2.

Table 20.2

		$c_j \rightarrow$	3	6	9	12	2	4	6	8	0	0
c_B	Basic Variables (B)	Solution Values $b(= x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1	w_{11}
0	s_1	0	(4)	16	36	64	−15	−12	−7	0	1	0 →
0	w_{11}	1	1	1	1	1	0	0	0	0	0	1
8	w_{25}	1	0	0	0	0	1	1	1	1	0	0
$f(x) = 8$		$c_j - z_j$	3 ↑	6	9	12	6	4	2	—	—	—

In Table 20.2 out of eligible variables $w_{12}, w_{13}, w_{14}, w_{15}$ to enter the basis, we decide to enter variable w_{12} into the basis in view of the additional conditions. The new solution after introducing variable w_{12} into the basis and dropping variable s_1 from the basis is shown in Table 20.3.

Table 20.3

		$c_j \rightarrow$	3	6	9	12	2	4	6	8	0	0
c_B	Basic Variables B	Solution Values $b\,(= x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1	w_{11}
3	w_{12}	0	1	4	9	16	– 15/4	– 3	– 7/4	0	1/4	0
0	w_{11}	1	0	– 3	– 8	– 15	15/4	3	(7/4)	0	– 1/4	1 →
8	w_{25}	1	0	0	0	0	1	1	1	1	0	0
$f(x) = 8$		$c_j - z_j$	—	– 6	– 18	– 36	21/4	5 ↑	13/4	—	– 3/4	—

To get the next best solution, we need to introduce the variable w_{24} into the basis and drop variable w_{11} from the basis in the solution, as shown in Table 20.4. The new solution is shown in Table 20.4.

Table 20.4

		$c_j \rightarrow$	3	6	9	12	2	4	6	8	0
c_B	Basic Variables B	Solution Values $b(= x_B)$	w_{12}	w_{13}	w_{14}	w_{15}	w_{22}	w_{23}	w_{24}	w_{25}	s_1
3	w_{12}	1	1	1	1	1	0	0	0	0	0
6	w_{24}	4/7	0	– 12/7	– 32/7	– 60/7	15/7	12/7	1	0	– 1/7
8	w_{25}	3/7	0	12/7	32/7	60/7	– 8/7	– 5/7	0	1	1/7
$f(x) = 69/7$		$c_j - z_j$	—	– 3/7	– 22/7	– 67/7	– 12/7	– 4/7	—	—	– 9/14

Since all $c_j - z_j \leq 0$, the optimal solution has been arrived at. The optimal solution shown in Table 20.4 is:

$$w_{12} = 1, w_{24} = 4/7, w_{25} = 3/7 \quad \text{and} \quad f(x) = 69/7$$

The optimal solution to the original NLP problem can be obtained by using the formula:

$$x_j = \sum_{k=1}^{5} a_{jk} w_{jk}; \;\; j = 1, 2$$

This gives

$$\begin{aligned} x_1 &= a_{11}w_{11} + a_{12}w_{12} + a_{13}w_{13} + a_{14}w_{14} + a_{15}w_{15} \\ &= 0(0) + 1(1) + 2(0) + 3(0) + 4(0) = 1 \\ x_2 &= a_{21}w_{21} + a_{22}w_{22} + a_{23}w_{23} + a_{24}w_{24} + a_{25}w_{25} \\ &= 0(0) + 1(0) + 2(0) + 3(4/7) + 4(3/7) = 24/7 \end{aligned}$$

Hence, the optimal solution to the given NLP problem is:

$$x_1 = 1, \;\; x_2 = 24/7 \;\; \text{and Max} \;\; f(x) = 3 + 2(24/7) = 69/7.$$

Example 20.4.2. Use separable programming algorithm to solve the non-linear programming problem:

$$\begin{aligned} &\text{Max } Z = x_1 + x_2^2 \\ &\text{subject to the constraints} \\ &\quad 3x_1 + 2x_2^2 \leq 9 \\ &\quad x_1, x_2 \geq 0. \end{aligned}$$

Solution. The objective function is already of the maximization form. Consider the following separable functions:

$$\begin{aligned} f_1(x) = x_1; \quad & f_2(x_2) = x_2^2 \\ g_{11}(x_1) = 3x_1; \quad & g_{12}(x_2) = 2x_2^2 \end{aligned}$$

Since $f_1(x_1)$ and $g_{11}(x_1)$ are in linear form, therefore these functions are left in their present form. Further, it may be observed that these functions satisfy concavity (convexity) conditions.

The constraints of the problem suggest the value of variables as: $x_1 \leq 3$ and $x_2 \leq \sqrt{9/2} = 2.13$. Therefore, we consider $t_1 = 3$ and $t_2 = 3$ as the upper limits for the variables x_1 and x_2 respectively. Thus, we divide the closed interval $[0, 3]$ into four breaking points of equal intervals for both x_1 and x_2. That is, the four breaking points $a_{jk}(j = 1, 2; k = 1, 2, 3, 4)$ will be $a_{j1} = 0, a_{j1} < a_{j2} < a_{j3} < a_{j4} = 3$.

We consider non-linear functions $f_2(x_2)$ and $g_{12}(x_2)$ and assume that there are four breaking points $(k = 4)$. Since the value of $x_2 \leq 3$, therefore the piece-wise linear approximations for $f_2(x_2)$ and $g_{12}(x_2)$ are computed as follows:

k	a_{jk}	$f_2(x_2 = a_{jk})$	$g_{12}(x_2 = a_{jk})$
1	0	0	0
2	1	1	2
3	2	16	8
4	3	81	18

This gives

$$\begin{aligned} f_2(x_2) &= w_{21} f_2(a_{21}) + w_{22} f_2(a_{22}) + w_{23} f_2(a_{23}) + w_{24} f_2(a_{24}) \\ &= w_{21}(0) + w_{22}(1) + w_{23}(16) + w_{24}(81) = w_{22} + 16w_{23} + 81w_{24} \\ g_{12}(x_2) &= w_{21} g_{12}(a_{21}) + w_{22} g_{12}(a_{22}) + w_{23} f_{12}(a_{23}) + w_{24} g_{12}(a_{24}) \\ &= w_{21}(0) + w_{22}(2) + w_{23}(8) + w_{24}(18) = 2w_{22} + 8w_{23} + 18w_{24} \end{aligned}$$

Using the above data, the approximating LP problem can now be stated as follows:

$$\begin{aligned} &\max \;\; f(x) = x_1 + w_{22} + 16w_{23} + 81w_{24} \\ &\text{subject to the constraints} \\ &\quad 2x_1 + 2w_{22} + 8w_{23} + 18w_{24} \leq 9 \\ &\quad w_{21} + w_{22} + w_{23} + w_{24} = 1 \\ &\quad \text{and} \;\; x_1, \; w_{21}, \; w_{22}, \; w_{23}, \; w_{24} \geq 0 \end{aligned}$$

with the two additional restricted basis conditions:

(i) for each j, no more than two w_{jk} are positive, and

(ii) if two w_{jk} are positive, they must correspond to adjacent points.

Treating w_{21} as the artificial variable (because coefficient in the objective function of reduced LP problem is zero) the given LP problem can be solved by using Two-Phase simplex method. The initial simplex table for Phase II is given in Table 20.5.

Table 20.5

		$c_j \rightarrow$	1	1	16	81	0	0
c_B	Basic Variables B	Solution Values $b(= x_B)$	x_1 Values	w_{22}	w_{23}	w_{24}	s_1	w_{21}
0	s_1	9	3	2	8	18	1	0
0	w_{21}	1	0	1	(1)	1	0	1 →
$Z = 0$		$c_j - z_j$	1	1	16 ↑	81	—	—

From $c_j - z_j$ row of Table 20.5, it appears that the variable w_{24} should enter the basis. Since w_{21} is artificial basic variable, it must be dropped before w_{24} enters the basis (restricted basis condition). By the feasibility conditions (minimum ratio rule), s_1 is the leaving variable. This means that w_{24} cannot enter the basis. Thus, we consider the next best entering variable w_{23} $[c_3 - z_3 = 16(< 81)]$. Again the artificial variable w_{21} must be dropped first. From the feasibility condition, w_{21} is the leaving variable. The new solution is shown in Table 20.6.

In Table 20.6, $c_j - z_j$ row values indicate that w_{24} is the entering variable. Because w_{23} is already in the

Table 20.6

		$c_j \rightarrow$	1	1	16	81	0
c_B	Basic Variables B	Solution Values $b\ (= x_B)$	x_1	w_{22}	w_{23}	w_{24}	s_1
0	s_1	1	3	− 6	0	(10)	1 →
16	w_{23}	1	0	1	1	1	0
$Z = 16$		$c_j - z_j$	1	− 15	—	65 ↑	—

basis, w_{24} is an admissible entering variable. From the feasibility condition, s_1 is the leaving variable. The new solution, so obtained, is shown in Table 20.7.

Table 20.7 shows that w_{22} should enter into the basis. But this is not possible because w_{24} cannot be dropped from the current solution. Thus, the procedure terminates at this point and the given solution is the optimal solution for the approximate LP problem $w_{23} = 9/10, w_{24} = 1/10$ and $\text{Max}\, f(x) = 22.5$.

Table 20.7

		$c_j \rightarrow$	1	1	16	81	0
c_B	Basic Variables B	Solution Values $b(= x_B)$	x_1	w_{22}	w_{23}	w_{24}	s_1
81	w_{24}	1/10	3/10	– 6/10	0	1	1/10
16	w_{23}	9/10	–3/10	16/10	1	0	– 1/10
Z = 22.5		$c_j - z_j$	–37/2	24	—	—	– 13/2

The optimal solution of the original non-linear programming problem in terms of x_1 and x_2 is obtained by using the relationship:

$$x_j = \sum_{k=1}^{4} a_{jk} w_{jk}; \quad j = 1, 2.$$

Therefore

$$\begin{aligned} x_2 &= a_{21}\, w_{21} + a_{22}\, w_{22} + a_{23}\, w_{23} + a_{24}\, w_{24} \\ &= 0(0) + 1(0) + 2(9/10) + 3(1/10) = 2.1. \\ x_1 &= 0 \text{ and } \operatorname{Max} f(x) = 22.5 \end{aligned}$$

Exercise 20.4.3. 1. What do you mean by separable and/or nonlinear convex programming? How will you solve the separable non-linear programming problem:

$$\text{Minimize } Z = \sum_{j=1}^{n} f_{0j}(x_j)$$

subject to the constraints

$$\sum_{j=1}^{n} f_{0j}(x_j) \geq b_i; \quad i = 1, 2, \ldots, m$$

2. Show that if $f_{0j}(x_j)$ is strictly convex and $f_{ij}(x_j)$ is concave for $i = 1, 2, \ldots, m$, then we can discard the additional restriction in the approximated separable non-linear programming problem of exercise 1 and solve the resulting LP problem to find an approximate solution to the given problem.

3. Show that the non-linear non-convex programming problem:

$$\text{Minimize } Z = a_0 + b_{01} x_1 + \left(\sum_{j=2}^{5} b_{0j} x_j \right) x_1$$

subject to the constraints

$$\begin{aligned} &0 \leq a_{i1} x_1 + \left(\sum j = 2^5 a_{ij} x_j \right) x_1 \leq b_i; \;\; i = 1, 2, \ldots, 5 \\ &l_i \leq x_i \leq u_i; \;\; i = 1, 2, \ldots, 5. \end{aligned}$$

can be transformed into a concave LP problem by setting

$$y_i = x_i x_1 \;\; (i = 1, 2, \ldots, 5) \quad \text{and} \quad y_1 = x_1$$

where $a_0,\ b_{0j},\ a_{ij},\ b_i,\ l_i$ and u_i are real constants.

4. Solve the following non-linear programming problem:

$$\begin{aligned}&\text{Max } Z = (x_1+1)^2 + (x_2-2)^2\\&\text{subject to the constraints}\\&\quad x_1 - 2 \le 0; \quad x_2 - 1 \le 0\\&\quad x_1, x_2 \ge 0.\end{aligned}$$

5. Solve the following non-linear programming problem:

$$\begin{aligned}&\text{Min } Z = x_1^2 + x_2^2 + 5\\&\text{subject to the constraints}\\&\quad 3x_1^4 + x_2 \le 243; \quad x_1 + 2x_2^2 \le 32\\&\quad x_1, x_2 \ge 0.\end{aligned}$$

References

1. Introduction to Classical Mechanics - *R. G. Takwale and P. S. Puranik*
2. Classical Mechanics - *H. Goldstein.*
3. Classical Mechanics - *N. C. Rana and P.S. Joag.*
4. Fundamentals of Abstract Algebra - *D. S. Malik, John M. Mordeson, and M. K. Sen.*
5. Abstract Algebra - *David S. Dummit and Richard M. Foote*
6. Contemporary Abstract Algebra - *Joseph A Gallian*
7. Abstract Algebra - *I. N. Herstein.*
8. Operations Research – *K. Swarup, P. K. Gupta and Man Mohan.*
9. Operations Research: Theory and Applications – *J. K. Sharma.*
10. Nonlinear and Dynamic Programming – *G. Hadley.*

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER II

SELF LEARNING MATERIAL

PAPER : COR 2.3
(Pure & Applied Streams)

Numerical Analysis (Theory)

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

Content Writer

Block - I : Numerical Analysis (Theory)	Dr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani

July, 2022

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 Message

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, 2020 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.) Manas Kumar Sanyal, 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 well 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 coordinated 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 written and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Board of Studies Members of Department of Mathematics, Directorate of Open and Distance Learning (DODL), University of Kalyani

Sl No.	Name & Designation	Role
1	Dr. Animesh Biswas, Professor & Head, Dept. of Mathematics, KU	Chairperson
2	Dr. Pulak Sahoo, Professor, Dept. of Mathematics, KU	Member
3	Dr. Sahidul Islam, Assistant Professor, Dept. of Mathematics, KU	Member
4	Dr. Sushanta Kumar Mohanta, Professor, Dept. of Mathematics, West Bengal State University	External Nominated Member
5	Dr. Biswajit Mallick, Assistant Professor (Cont.), Department of Mathematics, DODL, KU	Member
6	Ms. Audrija Choudhury, Assistant Professor (Cont), Department of Mathematics, DODL, KU	Member
7	Director, DODL, KU	Convener

Core Paper

PURE AND APPLIED STREAMS

COR 2.3
Marks : 50 (SEE : 40; IA : 10); Credit : 4

Numerical Analysis (Theory) (Marks : 50 (SEE: 40; IA: 10))

Syllabus

Block I

- **Unit 1:** Errors: Floating-point approximation of a number, Loss of significance and error propagation, Stability in numerical computation.
- **Unit 2:** Interpolation: Hermite's and spline interpolation. Interpolation by iteration –Aitken's and Neville's schemes.
- **Unit 3:** Approximation of Function: Least square approximation. Weighted least square approximation. Orthogonal polynomials,
- **Unit 4:** Gram –Schmidt orthogonalisation process, Chebysev polynomials, Mini-max polynomial approximation.
- **Unit 5:** Numerical Integration: Gaussian quadrature formula and its existence. Euler-MacLaurin formula
- **Unit 6:** Gregory-Newton quadrature formula. Romberg integration.
- **Unit 7:** Systems of Linear Algebraic Equations: Direct methods, Factorization method.
- **Unit 8:** Eigenvalue and Eigenvector Problems: Direct methods, Iterative method –Power method.
- **Unit 9:** Nonlinear Equations: Fixed point iteration method, convergence and error estimation.
- **Unit 10:** Modified Newton-Raphson method, Muller's method, Inverse interpolation method, error estimations and convergence analysis.
- **Unit 11:** Ordinary Differential Equations: Initial value problems–Picard's successive approximation method, error estimation.
- **Unit 12:** Single-step methods –Euler's method and Runge-Kutta method, error estimations and convergence analysis
- **Unit 13:** Multi-step method –Milne's predictor-corrector method, error estimation and convergence analysis.
- **Unit 14:** Partial Differential Equations: Finite difference methods for Elliptic and Parabolic differential equations.

Contents

Unit 1

Course Structure

- Errors: Floating-point approximation of a number, Loss of significance and error propagation, Stability in numerical computation.

1.1 Introduction

In any numerical computation, we come across following types of errors, viz.,

1. Round-off Error
2. Floating Point Arithmetic and Propagated Error,
3. Truncation Error,
4. Loss of Significance: Condition and Stability

There are several potential sources of errors in numerical computation. But, round-off and truncation errors can occur in any numerical computation.

1.2 Round-off Error

During the implementation of a numerical algorithm with computing devices mainly calculator and computer, we have to work with a finite number of digits in representing a number. The number of digits depends on the word length of the computing device and software. The scientific calculations are carried out in floating point arithmetic. It is necessary to have knowledge of floating point representations of numbers and the basic arithmetic operations performed by the computer $(+, -, *, /)$ in these representations.

Floating Point Representation of Numbers

To understand the major sources of error during the implementation of numerical algorithms, it is necessary to discuss how the computer stores the numbers.

An m-digits floating point number in the base β is of the following form

$$x = \pm(.d_1 d_2 d_3 \cdots d_m)_\beta \beta^n$$

where $(.d_1d_2d_3\cdots d_m)_\beta$ is called as a mantissa and the integer n is called the exponent. A nonzero number is said to be normalized if $d_1 \neq 0$.

All the real numbers are stored in normalized form in the computer to avoid wastage of computer memory on storing useless non-significant zeroes. For example, 0.002345 can be represented in a wasteful manner as $(0.002345)10^0$ which is wasting two important decimal points. However, the normalized form is $(0.2345)10^{-2}$, which eliminates these useless zeroes; also known as spurious zeroes.

If we want to enter the number 234.1205, then this number stored in the computer in normalized form, i.e., $(0.2341205)10^3$. Similarly, the number 0.00008671213 stored in the computer in normalized form $(0.8671213)10^{-4}$.

The digits used in mantissa to express a number are called as significant digits or significant figures. More precisely, *digits in the normalized form mantissa of a number are significant digits.*

a) All non-zero digits are significant. For examples, the numbers $3.1416,\ 4.7894$ and 34.211 have five significant digits each.

b) All zeroes between non-zero digits are significant. For examples, the numbers 3.0156 and 7.5608 have five significant digits each.

c) Trailing zeroes following a decimal point are significant. So, the numbers 3.5070 and 76.500 have five significant digits each.

d) Zeroes between the decimal point and preceding a non-zero digit are not significant.i.e., the numbers 0.0023401 and 0.00023401 have five significant digits each.

e) Trailing zeroes are significant if the decimal point is not present, i.e., the numbers 45067000 and 45000 have eight and five significant digits, respectively.

To compute the significant digits in a number, simply convert the number in the normalized form and then compute the significant digits.

Rounding and Chopping

Rounding and chopping are two commonly used ways of converting a given real number x into its m-digits floating point representation $fl(x)$. In the case of chopping, the number x is retained up to m-digits, and remaining digits are simply chopped off. For example, consider 6-digits floating point representation, then

$$\begin{aligned} x_1 = \tfrac{2}{3} \qquad & fl\,(x_1) = 0.666666 \\ x_2 = 3456789 \qquad & fl\,(x_2) = (.345678)10^7 \\ x_3 = -0.0011223344 \qquad & fl\,(x_3) = -(.112233)10^{-2} \end{aligned}$$

In rounding, the normalized floating point number $fl(x)$ is chosen such that it is nearest to the number x. In the case of a tie, some special rules such as symmetric rounding can be used. Rules to round off a number to m significant figures are as follows

i) Discard all digits to the right of m-th digit.

ii) If the last discarded number is

   a) less than half of base β in the $(m+1)$-th place, leave the m-th digit unchanged;

b) greater than half of base β in the $(m+1)$-th place, increase the m-th digit by unity;

c) exactly half of base β in the $(m+1)$-th place, increase the m-th digit by unity if it is odd, otherwise leave the m-th digit unchanged. It is known as symmetric rounding around even number. Similarly, we can have symmetric rounding about odd number.

1.3 Floating Point Arithmetic and Error Propagation

We have discussed the errors in number representations. These errors further propagate while performing basic arithmetic operations using a computer. The result of an arithmetic operation is usually not accurate to the same length as the numbers used for the operations. The floating point numbers are first converted into the normalized forms as soon as they enter in the computer.

Here we will explain the arithmetic operations with 6-significant digits numbers. For example, let us take numbers $x = 123.456$ and $y = 12.3456$ with six significant digits. The various arithmetic operations $(+, -, *, /)$ on these two numbers are as follows

$$\begin{aligned}
x+y &= (.123456)10^3 + (.123456)10^2 \quad \text{(Normalized form)} \\
&= (.123456)10^3 + (.012346)10^3 \quad \text{(Equal exponent using symmetric rounding)} \\
&= (.135802)10^3 \\
x-y &= (.123456)10^3 - (.123456)10^2 \\
&= (.123456)10^3 - (.012346)10^3 \quad \text{(Equal exponent using symmetric rounding)} \\
&= (.111110)10^3 \\
x*y &= (.123456)10^3 * (.123456)10^2 \\
&= (.123456) * (.123456)10^{3+2} \quad \text{(Add the exponents)} \\
&= (.015241)10^5 \\
&= (.152410)10^4 \\
x/y &= (.123456)10^3/(.123456)10^2 \\
&= (.123456)/(.123456)10^{3-2} \quad \text{(Subtract the exponents)} \\
&= (1.00000)10^1 \\
&= (0.100000)10^2
\end{aligned}$$

Note 1.3.1. If two floating point numbers are added or subtracted, first they are converted into the numbers with equal exponents. The results in various arithmetic operations are not correct up to six significant digits due to rounding errors.

It is worth mentioning here that the result of subtraction of two nearly equal numbers leads to a very serious problem, i.e., loss of significant digits. For example, consider six significant digits numbers $x = 123.456$ and $y = 123.432$, then

$$\begin{aligned}
x-y &= (.123456)10^3 - (.123432)10^3 \quad \text{(Normalized form)} \\
&= (.000024)10^3 \quad \text{(Result containing only two significant digits, four non-significant zeroes are appended)}
\end{aligned}$$

This subtraction of two nearly equal numbers is called as subtractive cancellation or loss of significance. It is a classical example of computer handling mathematics can create a numerical problem.

1.3.1 Propagated Error in Arithmetic Operations

Consider any two numbers x_1 and x_2. Let the errors in the numbers x_1 and x_2 be δx_1 and δx_2, respectively. Then errors in the addition, subtraction, multiplication, and division of these two numbers are as follows

i) **Addition**: Let $X = x_1 + x_2$ and the error in X is δX. Therefore,

$$\begin{aligned} X + \delta X &= x_1 + \delta x_1 + x_2 + \delta x_2 \\ \Rightarrow \delta X &= \delta x_1 + \delta x_2 \end{aligned}$$

$$\text{Absolute Error} = |\delta X| \leq |\delta x_1| + |\delta x_2|; \quad \text{Relative Error} = \frac{|\delta X|}{|X|} \leq \frac{|\delta x_1|}{X} + \frac{|\delta x_2|}{X} \tag{1.3.1}$$

ii) **Subtraction**: Similarly, the error in subtraction $X = x_1 - x_2$ is $\delta X = \delta x_1 - \delta x_2$.

$$\text{Absolute Error} = |\delta X| \leq |\delta x_1| + |\delta x_2|; \quad \text{Relative Error} = \frac{|\delta X|}{|X|} \leq \frac{|\delta x_1|}{X} + \frac{|\delta x_2|}{X} \tag{1.3.2}$$

iii) **Multiplication**: Let $X = x_1 x_2$, then

$$X + \delta X = (x_1 + \delta x_1)(x_2 + \delta x_2) = x_1 x_2 + x_1 \delta x_2 + x_2 \delta x_1 + \delta x_1 \delta x_2$$

Neglecting second order term $(\delta x_1 \delta x_2)$, the error in the multiplication of two numbers becomes $\delta X = x_2 \delta x_1 + x_1 \delta x_2$.

$$\text{Absolute Error} = |\delta X| \leq |x_2 \delta x_1| + |x_1 \delta x_2|; \quad \text{Relative Error} = \frac{|\delta X|}{|X|} \leq \frac{|\delta x_1|}{x_1} + \frac{|\delta x_2|}{x_2} \tag{1.3.3}$$

iv) **Division**: Let $X = \dfrac{x_1}{x_2}$, then

$$X + \delta X = \frac{x_1 + \delta x_1}{x_2 + \delta x_2} = \left(\frac{x_1 + \delta x_1}{x_2 + \delta x_2}\right)\left(\frac{x_2 - \delta x_2}{x_2 - \delta x_2}\right) = \frac{x_1 x_2 + x_2 \delta x_1 - x_1 \delta x_2 - \delta x_1 \delta x_2}{x_2^2 - \delta x_2^2}$$

On neglecting the second order terms ($\delta x_1 \delta x_2$ and δx_2^2), the error is given by $\delta X = \dfrac{x_2 \delta x_1 - x_1 \delta x_2}{x_2^2}$.

$$\text{Absolute Error} = |\delta X| \leq \frac{|\delta x_1|}{|x_2|} + \frac{|x_1 \delta x_2|}{x_2^2}; \quad \text{Relative Error} = \frac{|\delta X|}{|X|} \leq \frac{|\delta x_1|}{x_1} + \frac{|\delta x_2|}{x_2} \tag{1.3.4}$$

Example 1.3.2. Calculate the absolute and relative errors in the expression $a + \dfrac{5b}{c} - 3bc$, if the measurements of $a = 3.5435, b = .2588$ and $c = 1.0150$ are possibly correct up to four decimal points.

Solution. Let $x = a + \dfrac{5b}{c} - 3bc = A + 5B - 3C$, where $A = a, B = \frac{b}{c}$ and $C = bc$.

Value of $x = a + \dfrac{5b}{c} - 3bc = 4.03033$
Error in a, b and c is $\delta a = \delta b = \delta c = .00005$
Absolute error in $A = |\delta A| = .00005$
Absolute error in $B = |\delta B| = \dfrac{|c\delta b| + |b\delta c|}{c^2} = \dfrac{(1.015 + 0.2588) \times .00005}{(1.015)^2} = .00006182$
Absolute error in $C = |\delta C| = |c\delta b| + |b\delta c| = (1.015 + 0.2588) \times .00005 = .00006369$
Absolute error in $x = |\delta x| \leq |\delta A| + 5|\delta B| + 3|\delta C| = .00005 + 5(.00006182) + 3(.00006369) = .0005502$
Relative error in $x = \left|\dfrac{\delta x}{x}\right| = \dfrac{.0005502}{4.03033} = .0001365$
Percentage error in $x = 0.01365\%$

1.3.2 Error Propagation in Function of Single Variable

Let us consider a function $f(x)$ of a single variable, x. Assume that the variable x has some error and its approximate value is $\tilde{x}$. The effect of error in the value of x on the value of function $f(x)$ is given by

$$\Delta f(x) = |f(x) - f(\tilde{x})|$$

Evaluating $\Delta f(x)$ is difficult as the exact value of x is unknown and hence exact $f(x)$ is unknown. But if $\tilde{x}$ is close to x and the function $f(x)$ is infinitely differentiable in some interval containing the points $\tilde{x}$ and x, then Taylor series can be employed as follows

$$f(x) = f(\tilde{x}) + (x - \tilde{x})f'(\tilde{x}) + \frac{(x-\tilde{x})^2}{2!}f''(\tilde{x}) + \cdots$$

Since the difference $(x - \tilde{x})$ is very small, hence neglecting the second and higher order terms of $(x - \tilde{x})$ will give following relation

$$f(x) - f(\tilde{x}) \approx (x - \tilde{x})f'(x) \Rightarrow |\Delta f(x)| \approx |x - \tilde{x}||f'(\tilde{x})| \approx \Delta x|f'(\tilde{x})| \tag{1.3.5}$$

where $f(x) = f(x) - f(\tilde{x})$ represents the estimated error in the function value and $x = x - \tilde{x}$ is the estimated error of x.

1.3.3 Error Propagation in Function of More than One Variable

General Error Formula: The approach above can be generalized to the function of more than one independent variable. Let $y = f(x_1, x_2, \ldots, x_n)$ be a function of n-independent variables $x_1, x_2, \ldots, x_n$. Let $\delta x_1, \delta x_2, \ldots, \delta x_n$ be the errors in calculating the variables $x_1, x_2, \ldots, x_n$, respectively. Let error in y be δy, i.e.,

$$y + \delta y = f(x_1 + \delta x_1, x_2 + \delta x_2, \ldots, x_n + \delta x_n)$$

When the required partial derivatives exist, then Taylor's series expansion is given by

$$\begin{aligned} y + \delta y = f(x_1, x_2, \ldots, x_n) + \left(\frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\delta x_n\right) \\ +\text{terms involving second and higher powers of } \delta x_1, \delta x_2, \ldots, \delta x_n \end{aligned} \tag{1.3.6}$$

The errors in the numbers $x_1, x_2, \ldots, x_n$ are small enough to neglect the second and higher degree terms of the numbers $\delta x_1, \delta x_2, \ldots, \delta x_n$. We can obtain the following result from Eq. (1.3.6)

$$\delta y \approx \frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\delta x_n \tag{1.3.7}$$

Equation (1.3.7) is known as the general error formula. Since the error term may be of any sign, $(+)$ve or $(-)$ve, we can take absolute values of the terms in the expression.

$$|\delta y| \approx \left|\frac{\partial f}{\partial x_1}\right||\delta x_1| + \left|\frac{\partial f}{\partial x_2}\right||\delta x_2| + \cdots + \left|\frac{\partial f}{\partial x_n}\right||\delta x_n|$$

Example 1.3.3. The radius r and height h of a right circular cylinder are measured as .25 m and 2.4 m, respectively, with a maximum error of 5%. Compute the resulting percentage error in the volume of the cylinder. Assume the value of π is exact for calculation.

Solution. The value of π is exact for calculation, so the volume $V = \pi r^2 h$ is dependent only on radius r and height h of the cylinder i.e., $V = V(r, h)$. Therefore, the error $\delta V(r, h)$ in the volume is given by

$$\delta V(r,h) = \frac{\partial V}{\partial r}\delta r + \frac{\partial V}{\partial h}\delta h = (2\pi rh)\delta r + (\pi r^2)\delta h$$

The radius r and height h of the cylinder are measured with a maximum error of 5% i.e.,

$$\frac{\delta r}{r} = \frac{\delta h}{h} = 0.05$$

The relative error in volume $V(r, h)$ is given by

$$\text{R.E.} = \frac{\delta V(r,h)}{V} = \frac{1}{\pi r^2 h}\left((2\pi rh)\delta r + (\pi r^2)\delta h\right) = 2\frac{\delta r}{r} + \frac{\delta h}{h} = 2(0.05) + 0.05 = 0.15$$

Percentage error in the volume of cylinder = R.E. $\times 100 = 15\%$.

1.4 Truncation Error

An infinite power series (generally Taylor series) represents the local behavior of a given function $f(x)$ near a given point $x = a$. Approximation of an infinite power series with its finite number of terms, while neglecting remaining terms, leads to the *truncation error*. If we approximate the power series by the n-th order polynomial, then truncation error is of order $n + 1$.

Taylor series for the function $f(x)$ at the point $x = a$ is given by

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{(n)!}f^{(n)}(a) + \cdots$$

$$\text{Or, } f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^n}{(n)!}f^{(n)}(a) + R_n(x)$$

where $R_n(x) = \dfrac{(x-a)^{n+1}}{(n+1)!}f^{(n+1)}(\xi)$ for some ξ between a and x.

On replacing $x = a + h$, we get following form of the Taylor series

$$f(a+h) = f(a) + (h)f'(a) + \frac{(h)^2}{2!}f''(a) + \cdots\cdots + \frac{(h)^n}{(n)!}f^{(n)}(a) + R_n(x)$$

where $R_n(x) = \dfrac{(h)^{n+1}}{(n+1)!}f^{(n+1)}(\xi); a < \xi < a + h$. For a convergent series, $R_n(x) \to 0$ as $n \to \infty$. Since it is not possible to compute an infinite number of terms, we approximate the function $f(x)$ by first n-terms, and neglecting higher order terms. Then the error is given by remainder term $R_n(x)$. The exact value of ξ is not known, so the value of ξ is such that the error term considered is maximum.

Example 1.4.1. Calculate the number of terms required in Taylor series approximation of $\sin(x)$ to compute the value of $\sin(\pi/12)$ correct up to 4-decimal places.

Solution. Using Taylor series of $\sin(x)$ at point $x = 0$, we have

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{(x)^{2n-1}}{(2n-1)!}(-1)^{n-1} + R_{2n-1}(x)$$

If we retain only first $2n-1$ terms in this expression, then the error term is given by

$$R_{2n-1}(x) = \frac{(x)^{2n}}{(2n)!} f^{(2n)}(\xi); \quad 0 < \xi < x \quad \text{at} \quad x = \frac{\pi}{12} = 0.2618$$

The maximum value of $f^{(2n)}(\xi)$ is 1. The error term must be less than .00005 for 4-decimal points accuracy

$$R_{2n-1}(x) = \frac{(0.2618)^{2n}}{(2n)!} < .00005 \Rightarrow 2n \geq 5$$

Hence, 4-decimal points accuracy can be achieved by computing more than five terms of Taylor series.

1.5 Loss of Significance: Condition and Stability

In this section, we will study the two related concepts of condition and stability for function and process, respectively. The condition is used to describe the sensitivity of the function and stability is used to describe the sensitivity of the process.

Condition: The sensitivity of the function $f(x)$ with the change in the argument x is described by the condition number (CN). It is a relative change in the function $f(x)$ for per unit relative change in x. CN of the function $f(x)$ at any point x is given by

$$\text{CN} = \frac{\left|\frac{f(x)-f(\tilde{x})}{f(x)}\right|}{\left|\frac{x-\tilde{x}}{x}\right|} = \left|\frac{f(x)-f(\tilde{x})}{x-\tilde{x}}\right| \left|\frac{x}{f(x)}\right|$$

For small change in x, Lagrange mean value theorem gives

$$\frac{f(x)-f(x)}{x-\tilde{x}} \approx f'(x)$$

So, CN is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| \tag{1.5.1}$$

If $\text{CN} \leq 1$, then the function $f(x)$ is said to be well-conditioned. Otherwise, it is said to be ill-conditioned. The function with large CN is more ill-conditioned as compared to the function with small CN.

Note 1.5.1. Let us consider a mathematical model of any system, in which variable x gives input, and output is the function $f(x)$. If a small relative change in x (input) produces a large relative change in output $f(x)$, then the system is said to be a sensitive system as fluctuation in input may break the system. Mathematically, if CN is large, then the function is more sensitive to changes and function is ill-conditioned.

Example 1.5.2. Find the CNs of the functions $f(x) = x$ and x^3.

Solution. Using (1.5.1), we have

$$\text{CN of the function } \sqrt{x} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(\frac{1}{2}x^{\frac{-1}{2}}\right)}{\sqrt{x}}\right| = \frac{1}{2}$$

$$\text{CN of the function } x^3 = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(3x^2\right)}{x^3}\right| = 3$$

CN of the function $\sqrt{x}$ is less than 1 , so the function $\sqrt{x}$ is well conditioned. The function x^3 is an ill-conditioned function as $\text{CN} > 1$.

Example 1.5.3. Check the condition of the function $f(x) = \dfrac{1}{1-2x+x^2}$ at $x = 1.01$.

Solution.

$$f(x) = \frac{1}{1-2x+x^2} = \frac{1}{(1-x)^2}$$

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right|_{x=1.01} = \left|\frac{x\left(\frac{-2}{(1-x)^3}\right)}{\left(\frac{1}{(1-x)^2}\right)}\right|_{x=1.01} = 202$$

The function $f(x) = \dfrac{1}{1-2x+x^2}$ at $x = 1.01$ is highly ill-conditioned function. The function has a singular point $x = 1$, so near this point, there are sharp changes in the function value, which make the function highly ill-conditioned.

Example 1.5.4. Find the CN of the function $f(x) = \sqrt{x+1} - \sqrt{x}$ at point $x = 11111$.

Solution.

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(\frac{1}{2\sqrt{x+1}} - \frac{1}{2\sqrt{x}}\right)}{\sqrt{x+1}-\sqrt{x}}\right|_{x=11111} \approx \frac{1}{2}$$

Example 1.5.5. Compute the function $f(x) = \sqrt{x+1} - \sqrt{x} = \dfrac{1}{\sqrt{x+1}+\sqrt{x}}$ by using both the formulae at point $x = 11111$. Use six significant digits floating point rounding arithmetic.

Solution. We have two formulas $f(x) = \sqrt{x+1} - \sqrt{x}$ and $f(x) = \dfrac{1}{\sqrt{x+1}+\sqrt{x}}$ to compute the function $f(x)$ at point $x = 11111$. We will use both the formulas with six significant digits arithmetic, and see that both the processes will produce different results for the same function.

$$\text{Process-I: } f(x) = \sqrt{x+1} - \sqrt{x}: \quad f(11111) = \sqrt{11112} - \sqrt{11111} = 105.413 - 105.409 = .004$$

$$\text{Process-II: } f(x) = \frac{1}{\sqrt{x+1}+\sqrt{x}}: \quad f(11111) = \frac{1}{\sqrt{11112}+\sqrt{11111}} = \frac{1}{105.413+105.409}$$

$$= \frac{1}{210.822} = 0.00474334$$

Note that, the exact result up to 6 significant digits is .00474330.

Here, it is candidly seen that if we compute the function $f(x) = \sqrt{x+1} - \sqrt{x}$ directly, then it is error-prone. This is due to the fact that if we subtract two approximately equal numbers, then there is a loss of significant digits. For example in Process-I, when we subtract 105.413 and 105.409, then these two numbers are correct up to six significant digits, but the result $.004$ contains only one significant digit. Since there is a loss of five significant digits, so the result obtained is highly erroneous. This step can be avoided by rationalizing the function $f(x)$. The result obtained in Process-II after rationalization is correct up to five significant digits.

Stability of the Process:

It is clear from Example 1.5.5 that computation of the same function from two different processes can produce different results. There are following two major phases for computation of the function value $f(x)$:

i) First phase is to check the condition of the function by computing the CN of the function.

ii) Second phase is to check the stability of the process involved in the computation of the function. The stability of process can be checked by calculating the condition of each step in the process.

The function $f(x) = 1/(1-x^2)$ is ill-conditioned (CN $\gg 1$) near $x = \pm 1$. If the function is ill-conditioned then whatever process we will use, it tends to error. So every process will produce an error in computation of the function value $f(x) = 1/(1-x^2)$ near $x = \pm 1$.

The function $f(x) = \sqrt{x+1} - \sqrt{x}$ at $x = 11111$ is well conditioned (CN $\approx 1/2$, Example 1.5.4). If the function is well conditioned, then we have to compute the function value by the stable process. If even a single step of the process is ill-conditioned, then the whole process is an unstable process, and we have to switch over to any other alternate stable process.

Example 1.5.6. Discuss the stability of the Processes-I and II in Example 1.5.5. Hence, validate the results that the Processes-I yields erroneous result and Process-II produces a more accurate result for the same function $f(x)$.

Solution. We will calculate the CN of each step involved in both the Processes-I and II.

$$\begin{aligned} \text{Process-I:} \quad f(x) &= \sqrt{x+1} - \sqrt{x} \\ f(x) &= \sqrt{11112} - \sqrt{11111} \\ &= 105.413 - 105.409 \\ &= 0.004 \end{aligned}$$

Various computational steps in the process are as follows

$$\begin{array}{ll} x_1 = 11111 & (f(x) = \text{ Constant, CN} = 0) \\ x_2 = x_1 + 1 = 11112 & (f(x) = x+1, \text{CN} = 1) \\ x_3 = \sqrt{x_2} = 105.413 & (f(x) = \sqrt{x}, \text{CN} = 1/2) \\ x_4 = \sqrt{x_1} = 105.409 & (f(x) = \sqrt{x}, \text{CN} = 1/2) \\ x_5 = x_4 - x_3 = .004 & (f(x) = x - x_3 \text{ and } f(x) = x_4 - x, \text{CN} = 26352) \end{array}$$

In the last step $x_5 = x_4 - x_3$, we can assume the function $f(x)$ of variable x_3 or x_4. Let $f(x) = x_4 - x$, so condition for this step is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x(-1)}{x_4 - x}\right| = \left|\frac{105.409}{.004}\right| \approx 26352$$

This step is not a stable step as CN is very large. So the whole process is an unstable process due to this step. That's why the result obtained from this process is highly erroneous.

Process-II: $f(x) = \dfrac{1}{\sqrt{x+1} + \sqrt{x}}$

We will check the conditions of each step in Process-II, and conclude that each step in this process is well conditioned.

$$f(x) = \frac{1}{\sqrt{11112} + \sqrt{11111}} = \frac{1}{105.413 + 105.409} = \frac{1}{210.822} = 0.00474334$$

Various steps involved in this process are as follows

$$\begin{aligned} x_1 &= 11111 \\ x_2 &= x_1 + 1 = 11112 \\ x_3 &= \sqrt{x_2} = 105.413 \\ x_4 &= \sqrt{x_1} = 105.409 \\ x_5 &= x_4 + x_3 = 210.822 \\ x_6 &= \frac{1}{x_5} = 0.00474334 \end{aligned}$$

The first four steps in the process are well conditioned as discussed in Process-I. For the fifth step, let $f(x) = x + x_4$. The condition for this step is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x(1)}{x_4 + x}\right| = \left|\frac{105.409}{222.822}\right| = \frac{1}{2}$$

The last step is $f(x) = \frac{1}{x} = 1$, and the condition for this step is given by

$$\text{CN} = \left|\frac{xf'(x)}{f(x)}\right| = \left|\frac{x\left(-\frac{1}{x^2}\right)}{\frac{1}{x}}\right| = 1$$

From above discussion, it is clear that all the steps in Process-II are well conditioned, and hence this process is a stable process. Since the process is stable, so the result obtained is accurate to five significant digits.

Note 1.5.7. Even a single step in the process can make the whole process unstable. So we have to be extra careful during a large process, and must avoid the steps (if possible) with the loss of significant digits. We can use any alternate approach like rationalization, Taylor series expansion, etc. to avoid loss of significant digits.

Example 1.5.8. Discuss the stability of the function $f(x) = 1 - \cos(x)$, when x is nearly equal to zero. Find a stable way to compute the function $f(x)$.

Solution. If we directly compute the function $f(x) = 1 - \cos(x)$ at $x \approx 0$, then it will lead to subtraction of two nearly equal numbers and produce loss of significance. To avoid this loss, we can use any of the following three alternates

$$(i) \quad f(x) = 1 - \cos(x) = 1 - \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right) = \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots$$

$$(ii) \quad f(x) = 1 - \cos(x) = \frac{1 - \cos^2(x)}{1 + \cos(x)} = \frac{\sin^2(x)}{1 + \cos(x)}$$

$$(iii) \quad f(x) = 1 - \cos(x) = 2\sin^2\frac{x}{2}$$

1.6 Some Interesting Facts about Error

1. Let us assume we are doing six significant digits arithmetic on a hypothetical computer. If we want to add a small number $x = 0.000123$ to a large number $y = 123.456$ using this computer, then

$$\begin{aligned} x + y &= (.123456)10^3 + (.123000)10^{-3} \quad \text{(Normalized form)} \\ &= (.123456)10^3 + (.000000)10^3 \quad \text{(Equal exponent using symmetric rounding)} \\ &= (.123456)10^3 \quad \text{(Result, we missed the addition!)} \end{aligned}$$

This type of situations occurred commonly during the computations of infinite series. In these series, the initial terms are comparatively large. So, usually after adding some terms of the series, we are in a situation of adding a small term to a very large term. It may produce high rounding error in the computation. To avoid this kind of error, we can use backward sum of the series instead of forward sum, such that the each new term is compatible with the magnitude of accumulated sum.

2. In the case of series with mixed signs (like Taylor series of sin(x)), sometimes individual terms are larger than the summation itself. For example, in Taylor series of $\sin(2.13)$, the first term is 2.13. It is called as smearing, and we should use these kinds of series with extra care.

3. While performing arithmetic computations in a numerical method, the steps involving large number of arithmetic operations must be computed in double precisions. Such operations are error-prone to round-off error. For example, in Gauss-Seidel method for the solution of system of linear equations, the inner product
$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$
is a common operation, and such computations must be made in double precisions.

Exercise 1.6.1.

1. Define the terms error, absolute error, relative error and significant digits. The numbers $x = 1.28$ and $y = 0.786$ are correct to the digits specified. Find estimates to the relative errors in $x + y, x - y, xy$, and x/y.

2. Calculate the absolute and relative errors in the expression $3a - 2bc + \dfrac{b}{a}$ if the measurement of $a = 3.5435, b = .2588$ and $c = 1.0150$ are possible only to correct up to four decimal points.

3. Find the maximum possible error in the computed value of the hyperbolic sine function $\sinh(x) = \dfrac{e^x - e^{-x}}{2}$ at the point $x = 1$, if the maximum possible error in the value of x is $dx = 0.01$.

4. Let the function $u = 4x^2y^3/z^4$ and errors in the values of variables x, y, z are 0.001. Find the relative error in the function u at $x = y = z = 1$.

5. The radius r and height h of a right circular cylinder are measured as 2.5 m and 1.6 m, respectively, with a maximum error of 2%. Compute the resulting percentage error measured in the volume of the cylinder by the formula $V = \pi r^2 h$. Assume the value of π is exact for calculation.

6. Calculate the number of terms required in Taylor series approximation of the function $\cos(x)$ to compute the value of $\cos(\pi/12)$ correct up to 4-decimal places.

7. Find the number of terms of the Taylor series expansion of the function e^x required to compute the value of e correct to six decimal places.

8. Discuss the condition and stability of the function $f(x) = x - \sqrt{x^2 - 1}$ at $x = 11111$, using six significant digits floating point rounding arithmetic. Find a stable way to compute the function.

9. Discuss the condition and stability of the function $f(x) = x - \sin(x)$, when x is nearly equal to zero. Find a stable way to compute the function $f(x)$.

10. Discuss CN and stability of the function $y = \sec(x)$ in the interval $[0, \pi/2]$.

Unit 2

Course Structure

- Interpolation: Hermite's and spline interpolation. Interpolation by iteration –Aitken's and Neville's schemes.

2.1 Introduction

The statement $y = f(x),\ \ x_0 \leq x \leq x_n$ means: corresponding to every value of x in the range $x_0 \leq x \leq x_n$, 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.

2.2 Hermite Interpolation

We are familiar with the interpolating polynomial of degree $\leq n$; which passes through $(n+1)$ points $(x_i, f(x_i)); i = 0, 1, 2, \ldots, n$. Now, let us derive interpolating polynomial for a function $f(x)$ such that the values of the function $f(x)$ and its derivative $f'(x)$ match with this polynomial at $(n+1)$ points $x_i; i = 0, 1, 2, \ldots, n$. The polynomial of degree $\leq 2n+1$ is required to satisfies $2n+2$ conditions. Let us consider an interpolating polynomial of degree $\leq 2n+1$ which satisfies the following $2(n+1)$ restrictions at $(n+1)$ points $x_i; i = 0, 1, 2, \ldots, n$.

$$\left.\begin{array}{l} P_{2n+1}(x_i) = f(x_i) \\ P'_{2n+1}(x_i) = f'(x_i) \end{array}\right\}, \quad i = 0, 1, 2, \cdots, n \tag{2.2.1}$$

We have to express the polynomial $P_{2n+1}(x)$ in terms of $(n+1)$ points, $x_i; i = 0, 1, 2, \ldots, n$. Therefore, let the polynomial $P_{2n+1}(x)$ be of the following form

$$P_{2n+1}(x) = \sum_{i=0}^{n} u_i(x) P_{2n+1}(x_i) + \sum_{i=0}^{n} v_i(x) P'_{2n+1}(x_i) = \sum_{i=0}^{n} u_i(x) f(x_i) + \sum_{i=0}^{n} v_i(x) f'(x_i) \tag{2.2.2}$$

where $u_i(x)$ and $v_i(x)$ are polynomials of degree $\leq 2n+1$. Let us rewrite these polynomials in terms of Lagrange polynomial coefficients $l_i(x)$ as follows

$$u_i(x) = (a_i x + b_i)\, l_i^2(x) \quad \text{and} \quad v_i(x) = (c_i x + d_i)\, l_i^2(x), \quad \text{for } i = 0, 1, 2, \ldots, n \tag{2.2.3}$$

where a_i, b_i, c_i, d_i are constants to be determined. The coefficients $l_i(x)$ are given by

$$l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k} = \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}$$

with property

$$l_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{2.2.4}$$

The polynomial (2.2.2) is interpolating polynomial if it satisfies the conditions (2.2.1). For this, we have

$$\begin{aligned} &u_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad v_i'(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \\ &\text{and} \\ &v_i(x_j) = 0 \;\; \forall j = 0, 1, \cdots, n \qquad u_i'(x_j) = 0 \;\; \forall j = 0, 1, \cdots, n \end{aligned} \tag{2.2.5}$$

On using Eqs. (2.2.3) - (2.2.5) , we have

$$\begin{aligned} u_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad &\Rightarrow \quad (a_i x_j + b_i)\, l_i^2(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \\ &\qquad a_i x_i + b_i = 1 \\ v_i(x_j) = 0 \;\; \forall \;\; j = 0, 1, \cdots, n \quad &\Rightarrow \quad (c_i x_j + d_i)\, l_i^2(x_j) = 0 \\ &\qquad c_i x_i + d_i = 0 \\ v_i'(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad &\Rightarrow \quad (c_i x_j + d_i)\, 2 l_i(x_j)\, l_i'(x_j) + c_i l_i^2(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \\ &\qquad (c_i x_i + d_i) 2 l_i'(x_i) + c_i = 1 \\ u_i'(x_j) = 0 \forall j = 0, 1, \cdots, n \quad &\Rightarrow \quad (a_i x_j + b_i)\, 2 l_i(x_j)\, l_i'(x_j) + a_i l_i^2(x_j) = 0 \\ &\qquad (a_i x_i + b_i) 2 l_i'(x_i) + a_i = 1 \end{aligned}$$

Therefore, we have following four sets of equations in the variables, $a_i, b_i, c_i, d_i; i = 0, 1, \cdots, n$.

$$\left.\begin{array}{l} a_i x_i + b_i = 1 \\ c_i x_i + d_i = 0 \\ (c_i x_i + d_i)\, 2 l_i'(x_i) + c_i = 1 \\ (a_i x_i + b_i)\, 2 l_i'(x_i) + a_i = 0 \end{array}\right\} \Rightarrow \left\{\begin{array}{l} a_i = -2 l_i'(x_i) \\ b_i = 1 + 2 x_i l_i'(x_i) \\ c_i = 1 \\ d_i = -x_i \end{array}\right.$$

On using these values of constants, $a_i, b_i, c_i, d_i; i = 0, 1, \cdots, n$ in Eqs. (2.2.3), we get

$$\begin{aligned} u_i(x) &= (a_i x + b_i)\, l_i^2(x) = \left(-2 l_i'(x_i)\, x + 1 + 2 x_i l_i'(x_i)\right) l_i^2(x) \\ v_i(x) &= (c_i x + d_i)\, l_i^2(x) = (x - x_i)\, l_i^2(x) \end{aligned}$$

Use these values in Eq. (2.2.2) to get the following Hermite interpolating polynomial

$$\begin{aligned} P_{2n+1}(x) &= \sum_{i=0}^{n} u_i(x) f(x_i) + \sum_{i=0}^{n} v_i(x) f'(x_i) \\ &= \sum_{i=0}^{n} \left(-2l_i'(x_i)\, x + 1 + 2x_i l_i'(x_i)\right) l_i^2(x) f(x_i) + \sum_{i=0}^{n} (x - x_i)\, l_i^2(x) f'(x_i) \end{aligned} \tag{2.2.6}$$

Example 2.2.1. Compute the Hermite interpolating polynomial and then the value of the function $f(0.5)$ from the following data set.

x	-1	0	1	2
$f(x)$	2	2	2	26
$f'(x)$	2	0	2	68

Solution. We have $4(= n+1)$ points, $x_0 = -1, x_1 = 0, x_2 = 1, x_3 = 2$; therefore, the Hermite polynomial (2.2.6) is of degree $\leq 7(= 2n+1)$. It is given by

$$P(x) = \sum_{i=0}^{3} \left(-2l_i'(x_i)\, x + 1 + 2x_i l_i'(x_i)\right) l_i^2(x) f(x_i) + \sum_{i=0}^{3} (x - x_i)\, l_i^2(x) f'(x_i)$$

We have to calculate Lagrange coefficients polynomials $l_i(x)$ and their derivatives $l_i'(x_i)$ to compute the interpolating polynomial. On using $n = 3$ in the following formula, we have

$$l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k} = \frac{(x - x_0)(x - x_1) \cdots (x - x_{i-1})(x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)}$$

For $i = 0, 1, 2, 3$, we have

$$l_0(x) = \prod_{\substack{k=0 \\ k \neq 0}}^{3} \frac{x - x_k}{x_0 - x_k} = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} = \frac{-1}{6}(x - 0)(x - 1)(x - 2)$$

$$\Rightarrow l_0'(x_0) = \frac{-11}{6}$$

$$l_1(x) = \prod_{\substack{k=0 \\ k \neq 1}}^{3} \frac{x - x_k}{x_1 - x_k} = \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} = \frac{1}{2}(x + 1)(x - 1)(x - 2)$$

$$\Rightarrow l_1'(x_1) = \frac{-1}{2}$$

$$l_2(x) = \prod_{\substack{k=0 \\ k \neq 2}}^{3} \frac{x - x_k}{x_2 - x_k} = \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} = \frac{-1}{2}(x + 1)(x - 0)(x - 2)$$

$$l_2'(x_2) = \frac{1}{2}$$

$$l_3(x) = \prod_{\substack{k=0 \\ k \neq 3}}^{3} \frac{x - x_k}{x_3 - x_k} = \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} = \frac{1}{6}(x + 1)(x - 0)(x - 1)$$

$$l_3'(x_3) = \frac{11}{6}$$

The interpolating polynomial is given by

$$\begin{aligned}
P(x) &= \sum_{t=0}^{3}\left(-2l_t'(x_t)\,x+1+2x_tl_t'(x_t)\right)l_t^2(x)f(x_t)+\sum_{i=0}^{3}(x-x_t)\,l_t^2(x)f'(x_t)\\
&= \left(-2l_0'(x_0)\,x+1+2x_0l_0'(x_0)\right)l_0^2(x)f(x_0)+\left(-2l_1'(x_1)\,x+1+2x_1l_1'(x_1)\right)l_1^2(x)f(x_1)\\
&\quad+\left(-2l_2'(x_2)\,x+1+2x_2l_2'(x_2)\right)l_2^2(x)f(x_2)+\left(-2l_3'(x_3)\,x+1+2x_3l_3'(x_3)\right)l_3^2(x)f(x_3)\\
&\quad+(x-x_0)\,l_0^2(x)f'(x_0)+(x-x_1)\,l_1^2(x)f'(x_1)+(x-x_2)\,l_2^2(x)f'(x_2)+(x-x_3)\,l_3^2(x)f'(x_3)\\
&= \left(\frac{11}{3}x+\frac{14}{3}\right)\left(\frac{-1}{6}(x-0)(x-1)(x-2)\right)^2(2)+(x+1)\left(\frac{1}{2}(x+1)(x-1)(x-2)\right)^2\\
&\quad+(-x+2)\left(\frac{-1}{2}(x+1)(x-0)(x-2)\right)^2(2)+\left(\frac{-11}{3}x+\frac{25}{3}\right)\left(\frac{1}{6}(x+1)(x-0)(x-1)\right)^2\\
&\quad+(x+1)\left(\frac{-1}{6}(x-0)(x-1)(x-2)\right)^2(2)+(x-0)\left(\frac{1}{2}(x+1)(x-1)(x-2)\right)^2(0)\\
&\quad+(x-1)\left(\frac{-1}{2}(x+1)(x-0)(x-2)\right)^2(2)+(x-2)\left(\frac{1}{6}(x+1)(x-0)(x-1)\right)^2(68)\\
&= \left(\frac{28}{3}x+\frac{34}{3}\right)\left(\frac{-1}{6}(x-0)(x-1)(x-2)\right)^2+(2x+2)\left(\frac{1}{2}(x+1)(x-1)(x-2)\right)^2\\
&\quad+2\left(\frac{-1}{2}(x+1)(x-0)(x-2)\right)^2+\left(\frac{-82}{3}x+\frac{244}{3}\right)\left(\frac{1}{6}(x+1)(x-0)(x-1)\right)^2\\
&= x^5-x^3+2
\end{aligned}$$

Therefore, $P(0.5)=(0.5)^5-(0.5)^3+2=1.90625$

Note: We can easily verify that the polynomial satisfies all the conditions

x	-1	0	1	2
$P(x)=x^5-x^3+2$	2	2	2	26
$P'(x)=5x^4-3x^2$	2	0	2	68

The polynomial x^5-x^3+2 is a unique polynomial of degree $\leq 7(=2n+1)$, and it satisfies the conditions above. Again, it is worth to mentioning here that there are an infinite number of polynomials of degree >7 which satisfying above conditions.

Exercise 2.2.2. 1. From the following table, find $f(0.5),\ f'(1.5)$ using Hermite interpolation.

x	0	2	3
$f(x)$	2	16	80
$f'(x)$	-1	31	107

2. Find the Hermite polynomial of the third degree approximating the function $y(x)$ such that

$$\begin{aligned} y(0)&=1, \quad y'(0)=0\\ y(1)&=3, \quad y'(1)=5. \end{aligned}$$

3. Calculate $f(1.2)$ by approximating the following values with cubic polynomial $f(1)=0,\ f'(1)=1, f(2)=0.693147, f'(2)=0.5$

2.3 Spline Interpolation

A sequence of continuous curves that are connected to form a single continuous curve is called as a spline curve. Consider a given set of data points $(x_i, y_i),\ i = 0, 1, \ldots, n$ such that $x_i < x_{i+1}$ for all $i = 0, 1, \ldots, n-1$. In general, a m-th degree spline $P_s(x)$ for this data set is a piecewise polynomial of degree m, which satisfies following two conditions:

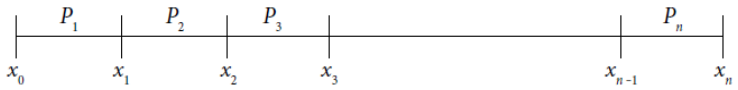
1. It is of degree $\leq m$ for each interval $(x_i, x_{i+1}), i = 0, 1, \ldots, n-1$ and of degree m in at least one such interval.

2. The spline $P_s(x)$ and its first $m-1$ derivatives are continuous at each node points $x_i, i = 1, \ldots, n-1$ in the interval (x_0, x_n).

2.3.1 Cubic Spline Interpolation

Let us approximate the function $f(x)$ by different cubic polynomials $P_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i, i = 1, 2, \ldots, n$ for each sub-interval $[x_{i-1}, x_i]$ in the given interval $[x_0, x_n]$.

$$P(x) = \begin{cases} P_1(x) = a_1 x^3 + b_1 x^2 + c_1 x + d_1 & x_0 \leq x \leq x_1 \\ P_2(x) = a_2 x^3 + b_2 x^2 + c_2 x + d_2 & x_1 \leq x \leq x_2 \\ \vdots & \\ P_n(x) = a_n x^3 + b_n x^2 + c_n x + d_n & x_{n-1} \leq x \leq x_n \end{cases}$$



In cubic spline approximation, the polynomials and their first and second derivatives are continuous at node points. A cubic spline polynomial $P(x)$ satisfies the following three properties.

1. On each subinterval $[x_{i-1}, x_i], 1 \leq i \leq n, P(x)$ is a third-degree polynomial, i.e.,

$$P_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i, i = 1, 2, \ldots, n$$

   We have to find $4n$ unknowns: $a_i, b_i, c_i, d_i; i = 1, 2, \ldots, n$.

2. The values of the cubic spline at node points equal the values of the function at these points.

$$P(x_i) = f_i, i = 0, 1, \ldots, n$$

3. The polynomials $P(x), P'(x)$ and $P''(x)$ are continuous throughout the interval (x_0, x_n).

On using the above second and third properties, we have following results

a) Continuity of $P(x)$: At each node point $x = x_p$, the values of two polynomials $P_i(x)$ and $P_{i+1}(x)$ must be equal, and also equals to the value of the function $f(x_i)$. At any node point $x = x_i, i = 1, 2, \ldots, n-1$, we can obtain following equations.

The polynomial value $P_i(x_i)$ for interval $[x_{i-1}, x_i]$, must equals the function value $f(x_i)$

$$P_i(x_i) = a_i x_i^3 + b_i x_i^2 + c_i x_i + d_i = f(x_i) = f_i$$

Similarly, polynomial for interval $[x_i, x_{i+1}]$ gives following equations

$$P_{i+1}(x_i) = a_{i+1} x_i^3 + b_{i+1} x_i^2 + c_{i+1} x_i + d_{i+1} = f(x_i) = f_i$$

So, we have following set of equations

$$\begin{aligned} a_i x_i^3 + b_i x_i^2 + c_i x_i + d_i &= f_i \\ a_{i+1} x_i^3 + b_{i+1} x_i^2 + c_{i+1} x_i + d_{i+1} &= f_i \end{aligned} \tag{2.3.1}$$

b) At the end points x_0 and x_n of the interval, the values of splines must be equal to the values of the function.

$$\begin{aligned} f_0 &= a_1 {x_0}^3 + b_1 {x_0}^2 + c_1 x_0 + d_1 \\ f_n &= a_n {x_n}^3 + b_n {x_n}^2 + c_n x_n + d_n \end{aligned} \tag{2.3.2}$$

c) Continuity of $P'(x)$ and $P''(x)$: At each node point $x = x_i$; the values of polynomials $P'(x)$ and $P'_{i+1}(x)$ are equal, and the values of polynomials $P_{ii}(x)$ and $P''_{i+1}(x)$ are also equal. At node points $x = x_p, i = 1, 2, \ldots, n-1$; we must have

$$3a_i x_i^2 + 2b_i x_i + c_i = 3a_{i+1} x_i^2 + 2b_{i+1} x_i + c_{i+1} \tag{2.3.3}$$

$$6a_i x_i + 2b_i = 6a_{i+1} x_i + 2b_{i+1} \tag{2.3.4}$$

We have $2(n-1)$ equations from system (2.3.1); two equations from system (2.3.2); and $2(n-1)$ equations from systems (2.3.3) and (2.3.4). So, we have total $4n-2$ equations, while the number of arbitrary constants to be determined is $4n\,(a_i, b_i, c_i, d_i; i = 1, 2, \ldots, n)$. Hence, we need two more equations for the polynomials to be unique.

Let us take the notation $P'(x_i) = m_i$ and $P''(x_i) = M_i$. In general, we assign some values to the polynomial $P''(x)$ at the end points, that is $P''(x_0) = M_0$ and $P''(x_n) = M_n$. If the end conditions are $M_0 = 0$ and $M_n = 0$, then our spline is called as a natural spline (As the drafting of the spline always behaves in this manner).

At last, we have $4n$ equations in $4n$ variables; which can be easily solved to obtain the required cubic spline. But to reduce the computational work, we use an alternative method to obtain the cubic spline interpolation described below.

Alternative Method for Cubic Spline: Since the function $P(x)$ is a cubic polynomial, so the function $P''(x)$ is linear function of x in the interval $x_{i-1} \le x \le x_i$ and can be written as

$$P''(x) = \frac{x_i - x}{x_i - x_{i-1}} P''(x_{i-1}) + \frac{(x - x_{i-1})}{x_i - x_{i-1}} P''(x_i)$$

Let us assume that the length of the interval (x_{i-1}, x_i) is h_i i.e. $h_i = x_i - x_{i-1}$. Also, assume $M_i = P''(x_i)$

$$P''(x) = \frac{x_i - x}{h_i} M_{i-1} + \frac{(x - x_{i-1})}{h_i} M_i$$

On integrating this equation twice on x, we have

$$P(x) = \frac{(x_i - x)^3}{6h_i} M_{i-1} + \frac{(x - x_{i-1})^3}{6h_i} M_i + k_1 x + k_2 \tag{2.3.5}$$

where k_1 and k_2 are arbitrary constants. The values of cubic spline polynomials must equal to function values at nodal points; therefore, we have

$$P(x_{i-1}) = f(x_{i-1}) = f_{i-1} \text{ and } P(x_i) = f(x_i) = f_i$$

On using these conditions in Eq. (2.3.5), we have

$$P(x_{i-1}) = f_{i-1} = \frac{1}{6} h^2 M_{i-1} + k_1 x_{i-1} + k_2$$
$$P(x_i) = f_i = \frac{1}{6} h^2 M_i + k_1 x_i + k_2$$

Solution of these two equations for k_1 and k_2 is given by

$$k_1 = \frac{1}{h_i}(f_i - f_{i-1}) - \frac{1}{6}(M_i - M_{i-1}) h_i$$
$$k_2 = \frac{1}{h_i}(x_i f_{i-1} - x_{i-1} f_i) - \frac{1}{6}(x_i M_{i-1} - x_{i-1} M_i) h_i$$

On substituting these values of k_1 and k_2 in Eq. (2.3.5), we have

$$\begin{aligned} P(x) &= \frac{1}{6h_i}(x_i - x)^3 M_{i-1} + \frac{1}{6h_i}(x - x_{i-1})^3 M_i + \frac{x}{h_i}(f_i - f_{i-1}) \\ &- \frac{x}{6}(M_i - M_{i-1}) h_i + \frac{1}{h_i}(x_i f_{i-1} - x_{i-1} f_i) - \frac{1}{6}(x_i M_{i-1} - x_{i-1} M_i) h_i \quad (x_{i-1} \le x \le x_i) \\ &= \frac{1}{6h_i}\left[(x_i - x)\left\{(x_i - x)^2 - h^2\right\}\right] M_{i-1} + \frac{1}{6h_i}\left[(x - x_{i-1})\left\{(x - x_{i-1})^2 - h^2\right\}\right] M_i \\ &+ \frac{1}{h_i}(x_i - x) f_{i-1} + \frac{1}{h_i}(x - x_{i-1}) f_i \quad i = 1, 2, \ldots, n \end{aligned} \tag{2.3.6}$$

To compute values of M_{i-1} and M_i, we will use continuity of the polynomial $P'(x)$. On differentiating the Eq. (2.3.6) w.r.t. x, we get

$$P'(x) = -\frac{(x_i - x)^2}{2h_i} M_{i-1} + \frac{(x - x_{i-1})^2}{2h_i} M_i - \frac{(M_i - M_{i-1}) h_i}{6} + \frac{f_i - f_{i-1}}{h_i} \quad x_{i-1} \le x \le x_i \tag{2.3.7}$$

Similarly, we can obtain $P'(x)$ for the interval $x_i \le x \le x_{i+1}$, by simply changing $i = i + 1$ in Eq. (2.3.7).

$$P'(x) = -\frac{(x_{i+1} - x)^2}{2h_{i+1}} M_i + \frac{(x - x_i)^2}{2h_{i+1}} M_{i+1} - \frac{1}{6}(M_{i+1} - M_i) h_{i+1} + \frac{f_{i+1} - f_i}{h_{i+1}} \quad (x_i \le x \le x_{i+1}) \tag{2.3.8}$$

The continuity of the derivatives implies that the derivatives $P'(x)$ in both the intervals $x_{i-1} \le x \le x_i$ and $x_i \le x \le x_{i+1}$ must be equal at the node point $x = x_i$. We have

$$\Rightarrow \frac{1}{2} h_i M_i - \frac{(M_i - M_{i-1}) h_i}{6} + \frac{f_i - f_{i-1}}{h_i} = -\frac{1}{2} h_{i+1} M_i - \frac{1}{6}(M_{i+1} - M_i) h_{i+1} + \frac{f_{i+1} - f_i}{h_{i+1}}$$

On rewriting this equation, we have

$$\frac{h_i}{6}M_{i-1}+\frac{h_i+h_{i+1}}{3}M_i+\frac{h_{i+1}}{6}M_{i+1}=\frac{1}{h_{i+1}}\left(f_{i+1}-f_i\right)-\frac{1}{h_i}\left(f_i-f_{i-1}\right) \quad i=1,2,\ldots,n-1;\ x_{i-1}\le x\le x_i \tag{2.3.9}$$

The system (2.3.9) will produce a linear system of $(n-1)$ equations in $(n+1)$ unknowns $M_0, M_1, \ldots, M_n$. We can use any two additional conditions for unique solution of the system. The spline is a natural spline in case of end conditions $M_0 = 0$ and $M_n = 0$.

We solve the system (2.3.9), and then use the values of $M_0, M_1, \ldots, M_n$ in system (2.3.6) to obtain the following cubic spline as desired.

$$\begin{aligned} P(x) \;=\; & \frac{1}{6h_i}\left(x_i-x\right)^3 M_{i-1}+\frac{1}{6h_i}\left(x-x_{i-1}\right)^3 M_i+\frac{x}{h_i}\left(f_i-f_{i-1}\right)-\frac{x}{6}\left(M_i-M_{i-1}\right)h_i \\ & +\frac{1}{h_i}\left(x_i f_{i-1}-x_{i-1}f_i\right)-\frac{1}{6}\left(x_i M_{i-1}-x_{i-1}M_i\right)h_i \quad i=1,2,\ldots,n \end{aligned} \tag{2.3.10}$$

2.3.2 Cubic Spline for Equi-spaced Points

The interval length of all the intervals is same in case of equi-spaced points, i.e.,

$$h_1 = h_2 = \ldots = h_n = h$$

So, our system of Eqs. (2.3.9) and (2.3.10) reduces to following Eqs. (2.3.11) and (2.3.12), respectively

$$M_{i-1}+4M_i+M_{i+1}=\frac{6}{h^2}\left(f_{i+1}-2f_i+f_{i-1}\right) \quad i=1,2,\ldots,n-1 \tag{2.3.11}$$

$$\begin{aligned} P(x) = & \frac{1}{6h}\left[\left(x_i-x\right)^3 M_{i-1}+\left(x-x_{i-1}\right)^3 M_i\right]+\frac{1}{h}(x_i-x)\left(f_{i-1}-\frac{h^2}{6}M_{i-1}\right) \\ & +\frac{1}{h}(x-x_{i-1})\left(f_i-\frac{h^2}{6}M_i\right), \quad i=1,2,\ldots,n \end{aligned} \tag{2.3.12}$$

These systems can be solved to obtain the desired cubic spline.

Example 2.3.1. Obtain cubic spline approximation for $e^{0.2}$ from the following values of e^x correct up to six significant digits. Use natural spline conditions.

x	0	0.1	0.3	0.4
e^x	1	1.10517	1.34986	1.49182

Solution.
Given

$$h_1 = 0.1, h_2 = 0.2, h_3 = 0.1$$
$$x_0 = 0, x_1 = 0.1, x_2 = 0.3, x_3 = 0.4$$
$$f_0 = 1, f_1 = 1.10517, f_2 = 1.34986, f_3 = 1.49182$$

Natural spline conditions are $M_0 = M_3 = 0$. Since the points are not equispaced; we have to use Eq. (2.3.9) for the values of M_1 and M_2

$$\frac{h_i}{6}M_{i-1}+\frac{h_i+h_{i+1}}{3}M_i+\frac{h_{i+1}}{6}M_{i+1}=\frac{1}{h_{i+1}}\left(f_{i+1}-f_i\right)-\frac{1}{h_i}\left(f_i-f_{i-1}\right)$$

$i = 1$

$$\begin{aligned} &\frac{h_1}{6}M_0 + \frac{h_1+h_2}{3}M_1 + \frac{h_2}{6}M_2 = \frac{1}{h_2}(f_2 - f_1) - \frac{1}{h_1}(f_1 - f_0) \\ &6M_1 + 2M_2 = 10.305 \end{aligned} \tag{2.3.13}$$

$i = 2$

$$\begin{aligned} &\frac{h_2}{6}M_1 + \frac{h_2+h_3}{3}M_2 + \frac{h_3}{6}M_3 = \frac{1}{h_3}(f_3 - f_2) - \frac{1}{h_2}(f_2 - f_1) \\ &2M_1 + 6M_2 = 11.769 \end{aligned} \tag{2.3.14}$$

On solving Eqs. (2.3.13) and (2.3.14) for M_1 and M_2, we get

$$M_1 = 1.196625 \text{ and } M_2 = 1.562625$$

We have to compute the value of $e^{0.2}$, that is in the interval (x_1, x_2). Therefore, we will use these values of M_1 and M_2 in the Eq. (2.3.10) for $i = 1$ to obtain the cubic spline approximation of the value $e^{0.2}$.

$$\begin{aligned} P(x) =& \frac{1}{6h_i}(x_i - x)^3 M_{i-1} + \frac{1}{6h_i}(x - x_{i-1})^3 M_i + \frac{x}{h_i}(f_i - f_{i-1}) \\ &- \frac{x}{6}(M_i - M_{i-1})h_i + \frac{1}{h_i}(x_i f_{i-1} - x_{i-1} f_i) - \frac{1}{6}(x_i M_{i-1} - x_{i-1} M_i)h_i \end{aligned}$$

$i = 1$

$$\begin{aligned} P(x) =& \frac{1}{6h_1}(x_1 - x)^3 M_0 + \frac{1}{6h_1}(x - x_0)^3 M_1 + \frac{x}{h_1}(f_1 - f_0) \\ &- \frac{x}{6}(M_1 - M_0)h_1 + \frac{1}{h_1}(x_1 f_0 - x_0 f_1) - \frac{1}{6}(x_1 M_0 - x_0 M_1)h_1 \end{aligned}$$

On using different values and $x = 0.2$, we obtain following cubic spline approximation for $e^{0.2}$

$$P(0.2) = 1.22088$$

While the exact value of $e^{0.2}$ is 1.22140. We can also compute the following cubic spline polynomials for the data set

$$\begin{cases} 1 + 1.032x + 1.994x^3 & 0 \le x \le 0.1 \\ 1.002 + 0.981x + 0.507x^2 + 0.305x^3 & 0.1 \le x \le 0.3 \\ 1.080 + 0.196x + 3.125x^2 - 2.604x^3 & 0.3 \le x \le 0.4 \end{cases}$$

Exercise 2.3.2. 1. Check the following functions, that they are splines or not

$$(i)\ f(x) = \begin{cases} 5x, & 0 \le x \le 1 \\ 11x - 6 & 1 \le x \le 2 \\ -4x + 10x, & 2 \le x \le 3 \end{cases} \quad (ii)\ f(x) = \begin{cases} 12x - 7x^2, & 0 \le x \le 1 \\ 1 + 10x - 6x^2 & 1 \le x \le 2 \end{cases}$$

2. Determine the cubic spline polynomial for the following data set and hence compute the values of $f(0.3)$ and $f(2.6)$.

x	0	1	2	3
$f(x)$	1	-8	-30	-59

3. Find the cubic spline fit for the following data points

x	-1	0	1
$f(x)$	2	5	9

Use natural spline conditions $f''(-1) = 0$ and $f''(1) = 0$.

4. Construct the cubic spline with the natural end conditions that passes through the points $(-1, 0), (0, 1), (2, 5)$ and $(3, 2)$.

2.4 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:

$$\begin{aligned}
[x_0, x_1] &= \frac{y_1 - y_0}{x_1 - x_0}, \\
[x_0, x_1, x_2] &= \frac{[x_1, x_2] - [x_0, x_1]}{x_2 - x_0}, \\
&\vdots \\
[x_0, x_1, \ldots, x_n] &= \frac{[x_1, x_2, \ldots, x_n] - [x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}.
\end{aligned} \tag{2.4.1}$$

2.5 Newton's General Interpolation Formula

By definition, we have

$$[x, x_0] = \frac{y - y_0}{x - x_0},$$

so that

$$y = y_0 + (x - x_0)[x, x_0] \tag{2.5.1}$$

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.(2.5.1), we obtain

$$y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x, x_0, x_1] \tag{2.5.2}$$

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

$$[x, x_0, x_1] = [x_0, x_1, x_2] + (x - x_2)[x, x_0, x_1, x_2] \tag{2.5.3}$$

Equation (2.5.2) now gives

$$\begin{aligned} y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ +(x - x_0)(x - x_1)(x - x_2)[x, x_0, x_1, x_2] \end{aligned} \tag{2.5.4}$$

Proceeding in this way, we obtain

$$\begin{aligned} y = y_0 + (x - x_0)[x_0, x_1] + (x - x_0)(x - x_1)[x_0, x_1, x_2] \\ +(x - x_0)(x - x_1)(x - x_2)[x_0, x_1, x_2, x_3] + \dots \\ +(x - x_0)(x - x_1)(x - x_2) \cdot (x - x_n)[x, x_0, x_1, \dots, x_n] \end{aligned} \tag{2.5.5}$$

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 2.5.1. 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.(2.5.5) gives

$$\log_{10} 301 = 2.4771 + 0.00145 + (-3)(-0.00001) = 2.4786$$

2.6 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

$$\Delta_{01}(x) = y_0 + (x - x_0)[x_0, x_1] = \frac{1}{x_1 - x_0}\begin{vmatrix} y_0 & x_0 - x \\ y_1 & x_1 - x \end{vmatrix} \tag{2.6.1}$$

Similarly, we can form $\Delta_{02}(x),\ \Delta_{03}(x), \ldots$ Next, we form Δ_{012} by considering the first-three points:

$$\Delta_{012}(x) = \frac{1}{x_2 - x_1}\begin{vmatrix} \Delta_{01}(x) & x_1 - x \\ \Delta_{02}(x) & x_2 - x \end{vmatrix} \tag{2.6.2}$$

Similarly, we obtain $\Delta_{013}(x), \Delta_{014}(x)$, etc. At the n-th stage of approximation, we obtain

$$\Delta_{0123\ldots n}(x) = \frac{1}{x_n - x_{n-1}}\begin{vmatrix} \Delta_{0123\ldots\overline{n-1}}(x) & x_{n-1} - x \\ \Delta_{0123\ldots\overline{n-2n}}(x) & x_n - x \end{vmatrix} \tag{2.6.3}$$

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 (2.5.1).

Example 2.6.1. 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 2.6.2. 1. Given $f(x) = \frac{1}{x^2}$. Find the divided differences $[a, b]$, and $[a, b, c]$.

2. 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.

Unit 3

Course Structure

- Approximation of Function: Least square approximation. Weighted least square approximation.

3.1 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.

3.2 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

$$e_i = y_i - f(x_i) \tag{3.2.1}$$

If we write

$$\begin{aligned} S &= [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \ldots + [y_m - f(x_m)]^2 \\ &= e_1^2 + e_2^2 + \ldots + e_m^2, \end{aligned} \tag{3.2.2}$$

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 $(x_i, y_i),\ i = 1, 2, \ldots, m$.

3.2.1 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.(3.2.2), we have

$$S = [y_1 - (a_0 + a_1 x)]^2 + [y_2 - (a_0 + a_1 x)]^2 + \ldots + [y_m - (a_0 + a_1 x_m)]^2 \tag{3.2.3}$$

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

$$\begin{aligned} & ma_0 + a_1(x_1 + x_2 + \ldots + x_m) = y_1 + y_2 + \ldots + y_m \\ \text{and } & a_0(x_1 + x_2 + \ldots + x_m) + a_1(x_1^2 + x_2^2 + \ldots + x_m^2) = x_1y_1 + x_2y_2 + \ldots + x_my_m \end{aligned} \quad (3.2.4)$$

or more compactly to

$$ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i \quad \text{and} \quad a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i \quad (3.2.5)$$

Equations (3.2.5) 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

$$a_1 = \frac{m \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \cdot \sum_{i=1}^{m} y_i}{m \sum x_i^2 - \left(\sum_{i=1}^{m} x_i \right)^2} \quad (3.2.6)$$

and then

$$a_0 = \overline{y} - a_1 \overline{x}. \quad (3.2.7)$$

Since $\frac{\partial^2 S}{\partial a_0^2}$ and $\frac{\partial^2 S}{\partial a_1^2}$ are both positive at the points a_0 and a_1, it follows that these values provide a *minimum* of S. In Eq.(3.2.7), $\overline{x}$ and $\overline{y}$ are the means of x and y, respectively. Form Eq.(3.2.7), 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

$$cc = \sqrt{\frac{S_y - S}{S_y}}, \quad \text{where} \quad S_y = \sum_{i=1}^{m} (y_i - \overline{y})^2 \quad \text{and} \quad S \text{ is defined by Eq.(3.2.3)} \quad (3.2.8)$$

If cc is close to 1, then the fit is considered to be good, although this is not always true.

Example 3.2.1. 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 = $\sqrt{\dfrac{16.10 - 0.2240}{16.10}} = 0.9930.$

3.3 Nonlinear Curve Fitting by Linearization of Data

There are some nonlinear curves, which are equivalent to linear fitting after some transformations in the dependent and independent variables. For example, if we want to fit a curve of the type $y = ae^{bx}$ to a data set. Then taking natural log on both sides, we have

$$\ln(y) = \ln(a) + bx$$

This expression is equivalent to following linear expression

$$Y = A + BX$$

where $Y = \ln(y), A = \ln(a), B = b, X = x$. We are summarizing some nonlinear curves in the following table, which with simple operations and transformations can be converted into linear curve fitting.

Sr. No.	Function $y=f(x)$	Operations	Linearization $Y=A+BX$	New Variables and Constants $Y=A+BX$
1	$y = ae^{bx}$ $y = axe^{bx}$ $y = ax^b$	Take Log	$\ln(y) = \ln(a) + bx$ $\ln\left(\frac{y}{x}\right) = \ln(a) + bx$ $\ln(y) = \ln(a) + b\ln(x)$	$Y = \ln(y),\ A = \ln(a), B = b,\ X = x$ $Y = \ln\left(\frac{y}{x}\right),\ A = \ln(a), B = b,\ X = x$ $Y = \ln(y),\ A = \ln(a), B = b,\ X = \ln(x)$
2	$y = \frac{x}{ax+b}$ $y = \frac{1}{a+bx}$ $y = \frac{a}{x+b}$	Inverse	$\frac{1}{y} = \frac{ax+b}{x} = a + b\frac{1}{x}$ $\frac{1}{y} = a + bx$ $\frac{1}{y} = \frac{1}{a}x + \frac{b}{a}$	$Y = \frac{1}{y}, A = a,\ B = b,\ X = \frac{1}{x}$ $Y = \frac{1}{y}, A = a,\ B = b,\ X = x$ $Y = \frac{1}{y}, A = \frac{b}{a},\ B = \frac{1}{a},\ X = x$
3	$y = \frac{C}{1+ae^{bx}}$	Inverse and Log	$\ln\left(\frac{C}{y} - 1\right) = \ln(a) + bx$	$Y = \ln\left(\frac{C}{y} - 1\right),\ A = \ln(a), B = b,\ X = x$
4	$y = (a+bx)^m$	m^{th} root (m fixed)	$(y)^{-m} = a + bx$	$Y = (y)^{-m},\ A = a, B = b,\ X = x$
5	$y = a + bg(x)$ $y = a + b\frac{1}{x}$ $y = a + b\ln(x)$ Etc.	Let $X = g(x)$ $X = \frac{1}{x}$ $X = \ln(x)$	$y = a + bg(x)$ $y = a + b\frac{1}{x}$ $y = a + b\ln(x)$	$Y = y,\ A = a, B = b,\ X = g(x)$ $Y = y,\ A = a, B = b,\ X = \frac{1}{x}$ $Y = y,\ A = a, B = b,\ X = \ln(x)$

Example 3.3.1. Fit a curve $y = ax^b$ to the following data

$x:$	1	2	3	5	6
$y:$	1	9	29	129	221

Solution. On taking log on both sides of the curve $y = ax^b$, we get

$$\ln(y) = \ln(a) + b\ln(x)$$

So, the curve fitting of type $y = ax^b$ is equivalent to fit a straight line $Y = A + bX$, where $Y = \ln(y), A = \ln(a), X = \ln(x)$. Normal equations for straight line $Y = A + bX$ fitting are as follows

$$\begin{aligned} nA + b\sum_{i=1}^{n} X_l &= \sum_{i=1}^{n} Y_l \\ A\sum_{i=1}^{n} X_t + b\sum_{i=1}^{n} X_l^2 &= \sum_{i=1}^{n} X_t Y_t \end{aligned} \tag{3.3.1}$$

On computing various terms in normal equations

x	$X = \ln(x)$	y	$Y = \ln(y)$	X^2	XY
1	0	1	0	0	0
2	0.693147181	9	2.197224577	0.480453014	1.523000021
3	1.098612289	29	3.36729583	1.206948961	3.699352578
5	1.609437912	129	4.859812404	2.590290394	7.821566331
6	1.791759469	221	5.398162702	3.210401996	9.672209137
	5.192956851		15.82249551	7.488094364	22.71612807

The normal Eqs. (3.3.1) are as follows

$$5A + 5.192956851b = 15.822495515.192956851A + 7.488094364b = 22.71612807$$

On solving these equations for A and b, we obtain

$$\begin{aligned} A &= 0.04931094359, b = 2.99943582 \\ a &= e^A = 1.050546961, b = 2.99943582 \end{aligned}$$

Hence, our curve is $y = ax^b = 1.050546961x^{2.99943582}$.

Example 3.3.2. Following are census details for the population of India from the year 1961 to 2011. Fit an exponential curve $y = ae^{bx}$ to these data, and hence find the approximate population in the years 1966, 1985, 1996 and 2009.

Year (x)	1961	1971	1981	1991	2001	2011
Population (in crores) (y)	43.9235	54.8160	68.3329	84.6421	102.8737	121.0193

Solution. We have to fit an exponential curve $y = ae^{bx}$ for years (x) from 1961 to 2011 . To avoid lengthy calculations (like $y = ae^{kx} = a(2.718)^{(1961)b}$), we can shift the origin and rescale the data as follows

$$X = \frac{x - 1981}{10}$$

Now, we have to fit the exponential curve $y = ae^{bx}$ to the following data

$X:$	-2	-1	0	1	2	3
$y:$	43.9235	54.8160	68.3329	84.6421	102.8737	121.0193

On taking logarithmic on both side to the equation $y = ae^{bx}$, we get

$$\ln y = \ln a + bX$$

By replacing $Y = \ln(y), A = \ln(a)$, we have following straight line

$$Y = A + bX$$

Normal equations for this straight line are as follows

$$nA + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$A\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

X	y	$Y = \ln(y)$	X^2	XY
-2	43.9235	3.78245	4	-7.5649
-1	54.8160	4.00398	1	-4.00398
0	68.3329	4.22439	0	0
1	84.6421	4.43843	1	4.43843
2	102.8737	4.63350	4	9.26700
3	121.0193	4.79595	9	14.38785
3		25.87870	19	16.52440

The normal equations are given by

$$6A + 3b = 25.87870$$
$$3A + 19b = 16.52440$$

On solving these equations, we have

$$A = 4.21069, b = 0.20486$$

Since $A = \ln(a) \Rightarrow a = 67.40281$ Hence, the fitted curve is given by

$$y = 67.40281e^{0.20486X}$$

Now, we have to compute populations in the years $(x) = 1966, 1985, 1996$ and 2009 . Corresponding to these years, the variable X is given by

$$X = \frac{x - 1981}{10} = -1.5, 0.4, 1.5., 2.8$$

So, the populations are given by

$$y(1966) = 67.40281e^{0.20486(-1.5)} = 49.570533$$
$$y(1985) = 67.40281e^{0.20486(0.4)} = 73.158664$$
$$y(1996) = 67.40281e^{0.20486(1.5)} = 91.649962$$
$$y(2009) = 67.40281e^{0.20486(2.8)} = 119.616951$$

Exercise 3.3.3. 1. 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.

2. Fit a straight line to the following data set

x	1	1.2	1.4	1.6	1.8	2.0
$f(x)$	0	0.182322	0.336472	0.470004	0.587787	0.693147

3. The growth of cell culture (optical density) at various pH levels are tabulated in the following table.

pH(x)	3.5	4	4.5	5	5.5	6	6.5	7
Optical density(y)	0.20	0.28	0.35	0.41	0.46	0.52	0.55	0.62

4. An empirical formula for the effect of temperature on viscosity of a liquid is given by Andrade's equation $\mu = ae^{bt}$ where μ is the dynamic viscosity of the liquid, T is absolute temperature, and $a,\ b$ are constants. Compute the best fitted Andrade's model to the following data for a given liquid.

T :	10	20	30	40	50
μ :	4.67	3.84	3.17	2.71	2.53

5. It is expected from theoretical consideration, that the rate of flow is proportional to some power of the pressure at the nozzle of a fire hose. Get the least squares values for exponent and proportionality factor.

Flow rate (F) :	90	110	130	150	170
Pressure (P) :	10	18	28	41	53

3.4 Curve Fitting by Polynomials

Let the polynomial of the n-th degree

$$Y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{3.4.1}$$

be fitted to the data points $(x_i, y_i),\ i = 1, 2, \ldots, m$. We then have

$$\begin{aligned} S \;=\; & \Big[y_1 - (a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n)\Big]^2 + \Big[y_2 - (a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_n x_2^n)\Big]^2 \\ & + \ldots\ldots + \Big[y_m - (a_0 + a_1 x_m + a_2 x_m^2 + \ldots + a_n x_m^n)\Big]^2 \end{aligned} \tag{3.4.2}$$

Equating to zero the first partial derivatives and simplifying, we obtain the normal equations:

$$\begin{aligned} m a_0 + a_1 \sum x_i + a_2 \sum x_i^2 + \ldots a_n \sum x_i^n &= \sum y_i, \\ a_0 \sum x_i + a_1 \sum x_i^2 + \ldots a_n \sum x_i^{n+1} &= \sum x_i y_i, \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots &\qquad \vdots \\ a_0 \sum x_i^n + a_1 \sum x_i^{n+1} + \ldots a_n \sum x_i^{2n} &= \sum x_i^n y_i, \end{aligned} \tag{3.4.3}$$

where the summations are performed from $i = 1$ to $i = m$. The system (3.4.3) constitutes $(n + 1)$ equations in $(n + 1)$ unknowns, and hence can be solved for $a_0,\ a_1,\ \ldots,\ a_n$. Equation (3.4.1) then gives the required polynomial of the n-th degree.

For larger values of n, system (3.4.3) 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.

Example 3.4.1. Fit a polynomial of the second degree to the data points (x, y) given by

$$(0,\ 1),\ (1,\ 6),\ \text{and}\ (2,\ 17)$$

Solution: For $n = 2$, Eq.(3.4.3) 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 3.4.2. 1. Fit a second degree parabola $y = a_0 + a_1x + a_2x^2$ to the data (x_i, y_i):

$$(1\ ,\ 0.63), (3\ ,\ 2.05), (4\ ,\ 4.08), (6\ ,\ 10.78)$$

2. Fit a quadratic curve to the following data, and compute the value of variable y at point $x = 3$

x	0	1	2	4	5
y	-2	0	10	78	148

3. Obtain the least squares fit of the form $y = ax^2 + bx + c$ for the following data set.

x :	10	12	15	23	20
y :	14	17	23	25	21

Solve the system of normal equations with the aid of Gauss elimination method.

3.5 Weighted Least Square Approximation

In the previous section, 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.(3.2.2), we have

$$\begin{aligned} S &= W_1\Big[y_1 - f(x_1)\Big]^2 + W_2\Big[y_2 - f(x_2)\Big]^2 + \ldots + W_m\Big[y_m - f(x_m)\Big]^2 \\ &= W_1e_1^2 + W_2e_2^2 + \ldots + W_me_m^2. \end{aligned} \tag{3.5.1}$$

In Eq.(3.5.1), 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.

3.5.1 Linear Weighted Least Squares Approximation

Let $Y = a_0 + a_1x$ be the straight line to be fitted to the given data points, viz. $(x_1,\ y_1), \ldots, (x_m, y_m)$. Then

$$S(a_0, a_1) = \sum_{i=1}^{m} W_i\Big[y_i - (a_0 + a_1x_i)\Big]^2. \tag{3.5.2}$$

For maxima or minima, we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = 0, \quad \text{which gives} \tag{3.5.3}$$

$$\frac{\partial S}{\partial a_0} = -2\sum_{i=1}^{m} W_i\Big[y_i - (a_0 + a_1x_i)\Big] = 0 \quad \text{and} \quad \frac{\partial S}{\partial a_1} = -2\sum_{i=1}^{m} W_i\Big[y_i - (a_0 + a_1x_i)\Big]x_i = 0.$$

Simplifying yields the system of equations for a_0 and a_1:

$$a_0\sum_{i=1}^{m} W_i + a_1\sum_{i=1}^{m} W_ix_i = \sum_{i=1}^{m} W_iy_i \quad \text{and} \quad a_0\sum_{i=1}^{m} W_ix_i + a_1\sum_{i=1}^{m} W_ix_i^2 = \sum_{i=1}^{m} W_ix_iy_i \tag{3.5.4}$$

which are the *normal equations* in this case and are solved to obtain a_0 and a_1.

Example 3.5.1. Suppose that in the data of Exercise (3.3.3), 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.(3.5.4) 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 3.5.2. 1. Consider Example (3.5.1) 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.

3.5.2 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

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{3.5.5}$$

be fitted to the given data points. We then have

$$S(a_0, a_1, \ldots, a_n) = \sum_{i=1}^{m} W_i \Big[y_i - (a_0 + a_1 x_i + \ldots + a_n x_i^n) \Big]^2. \tag{3.5.6}$$

If a minimum occurs at $(a_0, a_1, \ldots, a_n)$, then we have

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \frac{\partial S}{\partial a_2} = \ldots = \frac{\partial S}{\partial a_n} = 0. \tag{3.5.7}$$

These conditions yield the normal equations

$$\begin{aligned}
a_0 \sum_{i=1}^{m} W_i + a_1 \sum_{i=1}^{m} W_i x_i + \ldots + a_n \sum_{i=1}^{m} W_i x_i^n &= \sum_{i=1}^{m} W_i y_i \\
a_0 \sum_{i=1}^{m} W_i x_i + a_1 \sum_{i=1}^{m} W_i x_i^2 + \ldots + a_n \sum_{i=1}^{m} W_i x_i^{n+1} &= \sum_{i=1}^{m} W_i x_i y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad & \qquad \vdots \\
a_0 \sum_{i=1}^{m} W_i x_i^n + a_1 \sum_{i=1}^{m} W_i x_i^{n+1} + \ldots + a_n \sum_{i=1}^{m} W_i x_i^{2n} &= \sum_{i=1}^{m} W_i x_i^n y_i.
\end{aligned} \tag{3.5.8}$$

Equations (3.5.8) 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.

Note 3.5.3. In general, the least squares curves do not pass through any data point. The least squares curves have global effects (if we change the position of one point, it will change the whole curve).

Interpolating polynomial is best suitable for a data set having less number of points, as it has zero least squares error. But, for large data set, we have already discussed the disadvantages of higher order polynomials. Least squares fitting are suitable for large data set having global patterns like a straight line, exponential, parabolic, etc. But selection of appropriate curve is very difficult task, as it is not possible to predict the suitable curve by just looking at the data set. Scatter diagram will be helpful in this regard.

Unit 4

Course Structure

- Orthogonal polynomials, Gram –Schmidt orthogonalisation process, Chebysev polynomials, Mini-max polynomial approximation.

4.1 Orthogonal Polynomial approximation method

In the previous unit, we considered the least squares approximations of discrete data. We shall, in the present unit, 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:

$$Y(x) = a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x), \tag{4.1.1}$$

where $f_j(x)$ is a polynomial in x of degree j. Then we write

$$S(a_0, a_1, \ldots, a_n) = \int_0^a W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big]^2 dx. \tag{4.1.2}$$

For S to be minimum, we must have

$$\begin{aligned}
\frac{\partial S}{\partial a_0} &= 0 = -2\int_a^b W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big] f_0(x)\, dx \\
\frac{\partial S}{\partial a_1} &= 0 = -2\int_a^b W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big] f_1(x)\, dx \qquad (4.1.3) \\
&\vdots \\
\frac{\partial S}{\partial a_n} &= 0 = -2\int_a^b W(x)\Big[y(x) - \Big\{a_0 f_0(x) + a_1 f_1(x) + \ldots + a_n f_n(x)\Big\}\Big] f_n(x)\, dx
\end{aligned}$$

The system of normal equations can be written as

$$\begin{aligned}
&a_0\int_a^b W(x) f_0(x) f_j(x)\, dx \;+\; a_1\int_a^b W(x) f_1(x) f_j(x)\, dx \;+\; \ldots\ldots \\
&+ a_n\int_a^b W(x) f_n(x) f_j(x)\, dx = \int_a^b W(x) y(x) f_j(x)\, dx, \qquad j = 0, 1, 2, \ldots, n. \qquad (4.1.4)
\end{aligned}$$

In Eq.(4.1.4), we find products of the type $f_p(x) f_q(x)$ in the integrands, and if we assume that

$$\int_a^b W(x) f_p(x) f_q(x)\, dx = \begin{cases} 0, & p \neq q \\ \int_a^b W(x) f_p^2(x)\, dx, & p = q, \end{cases} \qquad (4.1.5)$$

Hence from Eq.(4.1.4), we obtain

$$a_j = \left[\int_a^b W(x) y(x) f_j(x)\, dx\right] \Bigg/ \left[\int_a^b W(x) f_j^2(x)\, dx\right], \qquad j = 0, 1, 2, \ldots, n. \qquad (4.1.6)$$

Substitution of $a_0, a_1, \ldots, a_n$ in Eq.(4.1.1) 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 (4.1.5), 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.

4.2 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

$$f_0(x) = 1 \tag{4.2.1}$$

we find that the linear polynomial $f_1(x)$, with leading term x, can be written as

$$f_1(x) = x + k_{1,0} f_0(x), \tag{4.2.2}$$

where $k_{1,0}$ is a constant to be determined. Since $f_1(x)$ and $f_0(x)$ are orthogonal, we have

$$\int_a^b W(x) f_0(x) f_1(x)\, dx = 0 = \int_a^b x W(x) f_0(x)\, dx + k_{1,0} \int_a^b W(x) f_0^2(x)\, dx \quad \text{[using Eqs.(4.1.5) and (4.2.1)]}$$

Now from the above, we obtain

$$k_{1,0} = -\left[\int_a^b x W(x) f_0(x)\, dx\right] \Big/ \left[\int_a^b W(x) f_0^2(x)\, dx\right], \tag{4.2.3}$$

and Eq.(4.2.2) gives

$$f_1(x) = x - \left[\left[\int_a^b x W(x) f_0(x)\, dx\right] \Big/ \left[\int_a^b W(x) f_0^2(x)\, dx\right]\right] f_0(x) \tag{4.2.4}$$

Now, the polynomial $f_2(x)$, of degree 2 in x and with leading term x^2, may be written as

$$f_2(x) = x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x), \tag{4.2.5}$$

where the constants $k_{0,2}$ and $k_{2,1}$ are to be determined by using the orthogonality conditions in Eq.(4.1.5). Since $f_2(x)$ is orthogonal to $f_0(x)$, we have

$$\int_a^b W(x) f_0(x) \Big[x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x)\Big] dx = 0. \tag{4.2.6}$$

Since $\displaystyle\int_a^b W(x) f_0(x) f_1(x)\, dx = 0$, the above equation gives

$$k_{2,0} = -\left[\int_a^b x^2 W(x) f_0(x)\, dx\right] \Big/ \left[\int_a^b W(x) f_0^2(x)\, dx\right] = -\left[\int_a^b x^2 W(x)\, dx\right] \Big/ \left[\int_a^b W(x)\, dx\right], \tag{4.2.7}$$

Again, since $f_2(x)$ is orthogonal to $f_1(x)$, we have

$$\int_a^b W(x) f_1(x) \Big[x^2 + k_{2,0} f_0(x) + k_{2,1} f_1(x)\Big] dx = 0. \tag{4.2.8}$$

Using the condition that $\int_a^b W(x)f_0(x)f_1(x)\,dx = 0$, the above yields

$$k_{2,1} = -\left[\int_a^b x^2 W(x)f_1(x)\,dx\right] \Big/ \left[\int_a^b W(x)f_1^2(x)\,dx\right], \tag{4.2.9}$$

Since $k_{2,0}$ and $k_{2,1}$ are known, Eq.(4.2.5) determines $f_2(x)$. Proceeding in this way, the method can be generalized and we write

$$f_j(x) = x^j + k_{j,0}f_0(x) + k_{j,1}f_1(x) + \ldots + k_{j,j-1}f_{j-1}(x), \tag{4.2.10}$$

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

$$k_{j,i} = -\left[\int_a^b x^j W(x)f_i(x)\,dx\right] \Big/ \left[\int_a^b W(x)f_i^2(x)\,dx\right], \tag{4.2.11}$$

Since the a_i and $f_i(x)$ in Eq.(4.1.1) are known, the approximation $Y(x)$ can now be determined. The following example illustrates the method of procedure.

Example 4.2.1. 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.(4.2.3) gives

$$k_{1,0} = -\left[\int_{-1}^1 x\,dx\right] \Big/ \left[\int_{-1}^1 dx\right] = 0,$$

We then obtain from Eq.(4.2.2), $f_1(x) = x$. Equations (4.2.7) and (4.2.9) 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.(4.2.5) 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.(4.1.5).

4.3 Chebyshev Polynomials Approximation

In this section, we are mainly concerned with the approximation by Chebyshev polynomials. To begin, here we will briefly discuss some basic aspects of Chebyshev polynomials. Chebyshev differential equation of degree n is given by

$$\left(1-x^2\right)\frac{d^2y}{dx^2} - x\frac{dy}{dx} + n^2y = 0 \quad -1 \le x \le 1$$

Two linearly independent solutions of this differential equation are Chebyshev polynomial of the first kind $T_n(x) = \cos\left(n\cos^{-1}x\right)$ and Chebyshev polynomial of the second kind $U_n(x) = \sin\left(n\cos^{-1}x\right)$. Here, we will concentrate only on Chebyshev polynomials of the first kind $T_n(x)$ and use its minimax property to obtain best lower approximation for a given polynomial.

Now, we will discuss important forms, recurrence relation and orthogonal property of Chebyshev polynomials of the first kind $T_n(x)$ of degree n.

1. Forms of Chebyshev polynomial

$$T_n(x) = \cos\left(n\cos^{-1}x\right), \quad -1 \le x \le 1$$

On replacing $x = \cos(\theta)$ or $\theta = \cos^{-1}x$, we have

$$T_n(\cos\theta) = \cos(n\theta)$$

On using the de Moivre's formula $(\cos\theta \pm i\sin\theta)^n = \cos(n\theta) \pm i\sin(n\theta)$, we have

$$\begin{aligned} T_n(\cos\theta) &= \cos(n\theta) \\ &= \frac{1}{2}\left[(\cos\theta + i\sin\theta)^n + (\cos\theta - i\sin\theta)^n\right] \\ &= \frac{1}{2}\left[\left(\cos\theta + i\sqrt{1-\cos^2\theta}\right)^n + \left(\cos\theta - i\sqrt{1-\cos^2\theta}\right)^n\right] \\ &= \frac{1}{2}\left[\left(\cos\theta + \sqrt{\cos^2\theta - 1}\right)^n + \left(\cos\theta - \sqrt{\cos^2\theta - 1}\right)^n\right] \end{aligned}$$

Substitute $x = \cos(\theta)$ to get the following form

$$T_n(x) = \frac{1}{2}\left[\left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n\right]$$

2. Chebyshev polynomial in terms of Gauss hypergeometric function

If we put $x = 1-2t$ in Chebyshev differential equation, then it will transform into the following differential equation

$$\left(t-t^2\right)\frac{d^2y}{dt^2} + \left(\frac{1}{2} - t\right)\frac{dy}{dt} + n^2y = 0$$

This differential equation is equivalent to following Gauss hypergeometric equation $\left(t-t^2\right)\frac{d^2y}{dt^2} + (\gamma - (\alpha + \beta + 1)t)\frac{dy}{dt} - \alpha\beta y = 0$, with $\alpha = n, \beta = -n$ and $\gamma = \frac{1}{2}$ Therefore, the Chebyshev polynomial can also be written in terms of Gauss hypergeometric function $F(\alpha, \beta; \gamma; t)$ as follows

$$T_n(x) = F\left(n, -n; \frac{1}{2}; \frac{1-x}{2}\right)$$

3. Polynomial expansion for Chebyshev polynomials

De Moivre's formula is given by

$$\cos(n\theta) + i\sin(n\theta) = (\cos\theta + i\sin\theta)^n = \sum_{m=0}^{n} {}^nC_m(\cos\theta)^{n-m}(i\sin\theta)^m$$

In this expansion, real terms exist only for $m = 2k$ as

$$(i\sin\theta)^m = (i\sin\theta)^{2k} = (-1)^k\left(\sin^2\theta\right)^k = \left(\cos^2\theta - 1\right)^k$$

On equating real terms, we have

$$\cos(n\theta) = \sum_{k=0}^{[n/2]} {}^nC_{2k}(\cos\theta)^{n-2k}\left(\cos^2\theta - 1\right)^k$$

Use $T_n(\cos\theta) = \cos(n\theta)$ and $x = \cos(\theta)$ to get following form

$$T_n(x) = \sum_{k=0}^{[n/2]} \frac{n!}{2k!(n-2k)!}x^{n-2k}\left(x^2 - 1\right)^k$$

4. Recurrence relation for Chebyshev polynomials

Consider the following trigonometric identities

$$\cos(n\theta) = \cos(\theta + (n-1)\theta) = \cos(\theta)\cos((n-1)\theta) - \sin(\theta)\sin((n-1)\theta)$$

$$\cos((n-2)\theta) = \cos(-\theta + (n-1)\theta) = \cos(\theta)\cos((n-1)\theta) + \sin(\theta)\sin((n-1)\theta)$$

On adding these equations, we get

$$\cos(n\theta) + \cos((n-2)\theta) = 2\cos(\theta)\cos((n-1)\theta)$$

On using $T_n(\cos\theta) = \cos(n\theta)$ and $x = \cos(\theta)$, we have

$$\begin{aligned} T_n(x) + T_{n-2}(x) &= 2xT_{n-1}(x) \\ \text{(or) } T_n(x) &= 2xT_{n-1}(x) - T_{n-2}(x) \end{aligned} \tag{4.3.1}$$

Since, we have $T_n(x) = \cos\left(n\cos^{-1}x\right), \Rightarrow T_0(x) = 1, T_1(x) = x.$

Higher degree Chebyshev polynomials can be obtained using recurrence relation (4.3.1), and Chebyshev polynomials up to degree six are listed in following table

Chebyshev Polynomial in Power of x	Power of x in Chebyshev Polynomial
$T_0(x) = 1$	$1 = T_0(x)$
$T_1(x) = x$	$x = T_1(x)$
$T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1$	$x^2 = \frac{1}{2}\left(T_0(x) + T_2(x)\right)$
$T_3(x) = 2xT_2(x) - T_1(x) = 4x^3 - 3x$	$x^3 = \frac{1}{4}\left(3T_1(x) + T_3(x)\right)$
$T_4(x) = 8x^4 - 8x^2 + 1$	$x^4 = \frac{1}{8}\left(3T_0(x) + 4T_2(x) + T_4(x)\right)$
$T_5(x) = 16x^5 - 20x^3 + 5x$	$x^5 = \frac{1}{16}\left(10T_1(x) + 5T_3(x) + T_5(x)\right)$
$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1$	$x^6 = \frac{1}{32}\left(10T_0(x) + 15T_2(x) + 6T_4(x) + T_6(x)\right)$
⋮	⋮

5. Orthogonal property of Chebyshev polynomials

Chebyshev polynomials $T_n(x)$ are orthogonal w.r.t. weight function $\frac{1}{\sqrt{1-x^2}}$ over the interval $[-1, 1]$, i.e.

$$\int_{-1}^{1} \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}dx = 0; n \neq m \tag{4.3.2}$$

When $n = m$, we have

$$\int_{-1}^{1} \frac{T_n^2(x)}{\sqrt{1-x^2}}dx = \begin{cases} \pi & n = 0 \\ \frac{\pi}{2} & n \neq 0 \end{cases}$$

6. Minimax property of Chebyshev polynomials

One of the most important properties of Chebyshev polynomials is minimax property. We will use this property to obtain lower order polynomial approximation for a given polynomial.

The coefficient of x^n in the polynomial $T_n(x)$ is 2^{n-1}, therefore $2^{1-n}T_n(x)$ is a polynomial with coefficient of x^n is 1 . It means leading coefficient in polynomial $2^{1-n}T_n(x)$ is 1 . Since $T_n(x) = \cos\left(n\cos^{-1}x\right)$ so its maximum absolute value is 1 .

$$\max_{-1\leq x\leq 1} |T_n(x)| = 1$$

On using these facts, we can state the minimax property of Chebyshev polynomial as follows

Consider polynomials with leading coefficients 1 (known as a monic polynomial) and of degree $n > 0$, i.e.,

$$P_n(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0$$

Then, following relations hold in the domain $-1 \leq x \leq 1$,

$$\max_{-1\leq x\leq 1} |P_n(x)| \geq \max_{-1\leq x\leq 1} \left|2^{1-n}T_n(x)\right| = 2^{1-n} \tag{4.3.3}$$

The minimax property implies that among all the monic polynomials of degree n, the $2^{1-n}T_n(x)$ has smallest least upper bound for its absolute value in the domain $-1 \leq x \leq 1$. Thus if we approximate a given polynomial by lower order polynomial, then by Chebyshev polynomial we can minimize the maximum absolute error.

Example 4.3.1. Use Chebyshev polynomials to compute the best lower order approximation for the polynomial $3x^4 + 5x^3 - x + 1$ in the domian $-1 \leq x \leq 1$. Also, compute the error bound in this approximation.

Solution. First, we replace the highest order term in the polynomial with the help of Chebyshev polynomial as follows

$$\begin{aligned} 3x^4 + 5x^3 - x + 1 &= \frac{3}{8}\left(3T_0(x) + 4T_2(x) + T_4(x)\right) + 5x^3 - x + 1 \\ &= \frac{3}{8}T_4(x) + \frac{3}{2}T_2(x) + \frac{9}{8}T_0(x) + 5x^3 - x + 1 \\ &= \frac{3}{8}T_4(x) + \frac{3}{2}\left(2x^2 - 1\right) + \frac{9}{8} + 5x^3 - x + 1 \\ &= \frac{3}{8}T_4(x) + 5x^3 + 3x^2 - x + \frac{5}{8} \end{aligned}$$

On neglecting the term $\frac{3}{8}T_4(x)$, the lower order approximation is as follows

$$3x^4 + 5x^3 - x + 1 = 5x^3 + 3x^2 - x + \frac{5}{8}$$

The maximum absolute error (4.3.3) in this approximation is given by

$$\frac{3}{2^3}T_4(x) = \frac{3}{2^3} = \frac{3}{8} = 0.375$$

Note that if we directly neglect the term $3x^4$ from the given polynomial, then the maximum possible error in the interval $-1 \leq x \leq 1$ is 3.

Example 4.3.2. Economize the Taylor series expansion $\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + O\left(x^8\right)$ to lower order approximation over the interval $-1 \leq x \leq 1$. Also, compute the error bound.

Solution. On using the value of x^6 in terms of Chebyshev polynomials, we have

$$\begin{aligned} 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{1}{6!}\left(\frac{1}{32}\left(10T_0(x) + 15T_2(x) + 6T_4(x) + T_6(x)\right)\right) \\ &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{1}{6!}\left(\frac{1}{32}\left(10 + 15\left(2x^2 - 1\right) + 6\left(8x^4 - 8x^2 + 1\right) + T_6(x)\right)\right) \end{aligned}$$

On neglecting the term containing $T_6(x)$, we have

$$\begin{aligned} 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} &\approx 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{1}{6!}\left(\frac{1}{32}\left(10 + 15\left(2x^2 - 1\right) + 6\left(8x^4 - 8x^2 + 1\right)\right)\right) \\ &= \frac{23039}{23040} - \frac{639}{1280}x^2 + \frac{19}{480}x^4 \end{aligned}$$

This polynomial is the lower order economized approximation for the function $\cos x$. The error in this approximation is given by

$$\frac{1}{6!}2^{1-n}T_n(x) = \frac{1}{6!}\frac{1}{2^5}T_6(x) = \frac{1}{6!}\frac{1}{2^5} = 0.000043402777$$

Example 4.3.3. Approximate the polynomial $x^3 + 5x^2 + 2x - 1$ to a quadratic polynomial with minimum error in the interval $(3, 4)$.

Solution. To apply the Chebyshev approximation, first of all, we have to change the variable x to variable t, such that the interval converts from $(3, 4)$ to $(-1, 1)$. Let our new variable be $t = ax + b$. At $x = 3$ and $x = 4$, we want $t = -1$ and $t = 1$ respectively, i.e.,

$$\begin{aligned} -1 &= 3a + b \\ 1 &= 4a + b \end{aligned}$$

On solving these two equations for a and b and using these values, we have

$$t = 2x - 7 \text{ or } x = \frac{1}{2}(t + 7)$$

On using this expression for x in given polynomial, we have

$$\begin{aligned} x^3 + 5x^2 + 2x - 1 &= \left(\frac{1}{2}(t+7)\right)^3 + 5\left(\frac{1}{2}(t+7)\right)^2 + 2\left(\frac{1}{2}(t+7)\right) - 1 \\ &= \frac{1}{8}t^3 + \frac{31}{8}t^2 + \frac{295}{8}t + \frac{881}{8} \end{aligned}$$

Now, we have to convert $\frac{1}{8}t^3 + \frac{31}{8}t^2 + \frac{295}{8}t + \frac{881}{8}$ to a quadratic polynomial over the domain $(-1, 1)$.

$$\begin{aligned} \frac{1}{8}t^3 + \frac{31}{8}t^2 + \frac{295}{8}t + \frac{881}{8} &= \frac{1}{8}\left(\frac{1}{4}\left(3T_1(t) + T_3(t)\right)\right) + \frac{31}{8}t^2 + \frac{295}{8}t + \frac{881}{8} \\ &= \frac{1}{32}T_3(t) + \frac{3}{32}T_1(t) + \frac{31}{8}t^2 + \frac{295}{8}t + \frac{881}{8} \\ &= \frac{1}{32}T_3(t) + \frac{3}{32}t + \frac{31}{8}t^2 + \frac{295}{8}t + \frac{881}{8} \\ &= \frac{1}{32}T_3(t) + \frac{31}{8}t^2 + \frac{1183}{32}t + \frac{881}{8} \end{aligned}$$

Lower order approximation is given by

$$\frac{31}{8}t^2 + \frac{1183}{32}t + \frac{881}{8} \quad \text{over the interval } (-1, 1)$$

(or) $\frac{31}{8}(2x-7)^2 + \frac{1183}{32}(2x-7) + \frac{881}{8} = 15.5x^2 - 34.5625x + 41.21875$ over the interval $(3, 4)$.

Exercise 4.3.4.

1. 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 $[-1, 1]$.
2. Define an orthogonal set of functions and show that the set $f(x) = \sin\frac{n\pi x}{l},\ n = 1, 2, \ldots$ is orthogonal on $[0, l]$.
3. Economize Taylor series expansion $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + O(x^5)$ to lower order Chebyshev approximation over the interval $-1 \le x \le 1$.
4. Use the Chebyshev polynomials to obtain the approximations of second degree for the following polynomials

   (i) $2x^4 + 3x^3 - x + 2$ on $[-1, 1]$ (ii) $x^3 + 2x^2 - 5x + 3$ on $[2, 3]$

Unit 5

Course Structure

- Numerical Integration: Gaussian quadrature formula and its existence. Euler-MacLaurin formula

5.1 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

$$I = \int_a^b y \, dx. \tag{5.1.1}$$

Different integration formulae can be obtained depending upon the type of interpolation formula used.

5.2 Gaussian quadrature formula

In numerical integration, the value of integral, $\int_a^b f(x)\, dx$, depends on the values of function $f(x)$ at suitable number of points. It can be written as follows

$$I = \int_a^b f(x)\, dx \approx \sum_{i=0}^{n} \lambda_i f(x_i),$$

where $x_0, x_1, \cdots, x_n$ are $(n+1)$ node points in the interval $[a, b]$, and λ_i's are weights given to the values of function $f(x)$ at these node points. A polynomial of degree n is used for approximation to compute these $(n+1)$ weights λ_i's. Let, there be no restriction on the points x_i's also, then there are total $2n+2$ arbitrary constants [$(n+1)$ weights λ_i's and $(n+1)$ node points x_i's]. For these $2n+2$ constants, a polynomial of degree $2n+1$ can be utilized to approximate the function. So, higher accuracy can be achieved by these formulas. These methods are known as Gauss quadrature methods. Here, we will discuss Gauss quadrature methods based on Gauss-Legendre formula.

In this method, we assume the integral is of the form, $\int_{-1}^{1} f(x)dx$. Note that any definite integral, $\int_{a}^{b} g(x)dx$ can be converted to the form, $\int_{-1}^{1} f(x)dx$ by substituting following formula

$$x = \frac{b-a}{2}t + \frac{b+a}{2}$$

Let the function $f(x)$ in the integral $\int_{-1}^{1} f(x)\,dx$ be approximated by the following polynomial of degree $2n+1$

$$f(x) \approx a_0 + a_1x + a_2x^2 + \cdots + a_{2n+1}x^{2n+1}$$

The integral is approximated by following expression

$$\int_{-1}^{1} f(x)\,dx = \sum_{i=0}^{n} \lambda_i\, f(x_i) \tag{5.2.1}$$

L.H.S. and R.H.S. of Eq. (5.2.1) are as follows

$$\begin{aligned}
\text{L.H.S.} &= \int_{-1}^{1} f(x)dx = \int_{-1}^{1} \left(a_0 + a_1x + a_2x^2 + \cdots + a_{2n+1}x^{2n+1}\right) dx \\
&= 2a_0 + \frac{2}{3}a_2 + \frac{2}{5}a_4 + \cdots \\
\text{R.H.S.} &= \sum_{n=1}^{n} \lambda_1 f\left(x_1\right) = \lambda_0 \left(a_0 + a_1x_0 + a_2x_0^2 + \cdots + a_{2n+1}{x_0}^{2n+1}\right) \\
&\quad + \lambda_1 \left(a_0 + a_1x_1 + a_2{x_1}^2 + \cdots + a_{2n+1}x_1^{2n+1}\right) \\
&\quad + \lambda_2 \left(a_0 + a_1x_2 + a_2{x_2}^2 + \cdots + a_{2n+1}x_2^{2n+1}\right) \\
&\quad \vdots \\
&\quad + \lambda_n \left(a_0 + a_1x_n + a_2x_n^2 + \cdots + a_{2n+1}x_n^{2n+1}\right)
\end{aligned}$$

On comparing both sides, we get

$$\begin{aligned}
&\lambda_0 + \lambda_1 + \lambda_2 + \cdots + \lambda_{\mathrm{n}} = 2 \\
&\lambda_0x_0 + \lambda_1x_1 + \lambda_2x_2 + \cdots + \lambda_nx_n = 0 \\
&\lambda_0x_0^2 + \lambda_1x_1^2 + \lambda_2x_2^2 + \cdots + \lambda_nx_n^2 = \frac{2}{3} \\
&\vdots \\
&\lambda_0x_0^{2n+1} + \lambda_1x_1^{2n+1} + \lambda_2x_2^{2n+1} + \cdots + \lambda_nx_n^{2n+1} = 0
\end{aligned} \tag{5.2.2}$$

In general, it is very difficult to solve these $2n+2$ nonlinear equations. But fortunately, the values of x_1's are zeroes of Legendre orthogonal polynomials (discussed later in the chapter). Once the values of x_1's are known, we can use these values in the system (5.2.2). We will get linear system for λ_l's, which can be easily

solved. Here, we are considering only some particular cases.

1-Point Formula $(n = 0)$ For $n = 0$, we have following two equations from system (5.2.2)

$$\lambda_0 = 2$$
$$\lambda_0 x_0 = 0$$

Solution is $\lambda_0 = 2, x_0 = 0$.

On using these values in Eq. (5.2.1), we have following Gauss-Legendre 1-point formula

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) = 2f(0) \tag{5.2.3}$$

2-Points Formula ($n = 1$)

For $n = 1$, we have following four Eqs. (5.2.2)

$$\lambda_0 + \lambda_1 = 2 \tag{5.2.4}$$
$$\lambda_0 x_0 + \lambda_1 x_1 = 0 \tag{5.2.5}$$
$$\lambda_0 {x_0}^2 + \lambda_1 x_1^2 = \tfrac{2}{3} \tag{5.2.6}$$
$$\lambda_0 {x_0}^3 + \lambda_1 x_1^3 = 0 \tag{5.2.7}$$

Eq. (5.2.7) $-x_1^2$ Eq. (5.2.5) implies

$$\lambda_0 x_0 \left({x_0}^2 - x_1^2\right) = 0 \quad \text{or} \quad \lambda_0 x_0 (x_0 - x_1)(x_0 + x_1) = 0$$

Now if we select $\lambda_0 = 0$, or $x_0 = 0$, or $x_0 = x_1$, then remaining equations do not hold. Therefore, we have $x_0 = -x_1$. On using this in Eq. (5.2.5) and solving Eqs. (5.2.4) and (5.2.5) simultaneously, we have

$$\lambda_0 = \lambda_1 = 1$$

On substituting the values $x_0 = -x_1$ and $\lambda_0 = \lambda_1 = 1$ in Eq. (5.2.6), we get

$$x_0 = \frac{1}{\sqrt{3}}, x_1 = \frac{-1}{\sqrt{3}}$$

Equation (5.2.1) provides the following Gauss-Legendre 2-points formula

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right) \tag{5.2.8}$$

3-Points Formula

For $n = 2$, we have following six equations from system (5.2.2)

$$\lambda_0 + \lambda_1 + \lambda_2 = 2$$
$$\lambda_0 x_0 + \lambda_1 x_1 + \lambda_2 x_2 = 0$$
$$\lambda_0 x_0^2 + \lambda_1 x_1^2 + \lambda_2 x_2^2 = \frac{2}{3}$$
$$\lambda_0 x_0^3 + \lambda_1 x_1^3 + \lambda_2 x_2^3 = 0$$
$$\lambda_0 x_0^4 + \lambda_1 x_1^4 + \lambda_2 x_2^4 = \frac{2}{5}$$
$$\lambda_0 x_0^5 + \lambda_1 x_1^5 + \lambda_2 x_2^5 = 0$$

Here, we are presenting solution directly without giving any computational details.

$$x_0 = -\sqrt{\frac{3}{5}},\ x_1 = 0,\ x_2 = \sqrt{\frac{3}{5}} \qquad \lambda_0 = \frac{5}{9},\ \lambda_1 = \frac{8}{9},\ \lambda_2 = \frac{5}{9}$$

Gauss–Legendre 3-points formula is given by

$$\int_{-1}^{1} f(x)\,dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right) \tag{5.2.9}$$

Example 5.2.1. Compute the integral $\int_{-1}^{1} \frac{1}{1+x^2} dx$ with the help of Gauss-Legendre 1, 2 and 3-points formulas. Compare the results with exact value.

Solution.

i) Gauss-Legendre 1-point formula (5.2.3)

$$\int_{-1}^{1} f(x)dx = 2f(0)$$
$$\int_{-1}^{1} \frac{1}{1+x^2} dx = 2(1) = 2$$

ii) Gauss-Legendre 2-points formula (5.2.8)

$$\int_{-1}^{1} f(x)dx = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$
$$\int_{-1}^{1} \frac{1}{1+x^2} dx = \frac{1}{1+\left(\frac{-1}{\sqrt{3}}\right)^2} + \frac{1}{1+\left(\frac{1}{\sqrt{3}}\right)^2} = \frac{3}{2} = 1.5$$

iii) Gauss-Legendre 3-points formula (5.2.9)

$$\int_{-1}^{1} f(x)dx = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$$
$$\int_{-1}^{1} \frac{1}{1+x^2} dx = \frac{5}{9} \frac{1}{1+\left(-\sqrt{\frac{3}{5}}\right)^2} + \frac{8}{9} \frac{1}{1+0} + \frac{5}{9} \frac{1}{1+\left(\sqrt{\frac{3}{5}}\right)^2} = \frac{114}{72} = 1.58333$$

Exact solution is given by

$$\int_{-1}^{1} \frac{1}{1+x^2} dx = \tan^{-1}(x)\Big|_{-1}^{1} = \frac{\pi}{2} = 1.571$$

Hence, 3-points formula gives better approximation.

Example 5.2.2. Solve the integral $\int_{0}^{0.5} \exp(-x^2)\,dx$ numerically with the help of Gauss Legendre 3-point formula.

Solution. To convert the interval $[0, 0.5]$ in to interval $[-1, 1]$, the transformation is given by

$$x = \frac{b-a}{2}t + \frac{b+a}{2} = \frac{1}{4}t + \frac{1}{4}$$

On using this expression in the given integral, we get

$$\int_0^{0.5} \exp(-x^2)\, dx = \int_{-1}^{1} \exp\left(-\left(\frac{t}{4}+\frac{1}{4}\right)^2\right)\frac{1}{4}dt = \frac{1}{4}\left(\int_{-1}^{1} \exp\left(-\left(\frac{t}{4}+\frac{1}{4}\right)^2\right)dt\right)$$

On applying Gauss–Legendre 3-points formula, we have

$$\begin{aligned}
&\frac{1}{4}\left(\int_{-1}^{1} \exp\left(-\left(\frac{t}{4}+\frac{1}{4}\right)^2\right)dt\right) = \frac{1}{4}\left(\frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f\left(0\right) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right)\right)\\
&= \frac{1}{4}\left(\frac{5}{9}\left(\exp\left(-\left(\frac{1}{4}\left(-\frac{3}{5}\right)+\frac{1}{4}\right)^2\right)\right) + \frac{8}{9}\left(\exp\left(-\left(\frac{1}{4}\right)^2\right)\right) + \frac{5}{9}\left(\exp\left(-\left(\frac{1}{4}\left(\sqrt{\frac{3}{5}}\right)+\frac{1}{4}\right)^2\right)\right)\right)\\
&= \frac{1}{4}\left(\frac{5}{9}(0.9968296200) + \frac{8}{9}(0.9394130628) + \frac{5}{9}(0.8213346963)\right)\\
&= 0.4612812800
\end{aligned}$$

Computation of Weights and Nodes using Legendre Polynomials

Since our major concern is to compute the weights λ_i and nodes x_i for Gauss–Legendre integration $\int_{-1}^{1} f(x)dx = \sum_{t=0}^{n} \lambda_t f(x_t)$ with the help of Legendre polynomials, hence here we are discussing the various properties of Legendre polynomials in brief only.

1. Legendre polynomial $L_n(x)$ of degree n is solution of following second order differential equation

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

2. *Rodrigues formula:* The Legendre polynomials can be obtained using Rodrigues formula

$$L_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left((x^2-1)^n\right)$$

3. *Recurrence relation for Legendre polynomials*: We have following recurrence relation for Legendre polynomials

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - nL_{n-1}(x)$$

From Rodrigues formula, we can easily compute following Legendre polynomials

$$L_0(x) = 1 \text{ and } L_1(x) = x$$

On using the recurrence relation for $n = 1$, we have

$$(2)\ L_2(x) = (3)xL_1(x) - L_0(x) = 3x^2 - 1$$

$$L_2(x) = \frac{1}{2}\left(3x^2 - 1\right)$$

Similarly, the recurrence relation provides higher order Legendre polynomials for $n = 2, 3, 4...$ The Legendre polynomials up to order 6 are as follows $L_0(x) = 1, \quad L_1(x) = x, \quad L_2(x) = \frac{1}{2}\left(3x^2 - 1\right)$, $L_3(x) = \frac{1}{2}\left(5x^3 - 3x\right), \quad L_4(x) = \frac{1}{8}\left(35x^4 - 30x^2 + 3\right), L_5(x) = \frac{1}{8}\left(63x^5 - 70x^3 + 15x\right), \quad L_6(x) = \frac{1}{16}\left(231x^6 - 315x^4 + 105x^2 - 5\right), \cdots$

4. Orthogonal property of Legendre polynomials Here, without going in details, we will only state the orthogonal property of Legendre polynomials. Legendre polynomials $\mathrm{L}_n(x)$ are orthogonal over the interval $[-1, 1]$

$$\int_{-1}^{1} L_m(x)L_n(x)dx = 0; n \neq m$$

When $n = m$, we have

$$\int_{-1}^{1} L_n^2(x)dx = \frac{2}{2n+1}$$

Theorem 5.2.3. Let us consider that orthogonal polynomials with weight functions $w(x)$ over the interval $[a, b]$. If x_i's $i = 0, 1, 2, \cdots, n$ are zeroes of orthogonal polynomials, then the integral $\int_a^b w(x)f(x)dx = \sum_{i=0}^{n} \lambda_i f\left(x_i\right)$ is exact for polynomials of degree $\leq (2n + 1)$.

Proof. Let the function $f(x)$ be a polynomial of degree $\leq (2n+1)$. Let $P_n(x)$ be the interpolating polynomial of degree $\leq n$ which agrees $f(x)$ at $(n + 1)$ points

$$P_n\left(x_t\right) = f\left(x_t\right), i = 0, 1, 2, \cdots, n$$

Therefore, the function $f(x) - P_n(x)$ has $(n+1)$ zeroes $x_i, i = 0, 1, 2, \cdots, n$. Let $Q_{n+1}(x)$ be polynomial of degree $(n + 1)$ having zeroes x_i's.

We can write $f(x) - P_n(x)$ as product of two polynomials $Q_{n+1}(x)$ and $R_n(x)$, where $R_n(x)$ is a polynomial of degree at most n, i.e.,

$$f(x) - P_n(x) = Q_{n+1}(x)R_n(x) \tag{5.2.10}$$

On multiplying (5.2.10) with $w(x)$ and then integrating from a to b, we have

$$\int_a^b w(x)f(x)dx - \int_a^b w(x)P_n(x)dx = \int_a^b w(x)Q_{n+1}(x)R_n(x)dx$$

The integral on right hand side is zero, if the function $Q_{n+1}(x)$ is orthogonal over the interval $[a, b]$ with respect to weight function, $w(x)$, to all polynomials of degree $\leq n$. Then, we have

$$\int_a^b w(x)f(x)dx = \int_a^b w(x)P_n(x)dx \tag{5.2.11}$$

Consider the interpolating polynomial $P_n(x)$ of Lagrange form

$$P_n(x) = \sum_{i=1}^{n} f\left(x_i\right) l_i(x)$$

From Eq. (5.2.11), we have

$$\int_a^b w(x)f(x)dx = \int_a^b w(x)P_n(x)dx = \int_a^b w(x)\left(\sum_{i=1}^{n} f(x_i)\, l_i(x)\right)dx = \sum_{i=0}^{n} \lambda_i f(x_i) \tag{5.2.12}$$

where $\lambda_i = \int_a^b w(x)l_i(x)dx$ are the weights. As we start with the assumption, that $f(x)$ is a polynomial of degree $\leq (2n+1)$. It proves that the formula has an accuracy of $(2n+1)$ degree polynomial. □

Note 5.2.4. We prove that if x_t's are zeroes of orthogonal polynomials, then the integral $\int_a^b w(x)f(x)dx = \sum_{i=0}^{n} \lambda_i f(x_i)$ is exact for polynomials of degree $\leq (2n+1)$.

Now, Legendre polynomials are orthogonal with respect to weight function $w(x) = 1$ over the interval $[a, b] = [-1, 1]$. So, we will use Legendre polynomials for the calculation of the weights λ_1 and nodes x_1 for Gauss-Legendre integration $\int_{-1}^{1} f(x)dx = \sum_{i=0}^{n} \lambda_i f(x_i)$

1-Point Formula ($n = 0$)

The nodes x_t's are zeroes of orthogonal polynomials. For $n = 0$, we have Legendre polynomial, $L_1(x) = x$. Therefore, the node x_0 is zero of this polynomial, and it follows $x_0 = 0$. Weight $\lambda_1 = \int_a^b w(x)l_l(x)dx$ is given by $\lambda_0 = \int_{-1}^{1} 1dx = 2$ Hence $\int_{-1}^{1} f(x)dx = \sum_{i=0}^{n} \lambda_t f(x_i) = \lambda_0 f(x_0) = 2f(0)$

2-Points Formula ($n = 1$)

For $n = 1$, nodes are zeroes of the Legendre polynomial $L_2(x) = \frac{1}{2}\left(3x^2 - 1\right)$ Nodes are $x_0 = \frac{-1}{\sqrt{3}}, x_1 = \frac{1}{\sqrt{3}}$. The weights are given by the formula $\lambda_1 = \int_a^b w(x)l_1(x)dx$, we have

$$\lambda_0 = \int_{-1}^{1} l_0(x)dx = \int_{-1}^{1} \frac{x - x_1}{x_0 - x_1}dx = \frac{-\sqrt{3}}{2}\int_{-1}^{1}\left(x - \frac{1}{\sqrt{3}}\right)dx = 1$$
$$\lambda_1 = \int_{-1}^{1} l_1(x)dx = \int_{-1}^{1} \frac{x - x_0}{x_1 - x_0}dx = \frac{\sqrt{3}}{2}\int_{-1}^{1}\left(x + \frac{1}{\sqrt{3}}\right)dx = 1$$

So, Gauss-Legendre 2-points formula is given by

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) = f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

3-Points Formula $(n = 2)$

For $n = 2$, we have following equation for Legendre polynomial

$$L_3(x) = \frac{1}{2}\left(5x^3 - 3x\right) = 0$$

Nodes are $x_0 = -\sqrt{\frac{3}{5}}, x_1 = 0, x_2 = \sqrt{\frac{3}{5}}$. Weights are given by

$$\lambda_0 = \int_{-1}^{1} l_0(x)dx = \int_{-1}^{1} \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}dx = \frac{5}{6}\int_{-1}^{1} x\left(x-\sqrt{\frac{3}{5}}\right)dx = \frac{5}{9}$$
$$\lambda_1 = \int_{-1}^{1} l_1(x)dx = \int_{-1}^{1} \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}dx = \frac{-5}{3}\int_{-1}^{1} \left(x^2-\frac{3}{5}\right)dx = \frac{8}{9}$$
$$\lambda_2 = \int_{-1}^{1} l_2(x)dx = \int_{-1}^{1} \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}dx = \frac{5}{6}\int_{-1}^{1} x\left(x+\sqrt{\frac{3}{5}}\right)dx = \frac{5}{9}$$

Gauss-Legendre 3-points formula is given by

$$\int_{-1}^{1} f(x)dx = \lambda_0 f(x_0) + \lambda_1 f(x_1) + \lambda_2 f(x_2) = \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right)$$

5.3 Euler-MacLaurin Formula

Euler–Maclaurin formula is used to compute numerical quadrature and to approximate the sum of finite and infinite series. Let us derive Euler-Maclaurin formula with the help of Binomial expansion and shift operator $(Ef(x) = f(x+h))$.

$$\begin{aligned}
\frac{1}{E-1}f(x) &= \frac{1}{e^{nDD}-1}f(x) \\
&= \frac{1}{hD + \frac{(hD)^2}{2!} + \frac{(hD)^3}{3!} + \frac{(hD)^4}{4!} + \cdots} f(x) \quad (\text{using } E = e^{nD}) \\
&= \frac{1}{hD\left(1 + \frac{hD}{2} + \frac{(hD)^2}{6} + \frac{(hD)^3}{24} + \cdots\right)} \\
&= \frac{1}{hD(1+z)}f(x)
\end{aligned}$$

where $z = \frac{hD}{2} + \frac{(hD)^2}{6} + \frac{(hD)^3}{24} + \cdots$ On using the expression, $\frac{1}{1+z} = 1 - z + z^2 - z^3 + \cdots$, we have

$$\begin{aligned}
\frac{1}{E-1}f(x) &= \frac{1}{hD}\left(1 - z + z^2 - z^3 + \cdots\right)f(x) \\
&= \frac{1}{hD}\left(1 - \left(\frac{hD}{2} + \frac{(hD)^2}{6} + \frac{(hD)^3}{24}\right) + \left(\frac{hD}{2} + \frac{(hD)^2}{6} + \frac{(hD)^3}{24}\right)^2\right. \\
&\quad \left. - \left(\frac{hD}{2} + \frac{(hD)^2}{6} + \frac{(hD)^3}{24}\right)^3 + \cdots\right)f(x) \\
&= \frac{1}{hD}\left(1 - \frac{hD}{2} + \frac{(hD)^2}{12} - \frac{(hD)^4}{720} + \frac{(hD)^6}{30240} - \cdots\right)f(x) \\
\Rightarrow \frac{1}{E-1}f(x) &= \left(\frac{1}{hD} - \frac{1}{2} + \frac{hD}{12} - \frac{(hD)^3}{720} + \frac{(hD)^5}{30240} - \cdots\right)f(x) \qquad (5.3.1)
\end{aligned}$$

Consider the following expression

$$\frac{E^n - 1}{E-1}f(x_0) = \frac{1}{E-1}\left(\left(E^n - 1\right)f(x_0)\right) = \frac{1}{E-1}\left(f(x_n) - f(x_0)\right)$$

On using Eq. (5.3.1) in this expression, we get

$$\begin{aligned}\frac{E^n-1}{E-1}f(x_0) &= \left(\frac{1}{hD}-\frac{1}{2}+\frac{hD}{12}-\frac{(hD)^3}{720}+\frac{(hD)^5}{30240}-\cdots\right)(f(x_n)-f(x_0))\\ &= \frac{1}{hD}(f(x_n)-f(x_0))-\frac{1}{2}(f(x_n)-f(x_0))+\frac{hD}{12}(f(x_n)-f(x_0)) \quad (5.3.2)\\ &\quad -\frac{(hD)^3}{720}(f(x_n)-f(x_0))+\cdots\\ &= \frac{1}{h}\int_{x_0}^{x_n} f(x)dx-\frac{1}{2}(f(x_n)-f(x_0))+\frac{h}{12}\left(f'(x_n)-f'(x_0)\right)\\ &\quad -\frac{h^3}{720}\left(f'''(x_n)-f'''(x_0)\right)+\cdots \quad (5.3.3)\end{aligned}$$

Also, we have the following expression

$$\begin{aligned}\frac{E^n-1}{E-1}f(x_0) &= \left(1+E+E^2+\cdots+E^{n-1}\right)f(x_0)\\ &= f(x_0)+f(x_1)+f(x_2)+\cdots+f(x_{n-1})\\ &= \sum_{i=0}^{n-1} f(x_t) \quad (5.3.4)\end{aligned}$$

On equating Eq. (5.3.3) and Eq. (5.3.4) , we have

$$\begin{aligned}\sum_{t=0}^{n-1} f(x_t) = &\frac{1}{h}\int_{x_0}^{x_n} f(x)dx-\frac{1}{2}(f(x_n)-f(x_0))+\frac{h}{12}\left(f'(x_n)-f'(x_0)\right)\\ &-\frac{h^3}{720}\left(f'''(x_n)-f'''(x_0)\right)+\frac{h^5}{30240}\left(f^{(v)}(x_n)-f^{(v)}(x_0)\right)-\cdots\\ \sum_{t=0}^{n} f(x_t) = &\frac{1}{h}\int_{x_0}^{x_n} f(x)dx+\frac{1}{2}(f(x_n)+f(x_0))+\frac{h}{12}\left(f'(x_n)-f'(x_0)\right)\\ &-\frac{h^3}{720}\left(f'''(x_n)-f'''(x_0)\right)+\frac{h^5}{30240}\left(f^{(v)}(x_n)-f^{(v)}(x_0)\right)-\cdots \quad (5.3.5)\end{aligned}$$

Equation (5.3.5) can be used to compute the series expansion. But to compute the integral value, rewrite the Eq. (5.3.5) as follows

$$\begin{aligned}\int_{x_0}^{x_n} f(x)dx = &\ h\left(\sum_{i=1}^{n-1} f(x_i)+\frac{1}{2}(f(x_n)+f(x_0))\right)-\frac{h^2}{12}\left(f'(x_n)-f'(x_0)\right)\\ &+\frac{h^4}{720}\left(f'''(x_n)-f'''(x_0)\right)-\frac{h^6}{30240}\left(f^{(v)}(x_n)-f^{(v)}(x_0)\right)+\cdots \quad (5.3.6)\end{aligned}$$

Note 5.3.1. It is worth mentioning here that from the integral formula (5.3.6), we can easily derive composite Trapezoidal and Simpson $1/3$ rules. In formula (5.3.6), on neglecting all the derivative terms, we have

$$\int_{x_0}^{x_n} f(x)dx = h\left(\sum_{i=1}^{n-1} f(x_i)+\frac{1}{2}(f(x_n)+f(x_0))\right)$$

It is nothing but the composite Trapezoidal rule. Similarly, we can derive Simpson $1/3$ rule.

Example 5.3.2. Find the sum of cubes of first n natural numbers using Euler-Maclaurin formula.

Solution. Euler-Maclaurin formula for the sum of finite series (Eq. (5.3.5)) is as follows

$$\sum_{i=0}^{n} f(x_i) = \frac{1}{h}\int_{x_0}^{x_n} f(x)dx + \frac{1}{2}\left(f(x_n) + f(x_0)\right) + \frac{h}{12}\left(f'(x_n) - f'(x_0)\right) - \frac{h^3}{720}\left(f'''(x_n) - f'''(x_0)\right) + \frac{h^5}{30240}\left(f^{(v)}(x_n) - f^{(v)}(x_0)\right) - \cdots$$

To find the sum of cubes of first n natural numbers, let $f(x) = x^3$ with $x_0 = 0, x_n = n$ and $x_i = i, i = 0, 1, 2, \ldots, n$. We have $f'(x) = 3x^2, f'''(x) = 6$ and higher derivatives terms are zeroes. Also, the step size is $h = 1$. Now, using all these values in Eq. (13.35), we have

$$\sum_{i=0}^{n} x_i^3 = \frac{x^4}{4}\Big|_{x_0}^{x_n} + \frac{1}{2}\left(n^3 + 0\right) + \frac{1}{12}\left(3x_n^2 - 3x_0^2\right) - \frac{1}{720}(6-6)$$

$$\sum_{i=0}^{n} x_i^3 = \frac{n^4}{4} + \frac{1}{2}\left(n^3\right) + \frac{1}{12}\left(3n^2\right) = \left(\frac{n(n+1)}{2}\right)^2$$

Example 5.3.3. Use Euler-Maclaurin formula to prove that

$$\cos(0) + \cos\left(\frac{\pi}{100}\right) + \cos\left(\frac{2\pi}{100}\right) + \cdots + \cos(2\pi) = 1$$

Solution. We have to prove that

$$\sum_{i=0}^{200} \cos\left(\frac{i\pi}{100}\right) = 1$$

The function is $f(x) = \cos(x)$ with step size $h = \frac{\pi}{100}$. On using the Euler-Maclaurin formula (5.3.5), we have

$$\sum_{i=0}^{n} f(x_1) = \frac{1}{h}\int_{x_0}^{x_n} f(x)dx + \frac{1}{2}\left(f(x_n) + f(x_0)\right) + \frac{h}{12}\left(f'(x_n) - f'(x_0)\right) - \frac{h^3}{720}\left(f^{m\prime\prime}(x_n) - f^{m\prime}(x_0)\right) + \frac{h^5}{30240}\left(f^{(v)}(x_n) - f^{(v)}(x_0)\right) - \cdots$$

$$\sum_{i=0}^{200} \cos\left(\frac{i\pi}{100}\right) = \frac{100}{\pi}\int_{0}^{2\pi} \cos(x)dx + \frac{1}{2}(\cos(2\pi) + \cos(0)) + \frac{\pi}{1200}(\sin(2\pi) - \sin(0)) - \frac{h^3}{720}(-\sin(2\pi) + \sin(0)) + \frac{h^5}{30240}(\sin(2\pi) - \sin(0)) - \cdots$$

$$\sum_{i=0}^{200} \cos\left(\frac{i\pi}{100}\right) = \frac{1}{2}(1+1) = 1$$

Example 5.3.4. Use Euler-Maclaurin formula to compute the value of the Integral, $\int_1^2 e^{-x^2}dx$. Divide the Interval into ten equal parts and use up to third derivative terms only.

Soltuion. The spacing is $h = 0.1$, and we have to compute the function $f(x) = e^{-x^2}$ at 11 node points $x_0 = 1, x_1 = 1.1, x_2 = 1.2, x_3 = 1.3, \ldots, x_{10} = 2$

x	$f(x) = e^{-x^2}$
1	0.367879
1.1	0.298197
1.2	0.236928
1.3	0.18452
1.4	0.140858
1.5	0.105399
1.6	0.077305
1.7	0.055576
1.8	0.039164
1.9	0.027052
2	0.018316

For $n = 10$, Euler-Maclaurin formula (5.3.6) is given by

$$\begin{aligned}\int_{x_0}^{x_{10}} f(x)dx &= h\left(\sum_{t=1}^{g} f(x_i) + \tfrac{1}{2}\left(f(x_{10}) + f(x_0)\right)\right) - \tfrac{h^2}{12}\left(f'(x_{10}) - f'(x_0)\right) \\ &\quad + \tfrac{h^4}{720}\left(f'''(x_{10}) - f^{m\prime}(x_0)\right) - \tfrac{h^6}{30240}\left(f^{(v)}(x_n) - f^{(v)}(x_0)\right) + \cdots \end{aligned} \tag{5.3.7}$$

The derivative terms up to third order are as follows

$$\begin{aligned} f(x) &= e^{-x^2} \\ f'(x) &= -2xe^{-x^2} \quad f'(1) = -0.735758 \quad f'(2) = -0.073264 \\ f''(x) &= -2e^{-x^2} + 4x^2e^{-x^2} \\ f'''(x) &= 12xe^{-x^2} - 8x^3e^{-x^2} \quad f'''(1) = 1.471516 \quad f'''(2) = -0.732640 \end{aligned}$$

On using these values of derivative terms and $h = 0.1$ in Eq. (5.3.7), we have

$$\begin{aligned}\int_{x_0}^{x_0} f(x)dx =& (0.1)\left(\sum_{i=1}^{g} f(x_t) + \frac{1}{2}(f(2) + f(1))\right) - \frac{(0.1)^2}{12}\left(f'(2) - f'(1)\right) \\ &+ \frac{(0.1)^4}{720}\left(f'''(2) - f'''(1)\right) \\ \int_{1}^{2} f(x)dx =& (0.1)\left(1.164999 + \frac{1}{2}(0.018316 + 0.367879)\right) - \frac{(0.1)^2}{12}(-0.073264 - (-0.735758)) \\ &+ \frac{(0.1)^4}{720}(-0.732640 - 1.471516) \\ \int_{1}^{2} f(x)dx =& 0.135810 - 0.000552 - 0.0000003 = 0.1352577 \end{aligned}$$

Exercise 5.3.5. 1. Evaluate the following integrals by using Gauss–Legendre 2-points and 3-points formulas.

$$a)\ \int_1^2 (x^2 - \ln x)\, dx, \quad b)\ \int_1^2 x^2 e^{-x^2}\, dx, \quad c)\ \int_1^{\pi/2} \sqrt{1+\sin^2 x}\, dx, \quad d)\ \int_0^1 \frac{e^x}{1+\sin x}\, dx$$

2. Calculate the value of $\log_e 2$ from the integral $\int_0^1 \frac{dx}{1+x}$ by using Euler-Maclaurin formula.

3. Use Euler–Maclaurin formula to prove that

$$\sin(0) + \sin\left(\frac{\pi}{100}\right) + \sin\left(\frac{2\pi}{100}\right) + \cdots + + \sin(2\pi) = 0.$$

4. Prove the following results with the help of Euler-Maclaurin formula

$$a)\ \sum_{x=1}^{n} x = \frac{n(n+1)}{2}, \qquad b)\ \sum_{x=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}$$

5. Use Euler-Maclaurin formula to compute the value of the series $\sum_{x=1}^{100} \frac{1}{x}$. Use derivative terms up to order 5.

Unit 6

Course Structure

- Gregory-Newton quadrature formula, Romberg integration.

6.1 Gregory-Newton quadrature formula

We explain here Gregory-Newton formula by replacing the derivatives of $f(x)$ at $x = x_0 = a$ and $x = x_n = b$ by the corresponding forward difference formula and backward difference formula respectively.

We have

$$hy_0' = \Delta y_0 - \frac{1}{2}\Delta^2 y_0 + \frac{1}{3}\Delta^3 y_0 - \frac{1}{4}\Delta^4 y_0 + \frac{1}{5}\Delta^5 y_0 - \cdots$$
$$h^3 y_0''' = \Delta^3 y_0 - \frac{3}{2}\Delta^4 y_0 + \frac{7}{4}\Delta^5 y_0 - \cdots$$

and

$$hy_n' = \Delta y_{n-1} + \frac{1}{2}\Delta^2 y_{n-2} + \frac{1}{3}\Delta^3 y_{n-3} + \frac{1}{4}\Delta^4 y_{n-4} + \frac{1}{5}\Delta^5 y_{n-5} + \cdots$$
$$h^3 y_n''' = \Delta^3 y_{n-3} + \frac{3}{2}\Delta^4 y_{n-4} + \frac{7}{4}\Delta^5 y_{n-5} \cdots .$$

Then substituting the values in the Euler-Maclaurin Summation formula, we get

$$\begin{aligned}\int_a^b f(x)dx &= \frac{h}{2}\left[(y_0 + y_n) + 2\left(y_1 + y_2 + \ldots + y_{n-1}\right)\right] - \frac{1}{6}\left(\Delta y_{n-1} - \Delta y_0\right) \\ &\quad -\frac{1}{12}\left(\Delta^2 y_{n-2} - \Delta^2 y_0\right) - \frac{19}{360}\left(\Delta^3 y_{n-3} - \Delta^3 y_0\right) \\ &\quad -\frac{3}{720}\left(\Delta^4 y_{n-4} - \Delta^4 y_0\right) - \frac{863}{30240}\left(\Delta^5_{n-5} - \Delta^5 y_0\right) - \ldots\Big]\end{aligned}$$

This is known as Gregory-Newton Quadrature formula.

Example 6.1.1. Find the value of the integral $\int_{31^\circ}^{36^\circ} f(x)dx$ by Gregory-Newton formula where $f(x)$ is given by the following table.

x	31°	32°	33°	34°	35°	36°
$f(x)$	2.4913617	2.5051500	2.5185139	2.5314789	2.5440680	2.5563025

Solution : We first construct the following table.

x	$y = y(x)$	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
31°	2.4913617					
		0.0137883				
32°	2.5051500		- 0.0004244			
		0.0133639		0.0000255		
33°	2.5185139		- 0.0003939		-25×10^{-7}	
		0.0129650		0.0000230		8×10^{-7}
34°	2.5314789		- 0.0003759		-17×10^{-7}	
		0.0125891		0.0000213		
35°	2.5440680		- 0.0003646			
		0.0122345				
36°	2.5563025					

Now, $h = 1^\circ$.

Hence by Gregory-Newton quadrature formula gives,

$$\begin{aligned}
\int_{31^\circ}^{36^\circ} f(x)\,dx &= \frac{(0.017453292)}{2}\Bigg[2.4913617 + 2.5563025 + 2(2.5051500 + 2.5185139 + 2.5314789 \\
&\quad +2.5440680) - \frac{1}{6}(0.0122345 - 0.0137883) - \frac{1}{12}(-0.0003646 - 0.000422) \\
&\quad -\frac{19}{360}(0.0000213 - 0.0000255)\Bigg] \quad \text{[Ignoring the other terms]} \\
&= 0.008726646[25.321663] + 0.0002589667 + 0.00006575 + 0.000000222] \\
&= 0.220976024 \\
&\simeq 0.220976 \quad \text{[Correct upto 6D]}
\end{aligned}$$

Exercise 6.1.2. 1. Evaluate the following integrals by Gregory-Newton quadrature formula

$$(i)\ \int_0^{0.3}(1-8x^3)^{1/2}dx \quad (ii)\ \int_0^1 \sqrt{\sin x}dx \quad (iii)\ \int_2^3 \sqrt{(2x^2+1)(x^2-2)}dx$$

6.2 Richardson Extrapolation

Richardson extrapolation techniques are used to improve the order of numerical techniques. We consider suitable numerical method with different spacing to improve the accuracy of the method. Here, we will discuss Richardson extrapolation for numerical integration. Consider a numerical method for the value of the integral $I = \int_a^b f(x)dx$ with spacing h has an accuracy of order k. A method is said to be of order k if the

order of error term is $k+1$. Let the approximate value computed by this method be I_1. Hence, we can write the method as follows

$$I = I_1 + a_1 h^{k+1} + a_2 h^{k+2} + \cdots \tag{6.2.1}$$

where a_i's are the asymptotic error constants. Suppose we use the same method with spacing $\frac{h}{2}$ and computed value is I_2. Then, in that case, we have

$$I = I_2 + a_1 \left(\frac{h}{2}\right)^{k+1} + a_2 \left(\frac{h}{2}\right)^{k+2} + \cdots \tag{6.2.2}$$

To increase the order of method, multiply Eq. (6.2.2) with 2^{k+1} and then subtract it from Eq. (6.2.1), we get

$$I = \frac{\left(2^{k+1} I_2 - I_1\right)}{\left(2^{k+1} - 1\right)} + b_2 h^{k+2} + b_3 h^{k+3} + \cdots$$

The value of the integral is given by

$$I = \int_a^b f(x)dx = \frac{\left(2^{k+1} I_2 - I_1\right)}{\left(2^{k+1} - 1\right)} \tag{6.2.3}$$

This scheme is of order at least $k+1$. This process of finding higher-order formula from two different spacing is called Richardson extrapolation.

6.3 Romberg Integration

Romberg integration technique is an iterative technique. It uses repeated applications of Richardson extrapolation for numerical integration. In Romberg integration, we use a numerical method with different spacing to improve the accuracy of the method. The Richardson formula (6.2.3) is given by

$$I = \int_a^b f(x)dx = \frac{\left(2^{k+1} I_2 - I_1\right)}{\left(2^{k+1} - 1\right)}$$

where I_1 is numerical integration with spacing h and I_2 is numerical integration with spacing $h/2$. In Romberg integration, we will use Richardson scheme successively to obtain further higher order scheme.

For example, we will compute the integral with spacing $h,\ h/2,\ h/4,\ h/8, \cdots$ from any method like Trapezoidal or Simpson method. Let these values be $I_1^0,\ I_2^0,\ I_3^0,\ I_4^0, \cdots$. Here the subscript denotes the integration with different spacing (subscript is 1 for spacing h, subscript 2 for spacing $h/2$, so on), and superscript denotes the iteration number (the superscript 0 denotes initial approximation for Romberg integration).

We apply Richardson scheme for each set $(h, h/2), (h/2, h/4), (h/4, h/8), (h/8, h/16) \ldots$ to obtained the values of $I_1^1, I_2^1, I_3^1, I_4^1, \ldots$. Then, Richardson scheme is applied further by using these obtained values. This process is repeated till only one value is remained.

For easy understanding and to keep all these computations at one place, we can build a table of the form

$$\begin{array}{ccccc} I_1^0 & & & & \\ I_2^0 & I_2^1 & & & \\ I_3^0 & I_3^1 & I_3^2 & & \\ I_4^0 & I_4^1 & I_4^2 & I_4^3 & \\ I_5^0 & I_5^1 & I_5^2 & I_5^3 & I_5^4 \end{array}$$

Let us discuss this scheme for composite Trapezoidal and Simpson schemes for numerical integrations.

Trapezoidal Rule: The composite Trapezoidal scheme is given by

$$I = I_T + a_1 h^2 + a_2 h^4 + \cdots$$

Let the computed values for the integral with spacing $h, h/2, h/4, h/8, \ldots$ using composite Trapezoidal scheme be $I_1^0, I_2^0, I_3^0, I_4^0, \ldots$. The Richardson scheme for composite trapezoidal rule provides first iteration as follows

$$I_{k+1}^1 = \frac{\left(4I_{k+1}^0 - I_k^0\right)}{(4-1)} + b_2 h^4 + \cdots, k = 1, 2, 3, \cdots$$

This expression provides the values of the first approximations $\left(I_1^1, I_2^1, I_3^1, \ldots\right)$ of Romberg integration. We can further use these values to obtain the higher approximations. In general, the j-th iteration is given by

$$I_{k+1}^J = \frac{\left(4^j I_{k+1}^{j-1} - I_k^{j-1}\right)}{(4^J - 1)}, k = 1, 2, 3, \ldots \tag{6.3.1}$$

Simpson Rule: The composite Simpson scheme is given by

$$I = I_S + a_1 h^4 + a_2 h^6 + \cdots$$

Proceeding in a similar manner as in Trapezoidal method, the j-th iteration of Romberg integration for composite Simpson rule is given by

$$I_{k+1}^j = \frac{\left(4^{j+1} I_{k+1}^{j-1} - I_k^{j-1}\right)}{(4^{j+1} - 1)}, k = 1, 2, 3, \ldots \tag{6.3.2}$$

Example 6.3.1. Compute the value of integral $I = \int_0^1 \frac{1}{1+x} dx$ with the help of Romberg integration. Use only four Initial values of integral with the Trapezoldal rule.

Solution. First, we will compute the four initlal approximation to the integral $I = \int_0^1 \frac{1}{1+x} dx$ by using Trapezoldal rule with spacing $h = 1, h/2 = 0.5, h/4 = 0.25, h/8 = 0.125$. These values are listed in following Table.

$n = 1$ $h = 1$	$I_1^0 = \frac{h}{2}\left(f(0)+f(1)\right) = \frac{1}{2}\left(1+\frac{1}{2}\right) = 0.75$
$n = 2$ $h/2 = 0.5$	$I_2^0 = \frac{h}{2}\left(f(0)+2f(0.5)+f(1)\right) = \frac{0.5}{2}\left(1+2\left(\frac{1}{1.5}\right)+\frac{1}{2}\right) = 0.708333$
$n = 4$ $h/4 = 0.25$	$I_3^0 = \frac{h}{2}\left(f(0)+2\left(f(0.25)+f(0.5)+f(0.75)\right)+f(1)\right) = 0.697024$
$n = 8$ $h/8 = 0.125$	$I_4^0 = 0.694122$

The Romberg integration formula (6.3.1) is given by

$$I_{k+1}^{j} = \frac{\left(4^1 I_{k+1}^{j-1} - I_k^{-1}\right)}{(4^{-1} - 1)}, k = 1, 2, 3, \ldots.$$

On using this formula and inithal values from the table, we can easily compute the Iterations of Romberg integration. For the first iteration, we have

$$I_2^1 = \frac{\left(4I_2^0 - I_1^0\right)}{3} = 0.694444$$
$$I_3^1 = \frac{\left(4I_3^0 - I_2^0\right)}{3} = 0.693254$$
$$I_4^1 = \frac{\left(4I_4^0 - I_3^0\right)}{3} = 0.693155$$

Similarly, the second iteration is given by

$$I_3^2 = \frac{\left(16I_3^1 - I_2^1\right)}{15} = 0.693175$$
$$I_4^2 = \frac{\left(16I_4^1 - I_3^1\right)}{15} = 0.693148$$

The last iteration is as follows

$$I_4^3 = \frac{\left(64I_4^2 - I_3^2\right)}{63} = 0.693147$$

This value of the integral is correct up to 6 decimal places. In table form, we can list the iterations as follows

Spacing	Value of integral using Trapezoidal rule	1^{st} iteration of Romberg Integration	2^{nd} iteration of Romberg Integration	3^{rd} iteration of Romberg Integration
$h = 1$	0.750000			
$h/2 = 0.5$	0.708333	0.694444		
$h/4 = 0.25$	0.697024	0.693254	0.693175	
$h/8 = 0.125$	0.694122	0.693155	0.693148	0.693147

Note that last two iterations $(0.693148$ and $0.693147)$ matches upto five decimal points. So the result 0.693147 is at least correct upto five decimal places.

Example 6.3.2. Use Simpson formula to compute the value of integral $\int_1^2 e^{-x^2} dx$ with $n = 2, 4$ and 8 . Then use the Romberg integration to improve these values.

Solution. First, we will compute initial values with the help of Simpson formula as follows

$n = 2$	$I_1^0 = \frac{0.5}{2}(f(1) + 4f(1.5) + f(2)) = 0.134632$
$n = 4$	$I_2^0 = \frac{0.25}{2}(f(1) + 4f(1.25) + 4f(1.75) + 2f(1.5) + f(2)) = 0.135210$
$n = 8$	$I_3^0 = \frac{0.125}{2}\left(\begin{array}{r} f(1) + 4(f(1.125) + f(1.375) + f(1.625) + f(1.8755)) \\ +2(f(1.25) + f(1.5) + f(1.75)) + f(2) \end{array}\right) = 0.135254$

The following table contains the iterations of the Romberg integration (6.3.2)

Spacing	Values of integral using Simpson rule	1^{st} iteration of Romberg Integration	2^{nd} iteration of Romberg Integration
$h = 0.5$	0.134632		
$h/2 = 0.25$	0.135210	0.135249	
$h/4 = 0.125$	0.135254	0.135257	0.135257

Hence, the value of integral $\int_1^2 e^{-x^2} dx$ from Romberg integration is 0.135257. It is correct up to six decimal points. Note that the correct value of the integral up to ten decimal points is 0.1352572579.

Exercise 6.3.3.

1. Use Trapezoidal formula to compute the value of integral $I = \int_0^1 e^{-x^2} dx$, with $n = 1, 2$ and 4. Then use the Romberg integration to improve these values.
2. Use Romberg integration to compute $I = \int_0^{2\pi} \sin x \; dx$, correct to three decimal places.
3. Use Romberg integration to compute $I = \int_0^4 x^5 \; dx$, correct to three decimal places.
4. Compute the value of integration $I = \int_0^1 \frac{1}{1+x^2} dx$ with help of Romberg integration. Use only 4 initial values of integral with Trapezoidal rule.

Unit 7

Course Structure

- Systems of Linear Algebraic Equations: Direct methods - Factorization method.

7.1 Introduction

The systems of linear equations arise in the modeling of many physical and engineering problems. The linear system of equations with m equations in n variables $x_1, x_2, \ldots, x_n$, has the following form.

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

or, equivalently

$$\sum_{j=1}^{n} a_{ij}x_j = b_i; \quad 1 \leq i \leq m$$

The matrix form of the system is given by

$$AX = B \tag{7.1.1}$$

where

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

The matrix A is a coefficient matrix, and vector X is a solution vector. If each element of vector B is zero, then the system is called homogeneous system. Otherwise, it is a non-homogeneous system. For any homogeneous system, zero solution is always a solution, and it is also known as trivial solution. The system of linear equations may have a unique solution, an infinite number of solutions, or no solution.

In this unit, linear systems with unique solutions have been discussed. There are many direct and iterative methods for the solutions of such systems. Both types of methods have some advantages and disadvantages. It depends on the size and structure of the coefficient matrix A, available computer resources, and solution strategies adopted. This unit deals with direct LU – decomposition method/factorization method/tri-angularization method.

7.2 LU Decomposition (or) Factorization (or) Triangularization Method

In this method, the coefficient matrix A is factorized into the product of two triangular matrices such that one matrix is lower triangular L and the other matrix is upper triangular U, i.e.,

$$A = LU$$

where $L = \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \dots & 0 \\ \vdots & & & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$ and $U = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \vdots & & & \\ 0 & 0 & \dots & u_{nn} \end{bmatrix}$ are lower and upper triangular matrices, respectively. The matrices L and U have to be computed, such that

$$\begin{aligned} A = LU &= \begin{bmatrix} l_{11} & 0 & \dots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & & & \\ l_{\mathrm{n1}} & l_{n2} & \dots & l_{\mathrm{nn}} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & & & \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix} \\ &= \begin{bmatrix} l_{11}u_{11} & l_{11}u_{12} & \cdots & l_{11}u_{1n} \\ l_{21}u_{11} & l_{21}u_{12} + l_{22}u_{22} & \cdots & l_{21}u_{1n} + l_{22}u_{2n} \\ \vdots & & & \\ l_{n1}u_{11} & l_{n1}u_{12} + l_{n2}u_{22} & \cdots & l_{n1}u_{1n} + l_{n2}u_{2\pi} + \cdots + l_{nn}u_{nn} \end{bmatrix} \end{aligned} \tag{7.2.1}$$

After comparing the elements of both the matrices, we get the following relations

$$l_{n1}u_{1j} + l_{12}u_{2j} + \cdots + l_{2s}u_{nj} = a_{ij} \quad 1 \le i, j \le n$$

where $l_{ij} = 0, \quad j > i$ and $u_{ij} = 0, i > j$

This set contains n^2 equations. But, the total number of variables is $\left(n^2 + n\right)$ in lower and upper triangular matrices. So, we have to predefine n variables for a unique solution. For convenience, let us consider

$$\text{either } l_{\mathrm{ii}} = 1 \text{ (or) } u_{ii} = 1; 1 \le i \le n$$

Accordingly, we have following two methods

7.2.1 Doolittle Method

In this method, we will consider

$$l_{ii} = 1; 1 \le i \le n$$

7.2.2 Crout Method

In this method, we will consider

$$u_{ii} = 1; 1 \leq i \leq n$$

Here, we will discuss the computation of lower and upper triangular matrices in Crout method i.e., we have $u_{ii} = 1,\ 1 \leq i \leq n$. The similar procedure can be used in Doolittle method. We have

$$l_{i1}u_{1j} + l_{i2}u_{2j} + \cdots + l_{in}u_{nj} = a_{ij} \quad 1 \leq i, j \leq n$$

where $u_{ii} = 1,\ 1 \leq i \leq n;\ l_{ij} = 0, j > i$; and $u_{ij} = 0, i > j$. From (7.2.1) and $u_{11} = 1$; it is clear that the first columns of matrix L and A are identical. So, we have

$$l_{i1} = a_{i1}, 1 \leq i \leq n$$

The first rows of both the matrices in (7.2.1) produce the first row of matrix U as follows

$$u_{11} = 1 \quad \text{and} \quad u_{1j} = \frac{a_{1j}}{l_{11}}, 2 \leq j \leq n$$

Now, we will compute second column of matrix L and second row of matrix U as follows

$$l_{i2} = a_{i2} - l_{i1}u_{12}, 2 \leq i \leq n$$
$$u_{22} = 1 \text{ and } u_{2j} = \frac{a_{2j} - l_{21}u_{1j}}{l_{22}}, 3 \leq j \leq n$$

In general, we can compute k-th column and k-th row of matrices L and U, respectively by using following equations

$$l_{ik} = a_{ik} - \sum_{j=1}^{k-1} l_{ij}u_{jk}, k \leq i \leq n$$

$$u_{kk} = 1 \quad \text{and} \quad u_{kj} = \frac{a_{kj} - \sum_{m=1}^{k-1} l_{km}u_{mj}}{l_{kk}}, k+1 \leq j \leq n$$

After computing the matrices L and U, the system of equations is given by

$$AX = B$$

$$LUX = B$$

Let $UX = Y$, then the above system reduces to

$$LY = B$$

The system $LY = B$ is the lower triangular system. So, the vector Y can be easily determined by using forward substitution. The vector X can be easily computed by using back substitution from the following upper triangular system

$$UX = Y$$

Example 7.2.1. Use Crout and Doolittle methods to calculate the solution of the following system of linear equations

$$3x_1 - x_2 + x_3 = 1$$
$$2x_1 + 3x_2 + x_3 = 4$$
$$3x_1 + x_2 - 2x_3 = 6$$

Solution. **Crout method**: First, we decompose the coefficient matrix A into the product of lower and upper triangular matrices with diagonal elements in upper triangular matrix as unity, i.e.

$$\begin{bmatrix} 3 & -1 & 1 \\ 2 & 3 & 1 \\ 3 & 1 & -2 \end{bmatrix} = \underbrace{\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}}_{U}$$

$$= \begin{bmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{bmatrix}$$

After equating the terms on both sides, we obtain following set of equations

$$\begin{aligned} &l_{11} = 3, l_{11}u_{12} = -1, l_{11}u_{13} = 1 \\ &l_{21} = 2, l_{21}u_{12} + l_{22} = 3, l_{21}u_{13} + l_{22}u_{23} = 1 \\ &l_{31} = 3, l_{31}u_{12} + l_{32} = 1, l_{31}u_{13} + l_{32}u_{23} + l_{33} = -2 \end{aligned}$$

The solution of this system produces the values of l_{ij} and u_{ij} as follows

First Column: $l_{11} = 3, l_{21} = 2, l_{31} = 3$

First Row: $u_{12} = -1/l_{11} = -1/3$ and $u_{13} = 1/l_{11} = 1/3$

Second Column: $l_{22} = 3 - l_{21}u_{12} = \frac{11}{3}$ and $l_{32} = 1 - l_{31}u_{12} = 2$

Second Row: $u_{23} = (1 - l_{21}u_{13})/l_{22} = \frac{1}{11}$

Third Column: $l_{33} = -2 - l_{31}u_{13} + l_{32}u_{23} = \frac{-35}{11}$

So, we can easily write the coefficient matrix A in terms of the matrices L and U as follows

$$\begin{bmatrix} 3 & -1 & 1 \\ 2 & 3 & 1 \\ 3 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 2 & \frac{11}{3} & 0 \\ 3 & 2 & \frac{-35}{11} \end{bmatrix} \begin{bmatrix} 1 & \frac{-1}{3} & \frac{1}{3} \\ 0 & 1 & \frac{1}{11} \\ 0 & 0 & 1 \end{bmatrix}$$

The system $LY = B$ is given by

$$\begin{bmatrix} 3 & 0 & 0 \\ 2 & \frac{11}{3} & 0 \\ 3 & 2 & \frac{-35}{11} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 6 \end{bmatrix}$$

This system of equations can be rewritten as follows

$$\begin{aligned} 3y_1 + 0y_2 + 0y_3 &= 1 \\ 2y_1 + \frac{11}{3}y_2 + 0y_3 &= 4 \\ 3y_1 + 2y_2 - \frac{35}{11}y_3 &= 6 \end{aligned}$$

From the first equation, we get

$$y_1 = \frac{1}{3}$$

On substituting this value in the second equation, we have $y_2 = \dfrac{10}{11}$, and from the last equation $y_3 = -1$.

On using these values of y_1, y_2 and y_3 in the system

$$UX = Y$$

we have

$$\begin{bmatrix} 1 & \frac{-1}{3} & \frac{1}{3} \\ 0 & 1 & \frac{1}{11} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{10}{11} \\ -1 \end{bmatrix}$$

From the last equation $x_3 = -1$, using this value in the second equation $x_2 = 1$, and the first equation gives $x_1 = 1$. So, the solution is given by

$$x_1 = 1, x_2 = 1, x_3 =| -1$$

Doolittle method: First, we decompose the coefficient matrix A in the product of lower and upper triangular matrices with diagonal elements in the lower triangular matrix as unity.

$$A = LU = \begin{bmatrix} 3 & -1 & 1 \\ 2 & 3 & 1 \\ 3 & 1 & -2 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}}_{U}$$

Proceeding in a similar manner as in Crout method, we obtain

$$\begin{bmatrix} 3 & -1 & 1 \\ 2 & 3 & 1 \\ 3 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ 1 & \frac{6}{11} & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 & 1 \\ 0 & \frac{11}{3} & \frac{1}{3} \\ 0 & 0 & \frac{-35}{11} \end{bmatrix}$$

First, we solve $LY = B$ by using forward substitution

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ 1 & \frac{6}{11} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 6 \end{bmatrix}$$

The solution is, $y_1 = 1,\ y_2 = \dfrac{10}{3},\ y_3 = \dfrac{35}{11}$. Now, we solve $UX = Y$ by using backward substituions

$$\begin{bmatrix} 3 & -1 & 1 \\ 0 & \frac{11}{3} & \frac{1}{3} \\ 0 & 0 & -\frac{35}{11} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{10}{3} \\ \frac{35}{11} \end{bmatrix}$$

On solving this system of equations, final soltuion is given by

$$x_1 = x_2 = 1 \text{ and } x_3 = -1$$

Example 7.2.2. Solve the following system of linear equations with the help of LU-decomposition method

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

Solution. **Crout method**: The coefficient matrix A can be written as the product of lower and upper triangular matrices as follows

$$\begin{bmatrix} 3 & -3 & 1 \\ -2 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}$$

After equating the terms on both sides, we obtain

$$\begin{aligned} &l_{11} = 3, l_{11}u_{12} = -3, l_{11}u_{13} = 1 \\ &l_{21} = -2, l_{21}u_{12} + l_{22} = 2, l_{21}u_{13} + l_{22}u_{23} = 1 \\ &l_{31} = 1, l_{31}u_{12} + l_{32} = 1, l_{31}u_{13} + l_{32}u_{23} + l_{33} = 2 \end{aligned}$$

The solution of these equations is as follows

$$\begin{aligned} &l_{11} = 3, u_{12} = -1, u_{13} = \frac{1}{3} \\ &l_{21} = -2, l_{22} = 0 \end{aligned}$$

Since the element $l_{22} = 0$, so we cannot solve the equation $l_{21}u_{13} + l_{22}u_{23} = 1$ for the variable u_{23}.

The method fails as the element $l_{22} = 0$ is zero.

Similarly, in Doolittle method, the element $u_{22} = 0$ is zero, so method fails again.

Note 7.2.3. Rather, the system has a unique solution $x_1 = 1, x_2 = 0, x_3 = 1$, but LU-decomposition method does not work here. The first two rows of the system are a linear multiple of each other till the first two terms. Hence, the pivot element (l_{22} and u_{22}) becomes zero and the method fails. The solution can be obtained by interchanging any of first two rows with the third row.

So far, we have discussed the LU decomposition method for a general coefficient matrix, but if the coefficient matrix A is a positive definite symmetric matrix, then the method becomes simpler and is known as a Cholesky method.

A square matrix A is a positive definite symmetric matrix if it is symmetric and $X^TAX > 0$ for each nonzero column vector X.

Example 7.2.4. Prove that matrix $A = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ is positive definite symmetric matrix.

Solution. Let nonzero vector be $X = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. Then, we have

$$\begin{aligned} X^T AX &= \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\ &= \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} 3a - b + c \\ -a + 3b + c \\ a + b + 2c \end{bmatrix} \\ &= 3a^2 - ab + ac - ab + 3b^2 + bc + ac + bc + 2c^2 \\ &= \left(a^2 + b^2 - 2ab\right) + \left(a^2 + c^2 + 2ac\right) + \left(b^2 + c^2 + 2bc\right) + a^2 + b^2 \\ &= (a-b)^2 + (a+c)^2 + (b+c)^2 + a^2 + b^2 \end{aligned}$$

which is always positive for each nonzero vector X. The matrix A is symmetric matrix and $X^T AX > 0$ for each nonzero X. So, the matrix A is positive definite symmetric matrix.

Example 7.2.5. Show that the matrix $A = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ is not positive definite symmetric matrix.

Soltuion.

$$X^\top AX = (a-b)^2 + (a+c)^2 + (b+c)^2$$

The scalar $X^T AX$ can be zero for a, b, c such that $a = b = -c$. For example $X = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$. The matrix A is symmetric, but it does not satisfy $X^\top AX > 0$ for each nonzero X. Hence, the matrix A is not positive definite symmetric matrix. A symmetric matrix A is positive definite symmetric matrix, if any one of the following properties holds

1. All its eigenvalues are positive
2. All its leading principal minors are positive
3. $a_{ii} > 0$ and $a_{ii} > \sum_{j \neq i} |a_{ij}|$ for each i
4. All pivots are positive

7.3 Cholesky Method

In case of positive definite symmetric matrix A, there exists a unique decomposition of matrix A, known as Cholesky decomposition

$$A = LL^{\mathrm{T}} \tag{7.3.1}$$

where L is a lower triangular matrix and L^{T} is its transpose. Therefore, the system $AX = B$ can be written as follows

$$LL^\top X = B$$

Let $L^{\mathrm{T}} X = Y$, then

$$LY = B \tag{7.3.2}$$

First we compute vector Y using forward substitution from Eq. (7.3.2) and then compute vector X from the equation

$$L^{\mathrm{T}}X = Y \tag{7.3.3}$$

The matrix A can also be decomposed as $A = UU^{\tau}$, where U is an upper triangular matrix.

Example 7.3.1. Solve the following system of linear equations with the aid of Cholesky method

$$\begin{aligned} 3x_1 - x_2 + x_3 &= 2 \\ -x_1 + 3x_2 + x_3 &= 6 \\ x_1 + x_2 + 2x_3 &= 5 \end{aligned}$$

Solution. The matrix $A = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ is a positive definite matrix. Using Eq. (7.3.1) to decompose the matrix, we get

$$A = LL^{\mathrm{T}}$$

where L is a lower triangular matrix and L^{T} is the transpose of L, i.e.

$$\begin{bmatrix} 3 & -1 & 1 \\ -1 & 3 & 1 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

On comparing both sides and solving the resulting equations, we get matrix L as follows

$$L = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ \frac{-\sqrt{3}}{3} & \frac{2\sqrt{6}}{3} & 0 \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{3} & 1 \end{bmatrix}$$

The system $AX = B$ can be written as follows

$$LL^{\mathrm{T}}X = B$$

Let $L^{\mathrm{T}}X = Y$, then $LY = B$. Compute vector Y from the equation $LY = B$.

$$\begin{bmatrix} \sqrt{3} & 0 & 0 \\ \frac{-\sqrt{3}}{3} & \frac{2\sqrt{6}}{3} & 0 \\ \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{3} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 5 \end{bmatrix}$$

The solution of this system of equations is given by

$$y_1 = \frac{2\sqrt{3}}{3}, y_2 = \frac{5\sqrt{6}}{3}, y_3 = 1$$

On computing the vector X from the equation $L^T X = Y$, we have

$$\begin{bmatrix} \sqrt{3} & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & \frac{2\sqrt{6}}{3} & \frac{\sqrt{6}}{3} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{2\sqrt{3}}{3} \\ \frac{5\sqrt{6}}{3} \\ 1 \end{bmatrix}$$

On solving this system of equations by back substitution, we get the following solution

$$x_1 = 1,\ x_2 = 2,\ x_3 = 1$$

Exercise 7.3.2. 1. 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 $B = [4\ \ 8\ \ 10]^T$

2. Show that the following system cannot be solved with the help of LU-decomposition method

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

3. Solve the following system of linear equations by Cholesky method

$$\begin{aligned} 2x_1 + x_2 - x_3 &= 6 \\ x_1 - 3x_2 + 5x_3 &= 11 \\ -x_1 + 5x_2 + 4x_3 &= 13 \end{aligned}$$

Unit 8

Course Structure

- Eigenvalue and Eigenvector Problems: Direct methods, Iterative method –Power method.

8.1 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

$$AX = \lambda X \tag{8.1.1}$$

In Eq.(8.1.1), λ is called the *eigenvalue* and X is called the corresponding *eigenvector*. The matrix Eq.(8.1.1), when written out in full, represents a set of homogeneous linear equations:

$$\begin{aligned} (a_{11}-\lambda)x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22}-\lambda)x_2 + \ldots + a_{2n}x_n &= 0 \\ \vdots \quad \vdots \quad \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + (a_{nn}-\lambda)x_n &= 0. \end{aligned} \tag{8.1.2}$$

A nontrivial solution exists only when the coefficient determinant in (8.1.2) vanishes. Hence, we have

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} & \ldots & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{nn}-\lambda \end{vmatrix} = 0. \tag{8.1.3}$$

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.(8.1.3) 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

$$AX_i = \lambda_i X_i. \tag{8.1.4}$$

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.(8.1.3). This method, which is demonstrated in Example (8.2.1), 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.

8.2 Direct Method

In this section we will recall, how to calculate eigenvalues and eigenvector a matrix by direct method. Let us consider the following example.

Example 8.2.1. 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 $X_3 = [1/\sqrt{2} \;\; 0 \;\; 1/\sqrt{2}]^T$.

8.3 Iterative method

We have discussed a direct method for computing eigenvalues of a square matrix. The eigenvalues can be obtained by using this method, but for higher order matrix, expanding the characteristic determinant and obtaining roots from the high-degree characteristic equation are very difficult. Also in the direct methods, the errors committed will remain in final results. In the case of higher order matrices, the numbers of operational counts are large, and the error propagation will cause great damage to the results obtained. Consequently, we require iterative procedures for the solution of eigenvalue problems.

In next subsection, we will discuss following iterative procedure to compute eigenvalues and eigenvectors for a square matrix.

8.3.1 Power Method

Power method is used to determine the largest eigenvalue (in magnitude) of matrix A of order n. Let $\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_n$ be the eigenvalues of the matrix A, such that

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$$

The aim is to determine the absolutely largest eigenvalue (λ_1). Let $X_1, X_2, \ldots, X_n$ be the eigenvectors corresponding to the eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively. It implies

$$AX_i = \lambda_i X_i; \quad 1 \leq i \leq n$$

If the matrix A has n-linearly independent eigenvectors, then we can write any vector X (from same vector space) as a linear combination of the vectors, $X_1, X_2, \ldots, X_n$. Therefore, for some scalars, $c_i; 1 \leq i \leq n$, we have

$$X = c_1X_1 + c_2X_2 + \ldots + c_nX_n$$

Pre-multiplying Eq. (8.3.1) with the matrix A, we get

$$\begin{aligned} AX &= A\left(c_1X_1 + c_2X_2 + \ldots + c_nX_n\right) \\ &= c_1AX_1 + c_2AX_2 + \ldots + c_nAX_n \quad (c_i; 1 \leq i \leq n \text{ are scalars}) \\ &= c_1\lambda_1X_1 + c_2\lambda_2X_2 + \ldots + c_n\lambda_nX_n \quad (AX_i = \lambda X_i;\ 1 \leq i \leq n) \\ &= \lambda_1\left(c_1X_1 + c_2\frac{\lambda_2}{\lambda_1}X_2 + \ldots + c_n\frac{\lambda_n}{\lambda_1}X_n\right) \end{aligned}$$

Again, pre-multiplying with matrix A, we get

$$\begin{aligned} A^2X &= \lambda_1\left(c_1AX_1 + c_2\frac{\lambda_2}{\lambda_1}AX_2 + \ldots + c_n\frac{\lambda_n}{\lambda_1}AX_n\right) (c_i, \lambda_i \text{ are scalars}) \\ &= \lambda_1^2\left(c_1X_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^2X_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^2X_n\right) \end{aligned}$$

Repeating this process k-times successively, we obtain

$$A^kX = \lambda_1^k\left(c_1X_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^kX_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^kX_n\right)$$

$$A^{k+1}X = \lambda_1^{k+1}\left(c_1X_1 + c_2\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}X_2 + \ldots + c_n\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}X_n\right)$$

Since $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$, it implies

$$\lim_{k\to\infty}\left(\frac{\lambda_i}{\lambda_1}\right)^k \to 0; 2 \le i \le n$$
$$\Rightarrow \lim_{k\to\infty}\frac{A^{k+1}X}{A^kX} = \lambda_1$$

It provides the largest eigenvalue λ_1.

Theoretically, the method is as follows: first, we take any initial vector X, then we multiply it by matrix A infinitely many times ($k \to \infty$). At last, we divide the last two vectors. Practically, it is not possible to repeat the process infinite times. So, we can multiply the vector X as many times as feasible, for example 50 times. Then the common ratio $\frac{A^{51}X}{A^{50}X}$ is the largest eigenvalue. But, it can create the problem of rounding error, as the elements of the vector $A^{51}X$ become very large. Therefore, the method is applied by taking the largest element (magnitude) common at each iteration (to minimize the round-off error), and then continue with the remaining vector.

A stepwise procedure is as follows

1. Let $X^{(0)}$ be any non-zero initial vector.
2. Multiply $X^{(0)}$ with the matrix A to obtain the vector $Y^{(0)}$ i.e. $Y^{(0)} = AX^{(0)}$.
3. Take the absolutely largest element $\left(\lambda^{(1)}\right)$ common from the vector $Y^{(0)}$. Let remaining vector be $X^{(1)}$.
$$Y^{(0)} = \lambda^{(1)}X^{(1)}$$
4. Repeat steps ii) and iii) till the last iteration has the desired accuracy.
$$Y^{(k)} = \lambda^{(k+1)}X^{(k+1)} \quad k = 0, 1, 2, \ldots$$
5. At last, $\lambda^{(k+1)}$ and $X^{(k+1)}$ are the approximations to the largest eigenvalue and eigenvector, respectively. Note that we cannot start with trivial initial vector, i.e., zero vector $X_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$.

Example 8.3.1. Determine the largest eigenvalue and corresponding eigenvector of the matrix

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

Start with the initial vector, $X^{(0)} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$. Perform the iterations till the eigenvalue and eigenvector are same up to two decimal places, in last two iterations.

Solution. The first iteration of the Power method is given by

$$Y^{(0)} = AX^{(0)} = \begin{bmatrix} 0 & 2 & 4 \\ 1 & 1 & -2 \\ -2 & 0 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \\ 3 \end{bmatrix}$$

On scaling the vector $Y^{(0)}$ with the absolutely largest element, we have

$$Y^{(0)} = 6\begin{bmatrix} 1 & 0 & 0.5 \end{bmatrix} = \lambda^{(1)}X^{(1)}$$

Similarly, the second iteration is computed as follows

$$Y^{(1)} = AX^{(1)} = \begin{bmatrix} 0 & 2 & 4 \\ 1 & 1 & -2 \\ -2 & 0 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0.5 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \\ 0.25 \end{bmatrix}$$

$\lambda^{(2)} = 2$ and $X^{(2)} = \begin{bmatrix} 1 & 0 & 0.25 \end{bmatrix}^T$.

Proceeding in a similar manner, the subsequent iterations of the Power method are as follows

$$\begin{aligned}
\lambda^{(3)} &= 1.000000, & X^{(3)} &= [1.000000 \quad 0.500000 \quad 0.7500003]^T \\
\lambda^{(4)} &= 5.750000, & X^{(4)} &= [-0.347826 \quad 0.521739 \quad -1.000000]^T \\
\lambda^{(5)} &= 4.304348, & X^{(5)} &= [-0.686869 \quad 0.505050 \quad -1.000000]^T \\
\lambda^{(6)} &= 3.626263, & X^{(6)} &= [-0.824513 \quad 0.501393 \quad -1.000000]^T \\
\lambda^{(7)} &= 3.350975, & X^{(7)} &= [-0.894431 \quad 0.500416 \quad -1.000000]^T \\
\lambda^{(8)} &= 3.211139, & X^{(8)} &= [-0.933989 \quad 0.500129 \quad -1.000000]^T \\
\lambda^{(9)} &= 3.132022, & X^{(9)} &= [-0.957765 \quad 0.500041 \quad -1.000000]^T \\
\lambda^{(10)} &= 3.084470, & X^{(10)} &= [-0.972588 \quad 0.500013 \quad -1.000000]^T \\
\lambda^{(11)} &= 3.054824, & X^{(11)} &= [-0.982044 \quad 0.500004 \quad -1.000000]^T \\
\lambda^{(12)} &= 3.035911, & X^{(12)} &= [-0.988168 \quad 0.500001 \quad -1.000000]^T \\
\lambda^{(13)} &= 3.023663, & X^{(13)} &= [-0.992173 \quad 0.500013 \quad -1.000000]^T
\end{aligned}$$

The difference in the values at last two iterations (twelfth and thirteenth) are less than 0.005. Therefore, the approximate eigenvalue and eigenvector are $\lambda^{(13)} = 3.023663$ and $X^{(13)} = \begin{bmatrix} -0.992173 & 0.500000 & -1.000000 \end{bmatrix}$, respectively.

Using direct method, the exact eigenvalue is 3 and eigenvector is $\begin{bmatrix} -1 & 0.5 & -1 \end{bmatrix}^T$.

Note 8.3.2. The differences between the largest eigenvalue $\lambda_3 = 3$ and other eigenvalues $\lambda_1 = 1, \lambda_2 = 2$ are relatively less. Therefore, a large number of iterations are required for higher accuracy. Note that the power method has following restrictions.

1. The largest (in magnitude) eigenvalue of the matrix must be distinct.
2. The matrix A has n-linearly independent eigenvectors.
3. The rate of convergence is proportional to the ratio, $\dfrac{|\lambda_2|}{|\lambda_1|}$, where λ_2 is the second largest (in magnitude) eigenvalue and λ_1 is the largest (in magnitude) eigenvalue of the matrix A.

8.3.2 Inverse Power Method

The inverse power method is used to compute the smallest (in magnitude) eigenvalue of a given square matrix A. Inverse power method is a variation of power method. It involves computing of the largest (in magnitude) eigenvalue of the inverse matrix, A^{-1}.

Theorem 8.3.3. Let λ_i be an eigenvalue of matrix A, then $\dfrac{1}{\lambda_i}$ is the eigenvalue of the matrix A^{-1}. The eigenvector X_i of matrix A^{-1} remains same as that of matrix A.

Proof. Let λ_i be an eigenvalue and X_i is the corresponding eigenvector of matrix A, then we have

$$AX_i = \lambda_i X_i$$

On pre-multiplying with the matrix A^{-1}, we have

$$A^{-1}(AX_i) = A^{-1}(\lambda_i X_i) \tag{8.3.1}$$

The matrix multiplication is associative, so we have

$$A^{-1}(AX_i) = \left(A^{-1}A\right)X_i = IX_i = X_i \tag{8.3.2}$$

where the matrix I is the identity matrix. Also, the eigenvalue λ_i is scaler quantity, so

$$A^{-1}(\lambda_i X_i) = \lambda_i\left(A^{-1}X_i\right) \tag{8.3.3}$$

Equations (8.3.1)-(8.3.3) provide the following result

$$X_i = \lambda_i A^{-1}X_i$$
$$\frac{1}{\lambda_i}X_i = A^{-1}X_i$$

It implies that $\dfrac{1}{\lambda_i}$ is the eigenvalue of A^{-1}, the eigenvector X_i of matrix A^{-1} remains same as that of matrix A. □

To find the smallest (in magnitude) eigenvalue of the matrix A, we find the largest eigenvalue (in magnitude) of the matrix A^{-1}, and then the inverse of that eigenvalue is the smallest eigenvalue of the matrix A.

Example 8.3.4. Determine the smallest eigenvalue and the corresponding eigenvector of the matrix

$$A = \begin{bmatrix} 10 & 6 & 7 \\ 1 & 7 & -2 \\ 2 & 2 & 2 \end{bmatrix}$$

Solution. The inverse of matrix A is given by

$$A^{-1} = \frac{1}{60}\begin{bmatrix} 18 & 2 & -61 \\ -6 & 6 & 27 \\ -12 & -8 & 64 \end{bmatrix}$$

To compute the smallest (in magnitude) eigenvalue of matrix A, first we find the largest eigenvalue (in magnitude) of A^{-1}, and then inverse of that eigenvalue is the smallest eigenvalue of A.

To compute the largest eigenvalue of matrix, A^{-1}, let us start the iterations with initial vector $X^{(0)} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$. The first iteration of Power method is given by

$$Y^{(0)} = A^{-1}X^{(0)} = \frac{1}{60}\begin{bmatrix} 18 & 2 & -61 \\ -6 & 6 & 27 \\ -12 & -8 & 64 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{60}\begin{bmatrix} -41 \\ 27 \\ 44 \end{bmatrix} = \frac{44}{60}\begin{bmatrix} -0.931818 \\ 0.613636 \\ 1 \end{bmatrix}$$

$$\lambda^{(1)} = \frac{44}{60} \text{ and } X^{(1)} = \begin{bmatrix} -0.931818 & 0.613636 & 1 \end{bmatrix}^T$$

The second iteration is as follows

$$Y^{(1)} = A^{-1}X^{(1)} = \frac{1}{60}\begin{bmatrix} 18 & 2 & -61 \\ -6 & 6 & 27 \\ -12 & -8 & 64 \end{bmatrix}\begin{bmatrix} -0.931818 \\ 0.613636 \\ 1 \end{bmatrix} = \frac{76.545456}{60}\begin{bmatrix} -1 \\ 0.473872 \\ 0.918052 \end{bmatrix}$$

$$\lambda^{(2)} = \frac{76.545456}{60} \quad \text{and} \quad X^{(2)} = \begin{bmatrix} -1 & 0.473872 & 0.918052 \end{bmatrix}^T$$

Other iterations are given by

$$\begin{aligned}
\lambda^{(3)} &= 73.053444/60 & X^{(3)} &= \begin{bmatrix} -1.000000 & 0.460357 & 0.9166493 \end{bmatrix}^T \\
\lambda^{(4)} &= 72.994881/60 & X^{(4)} &= \begin{bmatrix} -1.000000 & 0.459096 & 0.9176354 \end{bmatrix}^T \\
\lambda^{(5)} &= 73.057564/60 & X^{(5)} &= \begin{bmatrix} -1.000000 & 0.458963 & 0.9178505 \end{bmatrix}^T \\
\lambda^{(6)} &= 73.070930/60 & X^{(6)} &= \begin{bmatrix} -1.000000 & 0.458948 & 0.9178856 \end{bmatrix}^T \\
\lambda^{(7)} &= 73.073082/60 & X^{(7)} &= \begin{bmatrix} -1.000000 & 0.458946 & 0.9178907 \end{bmatrix}^T \\
\lambda^{(8)} &= 73.073402/60 & X^{(8)} &= \begin{bmatrix} -1.000000 & 0.458945 & 0.9178918 \end{bmatrix}^T \\
\lambda^{(9)} &= 73.073441/60 & X^{(9)} &= \begin{bmatrix} -1.000000 & 0.458945 & 0.9178919 \end{bmatrix}^T \\
\lambda^{(10)} &= 73.073448/60 & X^{(10)} &= \begin{bmatrix} -1.000000 & 0.458945 & 0.9178911 \end{bmatrix}^T
\end{aligned}$$

The approximate value of the largest eigenvalue of A^{-1} is $\lambda^{(10)} = 73.073448/60 = 1.2178908$
Hence, the smallest eigenvalue of A is $1/1.2178908 = 0.8210916775$.

8.3.3 Shifted Power Method

Shifted power method is another variation of power method. It is used to compute the eigenvalues which are farthest/nearest from a given scalar k.

Theorem 8.3.5. Let λ_i be an eigenvalue of matrix A, then $(\lambda_i - k)$ is an eigenvalue of the matrix $(A - kI)$ with the same eigenvector as that of matrix A.

Proof. Let λ_i be an eigenvalue and X_i is the corresponding eigenvector of matrix A, then we have

$$AX_i = \lambda_i X_i$$

To compute eigenvalues of $(A - kI)$, we have

$$\begin{aligned}
(A - kI)X_i &= AX_i - kIX_i \\
&= \lambda_i X_i - kX_i \\
&= (\lambda_i - k)\, X_i
\end{aligned}$$

It implies that if λ_i is an eigenvalue of matrix A, then $(\lambda_i - k)$ is an eigenvalue of a matrix, $(A - kI)$. The vector X_i is the corresponding eigenvector of matrix A as well as $(A - kI)$. □

Eigenvalue farthest to a given scalar: To compute eigenvalue of matrix A farthest to a given number k, first we find the largest eigenvalue (in magnitude) of the matrix, $(A - kI)$, and then that eigenvalue in addition with k is the desired eigenvalue of matrix A.

For example, let us assume the eigenvalues of a matrix A are $-5, 2$ and 8 and we want to compute the eigenvalue that is farthest from the scalar 5 . The eigenvalues of the matrix $(A - 5\,\mathrm{I})$ are $-10, -3$ and 3 . The computational procedure is to compute the largest (in magnitude) eigenvalue of a matrix $(A - 5I)$ (i.e. -10),

and then add scalar 5 to that eigenvalue to get the desired eigenvalue (i.e. -5).

Eigenvalue nearest to a given scalar: To compute eigenvalue of matrix A nearest to number k, first, we find the largest eigenvalue (in magnitude) of matrix $(A-kI)^{-1}$, and then inverse of that eigenvalue in addition with k is the desired eigenvalue of matrix A.

For example, let us assume the eigenvalues of a matrix A are $-1, 4.5$ and 7 and we want to compute the eigenvalue that is nearest to 4 . We have

Eigenvalues of matrix A are $-1, 4.5$ and 7

Eigenvalues of matrix $(A-4I)$ are $-5, 0.5$ and 3

Eigenvalues of matrix $(A-4I)^{-1}$ are $\frac{-1}{5}, 2$ and $\frac{1}{3}$

The computational procedure is to compute the largest (in magnitude) eigenvalue of a matrix $(A-4I)^{-1}$ (i.e. 2). Then reciprocal (0.5) of that eigenvalue in addition with $k(=4)$ is the desired eigenvalue (4.5) of matrix A.

Example 8.3.6. Determine the eigenvalue farthest to 4 for the matrix

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

Start the iterations with the initial vector $X^{(0)} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T$.

Solution. To compute the eigenvalue of matrix A which is farthest to 4 , we will find the largest (in magnitude) eigenvalue of the matrix $(A-4I)$, and then add scalar 4 to that eigenvalue.

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

Proceeding in a similar manner as in previous examples with an initial vector $X^{(0)} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T$ for matrix $A-4I$, the largest eigenvalue of this matrix is computed as follows

$$Y^{(0)} = (A-4I)X^{(0)} = \begin{bmatrix} -2 & 6 & -3 \\ 5 & -1 & -3 \\ 5 & -4 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -5 \\ 1 \\ 1 \end{bmatrix} = 5 \begin{bmatrix} -1 \\ 0.4 \\ 1 \end{bmatrix}$$

$\lambda^{(1)} = 5$ and $X^{(1)} = \begin{bmatrix} -1 & 0.4 & 1 \end{bmatrix}^T$

Other iterations are given by $\lambda^{(2)} = 8.4 \quad X^{(2)} = \begin{bmatrix} 0.166667 & -1.000000 & -0.7857142 \end{bmatrix}^T$

$\lambda^{(3)} = 4.833333 \quad X^{(3)} = \begin{bmatrix} -0.822660 & 0.866995 & 1.000000 \end{bmatrix}^T$

$\lambda^{(4)} = 7.980295 \quad X^{(4)} = \begin{bmatrix} 0.482099 & -1.000000 & -0.9500004 \end{bmatrix}^T$

$$\lambda^{(5)} = 6.410494 \quad X^{(5)} = \begin{bmatrix} -0.641791 & 0.976601 & 1.000000 \end{bmatrix}^T$$

$$\lambda^{(6)} = 7.185556 \quad X^{(10)} = \begin{bmatrix} 0.576599 & -1.000000 & -0.9902316 \end{bmatrix}^T$$

$$\lambda^{(7)} = 6.882997 \quad X^{(7)} = \begin{bmatrix} -0.607658 & 0.995742 & 1.000000 \end{bmatrix}^T$$

$$\lambda^{(8)} = 7.034031 \quad X^{(8)} = \begin{bmatrix} 0.595643 & -1.000000 & -0.9981848 \end{bmatrix}^T$$

$$\lambda^{(9)} = 6.978212 \quad X^{(9)} = \begin{bmatrix} -0.601405 & 0.999219 & 1.000000 \end{bmatrix}^T$$

$$\lambda^{(10)} = 7.006245 \quad X^{(10)} = \begin{bmatrix} 0.599198 & -1.000000 & -0.99966610 \end{bmatrix}^T$$

These iterations are converging to the eigenvalue 7. Since the elements of eigenvectors are changing the sign alternatively, so the eigenvalue is –7. The largest eigenvalue of matrix $(A - 4I)$ is -7, so the eigenvalue of matrix A is $-7 + 4 = -3$. The eigenvalue $\lambda = -3$ of matrix A is farthest from scalar 4.

Note that the eigenvalues of matrix A are $-3, 5$ and 7.

Example 8.3.7. Determine the eigenvalue nearest to 5 and the corresponding eigenvector of the matrix

$$A = \begin{bmatrix} 10 & 6 & 7 \\ 1 & 7 & -2 \\ 2 & 2 & 2 \end{bmatrix}.$$

Solution. First, we find the largest eigenvalue (in magnitude) of the matrix $(A - 5I)^{-1}$. Then, by adding number 5 to the inverse of that eigenvalue will produce an eigenvalue of the matrix A (nearest to number 5).

$$A - 5I = \begin{bmatrix} 5 & 6 & 7 \\ 1 & 2 & -2 \\ 2 & 2 & -3 \end{bmatrix}$$

$$(A - 5I)^{-1} = \frac{1}{30} \begin{bmatrix} 2 & -32 & 26 \\ 1 & 29 & -17 \\ 2 & -2 & -4 \end{bmatrix}$$

Now, we have to compute the largest eigenvalue of the matrix, $(A - 5I)^{-1}$. Let us start the iterations with the initial vector, $X^{(0)} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$. The first iteration is given by

$$Y^{(0)} = (A - 5I)^{-1} X^{(0)} = \frac{1}{30} \begin{bmatrix} 2 & -32 & 26 \\ 1 & 29 & -17 \\ 2 & -2 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{30} \begin{bmatrix} -4 \\ 13 \\ -4 \end{bmatrix} = \frac{13}{30} \begin{bmatrix} -0.307692 \\ 1 \\ -0.307692 \end{bmatrix}$$

Similarly, other iterations are as follows

$$\begin{aligned} \lambda^{(2)} &= 40.615383/30 \quad X^{(2)} = \begin{bmatrix} -1.000000 & 0.835227 & -0.034091 \end{bmatrix}^T \\ \lambda^{(3)} &= 29.613638/30 \quad X^{(3)} = \begin{bmatrix} -1.000000 & 0.835227 & -0.034091 \end{bmatrix}^T \\ \lambda^{(4)} &= 30.821949/30 \quad X^{(4)} = \begin{bmatrix} -1.0000000 & 0.789591 & -0.101554 \end{bmatrix}^T \\ \lambda^{(5)} &= 29.907299/30 \quad X^{(5)} = \begin{bmatrix} -1.000000 & 0.789926 & -0.106093 \end{bmatrix}^T \end{aligned}$$

The largest eigenvalue of the matrix $(A - 5I)^{-1}$ is $\dfrac{29.907299}{30} \approx 1$

The smallest eigenvalue of the matrix $(A - 5I)$ is $\frac{1}{1} = 1$

The eigenvalue of the matrix A nearest to 5 is $1 + 5 = 6$.

Exercise 8.3.8. 1. Determine the dominant eigenvalue of $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ by power method.

2. Determine the largest eigenvalue in magnitude and corresponding eigenvector of the following matrix by power method.

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

3. Using power method, find the largest eigenvalue and the corresponding eigenvector of the following matrices with initial approximation $[1, 1, 1]^t$. Perform only five iterations.

$$(i) \begin{bmatrix} 1 & 1 & -1 \\ 3 & 2 & 4 \\ -1 & 4 & 2 \end{bmatrix} \qquad (i) \begin{bmatrix} -9 & 2 & 6 \\ 5 & 0 & -3 \\ -16 & 4 & 11 \end{bmatrix}$$

Unit 9

Course Structure

- Nonlinear Equations: Fixed point iteration method, convergence and error estimation, Modified Newton-Raphson method, Muller's method.

9.1 Introduction

In scientific and engineering studies, a frequently occurring problem is to find the roots of equations of the form

$$f(x) = 0 \tag{9.1.1}$$

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.

9.2 Fixed point iteration method

In order to describe the method, we first rewrite the Eq. (9.1.1) in the form

$$x = \phi(x) \tag{9.2.1}$$

Now, let x_0 be an approximate root of Eq. (9.2.1). Then, substituting in Eq. (9.2.1), 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

$$x_{n+1} = \phi(x_n) \tag{9.2.2}$$

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

$$\xi = \phi(\xi), \tag{9.2.3}$$

which shows that ξ is a root of the equation $x = \phi(x)$.

Condition of Convergence

To establish the condition of convergence of Eq. (9.2.1), we proceed in the following way:

From Eq. (9.2.2), we have

$$x_1 = \phi(x_0) \tag{9.2.4}$$

From Eqs. (9.2.3) and Eq. (9.2.4), we get

$$\xi - x_1 = \phi(\xi) - \phi(x_0) = (\xi - x_0)\phi'(\xi_0), \quad x_0 < \xi_0 < \xi, \tag{9.2.5}$$

Similarly, we obtain

$$\begin{aligned}
\xi - x_2 &= (\xi - x_1)\phi'(\xi_1), \quad x_1 < \xi_1 < \xi && (9.2.6)\\
\xi - x_3 &= (\xi - x_2)\phi'(\xi_2), \quad x_2 < \xi_2 < \xi && (9.2.7)\\
&\vdots \\
\xi - x_{n+1} &= (\xi - x_n)\phi'(\xi_n), \quad x_n < \xi_n < \xi && (9.2.8)
\end{aligned}$$

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

$$|\xi - x_{n+1}| \leq k^{n+1}|\xi - x_0| \tag{9.2.9}$$

If $k < 1$, i.e., if $|\phi'(\xi_i)| < 1$, then the right side of Eq. (9.2.9) 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.*

Error Estimation

We shall fin the error in the root obtained. We have

$$\begin{aligned} |\xi - x_n| &\leq k|\xi - x_{n-1}| \\ \Rightarrow |\xi - x_n| &= k|\xi - x_n + x_n - x_{n-1}| \\ \Rightarrow |\xi - x_n| &\leq k[|\xi - x_n| + |x_n - x_{n-1}|] \\ \Rightarrow |\xi - x_n| &\leq \frac{k}{1-k}|x_n - x_{n-1}| = \frac{k}{1-k}k^{n-1}|x_1 - x_0| = \frac{k^n}{1-k}|x_1 - x_0|, \end{aligned} \tag{9.2.10}$$

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. (9.2.10) gives

$$|x_n - x_{n-1}| \leq \frac{1-k}{k}\varepsilon, \tag{9.2.11}$$

which can be used to find the difference between two successive approximation (or iterations) to achieve a prescribed accuracy. From (9.2.11), it is clear that the rate of convergence of the fixed point iteration method is linear.

9.3 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

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{9.3.1}$$

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. (9.3.1) is taken as

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)} \tag{9.3.2}$$

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. (9.3.2) 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.

Order of convergence : Simple Root

Consider the Newton-Raphson method (9.3.1) 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

$$\varepsilon_{n+1} = \varepsilon_n + \frac{f(\xi) - \varepsilon_n f'(\xi) + \frac{1}{2!}\varepsilon_n^2 f''(\xi) - \frac{1}{3!}\varepsilon_n^3 f'''(\xi) + \cdots}{f'(\xi) - \varepsilon_n f''(\xi) + \frac{1}{2!}\varepsilon_n^2 f'''(\xi) - \frac{1}{3!}\varepsilon_n^3 f^{iv}(\xi) + \cdots} \tag{9.3.3}$$

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.(9.3.3) with $f'(\xi)$, we get

$$\begin{aligned}
\varepsilon_{n+1} &= \varepsilon_n + \frac{-\varepsilon_n + \frac{1}{2!}\varepsilon_n^2 \frac{f''(\xi)}{f'(\xi)} - \frac{1}{3!}\varepsilon_n^3 \frac{f'''(\xi)}{f'(\xi)} + \cdots}{1 - \left(\varepsilon_n \frac{f''(\xi)}{f'(\xi)} - \frac{1}{2!}\varepsilon_n^2 \frac{f'''(\xi)}{f'(\xi)} + \frac{1}{3!}\varepsilon_n^3 \frac{f^{iv}(\xi)}{f'(\xi)} - \cdots\right)} \\
\Rightarrow \varepsilon_{n+1} &= \varepsilon_n + \left[-\varepsilon_n + \frac{1}{2!}\varepsilon_n^2 \frac{f''(\xi)}{f'(\xi)} - \frac{1}{3!}\varepsilon_n^3 \frac{f'''(\xi)}{f'(\xi)} + \cdots\right] \\
&\quad \cdot \left[1 - \left(\varepsilon_n \frac{f''(\xi)}{f'(\xi)} - \frac{1}{2!}\varepsilon_n^2 \frac{f'''(\xi)}{f'(\xi)} + \frac{1}{3!}\varepsilon_n^3 \frac{f^{iv}(\xi)}{f'(\xi)} - \cdots\right)\right]^{-1}
\end{aligned} \tag{9.3.4}$$

Let $z = \varepsilon_n \frac{f''(\xi)}{f'(\xi)} - \frac{1}{2!}\varepsilon_n^2 \frac{f'''(\xi)}{f'(\xi)} + \frac{1}{3!}\varepsilon_n^3 \frac{f^{iv}(\xi)}{f'(\xi)} - \cdots$. Since ε_n is the error term and as $\lim_{n\to\infty} \varepsilon_n \to 0$, so we have $z << 1$. On using the expansion $(1-z)^{-1} = 1 + z + z^2 + \cdot$ in the Eq. (9.3.4), we obtain

$$\begin{aligned}
\varepsilon_{n+1} &= \varepsilon_n + \left[-\varepsilon_n + \frac{\varepsilon_n^2}{2!}\frac{f''(\xi)}{f'(\xi)} + O(\varepsilon_n^3)\right]\left[1 + \varepsilon_n \frac{f''(\xi)}{f'(\xi)} + O(\varepsilon_n^2)\right] \\
\Rightarrow \varepsilon_{n+1} &= -\frac{\varepsilon_n^2}{2}\frac{f''(\xi)}{f'(\xi)} + O(\varepsilon_n^3) \Rightarrow \lim_{n\to\infty} \frac{|\varepsilon_{n+1}|}{|\varepsilon|^2} = \left|\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right|
\end{aligned}$$

This imply that, the order of convergence of Newton-Raphson method is 2 (quadratic convergence).

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. (9.3.3), 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 $\frac{(-1)^{m-1}\varepsilon_n^{m-1}}{(m-1)!} f^{(m)}(\xi)$, 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 9.3.1. 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.

9.4 Accelerated Newton-Raphson Method

Let the function $f(x)$ have a zero ξ of multiplicity m, then the function $f'(x)$ has zero ξ of multiplicity $m-1$. So, the function $g(x) = \dfrac{f(x)}{f'(x)}$ has a zero ξ of multiplicity 1, i.e., simple zero. So, the Newton-Raphson method will be applicable for the function $g(x)$ to compute this zero, i.e.,

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} \quad \text{where} \quad g(x) = \frac{f(x)}{f'(x)} \tag{9.4.1}$$

This method is known as the accelerated Newton-Raphson method.

Example 9.4.1. Solve the equation $x^3 - .642x^2 - 3.538959x + 3.490082 = 0$ with accelerated Newton Raphson method. Start with initial approximation 1.

$$\begin{aligned}
f(x) &= x^3 - .642x^2 - 3.538959x + 3.490082 \\
f'(x) &= 3x^2 - 1.284x - 3.538959 \\
g(x) &= \frac{f(x)}{f'(x)} = \frac{x^3 - .642x^2 - 3.538959x + 3.490082}{3x^2 - 1.284x - 3.538959} \\
g'(x) &= 1 - \frac{\left(x^3 - .642x^2 - 3.538959x + 3.490082\right)(6x - 1.284)}{(3x^2 - 1.284x - 3.538959)^2}
\end{aligned}$$

The accelerated Newton-Raphson method (9.4.1) is given by

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}$$

Let initial approximation be $x_0 = 1$, then we get

$$\begin{aligned} x_1 &= x_0 - \frac{g(x_0)}{g'(x_0)} \\ &= 1.302097145 \\ x_2 &= x_1 - \frac{g(x_1)}{g'(x_1)} \\ &= 1.320945602 \\ x_3 &= x_2 - \frac{g(x_2)}{g'(x_2)} \\ &= 1.320991553 \\ x_4 &= x_3 - \frac{g(x_3)}{g'(x_3)} \\ &= 1.321000 \end{aligned}$$

9.5 Muller Method

In this method, we will approximate the function $y = f(x)$ by a second-degree curve $p_2(x)$ in the neighborhood of the root. Fig. 9.5.1 Graphical representation of Muller method Let x_{i-2}, x_{i-1} and x_i be the

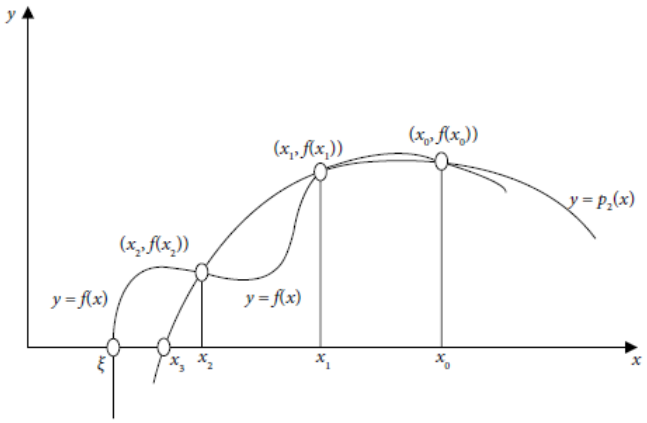


Figure 9.5.1: Graphical representation of Muller method

approximations to a root of the equation, $f(x) = 0$, then $y_{i-2} = f(x_{i-2}), y_{i-1} = f(x_{i-1})$ and $y_i = f(x_i)$ Let the approximating curve be a quadratic polynomial of the following form

$$y = A(x - x_i)^2 + B(x - x_i) + C \tag{9.5.1}$$

This parabola passes through the points $(x_{i-2}, y_{i-2}), (x_{i-1}, y_{i-1})$ and (x_i, y_i). So, we must have

$$\begin{aligned} y_{i-2} &= A(x_{i-2} - x_i)^2 + B(x_{i-2} - x_i) + C \\ y_{i-1} &= A(x_{i-1} - x_i)^2 + B(x_{i-1} - x_i) + C \\ y_i &= C \end{aligned} \tag{9.5.2}$$

Solving Eqs. (9.5.2) for the constants A, B and C, we have

$$\begin{aligned} A &= \frac{(x_{i-1}-x_i)(y_{i-2}-y_i)-(x_{i-2}-x_i)(y_{i-1}-y_i)}{(x_i-x_{i-1})(x_{i-1}-x_{i-2})(x_{i-2}-x_i)} \\ B &= \frac{(x_{i-2}-x_i)^2(y_{i-1}-y_i)-(x_{i-1}-x_i)^2(y_{i-2}-y_i)}{(x_i-x_{i-1})(x_{i-1}-x_{i-2})(x_{i-2}-x_i)} \\ C &= y_i \end{aligned} \tag{9.5.3}$$

We can obtain the quadratic polynomial (9.5.1) by using the values of constants A, B and C from Eqs. (9.5.3). The approximation to the root of equation, $f(x) = 0$ is given by the root of the following quadratic equation

$$A(x-x_i)^2 + B(x-x_{\mathrm{f}}) + C = 0$$

Let x_{i+1} be the next approximation to the root, i.e.

$$\begin{aligned} x_{i+1} - x_1 &= \frac{-B \pm \sqrt{B^2-4AC}}{2A} \\ x_{i+1} &= x_i + \frac{-2C}{B \pm \sqrt{B^2-4AC}} \end{aligned} \tag{9.5.4}$$

Note that the sign in the denominator of the Eq. (9.5.4) is chosen, so that the denominator becomes largest in the magnitude. It is to reduce the loss of significance in the approximation x_{i+1}. The method can be used to obtain the complex root, when $\sqrt{B^2-4AC} < 0$.

Example 9.5.1. Compute the approximate root of the equation $x^3 - 4x - 9 = 0$ correct to six decimal places. Use Muller method with initial approximations 2, 3 and 4.

Solution. Let $x_0 = 2, x_1 = 3$ and $x_2 = 4$ be the three Initial approximations for the root of the equation

$$y = f(x) = x^3 - 4x - 9 = 0$$

This implies

$$\begin{array}{lll} x_0 = 2 & x_1 = 3 & x_2 = 4 \\ y_0 = -9 & y_1 = 6 & y_2 = 39 \end{array}$$

Let $y = A(x-x_2)^2 + B(x-x_2) + C$ be the parabola passing through the points $(x_0, y_0), (x_1, y_1)$ and (x_2, y_2). We have

$$\begin{aligned} y_0 &= A(x_0-x_2)^2 + B(x_0-x_2) + C \\ y_1 &= A(x_1-x_2)^2 + B(x_1-x_2) + C \\ y_2 &= C \end{aligned}$$

Using the values of $(x_0, y_0), (x_1, y_1)$ and (x_2, y_2), and solving these equations for different constants, we have

$$\begin{aligned} C &= 39 \\ -9 &= 4A - 2B + 39 \\ 6 &= A - B + 39 \quad \Rightarrow A = 9, B = 42, C = 39 \end{aligned}$$

Let x_3 be the next approximation. From Eq. (9.5.4), we have

$$x_3 = x_2 + \frac{-2C}{B + \sqrt{B^2-4AC}} = 4 + \frac{-2(39)}{42 + \sqrt{(42)^2 - 4(9)(39)}} = 2.720759$$

Note that the +ve sign in the denominator as the value of B is +ve.

On using the following values of $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3)

$$\begin{array}{lll} x_1 = 3 & x_2 = 4 & x_3 = 2.720759 \\ y_1 = 6 & y_2 = 39 & y_3 = 0.257463 \end{array}$$

the next approximation of the Muller method (9.5.4) is as follows

$$x_4 = 2.706220$$

Similarly, we can obtain following approximations

$$x_5 = 2.706528$$
$$x_6 = 2.706528$$

Note 9.5.2.

1. It is easy to see that this method extracts a quadratic factor. So we can compute two roots simultaneously. This method can also be used to compute the complex roots of the nonlinear equation.
2. Muller method has high order of convergence 1.84, and generally gives the root with any initial approximation. The method can also be used to obtain complex roots.

Exercise 9.5.3.

1. Find the roots of the equation $f(x) = (x^2 - 3x + 1)^2 = 0$ using accelerated Newton–Raphson method, correct to four significant figures. Assuming multiplicity $m = 2$ and starting with the initial approximation $x_0 = 0, 2$.
2. Show that the equation $(1-x)\sin(1-x) = 0$ has a double root at the point $x = 1$. Compute the root by using the Newton–Raphson method and modified Newton–Raphson method with $m = 2$. Take initial approximation $x_0 = 0$ for both the methods.
3. Perform three iterations of the Muller method to compute the approximate root of the equation, $\cos x - 5x + 1 = 0$. Assume the first three initial approximations for the root are $0, 1$ and 2.
4. Compute all the three roots of the cubic equation $x^3 - 3x^2 - 5x + 1 = 0$ with the aid of Muller method, which are in the intervals $(-2, -1)$, $(0, 1)$ and $(4, 5)$.

Unit 10

Course Structure

- Inverse interpolation method, error estimations and convergence analysis.

10.1 Inverse Interpolation

Inverse interpolation is the process of finding the value of the argument corresponding to a given value of the function when the function is intermediate between two tabulated values. For this purpose, we shall assume that the function $y = f(x)$ has a unique inverse $x = f^{-1}(y) = F(y)$ (say) within the range of the table.

The problem of inverse interpolation can be solved by several methods, but in this unit we shall explain only two.

1. By Lagrange's formula: In the Lagrange's method we just interchange x and y and the method will be

$$x = \sum_{i=0}^{n} \frac{w(y) \cdot x_i}{(y - y_i)\, w'(y_i)} = w(y) \sum_{i=0}^{n} \frac{x_i}{\mathrm{D}_i}$$

where $\mathrm{D}'_i = (y - y_i) \times w'(y_i)$ for $i = 0, 1, \ldots n$ and we can use the table for inverse interpolation as

	y_0	y_1	$\cdots$	$\cdots$	$\cdots$	y_{n-1}	y_n	D'_i	x_i	x_i/D'_i
y_0	$\underline{y - y_0}$	$y_0 - y_1$	$\cdots$	$\cdots$	$\cdots$	$y_0 - y_{n-1}$	$y_0 - y_n$	D'_0	x_0	x_0/D'_0
y_1	$y_1 - y_0$	$\underline{y - y_1}$	$\cdots$	$\cdots$	$\cdots$	$y_1 - y_{n-1}$	$y_1 - y_n$	D'_1	x_1	x_1/D'_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
y_n	$y_n - y_0$	$y_n - y_1$	$\cdots$	$\cdots$	$\cdots$	$y_n - y_{n-1}$	$\underline{y - y_n}$	D'_n	x_n	x_n/D'_n

The product of the principal diagonal elements (underlined) is $w(y)$ and making product of row elements we get D''_1s $(i = 0, \ldots n)$.

2. By Divided difference formula, similarly as in Lagarange's method we just interchange x and y and the method will be

$$x = F(y_0) + F[y_0, y_1](y - y_0) + F[y_0, y_1, y_2](y - y_0) \times (y - y_1) + \ldots + F[y_0, y_1 \ldots y_n](y - y_0)(y - y_1) \ldots (y - y_{n-1})$$

and the computation for the same will be given as

y	$F(y)$	$\overline{\delta}$	$\overline{\delta}^2$	$\cdots$	$\overline{\delta}^n$
y_0	x_0				
		$F[y_0, y_1]$			
y_1	x_1		$F[y_0, y_1, y_2]$		
		$F[y_1, y_2]$			
y_2	x_2		$F[y_1, y_2, y_3]$	$\cdots$	$F(y_0, \ldots, y_{n-1}]$
$\vdots$	$\vdots$	$F[y_2, y_3]$	$\vdots$		
y_{n-1}	x_{n-1}	$\vdots$	$F[y_{n-2}, y_{n-1}, y_n]$		
		$F[y_n, y_{n-1}]$			
y_n	x_n				

Divided difference table for inverse interpolation

			y	$F(y)$	$\overline{\delta}$	$\overline{\delta}^2$	$\cdots$	$\overline{\delta}^n$
			y_0	x_0				
		$y_0 - y_1$			$F[y_0, y_1]$			
	$y_0 - y_2$		y_1	x_1				
$y_0 - y_3$		$y_2 - y_1$			$F[y_1, y_2]$			
$\cdots$	$\cdots$	$\cdots$	y_2	x_2	$\cdots$			
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$			
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$		$F[y_0, \ldots, y_n]$
$y_{n-3} - y_n$	$y_{n-2} - y_n$		y_{n-1}	x_{n-1}				
					$F[y_{n-1}, y_n]$			
		$y_{n-1} - y_n$	y_n	x_n				

10.2 An important application of Inverse Interpolation

The inverse interpolation can be fruitfully used in finding a real root of an equation $f(x) = 0$. Let $f(x)$ be a function which has a real root in some neighbourhood of $x = \alpha$, we have to find the values of x in the small interval $[\alpha - \delta, \alpha + \delta]$ and find the corresponding values of $f(x)$ [where $f(\alpha - \delta)f(\alpha + \delta) < 0$] and applying the inverse interpolation formula (either by Lagarange's or divided difference) we get the value of x when $f(x) = 0$.

Example 10.2.1. From the table compute the value of x for $y = 0.2$

x	$y = f(x)$
1.1	0.1047
1.2	0.1870
1.4	0.2412
1.6	0.3747
1.7	0.4353
1.9	0.5466

correct upto 4 decimal places.

Solution: We compute the problem by Lagrange's Inverse Interpolation formula

$$x = \sum_{i=0}^{5} \frac{w(y)x_i}{(y-y_i)\,w'(y_i)} = w(y)\sum_{i=0}^{5}\frac{x_i}{D_i}$$

where

$$w(y) = (y-y_0)(y-y_1)(y-y_2)(y-y_3)(y-y_4)(y-y_5)$$

$$D_i = (y-y_i)\,w'(y_i) = (y-y_i)\cdot(y-y_0)(y-y_1)(y-y_{i-1})(y-y_{i+1}),\ldots(y-y_n)$$

for $i = 0, 1, 2, 3, 4, 5$.

Computational Scheme for x when $y = 0.2$

	$y_0 =$ 0.1047	$y_1 =$ 0.0953	$y_2 =$ 0.2412	$y_3 =$ 0.3747	$y_4 =$ 0.4353	$y_5 =$ 0.5466	D_i	x_i	x_i/D_i'
$y_0 = 0.1047$	0.953	-0.0833	-0.1365	-0.2700	-0.3306	-0.4419	-0.0000427426	1.1	-25735.435
$y_1 = 0.0953$	0.0833	0.0130	-0.0542	-0.01877	-0.2483	-0.3596	0.0000009836	1.2	1219924.436
$y_2 = 0.2412$	0.1365	0.0542	-0.0412	-0.01335	-0.1941	-0.3054	0.0000241215	1.4	580394.178
$y_3 = 0.3747$	0.2700	0.1877	0.1335	-0.1747	-0.0606	-0.1719	-0.0000123126	1.6	-129947.902
$y_4 = 0.4353$	0.3306	0.2483	0.1941	0.0606	-0.2353	-0.1113	0.0000252855	1.7	67232.184
$y_5 = 0.5466$	0.4419	0.3596	0.3054	0.1719	0.1113	-0.3466	-0.0032181947	1.9	-5903.932

Therefore,

$$w(0.2) = (0.2-0.1047)\times(0.2-0.187)(0.2-0.2412)\times(0.2-0.3747)\times(0.2-4353)\times(0.2-5466) = 0.0000007272384$$

$$x = 0.0000007272384 \times 1705963.529 = 1.2406423 \approx 1.2406 \quad \text{upto } 4 \text{ decimal places}$$

Example 10.2.2. From the following table compute the value of x for $y = .7$

x	$y = f(x)$
1.15	.65468
1.16	.70108
1.17	.74727
1.18	.79325
1.19	.83902

Solution. Newton's Divided Difference formula for Inverse Interpolation :

$$x = x_0+(y-y_0)F[y_0,y_1]+(y-y_0)(y-y_1)F[y_0,y_1,y_2]+\cdots+\{(y-y_0)(y-y_1)\cdots(y-y_{n-1})\}F[y_0,y_1,\ldots,y_n]$$

y	x	$\delta(x)$	$\delta^2(x)$	$\delta^3(x)$	$\delta^4(x)$
0.65468	1.15				
		0.2155172			
0.70108	1.16		0.0105832		
		0.2164971		0.0010450	
0.74727	1.17		0.0107280		0.0001199
		0.2174859		0.0010671	
0.79325	1.18		0.0108752		
		0.2184837			
0.83902	1.19				

Computation for x for $y = 0.7$

$$\begin{aligned} x &= 1.15 + (0.7 - 0.65468)(0.2155172) + (0.7 - 0.65468)(0.7 - 0.70108)(0.0105832) \\ &\quad +(0.7 - 0.65468)(0.7 - 0.70108)(0.7 - 0.74727)(0.0010450) + \\ &\quad +(0.7 - 0.65468)(0.7 - 0.70108)(0.7 - 0.744727)(0.7 - 0.79325)(0.0001199) \\ &= 1.15 + 0.0097672 - 0.0000005 + 0.000000024 \\ &= 1.1597667 \end{aligned}$$

Therefore, $x \approx 1.159767$

Example 10.2.3. Find the root of the equation $x^2 + \log_k x - 1.6 = 0$ where $k = 7$ by interpolation formula.

Solution. We first find the values of $f(x) = x^2 + \log_k x - 1.6$

$$\begin{aligned} &f(0.1) = 0.01 + \frac{\log_e(0.1)}{\log_e 7} - 1.6 < 0 \\ &f(0.2) < 0, \\ &f(1) = -0.6000 < 0 \\ &f(1.2) = -0.0066 < 0 \quad \text{The root lies between (1.2 and 1.5)} \\ &f(1.5) = +0.8584 > 0 \\ &f(1.7) = +1.5627 > 0 \end{aligned}$$

Now, $f'(x) = 2x + \dfrac{1}{x \log_e 7} > 0$ for the range $[1, 1.7]$ hence we start to find the root of the equation by inverse interpolation formula.

1. By Lagrange's inverse interpolation formula : $y = 0$,

   The computation scheme

				D_i	x_i	x_i/D_i'
0.6000	-0.5934	-1.4584	-2.1627	-1.12297924	1	-0.0890489
0.5934	$\underline{+.0066}$	-0.8650	-1.5693	0.00531635	1.2	225.718773
1.4584	0.8650	$\underline{-0.8584}$	-0.7043	0.76267614	1.5	1.966759
2.1627	1.5693	0.7043	$\underline{-1.5627}$	-3.73538659	1.7	-0.455107
						$+227.141376$

   Therefore, $w'(0) = 0.00531203$

$$\therefore F(0) = w'(0) \times 227.141376 = 1.20658177$$

$$\therefore x = 1.20658 \text{ (upto six significant figure)}$$

   The root of the equation is 1.21 (upto three significant figure) and 1.20658 (upto six significant figure).

2. By Newton's divided difference (inverse) interpolation formula we estimate x for $y = 0$.

		y	x	$\overline{\delta}$	$\overline{\delta}^2$	$\overline{\delta}^3$
		-0.6000	1			
	0.5934			0.337040		
	1.4584	-0.0066	1.2		0.115718	
2.1627		0.8650		0.346821		-0.072025
	1.5693	+0.8584	1.5		-0.040050	
	0.7043			0.283970		
		1.5627	1.7			

Now,

$$\begin{aligned} F(0) &= 1 + (0 + 0.6000) \times 0.337040 + (0 + 0.6000) \times (0 + 0.0066) \times 0.115718 \\ &\quad + (0 + 0.6000) \times (0 + 0.0066)(0 - 0.8584) \times (-0.072025) \\ &= 1 + 0.202224 + 0.000416584 + 0.0002448 \\ &= 1.20289 \quad \text{(upto six significant figure)} \end{aligned}$$

$\therefore$ The root of the above equation is 1.20 (upto three significant figure) and 1.20289 (upto six significant figure).

Example 10.2.4. For what value of x is the value of the probability integral given in the following table equal to $\frac{1}{2}$

$$y = \frac{2}{\sqrt{\pi}} \int_0^{\pi} e^{-x^2} dx$$

y	x
0.4846555	0.46
0.4937452	0.47
0.5027498	0.48
0.5116683	0.49

Working Formulae:

1. Inverse Interpolation by Lagrange's formula:

$$x = \sum_{r=0}^{n} \frac{w(y) x_r}{(y - y_r)\, w'(y_r)} = w(y) \sum_{r=0}^{n} \frac{x_r}{\mathrm{D}_r}$$

where $\quad w(y) = (y - y_0)(y - y_1) \ldots (y - y_n)$

$$D_r = (y - y_r)\, w'(y_r)$$

2. Inverse Interpolation by divided difference formula:

$$\begin{aligned} x &= x_0 + (y - y_0) F(y_0, y_1) + (y - y_0)(y - y_1) F[y_0, y_1, y_2] + \ldots \\ &\quad + (y - y_0)(y - y_1) \ldots (y - y_{n-1}) F[y_0, y_1, y_2, \ldots, y_n] \end{aligned}$$

where $F[y_0, y_1, y_2, \ldots, y_r] = \dfrac{F[y_0, y_1, y_2, \ldots, y_{r-1}] - F[y_1, y_2, \ldots, y_r]}{y_0 - y_r}$

Computation (from 1st formula):

Row Product:
$(r \neq j) \quad (y_r - y_j)$ for $r = j(y - y_j)$

$y - y_0$	$y_0 - y_1$	$y_0 - y_2$	$y_0 - y_3$	D_r	x_r	x_r/D_r
0.0153445	−0.0090897	−0.0180943	−0.0270128	−0.00000006817	0.46	6747836.292
0.0090897	0.0062548	−0.0090046	−0.0179231	0.00000000918	0.47	51198257.081
0.0180943	0.0090043	0.0027498	−0.0089185	0.00000000400	0.48	120000000.000
0.0270128	0.0179231	0.0089185	−0.0116683	0.00000000038	0.49	−9726081.778
					Total	154724339.011

Therefore, $w(y) = 0.00000000308$

$$\therefore x = (154724339.011) \times (0.00000000308) = 0.47655096415 \simeq 0.4766 \text{ Ans.}$$

Computation (from 2nd formula):

			y	x	$\overline{\delta}$	$\overline{\delta}^2$	$\overline{\delta}^3$
			0.484655	0.46			
		-0.0090897			1.10014631946		
	-0.0180943		0.4937452	0.47		0.57461081777	
-0.0270128		-0.0090046			1.11054349998		0.87261555966
	-0.0179231		0.5027498	0.48		0.59818260736	
		-0.0089185			1.12126478667		
			0.5116683	0.49			

Coefficient	Multiplier	Positive Term	Negative Term
1	0.49	0.490000000000	
-0.0116683	1.12126478667		0.01308325391
0.00003208549	0.59818260736	1919298	
0.00000020068	0.87261555966	17511	
	Total =	0.49001936809	0.01308325391
	Difference =	0.47693611418 ≈ 0.4769	

Therefore, $x = 0.4769$.

Example 10.2.5. Compute a real root of the following equation lying in $(1, 2)$ correct to 4D by inverse interpolation;

$$x^3 + 1.6028x^2 + 7.8084x - 16.7664 = 0$$

Solution. We first find out the values of $f(x)$ in $(1, 2)$ an regular interval since the real root we have to find is in $(1, 2)$.

x	1	1.1	1.2	1.3	1.4	1.5	1.6
$f(x)$	−6.355200	−4.906772	−3.360288	−1.709748	0.050848	1.927500	3.926208

Since we have already found that $f(x)$ at $1.3 < 0$ and $f(x)$ at $1.4 > 0$, we concentrate in $(1.3, 1.4)$.

y	$x = F(y)$	$F[y_i, y_j]$	$F[y_i, y_j, y_k]$	$F[y_i, y_j, y_k, y_l]$	$F[y_i, y_j, y_k, y_l, y_m]$
-4.906772	1.1				
		0.064664			
-3.360288	1.2		-0.001275		
		0.060586		0.000033	
-1.709748	1.3		- 0.001110		- 0.010001
		0.56799		0.000027	
0.050848	1.4		- 0.000966		
		0.053286		0.000156	
1.927500	1.5		- 0.000084		
		-0.050032			
3.926208	1.6				

Coefficient	Multiplier	Positive Term	Negative Term
1	1.1	1.1	
4.906772	0.064663	0.317287	
16.488167	-0.001275		0.021022
28.190611	0.000033	0.000930	
-1.433436	- 0.000001	0.000001	
	Total =	1.418218	0.021022

Difference $= 1.418218 - 0.021022 = 1.397196$.
Therefore, the root is 1.3972

Example 10.2.6. Compute the real root of the following equation lying in $(1, 2)$ correct to 4D by inverse interpolation :

$$x^3 - px^2 + 8x - 7 = 0$$

where $p = 3.5 + \dfrac{2}{20} = 3.6$

Solution. $f(x) = x^3 - 3.6x^2 + 8x - 7$

We tabulate the values of $f(x)$ in the interval $(1, 2)$ with the step length 0.1.

x	1	1.1	1.2	1.3	1.4	1.5
$f(x)$	-1.6	-1.225	-0.856	-0.487	-0.112	0.275
x	1.6	1.7	1.8	1.9	2.0	
$f(x)$	0.68	1.109	1.568	2.063	2.6	

$f(x)$	x	δ	δ^2	δ^3
-0.856	1.2			
		0.26881720		
-0.112	1.4		-0.09212440	
		0.25839793		0.50536432
0.275	1.5		-0.01450044	
		0.24691358		
0.68	1.6			

Coefficient	Multiplier	+ ve	−ve
1	1.2	1.2	
0.856	0.26881720	0.230107523	
0.095872	-0.09212440		.00088321505
-0.0263648	-0.01543807	0.0013323829	
	Total =	1.431439906	0.0008832150

The required root $= 1.431439906 - 0.0008832150 = 1.430557 \approx 1.4306$.

Exercise 10.2.7.

1. Compute the real root of the following equation lying in $(1, 2)$ correct to 4D by inverse interpolation
$$x^3 - px^2 + 8x - 7 = 0 \text{ where } p = 3.5 + \frac{1}{20} = 3.55.$$
2. Find the root of $xe^x - 5.4 = 0$ in $(1, 2)$ by inverse interpolation.
3. Find a smallest positive root (correct upto three decimal places) of the equation $x + \log x = 0$ using inverse interpolation.
4. Applying inverse interpolation find a smallest positive root of the following equation (correct upto six significant figures)
$$e^{2.3x} + \ln(x^2 + 7.5) - 3.571 = 0$$
5. Calculate the real root of the following equation in $(1, 2)$ correct upto 4D by inverse interpolation
$$x^3 - px^2 + 8x - 7 = 0,$$
where $p = 3.5 + \frac{A}{20}$, where A represents the last digit of the AIN number of the student.

Unit 11

Course Structure

- Ordinary Differential Equations: Initial value problems–Picard's successive approximation method, error estimation, Single step method.

11.1 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

$$\frac{dy}{dx} = f(x, y) \text{ with the initial condition } y(x_0) = y_0 \tag{11.1.1}$$

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.

11.1.1 Picard's Successive Approximation Method

Integrating the differential equation given in Eq. (11.1.1), we obtain

$$y = y_0 + \int_{x_0}^{x} f(x, y)\, dx. \tag{11.1.2}$$

Equation (11.1.2), 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.(11.1.2), 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. (11.1.2) 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

$$y^{(n)} = y_0 + \int_{x_0}^{x} f(x, y^{(n-1)})\, dx \quad \text{with} \quad y^{(0)} = y_0 \tag{11.1.3}$$

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

$$|f(x, y) - f(x, \overline{y})| \leq K|y - \overline{y}|, \quad K \text{ being a constant} \tag{11.1.4}$$

then, the sequence $y^{(1)},\ y^{(2)}, \ldots$ converges to the solution of Eq. (11.1.1).

Example 11.1.1. 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 integration might become more and more difficult as we proceed to higher approximations.

Example 11.1.2. 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 11.1.3.

1. Use Picard's method to obtain a series solution the differential equation $\dfrac{dy}{dx} = 1 + xy,\ y(0) = 1$.
2. Use Picard's method to obtain $y(0.1)$ and $y(0.2)$ of the problem defined by
$$\frac{dy}{dx} = x + yx^4, \quad y(0) = 3$$
3. Using Picard's method, find $y(0.1)$, given that
$$\frac{dy}{dx} = \frac{y - x}{y + x}; \quad y(0) = 1$$

11.2 Single Step Methods

11.2.1 Euler's Method

Euler's method is the most elementary approximation technique for solving initial-value problems. Although it is seldom used in practice, the simplicity of its derivation can be used to illustrate the techniques involved in the construction of some of the more advanced techniques, without the cumbersome algebra that accompanies these constructions.

The object of Euler's method is to obtain approximations to the well-posed initial-value problem

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha. \tag{11.2.1}$$

A continuous approximation to the solution $y(t)$ will not be obtained; instead. approximations to y will be generated at various values, called mesh points, in the interval $[a, b]$. Once the approximate solution is obtained at the points, the approximate solution at other points in the interval can be found by interpolation.

We first make the stipulation that the mesh points are equally distributed throughout the interval $[a, b]$. This condition is ensured by choosing a positive integer N, setting $h = (b - a)/N$, and selecting the mesh points

$$t_i = a + ih, \qquad \text{for each } i = 0, 1, 2, \ldots, N.$$

The common distance between the points $h = t_{i+1} - t_i$ is called the step size. We will use Taylor's Theorem to derive Euler's method. Suppose that $y(t)$, the unique solution to (11.2.1), has two continuous derivatives on $[a, b]$, so that for each $i = 0, 1, 2, \ldots, N - 1$,

$$y\left(t_{i+1}\right) = y\left(t_1\right) + \left(t_{i+1} - t_i\right) y'\left(t_i\right) + \frac{\left(t_{i+1} - t_i\right)^2}{2} y''\left(\xi_i\right),$$

for some number ξ_i in (t_i, t_{i+1}). Because $h = t_{i+1} - t_i$, we have

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(\xi_i).$$

and, because $y(t)$ satisfies the differential equation (11.2.1),

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i). \tag{11.2.2}$$

Euler's method constructs $w_i \approx y(t_i)$, for each $i = 1, 2, \ldots, N$, by deleting the remainder term. Thus, Euler's method is

$$\begin{aligned} &w_0 = \alpha, \\ &w_{i+1} = w_i + hf(f_i, w_i), \qquad \text{for each } i = 0, 1, \ldots, N-1. \end{aligned} \tag{11.2.3}$$

Equation (11.2.3) is called the difference equation associated with Euler's method. In Euler method, we are using first order Taylor series. It means, we are approximating the solution curve $y(x)$ with the tangent at initial point $x = \alpha$.

Example 11.2.1. Use Euler's method algorithm to approximate the solution to

$$y' = y - t^2 + 1. \quad 0 \le t \le 2, \quad y(0) = 0.5.$$

at $t = 2$.

Solution. Here we will simply illustrate the steps in the technique when we have $h = 0.5$.

For this problem, $f(t, y) = y - t^2 + 1$; so,

$$\begin{aligned} w_0 &= y(0) = 0.5w_1 \\ w_1 &= w_0 + 0.5\left(w_0 - (0.0)^2 + 1\right) = 0.5 + 0.5(1.5) = 1.25 \\ w_2 &= w_1 + 0.5\left(w_1 - (0.5)^2 + 1\right) = 1.25 + 0.5(2.0) = 2.25 \\ w_3 &= w_2 + 0.5\left(w_2 - (1.0)^2 + 1\right) = 2.25 + 0.5(2.25) = 3.375 \end{aligned}$$

and

$$y(2) \approx w_4 = w_3 + 0.5(w_3 - (1.5)^2 + 1) = 3.375 + 0.5(2.125) = 4.4375.$$

11.2.2 Error Bounds for Euler's Method

Although Euler's method is not accurate enough to warrant its use in practice, it is sufficiently elementary to analyze the error that is produced from its application. The error analysis for the more accurate methods that we consider in subsequent sections follows the same pattern but is more complicated.

To derive an error bound for Euler's method, we need two computational lemmas.

Lemma 11.2.2. For all $x \ge -1$ and any positive m, we have $0 \le (1 + x)^m \le e^{mx}$.

Proof. Applying Taylor's Theorem with $f(x) = e^x, x_0 = 0$, and $n = 1$ gives

$$e^x = 1 + x + \frac{1}{2}x^2 e^{\xi}$$

where ξ is between x and zero. Thus,

$$0 \leq 1 + x \leq 1 + x + \frac{1}{2}x^2 e^{\xi} = e^x,$$

and, because $1 + x \geq 0$, we have

$$0 \leq (1 + x)^m \leq (e^x)^m = e^{mx}.$$

□

Lemma 11.2.3. If s and t are positive real numbers, $\{a_i\}_{i=0}^k$ is a sequence satisfying $a_0 \geq -t/s$, and

$$a_{i+1} \leq (1 + s)a_i + t, \quad \text{for each } i = 0, 1, 2, \ldots, k - 1, \tag{11.2.4}$$

then

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

Proof. For a fixed integer i, Inequality (11.2.4) implies that

$$\begin{aligned}
a_{i+1} &\leq (1 + s)a_i + t \\
&\leq (1 + s)\left[(1 + s)a_{i-1} + t\right] + t = (1 + s)^2 a_{i-1} + [1 + (1 + s)]t \\
&\leq (1 + s)^3 a_{i-2} + \left[1 + (1 + s) + (1 + s)^2\right] t \\
&\vdots \\
&\leq (1 + s)^{i+1} a_0 + \left[1 + (1 + s) + (1 + s)^2 + \cdots + (1 + s)^i\right] t.
\end{aligned}$$

But

$$1 + (1 + s) + (1 + s)^2 + \cdots + (1 + s)^i = \sum_{j=0}^{i} (1 + s)^j$$

is a geometric series with ratio $(1 + s)$ that sums to

$$\frac{1 - (1 + s)^{i+1}}{1 - (1 + s)} = \frac{1}{s}\left[(1 + s)^{i+1} - 1\right].$$

Thus,

$$a_{i+1} \leq (1 + s)^{i+1} a_0 + \frac{(1 + s)^{i+1} - 1}{s} t = (1 + s)^{i+1}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s},$$

and using Lemma 11.2.2 with $x = 1 + s$ gives

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}.$$

□

Theorem 11.2.4. Suppose f is continuous and satisfies a Lipschitz condition with constant L on

$$D = \{(t, y) \mid a \leq t \leq b \text{ and } -\infty < y < \infty\}$$

and that a constant M exists with

$$\left|y''(t)\right| \leq M, \quad \text{for all } t \in [a, b],$$

where $y(t)$ denotes the unique solution to the initial-value problem

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha.$$

Let $w_0, w_1, \ldots, w_N$ be the approximations generated by Euler's method for some positive integer N. Then, for each $i = 0, 1, 2, \ldots, N$,

$$|y(t_i) - w_i| \leq \frac{hM}{2L}\left[e^{L(t_i - a)} - 1\right] \tag{11.2.5}$$

Proof. When $i = 0$, the result is clearly true since $y(t_0) = w_0 = \alpha$. From Eq. (11.2.2), we have

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i),$$

for $i = 0, 1, \ldots, N - 1$, and from the equations in (11.2.3),

$$w_{i+1} = w_i + hf(t_i, w_i).$$

Using the notation $y_i = y(t_i)$ and $y_{i+1} = y(t_{i+1})$, we subtract these two equations to obtain

$$y_{i+1} - w_{i+1} = y_i - w_i + h[f(t_i, y_i) - f(t_i, w_i)] + \frac{h^2}{2}y''(\xi_i)$$

Hence,

$$|y_{i+1} - w_{i+1}| \leq |y_i - w_i| + h|f(t_i, y_i) - f(t_i, w_i)| + \frac{h^2}{2}|y''(\xi_i)|.$$

Now f satisfies a Lipschitz condition in the second variable with constant L, and $|y''(t)| \leq M$, so

$$|y_{i+1} - w_{i+1}| \leq (1 + hL)|y_i - w_i| + \frac{h^2 M}{2}$$

Referring to Lemma 11.2.3 and letting $s = hL, t = h^2M/2$, and $a_j = |y_j - w_j|$, for each $j = 0, 1, \ldots, N$, we see that

$$|y_{i+1} - w_{i+1}| \leq e^{(i+1)hL}\left(|y_0 - w_0| + \frac{h^2M}{2hL}\right) - \frac{h^2M}{2hL}.$$

Because $|y_0 - w_0| = 0$ and $(i + 1)h = t_{i+1} - t_0 = t_{i+1} - a$, this implies that

$$|y_{i+1} - w_{i+1}| \leq \frac{hM}{2L}\left(e^{(i_{i+1} - a)L} - 1\right)$$

for each $i = 0, 1, \ldots, N - 1$. □

The weakness of Theorem 11.2.4 lies in the requirement that a bound be known for the second derivative of the solution. Although this condition often prohibits us from obtaining a realistic error bound, it should be noted that if both $\frac{\partial f}{\partial t}$ and $\frac{\partial f}{\partial y}$ exist, the chain rule for partial differentiation implies that

$$y''(t) = \frac{dy'}{dt}(t) = \frac{df}{dt}(t, y(t)) = \frac{\partial f}{\partial t}(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \cdot f(t, y(t)).$$

So, it is at times possible to obtain an error bound for $y''(t)$ without explicitly knowing $y(t)$.

Example 11.2.5. The solution to the initial value problem

$$y' = y - t^2 + 1, \quad 0 \le t \le 2, \quad y(0) = 0.5,$$

was approximated in Example 11.2.1 using Euler's method with $h = 0.2$. Use the inequality in Theorem 11.2.4 to find a bound for the approximation errors and compare these to the actual errors.

Solution. Since $f(t, y) = y - t^2 + 1$, we have $\dfrac{\partial f(t, y)}{\partial y} = 1$ for all y, so $L = 1$. For this problem, the exact solution is $y(t) = (t + 1)^2 - 0.5e^t$, so $y''(t) = 2 - 0.5e^t$ and

$$|y''(t)| \le 0.5e^2 - 2, \quad \text{for all } t \in [0, 2].$$

Using the inequality in the error bound for Euler's method with $h = 0.2$, $L = 1$, and $M = 0.5e^2 - 2$ gives

$$|y_i - w_i| \le 0.1(0.5e^2 - 2)(e^{t_i} - 1).$$

Hence,

$$|y(0.2) - w_1| \le 0.1(0.5e^2 - 2)(e^{0.2} - 1) = 0.03752,$$
$$|y(0.4) - w_2| \le 0.1(0.5e^2 - 2)(e^{0.4} - 1) = 0.08334,$$

and so on. Table 11.2.5 lists the actual error found in Example 11.2.1 together with this error bound. Note that even though the true bound for the second derivative of the solution was used, the error bound is considerably larger than the actual error, especially for increasing values of t.

t_i	Actual Error	Error Bound
0.2	0.02930	0.03752
0.4	0.06209	0.08334
0.6	0.09854	0.13931
0.8	0.13875	0.20767
1.0	0.18268	0.29117
1.2	0.23013	0.39315
1.4	0.28063	0.51771
1.6	0.33336	0.66985
1.8	0.38702	0.85568
2.0	0.43969	1.08264

Note 11.2.6.

1. The principal importance of the error-bound formula given in Theorem 11.2.4 is that the bound depends linearly on the step size h. Consequently, diminishing the step size should give correspondingly greater accuracy to the approximations.

2. Neglected in the result of Theorem 11.2.4 is the effect that round-off error plays in the choice of step size. As h becomes smaller, more calculations are necessary, and more round-off error is expected. In actuality then, the difference-equation form

$$w_0 = \alpha,$$
$$w_{i+1} = w_i + hf(t_i, w_i), \quad \text{for each } i = 0, 1, \ldots, N - 1,$$

is not used to calculate the approximation to the solution y_i, at a mesh point t_i. We use instead an equation of the form

$$\begin{aligned} u_0 &= \alpha + \delta_0, \\ u_{i+1} &= u_i + hf(t_i, u_i) + \delta_{i+1}, \quad \text{for each } i = 0, 1, \ldots, N-1, \end{aligned} \tag{11.2.6}$$

where δ_i, denotes the round-off error associated with u_i. Using methods similar to those in the proof of Theorem 11.2.4, we can produce an error bound for the finite-digit approximations to y_i given by Euler's method.

Theorem 11.2.7. Let $y(t)$ denote the unique solution to the initial-value problem

$$y' = f(t, y), \quad a \le t \le b, \quad y(a) = \alpha, \tag{11.2.7}$$

and $u_0, u_1, \ldots, u_N$ be the approximations obtained using Eq. (11.2.6). If $|\delta_i| < \delta$ for each $i = 0, 1, \ldots, N$ and the hypotheses of Theorem 11.2.4 hold for Eq. (11.2.7), then

$$|y(t_i) - u_i| \le \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)\left[e^{L(t_i - a)} - 1\right] + |\delta_0|\, e^{L(t_i - a)}, \tag{11.2.8}$$

for each $i = 0, 1, \ldots, N$.

Note 11.2.8. The error bound (11.2.8) is no longer linear in h. In fact, since

$$\lim_{h \to 0}\left(\frac{hM}{2} + \frac{\delta}{h}\right) = \infty,$$

the error would be expected to become large for sufficiently small values of h. Calculus can be used to determine a lower bound for the step size h. Letting $E(h) = (hM/2) + (\delta/h)$ implies that $E'(h) = (M/2) - \left(\delta/h^2\right)$:

$$\begin{aligned} &\text{If } h < \sqrt{2\delta/M}, \text{ then } E'(h) < 0 \text{ and } E(h) \text{ is decreasing.} \\ &\text{If } h > \sqrt{2\delta/M}, \text{ then } E'(h) > 0 \text{ and } E(h) \text{ is increasing.} \end{aligned} \tag{11.2.9}$$

The minimal value of $E(h)$ occurs when

$$h = \sqrt{\frac{2\delta}{M}} \tag{11.2.10}$$

Decreasing h beyond this value tends to increase the total error in the approximation. Normally, however, the value of δ is sufficiently small that this lower bound for h does not affect the operation of Euler's method.

11.3 Modified (or) Improved Euler Method (or) Heun Method

Euler method involves the slope at an initial point, (x_0, y_0). In modified Euler method, we use the average value of slopes at the initial point (x_0, y_0) and last point, (x_1, y_1). It improves the estimate of the slope for the interval (x_0, x_1)

$$y_1 = y_0 + \frac{h}{2}\left(f(x_0, y_0) + f(x_1, y_1)\right)$$

The right-hand side of the equation involves the yet-to-be-determined value, y_1. To start, we can use y_1 obtained from Euler method and let it be initial approximation, $y_1^{(0)}$.

$$y_1^{(0)} = y_0 + hf(x_0, y_0)$$

The next approximate value of y_1 is computed by modified Euler method as follows

$$y_1^{(1)} = y_0 + \frac{h}{2}\left(f(x_0, y_0) + f\left(x_1, y_1^{(0)}\right)\right) \tag{11.3.1}$$

The formula (11.3.1) can be generalized in the following form

$$y_1^{(k+1)} = y_0 + \frac{h}{2}\left(f(x_0, y_0) + f\left(x_1, y_1^{(k)}\right)\right), \quad k = 0, 1, 2, \ldots \tag{11.3.2}$$

The process is repeated till the desired decimal points matches in two consecutive iterations. The formula (11.3.2) can be extended to compute, $y_{i+1}, i = 0, 1, \cdots, n$ as follows

$$\begin{aligned} y_{i+1}^{(0)} &= y_i + hf(x_i, y_i) \\ y_{i+1}^{(k+1)} &= y_i + \frac{h}{2}\left(f(x_i, y_i) + f\left(x_{i+1}, y_{i+1}^{(k)}\right)\right), \quad k = 0, 1, 2, \ldots \end{aligned} \tag{11.3.3}$$

Example 11.3.1. Use modified Euler method to compute $y(1)$ for the following IVP

$$\frac{dy}{dx} = x + y, \quad y(0) = 1$$

Use step size, $h = 0.1$.

Solution. We have

$$x_0 = 0, y_0 = 1, f(x, y) = x + y \text{ and } h = 0.1.$$

Value of $y(x_1) = y(0.1) = y_1$ Using Euler formula, we get following initial approximation $y_1^{(0)}$

$$y(x_1) = y(0.1) = y_1 = y_0 + hf(x_0, y_0) = y_0 + h(x_0 + y_0) = 1 + (0.1)(0 + 1) = 1.1 = y_1^{(0)}$$

Modified Euler method (11.3.2) can be used to improve the estimated value of $y(0.1)$ as follows

$$\begin{aligned} y_1^{(1)} &= y_0 + \frac{h}{2}\left(f(x_0, y_0) + f\left(x_1, y_1^{(0)}\right)\right) = y_0 + \frac{h}{2}\left((x_0 + y_0) + \left(x_1 + y_1^{(0)}\right)\right) \\ &= 1 + \frac{0.1}{2}((0 + 1) + (0.1 + 1.1)) = 1.11 \\ y_1^{(2)} &= y_0 + \frac{h}{2}\left(f(x_0, y_0) + f\left(x_1, y_1^{(1)}\right)\right) = y_0 + \frac{h}{2}\left((x_0 + y_0) + \left(x_1 + y_1^{(1)}\right)\right) \\ &= 1 + \frac{0.1}{2}((0 + 1) + (0.1 + 1.11)) = 1.1105 \\ y_1^{(3)} &= y_0 + \frac{h}{2}\left(f(x_0, y_0) + f\left(x_1, y_1^{(2)}\right)\right) = y_0 + \frac{h}{2}\left((x_0 + y_0) + \left(x_1 + y_1^{(2)}\right)\right) \\ &= 1 + \frac{0.1}{2}((0 + 1) + (0.1 + 1.1105)) = 1.110525 \end{aligned}$$

Value of $y(x_1) = y(0.1) = y_1 = 1.110525$

In these calculations, note that the superscripts are for the iterations of modified Euler method, while subscript denotes the variable.

Value of $y(x_2) = y(0.2) = y_2$

Using Euler formula, we have

$$y_2 = y(0.2) = y_1 + h\,(x_1 + y_1) = 1.110525 + (0.1)(0.1 + 1.110525) = 1.231578 = y_2^{(0)}$$

Modified Euler method (11.3.3) for $i = 1$ gives following iterations

$$\begin{aligned} y_2^{(1)} &= y_1 + \frac{h}{2}\left(f\,(x_1, y_1) + f\left(x_2, y_2^{(0)}\right)\right) = y_1 + \frac{h}{2}\left((x_1 + y_1) + \left(x_2 + y_2^{(0)}\right)\right) \\ &= 1.110525 + \frac{0.1}{2}((0.1 + 1.110525) + (0.2 + 1.231578)) = 1.242631 \\ y_2^{(2)} &= y_1 + \frac{h}{2}\left(f\,(x_1, y_1) + f\left(x_2, y_2^{(0)}\right)\right) = y_1 + \frac{h}{2}\left((x_1 + y_1) + \left(x_2 + y_2^{(1)}\right)\right) \\ &= 1.110525 + \frac{0.1}{2}((0.1 + 1.110525) + (0.2 + 1.242631)) = 1.243184 \\ y_2^{(3)} &= y_1 + \frac{h}{2}\left(f\,(x_1, y_1) + f\left(x_2, y_2^{(0)}\right)\right) = y_1 + \frac{h}{2}\left((x_1 + y_1) + \left(x_2 + y_2^{(2)}\right)\right) \\ &= 1.110525 + \frac{0.1}{2}((0.1 + 1.110525) + (0.2 + 1.243184)) = 1.243212 \end{aligned}$$

Value of $y\,(x_3) = y(0.3) = y_3$ The initial approximation for $y_3 = y(0.3)$ is given by

$$y_3^{(0)} = 1.387534$$

Using modified Euler formula, we get

$$\begin{aligned} y_3^{(1)} &= 1.399750 \\ y_3^{(2)} &= 1.400361 \\ y_3^{(3)} &= 1.400392 \end{aligned}$$

Value of $y\,(x_4) = y(0.4) = y_4$

$$\begin{aligned} y_4^{(0)} &= 1.570433 \\ y_4^{(1)} &= 1.583935 \\ y_4^{(2)} &= 1.584610 \\ y_4^{(3)} &= 1.584643 \end{aligned}$$

Similarly, we have the following values of y at $x = 0.5, 0.6, \ldots, 1$

$$y_5^{(0)} = 1.783110 \quad y_5^{(1)} = 1.798033 \quad y_5^{(2)} = 1.798779 \quad y_5^{(3)} = 1.798816$$

$$\begin{aligned} &y_6^{(0)} = 2.028700 \quad y_6^{(1)} = 2.045194 \quad y_6^{(2)} = 2.046019 \quad y_6^{(3)} = 2.046060 \\ &y_7^{(0)} = 2.310668 \quad y_7^{(1)} = 2.328898 \quad y_7^{(2)} = 2.329810 \quad y_7^{(3)} = 2.329856 \\ &y_8^{(0)} = 2.632844 \quad y_8^{(1)} = 2.652993 \quad y_8^{(2)} = 2.654000 \quad y_8^{(3)} = 2.654051 \end{aligned}$$

$$\begin{aligned} &y_9^{(0)} = 2.999459 \quad y_9^{(1)} = 3.021729 \quad y_9^{(2)} = 3.022842 \quad y_9^{(3)} = 3.022898 \\ &y_{10}^{(0)} = 3.415191 \quad y_{10}^{(1)} = 3.439806 \quad y_{10}^{(2)} = 3.441036 \quad y_{10}^{(3)} = 3.441098 \end{aligned}$$

It is worth mentioning here that all these iterations are obtained using C-Programs. It is very difficult and cumbersome to obtain all these manually or using a calculator. Hence it is advisable to solve these types of questions only for two or three iterations. For example, this question can be solved up to the value of $y(0.3)$.

Exercise 11.3.2. 1. Given the initial-value problem

$$y' = \frac{1}{t^2} - \frac{y}{t} - y^2, \quad 1 \le t \le 2, \quad y(1) = -1.$$

with exact solution $y(t) = -1/t$:

a. Use Euler's method with $h = 0.05$ to approximate the solution and compare it with the actual values of y.

b. Use the answers generated in part (a) and linear interpolation to approximate the following values of y and compare them to the actual values.

$$i.y(1.052) \quad ii.y(1.555) \quad iii.y(1.978)$$

c. Compute the value of h necessary for $|y(t_i) - w_i| \le 0.05$ using Eq. (11.2.5).

2. Given the initial-value problem

$$y' = -y + t + 1, \quad 0 \le t \le 5, \quad y(0) = 1,$$

with exact solution $y(t) = e^{-t} + t$:

a. Approximate $y(5)$ using Euler's method with $h = 0.2, h = 0.1$, and $h = 0.05$.

b. Determine the optimal value of h to use in computing $y(5)$, assuming that $\delta = 10^{-6}$ and that Eq. (11.2.10) is valid.

3. Consider the initial-value problem

$$y' = -10y, \quad 0 \le t \le 2, \quad y(0) = 1,$$

which has solution $y(t) = e^{-10t}$. What happens when Euler's method is applied to this problem with $h = 0.1$? Does this behavior violate Theorem 11.2.4?

Unit 12

Course Structure

- Runge-Kutta method, error estimations and convergence analysis; Multi-step method –Milne's predictor-corrector method, error estimation and convergence analysis.

12.1 Runge–Kutta (RK) Methods

In this section, we will derive certain higher order formulas known as Runge-Kutta methods, which do not involve the computations of derivative terms. Runge-Kutta methods (RK methods) are used to achieve the higher order accuracy of Taylor series without computing the higher order derivative terms. For this, we assume that the solution of the IVP

$$\frac{dy}{dx} = f(x, y), y\left(x_0\right) = y_0$$

is of the form

$$y_{i+1} = y_i + \lambda \tag{12.1.1}$$

where the general form of λ for an accuracy of $O\left(h^m\right)$ is given by the following expression

$$\lambda = w_1 k_1 + w_2 k_2 + w_3 k_3 + \cdots + w_m k_m \tag{12.1.2}$$

The aim is to determine the values of w_j 's and k_j 's in such a manner that we can achieve the desired accuracy. For this, let us assume k_j 's of the forms

$$\begin{aligned} k_1 &= hf\left(x_i, y_i\right) \\ k_2 &= hf\left(x_i + a_1 h, y_i + b_1 k_1\right) \\ k_3 &= hf\left(x_i + a_2 h, y_i + b_2 k_1 + b_3 k_2\right) \\ k_4 &= hf\left(x_i + a_3 h, y_i + b_4 k_1 + b_5 k_2 + b_6 k_3\right) \\ &\vdots \end{aligned} \tag{12.1.3}$$

where a_i 's and b_i 's are constants to determined.

First Order RK Method $(m = 1)$

Taylor series expansion is given by

$$y_{i+1} = y(x_i + h) = y(x_i) + hy'(x_i) + \frac{(h)^2}{2!}y''(x_i) + \cdots$$

By neglecting second and higher order terms, we get

$$y_{i+1} = y(x_i + h) = y(x_i) + hy'(x_i) = y(x_i) + hf(x_i, y_i)$$

On using $m = 1$ in Eqs. (12.1.1)-(12.1.3), we have

$$y_{i+1} = y_i + \lambda = y_i + \omega_1 k_1 = y_i + \omega_1 hf(x_i, y_i)$$

We get $w_1 = 1$ by comparing last two equations. So, RK method of order 1 is given by

$$y_{i+1} = y(x_i) + hf(x_i, y_i)$$

So, first order RK method is Euler method.

Second Order RK Method $(m = 2)$

Consider Eqs. (12.1.1)-(12.1.3) with $m = 2$, we have

$$y_{i+1} = y_i + \lambda, \text{ with } \lambda = w_1 k_1 + w_2 k_2$$

where k_1 and k_2 are given by

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + a_1 h, y_i + b_1 k_1) \end{aligned}$$

Accordingly, we have

$$\begin{aligned} y_{i+1} &= y_i + \lambda = y_i + w_1 k_1 + w_2 k_2 \\ &= y_i + w_1 hf(x_i, y_i) + w_2 hf(x_i + a_1 h, y_i + b_1 k_1) \end{aligned} \tag{12.1.4}$$

Expanding the term $f(x_i + a_1 h, y_i + b_1 k_1)$ by the Taylor series for the function of two variables

$$\begin{aligned} y_{i+1} &= y_i + w_1 hf(x_i, y_i) + w_2 h\left(f(x_i, y_i) + a_1 h \frac{\partial f}{\partial x}\Big|_{(x_i, y_i)} + b_1 k_1 \frac{\partial f}{\partial y}\Big|_{(x_i, y_i)} + \cdots\right) \\ &= y_i + w_1 hf(x_i, y_i) + w_2 h\left(f(x_i, y_i) + a_1 h \frac{\partial f}{\partial x}\Big|_{(x_i, y_i)} + b_1 (hf(x_i, y_i)) \frac{\partial f}{\partial y}\Big|_{(x_i, y_i)} + \cdots\right) \\ &= y_i + (w_1 + w_2) hf(x_i, y_i) + w_2 h^2 \left(a_1 \frac{\partial f}{\partial x}\Big|_{(x_i, y_t)} + b_1 (f(x_i, y_i)) \frac{\partial f}{\partial y}\Big|_{(x_i, y_1)} + \cdots\right) \end{aligned} \tag{12.1.5}$$

Since we have to achieve the accuracy up to $O\left(h^2\right)$, higher order terms can be avoided. Taylor series is given by

$$y_{i+1} = y(x_i + h) = y(x_i) + hy'(x_i) + \frac{(h)^2}{2!}y''(x_i) + \cdots \tag{12.1.6}$$

By using the given equation, $y' = f(x, y)$, we have

$$\begin{aligned} &y'(x_i) = f(x_i, y_i) \\ &y'' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} y' \\ &\Rightarrow y''(x_i) = \frac{\partial f}{\partial x}\Big|_{(x_i, y_i)} + \frac{\partial f}{\partial y}\Big|_{(x_i, y_i)} y'(x_i) = \frac{\partial f}{\partial x}\Big|_{(x_i, y_i)} + \frac{\partial f}{\partial y}\Big|_{(x_i, y_i)} f(x_i, y_i) \end{aligned} \tag{12.1.7}$$

Substituting the values of y' and y'' from Eqs. (12.1.7) in the Taylor series (12.1.6), we have

$$y_{i+1} = y(x_i + h) = y(x_i) + hf(x_i, y_i) + \frac{(h)^2}{2!}\left(\left.\frac{\partial f}{\partial x}\right|_{(x_i,y_i)} + \left.\frac{\partial f}{\partial y}\right|_{(x_i,y_i)} f(x_i, y_i)\right) + \cdots \quad (12.1.8)$$

Comparing the coefficients of $f(x_i, y_i)$, $\left.\frac{\partial f}{\partial x}\right|_{(x_i,y_i)}$, $\left.\frac{\partial f}{\partial y}\right|_{(x_i,y_i)}$ $f(x_i, y_i)$ from Eqs. (12.1.5) and (12.1.8), we have

$$\begin{aligned} w_1 + w_2 &= 1 \\ w_2 a_1 &= \frac{1}{2} \\ w_2 b_1 &= \frac{1}{2} \end{aligned} \quad (12.1.9)$$

The system (12.1.9) has three equations in four unknowns. One variable in system (12.1.9) can assume any value. Hence, infinite numbers of RK methods can be generated, here we are discussing only following two cases.

Case 1. $w_1 = \frac{1}{2}$ **(Modified Euler method)**

Let $w_1 = \frac{1}{2}$, then we have

$$w_2 = \frac{1}{2}, a_1 = b_1 = 1.$$

Using values $w_1 = w_2 = \frac{1}{2}, a_1 = b_1 = 1$, the formula (12.1.4) is given by

$$y_{i+1} = y_i + \lambda, \text{ with } \lambda = \frac{1}{2}(k_1 + k_2)$$

where k_1 and k_2 are given by

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf(x_i + h, y_i + k_i) \end{aligned}$$

It is easy to see that it is modified Euler method.

Case 2 $w_1 = \frac{1}{3}$ **(Ralston and Rabinowitz Method)**

For second order RK method, Ralston and Rabinowitz obtained that if we select $w_1 = \frac{1}{3}$, then truncation error has a minimum bound. For this case, we have

$$w_2 = \frac{2}{3}, a_1 = b_1 = \frac{3}{4}$$

On substituting the values $w_1 = \frac{1}{3}, w_2 = \frac{2}{3}, a_1 = b_1 = \frac{3}{4}$, the formula (12.1.4) produces following Ralston and Rabinowitz method for solution of IVP

$$y_{i+1} = y_i + \lambda, \text{ with } \lambda = \left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right)$$

where k_1 and k_2 are given by

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf\left(x_i + \frac{3}{4}h, y_i + \frac{3}{4}k_1\right) \end{aligned} \quad (12.1.10)$$

Third Order RK Method $(m = 3)$

For $m = 3$, the formulas (12.1.1)-(12.1.3) are given by

$$y_{i+1} = y_i + \lambda, \text{ with } \lambda = w_1k_1 + w_2k_2 + w_3k_3$$

where k_1, k_2 and k_3 are as follows

$$\begin{aligned} k_1 &= hf\left(x_i, y_i\right) \\ k_2 &= hf\left(x_i + a_1h, y_i + b_1k_1\right) \\ k_3 &= hf\left(x_i + a_2h, y_i + b_2k_1 + b_3k_2\right) \end{aligned} \tag{12.1.11}$$

Therefore, we have

$$\begin{aligned} y_{i+1} &= y_i + \lambda = y_i + w_1k_1 + w_2k_2 + w_3k_3 \\ &= y_i + w_1hf\left(x_i, y_i\right) + w_2hf\left(x_i + a_1h, y_i + b_1k_1\right) + w_3hf\left(x_i + a_2h, y_i + b_2k_1 + b_3k_2\right) \end{aligned}$$

Expanding the term $f\left(x_i + a_1h, y_i + b_1k_1\right)$ and $f\left(x_i + a_2h, y_i + b_2k_1 + b_3k_2\right)$ by the Taylor series of function of two variables

$$\begin{aligned} y_{i+1} \quad = \quad & y_i + w_1hf(x_i, y_i) + w_2h\left(f(x_i, y_i) + a_1h\frac{\partial f}{\partial x}\bigg|_{(x_i,y_i)} + b_1k_1\frac{\partial f}{\partial y}\bigg|_{(x_i,y_i)} + \cdots\right) \\ & w_3h\left(f(x_i, y_i) + a_2h\frac{\partial f}{\partial x}\bigg|_{(x_i,y_i)} + (b_2k_1 + b_3k_2)\frac{\partial f}{\partial y}\bigg|_{(x_i,y_i)} + \cdots\right) \end{aligned} \tag{12.1.12}$$

Taylor series expansion is given by

$$y_{i+1} = y(x_i + h) = y(x_i) + hy'(x_i) + \frac{h^2}{2!}y''(x_i) + \frac{h^3}{3!}$$

By using the given equation $y' = f(x, y)$ on a similar pattern as RK method of order 2 , we have

$$y_{i+1} = y\left(x_i + h\right) = y\left(x_i\right) + hf\left(x_i, y_i\right) + \frac{(h)^2}{2!}\left(\frac{\partial f}{\partial x}\bigg|_{(x_i,y_i)} + \frac{\partial f}{\partial y}\bigg|_{(x_i,y_t)} f\left(x_i, y_i\right)\right) + \cdots \tag{12.1.13}$$

Comparing the different coefficients in Eqs. (12.1.12) and (12.1.13), we get the following six equations

$$\begin{aligned} & w_1 + w_2 + w_3 = 1 \\ & b_1 - a_1 = 0 \\ & b_2 + b_3 - a_2 = 0 \\ & a_1w_2 + a_2w_3 = \frac{1}{2} \\ & a_1^2w_2 + a_2^2w_3 = \frac{1}{3} \\ & a_1b_3w_3 = \frac{1}{6} \end{aligned} \tag{12.1.14}$$

The system (**??**) has six equations in eight unknowns, so any two variables can be set as free variables to obtain infinite numbers of solutions. One solution is given by

$$w_1 = \frac{1}{4}, w_2 = w_3 = \frac{3}{8}, a_1 = a_2 = b_1 = \frac{2}{3}, b_2 = 0, b_3 = \frac{2}{3} \tag{12.1.15}$$

So, the RK method of order three is given by

$$y_{i+1} = y_i + \lambda, \quad \text{with} \quad \lambda = \frac{1}{8}\left(2k_1 + 3k_2 + 3k_3\right)$$

where k_1, k_2 and k_3 are as follows

$$\begin{aligned} k_1 &= hf\left(x_i, y_i\right) \\ k_2 &= hf\left(x_i + \tfrac{2}{3}h, y_i + \tfrac{2}{3}k_1\right) \\ k_3 &= hf\left(x_i + \tfrac{2}{3}h, y_i + \tfrac{2}{3}k_2\right) \end{aligned} \tag{12.1.16}$$

Fourth Order Runge-Kutta Method

The solution is assumed to be of the following form

$$y_{I+1} = y_1 + \lambda, \text{ with } \lambda = w_1k_1 + w_2k_2 + w_3k_3 + w_4k_4$$

where k_1, k_2, k_3 and k_4 are given by

$$\begin{aligned} k_1 &= hf\left(x_i, y_i\right) \\ k_2 &= hf\left(x_i + a_1h, y_i + b_1k_1\right) \\ k_3 &= hf\left(x_i + a_2h, y_i + b_2k_1 + b_3k_2\right) \\ k_4 &= hf\left(x_i + a_3h, y_i + b_4k_1 + b_5k_2 + b_6k_3\right) \end{aligned}$$

Proceeding in a similar manner as in previous methods, following 11 equations in 13 unknowns are obtained

$$\begin{aligned} &w_1 + w_2 + w_3 + w_4 = 1 \\ &b_1 - a_1 = 0 \\ &b_2 + b_3 - a_2 = 0 \\ &b_4 + b_5 + b_6 - a_3 = 0 \\ &a_1w_2 + a_2w_3 + a_3w_4 = \frac{1}{2} \\ &a_1^2w_2 + a_2^2w_3 + a_3^2w_4 = \frac{1}{3} \\ &a_1^3w_2 + a_2^3w_3 + a_3^3w_4 = \frac{1}{4} \\ &a_1b_3w_3 + a_1b_5w_4 + a_2b_6w_4 = \frac{1}{6} \\ &a_1^2b_3w_3 + a_1^2b_5w_4 + a_2^2b_6w_4 = \frac{1}{12} \\ &a_1b_3b_6w_4 = \frac{1}{24} \\ &a_1a_2b_3w_3 + a_1a_3b_5w_4 + a_2a_3b_6w_4 = \frac{1}{8} \end{aligned}$$

We can construct infinite numbers of 4^{th} order RK method from solution of this system. But most commonly used method is classical RK method or simply known as RK fourth order method with the following values

$$\begin{aligned} &w_1 = w_4 = \frac{1}{6}, w_2 = w_3 = \frac{1}{3} \\ &a_1 = a_2 = \frac{1}{2}, a_3 = 1 \\ &b_1 = b_3 = \frac{1}{2}, b_2 = b_4 = b_5 = 0, b_6 = 1 \end{aligned}$$

RK fourth order method with these values is given by

$$y_{i+1} = y_i + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)$$

where $k_1,\ k_2,\ k_3$ and k_4 are as follows

$$\begin{aligned} k_1 &= hf(x_i, y_i) \\ k_2 &= hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1\right) \\ k_3 &= hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2\right) \\ k_4 &= hf(x_i + h, y_i + k_3) \end{aligned} \tag{12.1.17}$$

Example 12.1.1. Use Runge-Kutta second order method with minimum bound on truncation error (Ralston and Rabinowitz method) to solve the following IVP

$$\frac{dy}{dx} = x + y, y(0) = 1$$

Compute $y(0.5)$ with step size $h = 0.1$.

Solution. Given that $x_0 = 0, y\left(x_0\right) = y_0 = 1, f(x, y) = x + y$ and $h = 0.1$. Ralston and Rabinowitz formula (12.1.10) is given by

$$y_{i+1} = y_i + \left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right)$$

where k_1 and k_2 are as follows

$$\begin{aligned} k_1 &= hf\left(x_i, y_i\right) \\ k_2 &= hf\left(x_i + \frac{3}{4}h, y_i + \frac{3}{4}k_1\right) \end{aligned}$$

Value of $y(0.1)$

$$\begin{aligned} k_1 &= hf\left(x_0, y_0\right) = 0.1\left(x_0 + y_0\right) = 0.1(0 + 1) = 0.1 \\ k_2 &= hf\left(x_0 + \frac{3}{4}h, y_0 + \frac{3}{4}k_1\right) = 0.1(0.075 + 1.075) = 0.115 \\ y(0.1) = y_1 &= y_0 + \left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right) = 1 + \left(\frac{1}{3}(0.1) + \frac{2}{3}(0.115)\right) = 1.11 \end{aligned}$$

Value of $y(0.2)$

$$\begin{aligned} k_1 &= hf\left(x_1, y_1\right) = 0.1\left(x_1 + y_1\right) = 0.1(0.1 + 1.11) = 0.121 \\ k_2 &= hf\left(x_1 + \frac{3}{4}h, y_1 + \frac{3}{4}k_1\right) = 0.1(.175 + 1.20075) = .137575 \\ y(0.2) = y_2 &= y_1 + \left(\frac{1}{3}k_1 + \frac{2}{3}k_2\right) = 1.11 + \left(\frac{1}{3}(0.121) + \frac{2}{3}(0.137575)\right) = 1.242050 \end{aligned}$$

Similarly, other iterations are as follows

$$\begin{array}{lll} k_1 = 0.144205 & k_2 = 0.162520 & y(0.3) = y_3 = 1.398465 \\ k_1 = 0.169847 & k_2 = 0.190085 & y(0.4) = y_4 = 1.581804 \\ k_1 = 0.198180 & k_2 = 0.220544 & y(0.5) = y_5 = 1.794894 \end{array}$$

Example 12.1.2. Use Runge-Kutta fourth order method with step size $h = 0.1$ for the IVP $\frac{dy}{dx} = x - y^2, y(1) = 2$, to compute $y(1.2)$

Solution. RK method of order 4 produces following iterations.

First Iteration

$$
\begin{aligned}
k_1 &= hf\left(x_0, y_0\right) = 0.1\left(x_0 - y_0^2\right) = 0.1\left(1 - 2^2\right) = -0.3 \\
k_2 &= hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right) = 0.1\left(1.05 - (1.85)^2\right) = -0.237250 \\
k_3 &= hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_2\right) = 0.1\left(1.05 - (1.881375)^2\right) = -0.248957 \\
k_4 &= hf\left(x_0 + h, y_0 + k_3\right) = 0.1\left(1.1 - (1.751043)^2\right) = -0.196615 \\
y(1.1) &= y_1 = y_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) \\
&= 2 + \frac{1}{6}(-0.3 + 2(-0.23725) + 2(-0.248957) - 0.196615) \\
&= 1.755162
\end{aligned}
$$

Second Iteration

$$
\begin{aligned}
k_1 &= hf\left(x_1, y_1\right) = 0.1\left(1.1 - (1.755162)^2\right) = -0.198059 \\
k_2 &= hf\left(x_1 + \frac{1}{2}h, y_1 + \frac{1}{2}k_1\right) = -0.159277 \\
k_3 &= hf\left(x_1 + \frac{1}{2}h, y_1 + \frac{1}{2}k_2\right) = -0.165738 \\
k_4 &= hf\left(x_1 + h, y_1 + k_3\right) = -0.132627 \\
y(1.2) &= y_2 = y_1 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = 1.591709
\end{aligned}
$$

Exercise 12.1.3.

1. Given the problem $\frac{dy}{dx} = f(x, y)$ 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 $k_1 = hf(x_0, y_0)$, $k_2 = hf\left(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}k_1\right)$ and $k_3 = hf(x_0 + h, y_0 + 2k_2 - k_1)$. Show that R_4 is of order h^4.

2. Use Runge-Kutta fourth order formula to find $y(0.2)$ and $y(0.4)$ given that

$$\frac{dy}{dx} = \frac{y^2 - x^2}{y^2 + x^2};\ \ y(0) = 1$$

3. Solve the initial value problem $\frac{dr}{d\theta} + r^2 = \sin 2\theta;\ r(0) = 0$ by using 4^{th} order Runge–Kutta method to compute the values of $r(0.2)$ and $r(0.4)$.

12.2 Milne's Predictor-Corrector Method

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.

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:

$$y' = y'_{n-3} + u \cdot \Delta y'_{n-3} + \frac{u(u-1)}{2!} \cdot \Delta^2 y'_{n-3} + \frac{u(u-1)(u-2)}{3!} \cdot \Delta^2 y'_{n-2} + \frac{u(u-1)(u-2)(u-3)}{4!} \cdot \Delta^4 y'_{n-3} + \cdots \tag{12.2.1}$$

where $u = \dfrac{x - x_{n-3}}{h}$ or $x = x_{n-3} + hu$. Therefore $dx = h\ du$. Let the differential equation be

$$\frac{dy}{dx} = f(x, y) \ \text{ with } \ y(x_{n-3}) = y_{n-3}. \tag{12.2.2}$$

Now integrating (12.2.2) over the range x_{n-3} to x_{n+1}, we get

$$\int_{x_{n-3}}^{x_{n+1}} dy = \int_{x_{n-3}}^{x_{n+1}} y'\, dx$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \int_0^4 \left[y'_{n-3} + u \cdot \Delta y'_{n-3} + \frac{u(u-1)}{2!} \cdot \Delta^2 y'_{n-3} + \frac{u(u-1)(u-2)}{6} \cdot \Delta^2 y'_{n-2} + \frac{u(u-1)(u-2)(u-3)}{24} \cdot \Delta^4 y'_{n-3} \right] du$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \left[4y'_{n-3} + 8\Delta y'_{n-3} + \frac{20}{3}\Delta^2 y'_{n-3} + \frac{8}{3}\Delta^3 y'_{n-3} + \frac{14}{45}\Delta^4 y'_{n-3} \right]$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \left[4y'_{n-3} + 8(E-1)y'_{n-3} + \frac{20}{3}(E-1)^2 y'_{n-3} + \frac{8}{3}(E-1)^3 y'_{n-3} \right] + \frac{14}{45} h \Delta^4 y'_{n-3}$$

$$\Rightarrow y_{n+1} - y_{n-3} = h \left[4y'_{n-3} + 8\{y'_{n-2} - y'_{n-3}\} + \frac{20}{3}\{y'_{n-1} - 2y'_{n-2} + y'_{n-3}\} + \frac{8}{3}\{y'_n - 3y'_{n-1} + 3y'_{n-2} - y'_{n-3}\} \right] + \frac{14}{45} h \Delta^4 y'_{n-3}$$

$$\Rightarrow y_{n+1} - y_{n-3} = \frac{4h}{3} \left[2y'_{n-2} - y'_{n-1} + 2y'_n \right] + \frac{14}{45} h \Delta^4 y'_{n-3}$$

$$\Rightarrow y_{n+1} = y_{n-3} + \frac{4h}{3} \left[2y'_{n-2} - y'_{n-1} + 2y'_n \right] + E_1$$

where $E_1 = \frac{14}{45}h\Delta^4 y'_{n-3} = \frac{14}{45}h^5 y^v(\xi_1), \ \ (x_{n-3} < \xi_1 < x_{n+1})$, assuming that $y^v(x)$ does not vary strongly in the small interval (x_{n-3}, x_{n+1}). Then the formula

$$y^{(p)}_{n+1} = y_{n-3} + \frac{4h}{3}\left[2y'_{n-2} - y'_{n-1} + 2y'_n\right] \tag{12.2.3}$$

is called the *Milne's extrapolation formula or Predictor formula* with the error

$$E_1 = \frac{14}{45}h\Delta^4 y'_{n-3} = \frac{14}{45}h^5 y^v(\xi_1), \ \ (x_{n-3} < \xi_1 < x_{n+1}) \tag{12.2.4}$$

To derive the corrector formula, we integrate Eq.(12.2.2) by the Newton's forward formula with starting node x_{n-1}, in terms of y' and u

$$\begin{aligned} y' = y'_{n-1} + u\cdot\Delta y'_{n-1} + \frac{u(u-1)}{2!}\cdot\Delta^2 y'_{n-1} + \frac{u(u-1)(u-2)}{3!}\cdot\Delta^2 y'_{n-1} \\ + \frac{u(u-1)(u-2)(u-3)}{4!}\cdot\Delta^4 y'_{n-1} + \cdots \end{aligned} \tag{12.2.5}$$

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:

$$\begin{aligned} \int_{x_{n-1}}^{x_{n+1}} dy &= \int_{x_{n-1}}^{x_{n+1}} y'\,dx \\ \Rightarrow y_{n+1} - y_{n-1} &= h\int_0^2 \left[y'_{n-1} + u\cdot\Delta y'_{n-1} + \frac{u(u-1)}{2!}\cdot\Delta^2 y'_{n-1} + \frac{u(u-1)(u-2)}{6}\cdot\Delta^2 y'_{n-1}\right. \\ &\qquad \left. + \frac{u(u-1)(u-2)(u-3)}{24}\cdot\Delta^4 y'_{n-1}\right] du \\ \Rightarrow y_{n+1} - y_{n-1} &= h\left[2y'_{n-1} + 2\Delta y'_{n-1} + \frac{1}{3}\Delta^2 y'_{n-1} - \frac{1}{90}\Delta^4 y'_{n-1}\right] \\ \Rightarrow y_{n+1} - y_{n-1} &= h\left[2y'_{n-1} + 2(E-1)y'_{n-1} + \frac{1}{3}(E-1)^2 y'_{n-1}\right] - \frac{h}{90}\Delta^4 y'_{n-1} \\ \Rightarrow y_{n+1} - y_{n-1} &= h\left[2y'_{n-1} + 2\{y'_n - y'_{n-1}\} + \frac{1}{3}\{y'_{n+1} - 2y'_n + y'_{n-1}\}\right] - \frac{h}{90}\Delta^4 y'_{n-1} \\ \Rightarrow y_{n+1} - y_{n-1} &= \frac{h}{3}\left[y'_{n-1} + 4y'_n + y'_{n+1}\right] - \frac{h}{90}\Delta^4 y'_{n-1} \end{aligned}$$

$$\Rightarrow y_{n+1} = y_{n-1} + \frac{h}{3}\left[y'_{n-1} + 4y'_n + y'_{n+1}\right] + E_2$$

where $E_2 = -\dfrac{h}{90}\Delta^4 y'_{n-4} = -\dfrac{h^5}{90}y^v(\xi_2), \ \ (x_{n-1} < \xi_2 < x_{n+1})$, assuming that $y^v(x)$ does not vary strongly in the small interval (x_{n-1}, x_{n+1}). Then the formula

$$y^{(c)}_{n+1} = y_{n-1} + \frac{h}{3}\left[y'_{n-1} + 4y'_n + y'_{n+1}\right] \tag{12.2.6}$$

is called the *Milne's corrector formula* with the error

$$E_2 = -\frac{h}{90}y^v_{\xi_2}, \quad (x_{n-1} < \xi_2 < x_{n+1}) \tag{12.2.7}$$

The value of y_{n+1} computed by (12.2.3) may be called it predicted value and that computed by (12.2.6) 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

$$D_{n+1} = \text{Corrected value } \ y_{n+1} - \text{Predicted value } \ y_{n+1} = E_1 - E_2 \approx -29E_2 \tag{12.2.8}$$

12.2.1 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 (12.2.3).

• 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 (12.2.6).

• 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 12.2.1. 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 (12.2.3) and in the corrector formula (12.2.6) we get, respectively,

$$y^{(p)}_4 = y_0 + \frac{4h}{3}[2y'_1 - y'_2 + 2y'_3] \tag{12.2.9}$$

$$y^{(p)}_4 = y_2 + \frac{h}{3}[y'_2 + 4y'_3 + y'_4] \tag{12.2.10}$$

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 (12.2.9), 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 (12.2.10), 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 (12.2.10), 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 12.2.2. 1. 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.

2. Compute $y(0.5)$, by Milne's predictor-corrector method from $\frac{dy}{dx} = 2e^x - y$ given that

$$y(0.1) = 2.0100, \quad y(0.2) = 2.0401, \quad y(0.3) = 2.0907, \quad y(0.4) = 2.1621$$

Unit 13

Course Structure

- Partial Differential Equations: Finite difference methods for Elliptic equations

13.1 Introduction

Most of the mathematical models of the physical systems give rise to a system of linear or nonlinear partial differential equations. Since analytical methods are not always available for solving these equations, we attempt to solve by numerical methods. The numerical methods can broadly be classified as finite element methods and finite difference methods. We shall be considering only the finite difference methods for solving some of these equations.

13.1.1 Finite difference method for elliptic partial differential equations

We know that a second order partial differential equation $Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu + G = 0$ is elliptic type if $B^2 - 4AC < 0$. For example, the Laplace's equation and Poisson's equation are the elliptic type partial differential equation. For this, we consider the solution of the following Dirichlet boundary value problems governed by the given partial differential equations along with suitable boundary conditions

- Laplace's equation: $u_{xx} + u_{yy} = \nabla^2 u = 0$, with $u(x, y)$ prescribed on the boundary, that is $u(x, y) = f(x, y)$ on the boundary.
- Poisson's equation: $u_{xx} + u_{yy} = \nabla^2 u = G(x, y)$, with $u(x, y)$ prescribed on the boundary, that is $u(x, y) = f(x, y)$ on the boundary.

Finite difference mehtod

We have a two dimensional domain $(x, y) \in R$. We superimpose on this domain R, a rectangular network or mesh of lines with step lengths h and k respectively, parallel to the x- and y-axis. The mesh of lines is called a grid. The points of intersection of the mesh lines are called nodes or grid points or mesh points. The grid points are given by (x_i, y_j) (see. Figs. 13.1.1 and 13.1.2), where the mesh lines are defined by

$$x_i = ih,\ i = 0, 1, 2, \ldots;\ y_j = jk,\ j = 0, 1, 2, \ldots$$

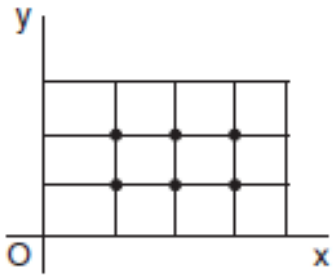


Figure 13.1.1: Nodes in a rectangle

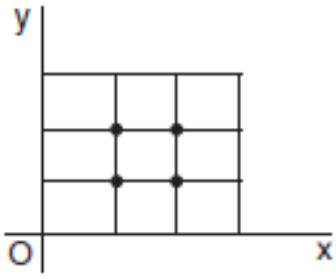


Figure 13.1.2: Nodes in a square

If $h = k$, then we have a uniform mesh. Denote the numerical solution at (x_i, y_i) by u_{ij}. At the nodes, the partial derivatives in the differential equation are replaced by suitable difference approximations. That is, the partial differential equation is approximated by a difference equation at each nodal point. This procedure is called discretization of the partial differential equation. We use the following central difference approximations.

$$(u_x)_{i,j} = \frac{1}{2h}(u_{i+1,j} - u_{i-1,j}), \qquad (u_y)_{i,j} = \frac{1}{2k}(u_{i,j+1} - u_{i,j-1}),$$
$$(u_{xx})_{i,j} = \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}), \quad (u_{yy})_{i,j} = \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}).$$

13.1.2 Solution of Laplace's equation

We apply the Laplace's equation at the nodal point (i, j). Inserting the above approximations in the Laplace's equation, we obtain

$$(u_{xx})_{i,j} + (u_{yy})_{i,j} = \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j} + \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = 0$$
$$\text{or } (u_{i+1,j} - 2u_{i,j} + u_{i-1,j} + p^2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = 0, \text{ where } p = h/k. \tag{13.1.1}$$

If $h = k$, that is, $p = 1$ (called the uniform mesh spacing), we obtain the difference approximation as

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0. \tag{13.1.2}$$

This approximation is called the *standard five point formula.* We can write this formula as

$$u_{i,j} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}). \tag{13.1.3}$$

We observe that $u_{i,j}$ is obtained as the mean of the values at the four neighbouring points in the x and y directions. The nodal points that are used in computations are given in Fig. 13.1.3.

Remark 13.1.1. The nodes in the mesh are numbered in an orderly way. We number them from left to right and from top to bottom or from bottom to top. A typical numbering is given in Figs. 13.1.4, 13.1.5.

System of equations governing the solutions: The difference approximation (13.1.2), to the Laplace equation $u_{xx} + u_{yy} = \nabla^2 u = 0$ is applied at all the nodes and the boundary conditions are used to simplify the equations. The resulting system is a linear system of algebraic equations $Au = d$.

Structure of the coefficient matrix: Let us write the system of equations that arise when we have nine nodes as given in Fig. 13.1.4. Since the boundary values are known, we have the following system of equation.

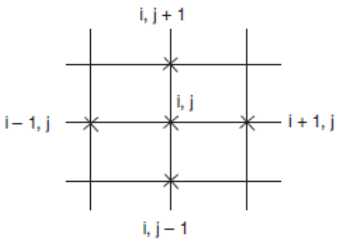


Figure 13.1.3: Standard five point formula

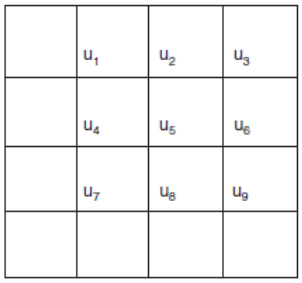


Figure 13.1.4: Numbering of nodes

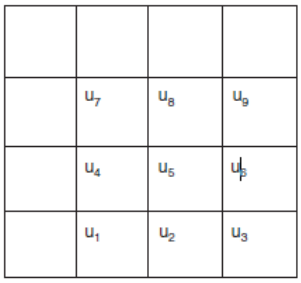


Figure 13.1.5: Nodes in a square

At 1:	$u_2 + u_4 - 4u_1 = b_1,$	or	$-4u_1 + u_2 + u_4 = b_1,$
At 2:	$u_1 + u_5 + u_3 - 4u_2 = b_2,$	or	$u_1 - 4u_2 + u_3 + u_5 = b_2,$
At 3:	$u_2 + u_6 - 4u_3 = b_3,$	or	$u_2 - 4u_3 + u_6 = b_3,$
At 4:	$u_1 + u_7 + u_5 - 4u_4 = b_4,$	or	$u_1 - 4u_4 + u_5 + u_7 = b_4,$
At 5:	$u_2 + u_4 + u_8 + u_6 - 4u_5 = 0,$	or	$u_2 + u_4 - 4u_5 + u_6 + u_8 = 0,$
At 6:	$u_3 + u_5 + u_9 - 4u_6 = b_6,$	or	$u_3 + u_5 - 4u_6 + u_9 = b_6,$
At 7:	$u_4 + u_8 - 4u_7 = b_7,$	or	$u_4 - 4u_7 + u_8 = b_7,$
At 8:	$u_5 + u_7 + u_9 - 4u_8 = b_8,$	or	$u_5 + u_7 - 4u_8 + u_9 = b_8,$
At 9:	$u_6 + u_8 - 4u_9 = b_2.$		

where $b_1, b_2, b_3, b_4, b_6, b_7, b_8, b_9$ are the contributions from the boundary values. We have the following linear algebraic system of equations,

$$\begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -4 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ 0 \\ b_6 \\ b_7 \\ b_8 \\ b_9 \end{bmatrix}$$

which is of the form $Au = d$. It is a band matrix system. The half band width is the number of nodal points on each mesh line, that is, 3. Therefore, the total band width of the matrix is $3 + 3 + 1 = 7$, that is, all the non-zero elements are located in this band.

13.1.3 Derivation of error in the approximation for the Laplace's equation

Consider the case of uniform mesh, that is, $h = k$. Using the Taylor series expansions in Eq. (13.1.1) with $h = k$, we obtain

$$
\begin{aligned}
&[\{u(x_{i+1}, y_j) - 2u(x_i, y_j) + u(x_{i-1}, y_j)\} + \{u(x_i, y_{j+1}) - 2u(x_i, y_j) + u(x_i, y_{j-1})\}] \\
&= \left[\left\{\left(u + h\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4} + \dots\right) - 2u\right.\right. \\
&\left.\left. + \left(u - h\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4} - \dots\right)\right\}\right. \\
&\left. + \left\{\left(u + h\frac{\partial u}{\partial y} + \frac{h^2}{2}\frac{\partial^2 u}{\partial y^2} + \frac{h^3}{6}\frac{\partial^3 u}{\partial y^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial y^4} + \dots\right) - 2u\right.\right. \\
&\left.\left. + \left(u - h\frac{\partial u}{\partial y} + \frac{h^2}{2}\frac{\partial^2 u}{\partial y^2} - \frac{h^3}{6}\frac{\partial^3 u}{\partial y^3} + \frac{h^4}{24}\frac{\partial^4 u}{\partial y^4} - \dots\right)\right\}\right]_{i,j} \\
&= \left[h^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \frac{h^4}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right) + \dots\right]_{i,j} \\
&= \frac{h^4}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right)_{i,j} + \dots \qquad \left(\because \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0\right)
\end{aligned}
$$

The truncation error of the method (13.1.1) when $h = k$, is given by

$$
T.E. = (u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + (u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = \frac{h^4}{12}\left(\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4}\right)_{i,j} + \dots
$$

Hence, the truncation error of the method is of order $O(h^4)$. The order of the formula (13.1.1) is defined as

$$
\text{Order} = \frac{1}{h^2}\,(T.E.) = O(h^2).
$$

We say that the method is of second order. When a method converges, it implies that the errors in the numerical solutions $\to 0$ as $h \to 0$. Suppose that a method is of order $O(h^2)$. Then, if we reduce the step length h by a factor, say 2, and recompute the numerical solution using the step length $h/2$, then the error becomes $O[(h/2)^2] = [O(h^2)]]/4$. Therefore, the errors in the numerical solutions are reduced by a factor of 4.

13.1.4 Solution of Poisson equation

Consider the solution of the Poisson's equation

$$
u_{xx} + u_{yy} = \nabla^2 u = G(x, y),
$$

with $u(x, y)$ prescribed on the boundary, that is, $u(x, y) = g(x, y)$ on the boundary. Eqs.(13.1.1)-(13.1.3) becomes

$$
(u_{xx})_{i,j} + (u_{yy})_{i,j} = \frac{1}{h^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + \frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = G_{i,j} \tag{13.1.4}
$$

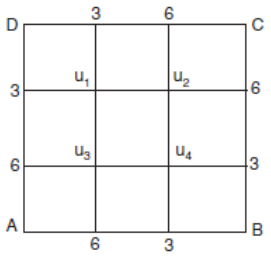
$$
(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}) + p^2(u_{i,j+1} - 2u_{i,j} + u_{i,j-1} = h^2 G_{i,j}, \tag{13.1.5}
$$

where $G_{i,j} = G(x_i, y_j)$ and $p = h/k$. If $h = k$, that is, $p = 1$, we obtain the difference approximation as

$$u_{i+1,j} + u_{i-1,j}) + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 G_{i,j}, \tag{13.1.6}$$

This approximation is called the standard five point formula for Poisson's equation. The formula (13.1.5) is of order $O(h^2 + k^2)$ and formula (13.1.6) is of order $O(h^2)$. We also call it a second order formula.

Example 13.1.2. Solve $u_{xx} + u_{yy} = 0$ numerically for the following mesh with uniform spacing and with boundary conditions as shown below.



Solution: We note that the partial differential equation and the boundary conditions are symmetric about the diagonals AC and BD. Hence, $u_1 = u_4$ and $u_2 = u_3$. Therefore, we need to solve for two unknowns u_1 and u_2. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

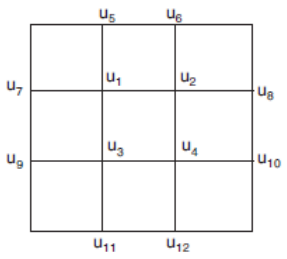
We obtain the following difference equations.

At 1: $u_2 + 3 + 3 + u_3 - 4u_1 = 0$, or $-4u_1 + 2u_2 = -6$, or $-2u_1 + u_2 = -3$,
At 2: $6+6+u_1 + u_4 - 4u_2 = 0$, or $2u_1 - 4u_2 = -12$.

Adding the two equations, we get $-3u_2 = -15$, or $u_2 = 5$. From the first equation, we get $2u_1 = u_2 + 3 = 5 + 3 = 8$, or $u_1 = 4$.

Example 13.1.3. Solve $u_{xx} + u_{yy} = 0$ numerically under the boundary conditions $u(x, 0) = 2x,\ u(0, y) = -y,\ u(x, 1) = 2x - 1,\ u(1, y) = 2 - y$ with square mesh of width $h = 1/3$.

Solution: The mesh is given in figure below. We need to find the values of the four unknowns $u_1,\ u_2,\ u_3$



and u_4. We use the standard five point formula

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0.$$

Using the boundary conditions, we get the boundary values as

$$u_5 = u\left(\frac{1}{3},1\right) = \frac{2}{3} - 1 = -\frac{1}{3}, \quad u_6 = u\left(\frac{2}{3},1\right) = \frac{4}{3} - 1 = \frac{1}{3}, \quad u_7 = u\left(0,\frac{2}{3}\right) = -\frac{2}{3},$$
$$u_8 = u\left(1,\frac{2}{3}\right) = 2 - \frac{2}{3} = \frac{4}{3}, \quad u_9 = u\left(0,\frac{1}{3}\right) = -\frac{1}{3}, \quad u_{10} = u\left(1,\frac{1}{3}\right) = 2 - \frac{1}{3} = \frac{5}{3},$$
$$u_{11} = u\left(\frac{1}{3},0\right) = \frac{2}{3}, \quad u_{12} = u\left(\frac{2}{3},0\right) = \frac{4}{3}.$$

At 1: $u_2 + u_5 + u_7 + u_3 - 4u_1 = 0$, or $-4u_1 + 2u_2 + u_3 = 1$
At 2: $u_8 + u_6 + u_1 + u_4 - 4u_2 = 0$, or $u_1 - 4u_2 + u_4 = -5/3$
At 3: $u_4 + u_1 + u_9 + u_{11} - 4u_3 = 0$, or $u_1 - 4u_3 + u_4 = -1/3$
At 4: $u_{10} + u_2 + u_3 + u_{12} - 4u_4 = 0$, or $u_2 + u_3 - 4u_4 = -3$.

We solve the system of equations using the Gauss elimination method. We use the augmented matrix $[A|d]$

$$\left[\begin{array}{cccc|c} -4 & 1 & 1 & 0 & 1 \\ 1 & -4 & 0 & 1 & -5/3 \\ 1 & 0 & -4 & 1 & -1/3 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right]; \xrightarrow{R_1/-4}, \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 1 & -4 & 0 & 1 & -5/3 \\ 1 & 0 & -4 & 1 & -1/3 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right];$$

$$\mathrm{R}_2 - R_1,\ R_3 - R_1 \rightarrow \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & -3.75 & 0.25 & 1 & -1.41667 \\ 0 & 0.25 & -3.75 & 1 & -0.08333 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right];$$

$$\mathrm{R}_2/-3.75 \rightarrow \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & 0.06667 & -0.26667 & 0.37778 \\ 0 & 0.25 & -3.75 & 1 & -0.08333 \\ 0 & 1 & 1 & -4 & -3 \end{array}\right];$$

$$\mathrm{R}_3 - 0.25R_2, R_4 - R_2 \rightarrow \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0 & -3.73333 & 1.06667 & -0.17778 \\ 0 & 0 & 1.06667 & -3.73333 & -3.377778 \end{array}\right];$$

$$\mathrm{R}_3/-3.73333 \rightarrow \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0 & 1 & -0.28572 & -0.04762 \\ 0 & 0 & 1.06667 & -3.73333 & -3.377778 \end{array}\right];$$

$$\mathrm{R}_4 - 1.06667R_3 \rightarrow \left[\begin{array}{cccc|c} 1 & -0.25 & -0.25 & 0 & -0.25 \\ 0 & 1 & -0.06667 & -0.26667 & 0.37778 \\ 0 & 0 & 1 & -0.28572 & -0.04762 \\ 0 & 0 & 0 & -3.42856 & -3.42857 \end{array}\right];$$

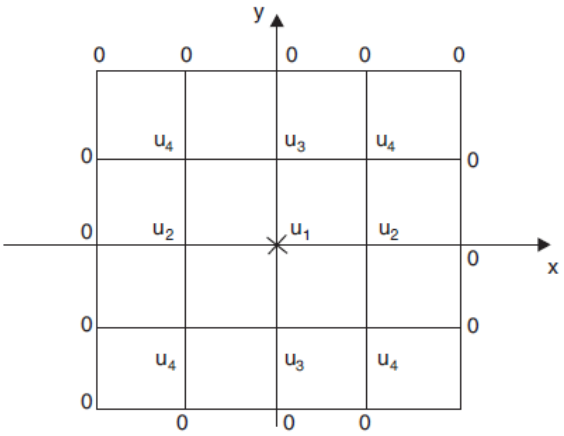
Last equation gives $u_4 = 1$. Substituting in the third equation, we get $u_3 = 0.044762 + 0.28572 = 0.33334$. Substituting in the second equation, we get $u_2 = 0.37778 + 0.06667(0.33334) + 0.26667 = 0.66667$. Substituting in the first equation, we get $u_1 = -0.25 + 0.25(0.66667 + 0.33334) = 0$).

Example 13.1.4. Solve the boundary value problem for the Poisson equation

$$u_{xx} + u_{yy} = x^2 - 1,\ |x| \leq 1,\ |y| \leq 1$$
$$u = 0 \text{ on the boundary of the square}$$

using the five point formula with square mesh of width $h = 1/2$.

Solution. The mesh is given in the figure below. The partial differential equation and the boundary conditions are symmetric about x-and y-axis. We need to find the values of the four unknowns u_1, u_2, u_3 and u_4. We use the standard five point formula



$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = h^2 G_{i,j} = 0.25\left(x_i^2 - 1\right)$$

We obtain the following difference equations.

At $1(0,0)$: $\quad u_2 + u_3 + u_2 + u_3 - 4u_1 = -0.25,$

$$\text{or} \quad -2u_1 + u_2 + u_3 = -0.125.$$

At $2(0.5,0)$: $\quad 0 + u_4 + u_1 + u_4 - 4u_2 = 0.25(0.25 - 1) = -0.1875,$

$$\text{or} \quad u_1 - 4u_2 + 2u_4 = -0.1875.$$

At $3(0,0.5)$: $\quad u_4 + 0 + u_4 + u_1 - 4u_3 = 0.25(0 - 1) = -0.25$, or

$$u_1 - 4u_3 + 2u_4 = -0.25.$$

At $4(0.5,0.5) : 0 + 0 + u_3 + u_2 - 4u_4 = 0.25(0.25 - 1) = -0.1875$, or

$$u_2 + u_3 - 4u_4 = -0.1875$$

We solve the system of equations using the Gauss elimination method. We use the augmented matrix [A|d].

$$\left[\begin{array}{cccc|c} -2 & 1 & 1 & 0 & -0.125 \\ 1 & -4 & 0 & 2 & -0.1875 \\ 1 & 0 & -4 & 2 & -0.25 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right]; \frac{R_1}{-2}, \left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 1 & -4 & 0 & 2 & -0.1875 \\ 1 & 0 & -4 & 2 & -0.25 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right]; R_2 - R_1, R_3 - R_1,$$

$$\left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & -3.5 & 0.5 & 2 & -0.25 \\ 0 & 0.5 & -3.5 & 2 & -0.3125 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right]; \frac{R_2}{-3.5}, \left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0.5 & -3.5 & 2 & -0.3125 \\ 0 & 1 & 1 & -4 & -0.1875 \end{array}\right];$$

$$R_3 - 0.5R_2, R_4 - R_2 \left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0 & -3.42857 & 2.28572 & -0.34822 \\ 0 & 0 & 1.14286 & -3.42857 & -0.25893 \end{array}\right]; \frac{R_3}{-3.42857},$$

$$\left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0 & 1 & -0.66667 & 0.10156 \\ 0 & 0 & 1.14286 & -3.42857 & -0.25893 \end{array}\right]; \quad R_4 - 1.14286R_3,$$

$$\left[\begin{array}{cccc|c} 1 & -0.5 & -0.5 & 0 & 0.0625 \\ 0 & 1 & -0.14286 & -0.57143 & 0.07143 \\ 0 & 0 & 1 & -0.66667 & 0.10156 \\ 0 & 0 & 0 & -2.66667 & -0.37500 \end{array}\right]$$

Last equation gives $u_4 = \dfrac{0.37500}{2.66667} = 0.14062.$
Substituting in the third equation, we get $u_3 = 0.10156 + 0.66667(0.14062) = 0.19531.$
Substituting in the second equation, we get

$$u_2 = 0.07143 + 0.14286(0.19531) + 0.57143(0.14062) = 0.17969.$$

Substituting in the first equation, we get $u_1 = 0.5(0.17969 + 0.19531) + 0.0625 = 0.25.$

Exercise 13.1.5. 1. Find the solution of the Laplace equation $u_{xx} + u_{yy} = 0$ in the region R is a square of side 3 units subject to the given boundary conditions $u(0, y) = 0,\ u(3, y) = 3 + y,\ u(x, 0) = x,\ u(x, 3) = 2x$, using the standard five point formula. Assume step length as $h = 1$.

2. Find the solution of the Laplace equation $u_{xx}+u_{yy} = 0$ in the region R is a square of side 1 units subject to the given boundary conditions $u(x, y) = x - y$, using the standard five point formula. Assume step length as $h = 1/3$.

3. Solve the boundary value problem for the Poisson equation

$$u_{xx} + u_{yy} = x^2 + y^2,\ 0 \le x \le 1,\ 0 \le y \le 1$$
$$u(x, y) = x^2 + y^2 \text{ on the boundary}$$

using the five point formula with square mesh of width $h = 1/3$.

Unit 14

Course Structure

- Partial Differential Equations: Finite difference methods for Parabolic partial differential equations.

14.1 Finite difference method for parabolic partial differential equations

Consider a thin homogeneous, insulated bar or a wire of length l. Let the bar be located on the x-axis on the interval $[0, l]$. Let the rod have a source of heat. For example, the rod may be heated at one end or at the middle point or has some source of heat. Let $u(x, t)$ denote the temperature in the rod at any instant of time t. The problem is to study the flow of heat in the rod. The partial differential equation governing the flow of heat in the rod is given by the parabolic equation

$$u_t = c^2 u_{xx}, \quad 0 \leq x \leq l, \;\; t > 0. \tag{14.1.1}$$

where c^2 is a constant and depends on the material properties of the rod. In order that the solution of the problem exists and is unique, we need to prescribe the following conditions.

(i) Initial condition At time $t = 0$, the temperature is prescribed,

$$u(x, 0) = f(x), 0 \leq x \leq l. \tag{14.1.2}$$

(ii) Boundary conditions Since the bar is of length l, boundary conditions at $x = 0$ and at $x = l$ are to be prescribed. These conditions are of the following types:
(a) Temperatures at the ends of the bar is prescribed

$$u(0, t) = g(t), \;\; u(l, t) = h(t), \;\; t > 0. \tag{14.1.3}$$

(b) One end of the bar, say at $x = 0$, is insulated. This implies the condition that

$$\frac{\partial u}{\partial x} = 0, \quad \text{at} \quad x = 0 \quad \text{for all time} \;\; t.$$

At the other end, the temperature may be prescribed, $u(l, t) = h(t), t > 0$. Alternatively, we may have the condition that the end of the bar at $x = l$ is insulated.

Since both initial and boundary conditions are prescribed, the problem is also called an initial boundary value problem. For our discussion, we shall consider only the boundary conditions given in (14.1.3).

Mesh Generation

Superimpose on the domain $0 \leq x \leq l, t > 0$, a rectangular network of mesh lines. Let the interval $[0, l]$ be divided into M equal parts. Then, the mesh length along the x-axis is $h = l/M$. The points along the x-axis are $x_i = ih,\ i = 0, 1, 2, ..., M$. Let the mesh length along the t-axis be k and define $t_j = jk$. The mesh points are (x_i, t_j) We call t_j as the j-th time level (see Fig.14.1.1). At any point (x_i, t_j), we denote the numerical solution by $u_{i,j}$ and the exact solution by $u(x_i, t_j)$. Finite difference methods are classified into two Finite difference methods are classified into two categories: explicit methods and implicit methods.

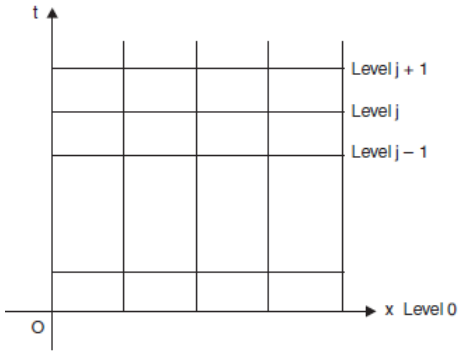


Figure 14.1.1: Nodes

14.1.1 Explicit Method

In explicit methods, the solution at each nodal point on the current time level is obtained by simple computations (additions, subtractions, multiplications and divisions) using the solutions at the previous one or more levels.

Using the relationship between the derivative and forward differences, we have the approximation

$$\left(\frac{\partial u}{\partial t}\right)_{i,j} \approx \frac{1}{k}\left[u_{i,j+1} - u_{i,j}\right]. \tag{14.1.4}$$

Using central differences, we also have the approximation

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j} \approx \frac{1}{h^2}\left[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right]. \tag{14.1.5}$$

Therefore, an approximation to the heat conduction equation (14.1.1) at the point (x_i, t_{j+1}), is

$$\begin{aligned}
&\frac{1}{k}\left[u_{i,j+1} - u_{i,j}\right] = \frac{c^2}{h^2}\left[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right] \\
\implies\ & u_{i,j+1} - u_{i,j} = \lambda\left[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right] \\
\implies\ & u_{i,j+1} = u_{i,j} + \lambda\left[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\right] \\
\implies\ & u_{i,j+1} = \lambda u_{i-1,j} + (1 - 2\lambda)u_{i,j} + \lambda u_{i+1,j}
\end{aligned}$$

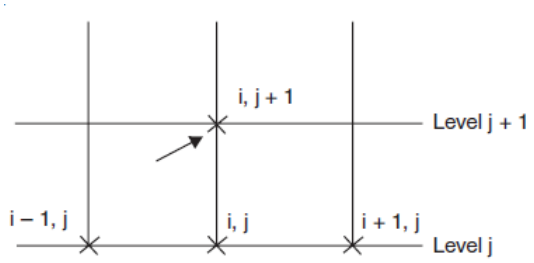


Figure 14.1.2: Schmidt method

where $\lambda = kc^2/h^2$, is called the mesh ratio parameter.

Note that the value $u_{i,j+1}$ at the node (x_i, t_{j+1}) is being obtained explicitly using the values on the previous time level t_j. The nodes that are used in the computations are given in Fig. 14.1.2 This method is called the *Schmidt method.* It is a two level method.

14.1.2 Truncation error of the Schmidt method

We have the method as

$$u_{i,j+1} - u_{i,j} = \lambda\big[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\big].$$

Expanding in Taylor's series, we obtain the left hand and right hand sides as

$$u(x_i, t_j + k) - u(x_i, t_j) = \left[\left\{u + k\frac{\partial u}{\partial t} + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2} + \ldots\right\} - u\right] = \left[k\frac{\partial u}{\partial t} + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2} + \ldots\right]$$

and

$$\begin{aligned}
& \lambda\big[u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\big] \\
= \; & \frac{kc^2}{h^2}\left[\left\{u + h\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3} + \ldots\right\} - 2u + \left\{u - h\frac{\partial u}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2} - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3} + \ldots\right\}\right] \\
= \; & \frac{kc^2}{h^2}\left[h^2\frac{\partial^2 u}{\partial x^2} + \frac{h^4}{12}\frac{\partial^4 u}{\partial x^4} + \ldots\right] = kc^2\left[\frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4} + \ldots\right]
\end{aligned}$$

where all the terms on the right hand sides are evaluated at (x_i, t_j). The truncation error is given by

$$\begin{aligned}
T.E. &= u(x_i, t_j + k) - u(x_i, t_j) - \lambda\big[u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)\big] \\
&= \left[k\frac{\partial u}{\partial t} + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2} + \ldots\right] - kc^2\left[\frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4} + \ldots\right] \\
&= \left[k\frac{\partial u}{\partial t} - c^2\frac{\partial^2 u}{\partial x^2} + \ldots\right] + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2} - \frac{kh^2c^2}{12}\frac{\partial^4}{\partial x^4}
\end{aligned}$$

Now, using the differential equation

$$\frac{\partial u}{\partial t} = c^2\frac{\partial^2 u}{\partial x^2}, \quad \text{and} \quad \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial t}\left(\frac{\partial u}{\partial t}\right) = c^2\frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial x^2}\right) = c^4\frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 u}{\partial x^2}\right) = c^4 \tag{14.1.6}$$

we obtain

$$T.E = \frac{k^2c^4}{2}\frac{\partial^4 u}{\partial x^4} - \frac{kh^2c^2}{12}\frac{\partial^4 u}{\partial x^4} + \ldots = \frac{kh^2c^2}{12}\left((6\lambda - 1)\frac{\partial^4 u}{\partial x^4} + \ldots\right) \tag{14.1.7}$$

The order of the method is given by

$$\text{order} = \frac{1}{k}(T.E) = O(h^2 + k). \tag{14.1.8}$$

Computational procedure

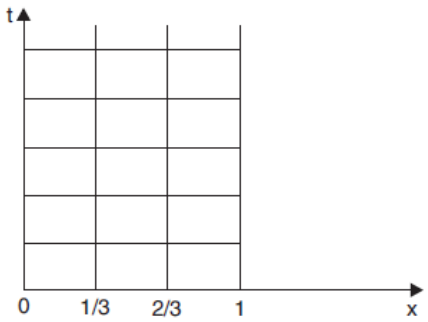
The initial condition $u(x, 0) = f(x)$ gives the solution at all the nodal points on the initial line (level 0). The boundary conditions $u(0, t) = g(t), u(l, t) = h(t), t > 0$ give the solutions at all the nodal points on the boundary lines $x = 0$ and $x = l$, (called boundary points), for all time levels. We choose a value for and h. This gives the value of the time step length k. Alternately, we may choose values for h and k. The solutions at all nodal points, (called interior points), on level 1 are obtained using the explicit method. The computations are repeated for the required number of steps. If we perform m steps of computation, then we have computed the solutions up to time $t_m = mk$. Let us illustrate the method through some problems.

Example 14.1.1. Solve the heat conduction equation $u_t = u_{xx}, 0 \leq x \leq 1$, with $u(x, 0) = \sin(\pi x),\ 0 \leq x \leq 1,\ u(0, t) = u(1, t) = 0$ using the Schmidt method. Assume $h = 1/3$. Compute with (i) $\lambda = 1/2$ for two time steps, (ii) $\lambda = 1/4$ for four time steps, (iii) $\lambda = 1/6$ for six time steps. If the exact solution is $u(x, t) = \exp(-\pi^2 t)\sin(\pi x)$, compare the solutions at time $t = 1/9$.

Solution. The Schmidt method is given by

$$u_{i,j+1} = \lambda u_{i-1,j} + (1 - 2\lambda)u_{i,j} + \lambda u_{i+1,j}$$

We are given $h = 1/3$. Hence, we have four nodes on each mesh line (see Figure below). We have to find the solution at the two interior points.



The initial condition gives the values

$$u\left(\frac{1}{3}, 0\right) = u_{1,0} = \sin\left(\frac{\pi}{3}\right) = \frac{\sqrt{3}}{2}$$
$$u\left(\frac{2}{3}, 0\right) = u_{2,0} = \sin\left(\frac{2\pi}{3}\right) = \frac{\sqrt{3}}{2} = 0.866025$$

The boundary conditions give the values $u_{0,j} = 0, u_{3,j} = 0$, for all j.

(i) We have $\lambda = 1/2, h = 1/3, k = \lambda h^2 = 1/18$. The computations are to be done for two time steps, that is, upto $t = 1/9$. For $\lambda = 1/2$, we get the method

$$u_{i,j+1} = \frac{1}{2}\left(u_{i-1,j} + u_{i+1,j}\right), \quad j = 0, 1; i = 1, 2.$$

We have the following values.

$$\begin{aligned} \text{For } j = 0 : i = 1 : & u_{1,1} = 0.5\left(u_{0,0} + u_{2,0}\right) = 0.5(0 + 0.866025) = 0.433013. \\ i = 2 : & u_{2,1} = 0.5\left(u_{1,0} + u_{3,0}\right) = 0.5(0.866025 + 0) = 0.433013. \end{aligned}$$

$$\begin{aligned} \text{For } j = 1 : i = 1 : & \quad u_{1,2} = 0.5\left(u_{0,1} + u_{2,1}\right) = 0.5(0 + 0.433013) = 0.216507. \\ i = 2 : & \quad u_{2,2} = 0.5\left(u_{1,1} + u_{3,1}\right) = 0.5(0.433013 + 0) = 0.216507. \end{aligned}$$

After two steps $t = 2k = 1/9$. Hence,

$$u\left(\frac{1}{3}, \frac{1}{9}\right) = u\left(\frac{2}{3}, \frac{1}{9}\right) = 0.216507.$$

(ii) We have $\lambda = 1/4, h = 1/3, k = \lambda h^2 = 1/36$. The computations are to be done for four time steps, that is, upto $t = 1/9$. For $\lambda = 1/4$, we get the method

$$u_{i,j+1} = \frac{1}{4}\left(u_{i-1,j} + 2u_{i,j} + u_{i+1,j}\right), \quad j = 0, 1, 2, 3; i = 1, 2.$$

We have the following values.

$$\begin{aligned} \text{For } j = 0 : i = 1 : & \quad u_{1,1} = 0.25\left(u_{0,0} + 2u_{1,0} + u_{2,0}\right) = 0.25[0 + 3(0.866025)] = 0.649519. \\ i = 2 : & \quad u_{2,1} = 0.25\left(u_{1,0} + 2u_{2,0} + u_{3,0}\right) = 0.25[3(0.866025) + 0] = 0.649519. \\ \text{For } j = 1 : i = 1 : & \quad u_{1,2} = 0.25\left(u_{0,1} + 2u_{1,1} + u_{2,1}\right) = 0.25[0 + 3(0.649519)] = 0.487139. \\ i = 2 : & \quad u_{2,2} = 0.25\left(u_{1,1} + 2u_{2,1} + u_{3,1}\right) = 0.25[3(0.649519) + 0] = 0.487139. \\ \text{For } j = 2 : i = 1 : & \quad u_{1,3} = 0.25\left(u_{0,2} + 2u_{1,2} + u_{2,2}\right) = 0.25[0 + 3(0.487139)] = 0.365354. \\ i = 2 : & \quad u_{2,3} = 0.25\left(u_{1,2} + 2u_{2,2} + u_{3,2}\right) = 0.25[3(0.487139) + 0] = 0.365354. \\ \text{For } j = 3 : i = 1 : & \quad u_{1,4} = 0.25\left(u_{0,3} + 2u_{1,3} + u_{2,3}\right) = 0.25[0 + 3(0.365354)] = 0.274016. \\ i = 2 : & \quad u_{2,4} = 0.25\left(u_{1,3} + 2u_{2,3} + u_{3,3}\right) = 0.25[3(0.365354) + 0] = 0.274016. \end{aligned}$$

After four steps $t = 4k = 1/9$. Hence,

$$u\left(\frac{1}{3}, \frac{1}{9}\right) = u\left(\frac{2}{3}, \frac{1}{9}\right) = 0.274016$$

(iii) We have $\lambda = 1/6, h = 1/3, k = \lambda h^2 = 1/54$. The computations are to be done for six time steps, that is, upto $t = 1/9$. For $\lambda = 1/6$, we get the method

$$u_{i,j+1} = \frac{1}{6}\left(u_{i-1,j} + 4u_{i,j} + u_{i+1,j}\right), \quad j = 0, 1, 2, 3, 4, 5; i = 1, 2.$$

We have the following values.

$$\text{For } j=0: i=1: \quad u_{1,1}=\frac{1}{6}\left(u_{0,0}+4u_{1,0}+u_{2,0}\right)=\frac{1}{6}[0+5(0.866025)]=0.721688.$$

$$i=2: \quad u_{2,1}=\frac{1}{6}\left(u_{1,0}+4u_{2,0}+u_{3,0}\right)=\frac{1}{6}[5(0.866025)+0]=0.721688.$$

$$\text{For } j=1: i=1: \quad u_{1,2}=\tfrac{1}{6}\left(u_{0,1}+4u_{1,1}+u_{2,1}\right)=\tfrac{1}{6}[0+5(0.721688)]=0.601407$$

$$i=2: \quad u_{2,2}=\frac{1}{6}\left(u_{1,1}+4u_{2,1}+u_{3,1}\right)=\frac{1}{6}[5(0.721688)+0]=0.601407.$$

$$\text{For } j=2: i=1: \quad u_{1,3}=\tfrac{1}{6}\left(u_{0,2}+4u_{1,2}+u_{2,2}\right)=\tfrac{1}{6}[0+5(0.601407)]=0.501173.$$

$$i=2: \quad u_{2,3}=\frac{1}{6}\left(u_{1,2}+4u_{2,2}+u_{3,2}\right)=\frac{1}{6}[5(0.601407)+0]=0.501173.$$

$$\text{For } j=3: i=1: \quad u_{1,4}=\tfrac{1}{6}\left(u_{0,3}+4u_{1,3}+u_{2,3}\right)=\tfrac{1}{6}[0+5(0.501173)]=0.417644.$$

$$i=2: \quad u_{2,4}=\frac{1}{6}\left(u_{1,3}+4u_{2,3}+u_{3,3}\right)=\frac{1}{6}[5(0.501173)+0]=0.417644.$$

$$\text{For } j=4: i=1: \quad u_{1,5}=\tfrac{1}{6}\left(u_{0,4}+4u_{1,4}+u_{2,4}\right)=\tfrac{1}{6}[0+5(0.417644)]=0.348037.$$

$$i=2: \quad u_{2,5}=\frac{1}{6}\left(u_{1,4}+4u_{2,4}+u_{3,4}\right)=\frac{1}{6}[5(0.417644)+0]=0.348037.$$

$$\text{For } j=5: i=1: \quad u_{1,6}=\tfrac{1}{6}\left(u_{0,5}+4u_{1,5}+u_{2,5}\right)=\tfrac{1}{6}[0+5(0.348037)]=0.290031.$$

$$i=2: \quad u_{2,6}=\frac{1}{6}\left(u_{1,5}+4u_{2,5}+u_{3,5}\right)=\frac{1}{6}[5(0.348037)+0]=0.290031.$$

After six steps $t=6k=1/9$. Hence,

$$u\left(\frac{1}{3},\frac{1}{9}\right)=u\left(\frac{2}{3},\frac{1}{9}\right)=0.290031.$$

The magnitudes of errors at $x=1/3$ and at $x=2/3$ are same. The exact solution at $t=1/9$ is

$$u\left(\frac{1}{3},\frac{1}{9}\right)=u\left(\frac{2}{3},\frac{1}{9}\right)=\exp\left(\frac{-\pi^2}{9}\right)\sin\left(\frac{\pi}{3}\right)=0.289250.$$

The magnitudes of errors are the following:

$$\begin{aligned}\lambda=1/2: &\quad |0.216507-0.289250|=0.072743.\\ \lambda=1/4: &\quad |0.274016-0.289250|=0.015234.\\ \lambda=1/6: &\quad |0.290031-0.289250|=0.000781.\end{aligned}$$

We note that the higher order method produced better results.

Exercise 14.1.2.

1. Solve $u_{xx}=32u_t, 0\le x\le 1$, taking $h=0.5$ and $u(x,0)=0, 0\le x\le 1, u(0,t)=0, u(1,t)=t, t>0$. Use an explicit method with $\lambda=1/2$. Compute for four time steps.
2. Solve $u_t=u_{xx}, 0\le x\le 1$, with $u(x,0)=x(1-x), 0\le x\le 1$ and $u(0,t)=u(1,t)=0$ for all $t>0$. Use explicit method with $h=0.25$ and $=0.25$. Compute for four time steps.
3. Solve $u_{xx}=16u_t, 0\le x\le 1$, with $u(x,0)=x(1-x), 0\le x\le 1$ and $u(0,t)=u(1,t)=0$ for all $t>0$. Use Schmidt method with $h=0.25$ and $\lambda=1/6$. Compute for four time steps.
4. Solve $u_{xx}=4u_t, 0\le x\le 1$, with $u(x,0)=2x$ for $x\in[0,1/2]$ and $2(1-x)$ for $x\in[1/2,1]$; and $u(0,t)=u(1,t)=0$ for all $t>0$. Use Schmidt method with $h=0.25$ and $\lambda=0.5$. Compute for four time steps.

14.1.3 Implicit method

Explicit methods have the disadvantage that they have a stability condition on the mesh ratio parameter λ. We have seen that the Schmidt method is stable for $\lambda \leq 0.5$. This condition severely restricts the values that can be used for the step lengths h and k. In most practical problems, where the computation is to be done up to large value of t, these methods are not useful because the time taken is too high. In such cases, we use the implicit methods. We shall discuss the most popular and useful method called the Crank-Nicolson method. There are a number of ways of deriving this method. We describe one of the simple ways. Denote ∇_t as the backward difference in the time direction. We write the relation

$$k\frac{\partial u}{\partial t} = -\log(1-\nabla_t)\,u = \left[\nabla_t + \frac{1}{2}\nabla_t^2 + \frac{1}{3}\nabla_t^3 + \ldots\right] u. \tag{14.1.9}$$

Now, approximate

$$k\frac{\partial u}{\partial t} \approx \left[\nabla_t + \frac{1}{2}\nabla_t^2\right] u \approx \left[\frac{\nabla_t}{1-(1/2)\nabla_t}\right] u. \tag{14.1.10}$$

If we expand the operator on the right hand side, we get

$$\frac{\nabla_t}{1-(1/2)\nabla_t} = \nabla_t\left[1-\frac{1}{2}\nabla_t\right]^{-1} = \nabla_t\left[1+\frac{1}{2}\nabla_t+\frac{1}{4}\nabla_t^2+\ldots\right]$$

which agrees with the first two terms on the right hand side of (14.1.9). Applying the differential equation at the nodal point $(i, j+1)$, (see Fig. 14.1.3), we obtain

$$\left(\frac{\partial u}{\partial t}\right)_{i,j+1} = c^2\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j+1}.$$

Using the approximation given in (14.1.10) to left hand side and the central difference approximation (14.1.5) to the right hand side, we obtain

$$\begin{aligned}
&\frac{1}{k}\left[\frac{\nabla_t}{1-(1/2)\nabla_t}\right] u_{i,j+1} = \frac{c^2}{h^2}\delta_x^2 u_{i,j+1}\\
\text{or,}\quad &\nabla_t u_{i,j+1} = \frac{kc^2}{h^2}\left(1-\frac{1}{2}\nabla_t\right)\delta_x^2 u_{i,j+1},\\
\text{or,}\quad &\nabla_t u_{i,j+1} = \lambda\left(\delta_x^2 u_{i,j+1} - \frac{1}{2}\nabla_t\delta_x^2 u_{i,j+1}\right),\\
\text{or,}\quad &\nabla_t u_{i,j+1} = \lambda\left(\delta_x^2 u_{i,j+1} - \frac{1}{2}\delta_x^2\nabla_t u_{i,j+1}\right).\\
\text{or,}\quad &\nabla_t u_{i,j+1} = \lambda\left(\delta_x^2 u_{i,j+1} - \frac{1}{2}\delta_x^2\left\{u_{i,j+1}-u_{i,j}\right\}\right),\\
\text{or,}\quad &\nabla_t u_{i,j+1} = \lambda\left(\delta_x^2 u_{i,j+1} - \frac{1}{2}\left\{\delta_x^2 u_{i,j+1}-\delta_x^2 u_{i,j}\right\}\right),\\
\text{or,}\quad &\nabla_t u_{i,j+1} = \frac{\lambda}{2}\left(\delta_x^2 u_{i,j+1} + \delta_x^2 u_{i,j}\right), \qquad (14.1.11)\\
\text{or,}\quad &u_{i,j+1} - u_{i,j} = \frac{\lambda}{2}\left(\delta_x^2 u_{i,j+1} + \delta_x^2 u_{i,j}\right),\\
\text{or,}\quad &u_{i,j+1} - \frac{\lambda}{2}\delta_x^2 u_{i,j+1} = u_{i,j} + \frac{\lambda}{2}\delta_x^2 u_{i,j}\\
\text{or,}\quad &u_{i,j+1} - \frac{\lambda}{2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} = u_{i,j} + \frac{\lambda}{2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}),\\
\text{or,}\quad &-\frac{\lambda}{2}u_{i-1,j+1} + (1+\lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,j+1} = \frac{\lambda}{2}u_{i-1,j} + (1-\lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j} \qquad (14.1.12)
\end{aligned}$$

where $\lambda = kc^2/h^2$. This method is called the Crank-Nicolson method. The nodal points that are used in the method are given in Fig. 14.1.3.

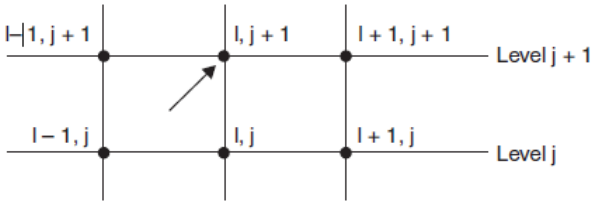


Figure 14.1.3: Nodes in Crank-Nicolson method.

Remark 14.1.3.

1. The order of the Crank-Nicolson method is $O(k^2 + h^2)$.

2. Implicit methods often have very strong stability properties. Stability analysis of the Crank-Nicolson method shows that the method is stable for all values of the mesh ratio parameter λ. This implies that there is no restriction on the values of the mesh lengths h and k. Depending on the particular problem that is being solved, we may use sufficiently large values of the step lengths. Such methods are called unconditionally stable methods.

3. The system of equations that is obtained if we apply the Crank- Nicolson method is a tri-diagonal system of equations. It uses the three consecutive unknowns $u_{i-1,j+1}, u_{i,j+1}$ and $u_{i+1,j+1}$ on the current time level. This is the advantage of the method.

Computational procedure

The initial condition $u(x, 0) = f(x)$ gives the solution at all the nodal points on the initial line (level 0). The boundary conditions $u(0, t) = g(t),\ u(l, t) = h(t), t > 0$ give the solutions at all the nodal points on the lines $x = 0$ and $x = l$ for all time levels. We choose a value for λ and h. This gives the value of the time step length k. Alternately, we may choose the values for h and k. The difference equations at all nodal points on the first time level are written. This system of equations is solved to obtain the values at all the nodal points on this time level. The computations are repeated for the required number of steps. If we perform m steps of computation, then we have computed the solutions up to time $t_m = mk$.

Example 14.1.4. Solve the equation $u_t = u_{xx}$ subject to the conditions

$$u(x, 0) = \sin(\pi x), 0 \le x \le 1, u(0, t) = u(1, t) = 0$$

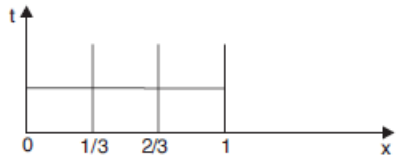
using the Crank-Nicolson method with, $h = 1/3, k = 1/36$. Do one time step.

Solution. We have

$$c^2 = 1, h = \frac{1}{3}, k = \frac{1}{36}, \lambda = \frac{kc^2}{h^2} = \frac{1}{36}(9) = \frac{1}{4}.$$

Crank-Nicolson method is given by

$$-\frac{\lambda}{2}u_{i-1,j+1} + (1 + \lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,j+1} = \frac{\lambda}{2}u_{i-1,j} + (1 - \lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j}$$



For $\lambda = 1/4$, we have the method as

$$-\frac{1}{8}u_{i-1,j+1} + \frac{5}{4}u_{i,j+1} - \frac{1}{8}u_{i+1,j+1} = \frac{1}{8}u_{i-1,j} + \frac{3}{4}u_{i,j} + \frac{1}{8}u_{i+1,j}$$

$$\text{or, } -u_{i-1,j+1} + 10u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + 6u_{i,j} + u_{i+1,j}, \quad j = 0; i = 1, 2.$$

The initial condition gives the values

$$u_{0,0} = 0, u_{1,0} = \sin(\pi/3) = (\sqrt{3}/2) = u_{2,0}, u_{3,0} = 0.$$

The boundary conditions give the values $u_{0,j} = 0 = u_{3,j}$ for all j,
We have the following equations.

For $j = 0, i = 1 : -u_{0,1} + 10u_{1,1} - u_{2,1} = u_{0,0} + 6u_{1,0} + u_{2,0}$

or $\quad 10u_{1,1} - u_{2,1} = \frac{6\sqrt{3}}{2} + \frac{\sqrt{3}}{2} = \frac{7\sqrt{3}}{2} = 6.06218.$

$i = 2 : \quad -u_{1,1} + 10u_{2,1} - u_{3,1} = u_{1,0} + 6u_{2,0} + u_{3,0}$

or $-u_{1,1} + 10u_{2,1} = u_{1,0} + 6u_{2,0} = \frac{\sqrt{3}}{2} + \frac{6\sqrt{3}}{2} = \frac{7\sqrt{3}}{2} = 6.06218.$

Subtracting the two equations, we get $11u_{1,1} - 11u_{2,1} = 0$.

Hence, $u_{1,1} = u_{2,1}$. The solution is given by

$$u_{1,1} = u_{2,1} = \frac{6.06218}{9} = 0.67358$$

Exercise 14.1.5. 1. Solve $u_{xx} = u_t$ in $0 < x < 2, t > 0$,

$$u(0,t) = u(2,t) = 0, t > 0 and u(x,0) = sin(\pi x/2), 0 \leq x \leq 2,$$

using $\delta x = 0.5, \delta t = 0.25$ for one time step by Crank-Nicolson implicit finite difference method.

2. Solve by Crank-Nicolson method the equation $u_{xx} = u_t$ subject to

$$u(x,0) = 0, u(0,t) = 0 \text{ and } u(1,t) = t,$$

for two time steps.

3. Solve the heat equation $u_t = u_{xx}, 0 \leq x \leq 1$, subject to the initial and boundary conditions

$$u(x,0) = sin(2\pi x), \;\; 0 \leq x \leq 1, \;\; u(0,t) = u(1,t) = 0$$

using the Crank-Nicolson method with, $h = 0.25, \lambda = 0.8$. Integrate for two time steps. If the exact solution of the problem is $u(x,t) = exp(-4\pi^2 t)\sin(2\pi x)$, find the magnitudes of the errors on the second time step.

4. Find the solution of the equation $4u_t = u_{xx}, 0 \leq x \leq 1$ subject to the conditions

$$u(x,0) = 3x, \;\; \text{for} \;\; x \in [0, 1/2] \;\; \text{and} \;\; 3(1-x), \; x \in [1/2, 1], \; u(0,t) = 0 = u(1,t)$$

using the Crank-Nicolson method with $h = 0.25, k = 1/32$. Integrate for two time steps.

References

1. Introductory Methods of Numerical Analysis, *S. S Sastry*
2. Numerical Methods, *S. R. K. Iyengar, R. K. Jain*
3. Numerical Methods: Problems and Solutions, *Mahinder Kumar Jain, S.R.K. Iyengar, R.K. Jain*
4. Numerical Analysis, *Sivarama Krishna Das*
5. Numerical Analysis: Mathematics of Scientific Computing, *David Kincaid*
6. Numerical Analysis, *L. Richard, J. Burden, Douglas Faires*

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER II

SELF LEARNING MATERIAL

PAPER : DSE 2.4
(Pure Stream)

Differential Geometry II

Topology II

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

Content Writers

Block - I : Differential Geometry II	Dr. Avijit Sarkar Assistant Professor, Department of Mathematics, University of Kalyani
Block - II : Topology II	Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani

July, 2022

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 Message

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, 2020 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.) Manas Kumar Sanyal, 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 well 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 coordinated 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 written and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Board of Studies Members of Department of Mathematics, Directorate of Open and Distance Learning (DODL), University of Kalyani

Sl No.	Name & Designation	Role
1	Dr. Animesh Biswas, Professor & Head, Dept. of Mathematics, KU	Chairperson
2	Dr. Pulak Sahoo, Professor, Dept. of Mathematics, KU	Member
3	Dr. Sahidul Islam, Assistant Professor, Dept. of Mathematics, KU	Member
4	Dr. Sushanta Kumar Mohanta, Professor, Dept. of Mathematics, West Bengal State University	External Nominated Member
5	Dr. Biswajit Mallick, Assistant Professor (Cont.), Department of Mathematics, DODL, KU	Member
6	Ms. Audrija Choudhury, Assistant Professor (Cont), Department of Mathematics, DODL, KU	Member
7	Director, DODL, KU	Convener

Discipline Specific Elective Paper

PURE STREAM

DSE 2.4
Marks : 100 (SEE : 80; IA : 20); Credit : 6

Differential Geometry II (Marks : 50 (SEE: 40; IA: 10))
Topology II (Marks : 50 (SEE: 40; IA: 10))

Syllabus

Block I

- **Unit 1:** Curves in the plane and space, surfaces in three-dimension, Smooth surface
- **Unit 2:** Tangents and derivatives, normal and orientability, Examples of surfaces.
- **Unit 3:** The first fundamental form, Length of curves on surfaces
- **Unit 4:** Isometries of surfaces, Conformal mapping of surfaces
- **Unit 5:** Curvature of surfaces, The second fundamental form, The Gauss and Weingarten map
- **Unit 6:** Normal and geodesic curvatures, Parallel transport and covariant derivative.
- **Unit 7:** Gaussian, mean and principal curvatures
- **Unit 8:** Gauss Theorem Egregium, Minimal surface
- **Unit 9:** The Gauss Bonnet Theorem. Abstract differentiable manifolds and examples, Tangent Spaces
- **Unit 10:** Continuation of Unit 9

Block II

- **Unit 11:** Connectedness: Examples, various characterizations and basic properties. Connectedness on the real line.
- **Unit 12:** Components and quasi components. Path connectedness and path components.
- **Unit 13:** Compactness: Characterizations and basic properties of compactness, Lebesgue, lemma. Sequential compactness
- **Unit 14:** BW Compactness and countable compactness. Local compactness and Baire Category Theorem.
- **Unit 15:** Identification spaces: Constructing a Mobius strip, identification topology, Orbit spaces.
- **Unit 16:** Continuation of Unit 15
- **Unit 17:** Some Matrix Lie Groups: Some elementary properties of topological groups.
- **Unit 18:** $GL(n, \mathbb{R})$ as a topological group and its subgroups.
- **Unit 19:** Fundamental groups, calculation of fundamental group of S.
- **Unit 20:** Continuation of Unit 19

Contents

Unit 1

Course Structure

- Curves in the plane and space
- Surfaces in three-dimension
- Smooth surface

1.1 Curves

What are curves? As we have read so far, the line $y - 2x = 1$ is a curve (even though its not curved). A curve can be said to be a generalisation of a straight line whose curvature is not zero. Say a circle $x^2 + y^2 = 1$, or the parabola, $y = x^2$. All these curves can be described by means of their Cartesian equation $f(x, y) = c$, where, f is a function of x and y and c is a constant. In this perspective, a curve may be considered as the set of points, namely

$$\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | f(x, y) = c\}$$

These are all curves in the $\mathbb{R}^2$. In $\mathbb{R}^3$, for example, the x-axis is the straight line given by $y = 0,\ z = 0$. More generally, a curve in $\mathbb{R}^3$ is defined as a pair of equations

$$f_1(x, y, z) = c_1,\ f_2(x, y, z) = c_2$$

There is another way to define curves which is more useful in many situations, which is viewed as the path traced out by a moving point in space.

Definition 1.1.1. A parametrized curve is a path in the xy-plane traced out by the point $(x(t), y(t))$ as the parameter t ranges over an interval I.

$$C = \{(x(t), y(t)) : t \in I\}$$

For example, the graph of the function $y = f(x)$, $x \in I$ is a curve C parametrised by the equations

$$x(t) = t,\ y(t) = f(t),\ t \in I$$

The above equations give a parametrisation of the curve $y = f(x)$. Also, a given curve may have more than one parametrisation. For example, the curve $y = 2x$, $x \in [1, 3]$ can be parametrised as

$$x(t) = t,\ y(t) = 2t,\ t \in [1, 3]$$

It can also be parametrised as

$$x(t) = t + 1,\ y(t) = 2t + 2, t \in [0, 2]$$

Example 1.1.2. Consider the parabola $x = 1 - y^2$, $-1 \leq y \leq 1$. Then we can think of two parametrisations of it as

1. When we set $y(t) = t$ and $x(t) = 1 - t^2$, $t \in [-1, 1]$. Changing the domain to all real t gives us the whole parabola.

2. When we set $y(t) = \cos t$ and $x(t) = 1 - cos^2 t$, $t \in [0, \pi]$. Changing the domain to all real t does not give us any more of the parabola.

Example 1.1.3. Now, consider the circle $x^2 + y^2 = 1$. If we take $x(t) = t$ and $y(t) = \sqrt{1 - t^2}$, then this will only represent the upper half of the circle since $\sqrt{1 - t^2} \geq 0$ always. Similarly, we could not have taken $y(t) = -\sqrt{1 - t^2}$ since that would have traced the lower half of the circle. So, let us take the parametrisation as $x(t) = \cos t$ and $y(t) = \sin t$, where $t \in [0, 2\pi]$. Then we see that the curve $(x(t), y(t))$ traces the whole of the circle as t traverses $[0, 2\pi]$.

There are other several examples which we will omit now since we have already studied those in DG I. We will straightaway move on to surfaces in three dimensions.

1.1.1 Surface in Three Dimensions

Just like a curve is the basic building block for figures in a plane, a surface is the basic building block for figures in space. A surface is essentially a curve with depth. Curves and surfaces are analogous in many ways. If you think of a curve as being the trace of the motion of a point in a plane, a surface is like the trace of the motion of a curve in space. Surfaces are continuous, meaning that given two points on a surface, you can start from one and reach the other without leaving that surface. Just like a curve is still one-dimensional, a surface, although it exists in three dimensions, is still two-dimensional. For example, when you build a curve by tracing the motion of a point, that curve, although it spans both length and width, has no width of its own. The curve doesn't have area, it only has length, one dimension. Similarly, a surface can span more than one plane, but it still does not have depth of its own. It only has two dimensions, length and width. We will work mostly with the simplest surface, a plane. Before formally defining a surface, we will see certain useful definitions.

Definition 1.1.4. A subset U of $\mathbb{R}^n$ is called Open if, whenever $\mathbf{a}$ is a point in U, there exists a positive ϵ such that

$$\mathbf{a} \in U \text{ and } ||\mathbf{u} - \mathbf{a}|| < \epsilon \text{ implies } \mathbf{u} \in U$$

The whole of $\mathbb{R}^n$ is an open set. Also, the open ball

$$\mathcal{D}_r(\mathbf{a}) = \{\mathbf{u} \in \mathbb{R}^n : ||\mathbf{u} - \mathbf{a}|| < r\},$$

with centre at $\mathbf{a}$ and radius $r > 0$. However, the set

$$\overline{\mathcal{D}}_r(\mathbf{a}) = \{\mathbf{u} \in \mathbb{R}^n : ||\mathbf{u} - \mathbf{a}|| \leq r\}$$

is not open.

Definition 1.1.5. If X and Y are subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, a map $f : X \to Y$ is said to be continuous at a point $\mathbf{a} \in X$ if given any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\mathbf{u} \in X \text{ and } ||\mathbf{u} - \mathbf{a}|| < \delta \text{ implies } ||f(\mathbf{u}) - f(\mathbf{a})|| < \epsilon$$

If f is continuous at every point of X, then it is said to be continuous on X. Composites of continuous maps are continuous.

Continuous maps can also be equivalently defined as follows:
f is continuous if and only if, for any open set $V \in \mathbb{R}^n$, there is an open set U in $\mathbb{R}^m$ such that $U \cap X = \{x \in X | f(x) \in V\}$.

Let us now formally define a surface in $\mathbb{R}^3$.

Definition 1.1.6. A subset S of $\mathbb{R}^3$ is a surface if, for every $\mathbf{p} \in S$, there is an open set U in $\mathbb{R}^2$ and an open set W in $\mathbb{R}^3$ containing $\mathbf{p}$ such that $S \cap W$ is homeomorphic to U. A subset of a surface S of the form $S \cap W$, where W is an open subset of $\mathbb{R}^3$, is called an open subset of S. A homeomorphism $\sigma : U \to S \cap W$ as in this definition is called a surface patch or parametrisation of the open subset $S \cap W$ of S. A collection of such surface patches whose images cover the whole of S is called an atlas of S.

A surface in three dimensions is often presented by a defining equation; one can also describe it by parameters, but one requires two of them: varying one parameter can only produce a one dimensional figure, some sort of curve.

Example 1.1.7. Every plane in $\mathbb{R}^3$ is a surface with an atlas consisting of a single surface patch. In fact, let **a** be a point on the plane, and let **p** and **q** be two unit vectors that are parallel to the plane and perpendicular to each other. If **v** is any point of the plane, **v-a** is parallel to the plane, and so

$$\textbf{v-a} = u\mathbf{p} + v\mathbf{q}$$

for some scalars u and v. Thus, the desired surface patch is

$$\sigma(u, v) = \mathbf{a} + u\mathbf{p} + v\mathbf{q}$$

and its inverse map is

$$\sigma^{-1}(\mathbf{v}) = ((\textbf{v-a}).\mathbf{p}, (\textbf{v-a}).\mathbf{q}).$$

These formulae make it clear that σ and σ^{-1} are continuous, and hence σ is a homeomorphism.

Example 1.1.8. A circular cylinder is the set of points of $\mathbb{R}^3$ that are at a fixed distance (the radius of the cylinder) from a fixed straight line (its axis). For example, the circular cylinder of radius 1 and axis the z–axis, which we shall call the unit cylinder, is

$$S = \{(x, y, z) \in \mathbb{R}^3 | x^2 + y^2 = 1\}$$

It can be parametrised as

$$\sigma(u, v) = (\cos u, \sin v, v)$$

Clearly, $\sigma(u, v) \in S$ for all $(u, v) \in \mathbb{R}^2$, and every point of S is of this form. Moreover σ is continuous. But it is not injective, and so is not a homeomorphism because $\sigma(u, v) = \sigma(u + 2\pi, v)$ for all (u, v). To get an injective map we can restrict u to lie in an interval of length $\leq 2\pi$, say $0 \leq u < 2\pi$. However, although the restriction $\sigma|_V$ to σ where

$$V = \{(u, v) \in \mathbb{R}^2 | 0 \leq u < 2\pi\}$$

is injective, V is not an open subset of $\mathbb{R}^2$ and so $\sigma|_V$ is not a surface patch. The largest open subset of $\mathbb{R}^2$ contained in V is

$$U = \{(u, v) \in \mathbb{R}^2 | 0 < u < 2\pi\}$$

and the restriction $\sigma|_U$ to U is a surface patch. However, $\sigma|_U$ does not cover the whole of S, but only the open subset obtained by removing the line $x = 1, y = 0$ from S.

To get an atlas for S, we therefore need at least one more surface patch. We can take $\sigma|_{\tilde{U}}$, where

$$\tilde{U} = \{(u, v) \in \mathbb{R}^2 | -\pi < u < \pi\}$$

which covers the open subset of S obtained by removing the line $x = -1, y = 0$. Every point of S is in the image of at least one of the surface patches $\sigma|_U$, $\sigma|_{\tilde{U}}$. So, $\{\sigma|_U, \sigma|_{\tilde{U}}\}$ is an atlas for S, and S is a surface.

1.1.2 Smooth Surfaces

In Differential Geometry we use calculus to analyse surfaces (and other geometric objects). We must be able to make sense of the statement that a function on a surface is differentiable, for example. For this, we have to consider surfaces with some extra structure.

First, if U is an open subset of $\mathbb{R}^m$, we say that a map $f : U \to \mathbb{R}^n$ is smooth if each of the n components of f, have continuous partial derivatives of all orders. For example, if $m = 2$ and $n = 3$, and

$$f(u, v) = (f_1(u, v), f_2(u, v), f_3(u, v)),$$

then

$$\frac{\partial f}{\partial u} = \left(\frac{\partial f_1}{\partial u}, \frac{\partial f_2}{\partial u}, \frac{\partial f_3}{\partial u}\right), \ \frac{\partial f}{\partial v} = \left(\frac{\partial f_1}{\partial v}, \frac{\partial f_2}{\partial v}, \frac{\partial f_3}{\partial v}\right)$$

and similarly for higher derivatives. We often use the following abbreviations:

$$\frac{\partial f}{\partial u} = f_u, \ \frac{\partial f}{\partial v} = f_v,$$

$$\frac{\partial^2 f}{\partial u^2} = f_{uu}, \ \frac{\partial^2 f}{\partial u \partial v} = f_{uv}, \ \frac{\partial^2 f}{\partial v \partial u} = f_{vu}, \ \frac{\partial^2 f}{\partial v^2} = f_{vv},$$

and so on. From advanced calculus we know that $f_{uv} = f_{vu}$, if f is smooth.

It now makes sense to say that a surface patch $\sigma : U \to \mathbb{R}^3$ is smooth. We have further definitions.

Definition 1.1.9. A surface patch $\sigma : U \to \mathbb{R}^3$ is called regular if it is smooth and the vectors σ_u and σ_v are linearly independent at all points $(u, v) \in U$. Equivalently, σ should be smooth and the vector product $\sigma_u \times \sigma_v$ should be non-zero at every point of U.

Definition 1.1.10. If S is a surface, an allowable surface patch for S is a regular surface patch $\sigma : U \to \mathbb{R}^3$ such that σ is a homeomorphism from U to an open subset of S. A **smooth surface** is a surface S such that, for any point $\mathbf{p} \in S$, there is an allowable surface patch σ such that $\mathbf{p} \in \sigma(U)$. A collection $\mathcal{A}$ of allowable surface patches for a surface S such that every point of S is in the image of at least one patch in $\mathcal{A}$ is called an atlas for the smooth surface S.

Example 1.1.11. The plane we have dealt with earlier is a smooth surface, since

$$\sigma(u, v) = \mathbf{a} + u\mathbf{p} + v\mathbf{q}$$

is clearly smooth and $\sigma_u = \mathbf{p}$ and $\sigma_v = \mathbf{q}$ are linearly independent because $\mathbf{p}$ and $\mathbf{q}$ were chosen to be perpendicular unit vectors.

Example 1.1.12. The unit cylinder we have previously dealt with is also a smooth surface since

$$\sigma(u, v) = (\cos u, \sin u, v)$$

is clearly smooth and

$$\sigma_u = (-\sin u, \cos u, 0),\ \sigma_v = (0, 0, 1)$$

are obviously linearly independent for all (u, v), so $\sigma|_U$ and $\sigma|_{\tilde{U}}$ are regular surface patches.

We have the following important result in this connection:

Theorem 1.1.13. Let U and $\tilde{U}$ be open subsets of $\mathbb{R}^2$ and let $\sigma : U \to \mathbb{R}^3$ be a regular surface patch. Let $\Phi : \tilde{U} \to U$ be a bijective smooth map with smooth inverse map $\Phi^{-1} : U \to \tilde{U}$. Then, $\tilde{\sigma} = \sigma \circ \Phi : \tilde{U} \to \mathbb{R}^3$ is a regular surface patch.

Proof. The patch $\tilde{\sigma}$ is smooth because any composite of smooth maps is smooth. As for regularity, let $(u, v) = \Phi(\tilde{u}, \tilde{v})$. By the chain rule,

$$\tilde{\sigma}_{\tilde{u}} = \frac{\partial u}{\partial \tilde{u}}\sigma_u + \frac{\partial v}{\partial \tilde{u}}\sigma_v,\ \tilde{\sigma}_{\tilde{v}} = \frac{\partial u}{\partial \tilde{v}}\sigma_u + \frac{\partial v}{\partial \tilde{v}}\sigma_v,$$

so,

$$\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}} = \left(\frac{\partial u}{\partial \tilde{u}}\frac{\partial v}{\partial \tilde{v}} - \frac{\partial u}{\partial \tilde{v}}\frac{\partial v}{\partial \tilde{u}}\right)\sigma_u \times \sigma_v \tag{1.1.1}$$

The scalar on the right-hand side of this equation is the determinant of the Jacobian matrix

$$J(\Phi) = \begin{bmatrix} \frac{\partial u}{\partial \tilde{u}} & \frac{\partial v}{\partial \tilde{u}} \\ \frac{\partial u}{\partial \tilde{v}} & \frac{\partial v}{\partial \tilde{v}} \end{bmatrix}$$

of Φ. We recall from calculus that, if ψ and $\tilde{\psi}$ are two smooth maps between open sets of $\mathbb{R}^2$,

$$J(\tilde{\psi} \circ \psi) = J(\tilde{\psi})J(\psi)$$

Taking $\psi = \Phi$ and $\tilde{\psi} = \Phi^{-1}$, we see that $J(\Phi^{-1}) = J(\Phi)^{-1}$. In particular, $J(\Phi)$ is invertible, so its determinant is non-zero and equation (1.1.1) shows that $\tilde{\sigma}$ is regular. □

If regular surface patches σ and $\tilde{\sigma}$ related as in this theorem, then $\tilde{\sigma}$ is said to be a reparametrisation of σ, and that Φ is a reparametrisation map.

Exercises

1. Show that the curcular cylinder $S = \{(x, y, z) \in R^3 | x^2 + y^2 = 1\}$ can be covered by a single surface patch, and so is a surface.

2. Define a surface patch $\sigma^x_{\pm} : U \to R^3$ for the unit sphere by solving the equation $x^2 + y^2 + z^2 = 1$ for x in terms of y and z :
$$\sigma^x_{\pm}(u, v) = (\pm\sqrt{1 - u^2 - v^2}, u, v),$$
defined on the open set $U = \{(u, v) \in R^2 < 1\}$. Define $\sigma^y_{\pm}$ and $\sigma^z_{\pm}$ respectively. Show that these six patches give the sphere the structure of a surface.

3. Show that if $f(x, y)$ is a smooth function, its graph $\{(x, y, z) \in R^3 | z = f(x, y)\}$ is a smooth surface with atlas consisting of the single regular surface patch $\sigma(u, v) = (u, v, f(u, v))$.

4. If S is a smooth surface, define the notion of a smooth function $S \to R$. Show that if S is a smooth surface, each component of the inclusion map $S \to R^3$ is a smooth function $S \to R$.

5. Show that translation and invertible linear transformations of R^3 take smooth surfaces to smooth surfaces.

Unit 2

Course Structure

- Tangents and derivatives
- Normal and orientability

2.1 Tangents and Derivatives

Definition 2.1.1. A tangent vector to a surface S at a point $\mathbf{p} \in S$ is the tangent vector at $\mathbf{p}$ of a curve in S passing through $\mathbf{p}$. The tangent space $T_{\mathbf{p}}S$ of S at $\mathbf{p}$ is the set of all tangent vectors to S at $\mathbf{p}$.

To understand the tangent space $T_{\mathbf{p}}S$, choose a surface patch $\sigma : U \to \mathbb{R}^3$ of S such that $\mathbf{p}$ is the image of σ, say $\sigma(u_0, v_0) = \mathbf{p}$. If a curve γ lies in S and passes through $\mathbf{p}$ when $t = t_0$, say, there are functions $u(t)$ and $v(t)$ such that

$$\gamma(t) = \sigma(u(t), v(t)) \tag{2.1.1}$$

for all values of t close to t_0, and $u(t_0) = u_0$, $v(t_0) = v_0$. The functions u and v are necessarily smooth; conversely, it is obvious that if $t \mapsto (u(t), v(t))$ is smooth, then equation (2.1.1) defines a curve lying in S.

Theorem 2.1.2. Let $\sigma : U \to \mathbb{R}^3$ be a patch of a surface S containing a point $\mathbf{p} \in S$, and let (u, v) be coordinates in U. The tangent space to S at $\mathbf{p}$ is the vector subspace of $\mathbb{R}^3$ spanned by the vectors σ_u and σ_v(the derivatives are evaluated at the point $(u_0, v_0) \in U$ such that $\sigma(u_0, v_0) = \mathbf{p}$).

Proof. Let γ be a smooth curve in S, say

$$\gamma(t) = \sigma(u(t), v(t)).$$

Denoting d/dt be a dot, we have, by the chain rule,

$$\dot{\gamma} = \sigma_u \dot{u} + \sigma_v \dot{v}.$$

Thus, $\dot{\gamma}$ is a linear combination of σ_u and σ_v.

Conversely, any vector in the vector subspace of $\mathbb{R}^3$, spanned by σ_u and σ_v is of the form $\lambda\sigma_u + \mu\sigma_v$ for some scalars λ and μ. Define

$$\gamma(t) = \sigma(u_0 + \lambda t, v_0 + \mu t).$$

Then, γ is a smooth curve in S and at $t = 0$, that is, at the point $\mathbf{p} \in S$, we have

$$\dot{\gamma} = \lambda\sigma_u + \mu\sigma_v$$

This shows that every vector in the span of σ_u and σ_v is the tangent vector at $\mathbf{p}$ of some curve in S. □

Since σ is assumed to be regular, σ_u and σ_v v are linearly independent so the tangent space is two-dimensional, and will be called the tangent plane from now on.

As a first application of the tangent plane to a smooth surface, we shall explain what is meant by the derivative of a smooth map between surfaces. Suppose then that $f : S \to \tilde{S}$ is such a map. The derivative of f at a point $\mathbf{p} \in S$ should measure how the point $f(\mathbf{p}) \in \tilde{S}$ changes when $\mathbf{p}$ moves to a nearby point, say $\mathbf{q}$ of S. If the points $\mathbf{p}$ and $\mathbf{q}$ are very close together, the straight line through them should be nearly tangent to S at $\mathbf{p}$. So we should expect that the derivative of f at $\mathbf{p}$ associates to any tangent vector to S at $\mathbf{p}$ a tangent vector to $\tilde{S}$ at $f(\mathbf{p})$, in other words, the derivative of f at $\mathbf{p}$ should be a map $D_{\mathbf{p}}f : T_{\mathbf{p}}S \to T_{f(\mathbf{p})}\tilde{S}$.

To give a precise definition of $D_{\mathbf{p}}f$ we let $\mathbf{w} \in T_{\mathbf{p}}S$ be a tangent vector to S at $\mathbf{p}$. By definition, $\mathbf{w}$ is the tangent vector at $\mathbf{p}$ of a curve γ in S passing through $\mathbf{p}$, say $\mathbf{w} = \dot{\gamma}(t_0)$. Then, $\tilde{\gamma} = f \circ \gamma$ is a curve in $\tilde{S}$ passing through $f(\mathbf{p})$ when $t = t_0$, so $\tilde{\mathbf{w}} = \dot{\tilde{\gamma}}(t_0) \in T_{f(\mathbf{p})}\tilde{S}$.

Definition 2.1.3. With the above notation, the derivative $D_{\mathbf{p}}f$ of f at the point $\mathbf{p} \in S$ is the map $D_{\mathbf{p}}f : T_{\mathbf{p}}S \to T_{f(\mathbf{p})}\tilde{S}$ such that $D_{\mathbf{p}}f(\mathbf{w}) = \tilde{\mathbf{w}}$ for any tangent vector $\mathbf{w} \in T_{\mathbf{p}}S$.

The first thing we must do now is to show that this definition makes sense, that is, that $D_{\mathbf{p}}f(\mathbf{w})$ depends only on $f, \mathbf{p}$, and $\mathbf{w}$: there are (infinitely) many curves γ with the correct tangent vector $\mathbf{w}$ at $\mathbf{p}$ and a priori $D_{\mathbf{p}}f(\mathbf{w})$ could depend on which curve is chosen.

Let $\sigma : U \to \mathbb{R}^3$ be a surface patch of S containing $\mathbf{p}$, say $\mathbf{p} = \sigma(u_0, v_0)$, and let α, β be the smooth functions on U such that

$$f(\sigma(u, v)) = \tilde{\sigma}(\alpha(u, v), \beta(u, v)).$$

Let $\mathbf{w} = \lambda\sigma_u + \mu\sigma_v$ be the tangent vector at $\mathbf{p}$ of a curve $\gamma(t) = \sigma(u(t), v(t))$, where u and v are smooth functions such that $\dot{u}(t_0) = \lambda, \dot{v}(t_0) = \mu$. Since the corresponding curve on $\tilde{S}$ is $\tilde{\gamma}(t) = \tilde{\sigma}(\tilde{u}(t), \tilde{v}(t))$, where $\tilde{u}(t) = \alpha(u(t), v(t))$ and $\tilde{v}(t) = \beta(u(t), v(t))$, we have

$$\begin{aligned} D_{\mathbf{p}}f(\mathbf{w}) &= \dot{\tilde{u}}\tilde{\sigma}_{\tilde{u}} + \dot{\tilde{v}}\tilde{\sigma}_{\tilde{v}} \\ &= (\dot{u}\alpha_u + \dot{v}\alpha_v)\tilde{\sigma}_{\tilde{u}} + (\dot{u}\beta_u + \dot{v}\beta_v)\tilde{\sigma}_{\tilde{v}}, \end{aligned}$$

the derivative of u and v being evaluated at t_0. Thus,

$$D_{\mathbf{p}}f(\mathbf{w}) = (\lambda\alpha_u + \mu\alpha_v)\tilde{\sigma}_{\tilde{u}} + \lambda\beta_u + \mu\beta_v)\tilde{\sigma}_{\tilde{v}}. \tag{2.1.2}$$

The RHS depends only on $\mathbf{p}$, f, λ and μ, that is, on $\mathbf{p}$, f and $\mathbf{w}$ as desired. The equation (2.1.2) also establishes the following theorem

Theorem 2.1.4. If $f : S \to \tilde{S}$ is a smooth map between surfaces and $\mathbf{p} \in S$, the derivative $D_{\mathbf{p}}f : T_{\mathbf{p}}S \to T_{f(\mathbf{p})}\tilde{S}$ is a linear map.

Theorem 2.1.5.

1. If S is a surface and $\mathbf{p} \in S$, the derivative at $\mathbf{p}$ of the identity map $S \to S$ is the identity map $T_{\mathbf{p}}S \to T_{\mathbf{p}}S$.

2. If S_1, S_2, S_3 are surfaces and $f_1 : S_1 \to S_2$ and $f_2 : S_2 \to S_3$ are smooth maps, then for all $\mathbf{p} \in S_1$,

$$D_{\mathbf{p}}(f_2 \circ f_1) = D_{f_1(\mathbf{p})}f_2 \circ D_{\mathbf{p}}f_1.$$

3. If $f : S_1 \to S_2$ is a diffeomorphism, then for all $\mathbf{p} \in S_1$, the linear map $D_{\mathbf{p}}f : T_{\mathbf{p}}S_1 \to T_{f(\mathbf{p})}S_2$ is invertible.

Proof. Part 1 is obvious. For 2, let $\mathbf{w} \in T_{\mathbf{p}}S_1$ be the tangent vector at $\mathbf{p}$ of a curve γ_1 on S_1. Then $\gamma_2 = f_1 \circ \gamma_1$ is a curve on S_2 with tangent vector $D_{\mathbf{p}}f_1(\mathbf{w})$ at $f_1(\mathbf{p})$, so $\gamma_3 = f_2 \circ \gamma_2 = (f_2 \circ f_1) \circ \gamma_1$ is a curve on S_3 with tangent vector $D_{f_1(\mathbf{p})}f_2(D_{\mathbf{p}}f_1(\mathbf{w}))$ at $f_2(f_1(\mathbf{p}))$. But the tangent vector of γ_3 at $\mathbf{p}$ is also $D_{\mathbf{p}}(f_2 \circ f_1)(\mathbf{w})$.

Finally, for 3, let $g : S_3 \to S_1$ 1 be the inverse map of f, so that $g \circ f$ and $f \circ g$ are the identity maps $S_1 \to S_1$ and $S_2 \to S_2$, respectively. Parts 1 and 2 show that $D_{f(\mathbf{p})}g$ g is the inverse of the linear map $D_{\mathbf{p}}f$. □

2.1.1 Normals and Orientability

Since the tangent plane $T_{\mathbf{p}}S$ of a surface S at point $\mathbf{p} \in S$ passes through the origin of $\mathbb{R}^3$, it is completely determined by giving a unit vector perpendicular to it, called a unit normal to S at $\mathbf{p}$. There are, of course, two such vectors, but by a previous theorem, choosing a surface patch $\sigma : U \to \mathbb{R}^3$ containing $\mathbf{p}$ leads to a definite choice, namely

$$N_\sigma = \frac{\sigma_u \times \sigma_v}{||\sigma_u \times \sigma_v||}$$

(with the derivatives evaluated at the point of U corresponding to p), for this is clearly a unit vector perpendicular to every linear combination of σ_u and σ_v. This is called the standard unit normal of the surface patch σ at $\mathbf{p}$. Unlike the tangent plane, however, N_σ is not quite independent of the choice of patch σ containing $\mathbf{p}$. In fact, if $\tilde{\sigma} : \tilde{U} \to \mathbb{R}^3$ is another surface patch in the atlas of S containing $\mathbf{p}$ that

$$\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}} = \det(J(\Phi))\sigma_u \times \sigma_v$$

where $J(\Phi)$ is the Jacobian matrix of the transition map Φ from σ to $\tilde{\sigma}$. So the the standard unit normal of $\tilde{\sigma}$ is

$$\begin{aligned} N_{\tilde{\sigma}} &= \frac{\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}}}{||\tilde{\sigma}_{\tilde{u}} \times \tilde{\sigma}_{\tilde{v}}||} \\ &= \pm\frac{\sigma_u \times \sigma_v}{||\sigma_u \times \sigma_v||} \\ &= \pm N_\sigma, \end{aligned}$$

where the sign is that of the determinant of $J(\Phi)$. This leads to the following definition.

Definition 2.1.6. A surface S is said to be orientable if there exists an atlas $\mathcal{A}$ for S with the property that, if Φ is the transition map between any two surface patches in $\mathcal{A}$, then $\det(J(\Phi)) > 0$, where Φ is defined.

The preceding discussion gives the following theorem.

Theorem 2.1.7. Let S be an orientable surface equipped with an atlas $\mathcal{A}$ as in the above definition. Then, there is a smooth choice of unit normal at any point of S; take the standard unit normal of any surface patch in $\mathcal{A}$.

An oriented surface is a surface S together with a smooth choice of unit normal N at each point, i.e., a smooth map $N : S \to \mathbb{R}^3$ (meaning that each of the three components of N is a smooth function $S \to \mathbb{R}$) such that, for all $\mathbf{p} \in S$, $N(\mathbf{p})$ is a unit vector perpendicular to $T_{\mathbf{p}}S$. Any oriented surface is orientable. To see this, start with the maximal atlas of S and retain a patch $\sigma(u, v)$ if $\sigma_u \times \sigma_v$ is a positive multiple of N at all points in the image of σ, otherwise discard it. The patches that remain form an atlas $\mathcal{A}$ satisfying the condition in the previous definition.

Example 2.1.8. The Möbius band is the surface obtained by rotating a straight line segment l around its midpoint $\mathbf{p}$ at the same time as $\mathbf{p}$ moves around a circle $\mathcal{C}$, in such a way as $\mathbf{p}$ moves once around $\mathcal{C}$, l makes a half-turn about $\mathbf{p}$. If we take $\mathcal{C}$ to be the circle $x^2 + y^2 = 1$ in the xy-plane, and l to be a segment of length 1 that is initially parallel to the z-axis with its midpoint $\mathbf{p}$ at $(1, 0, 0)$, and then after $\mathbf{p}$ has rotated by an angle θ around the z-axis, l should have rotated by $\theta/2$ around $\mathbf{p}$ in the plane containing $\mathbf{p}$ and the z-axis. The point of l initially at $(1, 0, t)$ is then at the point

$$\sigma(t, \theta) = \left(\left(1 - t \sin\frac{\theta}{2}\right)\cos\theta, \left(1 - t\sin\frac{\theta}{2}\right)\sin\theta, t\cos\frac{\theta}{2}\right).$$

We take the domain of definition of σ to be

$$U = \{(t, \theta) \in \mathbb{R}^2 | -1/2 < t < 1/2,\ 0 < \theta < 2\pi\}$$

We can define a second patch $\tilde{\sigma}$ by the same formula as σ but with domain of definition

$$\tilde{U} = \{(t, \theta) \in \mathbb{R}^2 | -1/2 < t < 1/2,\ -\pi < \theta < \pi\}$$

It can be checked that these two patches form an atlas for the Möbius band consisting of regular surface patches, making the Möbius band into a smooth surface S.

We compute the standard unit normal N_σ at points on the median circle (where $t = 0$). At such points, we have

$$\sigma_t = \left(-\sin\frac{\theta}{2}\cos\theta, -\sin\frac{\theta}{2}\sin\theta, \cos\frac{\theta}{2}\right),\ \sigma_\theta = (\sin\theta, \cos\theta, 0),$$

So,

$$\sigma_t \times \sigma_\theta = \left(-\cos\theta\cos\frac{\theta}{2}, -\sin\theta\cos\frac{\theta}{2}, -\sin\frac{\theta}{2}\right).$$

This is a unit vector, so it is equal to N_σ.

If the Möbius band was orientable, there would be a well-defined unit normal N defined at every point of S and varying smoothly over S. At a point $\sigma(0, \theta)$ on the median circle, we would have

$$N = \lambda(\theta) N_\sigma,$$

where $\lambda : (0, 2\pi) \to \mathbb{R}$ is smooth and $\lambda(\theta) = \pm 1$ for all θ. It follows that either $\lambda(\theta) = +1$ for all $\theta \in (0, 2\pi)$, or $\lambda(\theta) = -1$ for all $\theta \in (0, 2\pi)$. Replacing N by $-N$ if necessary, we can assume that $\lambda = 1$. At the point $\sigma(0, 0) = \sigma(0, 2\pi)$, we would have (since N is smooth)

$$N = \lim_{\theta \downarrow 0} N_\sigma = (-1, 0, 0) \text{ and also } N = \lim_{\theta \uparrow 2\pi} N_\sigma = (1, 0, 0).$$

This contradiction shows that the Möbius band is not orientable.

If a surface S is oriented, it is possible to give a sign to the angle between two tangent vectors at a point of S.

Let $\mathbf{p} \in S$ and let N be the chosen unit normal at $\mathbf{p}$. A rotation in the tangent plane $T_\mathbf{p}S$ is said to be in the positive sense, or anticlockwise, if rotation in this sense of a right-handed screw held perpendicular to $T_\mathbf{p}S$ would cause it to advance in the direction of N. Put another way, the choice of N enables us to distinguish the two 'sides' of $T_\mathbf{p}S$: the 'positive' side is the half-space into which N points. Then, if we view $T_\mathbf{p}S$ from a point on the positive side, a positive rotation in $T_\mathbf{p}S$ would be seen as anticlockwise in the usual sense.

If $\mathbf{v}$ and $\mathbf{w}$ are non-zero vectors in $T_\mathbf{p}S$, the oriented angle (which we shall sometimes just call the angle) between $\mathbf{v}$ and $\mathbf{w}$ is the angle through which $\mathbf{v}$ must be rotated in the positive sense in order for the resulting vector to be a positive scalar multiple of $\mathbf{w}$. We shall denote this angle by $\mathbf{v\hat{w}}$. Note that

$$\mathbf{w\hat{v}} = -\mathbf{v\hat{w}},$$

and that the sign of $\mathbf{v\hat{w}}$ will change if we change the choice of unit normal to $T_\mathbf{p}S$. Note also that $\mathbf{v\hat{w}}$ is determined only up to the addition of an integer multiple of 2π.

2.1.2 Examples of Surfaces

Level Surfaces

Surfaces are often given to us as level surfaces

$$\{(x, y, z) \in \mathbb{R}^3 | f(x, y, z) = 0\}$$

where f is a smooth function. The following result gives general conditions under which a level surface is a smooth surface.

Theorem 2.1.9. Let S be a subset of $\mathbb{R}^3$ with the following property: for each point $\mathbf{p} \in S$, there is an open subset W of $\mathbb{R}^3$ containing $\mathbf{p}$ and a smooth function $f : W \to \mathbb{R}$ such that

1. $S \cap W = \{(x, y, z) \in W | f(x, y, z) = 0\}$
2. $\nabla f = (f_x, f_y, f_z)$ of f does not vanish at $\mathbf{p}$.

Exercises

1. Find the equation of the tangent plane of the surface $\sigma(r, \theta) = (r \cosh \theta, r \sinh \theta, r^)$ at $(1, 0, 1)$.
2. Show that
$$\sigma(u, v) = (\text{sech}\, u \cos v, \text{sech}\, u \sin v, \tanh u)$$
is a regular surface patch for the unit sphere. Show that meredians and parallels on the sphere correspond under σ to perpendicular straightlines in the plane.
3. Consider the surface obtained by rotating the curve $x = \cosh z$ in the xz-plane around the z-axis. Describe an atlas for this surface.
4. Show that a Mobius band is not an orientable surface.
5. If $\sigma(u, v)$ is a surface patch, show that the set of linear combinations of σ_u and σ_v is unchanged when σ is reparametrized.

Unit 3

Course Structure

- The first Fundamental form
- Length of curves on surfaces
- Isometries of surfaces
- Conformal mapping of surfaces

Introduction

In our Euclidean plane we use the quadratic form $ds^2 = dx^2 + dy^2$ for the purpose of measurement. This formula is valid in $\mathbb{R}^2$ with rectangular cartesian coordinate system. The expression $ds = \sqrt{dx^2 + dy^2}$ is called Euclidean metric. Observe that $ds^2 = dx^2 + dy^2$ may be expressed as

$$ds^2 = \begin{bmatrix} dx & dy \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix}$$

The matrix in the above equation is called the Euclidean matrix.

But, if our coordinate system is not rectangular, then the metric is not Euclidean. On a surface we can't define rectangular cartesian coordinate system. The coordinate system on surface is a type of curvilinear coordinate system.

In this chapter we shall deduce a quadratic form on a surface which is known as the First Fundamental form and is used to measure lengths and angles on surfaces.

3.0.1 Lengths of curves on surfaces

If $\gamma(t) = \sigma(u(t), v(t))$ is a curve in a surface patch σ, its arc length starting at a point $\gamma(t_0)$ is given by

$$s = \int_{t_0}^{t} ||\dot{\gamma}(u)|| du$$

By chain rule,

$$\dot{\gamma}^2 = \sigma_u \dot{u} + \sigma_v \dot{v}$$

Hence

$$\begin{aligned} ||\dot{\gamma}||^2 &= (\sigma_u \dot{u} + \sigma_v \dot{v})(\sigma_u \dot{u} + \sigma_v \dot{v}) \\ &= (\sigma_u \sigma_u)\dot{u}^2 + 2(\sigma_u \sigma_v)\dot{u}\dot{v} + (\sigma_v \sigma_v)\dot{v}^2 \\ &= E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2 \end{aligned}$$

where $E = ||\sigma_u||^2$, $F = \sigma_u \sigma_v$, $G = ||\sigma_v||^2$. So

$$s = \int_{t_0}^{t} (E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)^{1/2} dt$$

Comparing the above equation with

$$s = \int_{t_0}^{t} (\sqrt{ds})^2$$

We have,

$$ds^2 = Edu^2 + 2Fdudv + Gdv^2$$

This is called the First Fundamental Quadratic form on surface.

3.0.2 First Fundamental form on Sphere

For the sphere in latitude longitude coordinates

$$\sigma(\theta, \phi) = (\cos\theta \cos\phi, \cos\theta \sin\phi, \sin\theta),$$

We have

$$\begin{aligned} \sigma_\theta &= (-\sin\theta \cos\phi, -\sin\theta \sin\phi, \cos\theta) \\ \sigma_\phi &= (-\cos\theta \sin\phi, \cos\theta \cos\phi, 0) \end{aligned}$$

Hence, $E = ||\sigma_\theta||^2 = 1$, $F = \sigma_\theta \sigma_\phi = 0$, $G = ||\sigma_\phi||^2 = \cos^2\theta$. So the First Fundamental form is $d\theta^2 + \cos^2\theta d\phi^2$.

Exercise 3.0.1.

1. Find the First Fundamental form on cylinder.
2. Calculate the First Fundamental form on $\sigma(u, v) = (u - v, u + v, u^2 + v^2)$.
3. Show that applying rigid motion to a surface does not change its First Fundamental form.

Definition 3.0.2. If S_1 and S_2 are surfaces, a diffeomorphism $f : S_1 \to S_2$ is called an isometry if it takes curves in S_1 to curves of same length in S_2. If an isometry $f : S_1 \to S_2$ exists, then we say that S_1 and S_2 are isomorphic.

Theorem 3.0.3. A diffeomorphism $f : S_1 \to S_2$ is an isometry if and only if, for any surface patch σ_1 of S_1, the patches σ_1 and $f \circ \sigma_1$ of S_1 and S_2 respectively, have the same first fundamental form.

Proof. Suppose the length of any curve can be computed as the sum of the lengths of curves each lying in a single surface patch, we can assume that S_1 and S_2 are covered by single surface patches. Moreover, since f is a diffeomorphism, we can assume that these patches are of the form $\sigma_1 : U \to \mathbb{R}^3$ (for S_1) and $f \circ \sigma_1 = \sigma_2$ (for S_2). We have to show that f is an isometry if and only if σ_1 and σ_2 have the same first fundamental form.

Suppose first that σ_1 and σ_2 have same first fundamental form. If $t \to (u(t), v(t))$ is any curve in U, say $\gamma_1(t) = \sigma_1(u(t), v(t))$ and $\gamma_2(t) = \sigma_2(u(t), v(t))$ are the corresponding curves in S_1 and S_2, then f takes γ_1 to γ_2, since

$$\begin{aligned} f_1(\gamma_1(t)) &= f(\sigma_1(u(t)), \sigma_1(v(t))) \\ &= \sigma_2(u(t), v(t)) \\ &= \gamma_2(t) \end{aligned}$$

It is clear that γ_1 and γ_2 have the same length, since both lengths are found by integrating the expression $(E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2)^{1/2}$, where $Edu^2 + 2Fdudv + Gdv^2$ is the common first fundamental form of S_1 and S_2.

Conversely suppose that f is an isometry. If $f \mapsto (u(t), v(t))$ is any curve in U defined for $t \to (\alpha, \beta)$, say, the curve $\gamma_1(t) = \sigma_1(u(t), v(t))$ and $\gamma_2(t) = (\sigma_2(u(t)), \sigma_2(v(t)))$ have the same length. Hence

$$\int_{t_0}^{t_1} (E_1\dot{u}^2 + 2F_1\dot{u}\dot{v} + G_1\dot{v}^2)^{1/2} dt = \int_0^{t_0} (E_2\dot{u}^2 + 2F_2\dot{u}\dot{v} + G_2\dot{v}^2)^{1/2} dt$$

for all $t_0, t_1 \in (\alpha, \beta)$, where E_1, F_1 and G_1 are the coefficients of the first fundamental form of σ_1 and E_2, F_2 and G_2 are those of σ_2. This implies that the two integrands are the same and hence that

$$E_1\dot{u}^2 + 2F_1\dot{u}\dot{v} + G_1\dot{v}^2 = E_2\dot{u}^2 + 2F_2\dot{u}\dot{v} + G_2\dot{v}^2$$

Fix $t_0 \in (\alpha, \beta)$ and let $u_0 = u(t_0), v_0 = v(t_0)$. We now apply the above equation for the following three choices of the curve $t \to (u(t), v(t))$ in U:

(i) $u = u_0 + t - t_0;\ v = v_0 \implies E_1 = E_2$

(ii) $v = v_0 + t - t_0;\ u = u_0 \implies G_1 = G_2$

(iii) $u = u_0 + t - t_0;\ v = v_0 + t - t_0$. This gives $E_1 + 2F_1 + G_1 = E_2 + 2F_2 + G_2$ and hence $F_1 = F_2$.

□

3.0.3 Conformal Mappings

A mapping from a surface σ_1 to another surface σ_2 is called conformal if angle between two intersecting curves is preserved under the mapping.

Angle between two curves on a surface

Suppose that two curves γ and $\tilde{\gamma}$ on a surface S intersect at a point **p** that lies in a surface patch σ on S. Then $\gamma(t) = \sigma(u(t), v(t))$ and $\tilde{\gamma}(t) = \sigma(\tilde{u}(t), \tilde{v}(t))$ for some smooth functions $u, v, \tilde{u}, \tilde{v}$ and for some parameter values t_0 and $\tilde{t}_0$, we have $\sigma(u(t_0), v(t_0)) = P = \sigma(\tilde{u}(t_0), \tilde{v}(t_0))$.

The angle θ between the curves γ and $\tilde{\gamma}$ is given by

$$\cos\theta = \frac{\dot{\gamma}.\dot{\tilde{\gamma}}}{||\dot{\gamma}||.||\dot{\tilde{\gamma}}||}$$

By chain rule

$$\begin{aligned} \dot{\gamma} &= \sigma_u\dot{u} + \sigma_v\dot{v} \\ \dot{\tilde{\gamma}} &= \sigma_u\dot{\tilde{u}} + \sigma_v\dot{\tilde{v}} \end{aligned}$$

So

$$\begin{aligned}\dot{\gamma}.\dot{\tilde{\gamma}} &= (\sigma_u\dot{u}+\sigma_v\dot{v})(\sigma_u\dot{\tilde{u}}+\sigma_v\dot{\tilde{v}})\\ &= (\sigma_u\sigma_u)\dot{u}\dot{\tilde{u}}+(\sigma_u\sigma_v)(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+(\sigma_v\sigma_v)\dot{v}\dot{\tilde{v}}\\ &= E\dot{u}\dot{\tilde{u}}+F(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+G\dot{v}\dot{\tilde{v}}\end{aligned}$$

Similarly, $||\dot{\gamma}|| = \sqrt{\dot{\gamma}.\dot{\gamma}}$ and $||\dot{\tilde{\gamma}}|| = \sqrt{\dot{\tilde{\gamma}}.\dot{\tilde{\gamma}}}$ can be evaluated and finally

$$\cos\theta = \frac{E\dot{u}\dot{\tilde{u}}+F(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+G\dot{v}\dot{\tilde{v}}}{(E\dot{u}^2+2F\dot{u}\dot{v}+G\dot{v}^2)^{1/2}(E\dot{\tilde{u}}^2+2F\dot{\tilde{u}}\dot{\tilde{v}}+G\dot{\tilde{v}}^2)^{1/2}}$$

Theorem 3.0.4. The parameter curves on a surface patch $\sigma(u,v)$ can be parametrized by $\gamma(t)=\sigma(a,t),\ \tilde{\gamma}=\sigma(t,b)$ receptively, where a is the constant value of u and b the constant value of v in the two cases. Thus $u(t)=a,\ v(t)=t,\ \tilde{u}(t)=t,\ \tilde{v}(t)=b,\ \dot{u}=0,\ \dot{v}=1,\ \dot{\tilde{u}}=1,\ \dot{\tilde{v}}=0$. These parameter curves intersect at the point $\sigma(a,b)$ of the surface and their angle of intersection θ is given by

$$\cos\theta = \frac{F}{\sqrt{EG}}$$

In particular, the parameter curves are orthogonal if and only if $F=0$.

Definition 3.0.5. If S_1 and S_2 are surfaces, a diffeomorphism $f: S_1 \to S_2$ is said to be conformal if, whenever f takes two intersecting curves γ_1 and $\tilde{\gamma}_1$ on S_1 to curves γ_2 and $\tilde{\gamma}_2$ on S_2, the angle of intersection of γ_1 and $\tilde{\gamma}_1$ is equal to the angle of intersection of γ_2 and $\tilde{\gamma}_2$.

Theorem 3.0.6. A diffeomorphism $f: S_1 \to S_2$ is conformal if and only if, for any surface patch σ_1 on S_1, the first fundamental forms of σ_1 and $f\circ\sigma_1$ are proportional.

Proof. Assume that S_1 and S_2 are covered by the single surface patches $\sigma_1: U\to\mathbb{R}^3$ and $\sigma_2 = f\circ\sigma_1$, respectively. Suppose that their first fundamental forms $E_1du^2+2F_1dudv+G_1dv^2$ and $E_2du^2+2F_2dudv+G_2dv^2$ are proportional, say

$$E_1du^2+2F_1dudv+G_1dv^2 = \lambda(E_2du^2+2F_2dudv+G_2dv^2)$$

for some smooth function $\lambda(u,v)$, where (u,v) are coordinates on U. Note that $\lambda>0$ everywhere, since (for example) E_1 and E_2 are both greater than 0. If $\gamma(t)=\sigma_1(u(t),v(t))$ and $\tilde{\gamma}(t)=\sigma_1(\tilde{u}(t),\tilde{v}(t))$ are curves in S_1, then f takes γ and $\tilde{\gamma}$ to the curves $\sigma_2(u(t),v(t))$ and $\sigma_2(\tilde{u}(t),\tilde{v}(t))$ in S_2, respectively. Now,

$$\begin{aligned}\cos\theta &= \frac{E_2\dot{u}\dot{\tilde{u}}+F_2(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+G_2\dot{v}\dot{\tilde{v}}}{(E_2\dot{u}^2+2F_2\dot{u}\dot{v}+G_2\dot{v}^2)^{1/2}(E_2\dot{\tilde{u}}^2+2F_2\dot{\tilde{u}}\dot{\tilde{v}}+G_2\dot{\tilde{v}}^2)^{1/2}}\\ &= \frac{\lambda E_1\dot{u}\dot{\tilde{u}}+\lambda F_1(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+\lambda G_1\dot{v}\dot{\tilde{v}}}{(\lambda E_1\dot{u}^2+2\lambda F_1\dot{u}\dot{v}+\lambda G_1\dot{v}^2)^{1/2}(\lambda E_1\dot{\tilde{u}}^2+2\lambda F_1\dot{\tilde{u}}\dot{\tilde{v}}+\lambda G_1\dot{\tilde{v}}^2)^{1/2}}\\ &= \frac{E_1\dot{u}\dot{\tilde{u}}+F_1(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+G_1\dot{v}\dot{\tilde{v}}}{(E_1\dot{u}^2+2F_1\dot{u}\dot{v}+G_1\dot{v}^2)^{1/2}(E_1\dot{\tilde{u}}^2+2F_1\dot{\tilde{u}}\dot{\tilde{v}}+G_1\dot{\tilde{v}}^2)^{1/2}}\end{aligned}$$

Hence f is conformal.

For the converse, we must show that if

$$\begin{aligned}&\frac{E_1\dot{u}\dot{\tilde{u}}+F_1(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+G_1\dot{v}\dot{\tilde{v}}}{(E_1\dot{u}^2+2F_1\dot{u}\dot{v}+G_1\dot{v}^2)^{1/2}(E_1\dot{\tilde{u}}^2+2F_1\dot{\tilde{u}}\dot{\tilde{v}}+G_1\dot{\tilde{v}}^2)^{1/2}} =\\ &\frac{E_2\dot{u}\dot{\tilde{u}}+F_2(\dot{u}\dot{\tilde{v}}+\dot{\tilde{u}}\dot{v})+G_2\dot{v}\dot{\tilde{v}}}{(E_2\dot{u}^2+2F_2\dot{u}\dot{v}+G_2\dot{v}^2)^{1/2}(E_2\dot{\tilde{u}}^2+2F_2\dot{\tilde{u}}\dot{\tilde{v}}+G_2\dot{\tilde{v}}^2)^{1/2}}\end{aligned} \tag{3.0.1}$$

for all pairs of intersecting curves

$$\begin{aligned} \gamma(t) &= \sigma_1(u(t), v(t)) \\ \text{and } \tilde{\gamma}(t) &= \sigma_1(\tilde{u}(t), \tilde{v}(t)) \end{aligned}$$

in S_1, then the first fundamental form of σ_1 and σ_2 are proportional. Fix $(a, b) \in U$ and consider the curves

$$\begin{aligned} \gamma(t) &= \sigma_1(a + t, b) \\ \tilde{\gamma}(t) &= \sigma_1(a + t\cos\phi, b + t\sin\phi) \end{aligned}$$

where ϕ is a constant, for which

$$\dot{u} = 1,\ \dot{v} = 0,\ \dot{\tilde{u}} = \cos\phi,\ \dot{\tilde{v}} = \sin\phi$$

Using these in (3.0.1), we get

$$\frac{E_1\cos\phi + F_1\sin\phi}{\sqrt{E_1(E_1\cos^2\phi + 2F_1\sin\phi\cos\phi + G_1\sin^2\phi)}} = \frac{E_2\cos\phi + F_2\sin\phi}{\sqrt{E_2(E_2\cos^2\phi + 2F_2\sin\phi\cos\phi + G_2\sin^2\phi)}} \quad (3.0.2)$$

Squaring both sides of equation (3.0.2) and writing

$$(E_1\cos\phi + F_1\sin\phi)^2 = E_1(E_1\cos^2\phi + 2F_1\sin\phi\cos\phi + G_1\sin^2\phi) - (E_1G_1 - F_1^2)\sin^2\phi,$$

we get

$$(E_1G_1 - F_1^2)E_2((E_2\cos^2\phi + 2F_2\sin\phi\cos\phi + G_2\sin^2\phi) = (E_2G_2 - F_2^2)E_1(E_1\cos^2\phi + 2F_1\sin\phi\cos\phi + G_1\sin^2\phi),$$

or setting

$$\lambda = \frac{(E_2G_2 - F_2^2)E_1}{(E_1G_1 - F_1^2)E_2},$$

we get

$$(E_2 - \lambda E_1)\cos^2\phi + (2F_2 - \lambda F_1)\sin\phi\cos\phi + (G_2 - \lambda G_1)\sin^2\phi = 0$$

Taking $\phi = 0$ and $\phi = \pi/2$ gives $E_2 = \lambda E_1,\ G_2 = \lambda G_1$. Then by the last equation, $F_2 = \lambda F_1$. □

Exercises

1. Calculate the first fundamental form of the surface
$$\sigma(u, v) = (\sinh u \sinh v, \sinh u \cosh v, \sinh u).$$
2. Show that applying a rigid motion to a surface does not change its first fundamental form.
3. Show that every isometry is a conformal map. Give an example of a conformal map that is not an isometry.
4. Show that every isometry is a conformal map. Give an example of a conformal map that is not an isometry.
5. Show that the map
$$\sigma(u, v) = (\text{sech}\, u\cos v, \text{sech}\, u\sin v, \tanh u)$$
is conformal.

Summary

In this unit, we have determined the expression of first fundamental form of surfaces. We have calculated first fundamental form of sphere. We have also discussed about conformal maps.

Unit 4

Course Structure

- Introduction
- Curvature of curves on surfaces
- Second Fundamental form

Introduction

To introduce the notion of curvature on surfaces, we start by finding a new interpretation of the curvature of a plane curve. Suppose that γ is a unit speed curve in $\mathbb{R}^2$. As the parameter t of γ changes to $t + \Delta t$, the curve moves away from its tangent line at $\gamma(t)$ by a distance $(\gamma(t + \Delta t) - \gamma(t)).\eta$, where η is the principal normal to γ at $\gamma(t)$. By Taylor's theorem,

$$\gamma(t + \Delta t) = \gamma(t) + \dot{\gamma}(t)\Delta t + \frac{1}{2} + \ddot{\gamma}(t)(\Delta t)^2 + \text{remainder},$$

where remainder/$(\Delta t)^2$ tends to 0 as $\Delta t \to 0$. Now, η is perpendicular to the unit tangent vector $t = \dot{\gamma}$ and $\ddot{\gamma} = \dot{t} = \kappa\eta$, where κ is the curvature of γ. Hence, $\ddot{\gamma}.\eta = \kappa$ and the derivative of γ forms its tangent line is

$$(\dot{\gamma}(t)\Delta t + \frac{1}{2} + \ddot{\gamma}(t)(\Delta t)^2 + \cdots).\eta = \frac{1}{2}\kappa(\Delta t)^2 + \text{remainder} \tag{4.0.1}$$

Now let σ be a surface patch in $\mathbb{R}^3$ with standard unit normal N (surface normal). As the parameters (u, v) of σ changes to $(u + \Delta u, v + \Delta v)$, the surface moves away from its tangent plane at $\sigma(u, v)$ by a distance $(\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v)).N$. By the two variable form of Taylor's theorem,

$$\sigma(u + \Delta u, v + \Delta v) - \sigma(u, v) = \sigma_u\Delta u + \sigma_v\Delta v + \frac{1}{2}(\sigma_{uu}(\Delta u)^2 + 2\sigma_{uv}\Delta u\Delta v + \sigma_{vv}(\Delta v)^2) + \text{remainder},$$

where (remainder)/$[(\Delta u)^2 + (\Delta v)^2]$ tends to zero as $(\Delta u)^2 + (\Delta v)^2 \to 0$. Now, σ_u and σ_v are tangents to the surface, hence perpendicular to N, so the derivation of σ from its tangent plane is

$$\frac{1}{2}(\mathcal{L}(\Delta u)^2 + 2\mathcal{M}\Delta u\Delta v + \mathcal{N}(\Delta v)^2) + \text{remainder} \tag{4.0.2}$$

where, $\mathcal{L} = \sigma_{uu}.N$, $\mathcal{M} = \sigma_{uv}.N$, $\mathcal{N} = \sigma_{vv}.N$. Comparing equation (4.0.2) with equation (4.0.1), we see that the expression is the analogue for the surface of the curvature term $\kappa(\Delta t)^2$ in the case of a curve. One calls the expression

$$\mathcal{L}du^2 + 2\mathcal{M}dudv + \mathcal{N}dv^2 \tag{4.0.3}$$

is called the second fundamental form of σ. As in the case of the first fundamental form we regard the expression (4.0.3) simply as a convenient way of keeping track of the three functions $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{N}$. We shall soon see that a knowledge of these functions (together with that of first fundamental form) will enable us to compute the curvature of any curve on the surface.

Note 4.0.1. $\mathcal{L} = \sigma_{uu}.N$, $\mathcal{M} = \sigma_{uv}.N$, $\mathcal{N} = \sigma_{vv}.N$, where

$$N = \frac{\sigma_u \times \sigma_v}{||\sigma_u \times \sigma_v||}.$$

Exercise 4.0.2. Calculate the second fundamental form of unit sphere and right circular cylinder.

4.0.1 Normal and geodesic curvature of a curve on a surface

If $\gamma(t) = \sigma(u(t), v(t))$ is a unit speed curve in a surface patch γ, then $\dot{\gamma}$ is a unit vector, and is by definition, a tangent vector to σ. Hence $\dot{\gamma}$ is perpendicular to the standard unit normal N of σ, so $\dot{\gamma}$, N and $N \times \dot{\gamma}$ are mutually perpendicular unit vectors. Again, since γ is unit speed, $\ddot{\gamma}$ is perpendicular to $\dot{\gamma}$ and hence is a linear combination of N and $N \times \dot{\gamma}$. So,

$$\ddot{\gamma} = \kappa_n N + \kappa N \times \dot{\gamma} \tag{4.0.4}$$

The scalars κ_n and κ_g are called the normal curvature and the geodesic curvature of γ respectively. Since N and $N \times \dot{\gamma}$ are perpendicular unit vectors, equation (4.0.4) implies $\kappa_n = \ddot{\gamma}.N$, $\kappa_g = \ddot{\gamma}.(N \times \dot{\gamma})$ and,

$$||\ddot{\gamma}||^2 = \kappa_n^2 + \kappa_g^2 \tag{4.0.5}$$

Hence the curvature $\kappa = ||\ddot{\gamma}||$ of γ is given by

$$\kappa^2 = \kappa_n^2 + \kappa_g^2$$

Moreover, if η is principal normal of γ, so that $\ddot{\gamma} = \kappa\eta$, we have

$$\kappa_\eta = \kappa\eta.N = \kappa \cos\psi$$

where ψ is the angle between η and N. Then from equation (4.0.5)

$$\kappa_g = \pm\kappa \sin\psi$$

It is clear from their definition that κ_η and κ_g either stay the same or both change sign when σ is reparametrized.

When $\psi = 0$, the curve is called a normal section. Then $\kappa_\eta = \pm\kappa$, $\kappa_g = 0$.

Theorem 4.0.3. If $\gamma(t) = \sigma(u(t), v(t))$ is a unit speed curve on a surface patch σ, its normal curvature is given by

$$\kappa_\eta = \mathcal{L}\dot{u}^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}\dot{v}^2,$$

where $\mathcal{L}\dot{u}^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}\dot{v}^2$ is the second fundamental form of σ.

Proof. We have, with N denoting the standard unit normal of σ,

$$\begin{aligned}
\kappa_\eta &= N.\ddot{\gamma} \\
&= N.\frac{d}{dt}(\dot{\gamma}) \\
&= N.\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v}) \\
&= N.(\sigma_u\ddot{u} + \sigma_v\ddot{v} + (\sigma_{uu}\dot{u} + \sigma_{uv}\dot{v})\dot{u} + (\sigma_{uv}\dot{u} + \sigma_{vv}\dot{v})\dot{v}) \\
&= \mathcal{L}\dot{u}^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}\dot{v}^2
\end{aligned}$$

□

Theorem 4.0.4. (**Meusnier's Theorem**) Let P be a point on a surface S and let V be a unit vector to S at P. Let π_θ be the plane containing the line through P parallel to V and making an angle θ with the tangent plane to S at P. Suppose that π_θ intersects S in a curve with curvature κ_θ. Then $\kappa_\theta \sin\theta$ is independent of θ.

Proof. Assume that γ_θ is a unit speed parametrisation of the curve of intersection of π_θ and S. Then at P, $\dot{\gamma}_\theta = \pm v$, so $\ddot{\gamma}_\theta$ is perpendicular to v and is parallel to π_θ. Thus if ψ is the angle between η and N, then $\psi = \pi/2 - \theta$. Again, we know

$$\kappa_\eta = \kappa\cos\psi$$

Hence,

$$\kappa_\eta = \kappa\sin\theta = \kappa_\theta\sin\theta$$

(since, here $\kappa = \kappa_\theta$ by notation). □

4.0.2 Matrix Representation of Normal Curvature

To analyse κ further, it is useful to use matrix notation. If $Edu^2 + 2Fdudv + Gdv^2$ and $\mathcal{L}du^2 + 2\mathcal{M}dudv + \mathcal{N}dv^2$ are the first and second fundamental forms of a surface σ, we introduce the following symmetric 2×2 matrices

$$F_1 = \begin{bmatrix} E & F \\ F & G \end{bmatrix},\ F_{II} = \begin{bmatrix} \mathcal{L} & \mathcal{M} \\ \mathcal{M} & \mathcal{N} \end{bmatrix}$$

Let

$$t_1 = \xi_1\sigma_u + \eta_1\sigma_v,\ t_2 = \xi_2\sigma_u + \eta_2\sigma_v$$

be two tangent vectors at some point of σ. Then

$$\begin{aligned}
t_1.t_2 &= (\xi_1\sigma_u + \eta_1\sigma_v)(\xi_2\sigma_u + \eta_2\sigma_v) \\
&= E\xi_1\xi_2 + F(\xi_1\eta_2 + \xi_2\eta_1) + G\eta_1\eta_2 \\
&= \begin{bmatrix}\xi_1 & \eta_1\end{bmatrix}\begin{bmatrix} E & F \\ F & G \end{bmatrix}\begin{bmatrix}\xi_2 \\ \eta_2\end{bmatrix}
\end{aligned}$$

Thus writing

$$T_1 = \begin{bmatrix}\xi_1 \\ \eta_1\end{bmatrix},\ T_2 = \begin{bmatrix}\xi_2 \\ \eta_2\end{bmatrix} \tag{4.0.6}$$

we get $t_1.t_2 = T_1^t F_1 T_2$. On the other hand, tangent vector

$$\dot{\gamma} = \dot{u}\sigma_u + \dot{v}\sigma_v$$

and if

$$T = \begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix}$$

then by using theorem 4.0.3, we see that

$$\kappa_\eta = T^t F_{II} T$$

Exercise 4.0.5. Compute the normal curvature of the circle $\gamma(t) = (\cos t, \sin t, 1)$ on the elliptic paraboloid $\sigma(u, v) = (u, v, u^2 + v^2)$.

Exercises

1. Compute the second fundamental form of the elliptic paraboloid

$$\sigma(u, v) = (u, v, u^2 + v^2).$$

2. Compute the normal curvature of the circle $\gamma(t) = (\cos t, \sin t, 1)$ on the elliptic paraboloid $\sigma(u, v) = (u, v, u^2 + v^2)$.

3. Show that if a curve on a surface has zero normal and geodesic curvature everywhere, it is part of a straightline.

4. Show that a curve on a surface has zero normal and geodesic curvature everywhere, it is part of a straightline.

5. Show that the normal curvature of any curve on a sphere of radius r is $\pm\frac{1}{r}$.

Unit 5

Course Structure

- Introduction
- Geodesic curvatures
- Gaussian and Mean Curvatures

5.1 Introduction

In the present unit, we shall study curvature of surfaces using curvature of curves defined on surfaces.

In the previous unit, we have seen the normal curvature of a curve defined on a surface is given by

$$\kappa_\eta = \mathcal{L}(\dot{u})^2 + 2\mathcal{M}\dot{u}\dot{v} + \mathcal{N}(\dot{v})^2$$

So,

$$\begin{aligned}
\kappa_\eta &= \mathcal{L}\left(\frac{du}{ds}\right)^2 + 2\mathcal{M}\frac{du}{ds}\frac{dv}{ds} + \mathcal{N}\left(\frac{dv}{ds}\right)^2 \\
&= \frac{\mathcal{L}(du)^2 + 2\mathcal{M}dudv + \mathcal{N}(dv)^2}{ds^2} \\
&= \frac{\mathcal{L}(du)^2 + 2\mathcal{M}dudv + \mathcal{N}(dv)^2}{E(du)^2 + 2Fdudv + G(dv)^2}
\end{aligned}$$

Hence

$$\mathcal{L}(du)^2 + 2\mathcal{M}dudv + \mathcal{N}(dv)^2 = \kappa_\eta(E(du)^2 + 2Fdudv + G(dv)^2) \tag{5.1.1}$$

From equation (5.1.1), after some calculations, we can show that κ_η has maximum or minimum value if

$$|F_{II} - \kappa_\eta F_I| = 0$$

The maximum or minimum values of κ_η are called principal curvatures of the surface (not curve).

Definition 5.1.1. The principal curvatures of a surface patch are the roots of the equation

$$\begin{vmatrix} \mathcal{L} - \kappa E & \mathcal{M} - \kappa F \\ \mathcal{M} - \kappa F & \mathcal{N} - \kappa G \end{vmatrix} = 0$$

If κ_1 and κ_2 are the roots, then $\kappa_1\kappa_2$ is called Gaussian curvature of the surface and $\frac{\kappa_1+\kappa_2}{2}$ is called mean curvature of the surface.

Theorem 5.1.2. Let κ_1 and κ_2 be the principal curvatures at a point P of a surface patch σ. Then

1. κ_1 and κ_2 are real numbers
2. If $\kappa_1 = \kappa_2 = \kappa$, say, then $F_{II} = \kappa F_I$ and hence every tangent vector to σ at P is principal vector
3. If $\kappa_1 \neq \kappa_2$, then any two non-zero principal vectors d_1 and d_2 corresponding to κ_1 and κ_2, respectively, are perpendicular.

(For case 2, P is called an umbilic).

Proof. For 1, let t_1 and t_2 be any two perpendicular unit tangent vectors to the surface at P (not yet known to be principal vectors). Define ξ_i, η_i, T_i for $i = 1, 2$ as done in the previous unit. Let

$$A = \begin{bmatrix} \xi_1 & \xi_2 \\ \eta_1 & \eta_2 \end{bmatrix}$$

By multiplying out the matrices, it is easy to check that

$$\begin{aligned} A^t F_I A &= \begin{bmatrix} T_1^t F_I T_1 & T_1^t F_I T_2 \\ T_2^t F_I T_1 & T_2^t F_I T_2 \end{bmatrix} \\ &= \begin{bmatrix} t_1.t_1 & t_1.t_2 \\ t_2.t_1 & t_2.t_2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

Since t_1 and t_2 are perpendicular unit vectors.
Let $G_{II} = A^t F_{II} A$. Then G_{II} is still (real and) symmetric because

$$\begin{aligned} G_{II}^t &= A^t F_{II} (A^t)^t \\ &= A^t F_{II} A \\ &= G_{II} \end{aligned}$$

From the theory of Linear algebra, we can say, there is an orthogonal matrix B such that

$$B^t G_{II} B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

for some real numbers λ_1 and λ_2. Let $C = AB$. Then

$$\begin{aligned} C^t F_I C &= B^t (A^t F_I A) B \\ &= B^t B \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \end{aligned} \tag{5.1.2}$$

because B is orthogonal, and

$$\begin{aligned} C^t F_{II} C &= B^t (A^t F_{II} A) B \\ &= B^t G_{II} B \\ &= \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \end{aligned} \tag{5.1.3}$$

Now, C is invertible. So,

$$\det(F_{II} - \kappa F_I) = 0 \text{ if and only if } \det(C^t(F_{II} - \kappa F_I)C) = 0,$$

hence,

$$\det(F_{II} - \kappa F_I) = 0 \text{ if and only if}$$

$$\det\left(\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} - \kappa \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = 0$$

Hence the principal curvatures are the roots of

$$\begin{vmatrix} \lambda_1 - \kappa & 0 \\ 0 & \lambda_2 - \kappa \end{vmatrix} = 0,$$

For 2,, suppose that the principal curvatures are equal to κ_1, say. Then $\lambda_1 = \lambda_2 = \kappa$, and equations (5.1.2) and (5.1.3) give

$$C^t F_I C = I,\ C^t F_{II} C = \kappa I$$

Hence

$$\begin{aligned} C^t(F_{II} - \kappa F_I)C &= 0 \\ \implies F_{II} - \kappa F_I &= 0 \\ \implies F_{II} &= \kappa F_I \end{aligned}$$

since C and C^t are invertible.

Obviously, if D is any 2×1 column matrix,

$$(F_{II} - \kappa F_I)D = 0$$

It follows that every tangent vector to σ at P is a principal vector.

Finally, for 3, let

$$\begin{aligned} d_i &= \alpha_i \sigma_u + \beta_i \sigma_v \\ D_i &= \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix}, \end{aligned}$$

for $i = 1, 2$. Then by $t_1.t_2 = T_1^t F_I T_2$, $d_1 d_2 = D_1^t F_I D_2$. From the definition of principal vector

$$F_{II} D_1 = \kappa_1 F_1 D_1,\ F_{II} D_2 = \kappa_2 F_1 D_2$$

Hence,

$$\begin{aligned} D_2^t F_{II} D_1 &= \kappa_1 (d_1 d_2) \\ D_2^t F_{II} D_2 &= \kappa_2 (d_1 d_2) \end{aligned} \tag{5.1.4}$$

But since $D_1^t F_{II} D_2$ is a 1×1 matrix, it is equal to its transpose

$$\begin{aligned} D_1^t F_{II} D_2 &= (D_1^t F_{II} D_2)^t \\ &= D_2^t F_{II}^t D_1 \\ &= D_2^t F_{II} D_1, \end{aligned}$$

the last equality comes from the fact that F_{II} is symmetric. Hence equation (5.1.4) gives

$$\kappa_1(d_1 d_2) = \kappa_2(d_1 d_2)$$

So, $\kappa_1 \neq \kappa_2$ implies $d_1 d_2 = 0$, that is, d_1 and d_2 are perpendicular. □

Example 5.1.3. Find the Gaussian curvature of the right circular cylinder.

We consider the circular cylinder of radius 1 and axis z-axis, parametrized in the usual way

$$\sigma(u, v) = (\cos v, \sin v, u).$$

Here,

$$E = \sigma_u \sigma_u = 1,\ F = \sigma_u \sigma_v = 0,\ G = \sigma_v \sigma_v = 1$$

Also

$$\mathcal{L} = 0,\ \mathcal{M} = 0,\ \mathcal{N} = 1.$$

So, the principal curvatures are the roots of

$$\begin{vmatrix} 0 - \kappa & 0 \\ 0 & 1 - \kappa \end{vmatrix} = 0$$

Hence

$$\kappa(\kappa - 1) = 0,\ \text{or}, \kappa = 0,\ \text{or}\ 1$$

Hence the principal curvatures are 0 and 1. Thus,

$$\text{Gaussian curvature} = 0 \times 1 = 0$$

$$\text{Mean Curvature} = \frac{0 + 1}{2} = \frac{1}{2}$$

Exercise 5.1.4. Find the Gaussian curvature of the unit sphere.

Note 5.1.5. In the following, we shall study Euler's theorem which will relate the curvature of a surface and curvature of a curve defined on a surface.

Theorem 5.1.6. (**Euler's Theorem**) Let γ be a curve on a surface patch σ, and let κ_1 and κ_2 be the principal curvatures of σ, with non-zero principal non-zero vectors t_1 and t_2. Then the normal curvature of γ is

$$\kappa_\eta = \kappa_1 \cos^2 \theta + \kappa_2 \sin^2 \theta$$

where θ is the angle between $\dot{\gamma}$ and t_1.

[thick,->] (0,0) – (6,0) node[anchor=north west] t_1; [thick,->] (0,0) – (0,6) node[anchor=south east] t_2; [thick,->] (0,0) – (4,5); (0.5,0) arc (0:50:0.5cm) node [anchor=east] θ; [dashed] (4,0) – (4,5);

Proof. We can assure that γ is a unit speed. Let t be the tangent vector of γ, and let

$$t = \xi\sigma_u + \eta\sigma_v,$$

$$T = \begin{bmatrix} \xi \\ \eta \end{bmatrix}$$

Suppose that

$$\kappa_1 = \kappa_2 = \kappa, \text{say.}$$

Now,

$$\begin{aligned} \kappa_\eta &= T^t F_{II} t \\ &= \kappa T^t F_I T \\ &= \kappa t.t \\ &= \kappa \end{aligned}$$

This agrees with the formula in the statement since

$$\begin{aligned} \kappa_1 \cos^2\theta + \kappa_2 \sin^2\theta &= \kappa(\cos^2\theta + \sin^2\theta) \\ &= \kappa \end{aligned}$$

Now, assume that $\kappa_1 \neq \kappa_2$, so that t_1 and t_2 are perpendicular. We might as well assume that t_1 and t_2 are unit vectors.

$$\begin{aligned} t_i &= \xi_i\sigma_u + \eta_i\sigma_v \\ T_i &= \begin{bmatrix} \xi_i \\ \eta_i \end{bmatrix} \end{aligned}$$

for $i = 1, 2$. Now

$$\dot{\gamma} = \cos\theta t_1 + \sin\theta t_2$$

So,

$$\cos\theta(\xi_1\sigma_u + \eta_1\sigma_v) + \sin\theta(\xi_2\sigma_u + \eta_2\sigma_v) = \xi\sigma_u + \eta\sigma_v$$

Hence

$$\begin{aligned} \xi_1\cos\theta + \xi_2\sin\theta &= \xi \\ \eta_1\cos\theta + \eta_2\sin\theta &= \eta \end{aligned}$$

Thus,

$$\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \cos\theta \begin{bmatrix} \xi_1 \\ \eta_1 \end{bmatrix} + \sin\theta \begin{bmatrix} \xi_2 \\ \eta_2 \end{bmatrix}$$

Therefore,

$$T = \cos\theta T_1 + \sin\theta T_2.$$

Hence

$$\begin{aligned}\kappa_\eta &= T^t F_{II} T \\ &= (\cos\theta T_1^t + \sin\theta T_2^t) F_{II} (\cos\theta T_1 + \sin\theta T_2) \\ &= \cos^2\theta T_1^t F_{II} T_1 + \cos\theta\sin\theta (T_1^t F_{II} T_2 + T_2^t F_{II} T_1) + \sin^2 T_2^t F_{II} T_2\end{aligned}$$

Now,

$$\begin{aligned}T_i^t F_{II} T_j = \kappa_i T_i^t F_{II} T_j &= \kappa_i, \text{ if } i = j \\ &= 0, \text{ otherwise}\end{aligned}$$

Hence

$$\kappa_\eta = \kappa_1 \cos^2\theta + \kappa_2 \sin^2\theta$$

□

Theorem 5.1.7. Let $\sigma(u, v)$ be a surface patch with first and second fundamental forms $Edu^2 + 2Fdudv + Gdv^2$ and $\mathcal{L}du^2 + 2\mathcal{M}dudv + \mathcal{N}dv^2$, respectively. Then

1. $\kappa = \frac{\mathcal{L}\mathcal{N} - \mathcal{M}^2}{EG - F^2}$
2. $H = \frac{\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E}{2(EG - F^2)}$
3. The principal curvatures are
$$H \pm \sqrt{H^2 - \kappa}.$$

Proof. The principal curvatures are the roots of

$$\begin{vmatrix} \mathcal{L} - \kappa E & \mathcal{M} - \kappa F \\ \mathcal{M} - \kappa F & \mathcal{N} - \kappa G \end{vmatrix} = 0$$

Hence

$$(EG - F^2)\kappa^2 - (\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E)\kappa + \mathcal{L}\mathcal{N} - \mathcal{M}^2 = 0, \tag{5.1.5}$$

The above equation is quadratic in κ. Hence

$$\kappa_1 + \kappa_2 = \frac{\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E}{(EG - F^2)}$$

Hence,

$$H = \frac{\kappa_1 + \kappa_2}{2} = \frac{\mathcal{L}G - 2\mathcal{M}F + \mathcal{N}E}{2(EG - F^2)} \tag{5.1.6}$$

and

$$K = \kappa_1\kappa_2 = \frac{\mathcal{L}\mathcal{N} - \mathcal{M}^2}{EG - F^2} \tag{5.1.7}$$

K is the Gaussian curvature. Then, from equations (5.1.5), (5.1.6) and (5.1.7),

$$\kappa^2 - 2H\kappa + K = 0.$$

Hence

$$H \pm \sqrt{H^2 - \kappa}$$

are the principal curvatures. □

Definition 5.1.8. A surface is called flat if its Gaussian curvature is zero.

Exercises

1. Compute Gaussian curvature of a sphere.
2. Calculate the principal curvatures of a helicoid.
3. Calculate the principal curvatures of catenoid.
4. Let $\gamma(t) = \sigma(u(t), v(t))$ be a regular, but not necessarily unit speed curve on a surface σ and denote d/dt by a dot. Prove that the normal curvature of γ is
$$\kappa_n = \frac{L\dot{u}^2 + 2M\dot{u}\dot{v} + N\dot{v}^2}{E\dot{u}^2 + 2F\dot{u}\dot{v} + G\dot{v}^2}.$$
5. Show that a curve on a surface is a line of curvature if and only if its geodesic torsion vanishes everywhere.

Summary

In this unit we have learnt how to calculate the Gaussian and mean curvatures of a surface.

Unit 6

Course Structure

- Introduction
- Geodesics on surfaces
- Differential equation for geodesics on surfaces

6.1 Introduction

In simple language a geodesic means the line of shortest distance between two points. In our Euclidean plane, a geodesic is a straight line. In this unit, we shall see that on a sphere, geodesics are great circles. Here, we will first deduce the differential equation of geodesics on a surface.

6.1.1 Geodesics on Surface

Definition 6.1.1. A curve γ on a surface S is called a geodesic if $\ddot{\gamma}(t)$ is zero or perpendicular to the surface at the point $\gamma(t)$, that is, parallel to its unit normal, for all values of the parameter t.

Theorem 6.1.2. Any geodesic has constant speed.

Proof. Let $\gamma(t)$ be a geodesic on a surface S. Then denoting $\frac{d}{dt}$ by a dot, we have

$$\frac{d}{dt}||\dot{\gamma}||^2 = \frac{d}{dt}(\dot{\gamma}.\dot{\gamma}) = 2\ddot{\gamma}.\dot{\gamma}$$

Since γ is a geodesic, $\ddot{\gamma}$ is perpendicular to the tangent plane and is therefore perpendicular to the tangent vector $\dot{\gamma}$. So, $\ddot{\gamma}.\dot{\gamma} = 0$ and the last equation shows that $||\ddot{\gamma}||$ is constant. □

Theorem 6.1.3. A curve on a surface is a geodesic if and only if its geodesic curvature is zero.

Proof. It is sufficient to consider a unit speed curve γ contained in a patch σ of the surface. Let N be the standard unit normal of σ, so that

$$\kappa g = \ddot{\gamma}.(N \times \dot{\gamma}) \tag{6.1.1}$$

If $\ddot{\gamma}$ is parallel to N, it is obviously perpendicular to $N \times \dot{\gamma}$, so by (6.1.1),

$$\kappa g = 0.$$

Conversely, suppose that $\kappa g = 0$. Then $\ddot{\gamma}$ is perpendicular to $N \times \dot{\gamma}$. But then, since $\dot{\gamma}, N$ and $N \times \dot{\gamma}$ are perpendicular unit vectors in $\mathbb{R}^3$, and since $\ddot{\gamma}$ is perpendicular to $\dot{\gamma}$, it follows that $\ddot{\gamma}$ is perpendicular to N. □

Exercise 6.1.4. 1. Prove that for normal section, $\kappa g = 0$.

2. All great circles on a sphere are geodesics.

6.1.2 Geodesic Equations

A curve γ on a surface S is a geodesic if and only if for any part $\gamma(t) = \sigma(u(t), v(t))$ of γ contained in a surface patch σ of S, the following two equations are satisfied

$$\begin{aligned} \frac{d}{dt}(E\dot{u} + F\dot{v}) &= \frac{1}{2}(E_u\dot{u}^2 + 2F_u\dot{u}\dot{v} + G_u\dot{v}^2) \\ \frac{d}{dt}(F\dot{u} + G\dot{v}) &= \frac{1}{2}(E_v\dot{u}^2 + 2F_v\dot{u}\dot{v} + G_v\dot{v}^2) \end{aligned}$$

where $Edu^2 + 2Fdudv + Gdv^2$ is the first fundamental form of σ.

Proof. Since $\{\sigma_u, \sigma_v\}$ is a basis of the tangent plane of σ, γ is a geodesic if and only if $\ddot{\gamma}$ is perpendicular to σ_u and σ_v. Since

$$\dot{\gamma} = \sigma_u\dot{u} + \sigma_v\dot{v},$$

this is equivalent to

$$\left(\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v})\right).\sigma_u = 0 \tag{6.1.2}$$

$$\left(\frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v})\right).\sigma_v = 0 \tag{6.1.3}$$

we show that these two equations are equivalent to the two geodesic equations. The left hand side of equation (6.1.2) is equal to

$$\begin{aligned} \frac{d}{dt}(\sigma_u\dot{u} + \sigma_v\dot{v}).\sigma_u) - (\sigma_u\dot{u} + \sigma_v\dot{v})).\frac{\sigma_u}{dt} &= \frac{d}{dt}(E\dot{u} + F\dot{v}) - (\sigma_u\dot{u} + \sigma_v\dot{v})(\sigma_{uu}\dot{u} + \sigma_{vv}\dot{v}) \\ &= \frac{d}{dt}(E\dot{u} + F\dot{v}) \\ &- (\dot{u}^2(\sigma_u\sigma_{uu})\dot{u}\dot{v}(\sigma_u\sigma_{uv} + \sigma_v\sigma_{uu}) + \dot{v}^2(\sigma_v\sigma_{uv})) \end{aligned} \tag{6.1.4}$$

Now,

$$\begin{aligned} E_u &= (\sigma_u.\sigma_u)_u \\ &= \sigma_{uu}\sigma_u + \sigma_u\sigma_{uu} \\ &= 2\sigma_u\sigma_{uu} \end{aligned}$$

So,

$$\sigma_u\sigma_{uu} = \frac{1}{2}E_u$$

Similarly,

$$\sigma_v \sigma_{uv} = \frac{1}{2} G_u$$

Finally,

$$\sigma_v \sigma_{uv} + \sigma_v \sigma_{uu} = (\sigma_u . \sigma_u)_{\dot{u}} = F_u$$

Substituting these values in equation (6.1.4), gives

$$\left(\frac{d}{dt}(\sigma_u \dot{u} + \sigma_v \dot{v}) \right) . \sigma_u = \frac{d}{dt}(E\dot{u} + F\dot{v}) - \frac{1}{2}(E_u \dot{u}^2 + 2F_u \dot{u}\dot{v} + G_u \dot{v}^2).$$

This shows that the equation (6.1.2) is equivalent to the first geodesic equation. Similarly the other. □

Example 6.1.5. Determine the geodesics on S^2 by solving the geodesic equations.

For the usual parametrisation by latitude and longitude coordinate

$$\sigma(u, v) = (\cos\theta \cos\phi, \cos\theta \sin\phi, \sin\theta),$$

we know that the first fundamental form of S^2 is

$$d\theta^2 + \cos^2\theta d\phi^2$$

We might as well restrict ourselves to unit speed curves

$$\gamma(t) = \sigma(\theta(t), \phi(t)),$$

so that

$$\dot{\theta}^2 + \dot{\phi}^2 \cos^2\theta = 1,$$

and if γ is a geodesic the second geodesic equation gives

$$\frac{d}{dt}(\dot{\phi} \cos^2\theta) = 0,$$

so that,

$$\dot{\phi} \cos^2\theta = \Omega,$$

where, Ω is a constant. If $\Omega = 0$, then $\dot{\phi} = 0$, so ϕ is constant and γ is part of a meridian. We assume that $\Omega \neq 0$ from now on.

The unit speed condition gives

$$\begin{aligned} \dot{\theta} &= 1 - \dot{\phi}^2 \cos^2\theta \\ \dot{\theta}^2 &= 1 - \frac{\Omega^2}{\cos^2\theta}, \end{aligned}$$

so along the geodesics we have

$$\left(\frac{d\phi}{d\theta} \right)^2 = \frac{\dot{\phi}^2}{\dot{\theta}^2} = \frac{1}{\cos^2\theta \left(\frac{\cos^2\theta}{\Omega^2} - 1 \right)}$$

and hence

$$\pm(\phi - \phi_0) = \int \frac{d\theta}{\cos\theta \sqrt{\frac{\cos^2\theta}{\Omega^2} - 1}},$$

where, ϕ_0 is a constant. The integral can be evaluated by making the substitution $u = \tan\theta$. This gives

$$\pm(\phi - \phi_0) = \int \frac{du}{\sqrt{\frac{1}{\Omega^2} - 1 - u^2}} \arcsin\left(\frac{u}{\sqrt{\frac{1}{\Omega^2} - 1}}\right),$$

and hence

$$\tan\theta = \pm\sqrt{\frac{1}{\frac{1}{\Omega^2} - 1}}\sin(\phi - \phi_0).$$

This implies that the coordinates $x = \cos\theta\cos\phi,\ y = \cos\theta\sin\phi,\ z = \sin\theta$ of $\gamma(t)$ satisfy $z = ax + by$.

Hence the geodesics are the intersection of the sphere and planes passing through the centre. So geodesics are the great circles.

Exercises

1. Let p be a point of a surface S, and let t be a unit tangent vector to S at p. Then show that there exists a unique unit speed geodesic γ on S which passes through p and has tangent vector t there.
2. Show that, if p and q are distinct points of a circular cylinder, there are either two or infinitely many geodesics on the cylinder joining p and q. Which pairs p, q have the former property?
3. Find the geodesics on circular cylinder by solving the geodesic equations.
4. Show directly that the parameter of any curve satisfying the geodesic equations is proportional to arc length.
5. Find geodesics on unit spheres.

Summary

In this unit, we have studied properties of geodesics and have deduced differential equations of geodesics and have seen examples for the same.

Unit 7

Course Structure

- Introduction
- Plateau's Problem
- Examples of minimal surface
- Exercises

7.1 Introduction

It is known that on a surface the curve with shortest length between two given points is a geodesic. Similarly, in this chapter we shall study the problem of finding a surface patch of minimal area with a fixed curve as its boundary. Such a problem is known as Plateau's Problem. We shall see that solution of Plateau's problem is a surface patch whose mean curvature is zero everywhere. A surface whose mean curvature is zero everywher is known as minimal surface.

7.2 Plateau's Problem

Consider a family of surface patches $\sigma^\tau : U \to R^3$, where U is an open subset of R^2 independent of τ, and τ lies in some open interval $(-\delta, \delta)$, for some $\delta > 0$. Consider $\sigma = \sigma^0$. The family is required to be smooth, i.e., the map $(u, v, \tau) \to \sigma^\tau(u, v)$ from the open subset $\{(u, v, \tau)|(u, v) \in U, \tau \in (-\delta, \delta)\}$ of R^3 to R^3 is smooth. The surface variation of the family is the function $\phi : U \to R^3$ given by

$$\phi = \dot{\sigma}^\tau|_{\tau=0},$$

where a dot denotes derivative with respect to τ.

Suppose π is a simple closed curve which is contained in U along with its interior. Then π corresponds to a simple closed curve $\gamma 6\tau = \sigma^\tau o\pi$ in the surface patch σ^τ, and we define the area function $A(\tau)$ to be the area of the part of σ^τ inside γ^τ :

$$A(\tau) = \int\int_{imt(\pi)} dA_{\sigma^\tau}.$$

Observe that if we are considering a family of surfaces with a fixed boundary curve γ, then $\gamma^\tau = \gamma$ for all τ and hence $\phi^\tau(u, v) = 0$, when (u, v) is a point on the curve π.

Theorem 7.2.1. If the surfac variation ϕ^τ vanishes along the boundary curve π, then

$$\dot{A}(0) = -2\int\int_{int(\pi)} H(EG - F^2)^{\frac{1}{2}}\alpha dudv,$$

where H is the mean curvature of σ, E, F and G are the coefficients of its first fundamental form, $\alpha = \phi.N$, and N is the standard unit normal of σ.

Proof. Let $\phi^\tau = \dot{\sigma}^\tau$, so that $\phi^0 = \phi$, and let N^τ be the standard unit normal of σ^τ. There are smooth functions α^τ, β^τ and γ^τ of (u, v, τ) such that

$$\phi^\tau = \alpha^\tau N^\tau + \beta 6\tau\sigma_u^\tau + \gamma^\tau\sigma_v^\tau,$$

so that $\alpha = \alpha^0$. To simplify the notation, we ommit the superscript τ in the remaining part of the proof. At the end of the proof we take $\tau = 0$.

We see that

$$A(\tau) = \int\int_{-nt(\pi)} ||\sigma_u \times \sigma_v||dudv = \int\int_{\int(\pi)} N.(\sigma_u \times \sigma_v)dudv,$$

so

$$\dot{A} = \int\int_{int(\pi)} \frac{\partial}{\partial\tau}(N.(\sigma_u \times \sigma_v))dudv. \tag{7.2.1}$$

Now,

$$\frac{\partial}{\partial\tau}(N.(\sigma_u \times \sigma_v)) = \dot{N}(\sigma_u \times \sigma_v) + \dot{N}(\dot{\sigma_u} \times \sigma_v) + N.(\sigma_u \times \dot{\sigma_v}). \tag{7.2.2}$$

Since N is a unit vector,

$$\dot{N}.(\sigma_u \times \sigma_v) = \dot{N}.N||\sigma_u \times \sigma_v|| = 0.$$

On the other hand,

$$\begin{aligned} N.(\dot{\sigma}_u \times \sigma_v) &= \frac{(\sigma_u \times \sigma_v).(\dot{\sigma}_u \times \sigma_v)}{||\sigma_u \times \sigma_v||} \\ &= \frac{(\sigma_u.\dot{\sigma}_u)(\sigma_v.\sigma_v) - (\sigma_v.\sigma_v)(\sigma_v.\dot{\sigma}_u)}{||\sigma_u \times \sigma_v||} \\ &= \frac{G(\sigma_u.\dot{\sigma}_u) - F(\sigma_v.\dot{\sigma}_u)}{(EG - F^2)^{\frac{1}{2}}}, \end{aligned}$$

since it is known that $||\sigma_u \times \sigma_v|| = (EG - F^2)^{\frac{1}{2}}$.

Similarly

$$N.(\sigma_u \times \dot{\sigma}_v) = \frac{E(\sigma_v.\dot{\sigma}_v) - F(\sigma_u.\dot{\sigma}_v)}{(EG - F^2)^{\frac{1}{2}}}.$$

Substituting these results in (7.2.2) we get

$$\frac{\partial}{\partial\tau}(N.(\sigma_u \times \sigma_v)) = \frac{E(\sigma_v.\dot{\sigma}_v) - F(\dot{\sigma}_u.\sigma_v + \sigma_u\dot{\sigma}_v) + G(\sigma_u.\dot{\sigma}_u)}{(EG - F^2)^{\frac{1}{2}}} \tag{7.2.3}$$

Now

$$\dot{\sigma}_u = \phi_u = \sigma_u N + \beta_u\sigma_u + \gamma_u\sigma_v + \alpha N_u + \beta\sigma_{uu} + \gamma\sigma_{uv}.$$

So

$$\sigma_u.\dot{\sigma}_u = E\beta_u + F\gamma_u + (\sigma_u.N_u)\alpha + (\sigma_u.\sigma_{uu})\beta + (\sigma_u.\sigma_{uv})\gamma.$$

Since $\sigma_u.N_u = -\sigma_{uu}.N = -L$, $\sigma_u.\sigma_{uu} = \frac{1}{2}E_u$ and $\sigma_u.\sigma_{uv} = \frac{1}{2}E_v$, we get

$$\sigma_u.\dot{\sigma}_u = E\beta_u + F\gamma_u - L\alpha + \frac{1}{2}E_u\beta + \frac{1}{2}E_v\gamma.$$

Similarly

$$\sigma_v.\dot{\sigma}_u = F\beta_u + G\gamma_u - M\alpha + (F_u - \frac{1}{2}E_v)\beta + \frac{1}{2}G_u\gamma.$$

$$\sigma_u.\dot{\sigma}_v = E\beta_v + F\gamma_v - M\alpha + \frac{1}{2}E_v\beta + (F_v - \frac{1}{2}G_u)\gamma.$$

$$\sigma_v.\dot{\sigma}_v = F\beta_v + G\gamma_v - N\alpha + \frac{1}{2}G_u\beta + \frac{1}{2}G_v\gamma.$$

Substituting these last four equations into the right hand side of equation (7.2.3), simplifying, and using the formula for the mean curvature H, we see that

$$\frac{\partial}{\partial\tau}(N.(\sigma_u \times \sigma_v)) = (\beta(EG - F^2)^{\frac{1}{2}})_u + (\gamma(EG - F^2)^{\frac{1}{2}})_v - 2\alpha H(EG - F^2)^{\frac{1}{2}}. \tag{7.2.4}$$

Comparing with equation (7.2.1) and reinstating the superscripts, we note that we must prove that

$$\int\int_{\int(\pi)}((\beta^0(EG - F^2)^{\frac{1}{2}})_u + (\gamma^0(EG - F^2)^{\frac{1}{2}})_v)dudv = 0.$$

But by Green's Theorem, this integral is equal to

$$\int_\pi (EG - F^2)^{\frac{1}{2}}(\beta^0 dv - \gamma^0 du),$$

and this obviously vanishes because $\beta^0 = \gamma^0 = 0$ along the boundary curve π. This completes the proof.

Example 7.2.2. The simplest minimal example is plane for which both principal curvatures are zero everywhere.

Example 7.2.3. A catenoid is obtained by rotating a curve $x = \frac{1}{a}\cosh az$ in the xz-plane around the z axis, where $a > 0$ is a constant. It can be shown that this is a minimal surface. The catenoud is a surface of revolution. In fact, apart from the plane, it is the only surface of revolution.

Proposition 7.2.4. Any minimal surface of revolution is either part of a plane or can be obtained by applying a rigid motion to part of a catenoid.

Proof. By applying a rigid motion, we can assume that the axis of the surface S is the z-axis and the profile curve lies in the xz-plane. We parametrise S in the usual way

$$\sigma(u, v) = (f(u)cosv, f(u)sinv, g(u)),$$

where the profile curve $u \to (f(u), 0, g(u))$ is assumed to be unit speed and $f > 0$. We can calculate the first and second fundamental form as $du^2 + f(u)^2dv^2$ and $(\dot{f}\ddot{g} - \ddot{f}\dot{g})du^2 + f\dot{g}dv^2$, respectively, a dot denotes derivative with respect to u. Using the formula for the mean curvature we get

$$H = \frac{1}{2}(\dot{f}\ddot{g} - \ddot{f}\dot{g} + \frac{\dot{g}}{f}).$$

Consider that for some value of u, say $u - u_0$, we have $\dot{g}(u_0) \neq 0$. We shall then have $\dot{g}(u) \neq 0$ for u in some open interval containing u_0. Let (α, β) be the largest such interval. Supporting now that $u \in (\alpha, \beta)$, the unit speed condition $\dot{f}^2 + \dot{g}^2 = 1$ gives

$$\dot{f}\ddot{g} - \ddot{f}\dot{g} = -\frac{\ddot{f}}{\dot{g}},$$

So, we get

$$H = \frac{1}{2}(\frac{\dot{g}}{f} - \frac{\ddot{f}}{\dot{g}}).$$

Since $\dot{g}^2 = 1 - \dot{f}^2$, S is minimal if and only if

$$f\ddot{f} = 1 - \dot{f}^2. \tag{7.2.5}$$

To solve the differential equation (7.2.5), put $h = \dot{f}$. We note that

$$\ddot{f} = \frac{dh}{dt} = \frac{dh}{df}\frac{df}{dt} = h\frac{dh}{df}.$$

Hence equation (7.2.5) becomes

$$fh\frac{dh}{df} = 1 = h^2.$$

We note that since $\dot{g} \neq 0$, we have $h^2 \neq 1$, so we can integrate this equation as follows:

$$\int \frac{hdh}{1-h^2} = \int \frac{df}{f}.$$

So

$$\frac{1}{\sqrt{1-h^2}} = af,$$

Thus, $h = \frac{\sqrt{a^2f^2-1}}{af}$, where a is a non-zero constant. Writing $h = \frac{df}{du}$ and integrating again

$$\int \frac{afdf}{\sqrt{a^2f^2-1}} = \int du,$$

So

$$f = \frac{1}{a}\sqrt{1 + a^2(u+b)^2},$$

where b is a constant . By change of parameter $u \to u + b$, we can assume that $b = 0$. So $f = \frac{1}{a}\sqrt{1 + a^2u^2}$. To compute g we have

$$\dot{g}^2 = 1 - \dot{f}^2 = 1 - h^2 = \frac{1}{a^2f^2}.$$

So

$$\frac{dg}{du} = \pm\frac{1}{\sqrt{1 + a^2u^2}}$$

So

$$g = \pm\frac{1}{a}sinh^{-1}(au) + c$$

So

$$au = \pm sinh(a(g - c)),$$

So

$$f = \frac{1}{a} cosh(a(g-c)).$$

Thus the profile curve of S is

$$x = \frac{1}{a} cosh(a(z-c)).$$

By the translation along the z-axis, we can assume that $c = 0$. So we have a catenoid.

Now, suppose $\beta < \infty$. Then if the profile curve is defined for values $u \geq \beta$, we must have $\dot{g}\beta = 0$, for otherwise $\dot{g}$ would be non-zero on an open interval containing β, which would contradict our assumption that (α, β) is the largest open interval containing u_0 on which $\dot{g} \neq 0$. But the formulas above show that

$$\dot{g}^2 = \frac{1}{1 + a^2 u^2}$$

if $u \in (\alpha, \beta)$, so, since $\dot{g}$ is a continuous function of u, $\dot{g}(\beta) = \pm(1 + \alpha^2\beta^2)^{-\frac{1}{2}} \neq 0$. This contradiction shows that the profile curve is not defined for values of $u \geq \beta$. Of course, this also holds trivially $if \beta = \infty$. A similar argument applies to α, and shows that (α, β) is the entire domain of definition of the profile curve. Hence the whole of S is part of a catenoid.

The only remaining case to consider is that in which $\dot{g}(u) = 0$ for all values of u for which the profile curve is defined. But then $g(u)$ is a constant, say,d, and S is part of the plane $z = d$.

Exercises

1. Show that any rigid motion of R^3 takes a minimal surface to another minimal surface, as does any dilation $(x, y, z) \to \alpha(x, y, z)$, where a is a non-zero constant.

2. Show that $z = f(x, y)$, where f is a smooth function of two variables, is a minimal surface if and only if
$$(1 + f_y^2) f_{xx} - 2 f_x f_y f_{xy} + (1 + f_x^2) f_{yy} = 0.$$

3. Show that every umbilic on a minimal surface is a planar point.

4. Show that the Gaussian curvature of a minimal surface is ≤ 0 everywhere and that it is zero everywhere if and only if the surface is part of a plane.

5. Show that there is no compact minimal surface.

6. Show that a ruled minimal surface is part of a plane or part of a helicoid.

7. Show that the surface $\sigma(u, v) = (u - \frac{1}{3}u^3 + uv^2, v - \frac{1}{3}v^3 + vu^2, u^2 - v^2)$ is minimal.

8. Show that the helecoid is a minimal surface.

9. Show that a generalised cylinder is a minimal surface only when the cylinder is part of a plane.

10. A translation surface is a surface of the form $z = f(x) + g(y)$, where f and g are smooth functions. Show that this is a minimal surface if and only if
$$\frac{d^2 f/dx^2}{1 + (df/dx)^2} = -\frac{d^2 g/dy^2}{1 + (dg/dy)^2}.$$

11. Test whether the Catalan's surface

$$\sigma(u,v) = (u - \sin u \cosh v, 1 - \cos u \cos v, -4\sin\frac{u}{2}\sinh\frac{v}{2})$$

is a conformally parametrised minimal surface.

Summary

In this chapter concept of minimal surfaces has been given. Some standard results associated with minimal surfaces have been established. Some examples have been provided. A list of exercieses has been given for further practice.

Unit 8

Course Structure

- Introduction
- Gausse's theorema Egregium
- Gauss-Bonnet Theorem
- Exercises

8.1 Introduction

'Egregium' means remarkable. So 'Gaussse's Theorema Egregium' means Gausse's remarkable theorem. In this unit, we shall study the remarkable theorem of Gauss. The statement of the theorem is following.

8.2 Gausse's remarkable theorem

Theorem 8.2.1. The Gaussian curvature of a surface is preserved by isometries.

(**Note.** Since isometry depends on only first fundamental form of a surface, the alternative statement of the above theorem is : The gaussian curvature of a surface depends only on the first fundamental form of the surface.)

Proof. A diffeomorphism $f : S_1 \to S_2$ is an isometry if and only if, for any surface patch σ_1 of S_1, the patches σ_1 and $fo\sigma_1$ of S_1 and S_2, respectively, have the same first fundamental form. So to prove the theorem, it is enough to consider the case of a surface patch σ on S_1 and to prove that, if σ and $fo\sigma$ have the same first fundamental forms, then they have the same Gaussian curvature. We know that the Gaussian curvature K is given by

$$K = \frac{LN - M^2}{EG - F^2}$$

So we see that the denominator of the above expression depends on the 1st fundamental form but the numerator depends on the 2nd fundamental form. So we have to show that the expression $LN - M^2$ can be expressed in terms of coefficients of 1st fundamental form. To prove the theorem, we shall make use of smooth orthonormal basis $\{e', e''\}$ of the tangent plane at each point of the surface patch, where 'smooth' means e' and e'' are

smooth functions of the surface parameters (u, v). Then $\{e', e'', N\}$ is an orthonormal basis of R^3. N is standard unit normal of σ. We shall assume it is right handed i.e., $N = e' \times e''$. This can always be achieved by interchanging by e' and e'' if necessary. Note that here dash is just a symbol not derivative.

We can express the partial derivatives of e' and e'' with respect to u and v interms of the orthonormal basis$\{e', e'', N\}$. Since both partial derivatives of e' are perpendicular to e' , the e' components of e'_u and e'_v are zero. Similar for e''. Thus,

$$\begin{aligned} e'_u &= \alpha e'' + \lambda' N, \\ e'_v &= \beta e'' + \mu' N, \\ e''_u &= -\alpha' e' + \lambda'' N, \\ e''_v &= -\beta' e' + \mu'' N, \end{aligned}$$

for some scalars $\alpha, \beta, \alpha', \beta', \lambda', \mu', \lambda'', \mu''$(which may depend on u and v.) Moreover, by differentiating the equation $e'.e'' = 0$ with respect to u, we see that $e'_u.e'' = -e'.e''_u$., i.e., $\alpha' = \alpha$ and similarly $\beta' = \beta$. Thus

$$\begin{aligned} e'_u &= \alpha e'' + \lambda' N, \\ e'_v &= \beta e'' + \mu' N, \\ e''_u &= -\alpha e' + \lambda'' N, \\ e''_v &= -\beta e' + \mu'' N. \end{aligned} \tag{8.0}$$

To complete the proof of the theorem let us fisrst state and prove the following lemma.

Lemma 8.2.2. If $\{e', e'', N\}$ is an orthonormal basis of R^3, then

$$\begin{aligned} e'_u e''_v - e''_u . e'_v &= \lambda' \mu'' - \lambda'' \mu' && (8.2.1) \\ &= \alpha_v - \beta_u && (8.2.2) \\ &= \frac{LN - M^2}{(EG - F^2)^{\frac{1}{2}}} && (8.2.3) \end{aligned}$$

Proof of the lemma. Equation (8.2.1) follows from (8.0), since e', e'' and N are perpendicular unit vectors. Next, we compute

$$\begin{aligned} \alpha_v - \beta_u &= \frac{\partial}{\partial u}(e', e''_v) - \frac{\partial}{\partial v}(e'.e''_u) \quad \text{by}(8.0) \\ &= e'_u.e''_v + e'.e''_v - e'_v e''_u - e' e''_{uv} \\ &= e'_u.e''_v - e'_v - e'_v e''_v. \end{aligned}$$

This proves (8.2.2). To prove equation (8.2.3), we use the formula

$$N_u \times N_v = K \sigma_u \times \sigma_v$$

Combining this with the formulas $N = \frac{\sigma_u \times \sigma_v}{||\sigma_u \times \sigma_v||}$, $\quad ||\sigma_u \times \sigma_v|| = (EG - F^2)^{\frac{1}{2}}$. we get

$$N_u \times N_v = \frac{LN - M^2}{(EG - F^2)^{\frac{1}{2}}} N$$

and hence

$$(N_u \times N_v).N = \frac{LN - M^2}{(EG - F^2)^{\frac{1}{2}}} \tag{8.2.4}$$

Since $N = e' \times e''$, we get

$$\begin{aligned}(N_u \times N_v).N &= (N_u \times N_v).(e' \times e'') \qquad (8.2.5)\\ &= (N_u.e')(N_v.e'') - (N_u.e'')(N_v.e')\\ &= (N.e'_u)(N.e''_v) - (N.e''_u)(N.e'_v)\\ &= \lambda'\mu'' - \lambda''\mu' \qquad (8.2.6)\end{aligned}$$

In the above equations, we used $N_u.e' = -N.e'_u$, $N_u.e'' = -N.e''_u$, $N_v, e' = -N.e'_v$, $N_v.e'' = -N.e''_v$. which follows from fifferentiating $N.e' = 0 = N.e''$ with respect to u and v. Putting equations (8.2.4) and (8.2.6) together shows that the right hand side of equation (8.2.1) and (8.2.3) are equal. Since equation (8.2.1) has already been established, this equation (8.2.3). This completes the proof of the lemma. Now we go th prove the main theorem in the following:

Combining equations (8.2.2) and (8.2.3), we get

$$K = \frac{\alpha_v - \beta_u}{(EG - F^2)^{\frac{1}{2}}}, \qquad (8.2.7)$$

so to prove the theorem it suffices to show that, for a suitable choice of $\{e', e''\}$ the scalars α and β depend only on E, F and G. We shall construct $\{e', e''\}$ by applying Gram-Schmidt orthogonalization process to the basis $\{\sigma_u, \sigma_v\}$ of the tangent plane. and will then show that they have the desired property. So, we first define

$$e' = \frac{\sigma_u}{||\sigma_u||} = \epsilon\sigma_u,$$

where $\epsilon = E^{-\frac{1}{2}}$. Now, we look for a vector $e'' = \gamma\sigma_u + \delta\sigma_v$, for some scalars γ, δ such that e'' is a unit vector perpendicular to e'. These conditions give

$$E^{-\frac{1}{2}}(\gamma E + \delta F) = 0, \;\; \gamma^2 E + 2\gamma\delta F + \delta^2 G = 1.$$

The first equation gives $\gamma = -\delta F/E$, and substituting in the second equation then gives

$$\delta^2\left(\frac{F^2}{E} - 2\frac{F^2}{E} + G\right) = 1,$$

So

$$\delta = \frac{E^{\frac{1}{2}}}{(EG - F^2)^{\frac{1}{2}}}, \;\; \gamma = -\frac{FE^{-\frac{1}{2}}}{(EG - F^2)^{\frac{1}{2}}}, \;\; \epsilon = E^{-\frac{1}{2}}. \qquad (8.2.8)$$

Thus

$$e' - \epsilon\sigma_u, \;\; e'' = \gamma\sigma_u + \delta\sigma_v, \qquad (8.2.9)$$

where γ, δ and ϵ depend only on E, F and G. We now compute α and β. First

$$\begin{aligned}\alpha &= e'_u.e''\\ &= (\epsilon_u\sigma_u + \epsilon\sigma_{uu}).(\gamma\sigma_u + \delta\sigma_v)\\ &= \frac{\epsilon_u}{\epsilon}(\epsilon\sigma_u).(\gamma\sigma_u + \delta_v) + \epsilon\gamma\sigma_{uu}.\sigma_v\\ &= \frac{\epsilon_u}{\epsilon}e'e'' + \frac{1}{2}\epsilon\gamma(\sigma_u.\sigma_u)_u + \epsilon\delta((\sigma_u.\sigma_v)_u - \sigma_u.\sigma_{uv})\\ &= \frac{1}{2}\epsilon\gamma E_u + \epsilon\delta(F_u - \frac{1}{2}E_v) \qquad (8.2.10)\end{aligned}$$

And finally,

$$\begin{aligned}\beta &= e'_v.e'' \\ &= (\epsilon_v\sigma_u + \epsilon\sigma_{uv}).(\gamma\sigma_u + \delta\sigma_v) \\ &= \frac{\epsilon_v}{\epsilon}e'.e'' + \epsilon\gamma\sigma_{uv}.\sigma_u + \epsilon\delta\sigma_{uv}.\sigma_v. \\ &= \frac{1}{2}\epsilon\gamma E_v + \frac{1}{2}\epsilon\delta G_u, \end{aligned} \tag{8.2.11}$$

which also depends on E, F and G. This completes the proof of Gausse's Theorema Egregium.

8.3 Gauss-Bonnet Theorem.

In the previous section, we have studied Gausse's Theorema Egrigium which tells us that the Gaussian curvature of a surface depents on the metrici.e. the first fundamental form of the surface. So Gaussian curvature of two isometric surfaces are same. On the other hand Gauss-Bonnet theorem gives the relation between the topological and geometric properties of a surface. In the following we shall prove the Gauss Bonnet theorem for simple closed curves and the only state the Gauss-Bonnet theorem for compact surfaces.

Definition 8.3.1. a curve $\gamma(t) = \sigma(u(t), v(t))$ on a surface patch $\sigma : U \to R^3$ is called a simple closed curve with period $\pi(t) = (u(t), v(t))$ is simple closed curve in R^2 enclosed by π is entirely contained in U. The curve γ is said to be positively oriented if π is positively oriented. Finally, the image of int(π) under the map σ is defined to be the interior int(γ) of γ.

Theorem 8.3.2. Let $\gamma(s)$ be a unit-speed simple closed curve on a surface σ of length $l(\gamma)$, and assume that γ is positively oriented. Then

$$\int_0^{l(\gamma)} k_g ds = 2\pi - \int\int_{int(\gamma)} K dA_\sigma,$$

where k_g is the Gaussian curvature of γ, K is the Gaussian curvature of σ and $dA_\sigma = (EG - F^2)^{\frac{1}{2}} dudv$ is the area element of σ.

Proof. Choose a smooth orthonormal basis $\{e', e''\}$ of the tangent plane of σ at each point such that $\{e', e'', N\}$ is right handed orthonormal basis of R^3, where N is the unit normal to σ. Consider the following integral

$$\begin{aligned} I &= \int_0^l (\gamma) e'.e'' ds \\ &= \int_0^l (\gamma) e'(e''_u \dot{u} + e''_v \dot{v}) ds \\ &= \int_\pi (e'.e'') du + (e'.e'') dv \end{aligned}$$

By Green's Theorem, this can be rewritten as double integral

$$I = \int\int_{int(\pi)} \{(e'.e''_v)_u - (e'.e''_u)v\} dudv.$$

Simplifying the above

$$I = \int\int_{int(\pi)} K dA_\sigma. \tag{8.3.1}$$

Now $\theta(s)$ be the angle between the unit tangent vector $\dot{\gamma}$ of γ at $\gamma(s)$ and the unit vector e'at the same point . More precisely, θ is the angle, uniquely determined up to a multiple of 2π such that

$$\dot{\gamma} = \cos\theta e' + \sin\theta e'' \tag{8.3.2}$$

Then,

$$N \times \dot{\gamma} = -\sin\theta e' + \cos\theta e'' \tag{8.3.3}$$

Now by equation (8.3.2)

$$\ddot{\gamma} = \cos\theta \dot{e}' + \sin\theta \dot{e}'' + \dot{\theta}(-\sin\theta e' + \cos\theta e'') \tag{8.3.4}$$

So by equation (8.3.3) and (8.3.4) the geodesic curvature of γ is $k_g = (N \times \dot{\gamma}).\ddot{\gamma}$ Simplifying we get

$$k_g = \dot{\theta} - e'.\dot{e}''.$$

Hence, by definition

$$I = \int_0^{l(\gamma)} (\dot{\theta} - k_g) ds.$$

Now by Hopf Umlaufstatz theorem $\int_0^{l(\gamma)} \dot{\theta} ds = 2\pi$. Hence the theorem follows by equation (8.3.1).

In the following, the statement of Gauss-Bonnet theorem for compact surfaces is given.

Theorem 8.3.3. Let S be a compact surface. Then

$$\int \int_S K dA = 2\pi\chi,$$

where χ is the Euler number of the surface.

Exercises

1. Show that any point of a surface of constant Gaussian curvature is contained in a patch that is isometric to part of a plane, a sphere or a pseudosphere.
2. If a surface patch has first fundamental form $e^{\lambda}(du^2 + dv^2)$, where λ is a smooth function of u and v, show that its Gaussian curvature K satisfies $\delta\lambda + 2Ke^{\lambda} = 0$.
3. Show that every compact surface whose Gaussian curvature is constant is a sphere.
4. A surface patch σ has Gaussian curvature ≤ 0 everywhere. prove that there are no simple closed geodesic in σ. How do you reconcile this with the fact that the parallels of a circular cylinder are geodesics?

Summary

In this unit we have proved Gausse's Theorema Egrigium and an elementary version of Gauss-Bonnet theorem associated with simple closed curve. Some exercises are given.

Unit 9

Course Structure

- Introduction
- Introduction to manifolds
- Examples of manifolds.
- Exercises.

9.1 Introduction

In the definition of surface, we have seen that for each point $\mathbf{p} \in S$, where, S is a surface, there is an open neighbourhood of $\mathbf{p}$ which is homeomorphic to an open subset of $\mathbb{R}^2$. Also, we have taken the surface S as a subset of $\mathbb{R}^3$.

If instead of taking Sas a subset of $\mathbb{R}^3$, we take S as a topological space and for if every point $\mathbf{p} \in S$, there is a neighbourhood which is homeomorphic to an open set of $\mathbb{R}^n$, then S will be called a topological manifold or simply a manifold.

Like the concept of smooth surface, there is a concept of smooth manifold or differentiable manifold. In the following, we give definition of smooth manifold.

9.1.1 Smooth Manifold

Definition 9.1.1. Let M be a topological space. If each point $p \in M$ has an open neighbourhood which is homeomorphic to an open subset of R^n, then M is called a topological manifold of dimension n. In order to ensure existence of metric on a topological manifold, we assume the topological manifold as Hausdorff and second countable.

Definition 9.1.2. A topological space is called locally Euclidean if each point $p \in M$ has an open neighbourhood which is homeomorphic to an open subset of R^n.

Definition 9.1.3. Let M be a locally Euclidean , Hausdorff, second countable topological manifold. let $p \in M$ has two neighbourhoods U and V which are homeomorphic to two different open subsets of R^n. Let $f(U \cap V) = A$ and $g(U \cap V) = B$. If the transitions maps $gof^{-1} : A \to B$ and $fog^{-1} : B \to A$ are differentiable, then the manifold is called differentiable manifold.

The pairs (U, f) and (V, g) are known as coordinate charts. The collection of all charts of the manifold is called an atlas of the manifold.

Example 9.1.4. A circle is a differentiable manifold of dimension 1. We know that a circle S^1 is a topological space by subspace topology of R^2. Define a map $f : (0, 2\pi) \to S^1$ by $f(t) = (cost, sint)$. Image of f is $S^1 - \{(1, 0)\} = A$ (say). Since the one pointic set $\{(1, 0)\}$ is closed, so A is open.

We see that

f is defined from an open set of R to an open set of S^1.

f is continuous.

f is bijective

$f^{-1}(cost, sint) = t$ is continuous.

Therefore f is a homeomorphism. Define another map $g : (-\pi, \pi) \to S^1$ by $g(t) = (cost, sint)$. The image of g is $S^1 - \{(-1, 0)\} = B$ say. Clearly B is an open set. Like f, g is also a homeomorphism. Hence, for any point of S^1, there exists atleast one open set containing the point which is homeomorphic to an open set of R. So S^1 is a topological manifold. Now we see that Image of f is A and image of g is B. So $A \cap B - S^1 - (1, 0), (-1, 0)$. Now $f^{-1}(A \cap B) = (0, \pi) \cup (\pi, 2\pi)$ and $g^{-1}(A \cap B) = (-\pi, 0) \cup (0, \pi)$. Hence the transition maps are $g^{-1}of$ and $f^{-1}og$. We see that $g^{-1}of$ is defined from $(0, \pi) \cup (\pi, 2\pi)$ to $(-\pi, 0) \cup (0, \pi)$ and $f^{-1}og$ is defined from $(-\pi, 0) \cup (0, \pi)$ to $(0, \pi) \cup (\pi, 2\pi)$.

See that

$$g^{-1}of(t) = g^{-1}f(t) = \begin{cases} t, & t \in (0, \pi) \\ t + 2\pi, & t \in (\pi, 2\pi) \end{cases}$$

And

$$f^{-1}og(t) = f^{-1}g(t) = \begin{cases} t, & t \in (0, \pi) \\ t + 2\pi, & t \in (-\pi, 0) \end{cases}$$

It is clear that the above maps are differentiable. Therefore circle is a differentiable manifold of dimension 1.

Example 9.1.5. A sphere is a differentiable manifold of dimension 2:

We know that the sphere S^2 is a Hausdorff, 2nd countable topological space by subspace topology of R^3. Let us define a map $f : R^2 \to S^2$ by

$$f(u, v) = \Big(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\Big)$$

We see that Image of f is $S^2 - \{(0, 0, 1)\}$ We see that f is continuous, f is bijective. We find the the inverse of f is given by

$$f^{-1}(x, y, z) = \Big(\frac{x}{1 - z}, \frac{y}{1 - z}\Big).$$

Let $N = (0, 0, 1)$ and $S = (0, 0, -1)$. Since both $S^2 - N$ and R^2 are open, so f is continuous. So f is homeomorphism. Let us define another map $g : R^2 \to S^2$ given by

$$g(u, v) = \Big(\frac{2u}{u^2 + v^2 + 1}, -\frac{2v}{u^2 + v^2 + 1}.$$

So as before, g is continious and image of g $S^2 - S$. Like, f, g is also a homeomorphism.

See that Image of $f \cap$ Image of g = $S^2 - \{N, S\}$. Denote Image of $f \cap$ Image of g by A. So

$$f^{-1}(A) = R^2 - \{(0,0)\}$$

and

$$g^{-1}(A) = R^2 - \{(0,0)\}.$$

Here the transition maps are $g^{-1}of$ and $f^{-1}og$ are defined from $R^2 - \{(0,0)\}$ to $R^2 - \{(0,0)\}$ by

$$f^{-1}og(u,v) = \Big(\frac{u}{u^2+v^2}, \frac{v}{u^2+v^2}\Big).$$

Hence S^2 is a differentiable manifold of dimension 2.

Example 9.1.6. The open interval is a differentiable manifold of dimension 1. Any open set of R^2 is a differentiable manifold of R^2. Any n-dimensional vector space is a differentiable manifold of dimension n.

Example 9.1.7. Let $P_3(R)$ be the set of all polynomials of degree less than three. Then it is a vector space of dimension three. So it is a differentiable manifold of dimension three.

Exercises

1. Show that a circle is a smooth manifold.
2. Show that a sphere is a smooth manifold.
3. Show that the set of all $n \times n$ matrices is a smooth manifold.
4. Show that the set of all polynomials with real coefficient and degree less than three is a smooth manifold.
5. Show that an open interval of $\mathbb{R}$ is a smooth manifold.
6. Show that $\mathbb{R}^n$ is a smooth manifold.

Unit 10

Course Structure

- Introduction
- Introduction to manifolds
- Examples of manifolds.
- Exercises.

10.1 Introduction

We all are familiar with tangent vector to a curve at any point of plane. The idea of tangent vecto to a curve is generalized as tangent plane of any surface at any point on the surface. This is further generalized as tangent space on arbitrary manifolds. In this unit se shall study tangent space on a manifolds. In the following, let us first have the idea of tangent vectors on manifolds.

Definition 10.1.1. Let $\gamma : (-\epsilon, \epsilon) \to R^n$ be a smooth curve. Let $\gamma(0) = (p_1, p_2, ..., p_n) := p$ By tangent vector on R^n at p, we mean $\gamma'(t)|_{t=0}$.

Example 10.1.2. Let $f : R^n \to R$ be a differentiable function. let $\alpha = (\alpha_1, ..., \alpha_n)$ be a given point on R^n. Show that there exists a curve $\gamma : (\epsilon, \epsilon) \to R^n$ passing through α such that

$$D_v f(\alpha) = \frac{d}{dt}(fo\gamma(t))|_{t=0}$$

Also show that $D_v f(\alpha) = Df.v$, where $v = \gamma'(t)|_{t=0}$.

Soln. Here $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$. By definition of directional derivative at α along a vector $v = (v_1, v_2, .., v_n)$, we get

$$\begin{aligned} D_v f(\alpha) &= \lim_{t\to 0} \frac{f(\alpha + tv) - f(\alpha)}{t} \\ &= \lim_{t\to 0} \frac{f(\alpha_1 + tv_1, \alpha_2 + tv_2, ..., \alpha_n + tv_n) - f(\alpha_1, .., \alpha_n)}{t} \end{aligned} \tag{10.1.1}$$

Define the curve $\gamma : (-\epsilon, \epsilon) \to R^n$ by $\gamma(t) = \alpha + tv$ From 10.1.1, we have

$$\begin{aligned} D_v f(\alpha) &= \lim_{t\to 0} \frac{f(\gamma(t)) - f(\gamma(0))}{t} \\ &= \frac{d}{dt}\{f(\gamma(t))\}|_{t=0} \\ &= \frac{\partial f}{\partial \alpha_1}\frac{d}{dt}(\gamma_1(t)) + ... + \frac{\partial f}{\partial n}\frac{d}{dt}(\gamma_n(t)) \\ &= \frac{\partial f}{\partial \alpha_1}v_1 + ... + \frac{\partial f}{\partial n}v_n \\ &= (\frac{\partial f}{\partial \alpha_1}, ..., \frac{\partial f}{\partial n})(v_1, v_2, ..., v_n). \\ &= Df.v \end{aligned}$$

Note: From the relation $D_v f(\alpha) = \frac{d}{dt}\{f(\gamma(t))\}|_{t=0}$ and the definition of tangent vectors in R^n we see that each directional derivative operator assigns a tangent vector to R^n. Hence we can identify directional derivative operators and tangent vectors in R^n.

Definition 10.1.3. Let p be a point of R^n. The set of all tangent vectors passing through p is called tangent of R^n at p and it is denoted by T_pR^n.

Example 10.1.4. Show that tangent space at any point of R^n is R^n itself.

Soln. Let $P = (p_1, p_2, .., p_n)$ be a given point of R^n. We have to show that $T_pR^n = R^n$. Let $v = (v_1, v_2, ..., v_n)$ be any arbitrary vector of R^n. Define a curve, $\gamma(t) = p + tv$ where $\gamma(0) = p$. Therefore γ passes through p.

$$\gamma(t) = (p_1 + tv_1, ..., p_n + tv_n)$$

$$\gamma'(t) = (v_1, v_2, .., v_n).$$

So $\gamma'(t)|_{t=0} = (v_1, v_2, .., v_n)$. Therefore any vector of R^n is tangent vector of R^n. Hence $T_pR^n = R^n$.

Example 10.1.5. Find the tangent space of $GL_2(R)$ at I_2.

soln. Let us first show that, if X is a 2×2 matrix such that $||X|| < 1$, then $I_2 + X \in GL_2(R)$. Let $A = I_2 + X$. Let us consider the matrix $B = I_2 - X + X^2 - X^3 + ...$ Now,

$$S_m = I_2 - X + X^2 - X^3 + ... + (-1)^{m-1}X^{m-1}$$

$$S_{m+k} = I_2 - X + X^2 - .... + (-1)^{m-1}X^{m-1} + ... + (-1)^{m+k-1}X^{m+k-1}$$

So

$$\begin{aligned} S_{m+k} - S_m &= (-1)^m X^m + (-1)^{m+1}X^{m+1} + ... + (-1)^{m+k-1}. \\ &= (-1)^m X^m (I_2 - X + X^2 - ... - +(-1)^{k-1}X^{k-1}) \end{aligned}$$

So

$$\begin{aligned} ||S_{m+k} - S_m|| &= ||X^m||||I_2 - X + X^2 - -... + (-1)^{k-1}X^{k-1}|| \\ &\leq ||X^m||(1 + ||X|| + ||X^2|| + ... + ||X^{k-1}||) \\ &\leq ||X^m||(1 + ||X|| + ||X^2|| + ... + ||X^{k-1}||) \\ &\to 0 \text{ as } ||X|| < 1 \text{ as} m \to \infty. \end{aligned}$$

Hence B exists. Now

$$\begin{aligned} AB &= (I_2 + X)(I_2 - X + X^2 - X^3 + ...) \\ &= (I_2 + X)(I_2 - X)^{-1} \\ &= I_2. \end{aligned}$$

So A is invertible, hence $A \in GL_2(R)$. Let M be any matrix of $M_2(R)$, $t \in (-\epsilon, \epsilon)$, $\epsilon \to 0$. So $||tM|| < 1$. So $I_2 + tM \in GL_2(R)$. Now define a curve $\gamma(t) = I_2 + tM$, $t \in (-\epsilon, \epsilon)$, $M \in M_2(R)$. For $t = t_0 = 0$, $\gamma(t_0) = I_2$. So γ passes through I_2. Now $\gamma'(t) = M$. So $\gamma'(t)|_{t=0} + M$. So tangent space of $GL_2(R)$ at I_2 is $M_2(R)$.

Definition 10.1.6. We have seen that directional derivative operators are identified with tangent vectors. Hence we shall call directional derivatives are tangent vectors of R^n. Generalizing this property we shall call any operator which satisfies properties of directional derivatives are tangent to any abstract manifolds.

Exercises

1. Let M be a differentiable manifold. If $p \in M$ has local coordinates $x_1, x_2, .., x_n$, show that $\{\frac{\partial}{\partial x_1}, ..., \frac{\partial}{\partial x_n}\}$ is a basis of T_pM.

2. Let M be a differentiable manifold. if $p \in M$ and g, h are two smooth functions defined from a neighbourhood of p to R and if γ is a smooth curve passing through p show that

$$\gamma'(0)(ag + bh) = a\gamma'(0)g + b\gamma'(0)h,$$

   where a, b are real numbers.

3. Find the tangent space of unit sphere at the point $(0, 0, 1)$.

Unit 11

Course Structure

- Introduction
- Objectives
- Connectedness : Examples, various characterizations and basic properties
- Connectedness on the real line

11.1 Introduction

The intermediate value theorem is a very important property of the continuous functions in the real line. It says that if a function : $[a, b] \to \mathbb{R}$ is continuous, and if r is a real number between $f(a)$ and $f(b)$, then there exists an element $c \in [a, b]$ such that $f(c) = r$. Can you see the geometric interpretation of this theorem? What if we define a continuous function on the interval $[1, 2] \cup [3, 4]$, then is the theorem true for this domain also? You will notice that its not. The intermediate value theorem is true for $[a, b]$ due to a particular property of the interval which is rightly termed as "connectedness" of the interval. As the name suggests, the interval is not "broken" or "separated" in some sense. In this unit, we will learn the concept of connectedness and its properties.

Objectives

After reading this unit, you will be able to

- define connected sets and differentiate between connected and separated sets.
- learn various equivalent definitions for connected spaces.
- learn the characteristics of connected sets.
- know the structure of connected sets in the real line.
- learn examples of connected and separated sets.

11.1.1 Connected Spaces

Roughly speaking, a topological space is said to be connected if it does not allow itself to be partitioned into two disjoint proper open subsets of itself. In an arbitrary topological space, we have we have only the open sets at our disposal. So we first define separated set as follows:

Definition 11.1.1. Let (X, τ) be a topological space. Then, X is said to be disconnected if there exists two open sets U and V, such that

1. $U \cap V = \phi$.
2. $U \cup V = X$.

Then the pair (U, V) is said to form a disconnection of X.

Definition 11.1.2. A topological space X is said to be connected if it does not admit of a disconnection.

Does the closed interval $[a, b]$ admit of a disconnection?

Connectedness is obviously a topological property since its defined in terms of open sets entirely. In other terms, we can say that if X is connected,then any space homeomorphic to X is also connected. In fact, if X and Y are homeomorphic, then there exists a homeomorphism $f : X \to Y$. If X is not connected, then it does not have a disconnection. If possible, let Y is disconnected. Then there exists disjoint open sets, say U and V. Since f is continuous, so $f^{-1}(U)$ and $f^{-1}(V)$ are open sets in X that forms a disconnection for X (verify!). Which is a contradiction. Hence Y is also connected.

Consider a subset A of a topological space X, which is both open and closed. We know that in any topological space X, the empty set $\emptyset$ and X are always both open and closed. Does there exist any other subset in X with the property? Consider the following example:

Example 11.1.3. The set $Y = [0, 1] \cup [3, 4]$ is a subspace of the real space with the usual topology. Then, the sets $[0, 1] = Y \cap (-1/2, 3/2)$ is open in Y. Similarly, $[3, 4] = Y \cap (3/2, 9/2)$ is open in Y. Also, see that $[0, 1] = Y \setminus [3, 4]$ and $[3, 4] = Y \setminus [0, 1]$. Hence, both are closed also. Thus the subspace Y has two proper subsets which are both open and closed.

As you can probably figure out, the subspace Y is not connected (the sets $[0, 1]$ and $[3, 4]$ forms a disconnection for Y). In fact, for connected space, we won't find any other proper subset which is both open and closed. We have the following equivalent definition for connectedness of a space X.

Theorem 11.1.4. A space X is said to be connected if and only if X and $\emptyset$ are the only subsets of X which are both open and closed.

Proof. Let X be connected and if possible, let A be a subset of X which is both open and closed. Since A is closed, $X \setminus A$ is open. Also, A and $X \setminus A$ are disjoint open sets whose union is X. Thus A and $X \setminus A$ forms a disconnection of X, a contradiction. Thus no proper subset of X can be both open and closed.
Conversely, suppose X is disconnected. Let U and V forms a disconnection of X. Then $X = U \cup V$. Which means, $U = X \setminus V$ and $V = X \setminus U$ implying that U and V are both open and closed. Hence the theorem. □

Example 11.1.5.
1. Let X be a two-point set with the indiscrete topology. Then X has no separation. In fact, any set with the indiscrete topology is always connected. But what happens if we take the discrete topology?

2. Consider $\mathbb{R}$ with co-finite topology. Then $\mathbb{R}$ is connected. If possible, let A and B be a separation of $\mathbb{R}$. Then $A \cap B = \emptyset$ and $\mathbb{R} \setminus A$ and $\mathbb{R} \setminus B$ are both finite sets. So, $(\mathbb{R} \setminus A) \cup (\mathbb{R} \setminus B)$ is a finite set. That is, $\mathbb{R} \setminus (A \cap B)$ is a finite set, that is, $\mathbb{R}$ is a finite set, since $A \cap B = \emptyset$. Thus, $\mathbb{R}$ is connected with the co-finite topology.

3. $\mathbb{R}$ with the usual topology is also connected.(can you prove it?)

Definition 11.1.6. Let A and B be two subsets of a topological space (X, τ). Then A and B are said to be separated if $\overline{A} \cap B = \emptyset$ and $A \cap \overline{B} = \emptyset$ simultaneously hold.

Definition 11.1.7. Let Y be a subspace of a topological space (X, τ). Then Y is said to be connected if it is connected with respect to the topology induced by τ, that is, if there does not exist disjoint open sets A and B in Y such that $Y = A \cup B$.

Is a subspace of a connected space always connected? Let's consider the following example:

Example 11.1.8. The subspace of all rationals is disconnected with the usual topology. Let $x, y \in \mathbb{Q}$ such that $x < y$. Then, there exists an irrational number a such that $x < a < y$. Then we can write

$$\mathbb{Q} = \{(-\infty, a) \cap \mathbb{Q})\} \cup \{(a, \infty) \cap \mathbb{Q}\}$$

The sets $\{(-\infty, a) \cap \mathbb{Q})\}$ and $\{(a, \infty) \cap \mathbb{Q}\}$ are non-empty since $x \in \{(-\infty, a) \cap \mathbb{Q})\}$ and $y \in \{(a, \infty) \cap \mathbb{Q}\}$ and open in the subspace topology. Thus $\mathbb{Q}$ is disconnected. Can you give other examples of disconnected subspaces of a connected topological space?

Note that, by choosing different $a \in \mathbb{Q}$, we can get a different disconnection for $\mathbb{Q}$. In fact, in this way, we can get infinite number of disconnections for $\mathbb{Q}$. We could similarly show that the set of irrationals is also disconnected. Also we see that the union of rationals and irrationals give us $\mathbb{R}$. Thus, we can conclude that two the union of two disconnected sets may be connected. Some additional properties are required for the union to be disconnected too. We have the following theorem in this direction:

Theorem 11.1.9. If A and B are two separated non-empty sets of (X, τ), then $A \cup B$ is disconnected.

Proof. Let A and B are two separated non-empty sets of (X, τ). Then we have $\overline{A} \cap B = \emptyset$ and $A \cap \overline{B} = \emptyset$. Put $G = X \setminus \overline{B}$ and $H = X \setminus \overline{A}$. Then G and H are disjoint non-empty sets in X. Then we have

$$\begin{aligned} (A \cup B) \cap G) &= (A \cap G) \cup (A \cap G) \\ &= A \cup \emptyset \\ &= A \end{aligned}$$

Similarly, we have

$$(A \cup B) \cap H = B$$

Thus,

$$A \cup B = ((A \cup B) \cap G) \cup ((A \cup B) \cap H)$$

Hence, $(A \cup B) \cap G$ and $(A \cup B) \cap H$ clearly forms a disconnection for $A \cup B$. Hence the result. □

Theorem 11.1.10. A subset Y of (X, τ) is disconnected if and only if Y is the union of two non-empty separated sets.

Proof. The sufficient condition immediately follows from the previous theorem.

For the necessary part, let Y be disconnected. Then Y can be partitioned as

$$Y = (Y \cap A) \cup (Y \cap B)$$

where A and B are open sets in X whose intersections with Y are non-empty. We will check for the separatedness of $Y \cap A$ and $Y \cap B$. If possible, let a be a limit point of $Y \cap A$ and $a \in Y \cap B$. Since B is an open set containing a, so

$$B \cap \{(Y \cap A) \setminus \{a\}\} \neq \emptyset$$

which is a contradiction since

$$B \cap (Y \cap A) = (Y \cap A) \cap (Y \cap B) = \emptyset.$$

Hence $Y \cap A$ and $Y \cap B$ are separated sets. □

Example 11.1.11.

1. We told that the set $Y = [0, 1] \cup [3, 4]$ is disconnected. But in light of the previous theorem, we can see that $[0, 1]$ and $[3, 4]$ forms a disconnection for Y in which neither of them contains a limit point of the other.
2. The set $Y = [-1, 0) \cup (0, 1]$ of the real line is disconnected since $[-1, 0)$ and $(0, 1]$ are separated sets whose union is Y (neither of them contains a limit point of the other).
3. The set $Y = [0, 2]$ of the real line is connected. The sets $[0, 1)$ and $[1, 2]$ are of course disjoint whose union gives Y. But $[1, 2]$ is not open in the subspace Y. So this does not form a disconnection for Y. In fact, by the previous theorem, we can say that since $[0, 1)$ and $[1, 2]$ are not separated, so it does not form a disconnection for Y.

Now we will see certain results for connectedness in series.

Lemma 11.1.12. If the sets C and D form a disconnection for X and Y is a connected subspace of X, then Y is either in C or D.

Proof. Since C and D are disjoint open sets in X, the sets $C \cap Y$ and $D \cap Y$ are open in Y. These two sets are disjoint and their union is Y. If they were both non-empty, they would constitute a separation for Y. Hence, one of them should be empty. Hence Y entirely lies either in C or D. □

Theorem 11.1.13. The union of a collection of connected subspaces of X that have a point in common is connected.

Proof. Let $\{A_\alpha\}$ be a collection of connected subspaces of a space X and let $\cap_\alpha A_\alpha = p$. We prove that the space $Y = \cup_\alpha A_\alpha$ is connected. If possible, let C and D form a separation for Y. Then p lies in either C or D. Without any loss of generality, let us assume that $p \in C$. For each α, A_α must lie entirely on C or D since it is connected and it can't be in D since $p \in C$. Since all $A_\alpha \in C$, so $\cup_\alpha A_\alpha \subset C$ and hence $Y \cap D = \emptyset$. We arrive at a contradiction. Hence Y is connected. □

Theorem 11.1.14. Let A be a connected subspace of X and if $A \subset B \subset \overline{A}$, then B is connected.

Proof. Let A be connected and let $A \subset B \subset \overline{A}$. Suppose that $B = C \cup D$ be a separation for B. Then, A must lie entirely in either C or D. Suppose $A \subset C$. Then $\overline{A} \subset \overline{C}$. Since $\overline{C}$ and D are disjoint, B can't intersect D. This contradicts the fact that D is a non-empty subset of B. □

Corollary 11.1.15. Closure of a connected set is connected in a Topological space

Theorem 11.1.16. The continuous image of a connected space is connected.

Proof. Let $f : X \to Y$ be a continuous function and let X be connected. We show that the space $Z = f(X)$ is connected. If not, then there exists a separation $Z = C \cup D$ of Z. Since f is continuous, so $f^{-1}(C)$ and $f^{-1}(D)$ are both disjoint open sets in X, whose union is X. Then they form a disconnection for X, a contradiction. Hence $Z = f(X)$ has to be connected. □

Let us consider a topological space consisting of two members $\{a, b\}$ with the discrete topology, $a, b \in \mathbb{R}$. Then the discrete two-point space is disconnected since $\{a, b\} = \{a\} \cup \{b\}$ is clearly a disconnection for the space. One can characterize a connected space as follows.

Theorem 11.1.17. A space X is disconnected if and only if there is a continuous function $f : X \to \{a, b\}$, which is onto.

Proof. Left as an exercise. □

Theorem 11.1.18. A topological space (X, τ) is connected if and only if given any two distinct points in X, there is a connected subspace of X containing both.

Proof. Let (X, τ) be connected. Given any two distinct points in X, then X itself is the required subspace containing both.

Conversely, let the given condition holds. If possible, let X is disconnected. Then there exist a disconnection $C,\ D$ for X such that $X = C \cup D$. Let $c \in C$ and $d \in D$. Since C and D are disjoint, so $c \neq d$. By the hypothesis, there exists a connected subspace of X, say G that contains both c and d. Clearly, $G \subset C$ or $G \subset D$. Let $G \subset C$. Then we get $d \in (C \cap D) = \emptyset$, which is absurd. Hence, X has to be connected. □

11.1.2 Connected Sets on the Real line

We will now deal with the connected sets in the real line with the usual topology. With all the examples that we have so far seen, you must have got an idea about the connected sets on the real line. Yes. They are intervals: $(a, b),\ [a, b],\ [a, b),\ (a, b],\ (a, \infty),\ [a, \infty),\ (-\infty, b)$, or $(-\infty, b]$, whatever may be the form. The following theorem guarantees the fact.

Theorem 11.1.19. A subset of $\mathbb{R}$ is connected if and only if it is an interval.

Proof. Suppose E is a connected subset of $\mathbb{R}$ without being an interval. Then we find a pair of distinct members $a, b \in E$ such that $[a, b] \not\subset E$. Thus, there is a member u such that $a < u < b$ and $u \notin E$. Then we write $E = ((-\infty, u) \cap E) \cup ((u, \infty) \cap E)$ and that is a disconnection for E, a contradiction. Hence E has to be an interval.

Conversely suppose that I be an interval of the reals. If possible, let I has a disconnection

$$I = A \cup B,$$

where, A and B are a pair of non-empty disjoint closed sets in I. Take $x \in A$ and $z \in B$. Since $A \cap B = \emptyset$, $x \neq z$, and without any loss of generality, assume that $x < z$. Because I is an interval we have the closed interval $[x, z] \subset I$. Thus

$$[x, z] \subset A \cup B$$

Put $y = \sup([x, z] \cap A)$. Then $x \leq y \leq z$, so $y \in I$. Since A is closed in I, we have,

$$y \in A \tag{11.1.1}$$

Therefore $y \neq z$ and we have $y < z$. By the property of supremum, for large natural numbers n, all numbers $y + \frac{1}{n}$ belong to B and since B is closed, taking limit as n tends to infinity, we have

$$\lim_{n\to\infty} \left(y + \frac{1}{n} \right) = y \in B$$

But this contradicts (11.1.1) because $A \cap B = \emptyset$. Hence proved. □

We can say from the theorem that $\mathbb{R}$ is connected.

Exercise 11.1.20.

1. Let τ and σ be two topologies on X. If $\sigma \supset \tau$, what does connectedness of one topology imply about the connectedness of the other?
2. Let $\{A_n\}$ be a sequence of connected subspaces of X, such that $A_n \cap A_{n+1} \neq \emptyset$ for all n. Show that $\bigcup_n A_n$ is connected.
3. Let $\{A_\alpha\}$ be a collection of connected subspaces of X and let A be a connected subspace X. If $A \cap A_\alpha \neq \emptyset$ for all α, then show that $A \cup (\bigcup_\alpha A_\alpha)$ is connected.

Unit 12

Course Structure

- Introduction
- Objectives
- Components and quasi components
- Path connectedness and path components

Introduction

Given any arbitrary space X, there is a natural way to break it into pieces as we shall see in this unit. Those pieces are called components. We will study various components of a topological space and further characterisations of connectedness of a topological space. Let us start with the idea of components of a topological space.

12.1 Components

Definition 12.1.1. Given a topological space X, define an equivalence relation on X as follows

$$x \sim y \text{ if and only if there is a connected subspace containing both of them.}$$

The equivalence classes are called the components of X.

The proof that '$\sim$' is an equivalence relation has been left as en exercise.

Theorem 12.1.2. The components of X are connected disjoint subspaces of X whose union in X; such that every non-empty connected subspace of X is contained in exactly one of them.

The statement of the theorem says that the components of X partition X into disjoint sets. The proof is as follows

Proof. Since the components are equivalence classes, they are disjoint and their union is X. For the second part, let A be a connected subspace of X. If possible, let A intersects tow of the components, say C and D at points, say c and d respectively. Then by the definition of $\sim$, we can say that $c \sim d$ and A is the connected subspace containing both. This can't be true unless $C = D$.

To show that a component C is connected, choose a point $x_0 \in C$. For each point $x \in C$, we know that $x_0 \sim x$, so there is a connected subspace A_x of X containing x_0 and x. By the result just proved, $A_x \subset C$. Hence,

$$C = \bigcup_{x \in C} A_x.$$

Since the subspace C is the union of connected subspaces containing a common point x_0, their union is connected. □

Theorem 12.1.3. A component is a maximal connected subspace of X, that is, it is not contained in a connected subspace of X.

Proof. If possible let a component C be properly contained in a connected subspace D of X. Then there exists a $d \in D$ such that $d \notin C$. Since X is partitioned into components, so d is contained in some other component,say E of X. Now, let $c \in C$. Since $C \subset D$, so $c \in D$. Thus, $c \sim d$ and D is the connected subspace of X containing both c and d. Also, D intersects both the components C and E which is a contradiction by the previous theorem. Hence C can't be contained in any connected subspace of X. □

Theorem 12.1.4. A connected set in X that is both open and closed is a component of X.

Proof. Let E be a connected set in X which is both open and closed. Now, since each connected set is contained in a component of X, so $E \subset C$ for some component C of X. Suppose E is a proper subset of C. Then we can write

$$C = (C \cap E) \cup (C \cap (X \setminus E)).$$

Since E is both open and closed, so the above decomposition yields a disconnection for C, a contradiction. Hence, E can't be a proper subset of C and so $E = C$. □

Theorem 12.1.5. Every component of X is a closed set.

Proof. Let Cbe a component of X without being closed. So $\overline{C}$ is strictly contains C. But, we know that the closure of connected set is connected and thus it contradicts the maximality of C. Hence C is closed. □

Definition 12.1.6. A topological space (X, τ) is said to be totally disconnected if for every pair of disjoint points $x, y \in X$, X has a disconnection $X = G \cup H$ with $x \in G$ and $y \in H$.

Does the above definition remind you of some similar definition that you have learnt earlier?

Of course a totally disconnected space is always disconnected.

Example 12.1.7. The real numbers with the upper limit topology is totally disconnected.

We know that the upper limit topology for $\mathbb{R}$ is generated by the left-open intervals like $(a, b]$ for $a, b \in \mathbb{R}$ with $a < b$. Let $x, y \in \mathbb{R}$ with $x < y$. Then we can write

$$\mathbb{R} = (-\infty, x] \cup (x, \infty)$$

where the sets in the r.h.s of the above equation are clearly disjoint open sets in the upper limit topology containing x and y respectively. So, $\mathbb{R}$ is totally disconnected.

Theorem 12.1.8. The components for a totally disconnected space are its singletons.

Proof. Let X be totally disconnected and C be a component of X. We show that C can't have more than one point. Let $x, y \in C$ with $x \neq y$. Since X is totally disconnected, it has a disconnection as

$$X = G \cup H$$

where G and H are disjoint non-empty open sets such that $x \in G$ and $y \in H$. We write

$$\begin{aligned} C &= C \cap X \\ &= C \cap (G \cup H) \\ &= (C \cap G) \cup ((C \cap H) \end{aligned}$$

each of which are disjoint open sets in C thus yielding a disconnection for C,a contradiction. Hence the theorem. □

Definition 12.1.9. A topological space X is said to be locally connected at $x \in X$, if every neighbourhood of x contains an open connected neighbourhood of x. And X is said to be locally connected if it is locally connected at each of its point.

It is interesting to note that neither local-connectedness implies connectedness nor conversely.

Example 12.1.10. If $X = (0, 1) \cup (2, 3)$ be a space with usual topology, then it is locally connected without being connected. Take a real number a with $1 < a < 2$. We write

$$X = ((-\infty, a) \cap X) \cup ((a, \infty) \cap X),$$

and this forms a disconnection for X. Also let $u \in X$ and let $0 < u < 1$, and given a neighbourhood N_u of u in X, we can find an open interval like $(u-\delta, u+\delta)$, $\delta > 0$ such that $(u-\delta, u+\delta) \subset N_u$. Since open intervals of reals with the usual topology are connected, so N_u contains an open neighbourhood of u, and hence X is locally connected at u. Since u is arbitrary, so X is locally connected at each point of $(0, 1)$. By similar argument, we can show that X is locally connected at each point of $(2, 3)$. Hence X is locally-connected without being connected.

Can you think of a space which is connected without being locally-connected?

Theorem 12.1.11. A topological space X is locally-connected if and only if the components of each open subspace of X are open in X.

Proof. Let X be locally-connected and let Y be an open subspace of X. Suppose C is a component of Y. Take $x \in C$ Since X is locally-connected at x, there is an open connected set U in X such that $x \in U \subset Y$. Now, $x \in C \cap U$, where U and C are connected and hence $C \cup U$ is connected and $C \cup U \subset Y$. Since C is a component, by maximality of C, we have $C \cup U = C$ or $U \subset C$. Hence, $x \in U \subset C$. Since x is arbitrary, so we conclude that C is open.

Conversely let the given condition holds. Let $x \in X$ and let N_x be an open neighbourhood of x in X. Take C as a component such that $x \in C \subset N_x$. By hypothesis, C is open. This shows that there is an open connected neighbourhood C of x such that $C \subset N_x$. Thus X is locally-connected at x. Since x is arbitrary, so X is locally-connected. □

12.1.1 Path Connectedness

Let us first define path in a topological space.

Definition 12.1.12. Let X be a topological space and let $x, y \in X$. A path in X from x to y is a continuous map $\gamma : [a, b] \to X$ satisfying $\gamma(a) = x$ and $\gamma(b) = y$. Here, $a, b \in \mathbb{R}$ and $a < b$.

Definition 12.1.13. We define another equivalence relation on X as follows:

$$x \sim y \text{ if and only if there is a path between the two.}$$

Hence the equivalence classes for '$\sim$' partitions X, and they are called the path-components of X.

The proof of the equivalence of the relation is easy and has been left as an exercise.

Definition 12.1.14. A topological space X is path-connected if for any $x, y \in X$, there exists a path from x to y.

Theorem 12.1.15. Let X be a path-connected space. Then X is connected.

Proof. Let X be path-connected. We will use paths in X to show that if X is not connected then $[0, 1]$ is not connected, which of course is a contradiction, so X has to be connected.

Suppose X is not connected, so we can write

$$X = U \cup V,$$

where U and V are non-empty disjoint open sets of X. Let $x \in U$ and $y \in V$. Then, there is a path $\gamma : [0, 1] \to X$ such that $\gamma(0) = x$ and $\gamma(1) = y$. The partition of X into U and V leads via this path to a partition of $[0, 1]$.

$$[0, 1] = A \cup B,$$

where, $A = \gamma^{-1}(U)$ and $B = \gamma^{-1}(V)$. Note that $0 \in A$ and $1 \in B$, so A and B are non-empty subsets of $[0, 1]$. Obviously A and B are disjoint, since no point in $[0, 1]$ can have a common γ value in both U and V. Since γ is continuous and U and V are both open in X, A and B are open in $[0, 1]$. Thus we get a disconnection for $[0, 1]$, a contradiction. Hence, X has to be connected. □

The converse does not hold in general.

Example 12.1.16. The space

$$A = \{(x, y) \in \mathbb{R}^2 | x > 0,\ y = \sin \frac{1}{x}\} \subset \mathbb{R}^2$$

is path-connected and hence connected. So, its closure

$$\overline{A} = A \cup (\{0\} \times [-1, 1])$$

is connected. But $\overline{A}$ is not path-connected.

Though the converse is sometimes true in special cases. We have the following theorem in this direction.

Theorem 12.1.17. If a non-empty open subset of $\mathbb{R}^n$ is connected, then it is path-connected.

We will omit the proof of the theorem.

Theorem 12.1.18. The path components of a topological space X are path connected disjoint subspaces of X whose union is X, such that every non-empty path-connected subspace intersects only one of them.

Proof. Since the path components are equivalence classes of the equivalence relation for paths, so each of the components are disjoint subspaces whose union is X. Let A be a non-empty path-connected subspace of X. If possible, let A intersects two path-components P and S of X, say at points p and s respectively. Then, since A is path-connected, and s and p are in A, so there exists a path between s and p. But by the definition of path-components, this is impossible unless $P = S$. Thus A can intersect only one of the path-components.

Let P be a path-component of X. Then by the definition of path components, for any two distinct points of P, there exists a path between the two. Then it is obviously path-connected. □

We saw in case of components that they are closed. But they need not be open. But in case of path-components, we can say nothing about this. They need be neither open nor closed.

Example 12.1.19. In the subspace of rationals $\mathbb{Q}$ of $\mathbb{R}$, each component consists of a single point and hence is not open in $\mathbb{Q}$.

Definition 12.1.20. A space X is said to be locally path-connected at a point $x \in X$ if for every neighbourhood U of x, there exists a path-connected neighbourhood V of x contained in U. If X is locally path-connected at each of its points, then it is called locally path-connected.

Theorem 12.1.21. A space X is locally path-connected if and only if for every open set U of X, each path component of U is open in X.

Proof. The proof is similar as that in the case of local connectedness and has been left as an exercise. □

Theorem 12.1.22. If X is a topological space, each path component of X lies in a component of X.

Proof. Let C be a component of X and let $x \in C$. Let P be the path component of X containing x. Since P is connected, so by maximality of C, $P \subset C$. Hence proved. □

Theorem 12.1.23. If X is locally path-connected, then the components and path-components are the same.

Proof. Let C be a component and P be a path-component as in the previous theorem. Then by the previous theorem , we have

$$P \subset C.$$

If possible, let $P \neq C$. Let Q denotes the union of all path-components of X that are different from P and intersect C; each of them necessarily lie in C, so that

$$C = P \cup Q.$$

Since X is locally path-connected, each path-component of X is open in X. Hence, P and Q (which is the union of path-components) are open in X, so they constitute a disconnection of C, a contradiction. Thus P has to be equal to C. Hence the theorem. □

12.1.2 Quasicomponents

Definition 12.1.24. Let X be a space. We define a relation '$\sim$' as follows: $x \sim y$ if and only if there is no separation

$$X = A \cup B,$$

where A and B are disjoint open sets in X such that $x \in A$ and $y \in B$. Then it is an equivalence relation on X and the equivalence classes are called quasicomponents of X.

Theorem 12.1.25. In a space X, each component lies in a quasicomponent of X.

Proof. If X is connected then the result is trivially true. Let X be not connected and let C be a component of X. Let $c \in C$ and also let Q be the quasicomponent of X containing c. We show that $C \subset Q$. If not, then there exists at least a pair of elements $x, y \in C$ for which there exists a disconnection

$$X = A \cup B$$

such that $x \in A$ and $y \in B$. Then, we will get

$$\begin{aligned} C &= (C \cap X) \\ &= C \cap (A \cup B) \\ &= (C \cap A) \cup (C \cap B) \end{aligned}$$

which yields a disconnection for C, a contradiction. Hence our assumption is wrong and C has to be in Q. □

Exercise 12.1.26.

1. Examine if the real number space with the lower limit topology is connected.
2. Show that the set of irrational numbers with the usual topology is disconnected.
3. Show that a totally disconnected space is T_2.
4. Show that continuous image of a locally connected space may not be so.
5. Show that any ball in $\mathbb{R}^2$ with usual topology is connected.

Unit 13

Course Structure

- Introduction
- Objectives
- Compactness
- Sequential Compactness

Introduction

As we have seen, the closed interval $[a, b]$ has certain crucial properties which has many applications such as the maximum value theorems. But for a long time it was not clear how to formulate it in any arbitrary topological space. Mathematicians formulated it in terms of open coverings of the space. This formulation is what is called compactness. We do have preliminary ideas about compactness when we learned real analysis and metric spaces. We are very familiar with the Heine-Borel theorem which says that any closed and bounded set is compact in the real line $\mathbb{R}$. In fact, the theorem is true for any $\mathbb{R}^n$. By virtue of this, it can be said that $[a, b]$ is compact in $\mathbb{R}$. Let's look at it in a more general way, that is, for any arbitrary topological space and lets find out whether we can connect with whatever we have learned earlier.

Objectives

After reading this unit, you will be able to

- learn compactness in a more general setting
- relate the definitions with the earlier ones that you have learnt
- learn various characterizations and properties of compact spaces
- see certain examples of compact sets in familiar spaces
- learn about sequential compactness
- learn about various applications of compact spaces

13.1 Compact Spaces

Before starting with the definitions of compact spaces, lets go with the tradition of defining open covers, and subcovers first.

Definition 13.1.1. A collection $\mathcal{A}$ of subsets of a space X is said to cover X if X is equal to the union of the elements of $\mathcal{A}$. And $\mathcal{A}$ is called open cover if each of the elements of $\mathcal{A}$ are open sets of X.

Definition 13.1.2. A finite subfamily of $\mathcal{A}$ that also covers X is called a finite subcover for X.

Example 13.1.3. For example, the family of open intervals $\mathcal{A} = \{(-n, n)|n \in \mathbb{N}\}$, forms an open cover for $\mathbb{R}$. Also the subfamily $\mathcal{A}_e = \{(-2n, 2n)|n \in \mathbb{N}\}$ forms a subcover for $\mathbb{R}$.

We now define compactness of a space X.

Definition 13.1.4. A topological space (X, τ) is said to be compact if each open cover of X has a finite subcover.

Example 13.1.5. Let X be an infinite set with the co-finite topology. Then X is compact. Let

$$\mathcal{A} = \{U_\alpha|\alpha \in I\}$$

be an open cover for X. Take any $U_\alpha \in \mathcal{A}$. Then, by the definition of co-finite topology, $X \setminus U_\alpha$is a finite set, say

$$X \setminus U_\alpha = \{a_1, a_2, \ldots, a_n\}$$

Since $\mathcal{A}$ covers X so for each $i = 1(1)n$, there exists a $U_{\alpha_i} \in \mathcal{A}$ such that $a_i \in U_{\alpha_i}$. So clearly, the finite subcollection $\{U_\alpha, U_{\alpha_1}, U_{\alpha_2}, \ldots, U_{\alpha_n}\}$ covers X. Hence X with the co-finite topology is compact.

Definition 13.1.6. A subset A of a topological space X is said to be compact if it is compact as a subspace of X with the topology induced by the topology on X.

Example 13.1.7.

1. The real line with the usual topology is not compact since the open cover

   $$\mathcal{A} = \{(n, n + 2)|n \in \mathbb{Z}\}$$

   does not have any finite subcover.

2. The interval $(0, 1]$ is not compact. The open covering

   $$\mathcal{A} = \{(1/n, 1]|n \in \mathbb{Z}_+\}$$

   contains no finite subcover for $(0, 1]$.

3. The closed interval $[0, 1]$ is compact and it is by the well known Heine Borel theorem for $\mathbb{R}$.

Lemma 13.1.8. Let A be a subspace of X. Then A is compact if and only if every open covering of A by the sets open in X contains a finite subcollection covering A.

Proof. Let A be compact and let $\mathcal{A} = \{A_\alpha|\alpha \in I\}$ be an open covering for X by open sets in X. Then the family

$$\{A_\alpha \cap A|\alpha \in I\}$$

is an open covering for X by open sets in A. Then there exists a finite subcollection $\{A_{\alpha_1} \cap A, A_{\alpha_2} \cap A, \ldots, A_{\alpha_n} \cap A\}$ which covers A. Thus, the subcollection $\{A_{\alpha_1}, A_{\alpha_2}, \ldots, A_{\alpha_n}\}$ covers A by means of open

sets in A.

Conversely, suppose that the given condition holds. We prove that A is compact. Let $\{B_\alpha | \alpha \in J\}$ be an open cover for A by open sets in A. For each α,choose a set A_α open in X such that

$$B_\alpha = A \cap A_\alpha$$

Then the collection $\{A_\alpha | \alpha \in J\}$ is an open cover for A by open sets in X. By hypothesis, there exists a finite subcollection of $\{A_\alpha | \alpha \in J\}$, say $\{A_{\alpha_1}, A_{\alpha_2}, \ldots, A_{\alpha_n}\}$ that covers A. Thus the subcollection $\{B_{\alpha_1}, B_{\alpha_2}, \ldots, B_{\alpha_n}\}$ covers A. Thus, A is compact. □

The usefulness of the above lemma lies in the fact that whenever we are dealing with the compactness of a subspace, then it becomes redundant to differentiate between open cover by the open sets in X or in A.

Theorem 13.1.9. Every closed subspace of a compact space is compact.

Proof. Let C be a closed subspace of a compact topological space X. Let $\mathcal{A}$ be an open cover of C by open sets in X. Then the family of open sets in $\mathcal{A}$ along with the open set $X \setminus C$ forms an open cover of X. Since X is compact, so the said family has a finite subcover say

$$\{G_1, G_2, \ldots, G_n\} \cup (X \setminus C)$$

of X. Thus, the subfamily

$$\{G_1, G_2, \ldots, G_n\}$$

clearly covers C. Thus, C is compact. □

Theorem 13.1.10. Every compact subspace of a Hausdorff space is closed.

Proof. Let Y be a compact subspace of a Hausdorff space X. We will prove that $X \setminus Y$ is open. Let $a \in X \setminus Y$. For each point $y \in Y$, we will get disjoint open neighbourhoods U_y and V_y of the points a and y respectively. The collection $\{V_y | y \in Y\}$ is a cover of Y by open sets in X, and hence we will get a finite subfamily of it, say $\{V_{y_1}, V_{y_2}, \ldots, V_{y_n}\}$ that cover Y. The open set

$$V = V_{y_1} \cup V_{y_2} \cup \ldots \cup V_{y_n}$$

contains Y and it is disjoint from the open set

$$U = U_{y_1} \cap U_{y_2} \cap \ldots \cap U_{y_n}$$

formed by taking the corresponding intersection of the open neighbourhoods of a. For if $z \in V$, then $z \in V_{y_i}$ for each i and hence $z \notin U_{y_i}$ for each i and hence $z \notin U$. Thus U is the required neighbourhood of a that is disjoint from Y, that is $U \subset X \setminus Y$. Hence $X \setminus Y$ is open and thus, Y is closed. □

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

Lemma 13.1.11. If Y is a compact subspace of a Hausdorff space X and $a \notin Y$. Then there exist disjoint open sets U and V of X containing a and Y respectively.

Theorem 13.1.12. The image of a compact space under a continuous map is compact.

Proof. Let $f : X \to Y$ be continuous and X be compact. Let $\mathcal{A}$ be an open covering of the set $f(X)$ by sets open in Y. Then the collection

$$\{f^{-1}(A)|A \in \mathcal{A}\}$$

is a collection of open sets covering X. Thus, due to the compactness of X, there is a finite subfamily, say

$$\{f^{-1}(A_i)|i = 1(1)n\}$$

that covers X. Then the subfamily $\{A_1, A_2, \ldots, A_n\}$ that cover $f(X)$. □

An important application of the above theorem can be stated as the following theorem.

Theorem 13.1.13. Let $f : X \to Y$ be a bijective continuous map where X is compact and Y is Hausdorff. Then f is a homeomorphism.

Proof. We will prove that the images of a closed sets of X under f are closed in Y. If A is closed in X, then its compact. Thus, by the previous theorem, $f(A)$ is compact. Since Y is Hausdorff, so $f(A)$ is closed in Y. □

We are now heading to prove another equivalent definition of compactness. First we will define the Finite Intersection Property (FIP) of a collection of subsets of a space X.

Definition 13.1.14. A collection $\mathcal{C}$ of subsets of X is said to have FIP, if the intersection of a finite subfamily of $\mathcal{C}$ has non-empty intersection.

The following theorem gives us an equivalent definition of compactness in terms of FIP.

Theorem 13.1.15. A topological space X is compact if and only if for every collection $\mathcal{C}$ of closed sets in X having FIP, the intersection $\bigcap\limits_{C\in\mathcal{C}} C$ is non-empty.

Proof. Given a collection $\mathcal{A}$ of subsets of X, let

$$\mathcal{C} = \{X - A|A \in \mathcal{A}\}$$

be the collection of their complements. Then the following statements hold:

- $\mathcal{A}$ is a collection of open sets if and only if $\mathcal{C}$ is a collection of closed sets.
- The collection $\mathcal{A}$ covers X if and only if the intersection $\cap_{C\in\mathcal{C}}C$ of all the elements of $\mathcal{C}$ is empty.
- The finite subcollection $\{A_1, A_2, ..., A_n\}$ of $\mathbb{A}$ covers X if and only if the intersection of the corresponding elements $C_i = X - A_i$ of $\mathcal{C}$ is empty.

The first statement is trivial, while the second and third follow from DeMorgan's law:

$$X - (\cup_{\alpha\in J}A_\alpha) = \cap_{\alpha\in J}(X - A_\alpha).$$

The proof of the theorem now proceeds in two easy steps: taking the contrapositive (of the theorem), and then the complement(of the steps).
The statement that X is compact is equivalent to saying: "Given any collection $\mathcal{A}$ of open subsets of X, if $\mathcal{A}$ covers X, then some finite subcollection of $\mathcal{A}$ covers X."
This statement is equivalent to its contrapositive, which is the following: "Given any collection $\mathcal{A}$ of open sets, if no finite subcollection of $\mathcal{A}$ covers X, then $\mathcal{A}$ does not cover X." Letting $\mathcal{C}$ be, as earlier, the collection $\{X - A|A \in \mathcal{A}\}$ and applying to all the three above bullets, we see that this statement is in turn equivalent to the following:
"Given any collection $\mathcal{C}$ if closed sets, if every finite intersection of elements of $\mathcal{C}$ is nonempty, then the intersection of all the elements of $\mathcal{C}$ is nonempty." This is just the condition of our theorem. □

A special case of this theorem occurs when we have a nested sequence $C_1 \supset C_2 \supset ... \supset C_n \supset C_{n+1} \supset ...$ of closed sets in a compact space X. If each of the sets C_n is nonempty, then the collection $\mathcal{C}$=$\{C_n\}_{n\in\mathbb{Z}_+}$ automatically has the finite intersection

$$\cap_{n\in\mathbb{Z}_+} C_n$$

is nonempty.
We shall use the closed set criterion for the compactness in the next section to prove the uncoutability of the set of real numbers.

13.1.1 Lebesgue Lemma

The concept of Lebesgue number is new and highly useful for an open covering of a metric space. First lets recapitulate some preliminary definitions.

Definition 13.1.16. Let (X, d) be a metric space and let A be a non-empty subset of X. Then for each $x \in X$, the distance of x from A is defined as

$$d(x, A) = \inf\{d(x, a)|a \in A\}$$

It is easy to show that the function $d(x, A)$ is continuous.

Also recall the diameter of a set A in X, which is defined as

$$\text{diam}A = \sup\{d(a_1, a_2)|a_1, a_2 \in A\}$$

Then a set A is called bounded if diam(A) is finite.

We will now state a lemma which is known as the Lebesgue number lemma.

Lemma 13.1.17. Let $\mathcal{A}$ be an open covering of the metric space (X, d). If X is compact, there is a $\delta > 0$ such that for each subset of X having diameter less than δ, there exists an element of $\mathcal{A}$ containing it.

This number δ is called the Lebesgue number for the covering $\mathcal{A}$.

Proof. Let $\mathcal{A}$ be an open cover of X. If X is itself an element of $\mathcal{A}$, then any positive number is a Lebesgue number for $\mathcal{A}$. So, we assume that X is not an element of $\mathcal{A}$. Choose a finite subcollection $\{A_1, A_2, \ldots, A_n\}$ of $\mathcal{A}$ that covers X. For each i, set

$$C_i = X \setminus A_i,$$

and define $f : X \to \mathbb{R}$ by letting

$$f(x) = \frac{d(x, C_1) + d(x, C_2) + \cdots + d(x, C_n)}{n}$$

We show that $f(x) > 0$ for all x. Given $x \in X$, choose i so that $x \in A_i$. Then choose ϵ so that the ϵ-neighbourhood of x lies in A_i. Then $d(x, C_i) \geq \epsilon$, so that $f(x) \geq \frac{\epsilon}{n}$.

Since f is continuous, it has a minimum value δ. We show that this δ is our required Lebesgue number. Let B be a subset of X of diameter less than δ. Choose a point a of B. Then B lies in the δ-neighbourhood of a. Now

$$\delta \leq f(a) \leq d(a, C_m),$$

where $d(a, C_m)$ is the largest of the numbers $d(a, C_i)$. Then the δ-neighbourhood of a is contained in the element $A_m = X \setminus C_m$ of the cover $\mathcal{A}$. □

13.1.2 Limit Point Compactness

Definition 13.1.18. A space X is said to be limit point compact if every infinite subset of X has a limit point.

Theorem 13.1.19. Compactness implies limit point compactness.

Proof. Let X be a compact set and let A be a subset of X. Suppose A has no limit point. Then we show that A is finite.

Since A has no limit point, so A is closed. Further, for each $a \in A$, we can choose a neighbourhood U_a of a that intersects A at the point a alone. The space X is covered by the open set $X \setminus A$ along with the open sets U_a. Since X is compact, it can be covered by finitely many of these sets. Since $X \setminus A$ does not intersect A, and each U_a contains only one point of A, the set A must be finite. Hence contrapositively, we have proved the theorem. □

The converse of the theorem is not necessarily true.

Example 13.1.20. Let Y consists of two points endowed with the indiscrete topology. The space $X = \mathbb{Z} \times Y$ is limit point compact, for every non-empty subset of X has a limit point. But it is not compact since for the covering of X by open sets of the form

$$U_n = \{n\} \times Y$$

there is no finite subcover for X.

We will now move to define the sequential compactness of a space. But before that, let us see the convergence of sequences in a topological space.

Definition 13.1.21.

1. Let X be a topological space and let $\{x_n\}$ be a sequence of points in X. Then $\{x_n\}$ is said to converge to a point $a \in X$, if for every open set U containing a, we can find $m \in \mathbb{N}$ such that $x_n \in U$ for all $n \geq m$.

2. If

$$n_1 < n_2 < \cdots < n_k < \cdots$$

is an increasing sequence of natural numbers, then the sequence $\{x_{n_k}\}$ is called a subsequence of the sequence $\{x_n\}$.

Definition 13.1.22. A topological space is said to be sequentially compact if every sequence of points of X has a convergent subsequence.

Example 13.1.23.

1. Let A be a finite subset of a space X. Let $\{x_n\}$ be a sequence of points in A. Since A is finite, then at least one element of the sequence, say x_k has to appear infinite number of times in the sequence. Thus, if we construct a subsequence as $\{x_k, x_k, \ldots, x_k, \ldots\}$, then it surely converges to the point x_k of A. Thus, A is sequentially compact.

2. The open interval $A = (0, 1)$ with the usual topology is not sequentially compact. Since the sequence of points $\{1/2, 1/3, 1/4, \ldots, 1/n, \ldots\}$ clearly does not have any subsequence that converges to a point of A.

Now, for any arbitrary metric space, the concept of sequential compactness and compactness are the same. But this is not true for any arbitrary topological space. In fact, for a topological space, neither implies the other.

Exercise 13.1.24.

1. Show that a finite union of compact subspaces of X is compact.
2. Show that if $f : X \to Y$ is a continuous map where X is compact and Y is Hausdorff, then f is a closed mapping.
3. Show that every subspace of $\mathbb{R}$ with the co-finite topology is compact.
4. Show that sequential compactness may not necessarily imply compactness.
5. Show that compactness may not necessarily imply sequential compactness.

Unit 14

Course Structure

- Introduction
- Objectives
- BW compactness and Countable compactness
- Local compactness
- Baire Category Theorem

Introduction

In this section we will learn still further types of compactness and their relationship with compactness and sequential compactness. In the previous unit, we saw that sequential compactness and compactness are equivalent in any arbitrary metric space but we can't conclude anything in case of any arbitrary topological space. In this unit also we will see the relationship between the definitions in a metric space as well as topological space.

Objectives

After reading this unit, you will be able to

- define countable compactness
- define BW compactness
- define locally compact sets
- relate them with compactness and sequential compactness
- understand the explicit relationships between these definitions in topological spaces
- relate them in case of metric spaces

- know various properties relating to them
- show various examples for each type
- state the Baire Category theorem

14.1 Countable Compactness

Definition 14.1.1. A topological space X is called countably compact if every countable open cover of X has a finite subcover. A family of sets $\mathcal{A}$ is called a countable open cover of X if

- it covers X
- it has countable number of elements
- each element of $\mathcal{A}$ is open

The definition can also be given as the following theorem.

Theorem 14.1.2. A topological space X is countably compact if and only if each sequence has an accumulation point.

From the definition, it is obvious that any compact space is always countably compact. Also, any sequentially compact space is also countably compact. But the converse is not true in both the cases.

Example 14.1.3. Consider the natural number space $\mathbb{N}$ with the topology generated by

$$\{1,2\},\{3,4\},\{5,6\},\ldots$$

Let A be a non-empty subset of $\mathbb{N}$ and let $m \in A$. If a is odd, then $a+1$ is a limit point of A and if a is even, then $a-1$ is a limit point of A. In either case, A has an accumulation point. So, $\mathbb{N}$ is countably compact. But it is not compact since the open cover

$$\{1,2\},\{3,4\},\{5,6\},\ldots$$

clearly has no finite subcover. Also, it is not sequentially compact since the sequence $\{1,2,3,\ldots\}$ contains no convergent subsequence.

Theorem 14.1.4. A countably compact space is limit point compact.

Proof. Without any loss of generality, assume that A be a countably infinite subset of a countably compact topological space X without any limit points. Then, A is closed in X. So for each $a \in A$, the open set U_a containing a intersects A only at a. Let $V_a = U_a \cup (X \setminus A)$. Then V_a is an open set. Still $V_a \cap A = \{a\}$. Also, $\{V_a | a \in A\}$ is a countable cover of X. Thus it has a finite subcover say $\{V_{a_1}, V_{a_2}, \ldots, V_{a_n}\}$ of X. But then

$$\begin{aligned} A &= \left(\bigcup_{i=1}^{n} V_{a_i}\right) \cap A \\ &= \bigcup_{i=1}^{n} (V_{a_i} \cap A) \\ &= \{a_1, a_2, \ldots, a_n\} \end{aligned}$$

which shows that A is finite, a contradiction. Hence the result. □

Theorem 14.1.5. A space that is T_1 and limit point compact is countably compact.

Proof. Suppose that X is not countably compact. So there is a countable open cover $\{U_n | n \in \mathbb{N}\}$ without a finite subcover. For each i, pick $x_i \in X \setminus (U_1 \cup U_2 \cup \ldots \cup U_i)$. This is possible since $\{U_n\}$ does not have a finite subcover for X. Let

$$A = \{x_1, x_2, \ldots\}$$

Suppose now that x is in X. Then $x \in U_N$ for some $N \in \mathbb{N}$(since $\{U_n\}$ covers X). This U_N can only contain x_i for $i < N$ by definition of $x_i's$. So, U_N is a neighbourhood of x that contains only finite number of points of A. Since X is T_1, this implies that x can't be a limit point of A. Since $x \in X$ was arbitrary, so no point of X can be a limit point of A, which is a contradiction since X is limit point compact. Hence, X has to be countably compact. □

We have another equivalent definition for countable compactness given in the form of the following theorem.

Theorem 14.1.6. A topological space is countably compact if and only if every nested sequence $C_1 \supset C_2 \supset \ldots$ of closed non-empty sets of X has a non-empty intersection.

Proof. Left as an exercise. □

We will now look at a new type of topological space called a Bolzano Weierstrass space which we define below.

Definition 14.1.7. A topological space X is called Bolzano-Weierstrass Space (often called BW-space) if every infinite subset of X has a limit point.

Example 14.1.8. The real line with the usual topology is not a BW-space. Since the subset $\mathbb{Z}$ contains no limit point in $\mathbb{R}$.

Theorem 14.1.9. If a space X is compact, then it is BW.

Proof. Let X be a compact space. Then every open cover of X has a finite subcover. If possible, let X is not a BW-space. Then there exists an infinite subset A of X that does not have a limit point, that is, $A' = \emptyset$. So A is closed. So, A is compact in X.Further, since A has no limit point, for each $a \in A$, there exists an open neighbourhood U_a of a, such that

$$A \cap U_a = \{a\}$$

So, $\{U_a | a \in A\}$ is an open cover for A. Since A is compact, there is a finite subcover,say $\{U_{a_1}, U_{a_2}, \ldots, U_{a_n}\}$ which covers A. But $\bigcup_{i=1}^{n} U_{a_i}$ contains only n points, which implies that A is a finite set, a contradiction. Hence X has to be a BW-space. □

Does this proof have any resemblance to that we did for countably compact spaces? So it certainly must have something to do with countably compact spaces. Lets look at the following theorem.

Theorem 14.1.10. If X is Hausdorff, then X is BW if and only if it is countably compact.

14.1.1 Local Compactness

Definition 14.1.11. A space X is said to be locally compact if each point in X has a compact neighbourhood.

If X is itself compact, then it is locally compact and in that case X itself acts as a compact neighbourhood of each of its points. If the topology is discrete and X is infinite, then X is locally compact without being compact (verify!).

Example 14.1.12. The real number space $\mathbb{R}$ with the usual topology is locally compact without being compact. Since, for $x \in \mathbb{R}$, the closed interval $[x - \delta, x + \delta]$, $\delta > 0$ is a neighbourhood of x which is compact.

Theorem 14.1.13. Let X be a Locally compact T_2-space

Proof. Let E be a closed set of X which is locally compact. We will show that E with the subspace topology induced by the topology on X, is locally compact. Let $x \in E$. Since X is locally compact, so we can find a compact neighbourhood N of x in X. Put $M = N \cap E$. So, M is a neighbourhood of x in E and as M is a closed set in a compact space N, so M is a compact subset of N, that is, M is a compact neighbourhood of x in E. Thus, E is locally compact. □

Compact Hausdorff spaces are one of the "good" spaces to work with since they have many useful properties. If a given space is not so, then the next best thing that one can hope for is that whether it is a subspace of one such space, that is to say, whether it is homeomorphic to the subspace of such a space. If so, then under what conditions is it possible to define such a homeomorphism. This section mainly deals with the problem of embedding a given space into a Compact Topological space. We start with the following theorem.

Theorem 14.1.14. A space X is locally compact Hausdorff if and only if there exists a space Y such that

1. X is a subspace of Y.
2. The set $Y \setminus X$ consists of a single point.
3. Y is a compact Hausdorff space.

If Y and Y' are two spaces satisfying the above conditions, then there is a homeomorphism of Y and Y' that equals the identity map on X.

If X itself happens to be compact, then the space Y is not so interesting since it is formed by adjoining a single isolated point to X. However, if X is not compact, then the point of $Y \setminus X$ is a limit point of X and $\overline{X} = Y$. If Y is a compact Hausdorff space and X is a proper subspace of Y whose closure equals Y, then Y is said to be a *compactification* of X and if $Y \setminus X$ is a singleton set, then Y is called a *one-point compactification* of X.

Theorem 14.1.15. Let X be a Hausdorff space. Then it is locally compact if and only if given $x \in X$ and a neighbourhood U of x, there is a neighbourhood V of x such that $\overline{V}$ is compact and $\overline{V} \subset U$.

Proof. Suppose X is locally compact and let $x \in X$ and U be a neighbourhood of X. Take the one-point compactification Y of X and let $C = Y \setminus U$. Then C is closed in Y and hence compact subspace of Y. Then, by a previous lemma, we can find disjoint open sets V and W containing x and C respectively. Then the closure $\overline{V}$ of V in Y is compact and moreover, $\overline{V}$ is disjoint from C, so that $\overline{V} \subset U$. Hence the result. □

14.1.2 Baire Spaces

Definition 14.1.16. We have learnt about the interior of a set A as the union of all the open sets contained in A. If A has an empty interior, then it contains no open set of X other than the emptyset. Remember that

$$\overline{(X \setminus A)} = X \setminus \text{Int}A$$

So, equivalently, A has empty interior if every point of A is a limit point of the complement of A, that is, if the complement of A is dense in X.

Definition 14.1.17. A space X is said to be a Baire space if the following condition holds: Given any countable collection $\{A_n\}$ of closed sets of X each of which has an empty interior in X, their union $\bigcup A_n$ also has an empty interior in X.

Example 14.1.18. The space $\mathbb{Q}$ of rationals is not a Baire space since each one point set in $\mathbb{Q}$ is closed and has an empty interior in $\mathbb{Q}$ and $\mathbb{Q}$ is the countable union of these one-point sets. Whereas, the set $\mathbb{Z}_+$ forms a Baire space. Every subset of $\mathbb{Z}_+$ is open and hence have non-empty interior. Thus, $\mathbb{Z}_+$ satisfies the Baire condition trivially.

Theorem 14.1.19. (**Baire Category Theorem**:) If X is a compact Hausdorff space, or a complete metric space, then X is a Baire space.

Proof. Given a countable collection $\{A_n\}$ of closed sets of X having empty interiors, we will show that $\bigcup A_n$ also has an empty interior in X. So, given the non-empty open set U_0 of X, we must find a point x of U_0 that does not lie in any of the sets A_n.

Consider the first set A_1. By hypothesis, A_1 does not contain U_0. Hence, we may choose a point y of U_0 that is not in A_1. Regularity of X, along with the fact that A_1 is closed, enables us to choose a neighbourhood U_1 of y such that

$$U_1 \cap A_1 = \emptyset,\ \overline{U}_1 \subset U_0.$$

If X is a metric, then we choose U_1 small enough that its diameter is less than 1.

In general, given the non-empty open set U_{n-1}, we choose a point of U_{n-1} that is not in the closed set A_n, and then we choose U_n to be a neighbourhood of this point such that

$$\overline{U}_n \cap A_n = \emptyset,\ \overline{U}_n \subset U_{n-1}$$

such that diam$U_n < 1/n$ in the metric case. We assert that the intersection $\cap\overline{U}_n$ is non-empty. From this fact, our theorem will follow. For, if $x \in \cap\overline{U}_n$, then x is in U_0 since $\overline{U}_1 \subset U_0$. And for each n, the point x is not in A_n because $\overline{U}_n$ is disjoint from A_n.

If X is a compact Hausdorff space, then we consider the nested sequence

$$\overline{U}_1 \supset \overline{U}_2 \supset \cdots$$

of non-empty subsets of X. The collection $\{\overline{U}_n\}$ has the finite intersection property. Since X is compact, $\cap\overline{U}_n$ must be non-empty.

If on the other hand, X is a complete metric, then we use the Cantor's Intersection theorem to complete the proof. □

Exercise 14.1.20.

1. Show that a sequentially compact space is countably compact.
2. Show that an infinite set with the discrete topology is compact without being locally compact.
3. Show that any open subspace of a Baire space is also a Baire space.
4. Show that a space X is countably compact if and only if every nested sequence
$$C_1 \supset C_2 \supset \ldots \supset C_n \ldots$$
of closed non-empty sets of X has a non-empty intersection.
5. Show that a space X is limit point compact if and only if each sequence in X has a limit point.
6. Let X be limit point compact space and let A be a closed subset of X. Is A limit point compact?

Unit 15

Course Structure

- Constructing a M*ö*bius strip
- Identification topology

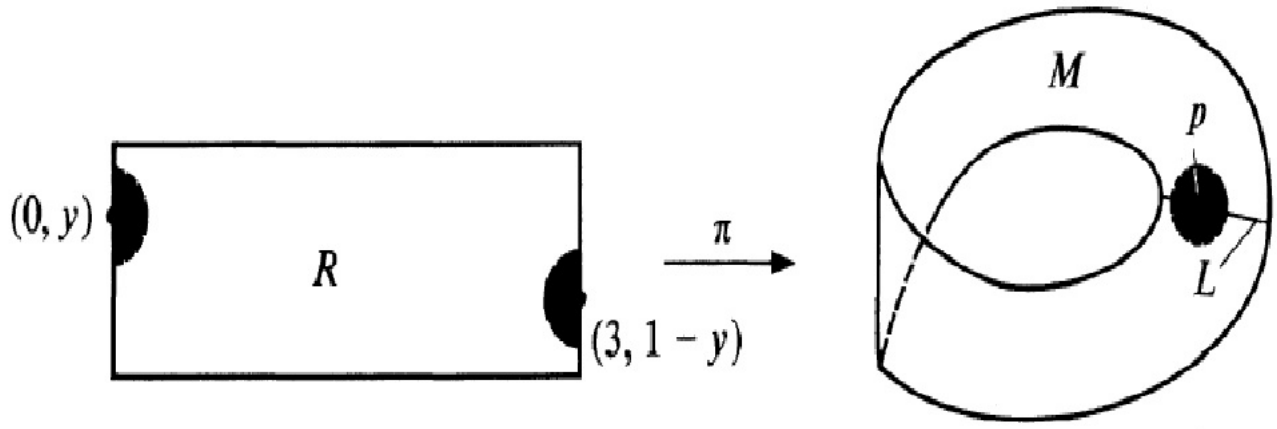
15.1 Constructing a M*ö*bius Strip

To construct a Möbius strip, one takes a rectangle and identifies a pair of opposite edges with a half twist. Our first job is to translate this process into precise mathematical language. For the rectangle take the subspace R of $\mathbb{R}^2$ consisting of the points (x, y) for which $0 \leq x \leq 3$ and $0 \leq 1$. To describe the identification of the vertical edges of R with a half twist, we partition R into disjoint non-empty sub sets in such a way that two points lie in the same subset if and only if we wish them to be identified. If we now take these sub sets as the points of our M*ö*bius strip, then we have made the required identifications. The appropriate partition of R consists of:

1. sets consisting of a pair of points of the form $(0, y)$, $(3, 1 - y)$, where $0 \leq y \leq 1$.
2. sets consisting of a single point (x, y), where $0 < x < 3$, $0 \leq y \leq 1$.

So far we have defined a set which we shall call M, its points being the subsets of the above partition of R. There is a natural function π from R onto M that sends each point of R to the subset of the partition in which it lies. The identification topology on M is defined to be the largest topology for which π is continuous. That is to say, a subset O is defined to be open in the identification topology on M if and only if $\pi^{-1}(O)$ is open in the rectangle R.

A glance at the figure above shows the sort of open sets we obtain. We represent the points of M in the usual way as a sub set of $\mathbb{R}^3$, and we label with the letter L the image under π of the two vertical edges of R. If we use R_* to denote R minus its vertical edges, then the restriction of π to R_* is one-one and is a homeomorphism of R_* with $M \setminus L$. Therefore, we know all about the neighbourhoods of points of $M \setminus L$; they are simply the images under π of neighbourhoods of the points of R_*. If p lies on the line L, then $\pi^{-1}(p)$ consists of two distinct points, situated on the vertical edges of R, of the form $(0, y),\ (3, 1 - y)$. The union of two open half-discs in R, centres $(0, y),\ (3, 1 - y)$ and of equal radius, maps via π to an open neighbourhood of p in the identification topology on M. Notice that if we take a single half-disc, its image in M is not a



neighbourhood of p and is not open, so π is not an open mapping. The points of L are in no sense special in the Möbius strip; they have the same sort of neighbourhoods in the identification topology as all the other points of M. In fact, it is easy to check that the identification topology coincides with that induced from $\mathbb{R}$ on our set M.

For convenience we have illustrated M in $\mathbb{R}^3$. However, we emphasize that the definition of the Möbius strip as an 'identification space' given in this section is entirely abstract, and in no way relies on a particular representation of the strip as a set of points in euclidean space.

15.1.1 The Identification Topology

Let X be a topological space and let $\mathcal{P}$ be a family of disjoint non-empty subsets of X such that $\cup\mathcal{P} = X$. Such a family is usually called a partition of X. We form a new space Y, called an identification space, as follows:

The points of Y are the members of $\mathcal{P}$ and, if $\pi : X \to Y$ sends each point of X to the subset of $\mathcal{P}$ containing it, the topology of Y is the largest for which π is continuous. Hence, a subset of Y is open if and only if $\pi^{-1}(O)$ is open in X. This topology is called the identification topology on Y. We think of Y as a space obtained from X by identifying each of the subsets of $\mathcal{P}$ to a single point.

Our construction of the Möbius strip was a special case of this procedure. We will now prove one or two general results on identification spaces. We begin with a theorem that is useful when checking the continuity of a function which has an identification space as domain.

Theorem 15.1.1. Let Y be an identification space defined as above and let Z be an arbitrary topological space. A function $f : Y \to Z$ is continuous if and only if the composition $f\pi : X \to Z$ is continuous.

Proof. Let U be an open subset of Z. Then $f^{-1}(U)$ is open in Y if and only if $\pi^{-1}(f^{-1}(U))$ is open in X, that is, if and only if $(f\pi)^{-1}(U)$ is open in X. □

Let $f : X \to Y$ be an onto map and suppose that the topology on Y is the largest for which f is continuous. Then we call f an identification map, the reason for our terminology being as follows. Any function $f : X \to Y$ gives rise to a partition of X whose members are the subsets $\{f^{-1}(y)\}$, where $y \in Y$. Let Y_* denote the identification space associated with this partition, and $\pi : X \to Y_*$ the usual map.

Theorem 15.1.2. If f is an identification map, then

1. the spaces Y and Y_* are homeomorphic;
2. a function $g : Y \to Z$ is continuous if and only if the composition $gf : X \to Z$ is continuous.

Proof. The proof of 2 is exactly that of the previous theorem since Y has the largest topology for which f is continuous. The points of Y_* are the sets $\{f^{-1}(y)\}$, where $y \in Y$. Define $h : Y_* \to Y$ by $h(\{f^{-1}(y)\}) = y$. Then h is a bijection and satisfies $h\pi = f$, $h^{-1}f = \pi$. By the previous theorem, h is continuous, and h^{-1} is continuous by 2. Hence h is a homeomorphism. □

Theorem 15.1.3. Let $f : X \to Y$ be an onto map. If f maps open sets of X to open sets of Y, or closed sets to closed sets, then f is an identification map.

Proof. Suppose f maps open sets to open sets. Let U be a subset of Y for which $f^{-1}(U)$ is open in X. Since f is onto, we have $f(f^{-1}(U)) = U$, and therefore U must be open in the given topology on Y. So this topology is the largest for which f is continuous, and f is an identification map. The proof for closed maps is similar. □

Corollary 15.1.4. Let $f : X \to Y$ be an onto map. If X is compact and Y is Hausdorff, then f is an identification map.

Proof. A closed subset of the compact space X is compact and its image under the continuous function f is therefore a compact subset of Y. But a compact subset of a Hausdorff space is closed. Therefore f takes closed sets to closed sets, and we can apply the previous theorem. □

We shall use the previous theorem and corollary in order to compare different descriptions of the same topological space. We begin with two methods of constructing a torus.

The Torus. Take X to be the unit square $[0, 1] \times [0, 1]$ in $\mathbb{R}^2$, with the subspace topology, and partition X into the following subsets:

1. the set $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$ of four corner points;
2. sets consisting of pairs of points $(x, 0), (x, 1)$, where $0 < x < 1$;
3. sets consisting of pairs of points $(0, y), (1, y)$, where $0 < y < 1$;
4. sets consisting of a single point (x, y), where $0 < x < 1$ and $0, < y < 1$;

The resulting identification space is the torus. An equally common description is to say that the torus is the product $S^1 \times S^1$ of two circles. Thinking of the points of S^1 as complex numbers, we can define a map $f : [0, 1] \times [0, 1] \to S^1 \times S^1$ by $f(x, y) = (e^{2\pi ix}, e^{2\pi iy})$. The partition of $[0, 1] \times [0, 1]$ which contains of the inverse images under f of the points of $S^1 \times S^1$ is exactly that given earlier. By the previous corollary, f is an identification map and hence our two descriptions of the torus are homeomorphic.

Let us recollect the gluing lemma.

Theorem 15.1.5. If X and Y are closed in $X \cup Y$, and if both f and g are continuous on X and Y respectively, with a common co-domain Z such that $f(x) = g(x)$ when $x \in X \cap Y$, then the map $h : X \cup Y \to Z$, defined as

$$\begin{aligned} h(x) &= f(x),\ x \in X \\ &= g(x),\ x \in Y \end{aligned}$$

is continuous.

The glueing lemma remains true if we ask that X and Y are both open in $X \cup Y$. The lemma is also true for open sets. But, it fails if no restrictions are placed on the sets X and Y.

As we shall see, the glueing lemma can be explained in terms of identification maps and interpreted as a special case of the previous theorem. In order to do this, we introduce the disjoint union $X + Y$ of the spaces X, Y and the function $j : X + Y \to X \cup Y$ which when restricted to either X or Y is just the inclusion in $X \cup Y$. The function is important for our purpose because

1. it is continuous;

2. the composition $hj : X + Y \to Z$ is continuous if and only if f and g are continuous;

By combining parts 1 and 2 of theorem 15.1.2, we have the following theorem.

Theorem 15.1.6. If j is an identification map, and if both $f : X \to Z$ and $g : Y \to Z$ are continuous, then $h(= f \cup g) : X \cup Y \to Z$ is continuous.

The glueing lemma is a special case of this result, since if both X and Y are closed in $X \cup Y$, then j sends closed sets to closed sets and is an identification map by theorem 15.1.3.

If j is an identification map, then then we can think of $X \cup Y$ as an identification space formed from the disjoint union $X + Y$ by identifying certain points of X with points of Y. In this case, we often say that $X \cup Y$ has the identification topology. The open(closed) sets of $X \cup Y$ are those sets A for which $A \cap X$ and $A \cap Y$ are open (closed) sets of X and Y respectively.

The above theorem generalizes to the case of an arbitrary union. Let X_α, $\alpha \in A$ be a family of subsets of a topological space and give each X_α, and the union $\cup X_\alpha$, the induced topology. Let Z be a space and suppose we are given maps $f_\alpha : X_\alpha \to Z$, one for each $\alpha \in A$, such that if $\alpha, \beta \in A$,

$$f_\alpha|_{X_\alpha \cap X_\beta} = f_\beta|_{X_\alpha \cap X_\beta}$$

Define a function $F : \cup X_\alpha \to Z$ by gluing together the f_α, that is,

$$F(x) = f_\alpha(x),\ x \in X_\alpha.$$

Let $\boxplus X_\alpha$ denote the disjoint union of the spaces X_α, and let $j : \boxplus X_\alpha \to X_\alpha$ be the function which when restricted to each X_α is the inclusion in $\cup X_\alpha$.

Theorem 15.1.7. If j is an identification map, and if each f_α is continuous, then F is continuous.

Proof. Observe that $Fj : \boxplus X_\alpha \to Z$ is continuous if and only if each f_α is continuous and apply part 2 of theorem 15.1.2. □

Attaching maps. As a final example of an identification space we formalize the notion of attaching one space to another by means of a continuous function.

Let X, Y be spaces, let A be a subspace of Y, and let $f : A \to X$ be a continuous function. Our aim is to attach Y to X using f and to form a new space which we shall denote by $X \cup_f Y$. We begin with the disjoint union $X + Y$ and define a partition so that two points lie in the same subset if and only if they are identified under f. precisely, the subsets of the partition are

1. pairs of points $\{a, f(a)\}$ where $a \in A$;

2. individual points of $Y \setminus A$;

3. the individual points of $X \setminus \text{image}(f)$.

The identification space associated with this partition is $X \cup_f Y$. The map f is called the attaching map.

One final comment: if Y is an identification space formed from X, then Y is the image of X under a continuous function and therefore inherits properties such as compactness, connectedness, and path-connectedness from X. However, X may be Hausdorff and yet Y not satisfy the Hausdorff axiom. As an example, take X to be the real line with the usual topology, and partition X so that real numbers r and s lie in the same elements of the partition if and only if $r - s$ is rational. Then the corresponding identification is an indiscrete space.

Unit 16

Course Structure

- Orbit Spaces

16.1 Orbit Spaces

We will now consider spaces which have, in addition to their topology, the structure of a group. A good example is the circle.

Definition 16.1.1. A topological group G is both a Hausdorff space and a group, the two structures being compatible in the sense that the group multiplication $m : G \times G \to G$, and the function $i : G \to G$ which sends each group element to its inverse, are continuous.

Example 16.1.2.

1. The real line, the group structure being addition of real numbers.
2. The circle S^1, thought of as the set of complex numbers of unit modulus. Then the two functions

$$S^1 \times S^1 \to S^1,\ (e^{i\theta}, e^{i\phi}) \mapsto e^{i(\theta+\phi)}$$

and

$$S^1 \to S^1,\ e^{i\theta} \mapsto e^{-i\theta}$$

are continuous.

3. Any abstract group with the discrete topology.
4. The Euclidean n-space.
5. The group of invertible $n \times n$ matrices with real entries. The group structure is matrix multiplication. For the topology we identify each $n \times n$ matrix $A = (a_{ij})$ with the corresponding point

$$(a_{11}, a_{12}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, \ldots, a_{n1}, \ldots, a_{nn})$$

of $\mathbb{R}^{n^2}$ and take the subspace topology. This topological group is called the general linear group, denoted by $GL(n, \mathbb{R})$.

6. The orthogonal group $O(n)$ consisting of $n \times n$ orthogonal matrices with real entries. $O(n)$ has both its topology and its group stricture induced from $GL(n, \mathbb{R})$.

The terms 'isomorphism' and 'subgroup' for topological groups require a few words of explanation. In each case we need to take into consideration both the topological and the algebraic structures. So an isomorphism between two topological groups is a homeomorphism which is also a group isomorphism. In the same spirit, a subset of a topological group is called a subgroup if it is algebraically a subgroup and in addition has the subspace topology. Hence, the integers $\mathbb{Z}$ with the discrete topology form a subspace of the real line $\mathbb{R}$.

Definition 16.1.3. A topological group G is said to act as a group of homeomorphisms on a space X if each group element induces a homeomorphism of the space in such a way that;

1. $hg(x) = h(g(x))$ for all $g, h \in G$ and all $x \in X$.
2. $e(x) = e$ for all $x \in X$, where e is the identity element of G.
3. the function $G \times X \to X$ defined by $(g, x) \mapsto g(x)$ is continuous.

If $x \in X$, then for each $g \in G$, the corresponding homeomorphism either fixes x or maps it to a new point $g(x)$. The subset of X consisting of all such images $g(x)$ as g varies over G is called the orbit of x and written as $O(x)$. If two orbits intersect then they must coincide: the relation defined by $x \sim y$ if and only if $x = g(y)$ for some $g \in G$ is an equivalence relation on X whose equivalence classes are precisely the orbits of the given action. So the orbits define a partition of X. The corresponding identification space is called the orbit space written as X/G. In constructing X/G, we 'divide' by G in the sense that we identify two points of X if and only if they differ by one of the homeomorphisms $x \mapsto g(x)$.

Example 16.1.4. The orbit of the real number x consists of all the points $x + n$ where $n \in \mathbb{Z}$. Therefore in forming $\mathbb{R}/\mathbb{Z}$, we identify two points of $\mathbb{R}$ if and only if they differ by an integer and we obtain the circle as orbit space.

Example 16.1.5. Taking the product of our first example with itself in the natural way gives an action of $\mathbb{Z}\times\mathbb{Z}$ on the plane. An ordered pair of integers $(m, n) \in \mathbb{Z} \times \mathbb{Z}$ sends the point $(x, y) \in \mathbb{R}^2$ to $(x + m, y + n)$. The orbit space is the product of two circles, in other words the torus. It may help to think of this action geometrically. Divide the plane into squares of unit side by drawing in all horizontal and vertical lines through the points with integer coordinates. The homeomorphisms of our group action preserve this pattern of squares, and any single square contains points from each orbit and therefore maps onto the torus under the identification map

$$\mathbb{R}^2 \xrightarrow{\pi} \mathbb{R}^2/\mathbb{Z} \times \mathbb{Z} = T$$

Each square has its sides identified by π in the usual way in order to form T.

Theorem 16.1.6. Let G act on X and suppose that both G and X/G are connected. Then X is connected.

Proof. Suppose X is the union of two disjoint non-empty open subsets U and V. Since the identification map $\pi : X \to X/G$ always takes open sets to open sets, and since X/G is connected, $\pi(U)$ and $\pi(V)$ can't be disjoint. Now, if $x \in \pi(U) \cap \pi(V)$, then both $U \cap O(x)$ and $V \cap O(x)$ are non-empty. These two sets decompose the orbit $O(x)$ as a disjoint union of two non-empty open sets. But $O(x)$ is the image of G under the continuous function $f : G \to X$ defined by $f(x) = g(x)$. $O(x)$ is therefore connected, and we have established the required contradiction. □

Unit 17

Course Structure

- Introduction
- Some Elementary properties of Topological groups

Introduction

This unit can be thought of as a continuation of the previous unit. It deals with primarily the topological groups along with certain properties of topological groups.

17.1 Elementary Properties of Topological Groups

Let G be a topological group and x an element of G. The function $L_x : G \to G$ defined by $L_x(g) = xg$ is called the left translation by the element x. It is clearly one-one and onto, and it is continuous because it is the composition

$$G \to G \times G \xrightarrow{m} G$$

as $g \mapsto (x, g) \mapsto xg$. The inverse of L_x is L_x^{-1} and thereforeL_x is a homeomorphism. Similarly the right translation $R_x : G \to G$ given by $R_x(g) = gx$ is also a homeomorphism. Thus, if we have $U \subset G$ and $x \in G$, then

$$U \text{ open} \iff tU \text{ open} \iff Ut \text{ open} \iff U^{-1} \text{ open}.$$

These translations show that a topological group has a certain 'homogeneity' as a topological space. For if x and y are any two points of a topological group G there is a homeomorphism of G that maps x to y, namely the translation L_{yx}^{-1}. Hence G exhibits the same topological structure locally near each point.

Theorem 17.1.1. Let G be a topological group and let K denote the connected component of G which contains the identity element. Then K is a closed normal subgroup of G.

Proof. Components are always closed as we have seen previously. For any $x \in K$, the set $Kx^{-1} = R_x^{-1}(K)$ is connected (since R_x^{-1} is a homeomorphism) and contains $e = xx^{-1}$. Since K is the maximal connected subset of G containing e, we must have $Kx^{-1} \subseteq K$. Hence $KK^{-1} = K$, and K is a subgroup of G. Normality follows in a similar manner. For any $g \in G$ the set $gKg^{-1} = R_g^{-1}L_g(K)$ is connected and contains e. Hence $gKg^{-1} \subseteq K$. □

Theorem 17.1.2. In a connected topological group any neighbourhood of the identity element is a set of generators for the whole group.

Proof. Let G be a connected topological group and let V be a neighbourhood of e in G. Let $H = \langle V \rangle$ be the subgroup of G generated by the elements of V. If $h \in H$ then the whole neighbourhood $hV = L_h(V)$ of h lies in H, so H is open. We claim that the complement of H is also open. For if $g \in G \setminus H$, consider the set gV. If $gV \cap H$ is non-empty, say $x \in gV \cap H$, then $x = gv$ for some $v \in V$. This gives $g = xv^{-1}$, which implies the contradiction $g \in H$ since both x and v^{-1} lie in H. Hence the neighbourhood $L_g(V) = gV$ of g lies in $G \setminus H$ and we see that $G \setminus H$ is an open set. Now, G is connected and so cannot be partitioned into two disjoint non-empty open sets. Since H is non-empty, we must have $G \setminus H = \emptyset$, that is, $G = H$. □

If G is a group and S and T are subsets of G, we let ST and S^{-1} denote

$$ST = \{st : \ s \in S, \ t \in T\} \text{ and } S^{-1} = \{s^{-1} : \ s \in S\}.$$

The subset S is called symmetric if $S^{-1} = S$. Also, we denote e as the identity element of G.

Theorem 17.1.3. Let G be a topological group. Every neighbourhood U of e contains an open symmetric neighbourhood V of e such that $VV \subset U$.

Proof. Let Int$U$ be the interior of U. Consider the multiplication map

$$\mu : \text{Int}U \times \text{Int}U \to G.$$

Since μ is continuous, then $\mu^{-1}(\text{Int}U)$ is open and contains (e, e). So, there are open sets $V_1, V_2 \in U$ such that $(e, e) \in V_1 \times V_2$, and $V_1V_2 \in U$. Let $V_3 = V_1 \cap V_2$, then $V_3V_3 \subset U$ and V_3 is an open neighbourhood of e. Finally, let $V = V_3 \cap V_3^{-1}$, which is open, contains e, is symmetric, and satisfies $VV \subset U$. □

Theorem 17.1.4. If G is a topological group, then every open subgroup of G is also closed.

Proof. Let H be an open subgroup of G. Then any coset xH is also open. So,

$$Y = \bigcup_{x \in G \setminus H} xH$$

is also open. From elementary group theory,

$$H = G \setminus Y,$$

and so H is closed. □

Theorem 17.1.5. If G is a topological group, and if K_1 and K_2 are compact subsets of G, then K_1K_2 is compact.

Proof. The set $K_1 \times K_2$ is compact in $G \times G$, and the multiplication mapping is continuous. Since the continuous image of a compact set is compact, so K_1K_2 is compact. □

Recall the closure of a set A in a topological space X. We have studied that a necessary and sufficient condition for $x \in \overline{A}$ is that, for every open neighbourhood U of x, the set $U \cap A = \not\emptyset$. It can also be seen as follows:

Theorem 17.1.6. If $x \notin \overline{A}$, then there is a closed set F which contains A, but $x \notin F$.

Proof. Left as an exercise. □

Theorem 17.1.7. If G is a topological group, and H is a subgroup of G, then $\overline{H}$ is a subgroup of G.

Proof. Let $g, h \in \overline{H}$, and let U be an open neighbourhood of the product gh. Let

$$\mu : G \times G \to G$$

be the multiplication map, which is continuous. So, $\mu^{-1}(U)$ is open in $G \times G$, and contains (g, h). So, there are open neighbourhoods V_1 of g and V_2 of h such that $V_1 \times V_2 \subset \mu^{-1}(U)$. Since $g, h \in \overline{H}$, then there are points $x \in V_1 \cap \overline{H}$ and $y \in V_2 \cap \overline{H}$. Since $x, y \in H$, we have $xy \in H$ and since $(x, y) \in \mu^{-1}(U)$, then $xy \in U$. Thus, $xy \in U \cap H$, and since U was an arbitrary open neighbourhood of gh, then we have $gh \in \overline{H}$.

Now, let

$$i : G \to G$$

denote the inverse map, and let W be an open neighbourhood of h^{-1}. Then, $i^{-1}(W) = W^{-1}$ is open and contains h. So there is a point $z \in H \cap W^{-1}$. Then $z^{-1} \in H \cap W$ and as before, this implies that $h^{-1} \in \overline{H}$. Thus, $\overline{H}$ forms a subgroup of G. □

In the course of proving the above theorem, we have shown that the closure of a symmetric neighbourhood of e is again symmetric.

Lemma 17.1.8. Let G be a topological group, F a closed subset of G, and K a compact subset of G, such that $F \cap K = \emptyset$. Then there is an open neighbourhood V of e such that $F \cap VK = \emptyset$ (and an open neighbourhood V' of e such that $F \cap KV' = \emptyset$).

Proof. Let $x \in K$, so $x \in G \setminus F$ and $G \setminus F$ is open. So, $(G \setminus F)x^{-1}$ is an open neighbourhood of e. By theorem 17.1.3, there is an open neighbourhood W_x of e such that $W_x W_x \subset (G \setminus F)x^{-1}$, Now,

$$K \subset \underset{x \in K_x}{W} x,$$

and K is compact, so there exists a finite number of points $x_1, x_2, \ldots, x_n \in K$, such that

$$K \subset \bigcup_{i=1}^{n} W_i x_i$$

where $W_i = W_{x_i}$. Now let

$$V = \bigcap_{i=1}^{n} W_i.$$

For any $x \in K$, $x \in W_i x_i$, for some i. Now we have

$$Vx \subset W_i x \subset W_i W_i x_i \subset G \setminus F.$$

In other words, $F \cap V_x = \emptyset$. Since this is true for any $x \in K$, we have,

$$F \cap VK = \emptyset.$$

□

From theorem 17.1.3, the neighbourhood V in above lemma can be taken to be symmetric.

Theorem 17.1.9. Let G be a topological group, K a compact subset of G, and F a closed subset of G. Then FK and KF are closed subsets of G.

Proof. If $FK = G$, then the result is trivial. So, let $y \in G \setminus FK$. This means that $F \cap yK^{-1} = \emptyset$. Since K is compact, yK^{-1} is compact. So, by the previous lemma, there is an open neighbourhood V of e such that $F \cap VyK^{-1} = \emptyset$, or $FK \cap Vy = \emptyset$. Since V_y is an open neighbourhood of y contained in $G \setminus FK$, we have FK is closed. Similarly, we can show that KF is closed. □

17.1.1 Separation properties and functions

You are quite familiar with the definitions of T_1 and T_2 spaces by now. We will investigate in this section, the relationship between these separation axioms and a topological group. If G is a topological group, then if G is T_1, then by homogeneity, $\{e\}$ is a closed set in G and conversely.

We have seen that, if a space X is T_2, then it is T_1 but the converse is not true in general. But, we see that the converse is true in case of topological groups.

Theorem 17.1.10. Let G be a T_1 topological group. Then G is Hausdorff.

Proof. Let g, h be distinct elements of G. By the T_1 axiom, let U be an open set containing e, such that $gh^{-1} \notin U$. By theorem 17.1.3, let V be an open symmetric neighbourhood containing e, such that $VV \subset U$. Now, Vg is open and contains g, and Vh is open containing h. We must have $Vg \cap Vh = \emptyset$, otherwise there are $v_1, v_2 \in V$ such that

$$v_1 g = v_2 h,$$

which would mean

$$gh^1 = v_2 v_1^{-1} \in VV^{-1} = VV \subset U,$$

while gh^{-1} was chosen to be an element not in U. Thus, G is Hausdorff. □

You might remember the regular and completely regular spaces as well as Tychonoff space. We learnt that a completely regular space is always regular. We will now see that any topological group which is T_1 is also completely regular, and thus, regular.

Theorem 17.1.11. Let G be a topological group, and let e_G denote the identity element in G and let F be a closed subset of G such that $e_G \notin F$. Then there is a continuous function $f : G \to [0, 1]$ such that

$$f(e_G) = 1, \text{ and } f(y) = 1,\ y \in F.$$

Theorem 17.1.12. If G is a topological group which is T_1, then G is completely regular and thus, regular.

Proof. Let $x \in G$ and let F be a closed subset of G such that $x \notin F$. Then $x^{-1}F$ is a closed subset of G not containing e_G, and thus from the previous theorem, there is a continuous function $f : G \to [0, 1]$ such that $f(e_G) = 0$ and $f(y) = 1$ for $y \in x^{-1}F$. Now the function $h(g) = f(x^{-1}g)$ is the desired continuous function, and since G is also T_1, G is completely regular, and hence is also regular. □

17.1.2 Connectedness

We have certain important properties of a topological group concerning connectedness. Let us study them in some details.

Theorem 17.1.13. The connected component of the identity in a topological group is a subgroup.

Proof. Let G_0 be the connected component of G containing the identity and $h, k \in G_0$ be arbitrary. The set $h^{-1}G_0$ is connected and contains the identity and so $G_0 \cup h^{-1}G_0$ is also connected. Since G_0 is a component, we have $G_0 \cup h^{-1}G_0 = G_0$ which implies that $h^{-1}G_0 \subset G_0$. In particular, $h^{-1}k$ belongs to G_0 from which we can conclude that G_0 is a subgroup. □

Theorem 17.1.14. Suppose that G is a topological group and K is a subgroup and the coset space G/K is given the quotient topology. Then

1. If K and G/K are connected, then G is connected.
2. If K and G/K are compact, then G is compact.

Proof. If G is connected then so is G/K since the quotient map $\eta : G \to G/K$ is a continuous surjection. To prove the converse, suppose that K and G/K are connected and $f : G \to \{0, 1\}$ be an arbitrary continuous map. We have to show that f is constant. The restriction of f to K must be constant and since each coset gK is connected, f must be constant on gK as well taking value $f(g)$. Thus, we have a well-defined map $f' : G/K \to \{0, 1\}$ such that $f' \circ \eta = f$. By the fundamental property of quotient spaces, it follows that f' is continuous and so must be constant since G/K is connected. □

Theorem 17.1.15. If G is a connected topological group and H is a subgroup which contains a neighbourhood of the identity then $H = G$. In particular, an open subgroup of G equals G.

Proof. Let U be the open neighbourhood of the identity that is contained in H and $h \in H$ be arbitrary. Since multiplication by h is a homeomorphism, the set $Uh = \{uh \mid u \in U\}$ is also open and also contained in H. Hence the set

$$L = \bigcup_{h \in H} Uh$$

is open and contained in H. Since U contains the identity, $H \subset L$ and we conclude that H is open. Our job will be over if we can show that H is closed as well. Let $x \in \overline{H}$ be arbitrary. Since the neighbourhood Ux of x contains a point $y \in H$, there exists $u \in U$ such that $y = ux$ which, in view of the fact that $U \subset H$, implies $x \in H$. Hence $\overline{H} = H$. □

Theorem 17.1.16. Suppose G is a connected topological group and H is a discrete normal subgroup of G then H is contained in the center of G.

Proof. Since H is discrete, the identity element is not a limit point of H and so there is a neighbourhood U of the identity such that $U \cap H = \{e\}$. We may assume U has the property that if $u_1, u_2 \in U$, then the product $u_1^{-1}u_2 \in U$. This follows from the continuity of the group operation and a detailed verification is left as an exercise. It is easy to see that if h_1 and h_2 are two distinct elements of H, then

$$Uh_1 \cap Uh_2 = \emptyset.$$

Fix $h \in H$ and consider now the set K given by

$$K = \{g \in G | gh = hg\}.$$

We shall show that the subgroup K contains a neighbourhood of the identity. Pick a neighbourhood V of the identity such that $V = V^{-1}$ and that $(hVh^{-1}V) \cap H = \{e\}$. Then for any $g \in V$, we have on the one hand

$$hgh^{-1}g^{-1} \in hVh^{-1}V$$

and on the other hand, $hgh^{-1}g^{-1} \in H$ since H is normal. Hence

$$hgh^{-1}g^{-1} \in (hVh^{-1}V) \cap H = \{e\}$$

which shows that g belongs to K and K contains a neighbourhood of the unit element. We may now invoke the previous theorem. □

Unit 18

Course Structure

- Introduction
- $GL(n, \mathbb{R})$ as a topological group
- Its subgroups

Introduction

Though we have already read a bit about $GL(n, \mathbb{R})$ previously, we will read about it in more details in this unit.

18.1 The Group $GL(n, \mathbb{R})$

We will first show that $GL(n, \mathbb{R})$ actually forms a topological group.

Theorem 18.1.1. The matrix group $GL(n, \mathbb{R})$ is a topological group.

Proof. Let M denote the set of all $n \times n$ matrices which have real entries, and let $A = (a_{ij})$ represent a typical element of M. We can identify M with euclidean space of dimension n^2 by associating $A = (a_{ij})$ with the point $(a_{11}, a_{12}, \ldots, a_{1n}, a_{21}, \ldots, a_{2n}, a_{n1}, \ldots, a_{nn})$. The identification gives us a topology on M and we claim that, with respect to this topology, matrix multiplication $m : M \times M \to M$ is continuous. To see this, we need only examine the well-known formula for the entries of a product matrix: If $A = (a_{ij})$ and $B = (b_{ij})$ then the ijth entry in the product $m(A, B)$ is $\sum_{k=1}^{n} a_{ik} b_{kj}$. Now, M has the topology of the product space

$$\mathbb{R} \times \mathbb{R} \times \cdots \mathbb{R} \ (n^2 \text{ copies}),$$

and for each i, j satisfying $1 \leq i, j \leq n$ we have a projection

$$\pi_{ij} : M \to \mathbb{R}$$

which sends a given matrix A to its ijth entry. By the property of continuous functions, we can say that m is continuous if and only if all the composite functions

$$M \times M \xrightarrow{m} M \xrightarrow{\pi_{ij}} \mathbb{R}$$

are continuous. But, $\pi_{ij}m(A, B) = \sum_{k=1}^{n} a_{ik}b_{kj}$, a polynomial in the entries of A and B. Hence $\pi_{ij}m$ is continuous.

The elements of $GL(n, \mathbb{R})$ are invertible matrices in M. If we give $GL(n, \mathbb{R})$ the subspace topology from M, then, by the above, matrix multiplication

$$GL(n, \mathbb{R}) \times GL(n, \mathbb{R}) \to GL(n, \mathbb{R})$$

is continuous. It remains to prove that the inverse function

$$i : GL(n, \mathbb{R}) \to GL(n, \mathbb{R})$$

is also continuous. We use the same technique:

$$i : GL(n, \mathbb{R}) \to GL(n, \mathbb{R}) \subseteq \mathbb{R} \times \mathbb{R} \times \cdots \mathbb{R}$$

is continuous if and only if all of the composite functions

$$GL(n, \mathbb{R}) \xrightarrow{i} GL(n, \mathbb{R}) \xrightarrow{\pi_{jk}} \mathbb{R}, \;\; 1 \leq j, k \leq n$$

are continuous. Now the composition of π_{jk} with i sends a matrix A to the jkth element of A^{-1}, that is, to $(1/\det A)(kj$th cofactor of $A)$. It should be clear that the determinant of A and the cofactors of A are polynomials in the entries of A. Since $\det A$ does not vanish on $GL(n, \mathbb{R})$, our composition $\pi_{jk}i$ is continuous. This completes the proof that $GL(n, \mathbb{R})$ is a topological group. □

We note in passing that $GL(n, \mathbb{R})$ is the inverse image of the nonzero real numbers under the determinant function

$$\det : M \to \mathbb{R}.$$

So, $GL(n, \mathbb{R})$ is not compact (it is an open subset of M), and is not connected (the matrices with positive and negative determinants partition $GL(n, \mathbb{R})$ into two disjoint non-empty open sets). Let us do it in some details.

Theorem 18.1.2. $GL(n, \mathbb{R})$ is open and unbounded.

Proof. The complement of $GL(n, \mathbb{R})$ in M is

$$\{A \in M \mid \det A = 0\}.$$

Since the determinant is continuous and $\{0\}$ is closed in $\mathbb{R}$, so the above set is closed in M, and hence its complement, that is, $GL(n, \mathbb{R})$ is open in M. Also, $kI_n \in GL(n, \mathbb{R})$ for all $k > 0$. So, $GL(n, \mathbb{R})$ is unbounded. Thus, proved. □

Theorem 18.1.3. $GL(n, \mathbb{R})$ is connected.

Proof. Note that,

$$\det : GL(n, \mathbb{R}) \to \mathbb{R} \setminus \{0\}$$

is a surjective continuous map and $\mathbb{R} \setminus \{0\}$ is not connected and since we know that a continuous map sends a connected set into a connected set, so $GL(n, \mathbb{R})$ can't be connected. □

In fact, as we have said earlier, we can partition $GL(n, \mathbb{R})$ into two disjoint open sets, namely,

$$\begin{aligned} GL_+(n, \mathbb{R}) &= \{A \in GL(n, \mathbb{R}) \; : \det A > 0\} \\ GL_-(n, \mathbb{R}) &= \{A \in GL(n, \mathbb{R}) \; : \det A < 0\} \end{aligned}$$

Since the determinant is a continuous function, so the above sets are open. Hence, there are two (path) components of $GL(n, \mathbb{R})$.

18.1.1 Subgroups of $GL(n, \mathbb{R})$

We have already seen that the set of all orthogonal matrices $O(n)$ forms a subgroup of $GL(n, \mathbb{R})$. In this section, we will study more about the subgroups of $GL(n, \mathbb{R})$. Let us list some of the subgroups of $GL(n, \mathbb{R})$ as follows:

1. $SL(n, \mathbb{R}) = \{A \in GL(n, \mathbb{R}) : \det A = 1\}$. This group is called the special linear group. Note that, since the determinant function is continuous, so $\det^{-1}(\{1\}) = SL(n, \mathbb{R})$ is a closed set in $GL(n, \mathbb{R})$.
2. $O(n) = \{A \in GL(n, \mathbb{R}) : AA^T = I_n = A^T A\}$. As we have already seen, this is the orthogonal group.
3. $SO(n) = \{A \in O(n) : \det A = 1\}$. This is called the special orthogonal group.

From the elementary properties of matrices, we can say that each of the above groups actually form a subgroup of $GL(n, \mathbb{R})$ and has been left as an exercise. If instead of $\mathbb{R}$, we take the field of complex, that is, if we consider the set $GL(n, \mathbb{C})$ then we have some additional subgroups, which are listed below:

1. $U(n) = \{U \in GL(n, \mathbb{C} : UU^* = I_n = U^*U\}$, where, U^* denotes the conjugate transpose of U. This is called the unitary group.
2. $SU(n) = \{U \in U(n) : \det(U) = 1\}$. This particular group is called the special unitary group.

Let us now see certain properties of these subgroups.

Theorem 18.1.4. The groups $O(n)$ and $SO(n)$ are compact.

Proof. Write any matrix $A \in O(n)$ as

$$A = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

where, each v_i is a row matrix. Then from the identity $AA^T = I_n$, we get,

$$v_i v_i^T = 1, \ \ 1 \leq i \leq n.$$

This implies that A is inside the unit ball of $\mathbb{R}^{n^2}$. Hence $O(n)$ is a bounded subset of the Euclidean space $\mathbb{R}^{n^2}$. Let $\{A_k\}$ be any sequence in $O(n)$ and let $A_k \to A$ in M. Taking limit as $k \to \infty$ in the relation

$$A_k A_k^T = A_k^T A_k = I_n,$$

by continuity of multiplication, we get

$$AA^T = A^T A = I_n$$

proving that $A \in O(n)$. Thus, $O(n)$ is closed too. Hence by Heine-Borel theorem, $O(n)$ is compact. If, in addition, each of the matrices A_k above have the determinant 1, then by the continuity of the determinant, we also see that$SO(n)$ is closed in $O(n)$. And since we know that a closed subset of a compact space is compact, so $SO(n)$ is compact. □

Before going into other details, let us see the following theorem.

Theorem 18.1.5. Let $P \in M$ be a positive semidefinite matrix. Then P is symmetric, $\det P \geq 0$ and there exists a unique positive demidefinite matrix $P^{1/2} \in M$ such that

$$(P^{1/2})^2 = P$$

and $P^{1/2}$ is invertible if and only if P is so.

Theorem 18.1.6. Let $A \in SL(n, \mathbb{R})$. Then there exists a rotation matrix $R \in SO(n)$ and a real, symmetric and positive semidefinite matrix $P \in SL(n, \mathbb{R})$ such that $A = RP$.

Proof. Let $P = (A^T A)^{1/2}$ and this forces R to be defined as

$$R = AP^{-1}.$$

Clearly, P is real, symmetric and positive semidefinite matrix. Further,

$$\begin{aligned} RR^T &= AP^{-1}P^{-1}A^T \\ &= AP^{-2}A^T \\ &= A(A^T A)^{-1}A^T \\ &= I_n \end{aligned}$$

and similarly, $R^T R = I_n$ implying that R is orthogonal. In particular,

$$1 = \det(R^T R) = \det(R^T) \det R = (\det R)^2$$

which implies that $\det R = 1$ or, -1. Now,

$$1 = \det A = \det R \det P$$

and so by the previous theorem, and the fact that P is invertible, we have $\det P > 0$. Hence, we must have $\det R = 1$, that is, $R \in SO(n)$. □

Theorem 18.1.7. Any matrix in $SO(n)$ is orthogonally similar to a block diagonal form $A_1 \boxplus A_2 \boxplus \cdots \boxplus A_r$, where, each A_r is $[1]$ or a rotation matrix of the type

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

for some $\theta \in \mathbb{R}$.

Theorem 18.1.8. $O(n)$ is not connected, whereas $SO(n)$ is path-connected.

Proof. Let $M \in O(n)$. Then, by the above theorem, $\det M \in \{+1, -1\}$. Let

$$O_{\pm}(n) = \{M \in O(n) : \det M = \pm 1\} = GL_{\pm}(n, \mathbb{R}) \cap O(n).$$

Then, $O_+(n)$, which is the same as the subgroup $SO(n)$ and $O_-(n)$ are open in the subspace topology and they form a disconnection of $O(n)$, implying that $O(n)$ is not connected.

Next we show that any matrix in $SO(n)$ is joined to I_n by a path. Since $SO(1) = \{|1|\}$, let us assume that $n \geq 2$. Let $R \in SO(n)$. Then by a previous theorem, there exists an orthogonal matrix $M \in O(n)$ such that

$$MRM^T = A_1 \boxplus A_2 \boxplus \cdots \boxplus A_r$$

where each A_i is $[1]$ or a 2×2 rotation matrix of the type

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

for some $\theta \in \mathbb{R}$. Without any loss of generality, assume that, for some $k \leq r$, for $1 \leq i \leq k$,

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

for some $\theta_i \in \mathbb{R}$ and that $A_i = [1]$ for $k \leq i \leq r$.

We can now look for an appropriate path. For each $1 \leq i \leq k$, consider the map

$$\phi_i : [0,1] \to SO(2)$$

as

$$\phi_i(t) = \begin{bmatrix} \cos(t\theta_i) & \sin(t\theta_i) \\ -\sin(t\theta_i) & \cos(t\theta_i) \end{bmatrix}.$$

Thus, each ϕ_i is a path in $SO(2)$ with end points I_2 and

$$\begin{bmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{bmatrix}.$$

Hence, the map $\phi : [0,1] \to SO(n)$ given by

$$\phi(t) = M^T(\phi_1(t) \boxplus \phi_2(t) \boxplus \cdots \boxplus \phi_k(t) \boxplus I_{n-2k})M$$

is a path in $SO(n)$ with end points $\phi(0) = I_n$ and $\phi(1) = R$, thereby establishing that $SO(n)$ is path-connected. □

We thus get the following corollary.

Corollary 18.1.9. $O(n)$ has precisely two path-components, namely $O_+(n)$ and $O_-(n)$.

Proof. By the previous theorem, we get that

$$O_+(n) = SO(n)$$

is path-connected. Now, let $A, B \in O_-(n)$ and fix a $C \in O_-(n)$. Then

$$AC, BC \in O_+(n)$$

and therefore, there exists a path ϕ in $O_+(n)$ joining AC and BC. Consider the map

$$\tilde{\phi} : [0,1] \to O_-(n)$$

given by

$$\tilde{\phi}(t) = \phi(t)C^{-1}.$$

This $\tilde{\phi}$ is a path in $O_-(n)$ joining A and B.

We also know that $O(n)$ is a disjoint union of $O_+(n)$ and $O_-(n)$, so these are the only two path-components of $O(n)$. □

Note that $SL(1,\mathbb{R}) = \{[1]\}$ is clearly path-connected and compact. However, in higher dimensions, this is not the case.

Corollary 18.1.10. $SL(n,\mathbb{R})$ is closed, path-connected and is not compact for $n \geq 2$.

Unit 19

Course Structure

- Introduction
- Objectives
- Fundamental Groups
- Preliminary Properties

Introduction

One of the main problems of topological spaces is to determine whether two spaces are homeomorphic or not. There is no method in general to solve this problem but we have some techniques to apply in particular cases. To find homeomorphism is to find bijective continuous maps, and constructing continuous maps is a problem for which there are techniques of solving.

However showing two space not homeomorphic is quite easy. This can be done with the help of topological properties. If a space has a topological property which is not satisfied by the other, then we can clearly say that they are not homeomorphic. But this technique also has very limited application so far. So we must introduce new properties and techniques. Hence comes the idea of fundamental groups which is a generalisation of all the ideas we have done so far. Two spaces that are homeomorphic, have their fundamental groups isomorphic. This unit deals in the fundamental groups and their basic properties.

Objectives

After reading this unit, you will be able to

- define homotopy of paths
- define the various properties related to homotopy and path homotopies
- define fundamental group of a topological space
- learn the properties of the fundamental groups

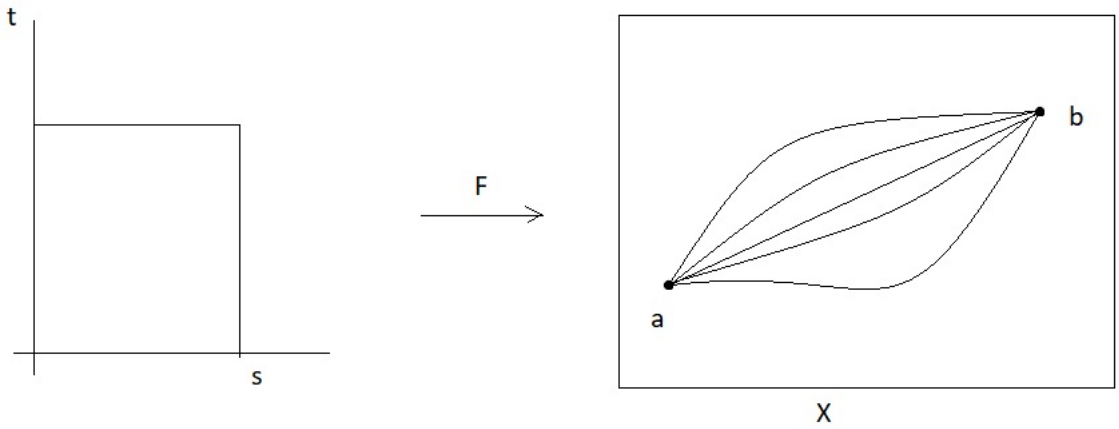


Figure 19.1.1: Path-Homotopy

19.1 Fundamental Group

Before defining the fundamental group of a space X, we shall consider paths on X and an equivalence relation called path homotopy between them.

Definition 19.1.1. If f and f' are continuous maps of the space X into the space Y, we say that f is homotopic to f' if there is a continuous map

$$F : X \times I \to Y$$

such that

$$F(x, 0) = f(x) \text{ and } F(x, 1) = f'(x)$$

for each x. Here, $I = [0, 1]$. The map F is called a homotopy between f and f'. If f is homotopic to f' then we write $f \simeq f'$. If $f \simeq f'$ and f' is a constant function, we say that f is nulhomotopic.

Now we consider a special case in which f is a path in X. Recall that if $f : [0, 1] \to X$ is a continuous map such that $f(0) = x_0$ and $f(1) = x_1$, we say that x_0 is the initial point and x_1, the final point, of the path f. We will now define a stronger relation, called path-homotopy between two paths.

Definition 19.1.2. Two paths f and f' mapping the interval $I = [0, 1]$ to f are said to be path-homotopic if they have the same initial point a and same final point b, where $a, b \in X$ and if there exists a continuous map $F : I \times I \to X$ such that

$$F(s, 0) = f(s) \text{ and } F(s, 1) = f'(s)$$

$$F(0, t) = a \text{ and } F(1, t) = b$$

for each $s \in I$ and each $t \in I$. We call F a path-homotopy between f and f'. If f and f' are path-homotopic, then we write $f \simeq_p f'$. (See Figure 1)

The first condition says that F is a homotopy between f and f' and the second condition says that for each t, the path f_t defined by the equation $f_t(s) = F(s, t)$ is a path from a to b. Said differently, the first condition says that F represents a continuous way of deforming the path f to f' and the second condition ensures that the end points remain fixed in this deformation.

Theorem 19.1.3. The relations $\simeq$ and $\simeq_p$ are equivalence relations.

If f is a path, then we shall denote its path-homotopy class by $[f]$.

Proof. Given f, it is trivial that $f \simeq f$; the map $F(x,t) = f(x)$ is the require homotopy. If f is a path, then F is a path-homotopy.

Given $f \simeq f'$, we show that $f' \simeq f$. Let F be a homotopy between f and f'. Then $G(x,t) = F(x, 1-t)$ is the homotopy between f' and f. If F is a path homotopy, then G is so.

Suppose $f \simeq f'$ and $f' \simeq f''$. Let F be a homotopy between f and f' and let F' be a homotopy between f' and f''. Define $G : X \times I \to Y$ by the equation

$$\begin{aligned} G(x,t) &= F(x,2t),\ t \in \left[0, \frac{1}{2}\right] \\ &= F'(x, 2t-1),\ t \in \left[\frac{1}{2}, 1\right] \end{aligned}$$

The map G is well-defined, since if $t = \frac{1}{2}$, we have

$$F(x,2t) = f'(x) = F'(x, 2t-1).$$

Because G is continuous on the two closed subsets $X \times \left[0, \frac{1}{2}\right]$ and $X \times \left[\frac{1}{2}, 1\right]$ of $X \times I$, it is continuous on all of $X \times I$, by the gluing lemma. Thus, G is the required homotopy between f and f''.

If F and F' are path-homotopies, then G is so. We only need to show that the initial and final points are fixed. Let f and f' are path-homotopic, then all the previous conditions remain same except in this case, we have their initial and final points same, say they are a and b respectively. Similarly, if f' and f'' are path-homotopic, then f'' also have the initial and final points as a and b respectively. We have,

$$\begin{aligned} G(0,t) &= F(0,2t) = a\ t \in \left[0, \frac{1}{2}\right] \\ &= F'(0, 2t-1) = a,\ t \in \left[\frac{1}{2}, 1\right] \end{aligned}$$

So, we are getting $G(0,t) = a$. Similarly, we can show that, $G(1,t) = b$, for $t \in I$. Hence, G is a path-homotopy. □

Example 19.1.4. Let C be a convex sub set of a euclidean space and let $f, g : X \to C$ maps, where X is an arbitrary topological space. For each point $x \in X$, the straight line joining $f(x)$ and $g(x)$ lies in C, and we can define a homotopy from f to g simply by sliding f along these straight lines. To be precise, define $F : X \times I \times C$ by

$$F(x,t) = (1-t)f(x) + tg(x).$$

Notice that if f and g happen to agree on a subset A of X then this homotopy is a homotopy relative to A. The homotopy F is called a straight-line homotopy.

We now introduce some algebra into this geometric definition.

Definition 19.1.5. If f is a path in X from a to b, and if g is a path in X from b to c, we define the product $f * g$ of f and g to be the path h given by the equations

$$\begin{aligned} h(s) &= f(2s),\ s \in \left[0, \frac{1}{2}\right], \\ &= g(2s-1),\ s \in \left[\frac{1}{2}, 1\right]. \end{aligned}$$

The function h is well-defined and continuous, by the pasting lemma; it is a path in X from a to c. We think of h as the path whose first half is the path f and second half is the path g.

The product operation on paths induces a well-defined operation on path-homotopy classes, defined by the equation

$$[f] * [g] = [f * g].$$

To verify this fact, let F be a path-homotopy between f and f' and let G be a path homotopy between g and g'. Define

$$\begin{aligned} H(s,t) &= F(2s,t),\ s \in \left[0, \frac{1}{2}\right], \\ &= G(2s-1,t),\ s \in \left[\frac{1}{2}, 1\right]. \end{aligned}$$

Because $F(1,t) = b = G(0,t)$ for all t, the map H is well-defined, it is continuous by the pasting lemma. It can be checked that H is the required path homotopy between $f * g$ and $f' * g'$.

The operation $*$ on path-homotopy classes turns out to satisfy properties that look very much like the axioms for a group. They are called the groupoid properties of $*$. One difference from the properties of a group is that $[f] * [g]$ is not defined for every pair of classes, but only for those pairs $[f]$, $[g]$ for which $f(1) = g(0)$.

Theorem 19.1.6. The operation $*$ has the following properties

1. If $[f] * ([g] * [h])$ is defined, so is $([f] * [g]) * [h]$, and they are equal.(associativity)
2. Given $x \in X$, let e_x denote the constant path $e_x : I \to X$ carrying all of I to the point x. If f is a path in X from a to b, then
$$[f] * [e_b] = [f] \text{ and } [e_a] * [f] = [f].$$
(right and left inverses)
3. Given the path f in X from a to b, let $\overline{f}$ be the path defined by $\overline{f}(s) = f(1-s)$. It is called the reverse of f. Then
$$[f] * [\overline{f}] = [e_a] \text{ and } [\overline{f}] * [f] = [e_b].$$
(inverse)

The set of path-homotopy classes of paths in a space X does not form a group under the operation $*$ because the product of two path-homotopy classes is not always defined. But if we pick out a point a of X serving as a base point and restrict ourselves to those beginning and ending at a, then those homotopy classes form a group under $*$, which is called the fundamental group.

Before starting with the definitions, we assume that the reader is well accustomed to the idea of group homomorphism, cosets, normal subgroups , factor or quotient groups. etc. Let us start with the definition of fundamental group relative to a base point a.

Definition 19.1.7. Let X be a space and let a be a point of X. A path that begins and ends at a is called a loop based at a. The set of path homotopy classes of loops based at a, with the operation $*$, is called the **fundamental group** of X relative to the base point a. It is denoted by $\pi_1(X, a)$.

It follows from the previous theorem that the operation $*$, when restricted to this set, satisfies the axioms for groups. Given two loops f and g based at a, the product $f * g$ is always defined and is a loop based at a. The other properties are immediate.

Sometimes this group is called the first homotopy group of X. From the term, it seems that there is also a second-homotopy group. There are in fact groups $\pi_n(X, a)$ for all $n \in \mathbb{N}$. But we shall omit them here and concentrate on the first homotopy group only.

Example 19.1.8. Let $\mathbb{R}^n$ denote the euclidean n-space. Then $\pi_1(\mathbb{R}^n, a)$ is the trivial group (the group consisting of the identity alone). For if f is a loop in $\mathbb{R}^n$ based at a, the straight-line homotopy is a path-homotopy between f and the constant path at a. More generally, if X is any convex subset of $\mathbb{R}^n$, then $\pi_1(X, a)$ is the trivial group. In particular, the open ball B^n in $\mathbb{R}^n$,

$$B^n = \{\mathbf{x} \mid x_1^2 + x_2^2 + \cdots + x_n^2 \leq 1\},$$

has a trivial fundamental group.

A valid question is that how much the fundamental group depends upon the base point. First let us see the following definition.

Definition 19.1.9. Let α be a path in X from a to b. We define a map

$$\hat{\alpha} : \pi_1(X, a) \to \pi_1(X, b)$$

by the equation

$$\hat{\alpha}([f]) = [\overline{\alpha}] * [f] * [\alpha].$$

The map $\hat{\alpha}$ is well-defined because the operation $*$ is well-defined. If f is a loop based at a, then $\overline{\alpha} * (f * \alpha)$ is a loop based at b. Hence $\hat{\alpha}$ maps $\pi_1(X, a)$ into $\pi_1(X, b)$ as desired. Note that, it only depends on the path-homotopy class of α.

Theorem 19.1.10. The map $\hat{\alpha}$ is a group isomorphism.

Proof. To show that $\hat{\alpha}$ is a homomorphism, we compute

$$\begin{aligned} \hat{\alpha}([f]) * \hat{\alpha}([g]) &= ([\overline{\alpha}] * [f] * [\alpha]) * ([\overline{\alpha}] * [g] * [\alpha]) \\ &= [\overline{\alpha}] * [f] * [g] * [\alpha] \\ &= \hat{\alpha}([f] * [g]). \end{aligned}$$

To show that $\hat{\alpha}$ is an isomorphism, we show that if β denotes the path $\overline{\alpha}$, which is the reverse of α, then $\hat{\beta}$ is an inverse for $\hat{\alpha}$. We compute for each element $[h]$ of $\pi_1(X, b)$,

$$\hat{\beta}([h]) = [\overline{\beta}] * [h] * [\beta] = [\alpha] * [f] * [\overline{\alpha}],$$

$$\hat{\alpha}(\hat{\beta}([h])) = [\overline{\alpha}] * ([\alpha] * [h] * [\overline{\alpha}]) * [\alpha] = [h].$$

A similar computation shows that

$$\hat{\beta}(\hat{\alpha}([f])) = [f]$$

for each $[f] \in \pi_1(X, a)$. □

Corollary 19.1.11. If X is path-connected and a and b are two points of X, then $\pi_1(X, a)$ and $\pi_1(X, b)$ are isomorphic.

Let X be a space and P be a path-component of it containing a. Then its easy to see that

$$\pi_1(X, a) = \pi_1(P, a),$$

since all the loops and homotopies in X are that are based at a must lie in P. Thus, $\pi_1(X, a)$ depends only on the path component of X containing a.

Also, if X is path-connected, all the groups $\pi_1(X, a)$ are isomorphic, so its tempting to try to identify all these groups with one another and to speak simply of the fundamental group of X, without reference to the base point. The difficulty in this approach is that there is no natural way to identify $\pi_1(X, a)$ with $\pi_1(X, b)$. Different paths α and β from a to b may give rise to different isomorphisms between these groups. For this reason, omitting the base point may lead to error.

We now move on to define the simple-connectedness of a space.

Definition 19.1.12. A space X is simply-connected if it is a path-connected space and if $\pi_1(X, a)$ is the trivial one-element group for some $a \in X$, and hence for every $a \in X$. We often express the fact that $\pi_1(X, a)$ is the trivial group by writing

$$\pi_1(X, a) = 0.$$

Lemma 19.1.13. In a simply connected space X, any two paths having the same initial and final points are path-homotopic.

Proof. Let α and β be two paths from a to b. Then $\alpha * \overline{\beta}$ is defined and is a loop on X based at a. Since X is simply connected, this loop is path-homotopic to the constant loop at a. Then

$$[\alpha * \overline{\beta}] * [\beta] = [e_a] * [\beta]$$

from which it follows that

$$[\alpha] = [\beta].$$

□

By now, it might be clear that fundamental groups are a topological invariant. But in order to prove it mathematically, one has to introduce the notion of the homomorphism induced by a continuous map.

Suppose that $h : X \to Y$ is a continuous map that carries the point a of X to the point a' of Y. We often denote this fact by writing

$$h : (X, a) \to (Y, a').$$

If f is a loop in X based at a, then the composite $h \circ f : I \to Y$ is a loop in Y based at a'. The correspondence $f \to h \circ f$ thus gives rise to a map carrying $\pi_1(X, a)$ into $\pi_1(Y, a')$. We define it formally as follows.

Definition 19.1.14. Let $h : (X, a) \to (Y, a')$ be a continuous map. Define

$$h_* : \pi_1(X, a) \to \pi_1(Y, a')$$

by the equation

$$h_*([f]) = [h \circ f].$$

The map h_* is called the homomorphism induced by h, relative to the base point a.

The map h_* is well-defined, for if F is a path-homotopy between the paths f and f', then $h \circ F$ is a path-homotopy between the paths $h \circ f$ and $h \circ f'$. The fact that h_* is a homomorphism follows from the equation

$$(h \circ f) * (h \circ g) = h \circ (f * g).$$

The homomorphism h_* depends not only on the map $h : X \to Y$ but also on the choice of the base point a (a' is determined by h). So some notational difficulty will arise if we consider several different base points for X.

The induced homomorphism has two properties that are crucial in the applications. They are called the functional properties and are given in the following theorem.

Theorem 19.1.15. If $h : (X, a) \to (Y, a')$ and $k : (Y, a') \to (Z, a'')$ are continuous, then $(k \circ h)_* = k_* \circ h_*$. If $i : (X, a) \to (X, a)$ is the identity map, then i_* is the identity homomorphism.

Proof. The proof is trivial. By definition,

$$(k \circ h)_*([f]) = [(k \circ h) \circ f],$$

$$\begin{aligned}(k_* \circ h_*)([f]) &= k_*(h_*([f])) \\ &= k_*([h \circ f]) \\ &= [k \circ (h \circ f)].\end{aligned}$$

Similarly,

$$i_*([f]) = [i \circ f] = [f].$$

Hence the theorem. □

Corollary 19.1.16. If $h : (X, a) \to (Y, a')$ is a homeomorphism of X with Y, then h_* is an isomorphism of $\pi_1(X, a)$ with $\pi_1(Y, a')$.

Proof. Let $k : (Y, a') \to (X, a)$ be the inverse of h. Then

$$k_* \circ h_* = (k \circ h)_* = i_*,$$

where i is the identity map of (X, a) and

$$h_* \circ k_* = (h \circ k)_* = j_*,$$

where j is the identity map of (Y, a'). Since i_* and j_* are the identity homomorphisms of the groups $\pi_1(X, a)$ and $\pi_1(Y, a')$ respectively, k_* is the inverse of h_*. □

Unit 20

Course Structure

- Calculation of fundamental group of S

20.1 Covering Spaces

Definition 20.1.1. Let $p : E \to B$ be a continuous surjective map. The open set U of B is said to be evenly covered by p if the inverse image $p^{-1}(U)$ can be written as the union of disjoint open sets V_α in E such that for each α, the restriction of p to V_α is a homeomorphism of V_α onto U. The collection $\{V_\alpha\}$ will be called a partition of $p^{-1}(U)$ into slices.

If U is an open set that is evenly covered by p, we often picture the set $p^{-1}(U)$ as a "stack of pancakes," each having the same size and shape as U, floating in the air above U; the map p squashes them all down onto U. See Figure 20.1.1. Note that if U is evenly covered by p and W is an open set contained in U, then W is also evenly covered by p.

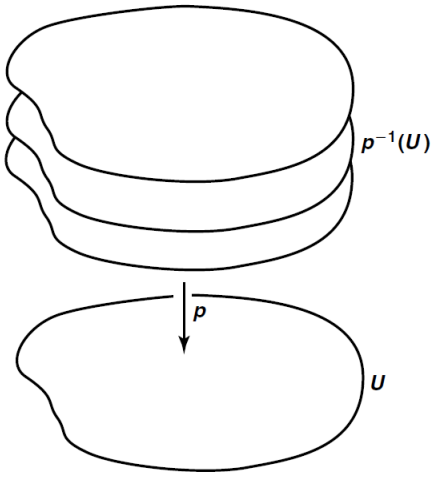


Figure 20.1.1: Slices of $p^{-1}(U)$

Definition 20.1.2. Let $p : E \to B$ be continuous and surjective. If every point b of B has a neighborhood U that is evenly covered by p, then p is called a covering map, and E is said to be a covering space of B.

Note that if $p : E \to B$ is a covering map, then for each $b \in B$ the subspace $p^{-1}(b)$ of E has the discrete topology. For each slice V_α is open in E and intersects the set $p^{-1}(b)$ in a single point; therefore, this point is open in $p^{-1}(b)$.

Note also that if $p : E \to B$ is a covering map, then p is an open map. For suppose A is an open set of E. Given $x \in p(A)$, choose a neighborhood U of x that is evenly covered by p. Let $\{V_\alpha\}$ be a partition of $p^{-1}(U)$ into slices. There is a point y of A such that $p(y) = x$; let V_β be the slice containing y. The set $V_\beta \cap A$ is open in E and hence open in V_β; because p maps V_β homeomorphically onto U, the set $p(V_\beta \cap A)$ is open in U and hence open in B; it is thus a neighborhood of x contained in $p(A)$, as desired.

Example 20.1.3. Let X be any space; let $i : X \to X$ be the identity map. Then i is a covering map (of the most trivial sort). More generally, let E be the space $X \times \{1, \ldots, n\}$ consisting of n disjoint copies of X. The map $p : E \to X$ given by $p(x, i) = x$ for all i is again a (rather trivial) covering map. In this case, we can picture the entire space E as a stack of pancakes over X.

In practice, one often restricts oneself to covering spaces that are path connected, to eliminate trivial coverings of the pancake-stack variety. An example of such a nontrivial covering space is the following:

Theorem 20.1.4. The map $p : \mathbb{R} \to S^1$ given by the equation

$$p(x) = (\cos 2\pi x, \sin 2\pi x)$$

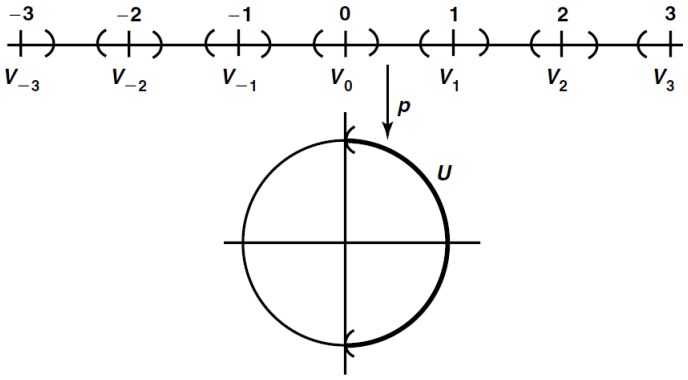
is a covering map.

One can picture p as a function that wraps the real line $\mathbb{R}$ around the circle S^1, and in the process maps each interval $[n, n + 1]$ onto S^1.

Proof. The fact that p is a covering map comes from elementary properties of the sine and cosine functions. Consider, for example, the subset U of S^f consisting of those points having positive first coordinate. The set $p^{-1}(U)$ consists of those points x for which $\cos 2\pi x$ is positive; that is, it is the union of the intervals

$$V_n = \left(n - \frac{1}{4}, n + \frac{1}{4}\right),$$

for all $n \in \mathbb{Z}$. See the figure below. Now, restricted to any closed interval $\bar{V}_n$, the map p is injective because



$\sin 2\pi x$ is strictly monotonic on such an interval. Furthermore, p carries $\bar{V}_n$ surjectively onto $\bar{U}$, and V_n to U, by the intermediate value theorem. Since $\bar{V}_n$ is compact, $p \mid V_n$ is a homeomorphism of $\bar{V}_n$ with $\bar{U}$. In

particular, $p \mid V_n$ is a homeomorphism of V_n with U. Figure 53.2 Similar arguments can be applied to the intersections of S^1 with the upper and lower open half-planes, and with the open left-hand half-plane. These open sets cover S^1, and each of them is evenly covered by p. Hence $p : \mathbb{R} \to S^1$ is a covering map. □

Example 20.1.5. The map $p : \mathbb{R}_+ \to S^1$ given by $p(x) = (\cos 2\pi x, \sin 2\pi x)$ is surjective, and it is a local homeomorphism. But it is not a covering map, for the point $b = (1, 0)$ has has no neighbourhood U that is evenly covered by p. The typical neighbourhood U of b has an inverse image consisting of small neighbourhoods V_n of each integer n for $n > 0$, alongwith the samll neighbourhood V_0 of the form $(0, \epsilon)$. Each of the intervals V_n for $n > 0$ is mapped homeomorphically onto U by the map p, but the interval V_0 is only imbedded in U by p.

The above example shows that the map obtained by restricting a covering map may not be a covering map. Here is one situation where it will be one.

Theorem 20.1.6. Let $p : E \to B$ be a covering map. If B_0 is a subspace of B, and if $E_0 = p^{-1}(B_0)$, then the map $p_0 : E_0 \to B_0$ obtained by restricting p is a covering map.

Proof. Given $b_0 \in B_0$, let U be an open set in B containing b_0 that is evenly covered by p; let $\{V_\alpha\}$ be a partition of $p^{-1}(U)$ into slices. Then $U \cap B_0$ is a neighborhood of bo in B_0, and the sets $V_\alpha \cap E_0$ are disjoint open sets in E_0 whose union is $p^{-1}(U \cap B_0)$, and each is mapped homeomorphically onto $U \cap B_0$ by p. □

Theorem 20.1.7. It $p : E \to B$ and $p' : E' \to B'$ are covering maps, then

$$p \times p' : E \times E' \to B \times B'$$

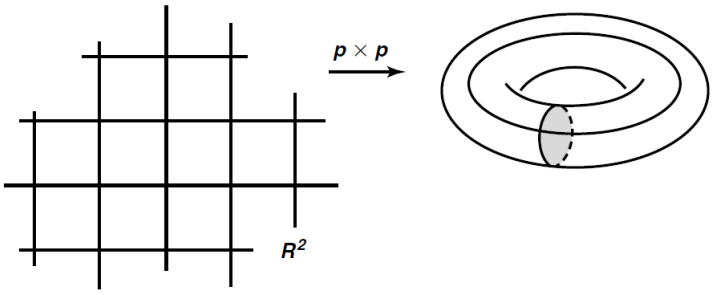
is a covering map.

Proof. Given $b \in B$ and $b' \in B'$, let U and U' be neighborhoods of b and b', respectively, that are evenly covered by p and p', respectively. Let $\{V_\alpha\}$ and $\left\{V'_\beta\right\}$ be partitions of $p^{-1}(U)$ and $(p')^{-1}(U')$, respectively, into slices. Then the inverse image under $p \times p'$ of the open set $U \times U'$ is the union of all the sets $V_\alpha \times V'_\beta$. These are disjoint open sets of $E \times E'$, and each is mapped homeomorphically onto $U \times U'$ by $p \times p'$. □

Example 20.1.8. Consider the space $T = S^1 \times S^1$; it is called the torus. The product map

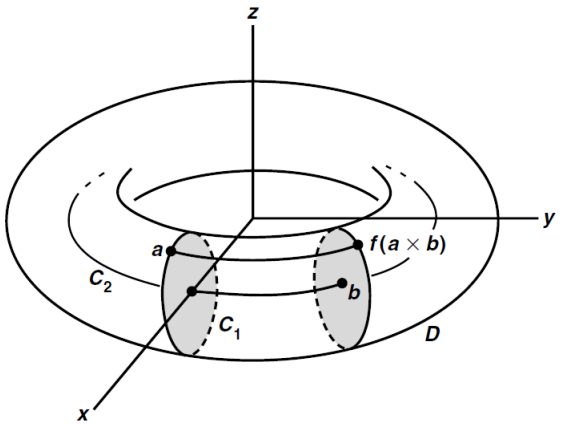
$$p \times p : \mathbb{R} \times \mathbb{R} \longrightarrow S^1 \times S^1$$

is a covering of the torus by the plane $\mathbb{R}^2$, where p denotes the covering map of Theorem 20.1.4. Each of the unit squares $[n, n+1] \times [m, m+1]$ gets wrapped by $p \times p$ entirely around the torus. See the figure below. In



this figure, we have pictured the torus not as the product $S^1 \times S^1$, which is a subspace of $\mathbb{R}^4$ and thus difficult to visualize, but as the familiar doughnut-shaped surface D in $\mathbb{R}^3$ obtained by rotating the circle C_1 in the xz-plane of radius $\frac{1}{3}$ centered at $(1, 0, 0)$ about the z-axis. It is not hard to see that $S^1 \times S^1$ is homeomorphic

with the surface D. Let C_2 be the circle of radius 1 in the xy-plane centered at the origin. Then let us map $C_1 \times C_2$ into D by defining $f(a \times b)$ to be that point into which a is carried when one rotates the circle C_1 about the z-axis until its center hits the point b. See Figure below. The map f will be a homeomorphism of $C_1 \times C_2$ with D, as you can check mentally. If you wish, you can write equations for f and check continuity, injectivity, and surjectivity directly. (Continuity of f^{-1} will follow from compactness of $C_1 \times C_2$)



20.2 Fundamental groups of the circle

Definition 20.2.1. Let $p : E \to B$ be a map. If f is a continuous mapping of some space X into B, a lifting of f is a map $\tilde{f} : X \to E$ such that $p \circ \tilde{f} = f$.

The existence of liftings when p is a covering map is an important tool in studying covering spaces and the fundamental group. First, we show that for a covering space, paths can be lifted; then we show that path homotopies can be lifted as well.

Example 20.2.2. Consider the covering $p : \mathbb{R} \to S^1$ of Theorem 20.1.4. The path $f : [0, 1] \to S^1$ beginning at $b_0 = (1, 0)$ given by $f(s) = (\cos \pi s, \sin \pi s)$ lifts to the path $\bar{f}(s) = s/2$ beginning at 0 and ending at $\frac{1}{2}$. The path $g(s) = (\cos \pi s, -\sin \pi s)$ lifts to the path $\bar{g}(s) = -s/2$ beginning at 0 and ending at $-\frac{1}{2}$. The path $h(s) = (\cos 4\pi s, \sin 4\pi s)$ lifts to the path $\bar{h}(s) = 2s$ beginning at 0 and ending at 2 . Intuitively, h wraps the interval $[0, 1]$ around the circle twice; this is reflected in the fact that the lifted path $\bar{h}$ begins at zero and ends at the number 2. These paths are pictured in the following figure.

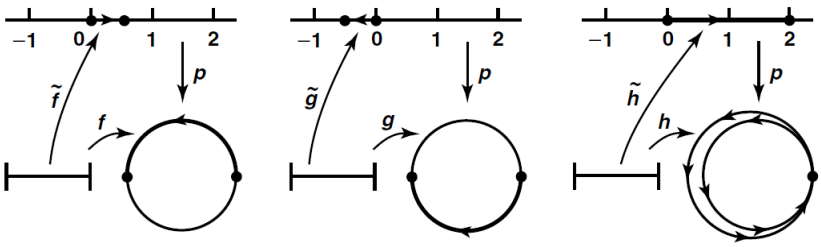


Figure 20.2.1: Caption

Lemma 20.2.3. Let $p : E \to B$ be a covering map, let $p(e_0) = b_0$. Any path $f : [0, 1] \to B$ beginning at bo has a unique lifting to a path $\bar{f}$ in E beginning at e_0.

Proof. Cover B by open sets U each of which is evenly covered by p. Find a subdivision of $[0,1]$, say $s_0, \ldots, s_n$, such that for each i the set $f([s_i, s_{i+1}])$ lies in such an open set U. (Here we use the Lebesgue number lemma.) We define the lifting $\tilde{f}$ step by step.

First, define $\tilde{f}(0) = e0$. Then, supposing $\tilde{f}(s)$ is defined for $0 \leq s \leq s_i$, we define $\tilde{f}$ on $[s_i, s_{i+1}]$ as follows: The set $f([s_i, s_{i+1}])$ lies in some open set $\mathcal{U}$ that is evenly covered by p. Let $\{V_\alpha\}$ be a partition of $p^{-1}(U)$ into slices; each set V_α is mapped homeomorphically onto U by p. Now $\tilde{f}(s_i)$ lies in one of these sets, say in V_0. Define $\tilde{f}(s)$ for $s \in [s_i, s_{i+1}]$ by the equation

$$\tilde{f}(s) = (p \mid V_0)^{-1}(f(s)).$$

Because $p \mid V_0 : V_0 \to U$ is a homeomorphism, $\tilde{f}$ will be continuous on $[s_i, s_{i+1}]$.

Continuing in this way, we define $\tilde{f}$ on all of $[0,1]$. Continuity of $\tilde{f}$ follows from the pasting lemma; the fact that $p \circ \tilde{f} = f$ is immediate from the definition of $\tilde{f}$.

The uniqueness of $\tilde{f}$ is also proved step by step. Suppose that $\tilde{\tilde{f}}$ is another lifting of f beginning at e_0. Then $\tilde{f}(0) = e_0 = \tilde{\tilde{f}}(0)$. Suppose that $\tilde{f}(s) = \tilde{\tilde{f}}(s)$ for all s such that $0 \leq s \leq s_i$. Let V_0 be as in the preceding paragraph; then for $s \in [s_i, s_{i+1}]$, $\tilde{f}(s)$ is defined as $(p \mid V_0)^{-1}(f(s))$. What can $\tilde{\tilde{f}}(s)$ equal? Since $\tilde{\tilde{f}}$ is a lifting of f, it must carry the interval $[s_i, s_{i+1}]$ into the set $p^{-1}(U) = \bigcup V_\alpha$. The slices V_α are open and disjoint; because the set $\tilde{f}([s_i, s_{i+1}])$ is connected, it must lie entirely in one of the sets V_α. Because $\tilde{\tilde{f}}(s_i) = \tilde{f}(s_i)$, which is in V_0, $\tilde{\tilde{f}}$ must carry all of $[s_i, s_{i+1}]$ into the set V_0. Thus, for s in $[s_i, s_{i+1}]$, $\tilde{f}(s)$ must equal some point y of V_0 lying in $p^{-1}(f(s))$. But there is only one such point y, namely, $(p \mid V_0)^{-1}(f(s))$. Hence $\tilde{\tilde{f}}(s) = \tilde{f}(s)$ for $s \in [s_i, s_{i+1}]$. □

Lemma 20.2.4. Let $p : E \to B$ be a covering map; let $p(e_0) = b_0$. Let the map $F : I \times I \to B$ be continuous, with $F(0,0) = b_0$. There is a unique lifting of F to a continuous map

$$\tilde{F} : I \times I \to E$$

such that $\tilde{F}(0,0) = e_0$. If F is a path homotopy, then $\tilde{F}$ is a path homotopy.

Proof of the lemma has been intentionally left.

Theorem 20.2.5. Let $p : E \to B$ be a covering map; let $p(e_0) = b_0$. Let f and g be two paths in B from b_0 to b_1; let $\tilde{f}$ and $\tilde{g}$ be thir respective liftings to paths in E beginning at e_0. If f and g are path homotopic, then $\tilde{f}$ and $\tilde{g}$ end at the same point of E and are path homotopic.

Proof. Let $F : I \times I \to B$ be the path homotopy between f and g. Then $F(0,0) = b_0$. Let $\tilde{F} : I \times I \to E$ be the lifting of F to E such that $\tilde{F}(0,0) = e_0$. By the preceding lemma, $\tilde{F}$ is a path homotopy, so that $\tilde{F}(0 \times I) = \{e_0\}$ and $\tilde{F}(1 \times I)$ is a one-point set $\{e_1\}$.

The restriction $\tilde{F} \mid I \times 0$ of $\tilde{F}$ to the bottom edge of $I \times I$ is a path on E beginning at e_0 that is a lifting of $F \mid I \times 0$. By uniqueness of path liftings, we must have $\tilde{F}(s,0) = \tilde{f}(s)$. Similarly, $\tilde{F} \mid I \times 1$ is a path on E that is a lifting of $F \mid I \times 1$, and it begins at e_0 because $\tilde{F}(0 \times I) = \{e_0\}$. By uniqueness of path liftings, $\tilde{F}(s,1) = \tilde{g}(s)$. Therefore, both $\tilde{f}$ and $\tilde{g}$ end at e_1, and $\tilde{F}$ is a path homotopy between them. □

Definition 20.2.6. Let $p : E \to B$ be a covering map; let $b_0 \in B$. Choose e_0 so that $p(e_0) = b_0$. Given an element $[f]$ of $\pi_1(B, b_0)$, let $\tilde{f}$ be the lifting of f to a path in E that begins at e_0. Let $\phi([f])$ denote the end point $\tilde{f}(1)$ of $\tilde{f}$. Then ϕ is a well-defined set map

$$\phi : \pi_1(B, b_0) \to p^{-1}(b_0).$$

We call ϕ the lifting correspondence derived from the covering map p. It depends of course on the choice of the point e_0.

Theorem 20.2.7. Let $p : E \to B$ be a covering map; let $p(e_0) = b_0$. If E is path connected, then the lifting correspondence

$$\phi : \pi_1(B, b_0) \to p^{-1}(b_0)$$

is surjective. If E is simply connected, it is bijective.

Proof. If E is path connected, then, given $e_1 \in p^{-1}(b_0)$, there is a path $\tilde{f}$ in E from e_0 to e_1. Then $f = p \circ \bar{f}$ is a loop in B at b_0, and $\phi([f]) = e_1$ by definition.

Suppose E is simply connected. Let $[f]$ and $[g]$ be two elements of $\pi_1(B, b_0)$ such that $\phi([f]) = \phi([g])$. Let $\tilde{f}$ and $\tilde{g}$ be the liftings of f and g, respectively, to paths in E that begin at e_0; then $\tilde{f}(1) = \tilde{g}(1)$. Since E is simply connected, there is a path homotopy $\bar{F}$ in E between $\bar{f}$ and $\tilde{g}$. Then $p \circ \bar{F}$ is a path homotopy in B between f and g. □

Theorem 20.2.8. The fundamental group of S^1 is isomorphic to the additive group of integers.

Proof. Let $p : \mathbb{R} \to S^1$ be the covering map of Theorem 20.1.4, let $e_0 = 0$, and let $b_0 = p(e_0)$. Then $p^{-1}(b_0)$ is the set $\mathbb{Z}$ of integers. Since $\mathbb{R}$ is simply connected, the lifting correspondence

$$\phi : \pi_1\left(S^1, b_0\right) \to \mathbb{Z}$$

is bijective. We show that ϕ is a homomorphism, and the theorem is proved.

Given $[f]$ and $[g]$ in $\pi_1(B, b_0)$, let $\tilde{f}$ and $\tilde{g}$ be their respective liftings to paths on $\mathbb{R}$ beginning at 0 . Let $n = \tilde{f}(1)$ and $m = \tilde{g}(1)$; then $\phi([f]) = n$ and $\phi([g]) = m$, by definition. Let $\tilde{g}$ be the path

$$\tilde{g}(s) = n + \tilde{g}(s)$$

on $\mathbb{R}$. Because $p(n + x) = p(x)$ for all $x \in \mathbb{R}$, the path $\tilde{g}$ is a lifting of g; it begins at n. Then the product $\tilde{f} * \tilde{g}$ is defined, and it is the lifting of $f * g$ that begins at 0 , as you can check. The end point of this path is $\tilde{g}(1) = n + m$. Then by definition,

$$\phi([f] * [g]) = n + m = \phi([f]) + \phi([g])$$

□

References

1. *Elementery Differential Geometry*, Andrew Pressley.
2. *Differential geometry of manifolds*, U. C. De and A. A. Shaikh, Narosa Pub.
3. *Introduction to manifolds*, J. M. Lee
4. *Differential Manifolds*, J. M. Lee.
5. *Differential Geometry*, J. M. Lee.
6. *Introduction to Toplogy and Modern Analysis*, G.F Simmons
7. *Topology*, James. R. Munkres
8. *Introduction to Topological Manifolds*, John. M. Lee
9. *Theory and Problems of General Topology*, Seymour Lipschutz

POST GRADUATE DEGREE PROGRAMME (CBCS)

M.SC. IN MATHEMATICS

SEMESTER II

SELF LEARNING MATERIAL

PAPER : GEC 2.5
(CBCS Paper)

History of Mathematics

Operations Research
Matrices and Linear Algebra
Theory of Dynamical Systems

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

Content Writers

Block - I : History of Mathematics	Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani
Block - II : Operations Research	Dr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani
Block - III : Matrices and Linear Algebra	Ms. Audrija Choudhury Assistant Professor (Cont.) DODL, University of Kalyani
Block - IV : Theory of Dynamical Systems	Dr. Biswajit Mallick Assistant Professor (Cont.) DODL, University of Kalyani

July, 2022

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 Message

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, 2020 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.) Manas Kumar Sanyal, 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 well 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 coordinated 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 written and collected from e-book, journals and websites.

Director

Directorate of Open and Distance Learning

University of Kalyani

Board of Studies Members of Department of Mathematics, Directorate of Open and Distance Learning (DODL), University of Kalyani

Sl No.	Name & Designation	Role
1	Dr. Animesh Biswas, Professor & Head, Dept. of Mathematics, KU	Chairperson
2	Dr. Pulak Sahoo, Professor, Dept. of Mathematics, KU	Member
3	Dr. Sahidul Islam, Assistant Professor, Dept. of Mathematics, KU	Member
4	Dr. Sushanta Kumar Mohanta, Professor, Dept. of Mathematics, West Bengal State University	External Nominated Member
5	Dr. Biswajit Mallick, Assistant Professor (Cont.), Department of Mathematics, DODL, KU	Member
6	Ms. Audrija Choudhury, Assistant Professor (Cont), Department of Mathematics, DODL, KU	Member
7	Director, DODL, KU	Convener

CBCS Paper

OTHER DEPARTMENTS

GEC 2.5
Marks : 100 (SEE : 80; IA : 20); Credit : 4

History of Mathematics (Marks 25 (SEE: 20; IA: 5))
Operations Research (Marks 25 (SEE: 20; IA: 5))
Matrices and Linear Algebra (Marks 25 (SEE: 20; IA: 5))
Theory of Dynamical Systems (Marks 25 (SEE: 20; IA: 5))

Syllabus

Block I

- **Unit 1:** Babylonian and Egyptian mathematics, Greek mathematics, Pythagoras, Euclid and the elements of geometry, Archimedes, Apollonius
- **Unit 2:** Development of Trigonometry, Development of Algebra, Development of Analytic Geometry
- **Unit 3:** Development of Calculus, Development of Selected Topics of Modern Mathematics.
- **Unit 4:** Development of Modern geometries, Modern algebra, Methods of real analysis.

Block II

- **Unit 5:** Formulation of linear programming models. Graphicalsolution. Basic solution (BS) and Basic Feasible Solution (BFS), Degenerate and non-degenerate BFS, Convex set, convex hull, convex polyhedron, extreme points, hyperplane.
- **Unit 6:** Standard form of LPP. Simplex method. Charnes' Big – M method.
- **Unit 7:** Transportation and assignment problems.
- **Unit 8:** A brief introduction to PERT and CPM, Components of PERT/CPM Network and precedence relationships, Critical path analysis.

Block III

- **Unit 9:** Matrix: definition, order, symmetric and skew symmetric matrices.
- **Unit 10:** Determinant of a matrix, elementary properties of determinants, inverse of a matrix, normal form of a matrix, rank of a matrix.
- **Unit 11:** Elementary concept of a vector space, linear dependence and independence of vectors, basis of a vector space, row space, column space, solution of system of linear equations, Cramar's rule.
- **Unit 12:** Eigen values and Eigen vectors of matrices, Cayley Hamilton Theorem, Diagonalization of matrices.

Block IV

- **Unit 13:** Linearization of dynamical systems: Two, three and higher dimension. Population growth. Lotka-Volterra system.
- **Unit 14:** Stability: Asymptotic stability (Hartman's theorem), Global stability (Liapunov's second method).
- **Unit 15:** Limit set, attractors, periodic orbits, limit cycles. Bendixon criterion, Dulac criterion, Poincare-Bendixon Theorem. Floquet's theorem.
- **Unit 16:** Stability and bifurcation: Routh-Hurwitz criterion for nonlinear systems. Saddle-Node, trans-critical and pitchfork bifurcations. Hopf- bifurcation.

Unit 1

Course Structure

- Introduction
- Babylonian and Egyptian mathematics
- Greek mathematics
- Pythagoras, Euclid and the elements of geometry
- Archimedes, Apollonius

Introduction

The area of study known as the history of mathematics is primarily an investigation into the origin of discoveries in mathematics and, to a lesser extent, an investigation into the mathematical methods and notation of the past. Before the modern age and the worldwide spread of knowledge, written examples of new mathematical developments have come to light only in a few locales. From 3000 BC the Mesopotamian states of Sumer, Akkad and Assyria, together with Ancient Egypt and Ebla began using arithmetic, algebra and geometry for purposes of taxation, commerce, trade and also in the field of astronomy and to formulate calendars and record time.

Babylonian and Egyptian mathematics

Our first knowledge of mankind's use of mathematics comes from the Egyptians and Babylonians. Both civilizations developed mathematics that was similar in scope but different in particulars. There can be no denying the fact that the totality of their mathematics was profoundly elementary, but their astronomy of later times did achieve a level comparable to the Greeks.

Babylonian Mathematics

The mathematics developed and practised by the people of Mesopotamia from the days of the early Sumerians to the fall of Babylon in 539 BCE. The Babylonian Mathematics can be categorised into two: one of the Old Babylonian Period (1830-1531 BCE) and the other mainly Seleucid from the last three of four centuries BCE.

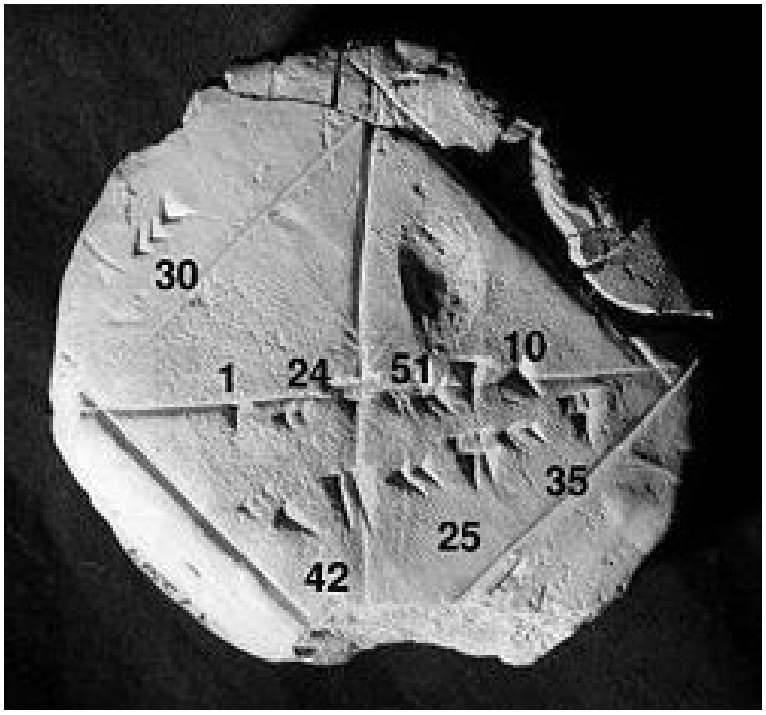


Figure 1.0.1: Babylonian clay tablet with annotations

The Babylonians were somewhat more advanced than the Egyptians in mathematics. The main features of the Babylonian Mathematics were:

- Their mathematical notation was positional but sexagesimal. From this we derive the modern day usage of 60 seconds in a minute, 60 minutes in an hour, and 360 degrees in a circle.
- They used no zero.
- More general fractions, though not all fractions, were admitted.
- They could extract square roots and solve linear systems.
- The Pythagorean triples were frequently used.
- They solved cubic equations with the help of tables.
- They studied circular measurement.
- They studied circular measurement.

Most clay tablets that describe Babylonian mathematics belong to the Old Babylonian, which is why the mathematics of Mesopotamia is commonly known as Babylonian mathematics. Some clay tablets contain mathematical lists and tables, others contain problems and worked solutions.

For enumeration the Babylonians used symbols for 1, 10, 60, 600, 3,600, 36,000, and 216,000, similar to the

earlier period. Below are some of the symbols. They did arithmetic in base 60, sexagesimal.

There is no clear reason why the Babylonians selected the sexagesimal system6 . It was possibly selected in

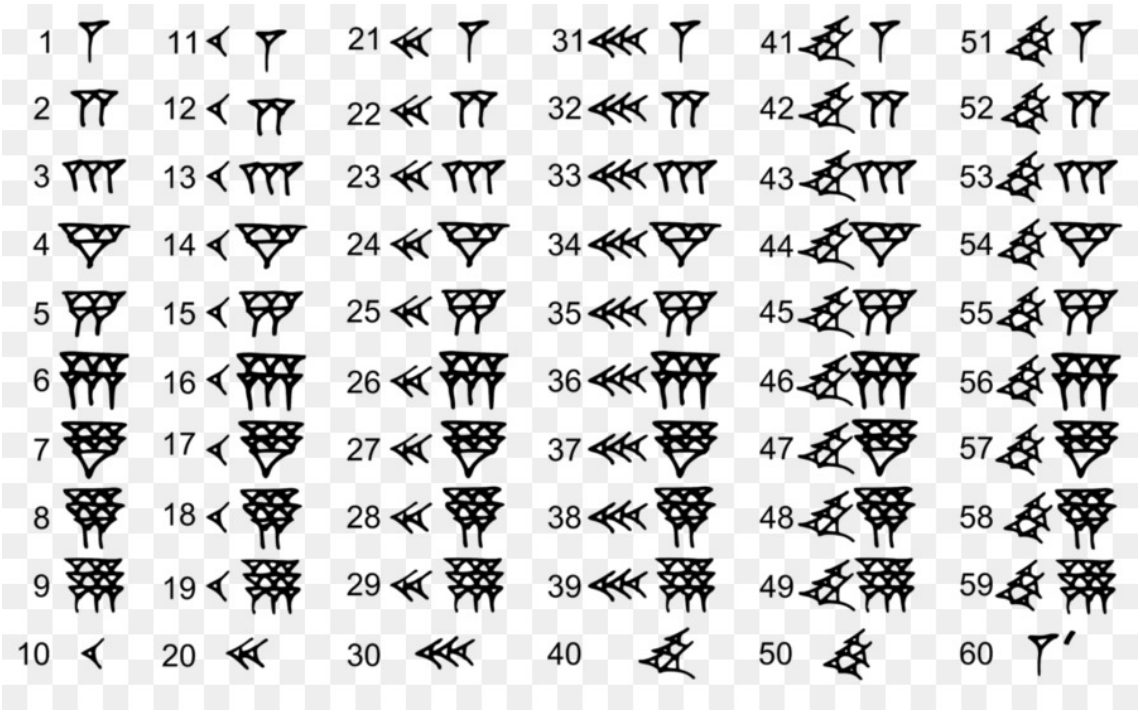


Figure 1.0.2: Babylonian numerals

the interest of metrology, this according to Theon of Alexandria, a commentator of the fourth century A.D.: i.e. the values 2,3,5,10,12,15,20, and 30 all divide 60. Remnants still exist today with time and angular measurement. However, a number of theories have been posited for the Babylonians choosing the base of 60. For example

1. The number of days, 360, in a year gave rise to the subdivision of the circle into 360 degrees, and that the chord of one sixth of a circle is equal to the radius gave rise to a natural division of the circle into six equal parts. This in turn made 60 a natural unit of counting. (Moritz Cantor, 1880)

2. The Babylonians used a 12 hour clock, with 60 minute hours. That is, two of our minutes is one minute for the Babylonians. (Lehmann-Haupt, 1889) Moreover, the (Mesopotamian) zodiac was divided into twelve equal sectors of 30 degrees each.

3. The number 60 is the product of the number of planets (5 known at the time) by the number of months in the year, 12. (D. J. Boorstin, 6Recall, the very early use of the sexagesimal system in China. There may well be a connection. 7See Georges Ifrah, The Universal History of Numbers, Wiley, New York, 2000. Babylonian Mathematics 8 1986)

Because of the large base, arithmetic was carried out with the aide of a pre-calculated table. For example, two tablets found at Senkerah on the Euphrates in 1854, dating from 2000 BC, give lists of the squares of numbers up to 59 and the cubes of numbers up to 32. The Babylonians used the lists of squares together with

Figure 1.0.3: Rhind Mathematical Papyrus

the formulae

$$\begin{aligned} ab &= \frac{(a+b)^2 - a^2 - b^2}{2} \\ ab &= \frac{(a+b)^2 - a^2 - b^2}{2} \\ ab &= \frac{(a+b)^2 - (a-b)^2}{4} \\ ab &= \frac{(a+b)^2 - (a-b)^2}{4} \end{aligned}$$

to simplify multiplication. The Babylonians did not have an algorithm for long division. Instead, they based their method on the fact that

$$\frac{a}{b} = a \times \frac{1}{b}$$

together with a table of reciprocals. Numbers whose only prime factors are 2, 3 or 5 (known as 5-smooth or regular numbers) have finite reciprocals in sexagesimal notation, and tables with extensive lists of these reciprocals have been found.

Egyptian Mathematics

Our sources of Egyptian mathematics are scarce. Indeed, much of our knowledge of ancient Egyptian mathematics comes not from the hieroglyphics (carved sacred letters or sacred letters) inscribed on the hundreds of temples but from two papyri containing collections of mathematical problems with their solutions.

1. The Rhind Mathematical Papyrus named for A.H. Rhind (1833- 1863) who purchased it at Luxor in 1858. Origin: 1650 BCE but it was written very much earlier. It is 18 feet long and 13 inches wide. It is also called the Ahmes Papyrus after the scribe that last copied it.

2. The Moscow Mathematical Papyrus purchased by V. S. Golenishchev (d. 1947). Origin: 1700 BC. It is 15 ft long and 3 inches wide. Two sections of this chapter offer highlights from these papyri.

The Egyptian counting system was decimal. Though non positional, it could deal with numbers of great scale. Yet, there is no apparent way to construct numbers arbitrarily large. The number system was decimal

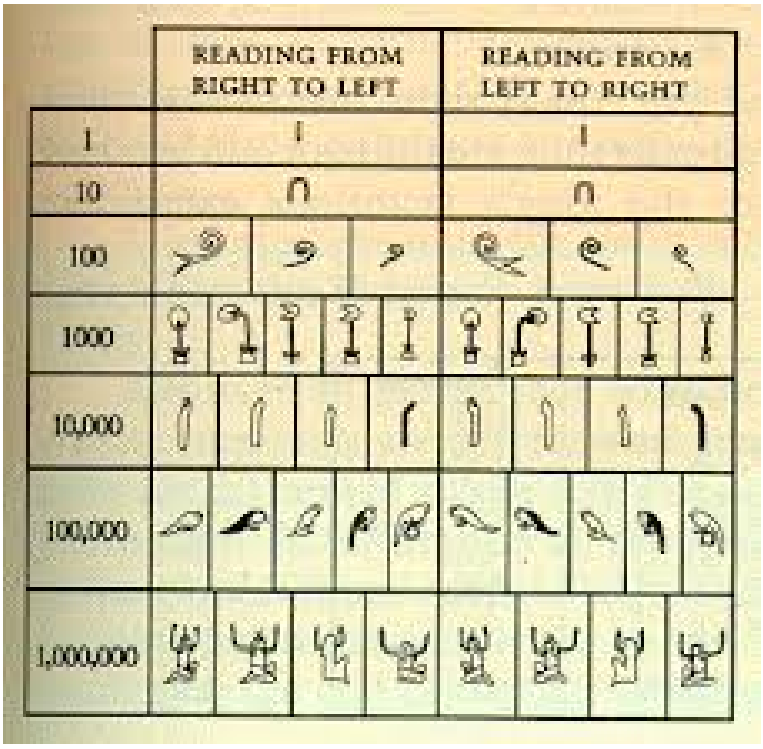


Figure 1.0.4: Mathematical Notations in Hieroglyphics

with special symbols for 1, 10, 100, 1,000, 10,000, 100,000, and 1,000,000. Addition was accomplished by grouping and regrouping. Multiplication and division were essentially based on binary multiples. Fractions were ubiquitous but only unit fractions, with two exceptions, were allowed. All other fractions were required to be written as a sum of unit fractions. Geometry was limited to areas, volumes, and similarity. Curiously, though, volume measures for the fractional portions of the hekat a volume measuring about 4.8 liters, were symbolically expressed differently from others. Simple algebraic equations were solvable, even systems of equations in two dimensions could be solved.

Greek mathematics

Greek mathematics refers to mathematics texts and advances written in Greek, developed from the 7th century BC to the 4th century AD around the shores of the Eastern Mediterranean. Our knowledge of Greek Mathematics is less reliable than our than that of the older Egyptian and Babylonian mathematics, because none of the original manuscripts are extant. There are two sources:

1. Byzantine Greek codices (manuscript books) written 500-1500 years after the Greek works were composed.

2. Arabic translations of Greek works and Latin translations of the Arabic versions.

Historians traditionally place the beginning of Greek mathematics proper to the age of Thales of Miletus (ca. 624–548 BC). The two earliest mathematical theorems, Thales' theorem and Intercept theorem are attributed to Thales. The former, which states that an angle inscribed in a semicircle is a right angle, may have been

1	α	alpha	10	ι	iota	100	ρ	rho
2	β	beta	20	κ	kappa	200	σ	sigma
3	γ	gamma	30	λ	lambda	300	τ	tau
4	δ	delta	40	μ	mu	400	υ	upsilon
5	ϵ	epsilon	50	ν	nu	500	ϕ	phi
6	ς	vau*	60	ξ	xi	600	χ	chi
7	ζ	zeta	70	o	omicron	700	ψ	psi
8	η	eta	80	π	pi	800	ω	omega
9	θ	theta	90	φ	koppa*	900	λ	sampi

*vau, koppa, and sampi are obsolete characters

Figure 1.0.5: Greek Numerals

learned by Thales while in Babylon but tradition attributes to Thales a demonstration of the theorem. It is for this reason that Thales is often hailed as the father of the deductive organization of mathematics and as the first true mathematician. It is also known that within two hundred years of Thales the Greeks had introduced logical structure and the idea of proof into mathematics.

Another important figure in the development of Greek mathematics is Pythagoras of Samos (ca. 580–500 BC) who had established an order called the Pythagoreans, which held knowledge and property in common and hence all of the discoveries by individual Pythagoreans were attributed to the order. And since in antiquity it was customary to give all credit to the master, Pythagoras himself was given credit for the discoveries made by his order.

Thales is supposed to have used geometry to solve problems such as calculating the height of pyramids based on the length of shadows, and the distance of ships from the shore. He is also credited by tradition with having made the first proof of two geometric theorems—the "Theorem of Thales" and the "Intercept theorem" described above. Pythagoras is widely credited with recognizing the mathematical basis of musical harmony and, according to Proclus' commentary on Euclid, he discovered the theory of proportionals and constructed regular solids. he Pythagoreans regarded numerology and geometry as fundamental to understanding the nature of the universe and therefore central to their philosophical and religious ideas. They are credited with numerous mathematical advances, such as the discovery of irrational numbers.

Euclid and the elements of geometry

Euclid, (born c. 300 BCE, Alexandria, Egypt), the most prominent mathematician of Greco-Roman antiquity, best known for his treatise on geometry, the **Elements**. Euclid compiled his Elements from a number of works of earlier men. Among these are Hippocrates of Chios (flourished c. 440 BCE), not to be confused with the physician Hippocrates of Cos (c. 460–375 BCE). The latest compiler before Euclid was Theudius, whose textbook was used in the Academy and was probably the one used by Aristotle (384–322 BCE).

The older elements were at once superseded by Euclid's and then forgotten. For his subject matter Euclid drew upon all his predecessors, but it is clear that the whole design of his work was his own, culminating in the construction of the five regular solids, now known as the Platonic solids. A brief survey of the Elements belies a common belief that it concerns only geometry. This misconception may be caused by reading no further than

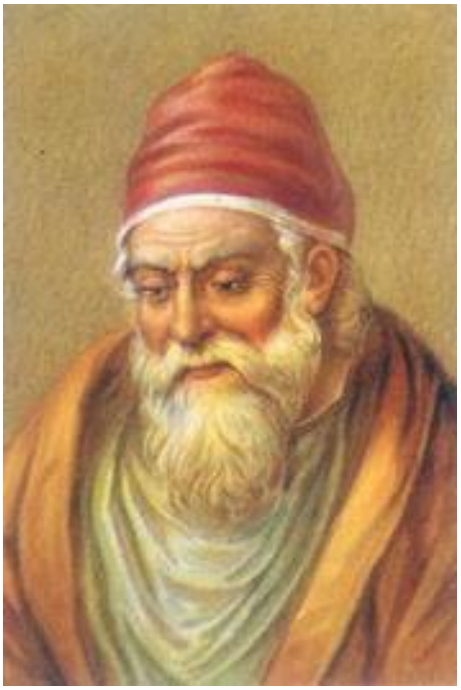


Figure 1.0.6: Euclid

Books I through IV, which cover elementary plane geometry. Euclid understood that building a logical and rigorous geometry (and mathematics) depends on the foundation—a foundation that Euclid began in Book I with 23 definitions (such as "a point is that which has no part" and "a line is a length without breadth"), five unproved assumptions that Euclid called postulates (now known as axioms), and five further unproved assumptions that he called common notions. (See the table of Euclid's 10 initial assumptions.) Book I then proves elementary theorems about triangles and parallelograms and ends with the Pythagorean theorem. (For Euclid's proof of the theorem, see Sidebar: Euclid's Windmill Proof.)

Archimedes, Apollonius

Archimedes was a Greek mathematician, philosopher and inventor who wrote important works on geometry, arithmetic and mechanics. In mechanics he defined the principle of the lever and is credited with inventing the compound pulley and the hydraulic screw for raising water from a lower to higher level. He is most famous for discovering the law of hydrostatics, sometimes known as 'Archimedes' principle', stating that a body immersed in fluid loses weight equal to the weight of the amount of fluid it displaces. During the Roman conquest of Sicily in 214 BC Archimedes worked for the state, and several of his mechanical devices were employed in the defence of Syracuse. Among the war machines attributed to him are the catapult and - perhaps legendary - a mirror system for focusing the sun's rays on the invaders' boats and igniting them. After Syracuse was captured, Archimedes was killed by a Roman soldier.

Apollonius of Perga (late 3rd – early 2nd centuries BC) was a Greek geometer and astronomer known for his theories on the topic of conic sections. Beginning from the theories of Euclid and Archimedes on the topic, he brought them to the state they were in just before the invention of analytic geometry. His definitions of the terms ellipse, parabola, and hyperbola are the ones in use today.

Apollonius worked on many other topics, including astronomy. Most of the work has not survived except in

fragmentary references in other authors. His hypothesis of eccentric orbits to explain the apparently aberrant motion of the planets, commonly believed until the Middle Ages, was superseded during the Renaissance.

Unit 2

Course Structure

- Introduction
- Development of Trigonometry
- Development of Algebra
- Development of Analytic Geometry

Introduction

The development of mathematics is intimately interwoven with the progress of civilization, influencing the course of history through its application to science and technology. Mathematics has changed with each step of advancement in civilizations. Even the mathematics of the 1800s can seem quite strange now, so greatly has mathematics evolved in the past 100 years and so thoroughly has it been reworked in the post-modern approach.

Despite its arcane appearance from the outside looking in, the present, abstract and highly specialized state of mathematics is the natural evolution of the subject, and there is much ahead that is exciting.

Development of Trigonometry

It began as a branch of geometry and was utilized extensively by early Greek mathematicians to determine unknown distances. The most notable examples are the use by Aristarchus (310-250 B.C.) to determine the distance to the Moon and Sun, and by Eratosthenes (c. 276-195 B.C.) to calculate the Earth's circumference. The general principles of trigonometry were formulated by the Greek astronomer, Hipparchus of Nicaea (162-127 B.C.), who is generally credited as the founder of trigonometry. His ideas were worked out by Ptolemy of Alexandria (A.D. c. 90-168), who used them to develop the influential Ptolemaic theory of astronomy. Much of the information we know about the work of Hipparchus and Ptolemy comes from Ptolemy's compendium, The Almagest, written around 150.

Trigonometry was initially considered a field of the science of astronomy. It was later established as a separate branch of mathematics—largely through the work of the mathematicians Johann Bernoulli (1667-1748) and Leonhard Euler (1707-1783). The major trigonometric functions, including sine, cosine, and tangent,

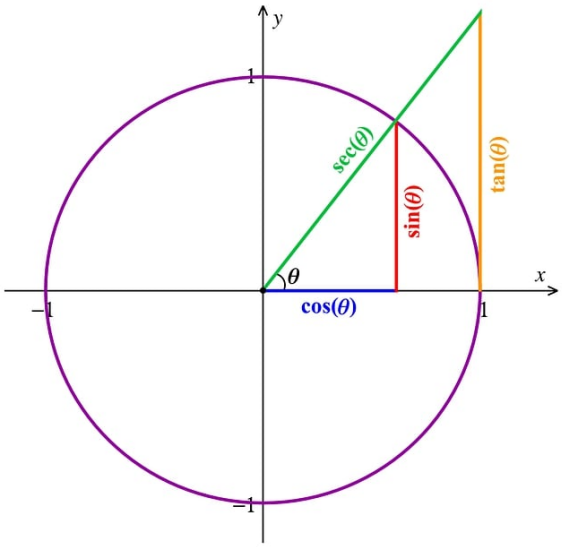


Figure 2.0.1: Trigonometric Functions Using Circle

were first defined as ratios of sides in a right triangle. Since trigonometric functions are intrinsically related, they can be used to determine the dimensions of any triangle given limited information. In the eighteenth century, the definitions of trigonometric functions were broadened by being defined as points on a unit circle. This allowed the development of graphs of functions related to the angles they represent which were periodic. Today, using the periodic nature of trigonometric functions, mathematicians and scientists have developed mathematical models to predict many natural periodic phenomena. Ancient Greek and Hellenistic mathematicians made use of the chord. Given a circle and an arc on the circle, the chord is the line that subtends the arc. A chord's perpendicular bisector passes through the center of the circle and bisects the angle. One half of the bisected chord is the sine of one half the bisected angle, that is,

$$\text{chord}\,\theta = 2\sin\frac{\theta}{2},\ \text{chord}\,\theta = 2\sin\frac{\theta}{2},$$

and consequently the sine function is also known as the half-chord. Due to this relationship, a number of trigonometric identities and theorems that are known today were also known to Hellenistic mathematicians, but in their equivalent chord form.

Some of the early and very significant developments of trigonometry were in India. Influential works from the 4th–5th century, known as the *Siddhantas* (of which there were five, the most important of which is the *Surya Siddhanta*) first defined the sine as the modern relationship between half an angle and half a chord, while also defining the cosine, versine, and inverse sine. Soon afterwards, another Indian mathematician and astronomer, Aryabhatta (476–550 AD), collected and expanded upon the developments of the Siddhantas in an important work called the Aryabhattya. The Siddhantas and the Aryabhattya contain the earliest surviving tables of sine values and versine values, to an accuracy of 4 decimal places. They used the words jya for sine, kojya for cosine, utkrama-jya for versine, and otkram jya for inverse sine. The words jya and kojya eventually became sine and cosine respectively after a mistranslation described above. Madhava (c. 1400) made early strides in the analysis of trigonometric functions and their infinite series expansions. He developed the concepts of the power series and Taylor series, and produced the power series expansions of sine, cosine, tangent, and

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

$$\sin x^2 = x^2 - \frac{\left(x^2\right)^3}{3!} + \frac{\left(x^2\right)^5}{5!} - \frac{\left(x^2\right)^7}{7!} + \ldots$$

$$\sin x^2 = x^2 - \frac{x^6}{3!} + \frac{x^{10}}{5!} - \frac{x^{14}}{7!} + \ldots$$

arctangent. Using the Taylor series approximations of sine and cosine, he produced a sine table to 12 decimal places of accuracy and a cosine table to 9 decimal places of accuracy. His works were expanded by his followers at the Kerala School up to the 16th century.

Development of Algebra

The roots of the word, algebra, can be found in Medieval Latin and the Arabic word, *al- jabr*, which means 'the reduction'. The word was first used as early as 1551. The history of algebra can be divided into three parts:

- The rhetorical, or written stage, where only words were used in equations
- The syncopated, or shortened, stage, where some symbols were used in equations
- The symbolic or modern stage

Early works of algebra of ancient Babylonians and Egyptians lack the abstract notation that algebra has today. The Babylonians had methods of solving quadratic equations, while the Egyptians used the symbol heap for the unknown.

In China, a treatise called Nine Chapters was compiled in the first century CE composed of 246 problems. The text shows methods of solving determinate and indeterminate equations. More sophisticated than the works of the Babylonians and Egyptians, the treatise is mostly what is known today as rhetorical algebra—problems and solutions are expressed in words rather than in algebraic notations.

Algebra to the ancient Greeks is an unknown science except for the Greek mathematician Diophantus of Alexandria (3rd century BCE). His work Arithmetica contains the first suggestions of algebraic notations and is probably the earliest treatise on algebra. He used algebraic equations and notations in presenting problems and solutions in Arithmetica.

In the 6th century CE in India, Aryabhata's works show knowledge in summing an arithmetic series, and solving quadratic equations and indeterminate linear equations. Shortly after, in the 7th century, Brahmagupta applied algebra to astronomy; his works gives the rules for using negative numbers, and solving quadratic

Figure 2.0.2: Khwarzimi

equations.

The Islamic scholars made several contributions to algebra as well—most notable of them is al-Khwārizmī. His treatise A short book on the calculus al-jabr and al-muqabalah, deals with the solution of quadratic equations. He did not use algebraic notations in his treatise but employed rhetorical algebra. This is one reason why some consider Diophantus as the "Father of Algebra" rather than Al-Khwārizmī.

Development of Analytic Geometry

Analytic geometry, also called coordinate geometry, mathematical subject in which algebraic symbolism and methods are used to represent and solve problems in geometry. Apollonius of Perga foreshadowed the development of analytic geometry by more than 1,800 years with his book **Conics**. He defined a conic as the intersection of a cone and a plane (see figure).

Further development of coordinate systems in mathematics emerged only after algebra had matured under Islamic and Indian mathematicians. With the power of algebraic notation, mathematicians were no longer completely dependent upon geometric figures and geometric intuition to solve problems. The more daring began to leave behind the standard geometric way of thinking in which linear (first power) variables corresponded to lengths, squares (second power) to areas, and cubics (third power) to volumes, with higher powers lacking "physical" interpretation. Two Frenchmen, the mathematician-philosopher René Descartes and the lawyer-mathematician Pierre de Fermat, were among the first to take this daring step.

Descartes and Fermat independently founded analytic geometry in the 1630s by adapting Viète's algebra to the study of geometric loci. They used letters to represent distances that are variable instead of fixed. Descartes used equations to study curves defined geometrically, and he stressed the need to consider general

algebraic curves—graphs of polynomial equations in x and y of all degrees. He demonstrated his method on a classical problem: finding all points P such that the product of the distances from P to certain lines equals the product of the distances to other lines.

Fermat emphasized that any relation between x and y coordinates determines a curve. Using this idea, he recast Apollonius's arguments in algebraic terms and restored lost work. Fermat indicated that any quadratic equation in x and y can be put into the standard form of one of the conic sections.

Their ideas gained general acceptance only through the efforts of other mathematicians in the latter half of the 17th century. In particular, the Dutch mathematician Frans van Schooten translated Descartes's writings from French to Latin. He added vital explanatory material, as did the French lawyer Florimond de Beaune, and the Dutch mathematician Johan de Witt. In England, the mathematician John Wallis popularized analytic geometry, using equations to define conics and derive their properties. He used negative coordinates freely, although it was Isaac Newton who unequivocally used two (oblique) axes to divide the plane into four quadrants.

Unit 3

Course Structure

- Development of Calculus
- Development of Selected Topics of Modern Mathematics

Development of Calculus

The discovery of calculus is often attributed to two men, Isaac Newton and Gottfried Leibniz, who independently developed its foundations. Although they both were instrumental in its creation, they thought of the fundamental concepts in very different ways. While Newton considered variables changing with time, Leibniz thought of the variables x and y as ranging over sequences of infinitely close values. He introduced dx and dy as differences between successive values of these sequences. Leibniz knew that dy/dx gives the tangent but he did not use it as a defining property. On the other hand, Newton used quantities x' and y', which were finite velocities, to compute the tangent. Of course neither Leibniz nor Newton thought in terms of functions, but both always thought in terms of graphs. For Newton the calculus was geometrical while Leibniz took it towards analysis. It is interesting to note that Leibniz was very conscious of the importance of good notation and put a lot of thought into the symbols he used. Newton, on the other hand, wrote more for himself than anyone else. Consequently, he tended to use whatever notation he thought of on that day. This turned out to be important in later developments. Leibniz's notation was better suited to generalizing calculus to multiple variables and in addition it highlighted the operator aspect of the derivative and integral. As a result, much of the notation that is used in Calculus today is due to Leibniz.

The development of Calculus can roughly be described along a timeline which goes through three periods: Anticipation, Development, and Rigorization. In the Anticipation stage techniques were being used by mathematicians that involved infinite processes to find areas under curves or maximaize certain quantities. In the Development stage Newton and Leibniz created the foundations of Calculus and brought all of these techniques together under the umbrella of the derivative and integral. However, their methods were not always logically sound, and it took mathematicians a long time during the Rigorization stage to justify them and put Calculus on a sound mathematical foundation.

In their development of the calculus both Newton and Leibniz used "infinitesimals", quantities that are infinitely small and yet nonzero. Of course, such infinitesimals do not really exist, but Newton and Leibniz found it convenient to use these quantities in their computations and their derivations of results. Although one could not argue with the success of calculus, this concept of infinitesimals bothered mathematicians. Lord

Bishop Berkeley made serious criticisms of the calculus referring to infinitesimals as "the ghosts of departed quantities".

Berkeley's criticisms were well founded and important in that they focused the attention of mathematicians on a logical clarification of the calculus. It was to be over 100 years, however, before Calculus was to be made rigorous. Ultimately, Cauchy, Weierstrass, and Riemann reformulated Calculus in terms of limits rather than infinitesimals. Thus the need for these infinitely small (and nonexistent) quantities was removed, and replaced by a notion of quantities being "close" to others. The derivative and the integral were both reformulated in terms of limits. While it may seem like a lot of work to create rigorous justifications of computations that seemed to work fine in the first place, this is an important development. By putting Calculus on a logical footing, mathematicians were better able to understand and extend its results, as well as to come to terms with some of the more subtle aspects of the theory.

When we first study Calculus we often learn its concepts in an order that is somewhat backwards to its development. We wish to take advantage of the hundreds of years of thought that have gone into it. As a result, we often begin by learning about limits. Afterward we define the derivative and integral developed by Newton and Leibniz. But unlike Newton and Leibniz we define them in the modern way – in terms of limits. Afterward we see how the derivative and integral can be used to solve many of the problems that precipitated the development of Calculus.

Development of Modern Mathematics

When we consider the history of modern Mathematics, two questions arise:

1. what limitations shall be placed upon the term mathematics?

2. what force shall be assigned to the word Modern?

Here, we mainly limit ourselves into the domain of pure science. Questions of applications of the various branches will be considered incidentally. Such great contributions as those of Newton in the realm of Mathematical Physics, of Laplace in celestial mechanics, of Lagrange and Cauchy in wave theory, belong rather to the filed of applications.

In particular, in the domain of numbers, reference will be made to certain of the contributions to the general theory, to the men who have placed the study of irrational and transcendental numbers upon a scientific foundation, and to those who have developed the modern theory of complex numbers and its elaboration in the filed of quarternions and Ausdehnungslehre. In the theory of equations, the manes of some of the leading investigators will be mentioned, together with a brief statement of the results which they secured. This phase of higher algebra will be followed by the theory of forms, or quantics. The later development of calculus, leading to differential equations and the theory of functions, will complete the algebraic side, save for a brief reference to the theory of probabilities. In the domain of geometry, some of the contributors to the later development of the analytic and synthetic fields will be mentioned, together with the most noteworthy results of their labours.

The term Modern Mathematics is not well defined. Algebra cannot be modern yet the theory of equations has received some of its most significant additions during the nineteenth century, while the theory of forms is a recent creation. Similarly with elementary geometry; the labours of Lobachevsky and Bolyai during the second quarter of the century threw a new light upon the whole subject, and more recently the study of the triangle has added another chapter to the theory. Thus, the history of modern mathematics must also be the modern history of ancient branches, while subjects which seem the product of late generations have root in other centuries than the present.

The nineteenth century has been a period of intense study of first principles, of the recognition of necessary limitations of various branches, of a great spread of mathematical knowledge, and of the opening of extensive

fields of applied mathematics. Especially influential has been the establishment of scientific schools and journals and university chairs. About the middle of the century, these schools began to exert a still greater influence through the custom of calling to them mathematicians of high repute, thus making Zurich, Karlsruhe, Munich, Dresden, and other cities well known as mathematical centres.

Unit 4

Course Structure

- Development of Modern geometries
- Development of Modern algebra
- Development of the methods of real analysis.

Development of Modern geometry

Descriptive, Projective, and Modern Synthetic Geometry are so interwoven in their historic development that it is even more difficult to separate them from one another than from the analytic geometry just mentioned. Monge had been in possession of his theory for over thirty years before the publication of his Géométrie Descriptive (1800), a delay due to the jealous desire of the military authorities to keep the valuable secret. It is true that certain of its features can be traced back to Desargues, Taylor, Lambert, and Frézier, but it was Monge who worked it out in detail as a science, although Lacroix (1795), inspired by Monge's lectures in the École Polytechnique, published the first work on the subject. After Monge's work appeared, Hachette (1812, 1818, 1821) added materially to its symmetry, subsequent French contributors being Leroy (1842), Olivier (from 1845), de la Gournerie (from 1860), Vallée, de Fourcy, Adhémar, and others. In Germany leading contributors have been Ziegler (1843), Anger (1858), and especially Fiedler (3d edn. 1883-88) and Wiener (1884-87). At this period Monge by no means confined himself to the descriptive geometry. So marked were his labors in the analytic geometry that he has been called the father of the modern theory. He also set forth the fundamental theorem of reciprocal polars, though not in modern language, gave some treatment of ruled surfaces, and extended the theory of polars to quadrics.

Monge and his school concerned themselves especially with the relations of form, and particularly with those of surfaces and curves in a space of three dimensions. Inspired by the general activity of the period, but following rather the steps of Desargues and Pascal, Carnot treated chiefly the metrical relations of figures. In particular he investigated these relations as connected with the theory of transversals, a theory whose fundamental property of a four-rayed pencil goes back to Pappos, and which, though revived by Desargues, was set forth for the first time in its general form in Carnot's Géométrie de Position (1803), and supplemented in his Théorie des Transversales (1806). In these works he introduced negative magnitudes, the general quadrilateral and quadrangle, and numerous other generalizations of value to the elementary geometry of to-day. But although Carnot's work was important and many details are now commonplace, neither the name

of the theory nor the method employed have endured. The present Geometry of Position (Geometrie der Lage) has little in common with Carnot's Géométrie de Position.

Projective Geometry had its origin somewhat later than the period of Monge and Carnot. Newton had discovered that all curves of the third order can be derived by central projection from five fundamental types. But in spite of this fact the theory attracted so little attention for over a century that its origin is generally ascribed to Poncelet. A prisoner in the Russian campaign, confined at Saratoff on the Volga (1812-14), "privé," as he says, "de toute espèce de livres et de secours, surtout distrait par les malheurs de ma patrie et les miens propres," he still had the vigor of spirit and the leisure to conceive the great work which he published (1822) eight years later. In this work was first made prominent the power of central projection in demonstration and the power of the principle of continuity in research. His leading idea was the study of projective properties, and as a foundation principle he introduced the anharmonic ratio, a concept, however, which dates back to Pappos and which Desargues (1639) had also used. Möbius, following Poncelet, made much use of the anharmonic ratio in his Barycentrische Calcül (1827), but under the name "Doppelschnitt-Verhältniss" (ratio bisectionalis), a term now in common use under Steiner's abbreviated form "Doppelverhältniss." The name "anharmonic ratio" or "function" (rapport anharmonique, or fonction anharmonique) is due to Chasles, and "cross-ratio" was coined by Clifford. The anharmonic point and line properties of conics have been further elaborated by Brianchon, Chasles, Steiner, and von Staudt. To Poncelet is also due the theory of "figures homologiques," the perspective axis and perspective center (called by Chasles the axis and center of homology), an extension of Carnot's theory of transversals, and the "cordes idéales" of conics which Plücker applied to curves of all orders, He also discovered what Salmon has called "the circular points at infinity," thus completing and establishing the first great principle of modern geometry, the principle of continuity. Brianchon (1806), through his application of Desargues's theory of polars, completed the foundation which Monge had begun for Poncelet's (1829) theory of reciprocal polars.

Among the most prominent geometers contemporary with Poncelet was Gergonne, who with more propriety might be ranked as an analytic geometer. He first (1813) used the term "polar" in its modern geometric sense, although Servois (1811) had used the expression "pole." He was also the first (1825-26) to grasp the idea that the parallelism which Maurolycus, Snell, and Viete had noticed is a fundamental principle. This principle he stated and to it he gave the name which it now bears, the Principle of Duality, the most important, after that of continuity, in modern geometry. This principle of geometric reciprocation, the discovery of which was also claimed by Poncelet, has been greatly elaborated and has found its way into modern algebra and elementary geometry, and has recently been extended to mechanics by Genese. Gergonne was the first to use the word "class" in describing a curve, explicitly defining class and degree (order) and showing the duality between the two. He and Chasles were among the first to study scientifically surfaces of higher order.

Steiner (1832) gave the first complete discussion of the projective relations between rows, pencils, etc., and laid the foundation for the subsequent development of pure geometry. He practically closed the theory of conic sections, of the corresponding figures in three-dimensional space and of surfaces of the second order, and hence with him opens the period of special study of curves and surfaces of higher order. His treatment of duality and his application of the theory of projective pencils to the generation of conics are masterpieces. The theory of polars of a point in regard to a curve had been studied by Bobillier and by Grassmann, but Steiner (1848) showed that this theory can serve as the foundation for the study of plane curves independently of the use of coordinates, and introduced those noteworthy curves covariant to a given curve which now bear the names of himself, Hesse, and Cayley. This whole subject has been extended by Grassmann, Chasles, Cremona, and Jonquières. Steiner was the first to make prominent (1832) an example of correspondence of a more complicated nature than that of Poncelet, Möbius, Magnus, and Chasles. His contributions, and those of Gudermann, to the geometry of the sphere were also noteworthy.

While Möbius, Plücker, and Steiner were at work in Germany, Chasles was closing the geometric era opened in France by Monge. His Aperçu Historique (1837) is a classic, and did for France what Salmon's works did for algebra and geometry in England, popularizing the researches of earlier writers and contributing

both to the theory and the nomenclature of the subject. To him is due the name "homographic" and the complete exposition of the principle as applied to plane and solid figures, a subject which has received attention in England at the hands of Salmon, Townsend, and H. J. S. Smith.

Von Staudt began his labors after Plücker, Steiner, and Chasles had made their greatest contributions, but in spite of this seeming disadvantage he surpassed them all. Joining the Steiner school, as opposed to that of Plücker, he became the greatest exponent of pure synthetic geometry of modern times. He set forth (1847, 1856-60) a complete, pure geometric system in which metrical geometry finds no place. Projective properties foreign to measurements are established independently of number relations, number being drawn from geometry instead of conversely, and imaginary elements being systematically introduced from the geometric side. A projective geometry based on the group containing all the real projective and dualistic transformations, is developed, imaginary transformations being also introduced. Largely through his influence pure geometry again became a fruitful field. Since his time the distinction between the metrical and projective theories has been to a great extent obliterated,59 the metrical properties being considered as projective relations to a fundamental configuration, the circle at infinity common for all spheres. Unfortunately von Staudt wrote in an unattractive style, and to Reye is due much of the popularity which now attends the subject.

Cremona began his publications in 1862. His elementary work on projective geometry (1875) in Leudesdorf's translation is familiar to English readers. His contributions to the theory of geometric transformations are valuable, as also his works on plane curves, surfaces, etc.

In England Mulcahy, but especially Townsend (1863), and Hirst, a pupil of Steiner's, opened the subject of modern geometry. Clifford did much to make known the German theories, besides himself contributing to the study of polars and the general theory of curves.

Development of Modern algebra

Modern algebra, also called abstract algebra, branch of mathematics concerned with the general algebraic structure of various sets (such as real numbers, complex numbers, matrices, and vector spaces), rather than rules and procedures for manipulating their individual elements.

Prior to the nineteenth century, algebra meant the study of the solution of polynomial equations. By the twentieth century algebra came to encompass the study of abstract, axiomatic systems such as groups, rings, and fields. This presentation provides an account of the history of the basic concepts, results, and theories of abstract algebra. The development of abstract algebra was propelled by the need for new tools to address certain classical problems that appeared unsolvable by classical means. A major theme of the approach in this book is to show how abstract algebra has arisen in attempts to solve some of these classical problems, providing context from which the reader may gain a deeper appreciation of the mathematics involved. Key features: Begins with an overview of classical algebra Contains separate chapters on aspects of the development of groups, rings, and fields Examines the evolution of linear algebra as it relates to other elements of abstract algebra Highlights the lives and works of six notables: Cayley, Dedekind, Galois, Gauss, Hamilton, and especially the pioneering work of Emmy Noether Offers suggestions to instructors on ways of integrating the history of abstract algebra into their teaching Each chapter concludes with extensive references to the relevant literature Mathematics instructors, algebraists, and historians of science will find the work a valuable reference. The book may also serve as a supplemental text for courses in abstract algebra or the history of mathematics.

During the second half of the 19th century, various important mathematical advances led to the study of sets in which any two elements can be added or multiplied together to give a third element of the same set. The elements of the sets concerned could be numbers, functions, or some other objects. As the techniques involved were similar, it seemed reasonable to consider the sets, rather than their elements, to be the objects of primary concern. A definitive treatise, Modern Algebra, was written in 1930 by the Dutch mathematician

Bartel van der Waerden, and the subject has had a deep effect on almost every branch of mathematics.

Development of Real Analysis

The study of real functions has played a fundamental role in the development of mathematics over the last three centuries. The discovery of calculus by eighteenth century mathematicians, notably Newton and Leibniz, was largely due to increased understanding of the behavior of real functions. The birth of analysis is often traced to the early nineteenth century work of Cauchy, who gave precise definitions of concepts such as continuity and limits for real functions. Convergence problems while approximating real functions by Fourier series gave rise to both the Riemann and Lebesgue integrals. Cantor developed his set theory in an effort to answer uniqueness questions about Fourier series.

During this time, different techniques have been used as the theory behind them became available. For example, after Cauchy, various limiting operations such as pointwise and uniform convergence were studied, giving rise to various approximation techniques. At the turn of this century, measure theoretic techniques were exploited, leading to stochastic convergence ideas in the 1920's. Also, at about the same time topology was developed, and its applications to analysis gave rise to functional analysis.

In recent years, a new research trend has appeared which indicates the emergence of a yet another branch of inquiry that could be called set theoretic real analysis. This area is the study of families of real functions using modern techniques of set theory. These techniques include advanced forcing methods, special axioms of set theory such as Martin's axiom (MA) and proper forcing axiom (PFA), as well as some of their weaker consequences like additivity of measure and category.

Set theoretic real analysis is closely allied with descriptive set theory, but the objects studied in the two areas are different. The objects studied in descriptive set theory are various classes of (mostly nice) sets and their hierarchies, such as Borel sets or analytic sets. Set theoretic real analysis uses the tools of modern set theory to study real functions and is interested mainly in more pathological objects. Thus, the results concerning subsets of the real line (like the series of studies on "small" subsets $\mathbb{R}$, or deep studies of the duality between measure and category) are considered only remotely related to the subject. (However, some of these duality studies spread to real analysis too.)

Set theoretic real analysis already has a long history. Its roots can be traced back to the 1920's, where powerful new techniques based on the Axiom of Choice (AC) and the Continuum Hypothesis (CH) can be seen in many papers from such journals as Fundamenta Mathematicae and Studia Mathematica. The most interesting consequences of the Continuum Hypothesis discovered in this period have been collected in 1934 monograph of Sierpinski, The influence of Sierpinski's results (and the monograph) on the set theoretic real analysis can be best seen in the next section.

The new emergence of the field was sparked by the discovery of powerful new techniques in set theory and can be compared to the parallel development of set theoretic topology during the late 1950's and 1960's. In fact, it is a bit surprising that the development of set theoretic analysis is so much behind that of set theoretic topology, since at the beginning of the century the applicability of set theory in analysis was at least as intense as in topology. This, however, can be probably attributed to the simple fact, that in the past half of a century there were many mathematicians that knew well both topology and set theory, and very few that knew well simultaneously analysis and set theory.

Unit 5

Simple Linear Programming Problems

When both the objective and all the constraints in Expression 1.5 are linear functions, then the optimization problem is called a *linear programming problem.* This has the general form:

$$(2.1)\qquad \begin{cases} \max\ z(x_1,\ldots,x_n) = c_1x_1 + \cdots + c_nx_n \\ s.t.\ a_{11}x_1 + \cdots + a_{1n}x_n \le b_1 \\ \qquad\qquad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \le b_m \\ h_{11}x_1 + \cdots + h_{n1}x_n = r_1 \\ \qquad\qquad \vdots \\ h_{l1}x_1 + \cdots + h_{ln}x_n = r_l \end{cases}$$

Definition 2.1 (Linear Function). A function $z : \mathbb{R}^n \to \mathbb{R}$ is *linear* if there are constants $c_1, \ldots, c_n \in \mathbb{R}$ so that:

$$(2.2)\qquad z(x_1,\ldots,x_n) = c_1x_1 + \cdots + c_nx_n$$

Lemma 2.2 (Linear Function). *If $z : \mathbb{R}^n \to \mathbb{R}$ is linear then for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all scalar constants $\alpha \in \mathbb{R}$ we have:*

$$(2.3)\qquad z(\mathbf{x}_1 + \mathbf{x}_2) = z(\mathbf{x}_1) + z(\mathbf{x}_2)$$

$$(2.4)\qquad z(\alpha\mathbf{x}_1) = \alpha z(\mathbf{x}_1)$$

Exercise 9. Prove Lemma 2.2.

For the time being, we will eschew the general form and focus exclusively on linear programming problems with two variables. Using this limited case, we will develop a graphical method for identifying optimal solutions, which we will generalize later to problems with arbitrary numbers of variables.

Example 2.3. Consider the problem of a toy company that produces toy planes and toy boats. The toy company can sell its planes for \$10 and its boats for \$8 dollars. It costs \$3 in raw materials to make a plane and \$2 in raw materials to make a boat. A plane requires 3 hours to make and 1 hour to finish while a boat requires 1 hour to make and 2 hours to finish. The toy company knows it will not sell anymore than 35 planes per week. Further, given the number of workers, the company cannot spend anymore than 160 hours per week

finishing toys and 120 hours per week making toys. The company wishes to maximize the profit it makes by choosing how much of each toy to produce.

We can represent the profit maximization problem of the company as a linear programming problem. Let x_1 be the number of planes the company will produce and let x_2 be the number of boats the company will produce. The profit for each plane is $\$10 - \$3 = \$7$ per plane and the profit for each boat is $\$8 - \$2 = \$6$ per boat. Thus the total profit the company will make is:

$$z(x_1, x_2) = 7x_1 + 6x_2 \tag{2.5}$$

The company can spend no more than 120 hours per week making toys and since a plane takes 3 hours to make and a boat takes 1 hour to make we have:

$$3x_1 + x_2 \leq 120 \tag{2.6}$$

Likewise, the company can spend no more than 160 hours per week finishing toys and since it takes 1 hour to finish a plane and 2 hour to finish a boat we have:

$$x_1 + 2x_2 \leq 160 \tag{2.7}$$

Finally, we know that $x_1 \leq 35$, since the company will make no more than 35 planes per week. Thus the complete linear programming problem is given as:

$$\begin{cases} \max & z(x_1, x_2) = 7x_1 + 6x_2 \\ s.t. & 3x_1 + x_2 \leq 120 \\ & x_1 + 2x_2 \leq 160 \\ & x_1 \leq 35 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{cases} \tag{2.8}$$

Exercise 10. A chemical manufacturer produces three chemicals: A, B and C. These chemical are produced by two processes: 1 and 2. Running process 1 for 1 hour costs \$4 and yields 3 units of chemical A, 1 unit of chemical B and 1 unit of chemical C. Running process 2 for 1 hour costs \$1 and produces 1 units of chemical A, and 1 unit of chemical B (but none of Chemical C). To meet customer demand, at least 10 units of chemical A, 5 units of chemical B and 3 units of chemical C must be produced daily. Assume that the chemical manufacturer wants to minimize the cost of production. Develop a linear programming problem describing the constraints and objectives of the chemical manufacturer. [Hint: Let x_1 be the amount of time Process 1 is executed and let x_2 be amount of time Process 2 is executed. Use the coefficients above to express the cost of running Process 1 for x_1 time and Process 2 for x_2 time. Do the same to compute the amount of chemicals A, B, and C that are produced.]

1. Modeling Assumptions in Linear Programming

Inspecting Example 2.3 (or the more general Problem 2.1) we can see there are several assumptions that must be satisfied when using a linear programming model. We enumerate these below:

Proportionality Assumption: A problem can be phrased as a linear program only if the contribution to the objective function *and* the left-hand-side of each constraint by each decision variable $(x_1, \ldots, x_n)$ is proportional to the value of the decision variable.

Additivity Assumption: A problem can be phrased as a linear programming problem only if the contribution to the objective function *and* the left-hand-side of each constraint by any decision variable x_i $(i = 1, \ldots, n)$ is completely independent of any other decision variable x_j $(j \neq i)$ and additive.

Divisibility Assumption: A problem can be phrased as a linear programming problem only if the quantities represented by each decision variable are infinitely divisible (i.e., fractional answers make sense).

Certainty Assumption: A problem can be phrased as a linear programming problem only if the coefficients in the objective function and constraints are known with certainty.

The first two assumptions simply assert (in English) that both the objective function and functions on the left-hand-side of the (in)equalities in the constraints are linear functions of the variables $x_1, \ldots, x_n$.

The third assumption asserts that a valid optimal answer could contain fractional values for decision variables. It's important to understand how this assumption comes into play– even in the toy making example. Many quantities can be divided into non-integer values (ounces, pounds etc.) but many other quantities cannot be divided. For instance, can we really expect that it's reasonable to make 1/2 a plane in the toy making example? When values must be constrained to true integer values, the linear programming problem is called an *integer programming problem*. These problems are outside the scope of this course, but there is a *vast* literature dealing with them For many problems, particularly when the values of the decision variables may become large, a fractional optimal answer could be obtained and then rounded to the nearest integer to obtain a reasonable answer. For example, if our toy problem were re-written so that the optimal answer was to make 1045.3 planes, then we could round down to 1045.

The final assumption asserts that the coefficients (e.g., profit per plane or boat) is known with absolute certainty. In traditional linear programming, there is no lack of knowledge about the make up of the objective function, the coefficients in the left-hand-side of the constraints or the bounds on the right-hand-sides of the constraints. There is a literature on *stochastic programming* that relaxes some of these assumptions, but this too is outside the scope of the course.

Exercise 11. In a short sentence or two, discuss whether the problem given in Example 2.3 meets all of the assumptions of a scenario that can be modeled by a linear programming problem. Do the same for Exercise 10. [Hint: Can you make 2/3 of a toy? Can you run a process for 1/3 of an hour?]

Exercise 12. Suppose the costs are not known with certainty but instead a probability distribution for each value of c_i $(i = 1, \ldots, n)$ is known. Suggest a way of constructing a linear program from the probability distributions. [Hint: Suppose I tell you that I'll give you a uniformly random amount of money between \$1 and \$2. How much money do you expect to receive? Use the same reasoning to answer the question.]

2. Graphically Solving Linear Programs Problems with Two Variables (Bounded Case)

Linear Programs (LP's) with two variables can be solved graphically by plotting the feasible region along with the level curves of the objective function. We will show that we can find a point in the feasible region that maximizes the objective function using the level curves of the objective function. We illustrate the method first using the problem from Example 2.3.

EXAMPLE 2.4 (Continuation of Example 2.3). Let's continue the example of the Toy Maker begin in Example 2.3. To solve the linear programming problem graphically, begin by drawing the feasible region. This is shown in the blue shaded region of Figure 2.1.

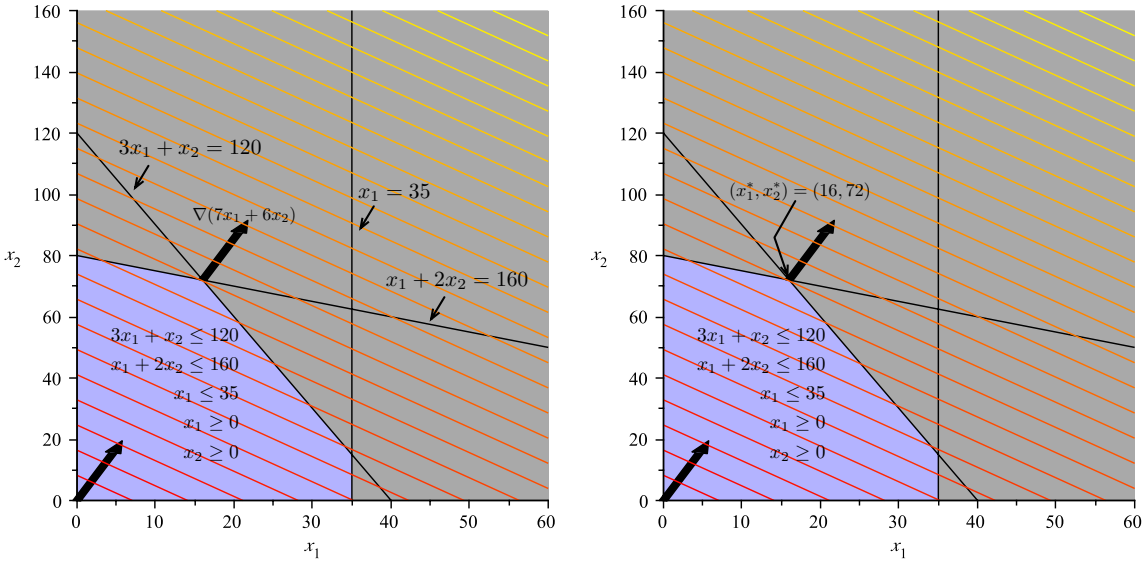


Figure 2.1. Feasible Region and Level Curves of the Objective Function: The shaded region in the plot is the feasible region and represents the intersection of the five inequalities constraining the values of x_1 and x_2. On the right, we see the optimal solution is the "last" point in the feasible region that intersects a level set as we move in the direction of increasing profit.

After plotting the feasible region, the next step is to plot the level curves of the objective function. In our problem, the level sets will have the form:

$$7x_1 + 6x_2 = c \implies x_2 = \frac{-7}{6}x_1 + \frac{c}{6}$$

This is a set of parallel lines with slope $-7/6$ and intercept $c/6$ where c can be varied as needed. The level curves for various values of c are parallel lines. In Figure 2.1 they are shown in colors ranging from red to yellow depending upon the value of c. Larger values of c are more yellow.

To solve the linear programming problem, follow the level sets along the gradient (shown as the black arrow) until the last level set (line) intersects the feasible region. If you are doing this by hand, you can draw a single line of the form $7x_1 + 6x_2 = c$ and then simply

draw parallel lines in the direction of the gradient $(7, 6)$. At some point, these lines will fail to intersect the feasible region. The last line to intersect the feasible region will do so at a point that maximizes the profit. In this case, the point that maximizes $z(x_1, x_2) = 7x_1 + 6x_2$, subject to the constraints given, is $(x_1^*, x_2^*) = (16, 72)$.

Note the point of optimality $(x_1^*, x_2^*) = (16, 72)$ is at a corner of the feasible region. This corner is formed by the intersection of the two lines: $3x_1 + x_2 = 120$ and $x_1 + 2x_2 = 160$. In this case, the constraints

$$3x_1 + x_2 \leq 120$$
$$x_1 + 2x_2 \leq 160$$

are both *binding*, while the other constraints are non-binding. In general, we will see that when an optimal solution to a linear programming problem exists, it will always be at the intersection of several binding constraints; that is, it will occur at a corner of a higher-dimensional polyhedron.

3. Formalizing The Graphical Method

In order to formalize the method we've shown above, we will require a few new definitions.

DEFINITION 2.5. Let $r \in \mathbb{R}$, $r \geq 0$ be a non-negative scalar and let $\mathbf{x}_0 \in \mathbb{R}^n$ be a point in $\mathbb{R}^n$. Then the set:

$$B_r(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^n | \ ||\mathbf{x} - \mathbf{x}_0|| \leq r\} \tag{2.9}$$

is called the *closed ball of radius r centered at point $\mathbf{x}_0$ in $\mathbb{R}^n$*.

In $\mathbb{R}^2$, a closed ball is just a disk and its circular boundary centered at $\mathbf{x}_0$ with radius r. In $\mathbb{R}^3$, a closed ball is a solid sphere and its spherical centered at $\mathbf{x}_0$ with radius r. Beyond three dimensions, it becomes difficult to visualize what a closed ball looks like.

We can use a closed ball to define the notion of boundedness of a feasible region:

DEFINITION 2.6. Let $S \subseteq \mathbb{R}^n$. Then the set S is *bounded* if there exists an $\mathbf{x}_0 \in \mathbb{R}^n$ and finite $r \geq 0$ such that S is totally contained in $B_r(\mathbf{x}_0)$; that is, $S \subset B_r(\mathbf{x}_0)$.

Definition 2.6 is illustrated in Figure 2.2. The set S is shown in blue while the ball of radius r centered at $\mathbf{x}_0$ is shown in gray.

We can now define an algorithm for identifying the solution to a linear programing problem in two variables with a *bounded* feasible region (see Algorithm 1):

The example linear programming problem presented in the previous section has a single optimal solution. In general, the following outcomes can occur in solving a linear programming problem:

(1) The linear programming problem has a unique solution. (We've already seen this.)
(2) There are infinitely many alternative optimal solutions.
(3) There is no solution and the problem's objective function can grow to positive infinity for maximization problems (or negative infinity for minimization problems).
(4) There is no solution to the problem at all.

Case 3 above can only occur when the feasible region is unbounded; that is, it cannot be surrounded by a ball with finite radius. We will illustrate each of these possible outcomes in the next four sections. We will prove that this is true in a later chapter.

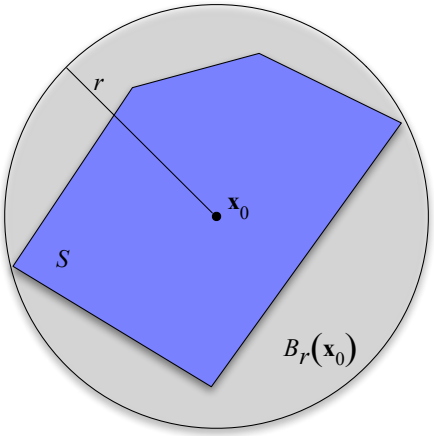


Figure 2.2. A Bounded Set: The set S (in blue) is bounded because it can be entirely contained inside a ball of a finite radius r and centered at some point $\mathbf{x}_0$. In this example, the set S is in $\mathbb{R}^2$. This figure also illustrates the fact that a ball in $\mathbb{R}^2$ is just a disk and its boundary.

Algorithm for Solving a Linear Programming Problem Graphically
Bounded Feasible Region, Unique Solution

(1) Plot the feasible region defined by the constraints.
(2) Plot the level sets of the objective function.
(3) For a maximization problem, identify the level set corresponding the greatest (least, for minimization) objective function value that intersects the feasible region. This point will be at a corner.
(4) The point on the corner intersecting the greatest (least) level set is a solution to the linear programming problem.

Algorithm 1. Algorithm for Solving a Two Variable Linear Programming Problem Graphically–Bounded Feasible Region, Unique Solution Case

Exercise 13. Use the graphical method for solving linear programming problems to solve the linear programming problem you defined in Exercise 10.

4. Problems with Alternative Optimal Solutions

We'll study a specific linear programming problem with an infinite number of solutions by modifying the objective function in Example 2.3.

EXAMPLE 2.7. Suppose the toy maker in Example 2.3 finds that it can sell planes for a profit of \$18 each instead of \$7 each. The new linear programming problem becomes:

$$(2.10)\quad \begin{cases} \max & z(x_1, x_2) = 18x_1 + 6x_2 \\ s.t. & 3x_1 + x_2 \leq 120 \\ & x_1 + 2x_2 \leq 160 \\ & x_1 \leq 35 \\ & x_1 \geq 0 \\ & x_2 \geq 0 \end{cases}$$

Applying our graphical method for finding optimal solutions to linear programming problems yields the plot shown in Figure 2.3. The level curves for the function $z(x_1, x_2) = 18x_1 + 6x_2$ are *parallel* to one face of the polygon boundary of the feasible region. Hence, as we move further up and to the right in the direction of the gradient (corresponding to larger and larger values of $z(x_1, x_2)$) we see that there is not *one* point on the boundary of the feasible region that intersects that level set with greatest value, but instead a side of the polygon boundary described by the line $3x_1 + x_2 = 120$ where $x_1 \in [16, 35]$. Let:

$$S = \{(x_1, x_2) | 3x_1 + x_2 \leq 120,\ x_1 + 2x_2 \leq 160,\ x_1 \leq 35,\ x_1, x_2 \geq 0\}$$

that is, S is the feasible region of the problem. Then for any value of $x_1^* \in [16, 35]$ and any value x_2^* so that $3x_1^* + x_2^* = 120$, we will have $z(x_1^*, x_2^*) \geq z(x_1, x_2)$ for all $(x_1, x_2) \in S$. Since there are infinitely many values that x_1 and x_2 may take on, we see this problem has an infinite number of alternative optimal solutions.

Based on the example in this section, we can modify our algorithm for finding the solution to a linear programming problem graphically to deal with situations with an infinite set of alternative optimal solutions (see Algorithm 2):

Algorithm for Solving a Linear Programming Problem Graphically
Bounded Feasible Region

(1) Plot the feasible region defined by the constraints.
(2) Plot the level sets of the objective function.
(3) For a maximization problem, identify the level set corresponding the greatest (least, for minimization) objective function value that intersects the feasible region. This point will be at a corner.
(4) The point on the corner intersecting the greatest (least) level set is a solution to the linear programming problem.
(5) **If the level set corresponding to the greatest (least) objective function value is parallel to a side of the polygon boundary next to the corner identified, then there are infinitely many alternative optimal solutions and any point on this side may be chosen as an optimal solution.**

Algorithm 2. Algorithm for Solving a Two Variable Linear Programming Problem Graphically–Bounded Feasible Region Case

EXERCISE 14. Modify the linear programming problem from Exercise 10 to obtain a linear programming problem with an infinite number of alternative optimal solutions. Solve

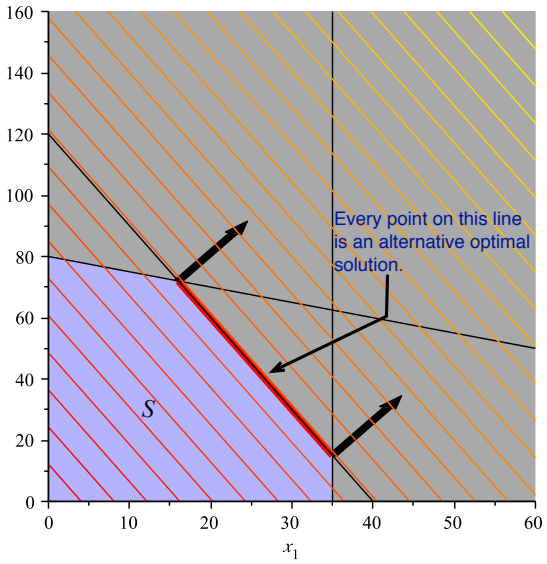


Figure 2.3. An example of infinitely many alternative optimal solutions in a linear programming problem. The level curves for $z(x_1, x_2) = 18x_1 + 6x_2$ are *parallel* to one face of the polygon boundary of the feasible region. Moreover, this side contains the points of greatest value for $z(x_1, x_2)$ inside the feasible region. Any combination of (x_1, x_2) on the line $3x_1+x_2 = 120$ for $x_1 \in [16, 35]$ will provide the largest possible value $z(x_1, x_2)$ can take in the feasible region S.

the new problem and obtain a description for the set of alternative optimal solutions. [Hint: Just as in the example, x_1 will be bound between two value corresponding to a side of the polygon. Find those values and the constraint that is binding. This will provide you with a description of the form for any $x_1^* \in [a, b]$ and x_2^* is chosen so that $cx_1^* + dx_2^* = v$, the point (x_1^*, x_2^*) is an alternative optimal solution to the problem. Now you fill in values for a, b, c, d and v.]

5. Problems with No Solution

Recall for *any* mathematical programming problem, the feasible set or region is simply a subset of $\mathbb{R}^n$. If this region is empty, then there is no solution to the mathematical programming problem and the problem is said to be *over constrained*. We illustrate this case for linear programming problems with the following example.

Example 2.8. Consider the following linear programming problem:

$$(2.11)\quad \begin{cases} \max\ z(x_1, x_2) = 3x_1 + 2x_2 \\ s.t.\ \ \dfrac{1}{40}x_1 + \dfrac{1}{60}x_2 \leq 1 \\ \qquad \dfrac{1}{50}x_1 + \dfrac{1}{50}x_2 \leq 1 \\ \qquad x_1 \geq 30 \\ \qquad x_2 \geq 20 \end{cases}$$

The level sets of the objective and the constraints are shown in Figure 2.4.

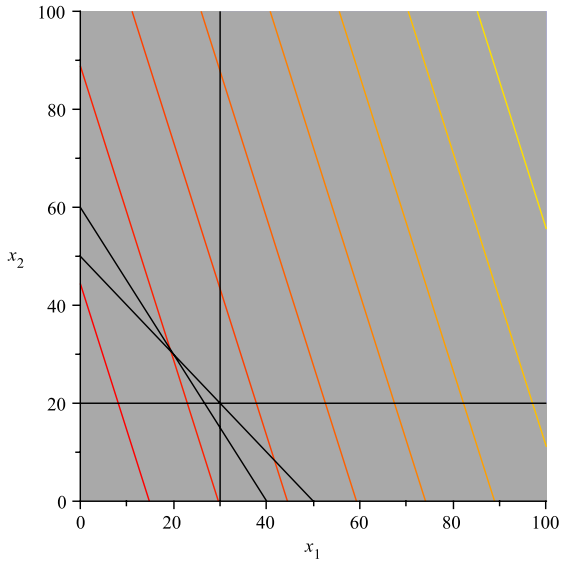


Figure 2.4. A Linear Programming Problem with no solution. The feasible region of the linear programming problem is empty; that is, there are no values for x_1 and x_2 that can simultaneously satisfy all the constraints. Thus, no solution exists.

The fact that the feasible region is empty is shown by the fact that in Figure 2.4 there is no blue region–i.e., all the regions are gray indicating that the constraints are not satisfiable.

Based on this example, we can modify our previous algorithm for finding the solution to linear programming problems graphically (see Algorithm 3):

Algorithm for Solving a Linear Programming Problem Graphically
Bounded Feasible Region

(1) Plot the feasible region defined by the constraints.
(2) **If the feasible region is empty, then no solution exists.**
(3) Plot the level sets of the objective function.
(4) For a maximization problem, identify the level set corresponding the greatest (least, for minimization) objective function value that intersects the feasible region. This point will be at a corner.
(5) The point on the corner intersecting the greatest (least) level set is a solution to the linear programming problem.
(6) **If the level set corresponding to the greatest (least) objective function value is parallel to a side of the polygon boundary next to the corner identified, then there are infinitely many alternative optimal solutions and any point on this side may be chosen as an optimal solution.**

Algorithm 3. Algorithm for Solving a Two Variable Linear Programming Problem Graphically–Bounded Feasible Region Case

6. Problems with Unbounded Feasible Regions

Again, we'll tackle the issue of linear programming problems with unbounded feasible regions by illustrating the possible outcomes using examples.

Example 2.9. Consider the linear programming problem below:

$$\text{(2.12)} \quad \begin{cases} \max \ z(x_1, x_2) = 2x_1 - x_2 \\ s.t. \ x_1 - x_2 \leq 1 \\ \quad\quad 2x_1 + x_2 \geq 6 \\ \quad\quad x_1, x_2 \geq 0 \end{cases}$$

The feasible region and level curves of the objective function are shown in Figure 2.5. The

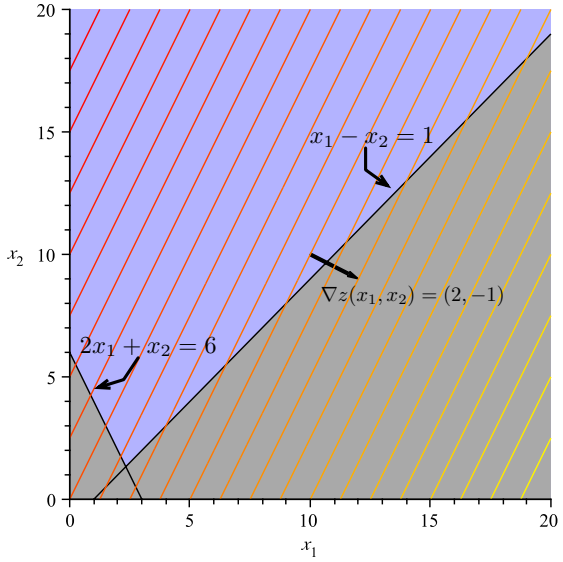


Figure 2.5. A Linear Programming Problem with Unbounded Feasible Region: Note that we can continue to make level curves of $z(x_1, x_2)$ corresponding to larger and larger values as we move down and to the right. These curves will continue to intersect the feasible region for any value of $v = z(x_1, x_2)$ we choose. Thus, we can make $z(x_1, x_2)$ as large as we want and still find a point in the feasible region that will provide this value. Hence, the optimal value of $z(x_1, x_2)$ subject to the constraints $+\infty$. That is, the problem is unbounded.

feasible region in Figure 2.5 is clearly unbounded since it stretches upward along the x_2 axis infinitely far and also stretches rightward along the x_1 axis infinitely far, bounded below by the line $x_1 - x_2 = 1$. There is no way to enclose this region by a disk of finite radius, hence the feasible region is not bounded.

We can draw more level curves of $z(x_1, x_2)$ in the direction of increase (down and to the right) as long as we wish. There will always be an intersection point with the feasible region because it is infinite. That is, these curves will continue to intersect the feasible region for any value of $v = z(x_1, x_2)$ we choose. Thus, we can make $z(x_1, x_2)$ as large as we want and still find a point in the feasible region that will provide this value. Hence, the largest value

$z(x_1, x_2)$ can take when (x_1, x_2) are in the feasible region is $+\infty$. That is, the problem is unbounded.

Just because a linear programming problem has an unbounded feasible region does not imply that there is not a finite solution. We illustrate this case by modifying example 2.9.

EXAMPLE 2.10 (Continuation of Example 2.9). Consider the linear programming problem from Example 2.9 with the new objective function: $z(x_1, x_2) = (1/2)x_1 - x_2$. Then we have the new problem:

$$\text{(2.13)} \qquad \begin{cases} \max & z(x_1, x_2) = \dfrac{1}{2}x_1 - x_2 \\ s.t. & x_1 - x_2 \leq 1 \\ & 2x_1 + x_2 \geq 6 \\ & x_1, x_2 \geq 0 \end{cases}$$

The feasible region, level sets of $z(x_1, x_2)$ and gradients are shown in Figure 2.6. In this case note, that the direction of increase of the objective function is *away* from the direction in which the feasible region is unbounded (i.e., downward). As a result, the point in the feasible region with the largest $z(x_1, x_2)$ value is $(7/3, 4/3)$. Again this is a vertex: the binding constraints are $x_1 - x_2 = 1$ and $2x_1 + x_2 = 6$ and the solution occurs at the point these two lines intersect.

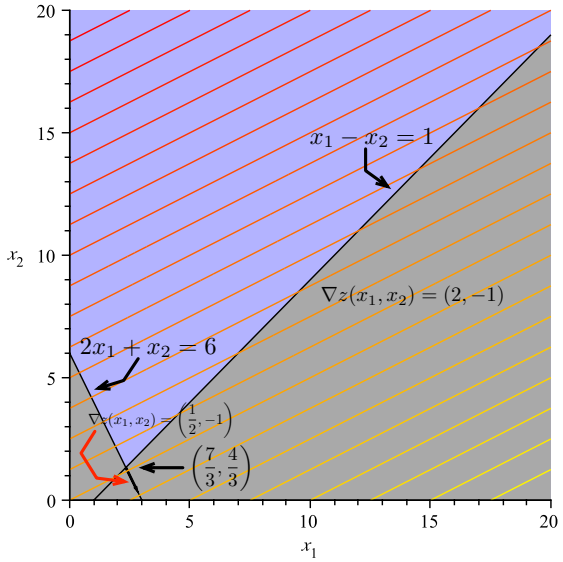


Figure 2.6. A Linear Programming Problem with Unbounded Feasible Region and Finite Solution: In this problem, the level curves of $z(x_1, x_2)$ increase in a more "southernly" direction that in Example 2.10–that is, *away* from the direction in which the feasible region increases without bound. The point in the feasible region with largest $z(x_1, x_2)$ value is $(7/3, 4/3)$. Note again, this is a vertex.

Based on these two examples, we can modify our algorithm for graphically solving a two variable linear programming problems to deal with the case when the feasible region is unbounded.

Algorithm for Solving a Two Variable Linear Programming Problem Graphically

(1) Plot the feasible region defined by the constraints.
(2) If the feasible region is empty, then no solution exists.
(3) If the feasible region is unbounded goto Line 8. Otherwise, Goto Line 4.
(4) Plot the level sets of the objective function.
(5) For a maximization problem, identify the level set corresponding the greatest (least, for minimization) objective function value that intersects the feasible region. This point will be at a corner.
(6) The point on the corner intersecting the greatest (least) level set is a solution to the linear programming problem.
(7) **If the level set corresponding to the greatest (least) objective function value is parallel to a side of the polygon boundary next to the corner identified, then there are infinitely many alternative optimal solutions and any point on this side may be chosen as an optimal solution.**
(8) (The feasible region is unbounded): Plot the level sets of the objective function.
(9) If the level sets intersect the feasible region at larger and larger (smaller and smaller for a minimization problem), then the problem is unbounded and the solution is $+\infty$ ($-\infty$ for minimization problems).
(10) Otherwise, identify the level set corresponding the greatest (least, for minimization) objective function value that intersects the feasible region. This point will be at a corner.
(11) The point on the corner intersecting the greatest (least) level set is a solution to the linear programming problem. **If the level set corresponding to the greatest (least) objective function value is parallel to a side of the polygon boundary next to the corner identified, then there are infinitely many alternative optimal solutions and any point on this side may be chosen as an optimal solution.**

Algorithm 4. Algorithm for Solving a Linear Programming Problem Graphically– Bounded and Unbounded Case

Exercise 15. Does the following problem have a bounded solution? Why?

$$\text{(2.14)} \quad \begin{cases} \textbf{min} \;\; z(x_1, x_2) = 2x_1 - x_2 \\ s.t. \;\; x_1 - x_2 \leq 1 \\ \qquad 2x_1 + x_2 \geq 6 \\ \qquad x_1, x_2 \geq 0 \end{cases}$$

[Hint: Use Figure 2.6 and Algorithm 4.]

Exercise 16. Modify the objective function in Example 2.9 or Example 2.10 to produce a problem with an infinite number of solutions.

Exercise 17. Modify the objective function in Exercise 15 to produce a **minimization** problem that has a finite solution. Draw the feasible region and level curves of the objective

Convex Sets, Functions and Cones and Polyhedral Theory

In this chapter, we will cover all of the geometric prerequisites for understanding the theory of linear programming. We will use the results in this section to prove theorems about the Simplex Method in other sections.

1. Convex Sets

DEFINITION 4.1 (Convex Set). Let $X \subseteq \mathbb{R}^n$. Then the set X is convex if and only if for all pairs $\mathbf{x}_1, \mathbf{x}_2 \in X$ we have $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in X$ for all $\lambda \in [0,1]$.

The definition of convexity seems complex, but it is easy to understand. First recall that if $\lambda \in [0,1]$, then the point $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ is on the line segment connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^n$. For example, when $\lambda = 1/2$, then the point $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ is the midpoint between $\mathbf{x}_1$ and $\mathbf{x}_2$. In fact, for every point $\mathbf{x}$ on the line connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ we can find a value $\lambda \in [0,1]$ so that $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. Then we can see that, convexity asserts that if $\mathbf{x}_1, \mathbf{x}_2 \in X$, then every point on the line connecting $\mathbf{x}_1$ and $\mathbf{x}_2$ is also in the set X.

DEFINITION 4.2 (Positive Combination). Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$. If $\lambda_1, \ldots, \lambda_m > 0$ and then

$$\mathbf{x} = \sum_{i=1}^{m} \lambda_i \mathbf{x}_i \tag{4.1}$$

is called a *positive combination* of $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

DEFINITION 4.3 (Convex Combination). Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$. If $\lambda_1, \ldots, \lambda_m \in [0,1]$ and

$$\sum_{i=1}^{m} \lambda_i = 1$$

then

$$\mathbf{x} = \sum_{i=1}^{m} \lambda_i \mathbf{x}_i \tag{4.2}$$

is called a *convex combination* of $\mathbf{x}_1, \ldots, \mathbf{x}_m$. If $\lambda_i < 1$ for all $i = 1, \ldots, m$, then Equation 4.2 is called a *strict convex combination*.

REMARK 4.4. If you recall the definition of linear combination, we can see that we move from the very general to the very specific as we go from linear combinations to positive combinations to convex combinations. A linear combination of points or vectors allowed us to choose any real values for the coefficients. A positive combination restricts us to positive values, while a convex combination asserts that those values must be non-negative and sum to 1.

EXAMPLE 4.5. Figure 4.1 illustrates a convex and non-convex set. Non-convex sets have

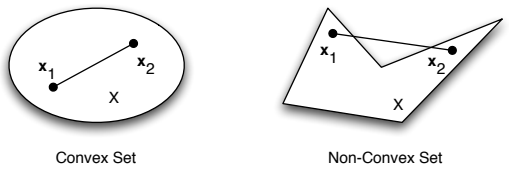


Figure 4.1. Examples of Convex Sets: The set on the left (an ellipse and its interior) is a convex set; every pair of points inside the ellipse can be connected by a line contained entirely in the ellipse. The set on the right is clearly not convex as we've illustrated two points whose connecting line is not contained inside the set.

some resemblance to crescent shapes or have components that look like crescents.

THEOREM 4.6. *The intersection of a finite number of convex sets in* $\mathbb{R}^n$ *is convex.*

PROOF. Let $C_1, \ldots, C_n \subseteq \mathbb{R}^n$ be a finite collection of convex sets. Let

$$C = \bigcap_{i=1}^{n} C_i \tag{4.3}$$

be the set formed from the intersection of these sets. Choose $\mathbf{x}_1, \mathbf{x}_2 \in C$ and $\lambda \in [0, 1]$. Consider $\mathbf{x} = \lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$. We know that $\mathbf{x}_1, \mathbf{x}_2 \in C_1, \ldots, C_n$ by definition of C. By convexity, we know that $\mathbf{x} \in C_1, \ldots, C_n$ by convexity of each set. Therefore, $\mathbf{x} \in C$. Thus C is a convex set. □

2. Convex and Concave Functions

DEFINITION 4.7 (Convex Function). A function $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function if it satisfies:

$$f(\lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1 - \lambda)f(\mathbf{x}_2) \tag{4.4}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$.

This definition is illustrated in Figure 4.2. When f is a univariate function, this definition can be shown to be equivalent to the definition you learned in Calculus I (Math 140) using first and second derivatives.

DEFINITION 4.8 (Concave Function). A function $f : \mathbb{R}^n \to \mathbb{R}$ is a concave function if it satisfies:

$$f(\lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2) \geq \lambda f(\mathbf{x}_1) + (1 - \lambda)f(\mathbf{x}_2) \tag{4.5}$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and for all $\lambda \in [0, 1]$ [1].

To visualize this definition, simply flip Figure 4.2 upside down. The following theorem is a powerful tool that can be used to show sets are convex. It's proof is outside the scope of the class, but relatively easy.

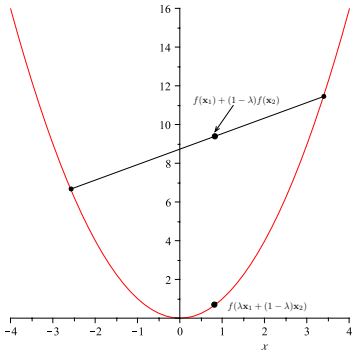


Figure 4.2. A convex function: A convex function satisfies the expression $f(\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2) \leq \lambda f(\mathbf{x}_1) + (1-\lambda)f(\mathbf{x}_2)$ for all $\mathbf{x}_1$ and $\mathbf{x}_2$ and $\lambda \in [0,1]$.

THEOREM 4.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the set $C = \{\mathbf{x} \in \mathbb{R}^n : f(x) \leq c\}$, where $c \in \mathbb{R}$, is a convex set.*

EXERCISE 40. Prove the Theorem 4.9. [Hint: Skip ahead and read the proof of Lemma 4.15. Follow the steps in that proof, but apply them to f.]

3. Polyhedral Sets

Important examples of convex sets are polyhedral sets, the multi-dimensional analogs of polygons in the plane. In order to understand these structures, we must first understand hyperplanes and half-spaces.

DEFINITION 4.10 (Hyperplane). Let $\mathbf{a} \in \mathbb{R}^n$ be a constant vector in n-dimensional space and let $b \in \mathbb{R}$ be a constant scalar. The set of points

$$H = \left\{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T\mathbf{x} = b\right\} \tag{4.6}$$

is a *hyperplane* in n-dimensional space. *Note the use of column vectors for* $\mathbf{a}$ *and* $\mathbf{x}$ *in this definition.*

EXAMPLE 4.11. Consider the hyper-plane $2x_1+3x_2+x_3 = 5$. This is shown in Figure 4.3. This hyperplane is composed of the set of points $(x_1, x_2, x_3) \in \mathbb{R}^3$ satisfying $2x_1+3x_2+x_3 = 5$. This can be plotted implicitly or explicitly by solving for one of the variables, say x_3. We can write x_3 as a function of the other two variables as:

$$x_3 = 5 - 2x_1 - 3x_2 \tag{4.7}$$

DEFINITION 4.12 (Half-Space). Let $\mathbf{a} \in \mathbb{R}^n$ be a constant vector in n-dimensional space and let $b \in \mathbb{R}$ be a constant scalar. The sets of points

$$H_l = \left\{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T\mathbf{x} \leq b\right\} \tag{4.8}$$

$$H_u = \left\{\mathbf{x} \in \mathbb{R}^n | \mathbf{a}^T\mathbf{x} \geq b\right\} \tag{4.9}$$

are the half-spaces defined by the hyperplane $\mathbf{a}^T\mathbf{x} = b$.

EXAMPLE 4.13. Consider the two dimensional hyperplane (line) $x_1 + x_2 = 1$. Then the two half-spaces associated with this hyper-plane are shown in Figure 4.4. A half-space is

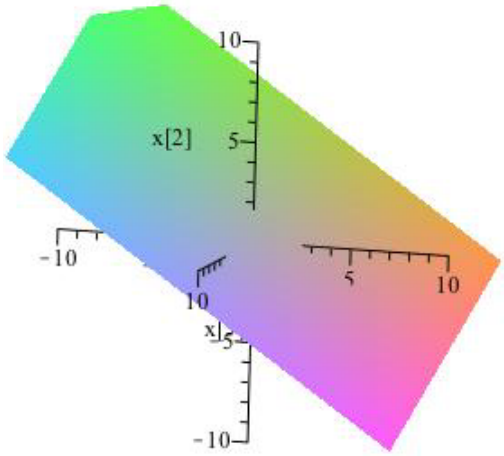
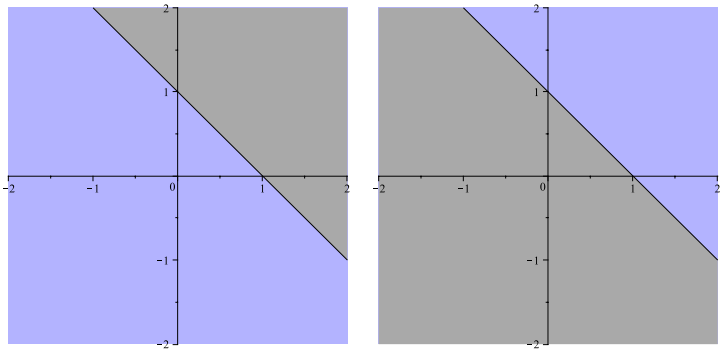


Figure 4.3. A hyperplane in 3 dimensional space: A hyperplane is the set of points satisfying an equation $\mathbf{a}^T\mathbf{x} = b$, where k is a constant in $\mathbb{R}$ and $\mathbf{a}$ is a constant vector in $\mathbb{R}^n$ and $\mathbf{x}$ is a variable vector in $\mathbb{R}^n$. The equation is written as a matrix multiplication using our assumption that all vectors are column vectors.



(a) H_l (b) H_u

Figure 4.4. Two half-spaces defined by a hyper-plane: A half-space is so named because any hyper-plane divides $\mathbb{R}^n$ (the space in which it resides) into two halves, the side "on top" and the side "on the bottom."

so named because the hyperplane $\mathbf{a}^T\mathbf{x} = b$ literally separates $\mathbb{R}^n$ into two halves: the half above the hyperplane and the half below the hyperplane.

LEMMA 4.14. *Every hyper-plane is convex.*

PROOF. Let $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ and let H be the hyperplane defined by $\mathbf{a}$ and b. Choose $\mathbf{x}_1, \mathbf{x}_2 \in H$ and $\lambda \in [0, 1]$. Let $\mathbf{x} = \lambda\mathbf{x}_1 + (1 - \lambda)\mathbf{x}_2$. By definition we know that:

$$\mathbf{a}^T\mathbf{x}_1 = b$$
$$\mathbf{a}^T\mathbf{x}_2 = b$$

Then we have:

$$\mathbf{a}^T\mathbf{x} = \mathbf{a}^T\left[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\right] = \lambda\mathbf{a}^T\mathbf{x}_1 + (1-\lambda)\mathbf{a}^T\mathbf{x}_2 = \lambda b + (1-\lambda)b = b \tag{4.10}$$

Thus, $\mathbf{x} \in H$ and we see that H is convex. This completes the proof. □

LEMMA 4.15. *Every half-space is convex.*

PROOF. Let $\mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Without loss of generality, consider the half-space H_l defined by $\mathbf{a}$ and b. For arbitrary $\mathbf{x}_1$ and $\mathbf{x}_2$ in H_l we have:

$$\mathbf{a}^T\mathbf{x}_1 \leq b$$
$$\mathbf{a}^T\mathbf{x}_2 \leq b$$

Suppose that $\mathbf{a}^T\mathbf{x}_1 = b_1 \leq b$ and $\mathbf{a}^T\mathbf{x}_2 = b_2 \leq b$. Again let $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. Then:

$$\mathbf{a}^T\mathbf{x} = \mathbf{a}^T\left[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2\right] = \lambda\mathbf{a}^T\mathbf{x}_1 + (1-\lambda)\mathbf{a}^T\mathbf{x}_2 = \lambda b_1 + (1-\lambda)b_2 \tag{4.11}$$

Since $\lambda \leq 1$ and $1-\lambda \leq 1$ and $\lambda \geq 0$ we know that $\lambda b_1 \leq \lambda b$, since $b_1 \leq b$. Similarly we know that $(1-\lambda)b_2 \leq (1-\lambda)b$, since $b_2 \leq b$. Thus:

$$\lambda b_1 + (1-\lambda)b_2 \leq \lambda b + (1-\lambda)b = b \tag{4.12}$$

Thus we have shown that $\mathbf{a}^T\mathbf{x} \leq b$. The case for H_u is identical with the signs of the inequalities reversed. This completes the proof. □

Using these definitions, we are now in a position to define polyhedral sets, which will be the subject of our study for most of the remainder of this chapter.

DEFINITION 4.16 (Polyhedral Set). If $P \subseteq \mathbb{R}^n$ is the intersection of a finite number of half-spaces, then P is a *polyhedral set*. Formally, let $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ be a finite set of constant vectors and let $b_1, \ldots, b_m \in \mathbb{R}$ be constants. Consider the set of half-spaces:

$$H_i = \{\mathbf{x} | \mathbf{a}_i^T\mathbf{x} \leq b_i\}$$

Then the set:

$$P = \bigcap_{i=1}^{m} H_i \tag{4.13}$$

is a *polyhedral set*.

It should be clear that we can represent any polyhedral set using a matrix inequality. The set P is defined by the set of vectors $\mathbf{x}$ satisfying:

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}, \tag{4.14}$$

where the *rows* of $\mathbf{A} \in \mathbb{R}^{m\times n}$ are made up of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$ and $\mathbf{b} \in \mathbb{R}^m$ is a column vector composed of elements $b_1, \ldots, b_m$.

THEOREM 4.17. *Every polyhedral set is convex.*

EXERCISE 41. Prove Theorem 4.17. [Hint: You can prove this by brute force, verifying convexity. You can also be clever and use two results that we've proved in the notes.]

Unit 6

The Simplex Method

1. Linear Programming and Extreme Points

In this section we formalize the intuition we've obtained in all our work in two dimensional linear programming problems. Namely, we noted that if an optimal solution existed, then it occurred at an extreme point. For the remainder of this chapter, assume that $A \in \mathbb{R}^{m\times n}$ with full row rank and $b \in \mathbb{R}^m$ let

$$X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} \leq \mathbf{b},\ \ \mathbf{x} \geq \mathbf{0}\} \tag{5.1}$$

be a polyhedral set over which we will maximize the objective function $z(x_1, \ldots, x_n) = \mathbf{c}^T\mathbf{x}$, where $\mathbf{c}, \mathbf{x} \in \mathbb{R}^n$. That is, we will focus on the linear programming problem:

$$P\begin{cases} \max\ \mathbf{c}^T\mathbf{x} \\ s.t.\ \ \mathbf{Ax} \leq \mathbf{b} \\ \qquad \mathbf{x} \geq \mathbf{0} \end{cases} \tag{5.2}$$

THEOREM 5.1. *If Problem P has an optimal solution, then Problem P has an optimal extreme point solution.*

PROOF. Applying the Cartheodory Characterization theorem, we know that any point $\mathbf{x} \in X$ can be written as:

$$\mathbf{x} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i + \sum_{i=1}^{l} \mu_i \mathbf{d}_i \tag{5.3}$$

where $\mathbf{x}_1, \ldots \mathbf{x}_k$ are the extreme points of X and $\mathbf{d}_1, \ldots, \mathbf{d}_l$ are the extreme directions of X and we know that

$$\begin{aligned} &\sum_{i=1}^{k} \lambda_i = 1 \\ &\lambda_i, \mu_i \geq 0\ \ \forall i \end{aligned} \tag{5.4}$$

We can rewrite problem P using this characterization as:

$$\begin{aligned} \max\ &\sum_{i=1}^{k} \lambda_i \mathbf{c}^T\mathbf{x}_i + \sum_{i=1}^{l} \mu_i \mathbf{c}^T\mathbf{d}_i \\ s.t.\ &\sum_{i=1}^{k} \lambda_i = 1 \\ &\lambda_i, \mu_i \geq 0\ \ \forall i \end{aligned} \tag{5.5}$$

If there is some i such that $\mathbf{c}^T\mathbf{d}_i > 0$, then we can simply choose μ_i as large as we like, making the objective as large as we like, the problem will have no finite solution.

Therefore, assume that $\mathbf{c}^T\mathbf{d}_i \leq 0$ for all $i = 1, \ldots, l$ (in which case, we may simply choose $\mu_i = 0$, for all i). Since the set of extreme points $\mathbf{x}_1, \ldots \mathbf{x}_k$ is finite, we can simply set $\lambda_p = 1$ if $\mathbf{c}^T\mathbf{x}_p$ has the largest value among all possible values of $\mathbf{c}^T\mathbf{x}_i$, $i = 1, \ldots, k$. This is clearly the solution to the linear programming problem. Since $\mathbf{x}_p$ is an extreme point, we have shown that if P has a solution, it must have an extreme point solution. □

COROLLARY 5.2. *Problem P has a finite solution if and only if $\mathbf{c}^T\mathbf{d}_i \leq 0$ for all $i = 1, \ldots l$ when $\mathbf{d}_1, \ldots, \mathbf{d}_l$ are the extreme directions of X.*

PROOF. This is implicit in the proof of the theorem. □

COROLLARY 5.3. *Problem P has alternative optimal solutions if there are at least two extreme points $\mathbf{x}_p$ and $\mathbf{x}_q$ so that $\mathbf{c}^T\mathbf{x}_p = \mathbf{c}^T\mathbf{x}_q$ and so that $\mathbf{x}_p$ is the extreme point solution to the linear programming problem.*

PROOF. Suppose that $\mathbf{x}_p$ is the extreme point solution to P identified in the proof of the theorem. Suppose $\mathbf{x}_q$ is another extreme point solution with $\mathbf{c}^T\mathbf{x}_p = \mathbf{c}^T\mathbf{x}_q$. Then every convex combination of $\mathbf{x}_p$ and $\mathbf{x}_q$ is contained in X (since X is convex). Thus every $\mathbf{x}$ with form $\lambda\mathbf{x}_p + (1-\lambda)\mathbf{x}_q$ and $\lambda \in [0, 1]$ has objective function value:

$$\lambda\mathbf{c}^T\mathbf{x}_p + (1-\lambda)\mathbf{c}^T\mathbf{x}_q = \lambda\mathbf{c}^T\mathbf{x}_p + (1-\lambda)\mathbf{c}^T\mathbf{x}_p = \mathbf{c}^T\mathbf{x}_p$$

which is the optimal objective function value, by assumption. □

EXERCISE 48. Let $X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}, \ \ \mathbf{x} \geq \mathbf{0}\}$ and suppose that $\mathbf{d}_1, \ldots \mathbf{d}_l$ are the extreme directions of X (assuming it has any). Show that the problem:

$$\begin{aligned} &\min \ \mathbf{c}^T\mathbf{x} \\ &s.t. \ \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ &\quad\ \ \mathbf{x} \geq \mathbf{0} \end{aligned} \tag{5.6}$$

has a finite optimal solution if (and only if) $\mathbf{c}^T\mathbf{d}_j \geq 0$ for $k = 1, \ldots, l$. [Hint: Modify the proof above using the Cartheodory characterization theorem.]

2. Algorithmic Characterization of Extreme Points

In the previous sections we showed that if a linear programming problem has a solution, then it must have an extreme point solution. The challenge now is to identify some simple way of identifying extreme points. To accomplish this, let us now assume that we write X as:

$$X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}, \ \ \mathbf{x} \geq \mathbf{0}\} \tag{5.7}$$

Our work in the previous sections shows that this is possible. Recall we can separate $\mathbf{A}$ into an $m \times m$ matrix B and an $m \times (n-m)$ matrix N and we have the result:

$$\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{N}\mathbf{x_N} \tag{5.8}$$

We know that $\mathbf{B}$ is invertible since we assumed that $\mathbf{A}$ had full row rank. If we assume that $\mathbf{x_N} = \mathbf{0}$, then the solution

$$\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} \tag{5.9}$$

was called a *basic solution* (See Definition 3.48.) Clearly any basic solution satisfies the constraints $\mathbf{Ax} = \mathbf{b}$ but it may not satisfy the constraints $\mathbf{x} \geq \mathbf{0}$.

DEFINITION 5.4 (Basic Feasible Solution). If $\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b}$ and $\mathbf{x_N} = \mathbf{0}$ is a basic solution to $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x_B} \geq 0$, then the solution $(\mathbf{x_B}, \mathbf{x_N})$ is called *basic feasible solution.*

THEOREM 5.5. *Every basic feasible solution is an extreme point of X. Likewise, every extreme point is characterized by a basic feasible solution of $\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$.*

PROOF. Since $\mathbf{Ax} = \mathbf{Bx_B} + \mathbf{Nx_N} = \mathbf{b}$ this represents the intersection of m linearly independent hyperplanes (since the rank of $\mathbf{A}$ is m). The fact that $\mathbf{x_N} = \mathbf{0}$ and $\mathbf{x_N}$ contains $n - m$ variables, then we have $n - m$ binding, linearly independent hyperplanes in $\mathbf{x_N} \geq 0$. Thus the point $(\mathbf{x_B}, \mathbf{x_N})$ is the intersection of $m + (n - m) = n$ linearly independent hyperplanes. By Theorem 4.31 we know that $(\mathbf{x_B}, \mathbf{x_N})$ must be an extreme point of X.

Conversely, let $\mathbf{x}$ be an extreme point of X. Clearly $\mathbf{x}$ is feasible and by Theorem 4.31 it must represent the intersection of n hyperplanes. The fact that $\mathbf{x}$ is feasible implies that $\mathbf{Ax} = \mathbf{b}$. This accounts for m of the intersecting linearly independent hyperplanes. The remaining $n - m$ hyperplanes must come from $\mathbf{x} \geq \mathbf{0}$. That is, $n - m$ variables are zero. Let $\mathbf{x_N} = \mathbf{0}$ be the variables for which $\mathbf{x} \geq \mathbf{0}$ are binding. Denote the remaining variables $\mathbf{x_B}$. We can see that $\mathbf{A} = [\mathbf{B}|\mathbf{N}]$ and that $\mathbf{Ax} = \mathbf{Bx_B} + \mathbf{Nx_N} = \mathbf{b}$. Clearly, $\mathbf{x_B}$ is the unique solution to $\mathbf{Bx_B} = \mathbf{b}$ and thus $(\mathbf{x_B}, \mathbf{x_N})$ is a basic feasible solution. □

3. The Simplex Algorithm–Algebraic Form

In this section, we will develop the simplex algorithm algebraically. The idea behind the simplex algorithm is as follows:

(1) Convert the linear program to standard form.
(2) Obtain an initial basic feasible solution (if possible).
(3) Determine whether the basic feasible solution is optimal. If yes, stop.
(4) If the current basic feasible solution is not optimal, then determine which non-basic variable (zero valued variable) should become basic (become non-zero) and which basic variable (non-zero valued variable) should become non-basic (go to zero) to make the objective function value better.
(5) Determine whether the problem is unbounded. If yes, stop.
(6) If the problem doesn't seem to be unbounded at this stage, find a new basic feasible solution from the old basic feasible solution. Go back to Step 3.

Suppose we have a basic feasible solution $\mathbf{x} = (\mathbf{x_B}, \mathbf{x_N})$. We can divide the cost vector $\mathbf{c}$ into its basic and non-basic parts, so we have $\mathbf{c} = [\mathbf{c_B}|\mathbf{c_N}]^T$. Then the objective function becomes:

$$\mathbf{c}^T\mathbf{x} = \mathbf{c}_\mathbf{B}^T\mathbf{x_B} + \mathbf{c}_\mathbf{N}^T\mathbf{x_N} \tag{5.10}$$

We can substitute Equation 5.8 into Equation 5.10 to obtain:

$$\mathbf{c}^T\mathbf{x} = \mathbf{c}_\mathbf{B}^T\left(\mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{Nx_N}\right) + \mathbf{c_N}\mathbf{x_N} = \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} + \left(\mathbf{c}_\mathbf{N}^T - \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N}\right)\mathbf{x}_N \tag{5.11}$$

Let $\mathcal{J}$ be the set of indices of non-basic variables. Then we can write Equation 5.11 as:

$$z(x_1, \ldots, x_n) = \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} + \sum_{j \in \mathcal{J}} \left(\mathbf{c}_j - \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{A}_{\cdot j}\right) x_j \tag{5.12}$$

Consider now the fact $x_j = 0$ for all $j \in \mathcal{J}$. Further, we can see that:

$$\frac{\partial z}{\partial x_j} = \mathbf{c}_j - \mathbf{c}_\mathbf{B}^T \mathbf{B}^{-1} \mathbf{A}_{\cdot j} \tag{5.13}$$

This means that if $\mathbf{c}_j - \mathbf{c}_\mathbf{B}^T \mathbf{B}^{-1} \mathbf{A}_{\cdot j} > 0$ and we *increase* x_j from zero to some new value, then we will *increase* the value of the objective function. For historic reasons, we actually consider the value $\mathbf{c}_\mathbf{B}^T \mathbf{B}^{-1} \mathbf{A}_{\cdot j} - \mathbf{c}_j$, called the *reduced cost* and denote it as:

$$-\frac{\partial z}{\partial x_j} = z_j - c_j = \mathbf{c}_\mathbf{B}^T \mathbf{B}^{-1} \mathbf{A}_{\cdot j} - \mathbf{c}_j \tag{5.14}$$

In a maximization problem, we chose non-basic variables x_j with negative reduced cost to become basic because, in this case, $\partial z / \partial x_j$ is *positive*.

Assume we chose x_j, a non-basic variable to become non-zero (because $z_j - c_j < 0$). We wish to know which of the basic variables will become zero as we *increase* x_j away from zero. We must also be very careful that *none* of the variables become negative as we do this.

By Equation 5.8 we know that the only current basic variables will be affected by increasing x_j. Let us focus explicitly on Equation 5.8 where we include only variable x_j (since all other non-basic variables are kept zero). Then we have:

$$\mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} - \mathbf{B}^{-1}\mathbf{A}_{\cdot j} x_j \tag{5.15}$$

Let $\overline{\mathbf{b}} = \mathbf{B}^{-1}\mathbf{b}$ be an $m \times 1$ column vector and let that $\overline{\mathbf{a}_j} = \mathbf{B}^{-1}\mathbf{A}_{\cdot j}$ be another $m \times 1$ column. Then we can write:

$$\mathbf{x_B} = \overline{\mathbf{b}} - \overline{\mathbf{a}}_j x_j \tag{5.16}$$

Let $\overline{\mathbf{b}} = [\overline{b}_1, \ldots \overline{b}_m]^T$ and $\overline{\mathbf{a}}_j = [\overline{a}_{j_1}, \ldots, \overline{a}_{j_m}]$, then we have:

$$\begin{bmatrix} x_{B_1} \\ x_{B_2} \\ \vdots \\ x_{B_m} \end{bmatrix} = \begin{bmatrix} \overline{b}_1 \\ \overline{b}_2 \\ \vdots \\ \overline{b}_m \end{bmatrix} - \begin{bmatrix} \overline{a}_{j_1} \\ \overline{b}_{j_2} \\ \vdots \\ \overline{b}_{j_m} \end{bmatrix} x_j = \begin{bmatrix} \overline{b}_1 - \overline{a}_{j_1} x_j \\ \overline{b}_2 - \overline{a}_{j_2} x_j \\ \vdots \\ \overline{b}_m - \overline{a}_{j_m} x_j \end{bmatrix} \tag{5.17}$$

We know (a priori) that $\overline{b}_i \geq 0$ for $i = 1, \ldots, m$. If $\overline{a}_{j_i} \leq 0$, then as we increase x_j, $\overline{b}_i - \overline{a}_{j_i} \geq 0$ no matter how large we make x_j. On the other hand, if $\overline{a}_{j_i} > 0$, then as we increase x_j we know that $\overline{b}_i - \overline{a}_{j_i} x_j$ will get smaller and eventually hit zero. In order to ensure that *all* variables remain non-negative, we cannot increase x_j beyond a certain point.

For each i ($i = 1, \ldots, m$) such that $\overline{a}_{j_i} > 0$, the value of x_j that will make x_{B_i} goto 0 can be found by observing that:

$$x_{B_i} = \overline{b}_i - \overline{a}_{j_i} x_j \tag{5.18}$$

and if $x_{B_i} = 0$, then we can solve:

$$0 = \overline{b}_i - \overline{a}_{j_i} x_j \implies x_j = \frac{\overline{b}_i}{\overline{a}_{j_i}} \tag{5.19}$$

Thus, the *largest possible value* we can assign x_j and ensure that all variables remain positive is:

$$\min \left\{ \frac{\overline{b}_i}{\overline{a}_{j_i}} : i = 1, \ldots, m \text{ and } a_{j_i} > 0 \right\} \tag{5.20}$$

Expression 5.20 is called the *minimum ratio test*. We are interested in which index i is the minimum ratio.

Suppose that in executing the minimum ratio test, we find that $x_j = \overline{b}_k/\overline{a}_{j_k}$. The variable x_j (which was non-basic) becomes basic and the variable $x_{\mathbf{B}_k}$ becomes non-basic. All other basic variables remain basic (and positive). In executing this procedure (of exchanging one basic variable and one non-basic variable) we have moved from one extreme point of X to another.

THEOREM 5.6. *If $z_j - c_j \geq 0$ for all $j \in \mathcal{J}$, then the current basic feasible solution is optimal.*

PROOF. We have already shown in Theorem 5.1 that if a linear programming problem has an optimal solution, then it occurs at an extreme point and we've shown in Theorem 5.5 that there is a one-to-one correspondence between extreme points and basic feasible solutions. If $z_j - c_j \geq 0$ for all $j \in \mathcal{J}$, then $\partial z/\partial x_j \leq 0$ for all non-basic variables x_j. That is, we cannot increase the value of the objective function by increasing the value of any non-basic variable. Thus, since moving to another basic feasible solution (extreme point) will not improve the objective function, it follows we must be at the optimal solution. □

THEOREM 5.7. *In a maximization problem, if $\overline{a}_{j_i} \leq 0$ for all $i = 1, \ldots, m$, and $z_j - c_j < 0$, then the linear programming problem is unbounded.*

PROOF. The fact that $z_j - c_j < 0$ implies that increasing x_j will improve the value of the objective function. Since $\overline{a}_{j_i} < 0$ for all $i = 1, \ldots, m$, we can increase x_j indefinitely without violating feasibility (no basic variable will ever go to zero). Thus the objective function can be made as large as we like. □

REMARK 5.8. We should note that in executing the exchange of one basic variable and one non-basic variable, we must be *very* careful to ensure that the resulting basis consist of m linearly independent columns of the original matrix $\mathbf{A}$. The conditions for this are provided in Lemma 3.39. Specifically, we must be able to write the column corresponding to x_j, the entering variable, as a linear combination of the columns of $\mathbf{B}$ so that:

$$\alpha_1 \mathbf{b}_1 + \ldots \alpha_m \mathbf{b}_m = \mathbf{A}_{\cdot j} \tag{5.21}$$

and further if we are exchanging x_j for $\mathbf{x}_{\mathbf{B}_i}$ $(i = 1, \ldots, m)$, then $\alpha_i \neq 0$.

We can see this from the fact that $\overline{\mathbf{a}}_j = \mathbf{B}^{-1}\mathbf{A}_{\cdot j}$ and therefore:

$$\mathbf{B}\overline{\mathbf{a}}_j = \mathbf{A}_{\cdot j}$$

and therefore we have:

$$\mathbf{A}_{\cdot j} = \mathbf{B}_{\cdot 1}\overline{\mathbf{a}}_{j_1} + \cdots + \mathbf{B}_{\cdot m}\overline{\mathbf{a}}_{j_m}$$

which shows how to write the column $\mathbf{A}_{\cdot j}$ as a linear combination of the columns of $\mathbf{B}$.

EXERCISE 49. Consider the linear programming problem given in Exercise 48. Under what conditions should a non-basic variable enter the basis? State and prove an analogous theorem to Theorem 5.6 using your observation. [Hint: Use the definition of reduced cost. Remember that it is $-\partial z/\partial x_j$.]

EXAMPLE 5.9. Consider the Toy Maker Problem (from Example 2.3). The linear programming problem given in Equation 2.8 is:

$$\begin{cases} \max \ z(x_1, x_2) = 7x_1 + 6x_2 \\ s.t. \ 3x_1 + x_2 \leq 120 \\ \quad x_1 + 2x_2 \leq 160 \\ \quad x_1 \leq 35 \\ \quad x_1 \geq 0 \\ \quad x_2 \geq 0 \end{cases}$$

We can convert this problem to standard form by introducing the slack variables s_1, s_2 and s_3:

$$\begin{cases} \max \ z(x_1, x_2) = 7x_1 + 6x_2 \\ s.t. \ 3x_1 + x_2 + s_1 = 120 \\ \quad x_1 + 2x_2 + s_2 = 160 \\ \quad x_1 + s_3 = 35 \\ \quad x_1, x_2, s_1, s_2, s_3 \geq 0 \end{cases}$$

which yields the matrices

$$\mathbf{c} = \begin{bmatrix} 7 \\ 6 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 3 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 120 \\ 160 \\ 35 \end{bmatrix}$$

We can begin with the matrices:

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \\ 1 & 0 \end{bmatrix}$$

In this case we have:

$$\mathbf{x_B} = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \quad \mathbf{x_N} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{c_B} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{c_N} = \begin{bmatrix} 7 \\ 6 \end{bmatrix}$$

and

$$\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 120 \\ 160 \\ 35 \end{bmatrix} \quad \mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \\ 1 & 0 \end{bmatrix}$$

Therefore:

$$\mathbf{c_B}^T\mathbf{B}^{-1}\mathbf{b} = 0 \quad \mathbf{c_B}^T\mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 0 & 0 \end{bmatrix} \quad \mathbf{c_B}^T\mathbf{B}^{-1}\mathbf{N} - \mathbf{c_N} = \begin{bmatrix} -7 & -6 \end{bmatrix}$$

Using this information, we can compute:

$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{A}_{\cdot 1} - \mathbf{c}_1 = -7$$
$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{A}_{\cdot 2} - \mathbf{c}_2 = -6$$

and therefore:

$$\frac{\partial z}{\partial x_1} = 7 \text{ and } \frac{\partial z}{\partial x_2} = 6$$

Based on this information, we could chose either x_1 or x_2 to enter the basis and the value of the objective function would increase. If we chose x_1 to enter the basis, then we must determine which variable will leave the basis. To do this, we must investigate the elements of $\mathbf{B}^{-1}\mathbf{A}_{\cdot 1}$ and the current basic feasible solution $\mathbf{B}^{-1}\mathbf{b}$. Since each element of $\mathbf{B}^{-1}\mathbf{A}_{\cdot 1}$ is positive, we must perform the minimum ratio test on each element of $\mathbf{B}^{-1}\mathbf{A}_{\cdot 1}$. We know that $\mathbf{B}^{-1}\mathbf{A}_{\cdot 1}$ is just the first column of $\mathbf{B}^{-1}\mathbf{N}$ which is:

$$\mathbf{B}^{-1}\mathbf{A}_{\cdot 1} = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$$

Performing the minimum ratio test, we see have:

$$\min\left\{\frac{120}{3}, \frac{160}{1}, \frac{35}{1}\right\}$$

In this case, we see that index 3 (35/1) is the minimum ratio. Therefore, variable x_1 will enter the basis and variable s_3 will leave the basis. The new basic and non-basic variables will be:

$$\mathbf{x_B} = \begin{bmatrix} s_1 \\ s_2 \\ x_1 \end{bmatrix} \quad \mathbf{x_N} = \begin{bmatrix} s_3 \\ x_2 \end{bmatrix} \quad \mathbf{c_B} = \begin{bmatrix} 0 \\ 0 \\ 7 \end{bmatrix} \quad \mathbf{c_N} = \begin{bmatrix} 0 \\ 6 \end{bmatrix}$$

and the matrices become:

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 0 & 1 \\ 0 & 2 \\ 1 & 0 \end{bmatrix}$$

Note we have simply swapped the column corresponding to x_1 with the column corresponding to s_3 in the basis matrix $\mathbf{B}$ and the non-basic matrix $\mathbf{N}$. We will do this repeatedly in the example and we recommend the reader keep track of which variables are being exchanged and why certain columns in $\mathbf{B}$ are being swapped with those in $\mathbf{N}$.

Using the new $\mathbf{B}$ and $\mathbf{N}$ matrices, the derived matrices are then:

$$\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 15 \\ 125 \\ 35 \end{bmatrix} \quad \mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} -3 & 1 \\ -1 & 2 \\ 1 & 0 \end{bmatrix}$$

The cost information becomes:

$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} = 245 \quad \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 7 & 0 \end{bmatrix} \quad \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N} - \mathbf{c_N} = \begin{bmatrix} 7 & -6 \end{bmatrix}$$

using this information, we can compute:

$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{A}_{.5} - \mathbf{c}_5 = 7$$
$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{A}_{.2} - \mathbf{c}_2 = -6$$

Based on this information, we can only choose x_2 to enter the basis to ensure that the value of the objective function increases. We can perform the minimum ration test to figure out which basic variable will leave the basis. We know that $\mathbf{B}^{-1}\mathbf{A}_{.2}$ is just the second column of $\mathbf{B}^{-1}\mathbf{N}$ which is:

$$\mathbf{B}^{-1}\mathbf{A}_{.2} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

Performing the minimum ratio test, we see have:

$$\min\left\{\frac{15}{1}, \frac{125}{2}\right\}$$

In this case, we see that index 1 (15/1) is the minimum ratio. Therefore, variable x_2 will enter the basis and variable s_1 will leave the basis. The new basic and non-basic variables will be: The new basic and non-basic variables will be:

$$\mathbf{x_B} = \begin{bmatrix} x_2 \\ s_2 \\ x_1 \end{bmatrix} \quad \mathbf{x_N} = \begin{bmatrix} s_3 \\ s_1 \end{bmatrix} \quad \mathbf{c_B} = \begin{bmatrix} 6 \\ 0 \\ 7 \end{bmatrix} \quad \mathbf{c_N} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and the matrices become:

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

The derived matrices are then:

$$\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 15 \\ 95 \\ 35 \end{bmatrix} \quad \mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} -3 & 1 \\ 5 & -2 \\ 1 & 0 \end{bmatrix}$$

The cost information becomes:

$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} = 335 \quad \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} -11 & 6 \end{bmatrix} \quad \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N} - \mathbf{c_N} = \begin{bmatrix} -11 & 6 \end{bmatrix}$$

Based on this information, we can only choose s_3 to (re-enter) the basis to ensure that the value of the objective function increases. We can perform the minimum ration test to figure out which basic variable will leave the basis. We know that $\mathbf{B}^{-1}\mathbf{A}_{.5}$ is just the fifth column of $\mathbf{B}^{-1}\mathbf{N}$ which is:

$$\mathbf{B}^{-1}\mathbf{A}_{.5} = \begin{bmatrix} -3 \\ 5 \\ 1 \end{bmatrix}$$

Performing the minimum ratio test, we see have:

$$\min\left\{\frac{95}{5}, \frac{35}{1}\right\}$$

In this case, we see that index 2 (95/5) is the minimum ratio. Therefore, variable s_3 will enter the basis and variable s_2 will leave the basis. The new basic and non-basic variables will be:

$$\mathbf{x_B} = \begin{bmatrix} x_2 \\ s_3 \\ x_1 \end{bmatrix} \quad \mathbf{x_N} = \begin{bmatrix} s_2 \\ s_1 \end{bmatrix} \quad \mathbf{c_B} = \begin{bmatrix} 6 \\ 0 \\ 7 \end{bmatrix} \quad \mathbf{c_N} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and the matrices become:

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 3 \\ 2 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}$$

The derived matrices are then:

$$\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 72 \\ 19 \\ 16 \end{bmatrix} \quad \mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 6/10 & -1/5 \\ 1/5 & -2/5 \\ -1/5 & 2/5 \end{bmatrix}$$

The cost information becomes:

$$\mathbf{c_B^T}\mathbf{B}^{-1}\mathbf{b} = 544 \quad \mathbf{c_B^T}\mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 11/5 & 8/5 \end{bmatrix} \quad \mathbf{c_B^T}\mathbf{B}^{-1}\mathbf{N} - \mathbf{c_N} = \begin{bmatrix} 11/5 & 8/5 \end{bmatrix}$$

Since the reduced costs are now positive, we can conclude that we've obtained an optimal solution because no improvement is possible. The final solution then is:

$$\mathbf{x_B}^* = \begin{bmatrix} x_2 \\ s_3 \\ x_1 \end{bmatrix} = \mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 72 \\ 19 \\ 16 \end{bmatrix}$$

Simply, we have $x_1 = 16$ and $x_2 = 72$ as we obtained in Example 2.3. The path of extreme points we actually took in traversing the boundary of the polyhedral feasible region is shown in Figure 5.1.

EXERCISE 50. Assume that a leather company manufactures two types of belts: regular and deluxe. Each belt requires 1 square yard of leather. A regular belt requires 1 hour of skilled labor to produce, while a deluxe belt requires 2 hours of labor. The leather company receives 40 square yards of leather each week and a total of 60 hours of skilled labor is available. Each regular belt nets \$3 in profit, while each deluxe belt nets \$5 in profit. The company wishes to maximize profit.

(1) Ignoring the divisibility issues, construct a linear programming problem whose solution will determine the number of each type of belt the company should produce.
(2) Use the simplex algorithm to solve the problem you stated above remembering to convert the problem to *standard form* before you begin.
(3) Draw the feasible region and the level curves of the objective function. Verify that the optimal solution you obtained through the simplex method is the point at which the level curves no longer intersect the feasible region in the direction following the gradient of the objective function.

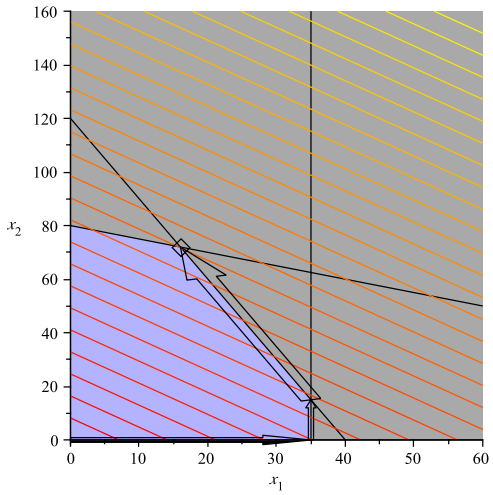


Figure 5.1. The Simplex Algorithm: The path around the feasible region is shown in the figure. Each exchange of a basic and non-basic variable moves us along an edge of the polygon in a direction that increases the value of the objective function.

4. Simplex Method–Tableau Form

No one executes the simplex algorithm in algebraic form. Instead, several representations (tableau representations) have been developed to lesson the amount of writing that needs to be done and to collect all pertinent information into a single table.

To see how a *Simplex Tableau* is derived, consider Problem P in standard form:

$$P\begin{cases} \max & \mathbf{c}^T\mathbf{x} \\ s.t. & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{cases}$$

We can re-write P in an unusual way by introducing a new variable z and separating $\mathbf{A}$ into its basic and non-basic parts to obtain:

$$\begin{aligned} \max \;& z \\ s.t. \;& z - \mathbf{c}_\mathbf{B}^T\mathbf{x}_\mathbf{B} - \mathbf{c}_\mathbf{N}^T\mathbf{x}_\mathbf{N} = 0 \\ & \mathbf{B}\mathbf{x}_\mathbf{B} + \mathbf{N}\mathbf{x}_\mathbf{N} = \mathbf{b} \\ & \mathbf{x}_\mathbf{B}, \mathbf{x}_\mathbf{N} \geq 0 \end{aligned} \tag{5.22}$$

From the second equation, it's clear

$$\mathbf{x}_\mathbf{B} + \mathbf{B}^{-1}\mathbf{N}\mathbf{x}_\mathbf{N} = \mathbf{B}^{-1}\mathbf{b} \tag{5.23}$$

We can multiply this equation by $\mathbf{c}_\mathbf{B}^\mathbf{T}$ to obtain:

$$\mathbf{c}_\mathbf{B}^T\mathbf{x}_\mathbf{B} + \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N}\mathbf{x}_\mathbf{N} = \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} \tag{5.24}$$

If we add this equation to the equation $z - \mathbf{c}_\mathbf{B}^T\mathbf{x}_\mathbf{B} - \mathbf{c}_\mathbf{N}^T\mathbf{x}_\mathbf{N} = 0$ we obtain:

$$z + \mathbf{0}^T\mathbf{x}_\mathbf{B} + \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N}\mathbf{x}_\mathbf{N} - \mathbf{c}_\mathbf{N}^T\mathbf{x}_\mathbf{N} = \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} \tag{5.25}$$

Here $\mathbf{0}$ is the vector of zeros of appropriate size. This equation can be written as:

$$z + \mathbf{0}^T\mathbf{x_B} + \left(\mathbf{c_B^T B^{-1} N} - \mathbf{c_N^T}\right)\mathbf{x_N} = \mathbf{c_B^T B^{-1} b} \tag{5.26}$$

We can now represent this set of equations as a large matrix (or tableau):

	z	$\mathbf{x_B}$	$\mathbf{x_N}$	RHS	
z	1	$\mathbf{0}$	$\mathbf{c_B^T B^{-1} N} - \mathbf{c_N^T}$	$\mathbf{c_B^T B^{-1} b}$	Row 0
$\mathbf{x_B}$	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{B^{-1} N}$	$\mathbf{B^{-1} b}$	Rows 1 through m

The augmented matrix shown within the table:

$$\left[\begin{array}{cc|c|c} 1 & \mathbf{0} & \mathbf{c_B^T B^{-1} N} - \mathbf{c_N^T} & \mathbf{c_B^T B^{-1} b} \\ \mathbf{0} & \mathbf{1} & \mathbf{B^{-1} N} & \mathbf{B^{-1} b} \end{array}\right] \tag{5.27}$$

is simply the matrix representation of the simultaneous equations described by Equations 5.23 and 5.26. We can see that the first row consists of a row of the first row of the $(m+1) \times (m+1)$ identity matrix, the reduced costs of the non-basic variables and the current objective function values. The remainder of the rows consist of the rest of the $(m+1) \times (m+1)$ identity matrix, the matrix $\mathbf{B^{-1} N}$ and $\mathbf{B^{-1} b}$ the current non-zero part of the basic feasible solution.

This matrix representation (or tableau representation) contains all of the information we need to execute the simplex algorithm. An entering variable is chosen from among the columns containing the reduced costs and matrix $\mathbf{B^{-1} N}$. Naturally, a column with a negative reduced cost is chosen. We then chose a leaving variable by performing the minimum ratio test on the chosen column and the right-hand-side (RHS) column. We pivot on the element at the entering column and leaving row and this transforms the tableau into a new tableau that represents the new basic feasible solution.

Example 5.10. Again, consider the toy maker problem. We will execute the simplex algorithm using the tableau method. Our problem in standard form is given as:

$$\begin{cases} \max \quad z(x_1, x_2) = 7x_1 + 6x_2 \\ \quad s.t. \quad 3x_1 + x_2 + s_1 = 120 \\ \qquad\quad x_1 + 2x_2 + s_2 = 160 \\ \qquad\qquad x_1 + s_3 = 35 \\ \qquad\quad x_1, x_2, s_1, s_2, s_3 \geq 0 \end{cases}$$

We can assume our initial basic feasible solution has s_1, s_2 and s_3 as basic variables and x_1 and x_2 as non-basic variables. Thus our initial tableau is simply:

$$\begin{array}{c} \\ z \\ s_1 \\ s_2 \\ s_3 \end{array} \left[\begin{array}{c|cccccc|c} z & x_1 & x_2 & s_1 & s_2 & s_3 & \text{RHS} \\ \hline 1 & -7 & -6 & 0 & 0 & 0 & 0 \\ \hline 0 & 3 & 1 & 1 & 0 & 0 & 120 \\ 0 & 1 & 2 & 0 & 1 & 0 & 160 \\ 0 & 1 & 0 & 0 & 0 & 1 & 35 \end{array}\right] \tag{5.28}$$

Note that the columns have been swapped so that the identity matrix is divided and $\mathbf{B^{-1} N}$ is located in columns 2 and 3. This is because of our choice of basic variables. The reduced cost vector is in Row 0.

Using this information, we can see that either x_1 or x_2 can enter. We can compute the minimum ratio test (MRT) next to the RHS column. If we chose x_2 as the entering variable, then the MRT tells us s_2 will leave. We put a box around the element on which we will pivot:

(5.29)

	z	x_1	x_2	s_1	s_2	s_3	RHS	MRT (x_2)
z	1	−7	−6	0	0	0	0	
s_1	0	3	1	1	0	0	120	120
s_2	0	1	$\boxed{2}$	0	1	0	160	80
s_3	0	1	0	0	0	1	35	−

If we pivot on this element, then we transform the column corresponding to x_2 into the identity column:

(5.30)
$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

This process will correctly compute the new reduced costs and $\mathbf{B}^{-1}$ matrix as well as the new cost information. The new tableau becomes:

(5.31)

	z	x_1	x_2	s_1	s_2	s_3	RHS
z	1	−4	0	0	3	0	480
s_1	0	2.5	0	1	−0.5	0	40
x_2	0	0.5	1	0	0.5	0	80
s_3	0	1	0	0	0	1	35

We can see that x_1 is a valid entering variable, as it has a negative reduced cost (−4). We can again place the minimum ratio test values on the right-hand-side of the matrix to obtain:

(5.32)

	z	x_1	x_2	s_1	s_2	s_3	RHS	MRT (x_1)
z	1	−4	0	0	3	0	480	
s_1	0	$\boxed{2.5}$	0	1	−0.5	0	40	16
x_2	0	0.5	1	0	0.5	0	80	160
s_3	0	1	0	0	0	1	35	35

We now pivot on the element we have boxed to obtain the new tableau[1]:

(5.33)

	z	x_1	x_2	s_1	s_2	s_3	RHS
z	1	0	0	1.6	2.2	0	544
x_1	0	1	0	0.4	−0.2	0	16
x_2	0	0	1	−0.2	0.6	0	72
s_3	0	0	0	−0.4	0.2	1	19

All the reduced costs of the non-basic variables (s_1 and s_2) are positive and so this is the optimal solution to the linear programming problem. We can also see that this solution agrees with our previous computations on the Toy Maker Problem.

5. Identifying Unboundedness

We have already identified a theorem for detecting unboundedness. Recall Theorem 5.7: *In a maximization problem, if $\overline{a}_{j_i} < 0$ for all $i = 1, \ldots, m$, and $z_j - c_j < 0$, then the linear programming problem is unbounded.*

This condition occurs when a variable x_j should enter the basis because $\partial z/\partial x_j > 0$ and there is no blocking basis variable. That is, we can arbitrarily increase the value of x_j without causing any variable to become negative. We give an example:

EXAMPLE 5.11. Consider the Linear programming problem from Example 2.9:

$$\begin{cases} \max & z(x_1, x_2) = 2x_1 - x_2 \\ s.t. & x_1 - x_2 \leq 1 \\ & 2x_1 + x_2 \geq 6 \\ & x_1, x_2 \geq 0 \end{cases}$$

We can convert this problem into standard form by adding a slack variable s_1 and a surplus variable s_2:

$$\begin{cases} \max & z(x_1, x_2) = 2x_1 - x_2 \\ s.t. & x_1 - x_2 + s_1 = 1 \\ & 2x_1 + x_2 - s_2 = 6 \\ & x_1, x_2, s_1, s_2 \geq 0 \end{cases}$$

This yields the matrices:

$$\mathbf{c} = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ s_1 \\ s_2 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} 1 & -1 & 1 & 0 \\ 2 & 1 & 0 & -1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 6 \end{bmatrix}$$

We have both slack and surplus variables, so the case when $x_1 = x_2 = 0$ is not a valid initial solution. We can chose a valid solution based on our knowledge of the problem. Assume that $s_1 = s_2 = 0$ and so we have:

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

In this case we have:

$$\mathbf{x_B} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{x_N} = \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \quad \mathbf{c_B} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \mathbf{c_N} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This yields:

$$\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 7/3 \\ 4/3 \end{bmatrix} \quad \mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 1/3 & -1/3 \\ -2/3 & -1/3 \end{bmatrix}$$

We also have the cost information:

$$\mathbf{c_B}\mathbf{B}^{-1}\mathbf{b} = \frac{10}{3} \quad \mathbf{c_B}\mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} \frac{4}{3} & -\frac{1}{3} \end{bmatrix} \quad \mathbf{c_B}\mathbf{B}^{-1}\mathbf{N} - \mathbf{c_N} = \begin{bmatrix} \frac{4}{3} & -\frac{1}{3} \end{bmatrix}$$

Based on this information, we can construct the tableau for this problem as:

$$(5.34)\qquad \begin{array}{c} \\ z \\ x_1 \\ x_2 \end{array} \left[\begin{array}{c|cccc|c} z & x_1 & x_2 & s_1 & s_2 & \text{RHS} \\ \hline 1 & 0 & 0 & \frac{4}{3} & \frac{-1}{3} & \frac{10}{3} \\ \hline 0 & 1 & 0 & \frac{1}{3} & \frac{-1}{3} & \frac{7}{3} \\ 0 & 0 & 1 & \frac{-2}{3} & \frac{-1}{3} & \frac{4}{3} \end{array}\right]$$

We see that s_2 should enter the basis because $\mathbf{c_B B^{-1} A}_{\cdot 4} - \mathbf{c}_4 < 0$. But the column corresponding to s_2 in the tabluau is all negative. Therefore there is no minimum ratio test. We can let s_2 become as large as we like and we will keep increasing the objective function without violating feasibility.

What we have shown is that the ray with vertex

$$\mathbf{x}_0 = \begin{bmatrix} 7/3 \\ 4/3 \\ 0 \\ 0 \end{bmatrix}$$

and direction:

$$\mathbf{d} = \begin{bmatrix} 1/3 \\ 1/3 \\ 0 \\ 1 \end{bmatrix}$$

is entirely contained inside the polyhedral set defined by $\mathbf{Ax} = \mathbf{b}$. This can be see from the fact that:

$$\mathbf{x_B} = \mathbf{B^{-1}b} - \mathbf{B^{-1}Nx_N}$$

When applied in this case, we have:

$$\mathbf{x_B} = \mathbf{B^{-1}b} - \mathbf{B^{-1}A}_{\cdot 4} s_2$$

We know that

$$-\mathbf{B^{-1}A}_{\cdot 4} = \begin{bmatrix} 1/3 \\ 1/3 \end{bmatrix}$$

We will be increasing s_2 (which acts like λ in the definition of ray) and leaving s_1 equal to 0. It's now easy to see that the ray we described is contained entirely in the feasible region. This is illustrated in the original constraints in Figure 5.2.

Based on our previous example, we have the following theorem that extends Theorem 5.7:

THEOREM 5.12. *In a maximization problem, if $\overline{a}_{j_i} \leq 0$ for all $i = 1, \ldots, m$, and $z_j - c_j < 0$, then the linear programming problem is unbounded furthermore, let $\overline{\mathbf{a}}_j$ be the j^{th} column of $\mathbf{B^{-1}A}_{\cdot j}$ and let $\mathbf{e}_k$ be a standard basis column vector in $\mathbb{R}^{m \times (n-m)}$ where k corresponds to the position of j in the matrix $\mathbf{N}$. Then the direction:*

$$(5.35)\qquad \mathbf{d} = \begin{bmatrix} -\overline{\mathbf{a}}_j \\ \mathbf{e}_k \end{bmatrix}$$

is an extreme direction of the feasible region $X = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{Ax} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\}$.

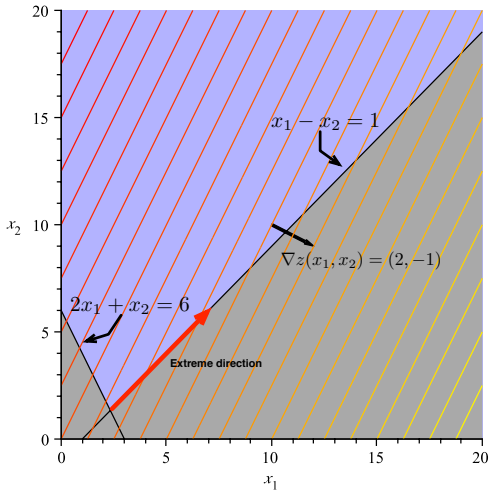


Figure 5.2. Unbounded Linear Program: The existence of a negative column $\overline{\mathbf{a}}_j$ in the simplex tableau for entering variable x_j indicates an unbounded problem and feasible region. The recession direction is shown in the figure.

PROOF. The fact that $\mathbf{d}$ is a direction is easily verified by the fact there is an extreme point $\mathbf{x} = [\mathbf{x_B}\ \mathbf{x_N}]^T$ and for all $\lambda \geq 0$ we have:

$$\mathbf{x} + \lambda\mathbf{d} \in X \tag{5.36}$$

Thus it follows from the proof of Theorem 4.24 that $\mathbf{Ad} \leq \mathbf{0}$. The fact that $\mathbf{d} \geq 0$ and $\mathbf{d} \neq 0$ follows from our assumptions. Now, we know that we can write $\mathbf{A} = [\mathbf{B}|\mathbf{N}]$. Further, we know that $\overline{\mathbf{a}}_j = \mathbf{B}^{-1}\mathbf{A}_{\cdot j}$. Let us consider $\mathbf{Ad}$:

$$\mathbf{Ad} = [\mathbf{B}|\mathbf{N}] \begin{bmatrix} -\overline{\mathbf{a}}_j \\ \mathbf{e}_k \end{bmatrix} = -\mathbf{B}\mathbf{B}^{-1}\mathbf{A}_{\cdot j} + \mathbf{N}\mathbf{e}_k \tag{5.37}$$

Remember, $\mathbf{e}_k$ is the standard basis vector that has have 1 precisely in the position corresponding to column $\mathbf{A}_{\cdot j}$ in matrix $\mathbf{N}$, so $\mathbf{A}_{\cdot j} = \mathbf{N}\mathbf{e}_j$. Thus we have:

$$-\mathbf{B}\mathbf{B}^{-1}\mathbf{A}_{\cdot j} + \mathbf{N}\mathbf{e}_k = -\mathbf{A}_{\cdot j} + \mathbf{A}_{\cdot j} = \mathbf{0} \tag{5.38}$$

Thus, $\mathbf{Ad} = \mathbf{0}$. We can scale $\mathbf{d}$ so that $\mathbf{e}^T\mathbf{d} = 1$. We know that $n - m - 1$ elements of $\mathbf{d}$ are zero (because of $\mathbf{e}_k$) and we know that $\mathbf{Ad} = \mathbf{0}$. Thus $\mathbf{d}$ can be made to represent the intersection of n-hyperplanes in $\mathbb{R}^n$. Thus, $\mathbf{d}$ is an extreme point of the polyhedron $D = \{\mathbf{d} \in \mathbb{R}^n : \mathbf{Ad} \leq \mathbf{0}, \mathbf{d} \geq \mathbf{0}, \mathbf{e}^T\mathbf{d} = 1\}$. It follows from Theorem 4.39, we know that $\mathbf{d}$ is an extreme direction of X. □

EXERCISE 51. Consider the problem

$$\begin{cases} \min & z(x_1, x_2) = 2x_1 - x_2 \\ s.t. & x_1 - x_2 + s_1 = 1 \\ & 2x_1 + x_2 - s_2 = 6 \\ & x_1, x_2, s_1, s_2 \geq 0 \end{cases}$$

Using the rule you developed in Exercise 49, show that the minimization problem has an unbounded feasible solution. Find an extreme direction for this set. [Hint: The minimum ratio test is the same for a minimization problem. Execute the simplex algorithm as we did in Example 5.11 and use Theorem 5.12 to find the extreme direction of the feasible region.]

6. Identifying Alternative Optimal Solutions

We saw in Theorem 5.6 that is $z_j - c_j > 0$ for all $j \in \mathcal{J}$ (the indices of the non-basic variables), then the basic feasible solution generated by the current basis was optimal. Suppose that $z_j - c_j \geq 0$. Then we have a slightly different result:

THEOREM 5.13. *In Problem P for a given set of non-basic variables $\mathcal{J}$, if $z_j - c_j \geq 0$ for all $j \in \mathcal{J}$, then the current basic feasible solution is optimal. Further, if $z_j - c_j = 0$ for at least one $j \in \mathcal{J}$, then there are alternative optimal solutions. Furthermore, let $\overline{\mathbf{a}}_j$ be the j^{th} column of $\mathbf{B}^{-1}\mathbf{A}_{.j}$. Then the solutions to P are:*

$$\begin{cases} \mathbf{x_B} = \mathbf{B}^{-1}\mathbf{b} - \overline{\mathbf{a}}_j x_j \\ x_j \in \left[0, \min\left\{\dfrac{\overline{b}_i}{\overline{a}_{j_i}} : i = 1, \ldots, m, \;\; \overline{a}_{j_i} > 0\right\}\right] \\ x_r = 0, \forall r \in \mathcal{J},\, r \neq j \end{cases} \tag{5.39}$$

PROOF. It follows from the proof of Theorem 5.6 that the solution must be optimal as $\partial z/\partial x_j \leq 0$ for all $j \in \mathcal{J}$ and therefore increasing and x_j will *not* improve the value of the objective function. If there is some $j \in \mathcal{J}$ so that $z_j - c_j = 0$, then $\partial z/\partial x_j = 0$ and we may increase the value of x_j up to some point specified by the minimum ratio test, while keeping other non-basic variables at zero. In this case, we will neither increase nor decrease the objective function value. Since that objective function value is optimal, it follows that the set of all such values (described in Equation 5.39) are alternative optimal solutions. □

EXAMPLE 5.14. Let us consider the toy maker problem again from Example 2.3 and 5.9 with our adjusted objective

$$z(x_1, x_2) = 18x_1 + 6x_2 \tag{5.40}$$

Now consider the penultimate basis from Example 5.9 in which we had as basis variables x_1, s_2 and x_2.

$$\mathbf{x_B} = \begin{bmatrix} x_1 \\ x_2 \\ s_2 \end{bmatrix} \quad \mathbf{x_N} = \begin{bmatrix} s_1 \\ s_3 \end{bmatrix} \quad \mathbf{c_B} = \begin{bmatrix} 18 \\ 6 \\ 0 \end{bmatrix} \quad \mathbf{c_N} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The matrices become:

$$\mathbf{B} = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \mathbf{N} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}$$

The derived matrices are then:

$$\mathbf{B}^{-1}\mathbf{b} = \begin{bmatrix} 35 \\ 15 \\ 95 \end{bmatrix} \quad \mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix} 0 & 1 \\ 1 & -3 \\ -2 & 5 \end{bmatrix}$$

The cost information becomes:

$$\mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{b} = 720 \quad \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N} = \begin{bmatrix}6 & 0\end{bmatrix} \quad \mathbf{c}_\mathbf{B}^T\mathbf{B}^{-1}\mathbf{N} - \mathbf{c}_\mathbf{N} = \begin{bmatrix}6 & 0\end{bmatrix}$$

This yields the tableau:

(5.41)

	z	x_1	x_2	s_1	s_2	s_3	RHS
z	1	0	0	6	0	**0**	720
s_1	0	1	0	0	0	1	35
s_2	0	0	1	1	0	-3	15
s_3	0	0	0	-2	1	5	95

Unlike example 5.9, the reduced cost for s_3 is 0. This means that if we allow s_3 to enter the basis, the objective function value will not change. Performing the minimum ratio test however, we see that s_2 will still leave the basis:

(5.42)

	z	x_1	x_2	s_1	s_2	s_3	RHS	MRT (s_3)
z	1	0	0	6	0	**0**	720	
x_1	0	1	0	0	0	1	35	35
x_2	0	0	1	1	0	-3	15	–
s_2	0	0	0	-2	1	$\boxed{5}$	95	19

Therefore any solution of the form:

$$s_3 \in [0, 19]$$

(5.43)
$$\begin{bmatrix}x_1\\x_2\\s_2\end{bmatrix} = \begin{bmatrix}35\\15\\95\end{bmatrix} - \begin{bmatrix}1\\-3\\5\end{bmatrix} s_3$$

is an optimal solution to the linear programming problem. This precisely describes the edge shown in Figure 5.3.

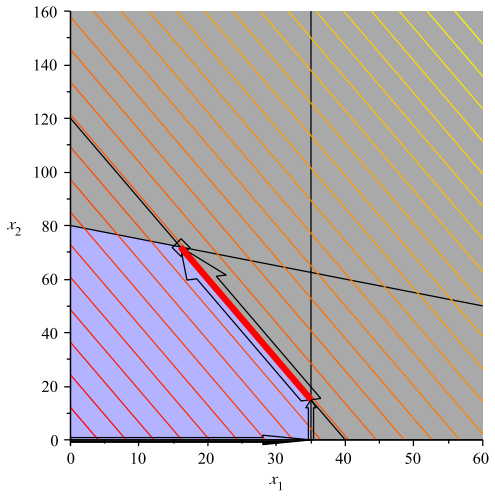


Figure 5.3. Infinite alternative optimal solutions: In the simplex algorithm, when $z_j - c_j \geq 0$ in a maximization problem with at least one j for which $z_j - c_j = 0$, indicates an infinite set of alternative optimal solutions.

Unit 7

Transportation and Assignment problems

Objectives

By the end of this unit you will be able to:

- formulate special linear programming problems using the transportation model.
- define a balanced transportation problem
- develop an initial solution of a transportation problem using the Northwest Corner Rule
- use the Stepping Stone method to find an optimal solution of a transportation problem
- formulate special linear programming problems using the assignment model
- solve assignment problems with the Hungarian method.

Introduction

In this unit we extend the theory of linear programming to two special linear programming problems, the **Transportation** and **Assignment Problems**. Both of these problems can be solved by the simplex algorithm, but the process would result in very large simplex tableaux and numerous simplex iterations.

Because of the special characteristics of each problem, however, alternative solution methods requiring significantly less mathematical manipulation have been developed.

The Transportation problem

The general transportation problem is concerned with determining an optimal strategy for distributing a commodity from a group of supply centres,such as factories, called *sources*, to various receiving centers, such as warehouses, called *destinations*, in such a way as to minimise total distribution costs.

Each source is able to supply a fixed number of units of the product, usually called the *capacity* or *availability*, and each destination has a fixed demand, often called the *requirement*.

Transportation models can also be used when a firm is trying to decide where to locate a new facility. Good financial decisions concerning facility location also attempt to minimize total transportation and production costs for the entire system.

4.3.1 Setting up a Transportation problem

To illustrate how to set up a transportation problem we consider the following example;

Example 4.1
A concrete company transports concrete from three plants, 1, 2 and 3, to three construction sites, A, B and C.

The plants are able to supply the following numbers of tons per week:

Plant	*Supply (capacity)*
1	*300*
2	*300*
3	*100*

The requirements of the sites, in number of tons per week, are:

Construction site	*Demand (requirement)*
A	*200*
B	*200*
C	*300*

The cost of transporting 1 ton of concrete from each plant to each site is shown in the figure 8 in Emalangeni per ton.

For computational purposes it is convenient to put all the above information into a table, as in the simplex method. In this table each row represents a source *and each column represents a* destination.

		Sites			
	To / *From*	*A*	*B*	*C*	*Supply (Availability)*
Plants	***1***	*4*	*3*	*8*	*300*
	2	*7*	*5*	*9*	*300*
	3	*4*	*5*	*5*	*100*
	Demand (requirement)	*200*	*200*	*300*	

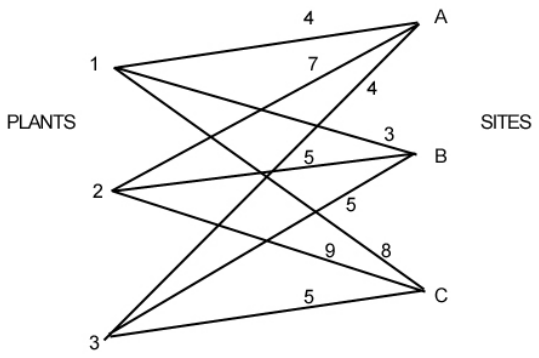


Figure 8: Constructing a transportation problem

4.3.2 Mathematical model of a transportation problem

Before we discuss the solution of transportation problems we will introduce the notation used to describe the transportation problem and show that it can be formulated as a linear programming problem.

We use the following notation;

$x_{ij} =$	the number of units to be distributed from source i to destination j ($i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$);
$s_i =$	supply from source i;
$d_j =$	demand at destination j;
$c_{ij} =$	cost per unit distributed from source i to destination j

With respect to Example 4.1 the decision variables x_{ij} are the numbers of tons transported from plant i (where $i = 1, 2, 3$) to each site j (where $j =$ A, B, C)

A basic assumption is that the distribution costs of units from source i to destination j is directly proportional to the number of units distributed. A typical **cost and requirements table** has the form shown on Table 4.

Let Z be total distribution costs from all the m sources to the n destinations. In example 4.1 each term in the objective function Z represents the total cost of tonnage transported on one route. For example, in the route $2 \longrightarrow C$, the term in $9x_{2C}$, that is:

$$(\text{Cost per ton} = 9) \quad \times \quad (\text{number of tons transported} = x_{2C})$$

		Destination				
		1	2	...	n	*Supply*
	1	c_{11}	c_{12}	...	c_{1n}	s_1
	2	c_{21}	c_{22}	...	c_{2n}	s_2
Source	⋮	⋮	⋮	...	⋮	⋮
	m	c_{m1}	c_{m2}	...	c_{mn}	s_m
Demand		d_1	d_2	...	d_n	

Table 4: *Cost and requirements table*

Hence the objective function is:

$$\begin{aligned} Z &= 4x_{1A} + 3x_{1B} + 8x_{1C} \\ &+ 7x_{2A} + 5x_{2B} + 9x_{2C} \\ &+ 4x_{3A} + 5x_{3B} + 5x_{3C} \end{aligned}$$

Notice that in this problem the total supply is 300 + 300 + 200 = 700 and the total demand is 200 + 200 + 300 = 700. Thus

$$\text{Total supply} = \text{total demand.}$$

In mathematical form this expressed as

$$\sum_{i=1}^{m} s_i = \sum_{j=1}^{n} d_j \tag{47}$$

This is called a **balanced problem** . In this unit our discussion shall be restricted to the balanced problems.

In a balanced problem all the products that can be supplied are used to meet the demand. There are no slacks and so all constraints are *equalities* rather than *inequalities* as was the case in the previous unit.

The formulation of this problem as a linear programming problem is presented as

$$\text{Minimise} \quad Z = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}\, x_{ij}, \tag{48}$$

subject to

$$\sum_{j=1}^{n} x_{ij} = s_i, \qquad \text{for} \quad i = 1, 2, \ldots, m \tag{49}$$

$$\sum_{i=1}^{n} x_{ij} = d_j, \qquad \text{for} \quad j = 1, 2, \ldots, n \tag{50}$$

and

$$x_{ij} \geq 0, \text{ for all } i \text{ and } j.$$

Any linear programming problem that fits this special formulation is of the transportation type, regardless of its physical context. For many applications, the supply and demand quantities in the model will have integer values and implementation will require that the distribution quantities also be integers. Fortunately, the unit coefficients of the unknown variables in the constraints guarantee an optimal solution with only integer values.

4.3.3 Initial solution - Northwest Corner Rule

The initial basic feasible solution can be obtained by using one of several methods. We will consider only the **North West corner rule** of developing an initial solution. Other methods can be found in standard texts on linear programming.

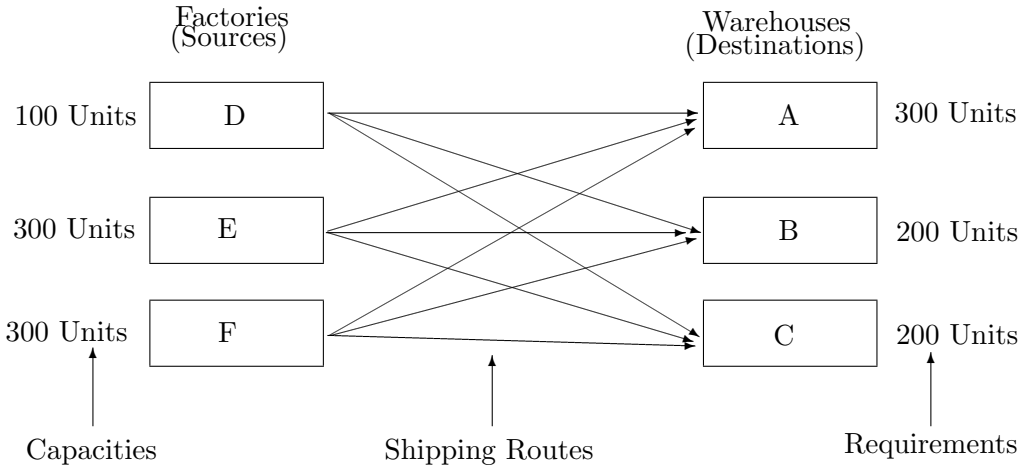
The procedure for constructing an initial basic feasible solution selects the basic variables one at a time. The North West corner rule begins with an allocation at the top left-hand corner of the tableau and proceeds systematically along either a row or a column and make allocations to subsequent cells until the bottom right-hand corner is reached, by which time enough allocations will have been made to constitute an initial solution.

The procedure for constructing an initial solution using the North West Corner rule is as follows:

NORTH WEST CORNER RULE

1. Start by selecting the cell in the most "North-West" corner of the table.
2. Assign the maximum amount to this cell that is allowable based on the requirements and the capacity constraints.
3. Exhaust the capacity from each row before moving down to another row.
4. Exhaust the requirement from each column before moving right to another column.
5. Check to make sure that the capacity and requirements are met.

Let us begin with an example dealing with Executive Furniture corporation, which manufactures office desks at three locations: D, E and F. The firm distributes the desks through regional warehouses located in A, B and C (see the Network format diagram below)



It is assumed that the production costs per desk are identical at each factory. The only relevant costs are those of shipping from each source to each destination. The costs are shown in Table 5

From \ To	A	B	C
D	\$5	\$4	\$3
E	\$8	\$4	\$3
F	\$9	\$7	\$5

Table 5: Transportation Costs per desk for Executive Furniture Corp.

We proceed to construct a transportation table and label its various components as show in Table 6.

We can now use the Northwest corner rule to find an initial feasible solution to the problem. We start in the upper left hand cell and allocate units to shipping routes as follows:

From \ To	A	B	C	Capacity
D	5	4	3	**100**
E	8	4	3	**300**
F	9	7	5	**300**
Requirements	**300**	**200**	**200**	**700**

Table 6: Transportation Table for Executive Furniture Corporation

1. Exhaust the supply (factory capacity) of each row before moving down to the next row.
2. Exhaust the demand (warehouse) requirements of each column before moving to the next column to the right.
3. Check that all supply and demand requirements are met.

The initial shipping assignments are given in Table 7

From \ To	A	B	C	Factory Capacity
D	100			100
E	200	100		300
F		100	200	300
Warehouse Requirements	300	200	200	700

Table 7: Initial Solution of the North West corner Rule

This initial solution can also be presented together with the costs per unit as shown in the Table 8.

We can compute the cost of this shipping assignment as follows;

Therefore, the initial feasible solution for this problem is $4200.

Example 4.2
Consider a transportation problem in which the cost, supply and demand values are presented in Table 10.

(a) Is this a balanced problem? Why?

From \ To	A	B	C	Capacity
D	5	4	3	
	100			**100**
E	8	4	3	
	200	100		**300**
F	9	7	5	
		100	200	**300**
Requirements	**300**	**200**	**200**	**700**

Table 8: Representing the initial feasible solution with costs

ROUTE		UNITS		PER UNIT		TOTAL
FROM	TO	SHIPPED	×	COST ($)	=	COST ($)
D	A	100		5		500
E	A	200		8		1600
E	B	100		4		400
F	B	100		7		700
F	C	200		5		1000
						Total 4200

Table 9: Calculation of costs of initial shipping assignments

(b) Obtain the initial feasible solution using the North-West Corner rule.

Solution:

(a) We calculate the total supply and total demand.
Total supply = 14 + 10 + 15 + 13 = 52
Total demand = 10 + 15 + 12 + 15 = 52
Since the total supply is equal to the total demand we conclude that the problem is balanced.

(b) The allocations according to the North-West corner rule are shown in Table 11 The initial feasible solution is

$$\textit{Total Cost} = 10 \cdot 10 + 4 \cdot 30 + 10 \cdot 15 + 1 \cdot 30 + 12 \cdot 20 + 2 \cdot 20 + 13 \cdot 45 = \$1265$$

Note that this is not necessarily equal to the optimal solution.

		Destination				
		1	2	3	4	Supply
Source	1	10	30	25	15	14
	2	20	15	20	10	10
	3	10	30	20	20	15
	4	30	40	35	45	13
	Demand	10	15	12	15	

Table 10: Supply and Demand values for Transportation problem

	1	2	3	4	Supply
1	10	4			14
2		10			10
3		1	12	2	15
4				13	13
Demand	10	15	12	15	

Table 11: Initial feasible solution

4.4 Exercises 4.1: Northwest Corner rule

In each of the following problems check whether the solution is balanced or not then use the North West Corner rule to find the basic feasible solution.

1.

TO / FROM	1	2	3	Supply
1	3	2	0	45
2	1	5	0	60
3	5	4	0	35
Demand	50	60	30	

2.

TO / FROM	1	2	3	Supply
1	5	4	3	100
2	8	4	3	300
3	9	7	5	300
Demand	300	200	200	

3.

FROM \ TO	1	2	3	4	Supply
A	12	13	4	6	500
B	6	4	10	11	700
C	10	9	12	4	800
Demand	400	900	200	500	

4.

FROM \ TO	1	2	3	4	Supply
1	10	30	25	15	14
2	20	15	20	10	10
3	10	30	20	20	15
4	30	40	35	45	13
Demand	10	15	12	15	

4.4.1 Optimality test - the Stepping Stone method

The next step is to determine whether the current allocation at any stage of the solution process is optimal. We will present one of the methods used to determine optimality of and improve a current solution. The method derives its name from the analogy of crossing a pond using stepping stones. The occupied cells are analogous to the stepping stones, which are used in making certain movements in this method.

The five steps of the **Stepping-Stone Method** are as follows:

STEPPING-STONE METHOD

1. Select an unused square to be evaluated.
2. Beginning at this square, trace a closed path back to the original square via squares that are currently being used (only horizontal or vertical moves allowed). You can only change directions at occupied cells!.
3. Beginning with a plus (+) sign at the unused square, place alternative minus (-) signs and plus signs on each corner square of the closed path just traced.
4. Calculate an **improvement index**, I_{ij} by adding together the unit cost figures found in each square containing a plus sign and then subtracting the unit costs in each square containing a minus sign.
5. Repeat steps 1 to 4 until an improvement index has been calculated for all unused squares.
   - If all indices computed are greater than or equal to zero, an optimal solution has been reached.
   - If not, it is possible to improve the current solution and decrease total shipping costs.

4.4.2 The optimality criterion

If all the cost index values obtained for all the currently unoccupied cells are nonnegative, then the current solution is optimal. If there are negative values the solution has to be improved. This means that an allocation is made to one of the empty cells (unused routes) and the necessary adjustments in the supply and demand effected accordingly.

To see how the Stepping-Stone method works we apply these steps to the Furniture Corporation example to evaluate the shipping routes.

Steps 1-3 Beginning with the D-B route, we first trace a closed path using only currently occupied squared (see Table 12) and then place alternate plus signs and minus signs in the corners of this path.

Step 4 An improvement index I_{ij} for the D-B route in now computed by adding unit costs in squares with plus signs and subtracting costs in squares with minus signs. Thus

$$I_{DB} = +4 - 5 + 8 - 4 = +3$$

This means that for every desk shipped via the D-B route, total transportation costs will increase by \$3 over their current level.

From \ To	A	B	C	Capacity
D	5	Start 4	3	
	100 - ←	+		**100**
E	↓ 8	↑ 4	3	
	200 + →	- 100		**300**
F	9	7	5	
		100	200	**300**
Requirements	**300**	**200**	**200**	**700**

Table 12: Evaluating the D-B route

Step 5 Next we consider the D-C unused route. The closed path we use is (see Table 13)

$$+DC - DA + EA - EB + FB - FC$$

The D-C improvement index is

$$I_{DC} = +3 - 5 + 8 - 4 + 7 - 5 = +4$$

From \ To	A	B	C	Capacity
D	5	4	Start 3	
	100 - ←	⟵ ⟵ ←	+	**100**
E	↓ 8	4	↑ 3	
	200 + →	- 100	↑	**300**
F	9	↓ 7	↑ 5	
		+ ⟶ →	-	
		100	200	**300**
Requirements	**300**	**200**	**200**	**700**

Table 13: Evaluating the D-C route

The other two routes may be evaluated in a similar fashion

E-C route: closed path = +EC - EB + FB - FC

$$I_{EC} = +3 - 4 + 7 - 5 = +1$$

FA route: closed path = +FA - FB + EB - EA

$$I_{FA} = +9 - 7 + 4 - 8 = -2$$

Because the I_{FA} index is negative, a cost saving may be attained by making use of the FA route i.e the FA cell can be improved. The Stepping-Stone path used to evaluate the route FA is shown in Table 14

From \ To	A	B	C	Capacity
D	5 100	4	3	**100**
E	8 200 - ←	4 + 100	3	**300**
F	↓ 9 Start + →	↑ 7 - 100	5 200	**300**
Requirements	**300**	**200**	**200**	**700**

Table 14: Stepping-Stone Path used to evaluate FA route

The next step, then is to ship the maximum allowable number of units on the new route (FA route). What is the maximum quantity that can be shipped on the money-saving route? The quantity is found by referring to the closed path of plus signs and minus signs drawn for the route and selecting the *smallest number* found in those squares containing minus signs. To obtain a new solution, that number is added to all squares on the closed path with plus signs and subtracted from all squares on the path assigned minus signs. All other squares are left unchanged. The new solution is shown in Table 15.

From \ To	A	B	C	Capacity
D	5 100	4	3	**100**
E	8 100	4 200	3	**300**
F	9 100	7	5 200	**300**
Requirements	**300**	**200**	**200**	**700**

Table 15: Improved solution: Second solution

The shipping cost for this new solution is

$$100 \cdot 5 + 100 \cdot 8 + 200 \cdot 4 + 100 \cdot 9 + 200 \cdot 4 = \$4000$$

This solution may or may not be optimal. To determine whether further improvement is possible, we return to the first five steps to test each square that is now unused. The four improvement indices - each representing an available shipping route are as follows:

$$\begin{aligned} \text{D to B} \quad &= \quad I_{DB} = 4 - 5 + 8 - 4 = +\$3 \\ &\qquad (\text{Closed path} : +DB - DA + EA - EB) \end{aligned}$$

$$\begin{aligned}
\text{D to C} &= I_{DC} = 3 - 5 + 9 - 5 = +\$2 \\
&\quad (\text{Closed path} : +DC - DA + FA - FC) \\
\text{E to C} &= I_{EC} = 3 - 8 + 9 - 5 = -\$1 \\
&\quad (\text{Closed path} : +EC - EA + FA - FC) \\
\text{F to B} &= I_{FB} = 7 - 4 + 8 - 9 = +\$2 \\
&\quad (\text{Closed path} : +FB - EB + EA - FA)
\end{aligned}$$

Hence, an improvement can be made by shipping the maximum allowable number of units from E to C (see Table 16).

From \ To	A	B	C	Capacity
D	5 → 100	4	3	**100**
E	8 (-) ↓ 100	4 ← ← 200	3 ← (+) Start	**300**
F	9 ↓ (+) 100 →	7 → → →	5 ↑ (-) 200 →	**300**
Requirements	**300**	**200**	**200**	**700**

Table 16: Path to evaluate the E-C route

The improved solution is shown in Table 17. The total cost for the third solution is

$$100 \cdot 5 + 200 \cdot 4 + 100 \cdot 3 + 200 \cdot 9 + 100 \cdot 5 = \$3900$$

To determine if the current solution is optimal we calculate the improvement indices - each

From \ To	A	B	C	Capacity
D	5 → 100	4	3	**100**
E	8	4 → 200	3 → 100	**300**
F	9 → 200	7	5 → 100	**300**
Requirements	**300**	**200**	**200**	**700**

Table 17: Improved solution: Third solution

representing an available shipping route - as follows:

$$\text{D to B} = I_{DB} = 4 - 5 + 9 - 5 + 3 - 4 = +\$2$$

$$
\begin{aligned}
& & & \text{(Closed path: } +DB - DA + FA - FC + EC - EB\text{)} \\
\text{D to C} & = & & I_{DC} = 3 - 5 + 9 - 5 = +\$2 \\
& & & \text{(Closed path: } +DC - DA + FA - FC\text{)} \\
\text{E to A} & = & & I_{EA} = 8 - 9 + 5 - 3 = +\$1 \\
& & & \text{(Closed path: } +EA - FA + FC - EC\text{)} \\
\text{F to B} & = & & I_{FB} = 7 - 5 + 3 - 4 = +\$1 \\
& & & \text{(Closed path: } +FB - FC + EC - EB\text{)}
\end{aligned}
$$

Table 17 contains the optimal solution because each improvement index for the Table is greater than or equal to zero.

4.5 Summary

In this section we discussed the formulation of transportation problems and their methods of solution. We used the North West corner rule to obtain the initial feasible solution and the Stepping-Stone method to find the optimal solution. We restricted focus to balanced transportation problems where it is assumed that the total supply is equal to total demand.

4.6 Exercises 4.2: Transportation problems

1. A company has factories at A, B and C which supply warehouses at D, E and F. Weekly factory capacities are 200, 160 and 90 units respectively. Weekly warehouse requirements (demands) are 180, 120 and 150 units respectively. Unit shipping costs (in Emalangeni) are as follows:

Factory	D	E	F	Capacity
A	16	20	12	**200**
B	14	8	18	**160**
C	26	24	16	**90**
Demand	**180**	**120**	**150**	**450**

   Determine the optimum distribution for this company to minimize shipping costs. [E5920]

2. A Timber company ships pine flooring to three building supply houses from its mills in Bhunya, Mondi and Pigg's Peak. Determine the best transportation schedule for the data given below using the Northwest corner rule and the Stepping Stone method. [E230]

TO / FROM	*Supply House* 1	*Supply House* 2	*Supply House* 3	*Mill Capacity* (*tons*)
Bhunya	3	3	2	**25**
Mondi	4	2	3	**40**
Pigg's Peak	3	2	3	**30**
Supply House Demand (*tons*)	**30**	**30**	**35**	**95**

3. A trucking company has a contract to move 115 truckloads of sand per week between three sand-washing plants W,X and Y, and three destinations, A,B and C. Cost and volume information is given below. Compute the optimal transportation cost.

To / From	Project A	Project B	Project C	Supply
Plant W	5	10	10	35
Plant X	20	30	20	40
Plant Y	5	8	12	40
Demand	45	50	20	

[C=1345]

4. In each of the following cases write down the North West corner solution and use the Stepping Stone method to find the minimal cost.

(a)

TO / FROM	D	E	F	Capacity
A	8	6	9	**20**
B	6	3	8	**30**
C	10	7	9	**70**
Demand	**90**	**20**	**10**	**120**

[E970]

(b)

TO / FROM	D	E	F	Capacity
A	2	2	3	**4**
B	2	1	6	**6**
C	1	3	4	**8**
Demand	**2**	**5**	**11**	**18**

[E48]

4.7 Assignment Problem

The **assignment problem** refers to the class of linear programming problems that involve determining the most efficient assignment of

- people to projects
- salespeople to territories
- contracts to bidders
- jobs to machines, etc.

The objective is most often to minimize total costs or total time of performing the tasks at hand.

One important characteristic of assignment problems is that only one job or worker is assigned to one machine or project. An example is the problem of a taxi company with 4 taxis and 4 passengers. Which taxi should collect which passenger in order to minimize costs?

Each assignment problem has associated with it a table, or matrix. Generally, the rows contain the objects or people we wish to assign, and the columns comprise the tasks or things we want them assigned to. The numbers in the table are the costs associated with each particular assignment.

An assignment problem can be viewed as a transportation problem in which

- the capacity from each source (or person to be assigned) is 1 and
- the demand at each destination (or job to be done) is 1.

As an illustration of the assignment problem, let us consider the case of a Fix-It-Shop, which has just received three new rush projects to repair: (1) a radio, (2) a toaster oven, and (3) a broken coffee table. Three repair persons, each with different talents and abilities, are available to do the jobs. The owner of the shop estimates what it will cost in wages to assign each of the workers to each of the three projects. The costs which are shown in Table 18 differ because the owner believes that each worker will differ in speed and skill on these quite varied jobs.

Table 19 summarizes all six assignment options. The table also shows that the least-cost solution would be to assign Cooper to project 1, Brown to project 2, and Adams to project 3, at a total cost of $25.

The owner's objective is to assign the three projects to the workers in a way that will result in the lowest cost to the shop. Note that the assignment of people to projects must be on a one-to-one basis; each project will be assigned exclusively to one worker only.

	PROJECT		
PERSON	1	2	3
Adams	$11	$14	$6
Brown	8	10	11
Cooper	9	12	7

Table 18: Repair costs of the Fix-It-Shop assignment problem

PROJECT ASSIGNMENT				
1	2	3	LABOUR COSTS ($)	TOTAL COSTS ($)
Adams	Brown	Cooper	11 + 10 +7	28
Adams	Cooper	Brown	11 + 12 +11	34
Brown	Adams	Cooper	8 + 14 + 7	29
Brown	Cooper	Adams	8 + 12 + 6	26
Cooper	Adams	Brown	9 + 14 + 11	34
Cooper	Brown	Adams	9 + 10 + 6	25

Table 19: Assignment alternatives and Costs of Fix-It-Shop assignment problem

Special algorithms exist to solve assignment problems. The most common is probably the **Hungarian** solution method. The Hungarian method of assignment provides us with an efficient means of finding the optimal solution without having to make a direct comparison of every assignment option. It operates on a principle of matrix reduction, which means that by subtracting and adding appropriate numbers in the cost table or matrix, we can reduce the problem to a matrix of *opportunity costs*. Opportunity costs show the relative penalties associated with assigning any person to a project as opposed to making the best or least-cost assignment. We would like to make assignments such that the opportunity cost for each assignment is zero.

The steps involved in the Hungarian method are outlined below.

THE HUNGARIAN METHOD

1. *Find the opportunity cost table by*
   (a) Subtracting the smallest number in each row of the original cost table or matrix from every number in that row.
   (b) Then subtracting the smallest number in each column of the table obtained in part (a) from every number in that column.
2. *Test the table resulting from step 1 to see whether an optimal assignment can be made.* The procedure is to draw the minimum number of vertical and horizontal straight lines necessary to cover all zeros in the table. If the number of lines equals either the number of rows or columns, an optimal assignment can be made. If the number of lines is less than the number of rows or columns, we proceed to step 3.
3. *Revise the present opportunity cost table.* This is done by subtracting the smallest number not covered by a line from every other uncovered number. This same smallest number is also added to any number(s) lying at the intersection of the horizontal and vertical lines. We then return to step 2 and continue the cycle until an optimal assignment is possible.

Let us now apply the three steps to the Fix-It-Shop assignment example.

The original cost table for the problem is given in Table 20

	PROJECT		
PERSON	1	2	3
Adams	11	14	6
Brown	8	10	11
Cooper	9	12	7

Table 20: Initial Table

	PROJECT		
PERSON	1	2	3
Adams	5	8	0
Brown	0	2	3
Cooper	2	5	0

Table 21: Row reduction (part a)

After the row reduction (Step 1 part a) we get the cost Table 21.

Taking the costs in Table 21 and subtracting the the smallest number in each column from each number in that column results in the total opportunity costs given in Table 22. This step is the column reduction of Step 1 part (b)

If we draw vertical and horizontal straight lines (Step 2) to cover all the zeros in Table 22 we get Table 23. Since the number of lines is less than the number of rows or columns an optimal assignment cannot be made.

Since Table 23 doesn't give an optimal solution we revise the table. This is accomplished by subtracting the smallest number not covered by a line from all numbers not covered by

	PROJECT		
PERSON	1	2	3
Adams	5	6	0
Brown	0	0	3
Cooper	2	3	0

Table 22: Column Reduction (Step 1 part b)

	PROJECT		
PERSON	1	2	3
Adams	5	6	0
~~Brown~~	~~0~~	~~0~~	~~3~~
Cooper	2	3	0

Table 23: Testing for an optimal solution

a straight line. This same smallest number is then added to every number (including zeros) *lying in the intersection* on any two lines. The smallest uncovered number in Table 23 is 2, so this value is subtracted from each of the four uncovered numbers. A 2 is also added to the number that is covered by the intersecting horizontal and vertical lines. The results of this step are shown in Table 24

To test now for an optimal assignment, we return to Step 2 and find the minimum number of lines necessary to cover all zeros in the revised opportunity cost table. Because it requires three lines to cover the zeros (see Table 25), an optimal assignment can be made.

	PROJECT		
PERSON	1	2	3
Adams	3	4	0
Brown	0	0	5
Cooper	0	1	0

Table 24: Revised opportunity cost table

	PROJECT		
PERSON	1	2	3
Adams	3	4	0
~~Brown~~	~~0~~	~~0~~	~~5~~
Cooper	0	1	0

Table 25: Optimality test on the revised table

Finally, we make the allocation. Note that only one assignment will be made from each row or column. We use this fact to proceed to making the final allocation as follows:

(a) Find a row or column with only one zero cell.

(b) Make the assignment corresponding to that zero cell.

(c) Eliminate that row and column from the table.

(d) Continue until all the assignments have been made.

For our Fix-It-Shop problem these steps are summarized in Table 26.

	FIRST ASSIGNMENT				SECOND ASSIGNMENT				THIRD ASSIGNMENT		
	1	2	3		1	2	3		1	2	3
Adams	3	4	[0]	Adams	~~3~~	~~4~~	~~[0]~~	Adams	~~3~~	~~4~~	~~[0]~~
Brown	0	0	5	Brown	0	0	~~5~~	Brown	0	[0]	5
Cooper	0	1	0	Cooper	[0]	1	~~0~~	Cooper	~~[0]~~	~~1~~	~~0~~

Table 26: Making the final assignment

To interpret the table we recall that our objective was to minimize costs, there is only one assignment that Adams can go to where the opportunity costs are \$0. That is to assign Adams Project 3. If Adams gets assigned to Project 3, then there is only one project left where the opportunity cost is \$0 for Cooper. Therefore Cooper gets assigned to Project 1. This leaves Brown being assigned to Project 2, where the opportunity costs are \$0.

The optimal allocation is to assign Adams to Project 3, Brown to Project 2, and Cooper to Project 1. The total labour cost of this assignment are computed from the original cost table (see Table 18). They are as follows:

ASSIGNMENT	COST (\$)
Adams to Project 3	6
Brown to Project 2	10
Cooper to Project 1	9
Total cost	25

Example 4.3 *Suppose we have to allocate 4 tasks (1,2,3,4) between 4 people (W,X,Y,Z). The costs are set out in the following table:*

	Task			
Person	*1*	*2*	*3*	*4*
W	*8*	*20*	*15*	*17*
X	*15*	*16*	*12*	*10*
Y	*22*	*19*	*16*	*30*
Z	*25*	*15*	*12*	*9*

The entries in the table denote the costs of assigning a task to a particular person.

Solution: Step 1 of the Hungarian method involves the following parts:

(a) subtract the minimum value from each column (see Table 27)

(b) subtract the minimum value from each column (see Table 28)

	Task			
Person	1	2	3	4
W	0	12	7	9
X	5	6	2	0
Y	6	3	0	14
Z	16	6	3	0

Table 27: Subtract the minimum value from each row

	Task			
Person	1	2	3	4
W	0	9	7	9
X	5	3	2	0
Y	6	0	0	14
Z	16	3	3	0

Table 28: subtract the minimum value from each column

The next step is to check whether optimal assignment can be made. This is done by finding the minimum *number of lines necessary to cross-out all the zero cells in the table. If this is equal to n (the number of people/tasks) then the solution has been found. The minimum number of lines necessary to cross through all the zeros (see Table 29)is 3 ¡ n = 4 so that an optimal allocation has not been found.*

*(*Note that there may be more than one way to draw the lines through the zero cells. It does not matter which way you choose as long as there is no alternative way involving fewer lines*)*

	Task			
Person	1	2	3	4
W	0	9	7	9
X	5	3	2	0
Y	6	0	0	14
Z	16	3	3	0

Table 29: Checking if an optimal assignment can been made

Next we revise the table by

(a) Finding the minimum uncovered cell. Table 29 shows that the minimum uncovered cell has a value of 2

(b) Subtracting the value obtained in (a) (i.e subtract 2) from all the uncovered cells.

(c) Adding to all the cells at the intersection of the two lines.

The result of the above steps is given in Table 30.

We then check if the revised allocation is optimal. This is done by finding the minimum number of lines required to cover all zeros (see Table 31).

This time the minimum number of lines necessary to cross through all the zeros is n = 4 so that an optimal allocation has been found.

To make the final allocation we use the following steps.

	Task			
Person	1	2	3	4
W	0	7	5	9
X	5	1	0	0
Y	8	0	0	16
Z	16	1	1	0

Table 30: Revising the Table

	Task			
Person	1	2	3	4
W	0	7	5	9
X	5	1	0	0
Y	~~8~~	~~0~~	~~0~~	~~16~~
Z	16	1	1	0

Table 31: Checking for optimality

- *Find a row or column with only one zero cell.*
- *Make the assignment corresponding to that zero cell.*
- *eliminate that row and column from the table.*
- *Continue until all assignments have been found.*

	Task			
Person	*1*	*2*	*3*	*4*
W	*0*	*7*	*5*	*9*
X	*5*	*1*	*0*	*0*
Y	*8*	*0*	*0*	*16*
Z	*16*	*1*	*1*	*0*

- *Assign person W to task 1 and eliminate row W and column 1.*
- *Assign person Y to task 2 and eliminate row Y and column 2.*
- *Assign person Z to task 4 and eliminate row Z and column 4.*
- *This leaves final person X assigned to remaining task 3.*

From the original cost table, we can determine the costs associated with the optimal assignment:

$$Total \;\; Cost = 48$$

4.8 Maximization Assignment Problems

Some assignment problems are phrased in terms of **maximizing** the payoff, profit, or effectiveness of an assignment instead of *minimization* costs. It is easy to obtain an equivalent minimization problem by converting all numbers in the table to opportunity costs; efficiencies to inefficiencies,etc. This is achieved through *subtracting every number in the original payoff table from the largest single number in the number.* The transformed entries represent opportunity costs; it turns out that minimizing the opportunity costs produces the same

assignment as the original maximization problem. Once the optimal assignment for this transformed problem has been computed, the total payoff or profit is found by adding the original payoffs of those cells that are in the original assignment.

Example. The British Navy wishes to assign four ships to patrol four sectors of the North Sea. In some areas ships are to be on the outlook for illegal fishing boats, and in other sectors to watch for enemy submarines, so the commander rates each ship in terms of its profitable efficiency in each sector. These relative efficiencies are illustrated in Tables 32. On the basis of the ratings shown, the commander wants to determine the patrol assignments producing the greatest overall efficiencies.

	SECTOR			
SHIP	A	B	C	D
1	20	60	50	55
2	60	30	80	75
3	80	100	90	80
4	65	80	75	70

Table 32: Efficiencies of British Ships in Patrol sectors

	SECTOR			
SHIP	A	B	C	D
1	80	40	50	45
2	40	70	20	25
3	20	0	10	20
4	35	20	25	30

Table 33: Opportunity Costs of British Ships

We start by converting the maximizing efficiency table into a minimization opportunity cost table. This is done by subtracting each rating from 100, the largest rating in the whole table. The resulting opportunity costs are given in Table 33.

Next, we follow steps 1 and 2 of the assignment algorithm. The smallest number is subtracted from every number in that row to give Table 34; and then the smallest number in each column is subtracted from every number in that column as shown in Table 35.

	SECTOR			
SHIP	A	B	C	D
1	40	0	10	5
2	20	50	0	5
3	20	0	10	20
4	15	0	5	10

Table 34: Row opportunity costs for the British Navy Problem

	SECTOR			
SHIP	A	B	C	D
1	25	0	10	0
2	5	50	0	0
3	5	0	10	15
4	0	0	5	5

Table 35: Total opportunity costs for the British Navy Problem

The minimum number of straight lines needed to cover all zeros in this total opportunity cost table is four. Hence an optimal assignment can be made. The optimal assignment is ship 1 to sector D, ship 2 to sector C, ship 3 to sector B, and ship 4 to sector A.

The overall efficiency, computed from the original efficiency data Table 32, can now be shown:

ASSIGNMENT	EFFICIENCY
Ship 1 to Sector D	55
Ship 2 to Sector C	80
Ship 3 to Sector B	100
Ship 4 to Sector A	65
Total Efficiency	300

4.9 Summary

In this section we discussed the Hungarian method for solving both maximization and minimization assignment problems.

4.10 Exercises 4.3: Minimization Assignment Problems

1. Three accountants, Phindile, Rachel and Sibongile, are to be assigned to three projects, 1, 2 and 3. The assignment costs in units of E1000 are given in the table below.

		Project		
		1	**2**	**3**
Accountant	**P**	15	9	12
	R	7	5	10
	S	13	4	6

2. Joy Taxi has four taxis, 1,2,3 and 4, and there are four customers, P, Q, R and S requiring taxis. The distance between the taxis and the customers are given in the table below, in Kilometres. The Taxi company wishes to assign the taxis to customers so that the distance traveled is a minimum.

		Customers			
		P	Q	R	S
Taxis	**1**	10	8	4	6
	2	6	4	12	8
	3	14	10	8	2
	4	4	14	10	8

3. Four precision components are to be shaped using four machine tools, one tool being assigned to each component. The machining times, in minutes, are given in the table below.

		Component			
		1	**2**	**3**	**4**
Machine Tool	**A**	21	20	39	36
	B	25	22	24	25
	C	36	22	36	26
	D	34	21	25	39

4. In a job shop operation, four jobs may be performed on any of four machines. The hours required for each job on each machine are presented in the following table. The plant supervisor would like to assign jobs so that total time in minimized. Use the assignment method to find the best solution.

	MACHINE			
JOB	W	X	Y	Z
A12	10	14	16	13
A15	12	13	15	12
B2	9	12	12	11
B9	14	16	18	16

Answer: A12 to W, A15 to Z, B2 to Y, B9 to Z, 50 hours.

4.11 Exercises 4.4: Maximization Assignment Problems

1. A head of department has four lecturers to assign to pure maths (1), mechanics (2), statistics (3) and Quantitative techniques (4). All of the teachers have taught the courses in the past and have been evaluated with a score from 0 to 100. The scores are shown in the table below.

	1	**2**	**3**	**4**
Peters	80	55	45	45
Radebe	58	35	70	50
Tsabedze	70	50	80	65
Williams	90	70	40	80

The head of department wishes to know the optimal assignment of teachers to courses that will maximize the overall total score. Use the Hungarian algorithm to solve this problem. [$P \to 1$, $R \to 3$, $T \to 4$, $W \to 2$ Max Score = 285]

2. A department store has leased a new store and wishes to decide how to place four departments in four locations so as to maximize total profits. The table below gives the profits, in thousands of emalangeni, when the departments are allocated to the various locations. Find the assignment that maximizes total profits.

		Location			
		1	**2**	**3**	**4**
Department	Shoes	20	16	22	18
	Toys	25	28	15	21
	Auto	27	20	23	26
	Housewares	24	22	23	22

3. The head of the business department, has decided to apply the Hungarian method in assigning lecturers to courses next semester. As a criterion for judging who should teach each course, the head of department reviews the past two years' teaching evaluations. All the four lecturers have taught each of the courses at one time or another during the two year period. The ratings are shown in the table below.
   Find the best assignment of lecturers to courses to maximize the overall teaching rating. Total Rating = 335

	COURSE			
LECTURER	STATISTICS	MANAGEMENT	FINANCE	ECONOMICS
Dlamini	90	65	95	40
Khumalo	70	60	80	75
Masuku	85	40	80	60
Nxumalo	55	80	65	55

Unit 8

Introduction to CPM / PERT Techniques
Applications of CPM / PERT
Basic Steps in PERT / CPM
Frame work of PERT/CPM
Network Diagram Representation
Rules for Drawing Network Diagrams
Common Errors in Drawing Networks
Advantages and Disadvantages
Critical Path in Network Analysis

Introduction to CPM / PERT Techniques

CPM/PERT or Network Analysis as the technique is sometimes called, developed along two parallel streams, one industrial and the other military.

CPM (Critical Path Method) was the discovery of M.R.Walker of E.I.Du Pont de Nemours & Co. and J.E.Kelly of Remington Rand, circa 1957. The computation was designed for the UNIVAC-I computer. The first test was made in 1958, when CPM was applied to the construction of a new chemical plant. In March 1959, the method was applied to maintenance shut-down at the Du Pont works in Louisville, Kentucky. Unproductive time was reduced from 125 to 93 hours.

PERT (Project Evaluation and Review Technique) was devised in 1958 for the POLARIS missile program by the Program Evaluation Branch of the Special Projects office of the U.S.Navy, helped by the Lockheed Missile Systems division and the Consultant firm of Booz-Allen & Hamilton. The calculations were so arranged so that they could be carried out on the IBM Naval Ordinance Research Computer (NORC) at Dahlgren, Virginia.

The methods are essentially **network-oriented techniques** using the same principle. PERT and CPM are basically time-oriented methods in the sense that they both lead to determination of a time schedule for the project. The significant difference between two approaches is that the time estimates for the different activities in CPM were assumed to be **deterministic** while in PERT these are described **probabilistically**. These techniques are referred as **project scheduling** techniques.

In **CPM** activities are shown as a network of precedence relationships using activity-on-node network construction

- Single estimate of activity time
- Deterministic activity times

USED IN: Production management - for the jobs of repetitive in nature where the activity time estimates can be predicted with considerable certainty due to the existence of past experience.

In **PERT** activities are shown as a network of precedence relationships using activity-on-arrow network construction

- Multiple time estimates
- Probabilistic activity times

USED IN: Project management - for non-repetitive jobs (research and development work), where the time and cost estimates tend to be quite uncertain. This technique uses probabilistic time estimates.

Benefits of PERT/CPM

- Useful at many stages of project management
- Mathematically simple

- Give critical path and slack time
- Provide project documentation
- Useful in monitoring costs

Limitations of PERT/CPM

- Clearly defined, independent and stable activities
- Specified precedence relationships
- Over emphasis on critical paths

Applications of CPM / PERT

These methods have been applied to a wide variety of problems in industries and have found acceptance even in government organizations. These include

- Construction of a dam or a canal system in a region
- Construction of a building or highway
- Maintenance or overhaul of airplanes or oil refinery
- Space flight
- Cost control of a project using PERT / COST
- Designing a prototype of a machine
- Development of supersonic planes

Basic Steps in PERT / CPM

Project scheduling by PERT / CPM consists of four main steps

1. **Planning**

- The planning phase is started by splitting the total project in to small projects. These smaller projects in turn are divided into activities and are analyzed by the department or section.
- The relationship of each activity with respect to other activities are defined and established and the corresponding responsibilities and the authority are also stated.
- Thus the possibility of overlooking any task necessary for the completion of the project is reduced substantially.

2. Scheduling

- The ultimate objective of the scheduling phase is to prepare a time chart showing the start and finish times for each activity as well as its relationship to other activities of the project.
- Moreover the schedule must pinpoint the critical path activities which require special attention if the project is to be completed in time.
- For non-critical activities, the schedule must show the amount of slack or float times which can be used advantageously when such activities are delayed or when limited resources are to be utilized effectively.

3. Allocation of resources

- Allocation of resources is performed to achieve the desired objective. A resource is a physical variable such as labour, finance, equipment and space which will impose a limitation on time for the project.
- When resources are limited and conflicting, demands are made for the same type of resources a systematic method for allocation of resources become essential.
- Resource allocation usually incurs a compromise and the choice of this compromise depends on the judgment of managers.

4. Controlling

- The final phase in project management is controlling. Critical path methods facilitate the application of the principle of management by expectation to identify areas that are critical to the completion of the project.
- By having progress reports from time to time and updating the network continuously, a better financial as well as technical control over the project is exercised.
- Arrow diagrams and time charts are used for making periodic progress reports. If required, a new course of action is determined for the remaining portion of the project.

The Framework for PERT and CPM

Essentially, there are six steps which are common to both the techniques. The procedure is listed below:

I. Define the Project and all of its significant activities or tasks. The Project (made up of several tasks) should have only a single start activity and a single finish activity.

II. Develop the relationships among the activities. Decide which activities must precede and which must follow others.

III. Draw the "Network" connecting all the activities. Each Activity should have unique event numbers. Dummy arrows are used where required to avoid giving the same numbering to two activities.

IV. Assign time and/or cost estimates to each activity

V. Compute the longest time path through the network. This is called the critical path.

VI. Use the Network to help plan, schedule, and monitor and control the project.

The Key Concept used by CPM/PERT is that a small set of activities, which make up the longest path through the activity network control the entire project. If these "critical" activities could be identified and assigned to responsible persons, management resources could be optimally used by concentrating on the few activities which determine the fate of the entire project.

Non-critical activities can be replanned, rescheduled and resources for them can be reallocated flexibly, without affecting the whole project.

Five useful questions to ask when preparing an activity network are:

- Is this a Start Activity?
- Is this a Finish Activity?
- What Activity Precedes this?
- What Activity Follows this?
- What Activity is Concurrent with this?

Network Diagram Representation

In a network representation of a project certain definitions are used

1. Activity

Any individual operation which utilizes resources and has an end and a beginning is called activity. An arrow is commonly used to represent an activity with its head indicating the direction of progress in the project. These are classified into four categories

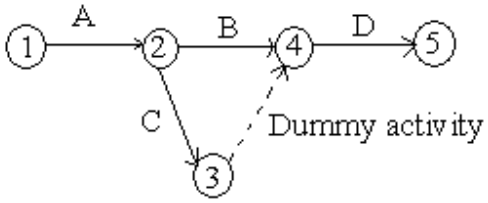
1. **Predecessor activity** – Activities that must be completed immediately prior to the start of another activity are called predecessor activities.

2. **Successor activity** – Activities that cannot be started until one or more of other activities are completed but immediately succeed them are called successor activities.
3. **Concurrent activity** – Activities which can be accomplished concurrently are known as concurrent activities. It may be noted that an activity can be a predecessor or a successor to an event or it may be concurrent with one or more of other activities.
4. **Dummy activity** – An activity which does not consume any kind of resource but merely depicts the technological dependence is called a dummy activity.

The dummy activity is inserted in the network to clarify the activity pattern in the following two situations

- To make activities with common starting and finishing points distinguishable
- To identify and maintain the proper precedence relationship between activities that is not connected by events.

For example, consider a situation where A and B are concurrent activities. C is dependent on A and D is dependent on A and B both. Such a situation can be handled by using a dummy activity as shown in the figure.



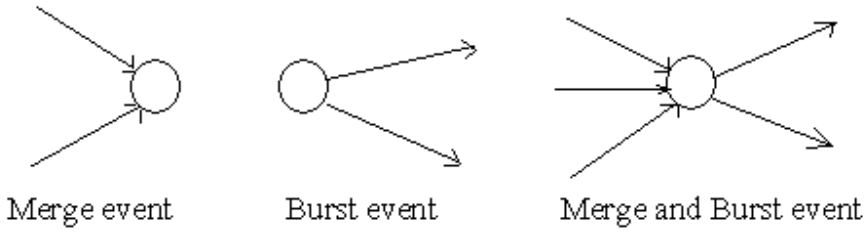
2. Event

An event represents a point in time signifying the completion of some activities and the beginning of new ones. This is usually represented by a circle in a network which is also called a node or connector.

The events are classified in to three categories

1. **Merge event** – When more than one activity comes and joins an event such an event is known as merge event.

2. **Burst event** – When more than one activity leaves an event such an event is known as burst event.
3. **Merge and Burst event** – An activity may be merge and burst event at the same time as with respect to some activities it can be a merge event and with respect to some other activities it may be a burst event.



3. Sequencing

The first prerequisite in the development of network is to maintain the precedence relationships. In order to make a network, the following points should be taken into considerations

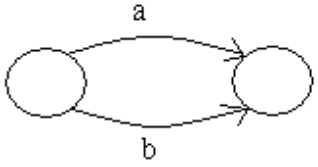
- What job or jobs precede it?
- What job or jobs could run concurrently?
- What job or jobs follow it?
- What controls the start and finish of a job?

Since all further calculations are based on the network, it is necessary that a network be drawn with full care.

Rules for Drawing Network Diagram

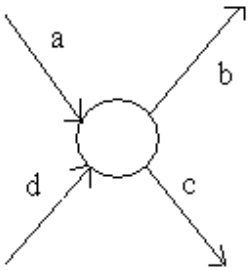
Rule 1

Each activity is represented by one and only one arrow in the network



Rule 2

No two activities can be identified by the same end events



Rule 3

In order to ensure the correct precedence relationship in the arrow diagram, following questions must be checked whenever any activity is added to the network

- What activity must be completed immediately before this activity can start?
- What activities must follow this activity?
- What activities must occur simultaneously with this activity?

In case of large network, it is essential that certain good habits be practiced to draw an easy to follow network

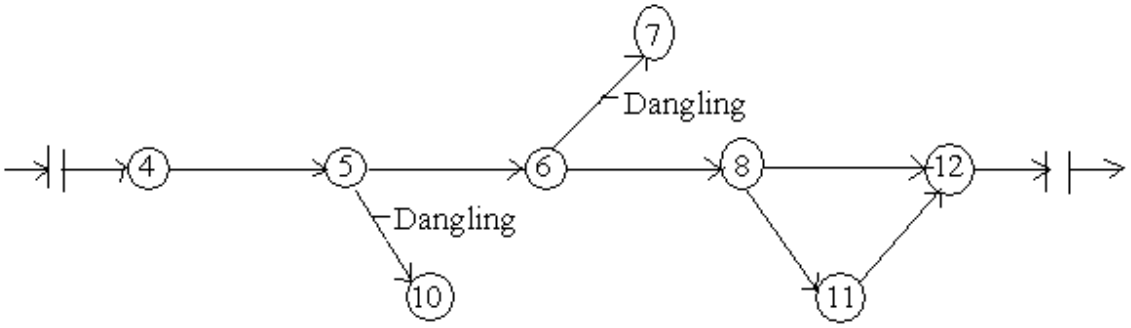
- Try to avoid arrows which cross each other
- Use straight arrows
- Do not attempt to represent duration of activity by its arrow length
- Use arrows from left to right. Avoid mixing two directions, vertical and standing arrows may be used if necessary.
- Use dummies freely in rough draft but final network should not have any redundant dummies.
- The network has only one entry point called start event and one point of emergence called the end event.

Common Errors in Drawing Networks

The three types of errors are most commonly observed in drawing network diagrams

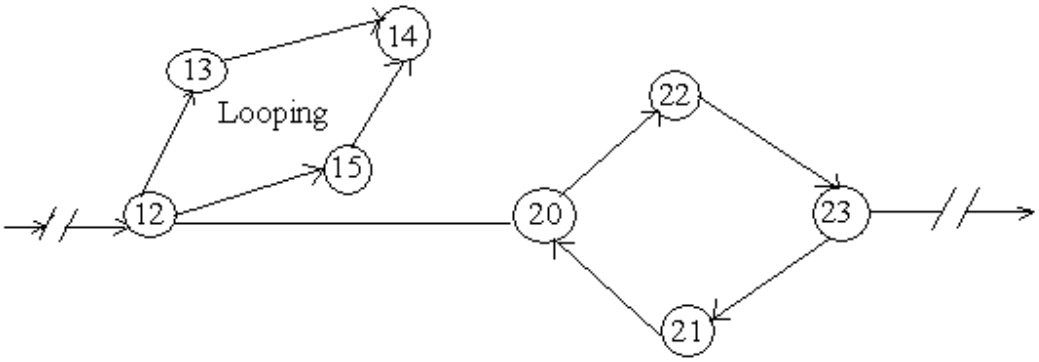
1. Dangling

To disconnect an activity before the completion of all activities in a network diagram is known as dangling. As shown in the figure activities (5 – 10) and (6 – 7) are not the last activities in the network. So the diagram is wrong and indicates the error of dangling



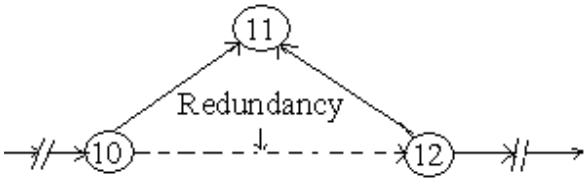
2. Looping or Cycling

Looping error is also known as cycling error in a network diagram. Drawing an endless loop in a network is known as error of looping as shown in the following figure.



3. Redundancy

Unnecessarily inserting the dummy activity in network logic is known as the error of redundancy as shown in the following diagram



Advantages and Disadvantages

PERT/CPM has the following advantages

- A PERT/CPM chart explicitly defines and makes visible dependencies (precedence relationships) between the elements,
- PERT/CPM facilitates identification of the critical path and makes this visible,
- PERT/CPM facilitates identification of early start, late start, and slack for each activity,
- PERT/CPM provides for potentially reduced project duration due to better understanding of dependencies leading to improved overlapping of activities and tasks where feasible.

PERT/CPM has the following disadvantages:

- There can be potentially hundreds or thousands of activities and individual dependency relationships,
- The network charts tend to be large and unwieldy requiring several pages to print and requiring special size paper,
- The lack of a timeframe on most PERT/CPM charts makes it harder to show status although colours can help (e.g., specific colour for completed nodes),
- When the PERT/CPM charts become unwieldy, they are no longer used to manage the project.

Critical Path in Network Analysis

Basic Scheduling Computations

The notations used are

(i, j) = Activity with tail event i and head event j

E_i = Earliest occurrence time of event i

L_j = Latest allowable occurrence time of event j

D_{ij} = Estimated completion time of activity (i, j)

$(Es)_{ij}$ = Earliest starting time of activity (i, j)

$(Ef)_{ij}$ = Earliest finishing time of activity (i, j)

$(Ls)_{ij}$ = Latest starting time of activity (i, j)

$(Lf)_{ij}$ = Latest finishing time of activity (i, j)

The procedure is as follows

1. **Determination of Earliest time (E_j): Forward Pass computation**

- **Step 1**

  The computation begins from the start node and move towards the end node. For easiness, the forward pass computation starts by assuming the earliest occurrence time of zero for the initial project event.

- **Step 2**

  i. Earliest starting time of activity (i, j) is the earliest event time of the tail end event i.e. $(Es)_{ij} = E_i$

  ii. Earliest finish time of activity (i, j) is the earliest starting time + the activity time i.e. $(Ef)_{ij} = (Es)_{ij} + D_{ij}$ or $(Ef)_{ij} = E_i + D_{ij}$

iii. Earliest event time for event j is the maximum of the earliest finish times of all activities ending in to that event i.e. E_j = max $[(Ef)_{ij}$ for all immediate predecessor of (i, j)] or E_j =max $[E_i + D_{ij}]$

2. Backward Pass computation (for latest allowable time)

- **Step 1**

  For ending event assume E = L. Remember that all E's have been computed by forward pass computations.

- **Step 2**

  Latest finish time for activity (i, j) is equal to the latest event time of event j i.e. $(Lf)_{ij} = L_j$

- **Step 3**

  Latest starting time of activity (i, j) = the latest completion time of (i, j) – the activity time or $(Ls)_{ij} = (Lf)_{ij} - D_{ij}$ or $(Ls)_{ij} = L_j - D_{ij}$

- **Step 4**

  Latest event time for event 'i' is the minimum of the latest start time of all activities originating from that event i.e. L_i = min $[(Ls)_{ij}$ for all immediate successor of (i, j)] = min $[(Lf)_{ij} - D_{ij}]$ = min $[L_j - D_{ij}]$

3. Determination of floats and slack times

There are three kinds of floats

- **Total float** – The amount of time by which the completion of an activity could be delayed beyond the earliest expected completion time without affecting the overall project duration time.

  Mathematically

  $(Tf)_{ij}$ = (Latest start – Earliest start) for activity (i – j)

  $(Tf)_{ij} = (Ls)_{ij} - (Es)_{ij}$ or $(Tf)_{ij} = (L_j - D_{ij}) - E_i$

- **Free float –** The time by which the completion of an activity can be delayed beyond the earliest finish time without affecting the earliest start of a subsequent activity.

  Mathematically

  $(Ff)_{ij}$ = (Earliest time for event j – Earliest time for event i) – Activity time for (i, j)

  $(Ff)_{ij} = (E_j - E_i) - D_{ij}$

- **Independent float** – The amount of time by which the start of an activity can be delayed without effecting the earliest start time of any immediately following activities, assuming that the preceding activity has finished at its latest finish time.

  Mathematically

  $(If)_{ij} = (E_j - L_i) - D_{ij}$

  The negative independent float is always taken as zero.

- **Event slack** - It is defined as the difference between the latest event and earliest event times.

  Mathematically

  Head event slack = $L_j – E_j$, Tail event slack = $L_i - E_i$

4. Determination of critical path

- **Critical event** – The events with zero slack times are called critical events. In other words the event i is said to be critical if $E_i = L_i$

- **Critical activity** – The activities with zero total float are known as critical activities. In other words an activity is said to be critical if a delay in its start will cause a further delay in the completion date of the entire project.

- **Critical path** – The sequence of critical activities in a network is called critical path. The critical path is the longest path in the network from the starting event to ending event and defines the minimum time required to complete the project.

Exercise

1. What is PERT and CPM?
2. What are the advantages of using PERT/CPM?
3. Mention the applications of PERT/CPM
4. Explain the following terms
    a. Earliest time
    b. Latest time
    c. Total activity slack
    d. Event slack
    e. Critical path
5. Explain the CPM in network analysis.
6. What are the rules for drawing network diagram? Also mention the common errors that occur in drawing networks.
7. What is the difference between PERT and CPM/

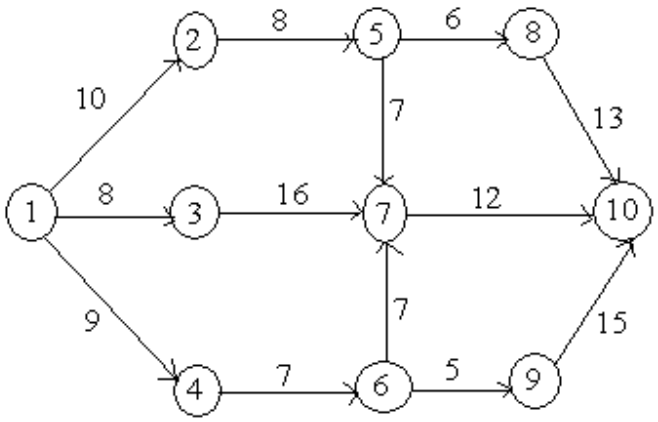
8. What are the uses of PERT and CPM?
9. Explain the basic steps in PERT/CPM techniques.
10. Write the framework of PERT/CPM.

Worked Examples on CPM
PERT
Worked Examples

Worked Examples on CPM

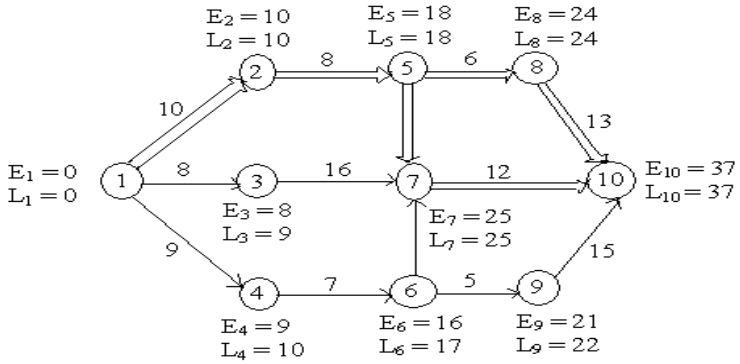
Example 1

Determine the early start and late start in respect of all node points and identify critical path for the following network.



Solution

Calculation of E and L for each node is shown in the network



Activity(i, j)	Normal Time (D_{ij})	Earliest Time		Latest Time		Float Time ($L_i - D_{ij}) - E_i$
		Start (E_i)	Finish ($E_i + D_{ij}$)	Start ($L_i - D_{ij}$)	Finish (L_i)	
(1, 2)	10	0	10	0	10	0
(1, 3)	8	0	8	1	9	1
(1, 4)	9	0	9	1	10	1
(2, 5)	8	10	18	10	18	0
(4, 6)	7	9	16	10	17	1

(3, 7)	16	8	24	9	25	1
(5, 7)	7	18	25	18	25	0
(6, 7)	7	16	23	18	25	2
(5, 8)	6	18	24	18	24	0
(6, 9)	5	16	21	17	22	1
(7, 10)	12	25	37	25	37	0
(8, 10)	13	24	37	24	37	0
(9, 10)	15	21	36	22	37	1

Network Analysis Table

From the table, the critical nodes are (1, 2), (2, 5), (5, 7), (5, 8), (7, 10) and (8, 10)

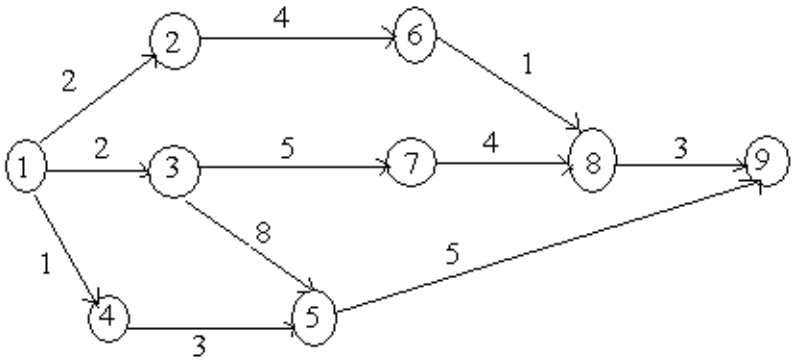
From the table, there are two possible critical paths

i. $1 \rightarrow 2 \rightarrow 5 \rightarrow 8 \rightarrow 10$

ii. $1 \rightarrow 2 \rightarrow 5 \rightarrow 7 \rightarrow 10$

Example 2

Find the critical path and calculate the slack time for the following network

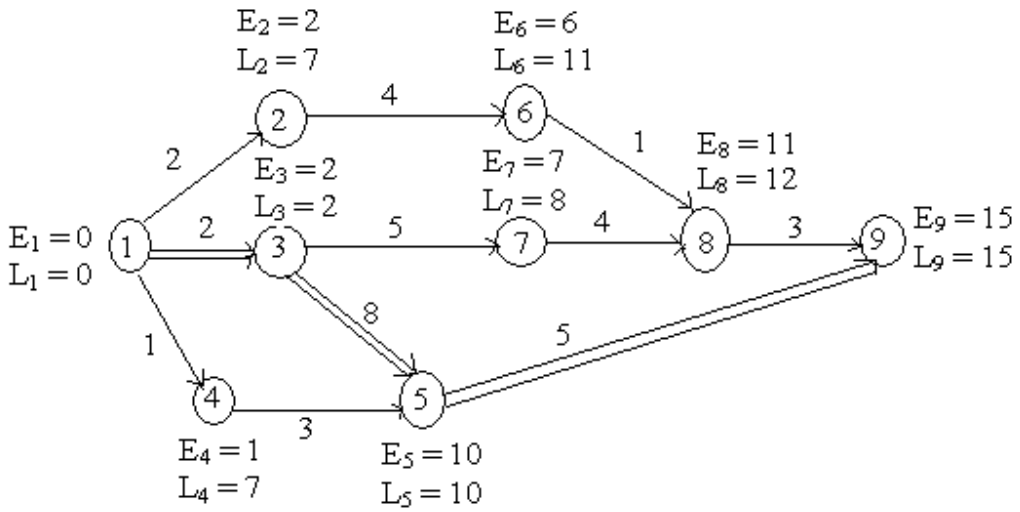


Solution

The earliest time and the latest time are obtained below

Activity(i, j)	Normal Time (D_{ij})	Earliest Time		Latest Time		Float Time ($L_i - D_{ij}) - E_i$
		Start (E_i)	Finish ($E_i + D_{ij}$)	Start ($L_i - D_{ij}$)	Finish (L_i)	
(1, 2)	2	0	2	5	7	5
(1, 3)	2	0	2	0	2	0
(1, 4)	1	0	1	6	7	6
(2, 6)	4	2	6	7	11	5
(3, 7)	5	2	7	3	8	1
(3, 5)	8	2	10	2	10	0
(4, 5)	3	1	4	7	10	6
(5, 9)	5	10	15	10	15	0
(6, 8)	1	6	7	11	12	5
(7, 8)	4	7	11	8	12	1
(8, 9)	3	11	14	12	15	1

From the above table, the critical nodes are the activities (1, 3), (3, 5) and (5, 9)



The critical path is $1 \rightarrow 3 \rightarrow 5 \rightarrow 9$

Example 3

A project has the following times schedule

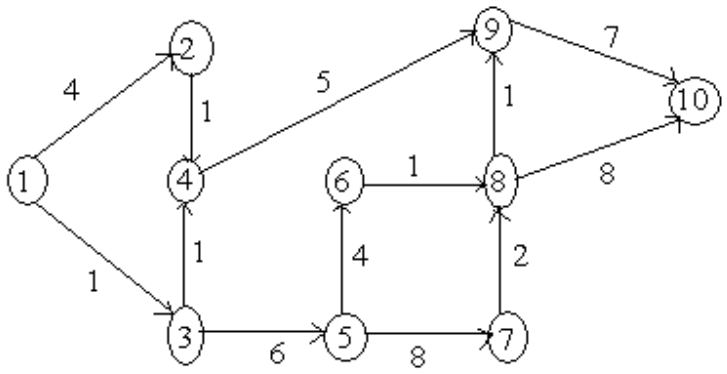
Activity	Times in weeks	Activity	Times in weeks
(1 – 2)	4	(5 – 7)	8
(1 – 3)	1	(6 – 8)	1
(2 – 4)	1	(7 – 8)	2
(3 – 4)	1	(8 – 9)	1
(3 – 5)	6	(8 – 10)	8
(4 – 9)	5	(9 – 10)	7
(5 – 6)	4		

Construct the network and compute

1. T_E and T_L for each event
2. Float for each activity
3. Critical path and its duration

Solution

The network is

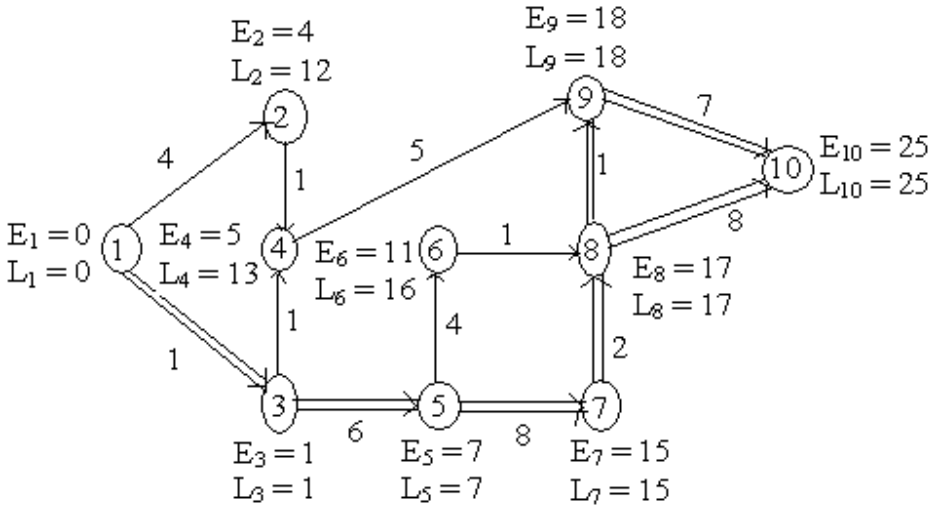


Event No.:	1	2	3	4	5	6	7	8	9	10
T_E:	0	4	1	5	7	11	15	17	18	25
T_L:	0	12	1	13	7	16	15	17	18	25

Float = T_L (Head event) – T_E (Tail event) – Duration

Activity	Duration	T_E (Tail event)	T_L (Head event)	Float
(1 – 2)	4	0	12	8
(1 – 3)	1	0	1	0
(2 – 4)	1	4	13	8
(3 – 4)	1	1	13	11
(3 – 5)	6	1	7	0
(4 – 9)	5	5	18	8
(5 – 6)	4	7	16	5
(5 – 7)	8	7	15	0
(6 – 8)	1	11	17	5
(7 – 8)	2	15	17	0
(8 – 9)	1	17	18	0
(8 – 10)	8	17	25	0
(9 – 10)	7	18	25	0

The resultant network shows the critical path



The two critical paths are

i. $1 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 8 \rightarrow 9 \rightarrow 10$

ii. $1 \rightarrow 3 \rightarrow 5 \rightarrow 7 \rightarrow 8 \rightarrow 10$

Project Evaluation and Review Technique (PERT)

The main objective in the analysis through PERT is to find out the completion for a particular event within specified date. The PERT approach takes into account the uncertainties. The three time values are associated with each activity

1. **Optimistic time** – It is the shortest possible time in which the activity can be finished. It assumes that every thing goes very well. This is denoted by t_0.
2. **Most likely time** – It is the estimate of the normal time the activity would take. This assumes normal delays. If a graph is plotted in the time of completion and the frequency of completion in that time period, then most likely time will represent the highest frequency of occurrence. This is denoted by t_m.
3. **Pessimistic time** – It represents the longest time the activity could take if everything goes wrong. As in optimistic estimate, this value may be such that

only one in hundred or one in twenty will take time longer than this value. This is denoted by t_p.

In PERT calculation, all values are used to obtain the percent expected value.

1. **Expected time** – It is the average time an activity will take if it were to be repeated on large number of times and is based on the assumption that the activity time follows Beta distribution, this is given by

$$t_e = (t_0 + 4\, t_m + t_p) / 6$$

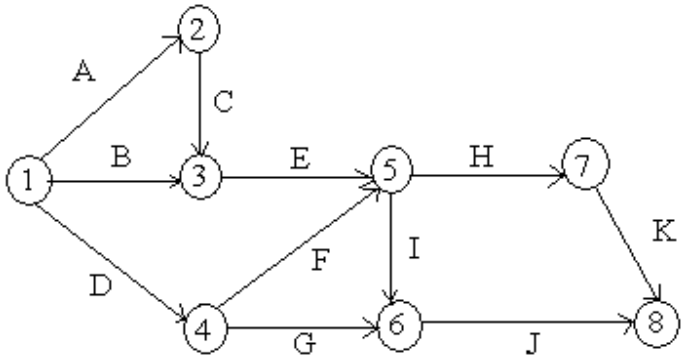
2. The **variance** for the activity is given by

$$\sigma^2 = [(t_p - t_o) / 6]^2$$

Worked Examples

Example 1

For the project



Task:	A	B	C	D	E	F	G	H	I	J	K
Least time:	4	5	8	2	4	6	8	5	3	5	6

Greatest time:	8	10	12	7	10	15	16	9	7	11	13
Most likely time:	5	7	11	3	7	9	12	6	5	8	9

Find the earliest and latest expected time to each event and also critical path in the network.

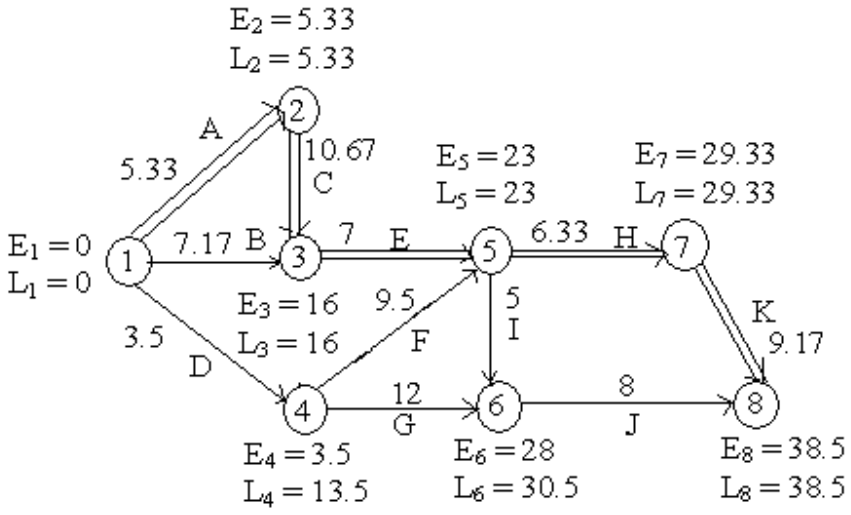
Solution

Task	Least time(t_0)	Greatest time (t_p)	Most likely time (t_m)	Expected time $(to + t_p + 4t_m)/6$
A	4	8	5	5.33
B	5	10	7	7.17
C	8	12	11	10.67
D	2	7	3	3.5
E	4	10	7	7
F	6	15	9	9.5
G	8	16	12	12
H	5	9	6	6.33
I	3	7	5	5
J	5	11	8	8
K	6	13	9	9.17

Task	Expected time (t_e)	Start		Finish		Total float
		Earliest	Latest	Earliest	Latest	
A	5.33	0	0	5.33	5.33	0
B	7.17	0	8.83	7.17	16	8.83
C	10.67	5.33	5.33	16	16	0
D	3.5	0	10	3.5	13.5	10
E	7	16	16	23	23	0

F	9.5	3.5	13.5	13	23	10
G	12	3.5	18.5	15.5	30.5	15
H	6.33	23	23	29.33	29.33	0
I	5	23	25.5	28	30.5	2.5
J	8	28	30.5	36	38.5	2.5
K	9.17	29.33	29.33	31.5	38.5	0

The network is



The critical path is A →C →E → H → K

Example 2

A project has the following characteristics

Activity	Most optimistic time (a)	Most pessimistic time (b)	Most likely time (m)
(1 – 2)	1	5	1.5
(2 – 3)	1	3	2
(2 – 4)	1	5	3

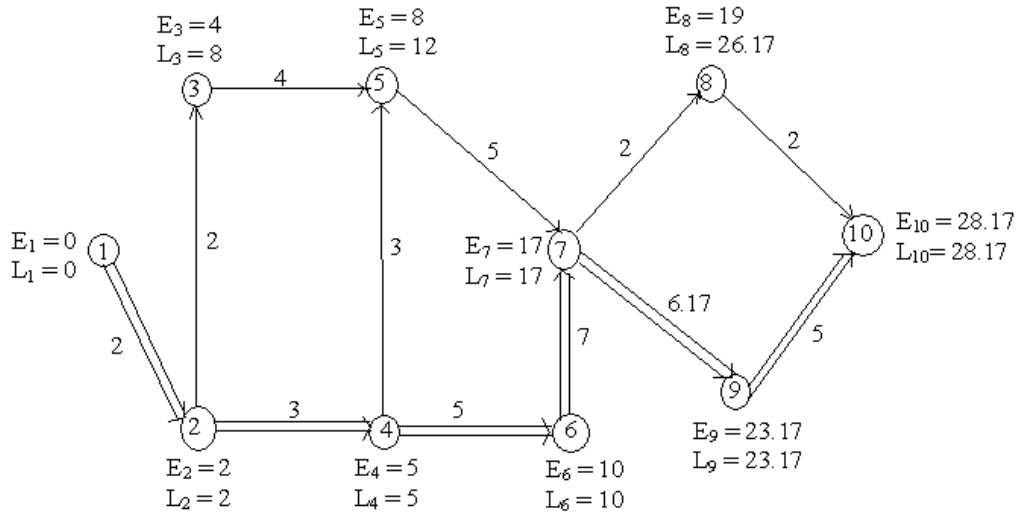
(3 – 5)	3	5	4
(4 – 5)	2	4	3
(4 – 6)	3	7	5
(5 – 7)	4	6	5
(6 – 7)	6	8	7
(7 – 8)	2	6	4
(7 – 9)	5	8	6
(8 – 10)	1	3	2
(9 – 10)	3	7	5

Construct a PERT network. Find the critical path and variance for each event.

Solution

Activity	(a)	(b)	(m)	(4m)	t_e $(a + b + 4m)/6$	v $[(b – a) / 6]^2$
(1 – 2)	1	5	1.5	6	2	4/9
(2 – 3)	1	3	2	8	2	1/9
(2 – 4)	1	5	3	12	3	4/9
(3 – 5)	3	5	4	16	4	1/9
(4 – 5)	2	4	3	12	3	1/9
(4 – 6)	3	7	5	20	5	4/9
(5 – 7)	4	6	5	20	5	1/9
(6 – 7)	6	8	7	28	7	1/9
(7 – 8)	2	6	4	16	4	4/9
(7 – 9)	5	8	6	24	6.17	1/4
(8 – 10)	1	3	2	8	2	1/9
(9 – 10)	3	7	5	20	5	4/9

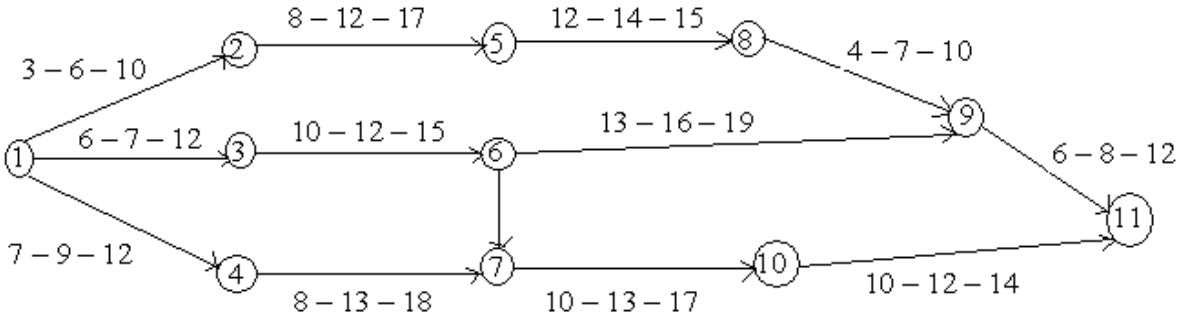
The network is constructed as shown below



The critical path = 1 → 2 → 4 → 6 → 7 →9 →10

Example 3

Calculate the variance and the expected time for each activity



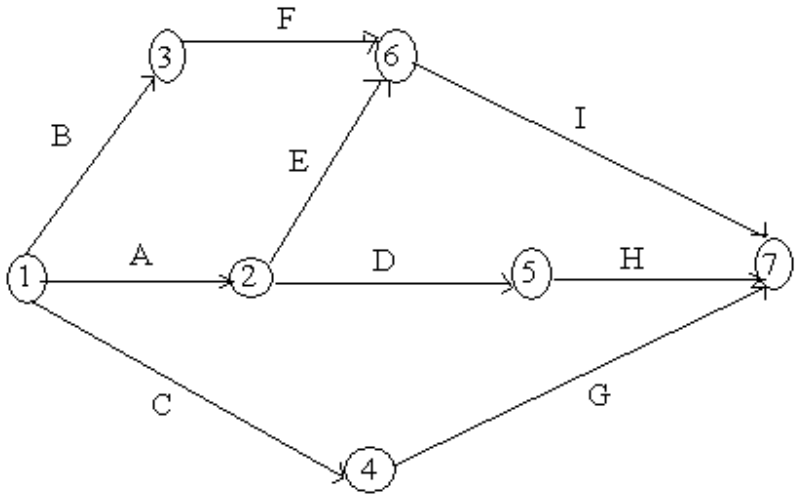
Solution

Activity	(t_o)	(t_m)	(t_p)	t_e $(t_o + t_p + 4t_m)/6$	v $[(t_p - t_o) / 6]^2$
(1 – 2)	3	6	10	6.2	1.36
(1 – 3)	6	7	12	7.7	1.00
(1 – 4)	7	9	12	9.2	0.69

(2 – 3)	0	0	0	0.0	0.00
(2 – 5)	8	12	17	12.2	2.25
(3 – 6)	10	12	15	12.2	0.69
(4 – 7)	8	13	19	13.2	3.36
(5 – 8)	12	14	15	13.9	0.25
(6 – 7)	8	9	10	9.0	0.11
(6 – 9)	13	16	19	16.0	1.00
(8 – 9)	4	7	10	7.0	1.00
(7 – 10)	10	13	17	13.2	1.36
(9 – 11)	6	8	12	8.4	1.00
(10 – 11)	10	12	14	12.0	0.66

Example 4

A project is represented by the network as shown below and has the following data



Task:	A	B	C	D	E	F	G	H	I
Least time:	5	18	26	16	15	6	7	7	3

Greatest time:	10	22	40	20	25	12	12	9	5
Most likely time:	15	20	33	18	20	9	10	8	4

Determine the following

1. Expected task time and their variance
2. Earliest and latest time

Solution

1.

Activity	Least time (t_0)	Greatest time (t_p)	Most likely time (t_m)	Expected time $(to + t_p + 4t_m)/6$	Variance (σ^2)
(1-2)	5	10	8	7.8	0.69
(1-3)	18	22	20	20.0	0.44
(1-4)	26	40	33	33.0	5.43
(2-5)	16	20	18	18.0	0.44
(2-6)	15	25	20	20.0	2.78
(3-6)	6	12	9	9.0	1.00
(4-7)	7	12	10	9.8	0.69
(5-7)	7	9	8	8.0	0.11
(6-7)	3	5	4	4.0	0.11

2.

Earliest time

$E_1 = 0$

$E_2 = 0 + 7.8 = 7.8$

$E_3 = 0 + 20 = 20$

$E_4 = 0 + 33 = 33$

$E_5 = 7.8 + 18 = 25.8$

$E_6 = \max [7.8 + 20, 20 + 9] = 29$

$E_7 = \max [33 + 9.8, 25.8 + 8, 29 + 4] = 42.8$

Latest time

$L_7 = 42.8$

$L_6 = 42.8 – 4 = 38.8$

$L_5 = 42.8 – 8 = 34.3$

$L_4 = 42.8 – 9.8 = 33$

$L_3 = 38.8 – 9 = 29.8$

$L_2 = \min [34.8 – 18, 38.8 – 20] = 16.8$

$L_1 = \min [16.8 – 7.8, 29.8 – 20, 33 - 33] = 0$

Exercise

1. What is PERT?

2. For the following data, draw network. Find the critical path, slack time after calculating the earliest expected time and the latest allowable time

Activity	Duration	Activity	Duration
(1 – 2)	5	(5 – 9)	3
(1 – 3)	8	(6 – 10)	5
(2 – 4)	6	(7 – 10)	4
(2 – 5)	4	(8 – 11)	9
(2 – 6)	4	(9 – 12)	2
(3– 7)	5	(10 – 12)	4
(3 – 8)	3	(11 – 13)	1
(4 – 9)	1	(12 – 13)	7

[Ans. Critical path: 1 → 3 → 7 → 10 → 12 →13]

3. A project schedule has the following characteristics

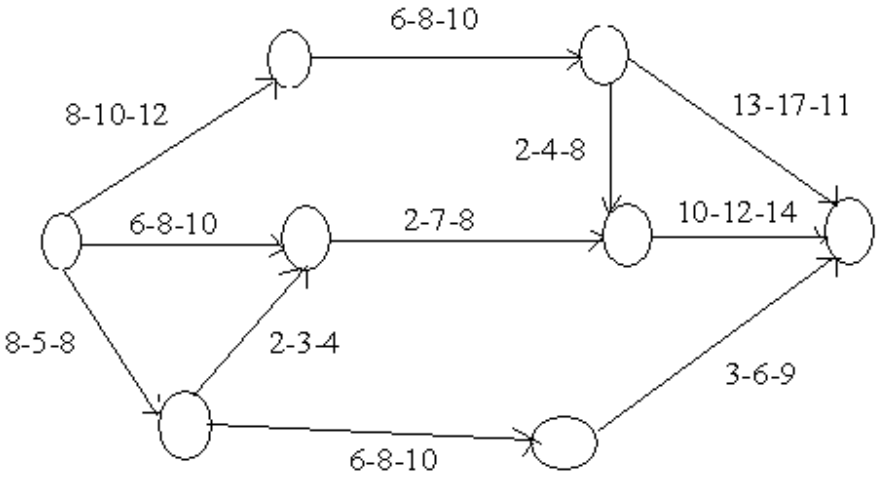
Activity	Most optimistic time	Most likely time	Most pessimistic time
(1 – 2)	1	2	3
(2 – 3)	1	2	3
(2 – 4)	1	3	5
(3 – 5)	3	4	5
(4 – 5)	2	5	4
(4 – 6)	3	5	7
(5 – 7)	4	5	6
(6 – 7)	6	7	8
(7 – 8)	2	4	6

(7 – 9)	4	6	8
(8 – 10)	1	2	3
(9 – 10)	3	5	7

Construct a PERT network and find out

a. The earliest possible time
b. Latest allowable time
c. Slack values
d. Critical path

4. Explain the following terms
    a. optimistic time
    b. Most likely time
    c. Pessimistic time
    d. Expected time
    e. Variance
5. Calculate the variance and the expected time for each activity



Unit 9

Course Structure

- Introduction
- Objectives
- Matrix: definition, order
- Symmetric and skew-symmetric matrices

Introduction

The matrix has a long history of application in solving linear equations. They were known as arrays until the 1800 ‘s. The term “matrix” (Latin for “womb”, derived from mater—mother) was coined by James Joseph Sylvester in 1850 , who understood a matrix as an object giving rise to a number of determinants today called minors, that is to say, determinants of smaller matrices that are derived from the original one by removing columns and rows. An English mathematician named Cullis was the first to use modern bracket notation for matrices in 1913 and he simultaneously demonstrated the first significant use of the notation $A = a_{ij}$ to represent a matrix where a_{ij} refers to the element found in the ith row and the jth column. Matrices can be used to compactly write and work with multiple linear equations, referred to as a system of linear equations, simultaneously. Matrices and matrix multiplication reveal their essential features when related to linear transformations, also known as linear maps.

Objectives

After reading this unit you will be able to:

- describe the parts of a matrix and what they represent
- add, subtract and multiply matrices
- distinguish the symmetric and skew-symmetric matrices
- come across various examples of matrices along with their applications

Matrix: Definition

In mathematics, a matrix (plural matrices) is a rectangular array of numbers, symbols, or expressions, arranged in rows and columns. Matrices are commonly written in box brackets. Unless specified, we will consider the entries of matrix to be real or complex numbers. The horizontal and vertical lines of entries in a matrix are called rows and columns, respectively. The size of a matrix is defined by the number of rows and columns that it contains. A matrix with m rows and n columns is called an $m \times n$ matrix , while m and n are called its dimensions.The dimensions of the following matrix is 3×2, because there are three rows and two columns.

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

The individual items (numbers, symbols or expressions) in a matrix are called its elements or entries. The elements in a row together form a row vector and the elements in a column together forms a column vector. Provided that they are the same size (have the same number of rows and the same number of columns), two matrices can be added or subtracted element by element. The rule for matrix multiplication, however, is that two matrices can be multiplied only when the number of columns in the first equals the number of rows in the second. Any matrix can be multiplied element-wise by a scalar from its associated field.

Matrices which have a single row are called row vectors, and those which have a single column are called column vectors. A matrix which has the same number of rows and columns is called a square matrix. In some contexts, such as computer algebra programs, it is useful to consider a matrix with no rows or no columns, called an empty matrix.

Operations of Matrices

Matrix addition, subtraction, and scalar multiplication are types of operations that can be applied to modify matrices. There are a number of operations that can be applied to modify matrices, such as matrix addition, subtraction, and scalar multiplication. These form the basic techniques to work with matrices. Matrix addition and subtraction requires both the matrices to be of equal dimensions. That is, if A and B are both of $m \times n$ order, then only matrix addition and subtraction are defined. Suppose,

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},\ B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

Then, the matrix addition and subtraction are defined as

$$\begin{aligned} A + B &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} \\ \text{and}\quad A - B &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} \\ a_{21} - b_{21} & a_{22} - b_{22} \end{bmatrix} \end{aligned}$$

The matrix addition and subtraction of any $n \times n$ matrix are analogously defined. Next, we define matrix multiplication. The first condition that needs to be satisfied for any two matrices $A_{m \times n}$ and $B_{p \times q}$ to get

multiplied is $n = p$. Then the resulting matrix AB is of order $m \times q$. Let us illustrate a matrix multiplication of 2×3 and 3×2 matrices

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 0 & 3 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 1 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1\times 3+2\times 2+1\times 0 & 1\times 2+2\times 1+1\times 3 \\ 2\times 3+0\times 2+3\times 0 & 2\times 2+0\times 1+3\times 3 \end{bmatrix}$$
$$= \begin{bmatrix} 7 & 7 \\ 6 & 13 \end{bmatrix}$$

The next thing that we come across is the multiplication of a matrix by a scalar, that is a real number, say k. The scalar multiplication is defined as

$$k.A = k \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$
$$= \begin{bmatrix} k.a_{11} & k.a_{12} \\ k.a_{21} & k.a_{22} \end{bmatrix}$$

Identity Matrix

We know that whenever a real number, say a, is multiplied by 1, we always get a. That is, $a \times 1 = 1 \times a = a$. The same idea is extendible in case of matrices, particularly those matrices $A_{m\times n}$, where $m = n$ (such matrices are called square matrices). Then the identity matrix of order $n \times n$ is defined as

$$I_{n\times n} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

such that for any matrix $A_{n\times n}$, we will have,

$$A_{n\times n} \times I_{n\times n} = I_{n\times n} \times A_{n\times n} = A_{n\times n}$$

Inverse of a Matrix

Let $A_{n\times n}$ be a matrix. A matrix $B_{n\times n}$ is said to be the inverse of the matrix A if it satisfies the following condition:

$$A \times B = B \times A = I$$

where, I is the identity matrix of order n. We will learn more about inverse of a matrix in the next unit.

Transpose of a matrix

Let A be an $m \times n$ matrix defined as

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Then the transpose of A is denoted by A^T and defined as

$$A^T = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{nm} \end{bmatrix}$$

Note that the resulting matrix is of order $n \times m$.

Exercise 9.1. 1. Find the transpose of the matrices

i.
$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

ii.
$$\begin{bmatrix} 1 & 3 \\ -1 & 0 \\ 7 & 10 \end{bmatrix}$$

iii.
$$\begin{bmatrix} 3 & 5 \\ 8 & 9 \end{bmatrix}$$

2. Let v be an $n \times 1$ matrix. Prove that vv^T is a symmetric matrix.
3. If A and B are two $n \times n$ symmetric matrices, then show that $A + B$ is also symmetric.

Few properties of transpose are:

- $(A^T)^T = A$
- $(A \pm B)^T = A^T \pm B^T$
- $(kA)^T = kA^T$
- $(AB)^T = A^T B^T$

for any two matrices A and B and a constant k.

Symmetric and Skew-Symmetric Matrices

Let A be an $m \times n$ matrix. Then A is said to be a symmetric matrix if $A = A^T$. Suppose,

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Then, $A = A^T$ implies that,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{nm} \end{bmatrix}$$

Then from the above equality, we can immediately conclude the following:

1. $m = n$
2. $x_{ij} = x_{ji}$ for all $1 \leq i, j \leq n$

Example 9.2. i. For all n, the identity matrix is symmetric.

ii.
$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 7 \\ 3 & 7 & 1 \end{bmatrix}$$
are two symmetric matrices.

iii. The matrices
$$\begin{bmatrix} 1 & 5 & 8 \\ 2 & 7 & 4 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 5 & 8 \\ 2 & 7 & 4 \\ 4 & 7 & 0 \end{bmatrix}$$
are not symmetric.

Skew-Symmetric Matrices

In a similar way, we can define skew-symmetric matrices. Let A be an $m \times n$ matrix. Then A is said to be a skew-symmetric matrix if $A = -A^T$. Suppose,

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix}$$

Then, $A = -A^T$ implies that,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{bmatrix} = \begin{bmatrix} -x_{11} & -x_{21} & -x_{31} & \dots & -x_{n1} \\ -x_{12} & -x_{22} & -x_{32} & \dots & -x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ -x_{1n} & -x_{2n} & -x_{3n} & \dots & -x_{nm} \end{bmatrix}$$

Then from the above equality, we can immediately conclude the following:

1. $m = n$
2. $x_{ij} = -x_{ji}$ for all $1 \leq i, j \leq n$ and $i \neq j$.
3. $x_{ii} = -x_{ii}$ for all $1 \leq i \leq n$, which implies that $x_{ii} = 0, \ \forall \ i$.

Thus, any skew-symmetric matrix looks like

$$\begin{bmatrix} 0 & x_{12} & x_{13} & x_{14} & \dots & x_{1n} \\ -x_{12} & 0 & x_{23} & x_{24} & \dots & x_{2n} \\ -x_{13} & -x_{23} & 0 & x_{34} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -x_{1n} & -x_{2n} & -x_{3n} & -x_{4n} & \dots & 0 \end{bmatrix}$$

Theorem 9.3. For any matrix A with real entries, $A + A^T$ is symmetric and $A - A^T$ is skew-symmetric.

Proof. Let A be a matrix and let $B = A + A^T$ and $C = A - A^T$. Then we see that,

$$\begin{aligned} B^T &= (A + A^T)^T \\ &= A^T + (A^T)^T \\ &= A^T + A \\ &= A + A^T \\ &= B \end{aligned}$$

Similarly,

$$\begin{aligned} C^T &= (A - A^T)^T \\ &= A^T - (A^T)^T \\ &= A^T - A \\ &= -(A - A^T) \\ &= -C \end{aligned}$$

Hence, B is symmetric and C is skew-symmetric. □

Theorem 9.4. Any square matrix can be written as the sum of a symmetric and a skew symmetric matrix.

Proof. Let A be a given matrix. Then A can be written as,

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$$

From the previous theorem, we can say that $\frac{1}{2}(A + A^T)$ is symmetric and $\frac{1}{2}(A - A^T)$ is skew-symmetric. Hence the theorem. □

Exercise 9.5. Let A and B be two $n \times n$ skew-symmetric matrices.

1. Prove that $A + B$ is skew-symmetric.
2. Prove that cA is skew-symmetric for any scalar c.
3. Let P be any $m \times n$ matrix. Prove that $P^T AP$ is skew-symmetric.
4. Prove that, if $AB = -BA$, then AB is a skew-symmetric matrix.

Unit 10

Course Structure

- Introduction
- Objectives
- Determinant of a matrix, elementary properties of determinants
- Inverse of a matrix
- Normal form of a matrix, rank of a matrix

Introduction

Determinant of a square matrix is a real number that is assigned to every square matrix. It is a scalar property of the matrix , which can be thought of as the volume enclosed by the row vectors of the matrix. Note that determinant is not defined for any arbitrary matrix. Now, Determinants are mathematical objects that are very useful in the analysis and solution of systems of linear equations. Determinants also have wide applications in engineering, science, economics and social science as well. In this unit, we will mainly study determinant of a matrix and its basic properties and few applications.

Objectives

After reading this unit, you will be able to:

- find the determinant of matrices (of lower order)
- know the properties of determinants
- learn few applications of determinant such as finding inverse of a matrix
- know about cofactors and minors of a matrix
- find the rank of a matrix

Determinant of a Matrix

Let A be a square matrix with real entries. Then, the determinant of A is denoted by $|A|$, or $\det A$. Finding $|A|$ depends upon the dimension of the matrix. For 2×2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

the determinant is defined as

$$|A| = a_{11}a_{22} - a_{12}a_{21}$$

Also, the determinant of a 3×3 matrix is

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

In general, the determinant of any $n \times n$ matrix is calculated as

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

where C_{ij} are referred to as the **cofactors** of A and are computed as

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

The term M_{ij} is known as the **Minor Matrix**, which is the matrix obtained by eliminating the ith row and jth column of the matrix A. So to find the determinant of e.g. a 4×4 matrix, you end up calculating a bunch of 3×3 matrix determinants which is much easier. Let us illustrate it for a 4×4 matrix.

Let

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

We intend to find the determinant of A. For that, we will first find the cofactors. We will set up a checkerboard of positive and negative signs as follows: for ith row and jth column, we write the sign as $(-1)^{i+j}$ as a superscript of the corresponding entries.

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

Next we will pick a row or column to expand on. It is easier to choose the row or column having maximum zeros, since it will make the calculations easier. We are using column 3.

$$\begin{aligned} |A| &= (1)(3)\begin{vmatrix} -3 & 1 & 4 \\ -2 & 1 & 1 \\ -1 & 3 & -2 \end{vmatrix} + (-1)(0)\begin{vmatrix} 2 & -1 & 0 \\ -2 & 1 & 1 \\ -1 & 3 & -2 \end{vmatrix} + (1)(4)\begin{vmatrix} 2 & -1 & 0 \\ -3 & 1 & 4 \\ -1 & 3 & -2 \end{vmatrix} + (-1)(0)\begin{vmatrix} 2 & -1 & 0 \\ -3 & 1 & 4 \\ -2 & 1 & 1 \end{vmatrix} \\ &= 3\begin{vmatrix} -3 & 1 & 4 \\ -2 & 1 & 1 \\ -1 & 3 & -2 \end{vmatrix} + 4\begin{vmatrix} 2 & -1 & 0 \\ -3 & 1 & 4 \\ -1 & 3 & -2 \end{vmatrix} \\ &= 3(-10) + 4(-18) \\ &= 102 \end{aligned}$$

Exercise 10.1. Find the determinant of the following matrices:

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

2. $$\begin{bmatrix} 1 & 5 & 7 \\ 5 & 2 & -6 \\ 2 & 1 & 0 \end{bmatrix}$$

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

Let us now discuss a few important properties of determinant. Let A be an $n \times n$ square matrix. Then,

- If two rows of A are equal or, any row of A is expressible as a linear combination of other rows, then $|A| = 0$. This is true for columns as well.
- $|A| = |A^T|$
- The value of the determinant remains unchanged if both rows and columns are interchanged.
- If any two rows (or columns) of a determinant are interchanged, then sign of determinant changes.
- If each element of a row (or a column) of a determinant is multiplied by a constant k, then its value gets multiplied by k. In other words, $|kA| = k^n|A|$.
- If some or all elements of a row or column of a determinant are expressed as the sum of two (or more) terms, then the determinant can be expressed as the sum of two (or more) determinants.
- If all the elements of a row (or a column) are zero, then the determinant is zero.

Let us work out some examples using the properties of determinant.

Example 10.2. Let us find the value of the determinant

$$\begin{vmatrix} \sin^2 x & \cos^2 x & 1 \\ \cos^2 x & \sin^2 & 1 \\ -10 & 12 & 2 \end{vmatrix}$$

By $C_1 = C_1 + C_2$, the above determinant transforms to

$$\begin{vmatrix} 1 & \cos^2 x & 1 \\ 1 & \sin^2 & 1 \\ 2 & 12 & 2 \end{vmatrix} = 0$$

since the columns C_1 and C_2 are identical.

Example 10.3. Solve for x

$$\begin{vmatrix} x & 2 & -1 \\ 2 & 5 & x \\ -1 & 2 & x \end{vmatrix} = 0$$

By $R_3 = R_3 - R_2$, the L.H.S. gets reduced to

$$\begin{aligned} \begin{vmatrix} x & 2 & -1 \\ 2 & 5 & x \\ -3 & -3 & 0 \end{vmatrix} &= 0 \\ \text{or, } x(5 \times 0 + 3x) - 2(2 \times 0 + 3x) - 1(-6 + 15) &= 0 \\ \text{or, } 3x^2 - 6x - 9 &= \\ \text{or, } x^2 - 2x - 3 &= \\ \text{or, } (x-3)(x+1) &= 0 \\ \text{or, } x &= 3, -1 \end{aligned}$$

Example 10.4. We show that

$$\begin{vmatrix} 115 & 106 & 97 \\ 10 & 1 & -8 \\ 106 & 97 & 88 \end{vmatrix} = 0$$

Operating $C_2 = C_2 - \frac{1}{2}(C_1 + C_3)$, we get,

$$\begin{vmatrix} 115 & 0 & 97 \\ 10 & 0 & -8 \\ 106 & 0 & 88 \end{vmatrix} = 0$$

Exercise 10.5. 1. Without expanding, prove that,

$$\begin{vmatrix} 1 & a & a^2 - bc \\ 1 & b & b^2 - ca \\ 1 & c & c^2 - ab \end{vmatrix} = 0$$

2. Evaluate

$$\begin{vmatrix} 1 & \log_x y & \log_x z \\ \log_y x & 1 & \log_y z \\ \log_z x & \log_z y & 1 \end{vmatrix}$$

where x, y, z are positive.

Inverse of a Matrix

As mentioned in the previous unit, an $n \times n$ matrix is said to be invertible if there exists another $n \times n$ matrix B such that

$$AB = BA = I$$

where I is the identity matrix of order n.

We can also define invertible matrix as follows:

Definition 10.6. An $n \times n$ matrix A is said to be *invertible* or *non-singular* if $|A| \neq 0$. The matrix whose determinant is zero is called *singular* matrix. The set of all non-singular matrices of order n and entries from a set X is called the General Linear Group of order n and denoted by $GL(n, X)$. If X is the set of real numbers, then it is denoted by $GL(n, \mathbb{R})$.

But why do we need to find the inverse of a matrix in the first place? To understand this, let's see the following equation:

$$ax = b$$

where a and b are real numbers, and we intend to find x. What we do is simply divide both sides by a and get the solution as

$$x = \frac{b}{a}$$

provided that $a \neq 0$. But what happens if we replace the real numbers a, b, x by real matrices?

$$AX = B$$

where A, B, X are matrices. We could not do it like below

$$X = \frac{B}{A}$$

since the division of matrix is as such not defined. This is where, the matrix inverse comes into play. What we do is multiply both sides by the inverse of A, provided the inverse exists, that is, $|A| \neq 0$(which is analogous to the condition $a \neq 0$ in the previous case). What we get in that case is,

$$\begin{aligned} A^{-1}AX &= A^{-1}B \\ \text{or, } IX &= A^{-1}B \\ \text{or, } X &= A^{-1}B \end{aligned}$$

and hence, we are done.

How to find the inverse of a given matrix?

Suppose we are given an $n \times n$ matrix A as:

$$A = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & x_{24} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & x_{34} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \dots & x_{nn} \end{bmatrix}$$

First check whether $|A| = 0$ or not. If so, then the inverse of A does not exist and we are done. If not, then proceed to the next step. Find out the cofactor for each term of the matrix. Let C_{ij} be the cofactor of each term a_{ij} of the matrix A. Then the inverse is given by

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & \dots & C_{1n} \\ C_{21} & C_{22} & C_{23} & C_{24} & \dots & C_{2n} \\ C_{31} & C_{32} & C_{33} & C_{34} & \dots & C_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ C_{nn} & C_{n2} & C_{n3} & C_{n4} & \dots & C_{nn} \end{bmatrix}^T$$

Note that the transpose of the cofactor matrix is called *adjoint* of the matrix.

Illustration

Let A be a 3×3 matrix given below:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 5 \\ 5 & 6 & 0 \end{bmatrix}$$

First of all find the determinant of A by the said method. We see that $|A| = 5$. Then, we find the cofactor of each term. For example, let us find C_{11}. For this, first exclude the first row and fist column from the matrix. Then find the determinant of the remaining matrix and multiply it by $(-1)^{1+1} = (-1)^2 = 1$, that is,

$$\begin{aligned} C_{11} &= 1. \begin{vmatrix} 1 & 5 \\ 6 & 0 \end{vmatrix} \\ &= 1.0 - 5.6 \\ &= -30 \end{aligned}$$

Similarly, we find $C_{12} = 25$, $C_{13} = -5$, $C_{21} = 18$, $C_{22} = -15$, $C_{23} = 4$, $C_{31} = 7$, $C_{32} = -5$, $C_{33} = 1$. Hence the adjoint is given by,

$$C = \begin{bmatrix} -30 & 18 & 7 \\ 25 & -15 & -5 \\ -5 & -4 & 1 \end{bmatrix}$$

Hence, the inverse is

$$\begin{aligned} A^{-1} &= \frac{1}{5} \begin{bmatrix} -30 & 18 & 7 \\ 25 & -15 & -5 \\ -5 & -4 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -6 & 18/5 & 7/5 \\ 5 & -3 & -1 \\ -1 & -4/5 & 1/5 \end{bmatrix} \end{aligned}$$

Exercise 10.7. Find the inverse of the following matrices(if exists):

1. $$\begin{bmatrix} 3 & 4 \\ 6 & 8 \end{bmatrix}$$

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

Unit 11

Course Structure

- Introduction
- Objectives
- Elementary concept of a vector space
- Linear dependence and independence of vectors, basis of a vector space, row space, column space
- Solution of system of linear equations, Cramar's rule

Introduction

Consider arrows in a fixed plane starting at one fixed point, say the origin. Given any two such arrows, v and w, the parallelogram spanned by these two arrows contains one diagonal arrow that starts at the origin, too. This new arrow is called the sum of the two arrows and is denoted $v + w$. In the special case of two arrows on the same line, their sum is the arrow on this line whose length is the sum or the difference of the lengths, depending on whether the arrows have the same direction. Another operation that can be done with arrows is scaling: given any positive real number a, the arrow that has the same direction as v, but is dilated or shrunk by multiplying its length by a, is called multiplication of v by a. It is denoted av. When a is negative, av is defined as the arrow pointing in the opposite direction, instead.

Objectives

After studying this unit, you will be able to:

- know the definition of vector spaces
- learn various standard vector spaces
- visualize linearly independent and dependent vectors
- find out basis of a vector space
- solve system of linear equations independently and by Cramer's Rule

Vector Space

As the name suggests, a vector space is a non-empty set V, say, with two binary compositions, one external between the elements of V and the associated field F, and one internal composition between two elements of V, defined as follows:

1. for $x, y \in V$, we have, $x + y \in V$(closure property)
2. for $x, y, z \in V$, we have, $(x + y) + z = x + (y + z)$(associativity)
3. for all $v \in V$, we have, $v + 0 = 0 + v = v$(identity property)
4. for all $v \in V$, there exists $-v \in V$ such that, $v + (-v) = (-v) + v = 0$(additive inverse property)
5. for $x \in V$ and $s \in F$, we will always have $sx \in V$
6. for $x \in V$ and $r, s \in F$, we have, $r(sx) = (rs)x$
7. for $x \in V$ and $r, s \in F$, we have, $(r + s)x = rx + sx$
8. for $x, y \in V$, $r \in F$, we have, $r(x + y) = rx + ry$
9. for $x \in V$, we always have $1.x = x.1 = x$, where 1 is the multiplicative identity in V

Then, $(V, +, .)$ forms a vector space over the field of scalars F. A vector space always has at least one element, the zero vector. Hence the set $(\{0\}, +, .)$, forms a vector space, called the trivial space.

Example 11.1.
1. The set of real numbers $\mathbb{R}$ is a vector space over the field $\mathbb{R}$ with respect to addition and multiplication.
2. The set of n-tuples of real numbers, $\mathbb{R}^n$, where n is a positive integer, is a vector space over the field $\mathbb{R}$.
3. The set of complex numbers $\mathbb{C}$ forms a vector space over both the fields $\mathbb{R}$ and $\mathbb{C}$.
4. The set of all matrices form a vector space over the field $\mathbb{R}$ and $\mathbb{C}$.

In the study of any algebraic structure, it is of interest to examine subsets that possess the same structure as the set under consideration. We define the apt substructure of a vector space as follows:

Definition 11.2. A subset W of a vector space V over the field F, is said to be a subspace of V if W forms a vector space over the field F with respect to the operations of addition and multiplication defined over V.

In any vector space V, the zero set $\{0\}$ and V itself are always a vector subspace of V. These are called the trivial subspaces of V.

Example 11.3.
1. The subset $\{(x, y, z) \in \mathbb{R}^3 : x = y + z\}$ is a subspace of $\mathbb{R}^3$ while the subset $\{(x, y, z) \in \mathbb{R}^3 : x = y + z + 1\}$ is not a subspace of $\mathbb{R}^3$.
2. Check whether the subset $\{(x, y, z) \in \mathbb{R}^3 : xyz = 0\}$ is a subspace of $\mathbb{R}^3$.

Example 11.4. Let $u, v \in V$, where V is a vector space over a field F. Define the set W as

$$W = \{ru + sv : r, s \in F\}$$

Check that W forms a vector space over F. This set W is said to be *spanned* by the vectors u and v and is denoted by $W = L(u, v)$.

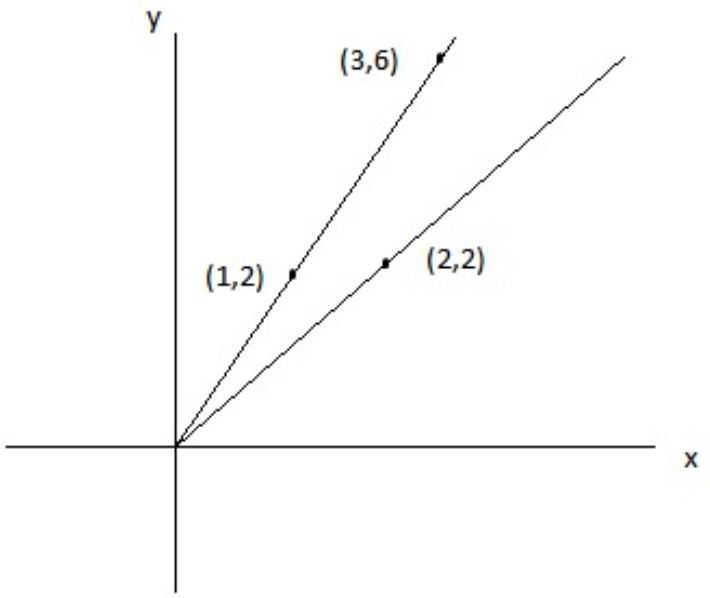


Figure 11.6.1: Representation of vectors in the xy-plane

In a similar way, the set

$$W = \{c_1v_1 + c_2v_2 + \cdots + c_nv_n : c_1, c_2, \ldots, c_n \in F\}$$

where $v_1, v_2, \ldots, v_n \in V$ forms a vector subspace of V.

Definition 11.5. Let V be a vector space and let $v_1, v_2, \ldots, v_n \in V$, then we define a *linear combination* of $v_1, v_2, \ldots, v_n \in V$ as

$$v = c_1v_1 + c_2v_2 + \cdots + c_nv_n$$

where $c_1, c_2, \ldots, c_n \in F$.

Definition 11.6. The subset W of V is called the *Linear Span* of $v_1, v_2, \ldots, v_n$ which is represented as $LS(v_1, v_2, \ldots, v_n)$.

Linear Dependence and Independence

Consider the vector space $\mathbb{R}^2$ and a vector say $(1, 2)$ in $\mathbb{R}^2$. Consider two other vectors $(3, 6)$ and $(2, 2)$. We see that the vector $(3, 6)$ can be written as $(3, 6) = 3(1, 2)$, that is, $(3, 6)$ lies on the line joining $(0, 0)$ and $(1, 2)$. But this is not the case with $(2, 2)$. Then $(1, 2)$ and $(3, 6)$ are said to be linearly dependent and $(1, 2)$ and $(2, 2)$ are said to be linearly independent. We formally define linear dependence and independence as follows:

Definition 11.7. The vectors $v_1, v_2, \ldots, v_n$ in a vector space V are said to be linearly independent if

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

can be satisfied only when $c_i = 0,\ \forall i$. If $v_1, v_2, \ldots, v_n$ are not linearly independent, then they are said to be linearly dependent. In other words, if $v_1, v_2, \ldots, v_n$ are linearly dependent, then any equation as

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

can be satisfied when at least one of c_i is non-zero.

For the above example, the given vectors can be written as

$$(3, 6) + (-3)(1, 2) = 0$$

We can also say that $(3, 6)$ is a linear combination of the set of vectors $\{(1, 2)\}$.

Now, let $\{v_1, v_2, \ldots, v_n\}$ are a set of linearly dependent vectors. Then for any equation

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

we must have at least one value of c, say $c_k \neq 0$. Then the above equation can be written as

$$\begin{aligned} c_1v_1 + c_2v_2 + \cdots + c_kv_k + \cdots + c_nv_n &= 0 \\ \text{or, } c_1v_1 + c_2v_2 + \cdots + c_nv_n &= -c_kv_k \\ \text{or, } -\frac{c_1}{c_k}v_1 - \frac{c_2}{c_k}v_2 - \cdots - \frac{c_n}{c_k}v_n &= v_k \end{aligned}$$

From the last line, we can say that v_k is a linear combination of the set of vectors $\{v_1, v_2, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}$. Thus, if you have a set of linearly dependent vectors, then you can remove the vector(s) that have a dependence and not change the possible things that the other vectors sum to. We thus have

$$\text{LS}\{v_1, v_2, \ldots, v_n\} = \text{LS}\{v_1, v_2, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n\}$$

Now, consider the vector space V. Then of course V spans itself. Then we choose an element say v which can be written as the linear combination of some other elements of V. Since deletion of a linearly dependent element does not affect the linear span, so we delete v from V. Again we select an element from $V \setminus \{v\}$ which can be written as the linear combination of some elements of $V \setminus \{v\}$. We continue the process and obtain a minimal set that spans the whole set V. Let the set be $\{v_1, v_2, \ldots\}$. Of course this set is linearly independent.

Now, consider an element v_1 of a vector space V. Then the linear span of v_1 is $\{av_1 : a \in F\}$ is of course a subspace of V. If $\text{LS}(v_1) = V$, then we stop here and say that v_1 spans the whole set V. If not, then we can find an element, say $v_2 \notin LS(v_1)$. Thus v_1, v_2 are linearly independent and again $LS(v_1, v_2)$ is a subspace of V. Again if $LS(v_1, v_2) = V$ then we say that V is spanned by $\{v_1, v_2\}$. If not, then we continue the process and get a linearly independent set $\{v_1, v_2, \ldots\}$ that spans V. Such a set is called the *basis* of V. We define the basis of a vector space V as

Definition 11.8. Let V be a vector space. Then a set of elements of V $\{v_1, v_2, \ldots\}$ is said to be a basis of V if

- it spans V, that is, every element of V can be written as a linear combination of the elements of the set
- it is linearly independent

Note that, a basis is the maximal linearly independent set and minimal spanning set of V. That is, if we add an element in the basis, then it no longer remains linearly independent. Also, if we exclude an element from the basis, then it no longer spans V. The number of elements in the basis of a vector space is called the **dimension** of V. If the basis of V contains infinite number of elements, then the dimension of V is infinite. Otherwise, the dimension of V is finite.

Some results related to the basis and dimension of a vector space

Let V be a vector space and let $S = \{v_1, v_2, \ldots, v_n\}$ be a basis of V. Then we have the following results:

1. Every element of V can be written uniquely as a linear combination of the elements of S.

   Proof. Let $v \in V$. Then there exists $c_1, c_2, \ldots, c_n$ such that

   $$v = c_1v_1 + c_2v_2 + \cdots + c_nv_n$$

   Again, let v has another representation as

   $$v = d_1v_1 + d_2v_2 + \cdots + d_nv_n$$

   Then, subtracting the two, we get

   $$0 = (c_1 - d_1)v_1 + (c_2 - d_2)v_2 + \cdots + (c_n - d_n)v_n$$

   Since $\{v_1, v_2, \ldots\}$ is linearly independent, so we will get, $c_i = d_i,\ \forall i$. Hence proved. □

   This result is also true for infinite dimensional vector spaces.

2. Any subset of V with more than n elements is dependent.

3. Any subset of V with fewer than n elements cannot span V.

4. Any linearly independent set having exactly $\dim V$ number of elements is a basis of V.

5. If W is a subspace of V such that $\dim V = \dim W$, then $V = W$.

Example 11.9.

1. $\mathbb{R}$ is a one-dimensional vector space over the field $\mathbb{R}$ with basis $\{1\}$. What happens if we replace the field by the set of all rational numbers $\mathbb{Q}$?

2. $\mathbb{R}^n$ is an n-dimensional vector space over the field $\mathbb{R}$ with basis $\{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \cdots, 1)\}$.

3. $\mathbb{C}$ is a one-dimensional vector space over the field $\mathbb{C}$ whose basis is $\{1\}$. It is also a 2-dimensional vector space if we replace the field by the field of real numbers $\mathbb{R}$. In that case, the basis is $\{1, i\}$.

4. The set of all polynomials of degree less than or equal to n is a vector space over the field $\mathbb{R}$. The basis in that case is $\{1, x, x^2, \ldots, x^n\}$.(verify!)

5. Consider the vector space $\mathbb{R}^2$. Is the set $\{(1, 1), (1, 0)\}$ linearly independent? Is it a basis of $\mathbb{R}^2$?

6. The matrices
   $$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$
   is a basis of the set of all 2×2 matrices over the field $\mathbb{R}$. Similarly, we can find the basis of the set of all $n \times n$ matrices over the field $\mathbb{R}$. (Find it)

7. Let A be an $m \times n$ matrix with real entries. Then the set spanned by the row vectors(or linearly independent row vectors) of A is called the row space of A. It is a subspace of $\mathbb{R}^n$. We can similarly define the column space of A which would be a subspace of $\mathbb{R}^m$.

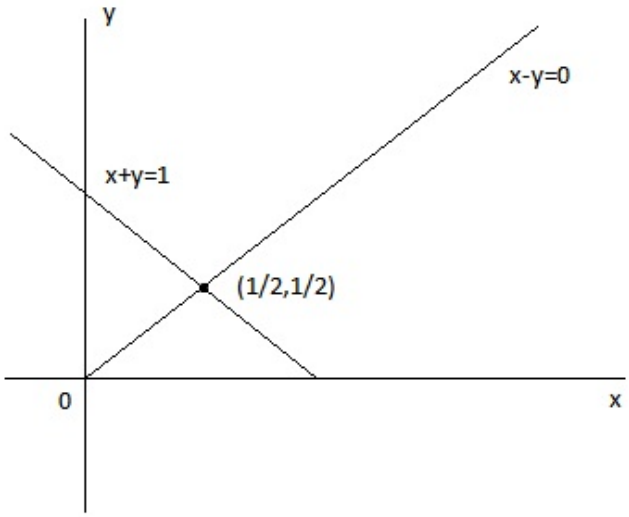


Figure 11.9.1: Geometrical Interpretation of the solution of the system of linear equations

System of Linear Equations

Consider the system of linear equations

$$\begin{aligned} x + y &= 1 \\ x - y &= 0 \end{aligned}$$

These two are a set of linear equations in $\mathbb{R}^2$. Any solution of these two equations are straight lines in the xy-plane. Solution of this system is the point where these two lines intersect. For the given system, $(1/2, 1/2)$ is the unique solution. Now, what happens if the second equation is replaced by $x + y = 4$? Then the equations represent two parallel straight lines which never intersects. Hence such system will have no solution. If instead, we have the second equation as $2x + 2y = 2$, then clearly, it is a linear combination of the first equation. Hence in such case the equation has infinitely many solutions.

General Form

The most general form of a system of linear equations is

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

which can be written in the matrix form as

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

or,

$$Ax = b$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \; x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

When the elements $b_i = 0, \; \forall \; i$, then the system is called homogeneous. Otherwise, it is called non-homogeneous system of linear equations. Notice that, $(0, 0, \dots, 0)$ is always a solution of the homogeneous system. This is called the trivial solution. Hence, if the solution set is S_h, then it is always non-empty. Now, if u, v are two solutions of the homogeneous system, then $cu + dv$ is always a solution of it, for real constants c, d (verify). Hence,the solution set S_h forms a vector space over the field $\mathbb{R}$. If $\{x_1, x_2, \dots, x_n\}$ is the basis of S_h, then the most general solution of the homogeneous system is $c_1x_1 + c_2x_2 + \dots + c_nx_n$, where c_i are real constants. If x_1, x_0 are solutions of the homogeneous and non-homogeneous systems respectively, then check that $x_1 + x_0$ is also a solution of the non-homogeneous system. Thus, the most general solution of the non-homogeneous system is $c_1x_1 + c_2x_2 + \dots + c_nx_n + x_0$, where c_i are real constants.

Solution of System of Linear Equations

A system of linear equations can be solved in many ways. One of the ways is by finding the inverse of the corresponding coefficient matrix, if it exists. The solution in such case, is given by

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

Illustration

Let a system of linear equation be given as

$$\begin{aligned} 2x + y &= 1 \\ 3x - y &= 0 \end{aligned}$$

The corresponding coefficient matrix is

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

Hence, the solution is given by

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} 2 & 1 \\ 3 & -1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= -5 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -5 \\ 0 \end{bmatrix} \end{aligned}$$

which is the required solution.

Note that the above process can only be used when the number of equations is equal to the number of unknowns, that is, the coefficient matrix is a square matrix.

The solution of the above system can also be found by elimination. For example, adding the two equations yield

$$\begin{aligned} 5x &= 1 \\ \text{or, } x &= \frac{1}{5} \end{aligned}$$

Putting the value of x in the second equation yields

$$y = \frac{3}{5}$$

Notice that the solutions obtained by the two methods are different. In fact, it is not mandatory for system of linear equations to have unique solution. They can have more than one solution by virtue of the description given in the previous section.

Cramer's Rule

Cramer's Rule is an explicit way of finding the solution of a system of linear equations whose number of equations and number of unknowns are same. Instead of solving the entire system of equations, you can use Cramer's to solve for just one single variable. Let's illustrate it with an example.

Illustration

Let the following be a linear system:

$$\begin{aligned} 2x + y + z &= 3 \\ x - y - z &= 0 \\ x + 2y + z &= 0 \end{aligned}$$

First we find the determinant of the coefficient matrix which is

$$D = \begin{vmatrix} 2 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 1 \end{vmatrix} = 3 \text{ (verify)}$$

Define D_1, D_2, D_3 as the determinant of the coefficient matrix in which the first, second and third columns respectively are replaced by the matrix b. That is,

$$D_1 = \begin{vmatrix} 3 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 2 & 1 \end{vmatrix} = 3,\ D_2 = \begin{vmatrix} 2 & 3 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{vmatrix} = -6,\ D_3 = \begin{vmatrix} 2 & 1 & 3 \\ 1 & -1 & 0 \\ 1 & 2 & 0 \end{vmatrix} = 9$$

The solutions are given by

$$\begin{aligned} x &= \frac{D_1}{D} = 1 \\ y &= \frac{d_2}{D} = -2 \\ z &= \frac{D_3}{D} = 3 \end{aligned}$$

There are other ways to find the solution of linear systems in which the number of equations and the number of unknowns are not same. But in this unit, we restrict ourselves to the linear system having same number of equations as the number of unknowns.

Exercise 11.10. Find the solution of the following systems:

a.

$$\begin{aligned} 2x + y + z &= 1 \\ x - y + 4z &= 0 \\ x + 2y - 2z &= 3 \end{aligned}$$

b.

$$\begin{aligned} 3x - y &= 7 \\ -5x + 4y &= -2 \end{aligned}$$

c.

$$\begin{aligned} -6x + 8y &= 17 \\ 13x - 2y &= -4 \end{aligned}$$

d.

$$\begin{aligned} 2x - y + 6z &= 10 \\ -3x + 4y - 5z &= 11 \\ 8x - 7y - 9z &= 12 \end{aligned}$$

e.

$$\begin{aligned} 2x - 3y &= 4 \\ -x + 4y - z &= 11 \\ 6x - 5y + 2z &= -3 \end{aligned}$$

Unit 12

Course Structure

- Introduction
- Objectives
- Eigen values and Eigen vectors of matrices
- Cayley Hamilton Theorem
- Diagonalization of matrices

Introduction

Consider the matrix

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

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. We are now in a position to formally define eigen values and eigen vectors of a matrix.

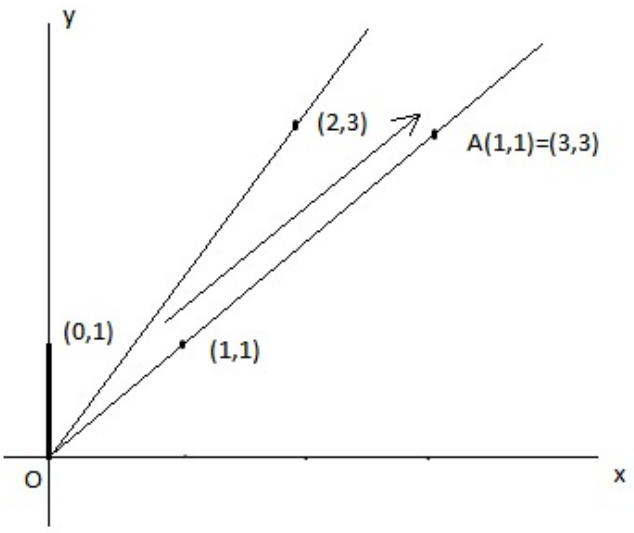


Figure 12.0.1: Eigen Values and Eigen Vectors Geometrically

Objectives

After reading this unit, you will be able to:

- find the eigen values and eigen vectors of a matrix
- learn the Cayley Hamilton theorem
- find the characteristic polynomial of a matrix
- find the eigen values from the characteristic polynomial
- define diagonalizability of a matrix
- find out when a matrix will be called diagonalizable

Eigen values and Eigen vectors of a matrix

Let A be an $n \times n$ matrix. The number c is called an eigen value of A if there exists a non-zero vector v such that

$$Av = cv$$

In such a case v is called the eigen vector corresponding to the eigen value c.

Computation of Eigen values and eigen vectors

The equation

$$Av = cv$$

can be re-written as

$$(A - cI)v = 0$$

Now, for non-trivial solution of the above equation, $A - cI$ must not be invertible, because in that case, we would have

$$\begin{aligned}(A-cI)^{-1}(A-cI)v &= (A-cI)^{-1}0 \\ \text{or, } v &= 0\end{aligned}$$

So $\det(A-cI) = 0$. We call $f(c) = \det(A-cI)$ as the characteristic polynomial of A. Solving this for c, we get the eigen values. Hence, the roots of the characteristic polynomial of A gives the eigen values of A. Also the power to which the factor of any eigen value is raised, is called its algebraic multiplicity, that is, if λ is an eigen value of A, and d is the highest power to which $(x-\lambda)$ is raised, then d is the algebraic multiplicity of λ.

To find the eigen vectors of A, we need to simply solve the system of linear equations given by

$$(A-cI)v = 0$$

Let's illustrate it by an example.

Illustration

Let the given matrix be

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

Let c be an eigen value of A and (v_1, v_2) be the corresponding eigen vector. Then

$$(A-cI)v = 0$$

will give non-trivial solution if $\det(A-cI) = 0$. Solving this for c will give us all the eigen values. So,

$$\begin{aligned}\begin{vmatrix}1-c & 2\\ 2 & 4-c\end{vmatrix} &= 0 \\ \text{or, } (1-c)(4-c)-4 &= 0 \\ \text{or, } c(c-5) &= 0 \\ \text{or } c &= 0, 5\end{aligned}$$

Hence, 5 and 0 are the eigen values of A. Now, we will find the corresponding eigen vectors.
For $c = 0$,

$$\begin{aligned}\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} &= 0\begin{bmatrix}v_1\\ v_2\end{bmatrix} \\ &= \begin{bmatrix}0\\ 0\end{bmatrix}\end{aligned}$$

This gives us the homogeneous system of linear equations

$$\begin{aligned}v_1 + 2v_2 &= 0 \\ 2v_1 + 4v_2 &= 0\end{aligned}$$

Solving, we get $v_1 = -2v_2$. If we take $v_2 = k$, then the eigen vector is given by $(-2k, k) = k(-2, 1)$, where k is any real constant. Thus any element of the subspace $S = \{k(-2,1) : k \in \mathbb{R}\}$ is an eigen vector of A corresponding to the eigen value 0.Taking $k = 1$ we get a particular eigen vector $(-2, 1)$ corresponding

to the eigen value 0. We can also say that $(-2, 1)$ spans the set S. This set is called the *eigen space* of A corresponding to the eigen value 0.
For $c = 5$, we will have

$$\begin{aligned}\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}\begin{bmatrix}v_1\\ v_2\end{bmatrix} &= 5\begin{bmatrix}v_1\\ v_2\end{bmatrix}\\ &= \begin{bmatrix}5v_1\\ 5v_2\end{bmatrix}\end{aligned}$$

which gives us the homogeneous system

$$\begin{aligned}-4v_1 + 2v_2 &= 0\\ 2v_1 - v_2 &= 0\end{aligned}$$

Solving, we get $2v_1 = v_2$. If we take $v_1 = k$, then the eigen vectors will be given by $(k, 2k) = k(1, 2)$ for any real constant k. Thus the eigen space, say T corresponding to the eigen value 5 is given by $\{k(1, 2) : k \in \mathbb{R}\}$. Putting $k = 1$, a particular eigen vector corresponding to the eigen value 5 is $(1, 2)$.

Remark: The dimension of the eigen space corresponding to the eigen value c of a matrix A is called its geometric multiplicity.

Exercise 12.1. Find the eigen values and eigen vectors of the following matrices:

$$\begin{bmatrix}2 & 3\\ 2 & 1\end{bmatrix}, \begin{bmatrix}0 & 1\\ -2 & -3\end{bmatrix}, \begin{bmatrix}3 & 6 & -8\\ 0 & 0 & 6\\ 0 & 0 & 2\end{bmatrix}$$

Theorem 12.2. If $v_1, v_2, \ldots, v_k$ are eigen vectors of a matrix A, corresponding to its distinct eigen values $c_1, c_2, \ldots, c_k$, then the set $\{v_1, v_2, \ldots, v_k\}$ is linearly independent.

Proof. Let us consider the equation

$$r_1v_1 + r_2v_2 + \cdots + r_kv_k = 0$$

for real constants $r_1, r_2, \ldots, r_k$. We have to show that these constants are zero. For any positive $i \leq k$, if we operate the matrix $(A - c_iI)$ on both sides of the equation □

Theorem 12.3. The determinant of an $n \times n$ matrix is 0 if and only if 0 is an eigen value of A.

Proof. Let 0 be an eigen value of A. Then

$$(A - oI)v = 0$$

which yields $\det A = 0$ for non-trivial solution. Hence we are done.
Conversely, let $\det A = 0$. This implies that

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

which implies that the equation

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

has non-trivial solution. Hence, 0 is an eigen value of A. □

Remark: Note that the zero vector can never be an eigen vector of any matrix.

Definition 12.4. Two $n \times n$ matrices A and B are said to be similar if there exists an invertible matrix P such that

$$A = PBP^{-1}$$

Two similar matrices always have the same characteristic polynomial. So they have the same eigen values.

Diagonalizability

An $n \times n$ matrix A is said to be diagonalizable if it has n linearly independent eigen vectors. If A has n distinct eigen values then the corresponding eigen vectors are linearly independent. So, the existence of n distinct eigen values guarantees the diagonalizability of A. If A does not have distinct eigen values then we check the dimension of the eigen spaces. For example, if λ is a repeated root of the characteristic polynomial of A, then we consider the eigen space corresponding to λ. If the geometric multiplicity of every eigen value of A is equal to its algebraic multiplicity, then A is diagonalizable. Let's illustrate it with an example.

Illustration

Let us consider the matrix

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

First we find the eigen values of A by

$$\det(A - cI) = 0$$

which gives the characteristic polynomial A as below,

$$c^2 = 0$$

and hence, the only eigen value of A is 0 with algebraic multiplicity 2. We will consider the eigen space of 0. Let us find the eigen vector of A corresponding to 0. For that, consider the system of equations

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

which gives $v_2 = 0$ and v_1, undetermined. So the eigen space is the set $\{k(1, 0) : k \in \mathbb{R}\}$. Thus the eigen space is spanned by only one element and hence, the geometric multiplicity of 0 is $1 < 2$, the algebraic multiplicity of 0. Hence, A is not diagonalizable.

Exercise 12.5. Show that the following matrices are not diagonalizable:

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

Let us have another example of a matrix which is diagonalizable. Consider the matrix

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

We find the eigen values of A by the equation

$$\det(A - cI) = 0$$

which gives us the characteristic polynomial of A as

$$(2 - c)^2(1 - c) = 0$$

Thus, the eigen values of A are 1 and 2 of algebraic multiplicities 1 and 2 respectively. Next we find the corresponding eigen spaces.

For $c = 1$, let $v = (v_1, v_2, v_3)$ be the eigen vector of A. Then we solve the system of linear equations as

$$(A - 1I)v = 0$$

which gives

$$\begin{aligned} 2v_1 + 0v_2 + 0v_3 &= 0 \\ v_1 + v_2 + v_3 &= 0 \\ -v_1 + 0v_2 + 0v_3 &= 0 \end{aligned}$$

Solving, we get $v_1 = 0, v_3 = -v_2$. Thus, if we take $v_2 = k$, for some real constant k, then the eigen space will be $\{k(0, 1, -1) : k \in \mathbb{R}\}$. Thus, the eigen space is spanned by $(0, 1, -1)$, and hence the geometric multiplicity of 1 is 1.

For $c = 2$, let $v = (v_1, v_2, v_3)$ be the eigen vector of A. Then we solve the system of linear equations as

$$(A - 2I)v = 0$$

which gives

$$\begin{aligned} 0v_1 + 0v_2 + 0v_3 &= 0 \\ v_1 + 0v_2 + v_3 &= 0 \\ -v_1 + 0v_2 - v_3 &= 0 \end{aligned}$$

solving, we get $v_3 = -v_1$ and v_2 remains undetermined. So, if we take $v_1 = k$ and $v_2 = l$, then we get the eigen space as $\{(k, l, -k) : k, l \in \mathbb{R}\} = \{k(1, 0, -1) + l(0, 1, 0) : k, l \in \mathbb{R}\}$. Hence, the eigen space corresponding to 2 is spanned by the vectors $(1, 0, -1)$ and $(0, 1, 0)$, hence the geometric multiplicity of 2 is 2.

Exercise 12.6. Are the following matrices diagonalizable?

$$\begin{bmatrix} 2 & 4 & 6 \\ 0 & 2 & 2 \\ 0 & 0 & 4 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 & 0 \\ 2 & 6 & 0 \\ 3 & 2 & 1 \end{bmatrix}$$

Cayley Hamilton Theorem

Cayley Hamilton theorem is an important tool to find the inverse and powers of a matrix in an easier way. We will learn to do a few of such applications. First let us state the theorem as follows:

Theorem 12.7. Let A be an $n \times n$ matrix and $f(t)$ be its characteristic polynomial. Then $f(A)$ is the $n \times n$ zero matrix.

Let us verify the Cayley-Hamilton theorem for the matrix

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

First find the characteristic polynomial of A. Verify that, it is $c^2 - 5c + 6 = 0$. Now we will find A^2.

$$\begin{aligned} A^2 &= \begin{bmatrix} 6 & -2 \\ 6 & -1 \end{bmatrix} \begin{bmatrix} 6 & -2 \\ 6 & -1 \end{bmatrix} \\ &= \begin{bmatrix} 24 & -10 \\ 30 & -11 \end{bmatrix} \end{aligned}$$

So,

$$\begin{aligned} A^2 - 5A + 6I &= \begin{bmatrix} 24 & -10 \\ 30 & -11 \end{bmatrix} - 5\begin{bmatrix} 6 & -2 \\ 6 & -1 \end{bmatrix} + 6\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 24 & -10 \\ 30 & -11 \end{bmatrix} - \begin{bmatrix} 30 & -10 \\ 30 & -5 \end{bmatrix} + \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \end{aligned}$$

Hence, A satisfies its characteristic polynomial.

Exercise 12.8. Verify Cayley-Hamilton theorem for the following matrices:

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

Applications of the Cayley-Hamilton theorem

Example 12.9. Let us find the inverse of

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

The characteristic equation of A is $f(t) = t^2 - 4t + 3$. Then by Cayley-Hamilton theorem,

$$A^2 - 4A + 3I = 0$$

Multiplying both sides by A^{-1}, we get

$$A^{-1}(A^2 - 4A + 3I) = A^{-1}.0$$

which gives

$$\begin{aligned} A - 4I + 3A^{-1} &= 0 \\ A^{-1} &= \frac{1}{3}(4I - A) \\ &= \frac{1}{3}\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} - \frac{1}{3}\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -2/3 \\ 0 & 1/3 \end{bmatrix} \end{aligned}$$

Exercise 12.10. a. Calculate and simplify the expression

$$-T^3 + 4T^2 + 5T - 2I$$

where, I is the 3×3 identity matrix and

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

b. Find the inverse of the matrix A using Cayley-Hamilton theorem, where

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 9 & 2 & 0 \\ 5 & 0 & 3 \end{bmatrix}$$

c. Find the inverse of the matrix B using Cayley-Hamilton theorem, where

$$B = \begin{bmatrix} 7 & 2 & -2 \\ -6 & -1 & 2 \\ 6 & 2 & -1 \end{bmatrix}$$

Unit 13

Course Structure

- Linearization of a dynmaical system
- Minimum Variance Unbiased Estimator
- Method of Maximum Likelihood for Estimation of a parameter

13.1 Linearization of a dynamical system

In mathematics, linearization is finding the linear approximation to a function at a given point. The linear approximation of a function is the first order Taylor expansion around the point of interest. In the study of dynamical systems, linearization is a method for assessing the local stability of an equilibrium point of a system of nonlinear differential equations or discrete dynamical systems. This method is used in fields such as engineering, physics, economics, and ecology.

A two dimensional dynamical system may be written as $\dot{x} = f(x)$ where $x = (x_1, x_2)$ and $f(x) = (f(x_1), f(x_2))$.

Existence and Uniqueness Theorem: Consider the initial value problem $\dot{x} = f(x),\ x(0) = x_0$. Suppose that f is continuous and that all its partial derivatives $\dfrac{\partial f_i}{\partial x_j},\ i, j = 1, \cdot, n$ are continuous for x in some open connected set $D \subset \mathbb{R}^n$. Then for $x_0 \in D$, the initial value problem has a solution $x(t)$ on some time interval $(-\tau, \tau)$ about $t = 0$, and the solution is unique.

Corollary: Different trajectories never intersect.

In this section, we first discuss the linearization technique for two dimensional dynamical system. Consider the system

$$\dot{x} = f(x, y)$$
$$\dot{y} = g(x, y)$$

and suppose that (x^*, y^*) is the fixed point, i.e.,

$$f(x^*, y^*) = 0 \quad \text{and} \quad g(x^*, y^*) = 0.$$

Let $u = x - x^*$, $v = y - y^*$ denote the components of a small disturbance from the fixed point. To see whether the disturbance grows or decays, we need to derive differential equations for u and v. Let us do u-equation first.

We have $u = x - x^*$. Differentiating with respect to time t,

$$\begin{aligned}
\dot{u} &= \dot{x} \quad (\text{since } x^* \text{ is a constant}) \\
&= f(x^* + u, y^* + v) \quad (\text{By substitution}) \\
&= f(x^*, y^*) + u\frac{\partial f}{\partial x} + v\frac{\partial f}{\partial y} + O(u^2, v^2, uv) \quad (\text{Expanding in Taylor series}) \\
&= u\frac{\partial f}{\partial x} + v\frac{\partial f}{\partial y} + O(u^2, v^2, uv) \quad (\text{Since } f(x^*, y^*) = 0).
\end{aligned}$$

To simplify the notation, we have written $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, but remember these partial derivatives are to be evaluated at the fixed point (x^*, y^*), thus they are numbers, not functions. Also the shorthand notation $O(u^2, v^2, uv)$ denotes quadratic terms in u and v. Since u and v are small, these quadratic terms are extremely small.

Similarly, we find

$$\dot{v} = \frac{\partial g}{\partial x} + v\frac{\partial g}{\partial y} + O(u^2, v^2, uv).$$

Hence the disturbance (u, v) evolves according to

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} + \text{Quadratic terms} \tag{13.1.1}$$

The matrix $A = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix}_{x^*, y^*}$ is called the Jacobian matrix at the fixed point (x^*, y^*). Now since the quadratic terms in Eq. (13.1.1) are tiny, it is tempting to neglect them. If we do that, we obtain the linearized system

$$\begin{bmatrix} \dot{u} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

whose dynamics can be analysed as before.

Effect of small nonlinear terms:

Is it really safe to neglect the quadratic terms? In other words, does the linearized system give a qualitatively correct picture near (x^*, y^*)?

The answer is yes, as long as the fixed point for the linearized system is not of the borderline case (centers, degenerate nodes, stars or non-isolated fixed pints). In other words, if the linearized system predicts a saddle, node, or spiral for the original nonlinear equations.

Find all the fixed points of the system

$$\begin{aligned} \dot{x} &= -x + x^3 \\ \dot{y} &= -2y \end{aligned}$$

and use linearization to classify them.

Solution: Fixed points occurs where $\dot{x} = 0$ and $\dot{y} = 0$ simultaneously, which give us $x = 0$ or $x = \pm 1$ and $y = 0$.

Thus there are three fixed points, viz $(0, 0),\ (1, 0)$, and $(-1, 0)$. The Jacobian matrix at the general point (x, y) is

$$A = \begin{bmatrix} \frac{\partial}{\partial x}(-x + x^3) & \frac{\partial}{\partial y}(-x + x^3) \\ \frac{\partial}{\partial x}(-2y) & \frac{\partial}{\partial y}(-2y) \end{bmatrix} = \begin{bmatrix} -1 + 3x^2 & 0 \\ 0 & -2 \end{bmatrix}$$

Next we evaluate A at the fixed points.

At the point $(0, 0)$, we find $A = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$ which gives two negative eigenvalues, viz $\lambda_1 = -1$ and $\lambda_2 = -2$. Therefore, the fixed point $(0, 0)$ is a stable node.

At $(\pm 1, 0)$, $A = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$, which gives two eigenvalues of opposite sign. So both the fixed points $(1, 0)$ and $(-1, 0)$ are saddle point.

Now since stable nodes and saddle points are not borderline cases, it is certain that the fixed points for the given nonlinear system has been predicted correctly.

Consider the system

$$\begin{aligned} \dot{x} &= -y + ax(x^2 + y^2) \\ \dot{y} &= x + ay(x^2 + y^2) \end{aligned}$$

where a is a parameter. Show that the linearized system incorrectly predicts that the origin is a center for all values of a, whereas in fact the origin is a stable spiral if $a < 0$ and unstable spiral if $a > 0$.

Solution: To obtain the linarization about the origin, i.e. about $(x^*, y^*) = (0, 0)$, we can either compute the Jacobain matrix directly form the definition, or we can take the following shortcut.

For any system with a fixed point at the origin, x and y represent deviations from the fixed point, since $u = x - x^* = x$ and $v = y - y^* = y$; hence we can linearize by simply omitting the nonlinear terms in x and y. Thus the linearized system is given by

$$\begin{aligned} \dot{x} &= -y \\ \dot{y} &= x \end{aligned}$$

The Jacobain at the fixed point $(0, 0)$ is $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ which has $\tau = 0, \Delta = 1 > 0$, so the origin is always a center.

To analyze the nonlinear system, we change variables to polar coordinates. Let

$$x = r\cos\theta$$
$$y = r\sin\theta.$$

To derive a differential equation for r, we note $x^2 + y^2 = r^2$, so on differentiation we obtain

$$\begin{aligned} & x\dot{x} + y\dot{y} = r\dot{r} \\ \Rightarrow\ & r\dot{r} = x\{-y + ax(x^2+y^2)\} + y\{x + ay(x^2+y^2\} \\ \Rightarrow\ & r\dot{r} = a(x^2+y^2)^2 \\ \Rightarrow\ & r\dot{r} = ar^4 \\ \Rightarrow\ & \dot{r} = ar^3 \end{aligned}$$

Now since $\theta = \tan^{-1}\left(\frac{y}{x}\right)$, we have

$$\begin{aligned} & \dot{\theta} = \frac{1}{1+\frac{y^2}{x^2}}\left[\frac{x\dot{y} - y\dot{x}}{x^2}\right] = \frac{x\dot{y} - y\dot{x}}{x^2+y^2} \\ \Rightarrow\ & \dot{\theta} = \frac{1}{r^2}[x\{-y + ax(x^2+y^2)\} - y\{x + ay(x^2+y^2)\}] \\ \Rightarrow\ & \dot{\theta} = \frac{x^2+y^2}{r^2} = \frac{r^2}{r^2} \\ \Rightarrow\ & \dot{\theta} = 1 \end{aligned}$$

Thus in polar coordinates the original system becomes

$$\begin{aligned} \dot{r} &= ar^3 \\ \dot{\theta} &= 1 \end{aligned}$$

The system is easy to analyse in this form, because the radial and angular motions are independent. All trajectories rotate about the origin with constant angular velocity $\dot{\theta} = 1$.

If $a < 0$, then $r(t) \to 0$ monotonically as $t \to \infty$. In this case, the origin is a stable spiral.

If $a = 0$, then $r(t) = r_0$ for all t and the origin is a center.

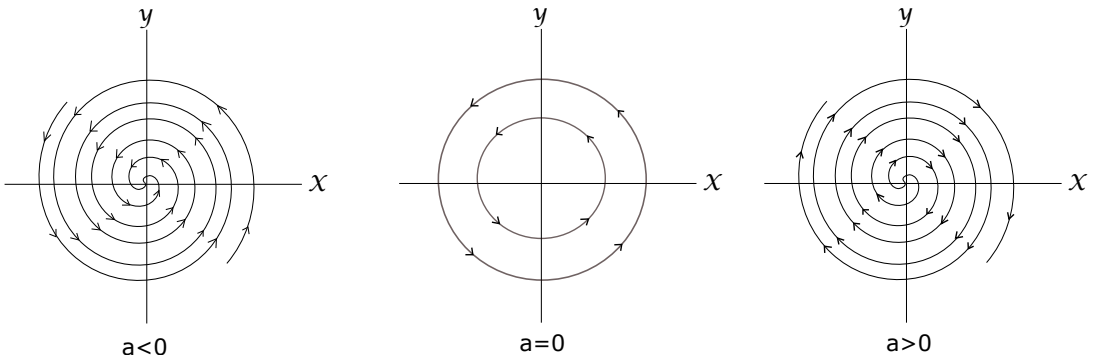
Finally if $a < 0$, then then $r(t) \to \infty$ monotonically and the origin is a unstable spiral.

13.1.1 A general interaction model for two population

In order to explain mathematical modelling with systems of differential equations, we investigate the following general two species interaction model:

$$\begin{aligned} \dot{x} &= \alpha x + \beta xy \\ \dot{y} &= \gamma y + \delta xy \end{aligned} \tag{13.1.2}$$

where $x(t)$ and $y(t)$ denote the concentration (or number) of two populations and $\alpha, \beta, \gamma, \delta$ are constant real numbers.



The linear terms αx and γy describe the growth or decay of the corresponding population x and y in isolation. For example, if $\alpha > 0$ and $\beta = 0$, the population x will grow like $e^{\alpha t}$; if $\alpha < 0$, it will decay exponentially. Similarly, if $\delta = 0$, then the sign of γ decides whether $y(t)$ is exponentially growing or decaying.

We begin by writing (13.1.2) in vector notation:

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f_1(x,y) \\ f_2(x,y) \end{bmatrix}$$

with $f_1(x,y) = \alpha x + \beta xy$ and $f_2(x,y) = \gamma x + \delta xy$. To find the x-nullclines, say η_x, we set $f_1(x,y) = 0$. Hence, the x-nullclines are $\eta_x = \left\{(x,y) : x = 0 \text{ or } y = -\frac{\alpha}{\beta}\right\}$. Similary, the y-nullclines are $\eta_y = \left\{(x,y) : y = 0 \text{ or } x = -\frac{\gamma}{\delta}\right\}$. The steady stated (x^*, y^*) are intersection points of the nullclines and they satisfy $f_1(x^*, y^*) = 0$ and $f_2(x^*, y^*) = 0$. We have two steady states, namely,

$$P_1 = (0,0) \quad \text{and} \quad P_2 = \left(-\frac{\gamma}{\delta}, -\frac{\alpha}{\beta}\right).$$

The linearization of the given system (13.1.2) is given by

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = Df(x^*, y^*)\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

where $Df(x^*, y^*) = \begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{bmatrix}_{(x^*,y^*)} = \begin{bmatrix} \alpha + \beta y & \beta x \\ \delta y & \gamma + \delta x \end{bmatrix}_{(x^*,y^*)}$. We evaluate this matrix at the two steady states, P_1 and P_2. For P_1, we find

$$Df(0,0) = \begin{bmatrix} \alpha & 0 \\ 0 & \gamma \end{bmatrix}$$

which has two eigenvalues $\lambda_1 = \alpha$ and $\lambda_2 = \gamma$. Similarly, for P_2, we find

$$Df\left(-\frac{\gamma}{\delta}, -\frac{\alpha}{\beta}\right) = \begin{bmatrix} 0 & -\frac{\beta\gamma}{\delta} \\ -\frac{\alpha\delta}{\beta} & 0 \end{bmatrix} = A, \quad \text{say.}$$

Since $\text{trace}(A) = 0$ and $\det(A) = -\alpha\gamma$, hence the eigenvalues are $\lambda_{1,2} = \pm\sqrt{\alpha\gamma}$. To identify the type of steady states, we nee to have more information. In particular, we need to know the signs of the parameters

$\alpha,\ \beta,\ \gamma,$ and δ. Analysis of three specific cases follows:

Case I: *A prey-predator model:*

We assume that $\alpha < 0,\ \beta > 0,\ \gamma > 0$ and $\delta < 0$. Hence, we see that one eigenvalue is negative $(\lambda_1 = \alpha < 0)$ and the other eigenvalue is positive $(\lambda_2 = \gamma > 0)$. Hence $P_1(0,0)$ is a saddle point. Before we study $P_2 = \left(-\frac{\gamma}{\delta}, -\frac{\alpha}{\beta}\right)$ we have to ensure that it is biologically relevant, i.e., $-\frac{\gamma}{\delta}, -\frac{\alpha}{\beta}$ both are positive. The product $\alpha\gamma < 0$, so that the eigenvalues are purely imaginary, namely $\lambda_{1,2} = \pm i\sqrt{|\alpha\gamma|}$. Hence the critical point $\left(-\frac{\gamma}{\delta}, -\frac{\alpha}{\beta}\right)$ is a center. Thus P_2 is not hyperbolic and the Hartman-Grobman theorem can not be applied.

Case II: *Mutualism of two species:*

We assume two species which cannot survive alone. For this $\alpha < 0$ and $\gamma < 0$. The eigenvalues of $Df(0,0)$ are $\alpha < 0$ and $\gamma < 0$. Hence $(0,0)$ is a stable node. Also, $-\frac{\alpha}{\beta} > 0$ and $-\frac{\gamma}{\delta} > 0$ and hence P_2 is biologically relevant. The product $\alpha\gamma > 0$. Hence the eigenvalues are $\lambda_{1,2} = \pm\sqrt{\alpha\gamma}$. Therefore, P_2 is a saddle point.

Case III: *A competition model:*

In this case, we assume that $\alpha > 0$ and $\beta < 0$, thus the critial point $(0,0)$ is a saddle point. But P_2 is not bilogically relevant because $-\frac{\gamma}{\delta} < 0$. Thus the population y goes extinct while population x can grow without competition.

A basic epidemic Model: We consider the spread of an infectious disease in a host population. Let $S,\ I$ and R denote the number of susceptible, infectious, and recovered individuals respectively.

If the disease is transmitted through direct contact, then the rate of new incidences, βIS, is in proportion to the number of susceptible and to the number of infectious individuals. With these assumptions, the disease process is descrbed by the following classical SIR (Susceptibles-Infected-Recovered) model which is given by

$$\begin{aligned} \dot{S} &= -\beta IS + \gamma R \\ \dot{I} &= \beta IS - \alpha I \\ \dot{R} &= \alpha I - \gamma R \end{aligned} \tag{13.1.3}$$

For simplicity, we assume $\gamma = 0$. This can be understood as assuming the mean immune period $\frac{1}{\gamma} \to \infty$; the disease incurs permanent immunity. The simplified model is known as the *Kermack-Mckendric model* which is given by

$$\begin{aligned} \dot{S} &= -\beta IS \\ \dot{I} &= \beta IS - \alpha I \end{aligned} \tag{13.1.4}$$

Qualitative Analysis of the epidemic model:

Let us analyse the epidemic model given in (13.1.4). To find the steady states, we set $\dot{S} = 0$ and $\dot{I} = 0$.

If $\dot{S} = 0$, then eigher $S = 0$ or $I = 0$ and if $\dot{I} = 0$, then either $I = 0$ or $S = \alpha/\beta$. Therefore, the system (13.1.4) has a ray of steady states along the positive S-axis, $\{(S,0) :\ S > 0\}$.

To find the stability of each steady state $(\overline{S}, 0)$, we examine the Jacobian matrix,

$$\begin{bmatrix} -\beta I & -\beta S \\ \beta I & \beta S - \alpha \end{bmatrix}_{S=\overline{S}, I=0} = \begin{bmatrix} 0 & -\beta\overline{S} \\ 0 & \beta\overline{S} - \alpha \end{bmatrix}$$

The two eigenvalues of this Jacobian matrix are $\lambda_1 = 0$ and $\lambda_2 = \beta\overline{S} - \alpha$. The eigenvalue $\lambda_1 = 0$ corresponds to the neutrally stable direction along the ary of steady states. The second eigenvalue $\lambda_2 = \beta\overline{S} - \alpha$ is positive if $\overline{S} > \dfrac{\alpha}{\beta}$ and negative if $\overline{S} < \dfrac{\alpha}{\beta}$.

To construct the phase portrait, we write one unknown, I, as a function of the other, S. This way, we still follow the trajectory of an epidemic, but we forget about the time course for a moment. To achieve this, we use the chain rule. In particular if $I = I(S(t))$, then

$$\frac{dI}{dt} = \frac{dI}{dS} \cdot \frac{dS}{dt}.$$

Hence,

$$\frac{dI}{dS} = \frac{\dot{I}}{\dot{S}} = \frac{\beta IS - \alpha I}{-\beta IS} = -1 + \frac{\alpha}{\beta S}.$$

If we regard I as a function of S, and integrate the above equation from S_0 to S, then we obtain

$$I(S) - I(S_0) = -(S - S_0) + \frac{\alpha}{\beta}(\ln S - \ln S_0)$$

$$\Rightarrow I(S) = \frac{\alpha}{\beta}\ln S - S + c_1$$

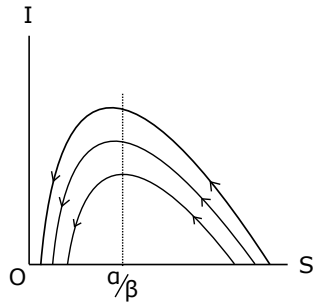


Figure 13.1.1

where the constant c_1 is determined by the initial condition $S(t) = S_0$, $I(t) = I_0$ at $t = 0$, so that $c_1 = I(S_0) + S_0 - \frac{\alpha}{\beta}\ln S_0$.

As shown in Fig. 13.1.1 the steady state to the right of $\dfrac{\alpha}{\beta}$, namely, $\overline{S} > \dfrac{\alpha}{\beta}$ are unstable in the direction away from the S-axis, and those to the left of $\dfrac{\alpha}{\beta}$ are stable.

Biologically, $\dfrac{\alpha}{\beta}$ represents the critical population size to sustain an epidemic. If the initial susceptible population is below $\dfrac{\alpha}{\beta}$, then no epidemic is possible and the number of infections decreases, whereas if $S_0 > \dfrac{\alpha}{\beta}$, then the number of infection initially increases, reaching its maximum when $S = \dfrac{\alpha}{\beta}$ and then declines.

Unit 14

Course Structure

- Stability and Liapunov Functions

14.1 Stability and Liapunov Functions

Here we discuss the stability of the equilibrium points of the non-linear system

$$\dot{x} = f(x). \tag{14.1.1}$$

The stability of any hyperbolic equilibrium point x_0 of (14.1.1) is determined by the sign of real parts of the eigen values λ_j of the matrix $Df(x_0)$. A hyperbolic equilibrium point x_0 is asymptotically stable if and only if $\text{Re}(\lambda_j) < 0$ for $j = 1, \ldots, n$, while it is unstable if and only if it is saddle or $\text{Re}(\lambda_j) > 0$ for $j = 1, \ldots, n$. The stability of non-hyperbolic equilibrium points is typically more difficult to determine. A method due to Liapunov, that is very useful for deciding the stability of non-hyperbolic equilibrium points. Consider the non-linear autonomous systen

$$\frac{dx}{dt} = P(x, y) \tag{14.1.2}$$

$$\frac{dy}{dt} = Q(x, y). \tag{14.1.3}$$

Assume that this system has an isolated critical point at the origin $(0, 0)$ and that P and Q have continuous first order partial derivatives for all (x, y). Let $E(x, y)$ be positive definite for all (x, y) in a domain D containing the origin and such that the derivative $\dot{E}(x, y)$ of E with respect to the above system is negative semi-definite for all $(x, y) \in D$. Then E is called a Liapunov function for the system in D.

Show that $E(x, y) = x^2 + y^2$ is a Liapunov function for the non-linear system

$$\frac{dx}{dt} = -x + y^2$$

$$\frac{dy}{dt} = -y + x^2.$$

Solution. Here the critical point is given by $(0,0)$. Now,

$$\begin{aligned}\dot{E} &= \frac{\partial E}{\partial x}\frac{dx}{dt} + \frac{\partial E}{\partial y}\frac{dy}{dt} \\ &= 2x(-x+y^2) + 2y(-y+x^2) \\ &= -2x^2 + 2xy^2 - 2y^2 + 2x^2y \\ &= -2(x^2+y^2) + 2(x^2y + xy^2).\end{aligned}$$

Here, $E(0,0) = 0$ and $E(x,y) = x^2 + y^2 > 0$ for all $x,\ y \neq 0$. Hence $E(x,y)$ is positive definite in any domain D containing the origin $(0,0)$. Now clearly $\dot{E}(0,0) = 0$ and if $x < 1$ and $y \neq 0$, then $xy^2 < y^2$. Also, if $y < 1$ and $x \neq 0$, then $x^2y < x^2$. Thus, if $x < 1$, $y < 1$ and $(x,y) \neq (0,0)$, then

$$x^2y + xy^2 < x^2 + y^2.$$

Hence,

$$\dot{E} = -2(x^2+y^2) + 2(x^2y + xy^2) < -2(x^2+y^2) + 2(x^2+y^2) = 0.$$

Hence, $\dot{E} < 0$. Thus, in every domain D containing $(0,0)$ and such that $x < 1$ and $y < 1$, $\dot{E}(x,y)$ is a negative definite function and hence negative semi-definite.

Therefore, $E = x^2 + y^2$ is a Liapunov function for the given system. ■

Theorem 14.2. Consider the system

$$\begin{aligned}\frac{dx}{dt} &= P(x,y) \\ \frac{dy}{dt} &= Q(x,y).\end{aligned}$$

Assume that this system has an isolated critical point at the origin $(0,0)$ and that P and Q have continuous first order partial derivatives for all (x,y). If there exists a Liapunov function E for the above system in some domain D containing $(0,0)$, then the critical point $(0,0)$ of the above system is stable.

1. If $\dot{E} < 0$ for all $x \neq 0$, then $(0,0)$ is asymptotically stable.
2. If $\dot{E} > 0$ for all $x \neq 0$, then $(0,0)$ is unstable.
3. If $\dot{E} = 0$ for all $x \in \mathbb{R}^2$, then $(0,0)$ is a stable equilibrium point which is not asymptotically stable and solution curves lie on circles centered at the origin.

Use the Liapunov function $v(x) = x_1^2 + x_2^2$ to establish the following results.

1. The origin is an asymptotically stable equilibrium point of
$$\dot{X} = \begin{bmatrix}0 & -1 \\ 1 & 0\end{bmatrix} X + \begin{bmatrix}-x_1^3 - x_1x_2^2 \\ -x_2^3 - x_2x_1^2\end{bmatrix}.$$
2. The origin is an unstable equilibrium point of
$$\dot{X} = \begin{bmatrix}0 & -1 \\ 1 & 0\end{bmatrix} X + \begin{bmatrix}x_1^3 + x_1x_2^2 \\ x_2^3 + x_2x_1^2\end{bmatrix}.$$
3. The origin is a stable equilibrium point which is not asymptotically stable for
$$\dot{X} = \begin{bmatrix}0 & -1 \\ 1 & 0\end{bmatrix} X + \begin{bmatrix}-x_1x_2 \\ x_1^2\end{bmatrix}.$$

Solution. Here, $v(x) = x_1^2 + x_2^2$. Differentiating with respect to time t, we have

$$\dot{v}(x_1, x_2) = 2x_1\dot{x}_1 + 2x_2\dot{x}_2. \tag{14.2.1}$$

1. The system is given by

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_1^3 - x_1x_2^2 \\ \dot{x}_2 &= x_1 - x_2^3 - x_2x_1^2. \end{aligned}$$

From (14.2.1),

$$\begin{aligned} \dot{v}(x_1, x_2) &= 2x_1[-x_2 - x_1^3 - x_1x_2^2] + 2x_2[x_1 - x_2^3 - x_2x_1^2] \\ &= -2x_1x_2 - 2x_1^4 - 2x_1^2x_2^2 + 2x_1x_2 - 2x_2^4 - 2x_2^2x_1^2 \\ &= -2[x_1^4 + x_2^4 + 2x_1^2x_2^2] \\ &= -2(x_1^2 + x_2^2). \end{aligned}$$

Hence $\dot{v}(0, 0) = 0$ and $\dot{v}(x_1, x_2) < 0$ for all $x_1, x_2 \in \mathbb{R}$. Thus the origin is an aysmptotically stable equilibrium point.

2. The system is equivalent to

$$\begin{aligned} \dot{x}_1 &= -x_2 + x_1^3 + x_1x_2^2 \\ \dot{x}_2 &= x_1 + x_2^3 + x_2x_1^2. \end{aligned}$$

Now from (14.2.1),

$$\begin{aligned} \dot{v}(x_1, x_2) &= 2x_1\dot{x}_1 + 2x_2\dot{x}_2 \\ &= 2x_1(-x_2 + x_1^3 + x_1x_2^2) + 2x_2(x_1 + x_2^3 + x_2x_1^2) \\ &= -2x_1x_2 + 2x_1^4 + 2x_1^2x_2^2 + 2x_1x_2 + 2x_2^4 + 2x_1^2x_2^2 \\ &= 2x_1^4 + 2x_2^4 + 4x_1^2x_2^2 \\ &= 2(x_1^2x_2^2)^2. \end{aligned}$$

Hence $\dot{v}(0, 0) > 0$ for all $(x_1, x_2) \in \mathbb{R}^2$. Thus, $(0, 0)$ is unstable critical point.

3. The system is given by

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_1x_2 \\ \dot{x}_2 &= x_1 + x_1^2. \end{aligned}$$

Now from (14.2.1),

$$\begin{aligned} \dot{v}(x_1, x_2) &= 2x_1\dot{x}_1 + 2x_2\dot{x}_2 \\ &= 2x_1(-x_2 - x_1x_2) + 2x_2(x_1 + x_1^2) \\ &= -2x_1x_2 - 2x_1^2x_2 + 2x_1x_2 + 2x_1^2x_2 \\ &= 0. \end{aligned}$$

Thus the origin is stable equilibrium point which is not asymptotically stable.

■

Show that the stable equilibrium point $E^0(1,0)$ pf the SIR epidemic model

$$\begin{aligned}\frac{dS}{dt} &= \mu(1-S) - \beta SI \\ \frac{dI}{dt} &= \beta SI - (\mu+\gamma)I\end{aligned}$$

is globally asymptotically stable if $R_0 = \dfrac{\beta}{\mu+\gamma} < 1$, where S and I are proportions of the susceptibles and infectives at time t respectively. Use the Liapunov function $v = I + S - 1 + \ln S$.

Solution. It is easy to verify that $(1,0)$ is a critical point. Now the Liapunov function is given by

$$v = I + S - 1 + \ln S.$$

Differentiating with respect to time t,

$$\begin{aligned}\frac{dv}{dt} &= \frac{dI}{dt} + \frac{dS}{dt} - \frac{1}{S}\frac{dS}{dt} \\ &= \mu(1-S) - \beta SI + \beta SI - (\mu+\gamma)I - \frac{1}{S}[\mu(1-S) - \beta SI] \\ &= \mu(1-S) - \frac{1}{S}\mu(1-S) + \beta I - (\mu+\gamma)I \\ &= -\frac{\mu(S-1)^2}{S} + I[\beta - (\mu+\gamma)] \\ &= -\frac{\mu(S-1)^2}{S} + (\mu+\gamma)I\left[\frac{\beta}{\mu+\gamma} - 1\right] \\ &= -\frac{\mu(S-1)^2}{S} + (\mu+\gamma)I(R_0 - 1).\end{aligned}$$

Thus, $\dfrac{dv}{dt} < 0 \Rightarrow R_0 - 1 < 0 \Rightarrow R_0 < 1 \Rightarrow \dfrac{\beta}{\mu+\gamma} < 1$. Hence the critical point $E^0(1,0)$ is asymptotically stable if $R_0 = \dfrac{\beta}{\mu+\gamma} < 1$. ■

Unit 15

Course Structure

- Limit cycles and periodic solutions
- Existence and Non-existence of limit cycles
- Bendixon's Non-existence criterion, Dulac's criterion

15.1 Limit Cycles and Periodic solutions

Given an autonomous system

$$\begin{aligned} \frac{dx}{dt} &= P(x,y) \\ \frac{dy}{dt} &= Q(x,y). \end{aligned} \tag{15.1.1}$$

One is often most interested in determining the existence of periodic solution of the system. If $x = f_1(t)$, $y = g_1(t)$, where f_1 and g_1 ate not both constant functions, is a periodic solution of the above system, then the path which the solution defines is a closed path. On the other hand, let C be a closed path of the above system defined by a solution $x = f(t)$, $y = g(t)$, and suppose $f(t_0) = x_0$, $g(t_0) = y_0$. Since C is a closed path, there exists a value $t_1 = t_0 + T$ where $T > 0$, such that $f(t_0) = x_0$, $g(t_0) = y_0$. Now the pair

$$\begin{aligned} x &= f(t+T) \\ y &= g(t+T) \end{aligned}$$

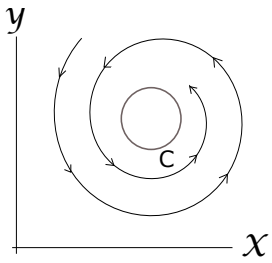
is a solution of (15.1.1). In other words, $f(t+T) = f(t)$, $g(t+T) = g(t)$ for all t, and so the solution $x = f(t)$, $y = g(t)$ defining the closed path C is a periodic solution.

A closed path C of the system (15.1.1) which is approached spirally from either the inside or the outside by a non-closed path C_1 of (15.1.1) either as $t \to +\infty$ or $t \to -\infty$ is called a limit cycle of (15.1.1).

The following example of a system having a limit cycle will illustrate the above discussion and definition.

Consider the following system.

$$\begin{aligned} \frac{dx}{dt} &= y + x(1 - x^2 - y^2) \\ \frac{dy}{dt} &= -x + y(1 - x^2 - y^2). \end{aligned} \tag{15.1.2}$$



To study this system, we shall introduce polar coordinates (r, θ), where

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta. \end{aligned}$$

From these relations, we find that

$$\begin{aligned} x\frac{dx}{dt} + y\frac{dy}{dt} &= r\cos\theta\left[-r\sin\theta\frac{d\theta}{dt} + \frac{dr}{dt}\cos\theta\right] + r\sin\theta\left[r\cos\theta\frac{d\theta}{dt} + \frac{dr}{dt}\sin\theta\right] \\ &= r\frac{dr}{dt}. \end{aligned} \tag{15.1.3}$$

Similarly,

$$x\frac{dy}{dt} - y\frac{dx}{dt} = r^2\frac{d\theta}{dt}. \tag{15.1.4}$$

Now, from (15.1.2),

$$\begin{aligned} & x\frac{dx}{dt} + y\frac{dy}{dt} = (x^2+y^2)(1-x^2-y^2) \\ \Rightarrow\quad & r\frac{dr}{dt} = r^2(1-r^2) \quad \text{[Using (15.1.3)]} \\ \Rightarrow\quad & \frac{dr}{dt} = r(1-r^2). \end{aligned}$$

Again, from (15.1.2),

$$\begin{aligned} & y\frac{dx}{dt} - x\frac{dy}{dt} = y^2 + x^2 \\ \Rightarrow\quad & -r^2\frac{d\theta}{dt} = r^2 \quad \text{[Using (15.1.4)]} \\ \Rightarrow\quad & \frac{d\theta}{dt} = -1. \end{aligned}$$

Thus in polar coordinate system, we have

$$\frac{dr}{dt} = r(1-r^2) \tag{15.1.5}$$

$$\frac{d\theta}{dt} = -1. \tag{15.1.6}$$

Integrating (15.1.6), we have, $\theta = -t + t_0$, t_0 is constant. From (15.1.5),

$$\begin{aligned} & \frac{dr}{r(1-r^2)} = dt \\ \Rightarrow \quad & \frac{r^2 + (1-r^2)}{r(1-r^2)} dr = dt \\ \Rightarrow \quad & \frac{r\,dr}{1-r^2} + \frac{dr}{r} = dt \\ \Rightarrow \quad & \frac{2r\,dr}{1-r^2} + 2\frac{dr}{r} = 2dt. \end{aligned}$$

Integrating, we get

$$\begin{aligned} & \ln r^2 - \ln|1-r^2| = 2t + \ln|C_0| \\ \Rightarrow \quad & \frac{r^2}{1-r^2} = C_0\,\mathrm{e}^{2t} \\ \Rightarrow \quad & r^2 = (1-r^2)C_0\,\mathrm{e}^{2t} \\ \Rightarrow \quad & (1 + C_0\,\mathrm{e}^{2t})r^2 = C_0\,\mathrm{e}^{2t} \\ \Rightarrow \quad & r^2 = \frac{C_0\,\mathrm{e}^{2t}}{1 + C_0\,\mathrm{e}^{2t}} \\ \Rightarrow \quad & r = \frac{1}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}, \end{aligned}$$

where $C = \dfrac{1}{C_0}$. Thus the solution of the system may be written as

$$\begin{aligned} r &= \frac{1}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}, \\ \theta &= -t + t_0, \end{aligned}$$

where C and t_0 are arbitrary constants. We may choose $t_0 = 0$. Then $\theta = -t$, and hence

$$x = \frac{\cos t}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}, \quad \text{and} \quad y = \frac{\sin t}{\sqrt{1 + C\,\mathrm{e}^{-2t}}}. \tag{15.1.7}$$

If $C = 0$, the path defined by (15.1.7) is the circle $x^2 + y^2 = 1$. If $C \neq 0$, the path defined by (15.1.7) are not closed paths but rather paths having a spiral behaviour. If $C > 0$, the paths are spirals lying inside the circle $x^2 + y^2 = 1$. As $t \to \infty$, they approach this circle, while as $t \to -\infty$, they approach the critical point $(0, 0)$. If $C < 0$, the paths lied outside the circle $x^2 + y^2 = 1$.

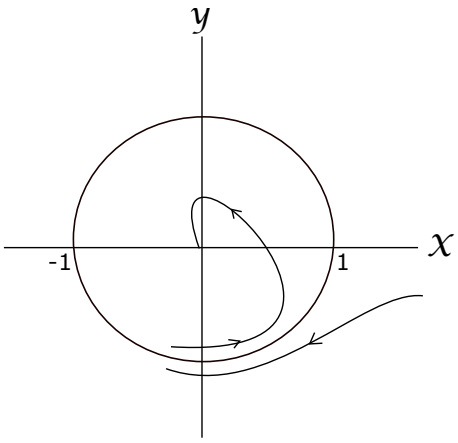
Since the closed path $x^2 + y^2 = 1$ approached spirally, both the inside and outside by non-closed paths as $t \to +\infty$, we conclude that this cycle is a limit cycle of the given system.

15.1.1 Existence and Non-existence of Limit cycles

Bendixon's Non-existence criterion

Let D be a domain in the xy-plane. Consider the autonomous system

$$\begin{aligned} \frac{dx}{dt} &= P(x, y) \\ \frac{dy}{dt} &= Q(x, y) \end{aligned} \tag{15.1.8}$$



where P and Q have continuous first order partial derivatives in D. Suppose that $\frac{\partial P(x,y)}{\partial x} + \frac{\partial Q(x,y)}{\partial y}$ has the same sign throughout D. Then the system (15.1.8) has no closed path in the domain D.

Proof. Let C be a closed curve in D. Let R be the region bounded by C and apply Green's theorem in the plane. We have

$$\int_C [P(x,y)dy - Q(x,y)dx] = \int\int_R \left[\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right] dxdy$$

where the line integral is taken in the positive sense. Now assume that C is a closed path of (15.1.8). Let $x = f(t)$, $y = g(t)$ be an arbitrary solution of (15.1.8), defining C parametrically and let T denote the period of this solution. Then

$$\frac{df(t)}{dt} = P[f(t), g(t)] \quad \text{and} \quad \frac{dg(t)}{dt} = Q[f(t), g(t)]$$

along C and we have

$$\begin{aligned}
\int_C [P(x,y)dy - Q(x,y)dx] &= \int_0^T \left\{ P[f(t),g(t)]\frac{dg(t)}{dt} - Q[f(t),g(t)]\frac{df(t)}{dt} \right\} dt \\
&= \int_0^T \{P[f(t),g(t)]Q[f(t),g(t)] - Q[f(t),g(t)]P[f(t),g(t)]\}\, dt \\
&= 0.
\end{aligned}$$

Thus,

$$\int\int_R \left[\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}\right] dx\, dy = 0.$$

But this double integral can be zero only if $\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}$ changes sign. This is a contradiction. Thus C is not a path of (15.1.8) and hence (15.1.8) possesses no closed path in D. □

Show that the following system has no closed path

$$\begin{aligned}
\frac{dx}{dt} &= 2x + y + x^3 \\
\frac{dy}{dt} &= 3x - y + y^3.
\end{aligned}$$

Solution. Here,

$$\begin{aligned} P(x,y) &= 2x + y + x^3 \\ Q(x,y) &= 3x - y + y^3. \end{aligned}$$

Now,

$$\frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y} = 3(x^2 + y^2) + 1.$$

Since this expression is positive throughout every domain D in the xy-plane, the given system has no closed path in any such domain. ■

By constructing a Liapunov function, show that the system

$$\begin{aligned} \dot{x} &= -x + 4y \\ \dot{y} &= -x + y^3 \end{aligned}$$

has no closed orbit.

Solution. Consider $v(x,y) = x^2 + ay^2$, where a is a parameter to be chosen later. Then

$$\begin{aligned} \dot{v} &= 2x\dot{x} + 2ay\dot{y} \\ &= 2x(-x + 4y) + 2ay(-x + y^3) \\ &= -2x^2 + (8 - 2a)xy - 2ay^4. \end{aligned}$$

If we choose $a = 4$, the xy term disappears and

$$\dot{v} = -2x^2 - 8y^4.$$

By inspection, $v > 0$ and $\dot{v} < 0$ for all $(x,y) \neq (0,0)$. Hence, $v = x^2 + y^2$ is u is a Liapunov function and so there are no closed orbits. In fact, all trajectories approach the origin as $t \to \infty$. ■

1. Show that the system $\dot{x} = y - x^3$, $\dot{y} = -x - y^3$ has no closed orbit, by constructing a Liapunov function $v = ax^2 + by^2$ with a suitable a, b.
2. Show that $v = ax^2 + 2bxy + cy^2$ is positive definite if and only if $a > 0$ and $ac - b^2 > 0$.
3. Show that $\dot{x} = -x + 2y^3 - 2y^4$, $\dot{y} = -x - y + xy$ has no periodic solution. [Hint: Choose a, m and n such that $v = x^m + ay^n$ is a Liapunov function]

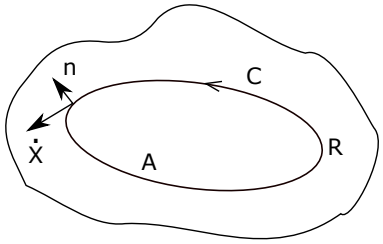
15.1.2 Dulac's Criterion

This is a method for ruling out closed orbits is based on Green's theorem, and is known as Dulac's criterion.

Theorem 15.2. Let $\dot{x} = f(x)$ be a continuously differentiable vector field defined on a simply connected subset of R of the plane. If there exists a continuously differentiable real valued function $g(x)$ such that $\nabla \cdot (g\dot{x})$ has one sign throughout R, then there are no closed orbits lying entirely in R.

Proof. Suppose there were a closed orbit C lying entirely in the region R. Let A denote the region inside C. Then Green's theorem yields

$$\int\int_A \nabla \cdot (g\dot{x}) dA = \oint_C g\dot{x} \cdot \eta dl$$



where η is the outward normal and dl is the element of arc length along C. Since $\nabla \cdot (g\dot{x})$ has one sign in R, hence the double integral on the left side must be non-zero. On the other hand, the line integral on the right equals zero. Since $\dot{x} \cdot \eta = 0$ everywhere, by assumption that C is a trajectory (the tangent vector $\dot{x}$ is orthogonal to η). This contradiction implies that no such C can exist. □

Dulac's criterion suffers from the same drawback as Liapunov's method; there is no algorithm for finding $g(x)$. Most commonly used $g(x)$ are

$$g = 1,\ \frac{1}{x^\alpha y^\beta},\ \mathrm{e}^{ax},\ \text{ and }\ \mathrm{e}^{ay}.$$

Show that the system $\dot{x} = x(2 - x - y)$, $\dot{y} = y(4x - x^2 - 3)$ has no closed orbit om the positive quadrant $x, y > 0$.

Solution. Let us choose $g = \dfrac{1}{xy}$. Then

$$\begin{aligned}
\nabla \cdot (g\dot{x}) &= \frac{\partial}{\partial x}(g\dot{x}) + \frac{\partial}{\partial y}(g\dot{y}) \\
&= \frac{\partial}{\partial x}\left(\frac{2 - x - y}{y}\right) + \frac{\partial}{\partial y}\left(\frac{4x - x^2 - 3}{x}\right) \\
&= -\frac{1}{y} < 0.
\end{aligned}$$

Since the region $x, y > 0$ is simply connected and g and f satisfy the required smoothness conditions. Hence Dulac's criterion implies that there are no closed orbits in the positive quadrant. ■

Show that the system $\dot{x} = y$, $\dot{y} = -x - y - x^2 + y^2$ has no closed orbits.

Solution. Let $g = \mathrm{e}^{-2x}$. Then

$$\nabla \cdot (g\dot{x}) = -2\,\mathrm{e}^{-2x}\,y + \mathrm{e}^{-2x}(1 - 2y) = -\,\mathrm{e}^{-2x} < 0.$$

By Dulac's criterion, there are no closed orbits. ■

1. Using Dulac's criterion with weight function $g = (N_1 N_2)^{-1}$, show that the system

$$\begin{aligned}
\dot{N}_1 &= r_1 N_1 \left(1 - \frac{N_1}{K_1}\right) - b_1 N_1 N_2 \\
\dot{N}_2 &= r_2 N_2 \left(1 - \frac{N_2}{K_2}\right) - b_2 N_1 N_2
\end{aligned}$$

has no periodic orbits in the first quadrant $N_1, N_2 > 0$.

2. Using Dulac's criterion, show that the system

$$\begin{aligned} \dot{x} &= -x + y^2 \\ \dot{y} &= y(2 + 2x - y^2) \end{aligned}$$

has no closed orbits. You may use a Dulac's function $g(x, y) = \dfrac{1}{y}$.

3. Use the Dulac's function $B(x, y) = b\,\mathrm{e}^{-2\beta x}$ to show that the system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -ax - by + \alpha x^2 + \beta y^2 \end{aligned}$$

has no limit cycle in $\mathbb{R}^2$.

Unit 16

Course Structure

- Bifurcation

16.1 Bifurcation

The dynamics of vector fields is very limited. All solutions either settle down to equilibrium or head out to $\pm\infty$. The most interesting fact of a dynamical system is the parametric dependence. Mathematical models often rise to differential equations that have many parameters. When the parameter values are changed, we may expect a change in the behaviour of the solution of the differential equation. If the variation of a parameter changes the qualitative behaviour of the solution, we call it bifurcation.

For example, consider the equation for linear growth or linear decay.

$$x' = \mu x.$$

If $\mu > 0$, solution grows exponentially; if $\mu < 0$, all solutions tend to zero.

The qualitative behaviour of solutions for $\mu < 0$ and $\mu > 0$ are quite different, whereas the behaviour of solution for $\mu = 1$ and $\mu = 2$ are very similar. For this example, $\mu = 0$ is a bifurcation value.

To understand a mathematical model properly, it is important to know when and how a bifurcation occurs. In this unit, we introduce four common bifurcations that occur at the equilibria.

We consider a scalar differential equation depending on a scalar parameter

$$x' = f(x, \mu),\ x \in \mathbb{R},\ \mu \in \mathbb{R},$$

where μ is the parameter, and $f : \mathbb{R}^2 \to \mathbb{R}$ is continuously differentiable.

We say that x^* is a bifurcation point and μ^* is a bifurcation value if

$$f(x^*, \mu^*) = 0, \quad \text{and} \quad \frac{\partial}{\partial x} f(x^*, \mu^*) = 0,$$

where $\dfrac{\partial}{\partial x}$ denotes the partial derivative with respect to x. Note that, $f(x^*, \mu^*) = 0$ implies x^* is a steady state of the differential equation $x' = f(x, \mu^*)$. We know that, x^* is a hyperbolic steady state if $f_x(x^*, \mu^*) \neq 0$. Thus, bifurcation points must be non-hyperbolic steady states.

We now discuss the normal forms of the four most common bifurcations. The first three (saddle-node, transcritical and pitchfork) can be exhibited in scalar equations. The Hopf bifurcation can occur in system having dimension atleast 2.

16.1.1 Saddle-Node Bifurcation

The saddle-node bifurcation is the basic mechanism by which fixed points are created and destroyed. As a parameter is varied, two fixed points move towards each other, collide and mutually annihilate.

The prototypical example of a saddle-node bifurcation is given by the first order system

$$\dot{x} = r + x^2$$

where r is a parameter, which may be positive, negative or zero. When r is negative, there are two fixed points, one is stable and one unstable.

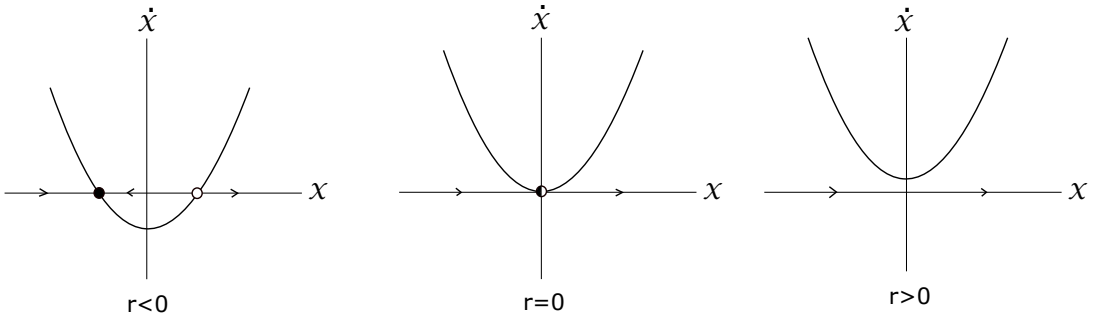


Figure 16.1.1: Saddle-Node Bifurcation

- As r approaches 0 from below, the parabola moves up and the two fixed points move towards each other.
- When $r = 0$, the fixed points coalesce into a half stable fixed point at $x^* = 0$. This type of fixed point is extremely delicate: it vanishes as soon as $r \to 0$ and now there are no fixed point at all.

In this case, we say that a bifurcation occurred at $r = 0$. Since the vector field for $r < 0$ and $r > 0$ are qualitatively different.

Graphical Conventions

We now show a stack of vector fields for discrete values of r. This representation emphasizes the dependence of the fixed points on R. In the limit of a continuous stack of vector fields, we have a picture like figure 16.1.2. The curve shown is $r = -x^2$, that is, $\dot{x} = 0$, which gives the fixed points for different r. To distinguish between stable and unstable fixed points, we use a solid line for fixed points and a broken line for unstable ones.

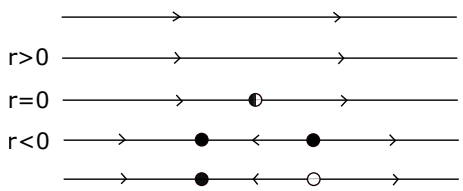


Figure 16.1.2

The most common way to depict the bifurcation is to invert the axis of Figure 16.1.3. The rationale is that r plays the role of an independent variable, and so should be plotted horizontally (Figure 16.1.4). The drawback is that now the x-axis has to be plotted vertically, which looks strange at first. Arrows are sometimes included in the graph, but not always. This picture is called the bifurcation diagram for the saddle-node bifurcation.

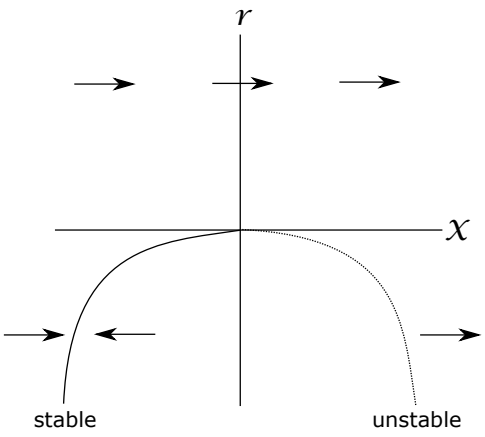


Figure 16.1.3

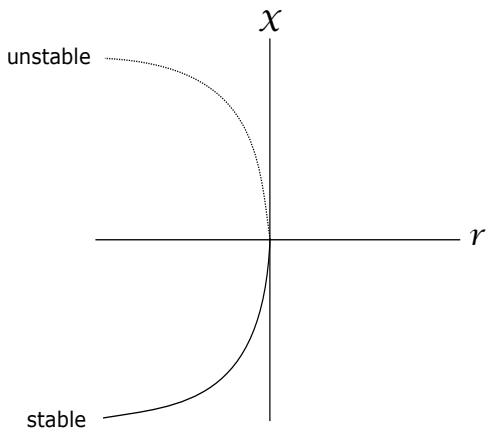


Figure 16.1.4

Show that the first order system $\dot{x} = r - x - \mathrm{e}^{-x}$ undergoes a saddle-node bifurcation as r varied, and find the value of r at the bifurcation point.

Solution. Using the Taylor series expansion for e^{-x} about $x = 0$, we have

$$\mathrm{e}^{-x} = 1 - x + \frac{x^2}{2!} - \dots$$

Now,

$$\begin{aligned} \dot{x} &= r - x - \mathrm{e}^{-x} \\ &= r - x - \left[1 - x + \frac{x^2}{2!} - \dots\right] \\ &= (r-1) - \frac{x^2}{2!} + \dots \end{aligned}$$

If we ignore the higher order terms, then we have

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

This is equivalent to the normal form of saddle-node bifurcation, $\dot{x} = r + x^2$. Thus, the given system undergoes a saddle-node bifurcation. The bifurcation point is given by

$$r - 1 = 0 \Rightarrow r = 1.$$

Differentiating partially with respect to x, we have

$$\frac{\partial f}{\partial x} = -1 + \mathrm{e}^{-x}.$$

Hence the critical point is given by

$$\begin{aligned}\frac{\partial f}{\partial x} = 0 \quad &\Rightarrow \quad -1 + \mathrm{e}^{-x} = 0 \\ &\Rightarrow \quad \mathrm{e}^{-x} = 1 \\ &\Rightarrow \quad -x \ln|\,\mathrm{e}\,| = \ln(1) \\ &\Rightarrow \quad -x = 0 \\ &\Rightarrow \quad x = 0.\end{aligned}$$

Thus the critical point is $x^* = 0$ and bifurcation point is given by $r^* = 1$. ■

16.1.2 Transcritical Bifurcation

There are certain situations where a fixed point exists for all values of a parameter and can never be destroyed. However, such a fixed point may change its stability as the parameter is varied. The transcritical bifurcation is the standard mechanism for such changes in stability. The normal form for a transcritical bifurcation is

$$\dot{x} = rx - x^2.$$

The following figure shows the vector field as r varies. Note that there is a fixed point at $x^* = 0$ for all values of r.

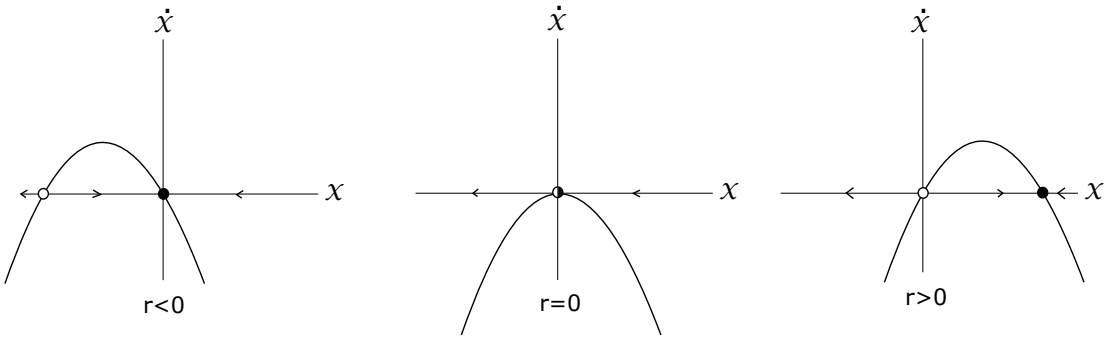


Figure 16.1.5: Transcritical Bifurcation

For $r < 0$, there is an unstable fixed point at $x^* = r$ and a stable fixed point at $x^* = 0$. As r increases, the unstable fixed point approaches the origin and coalesces with it when $r = 0$. Finally, when $r > 0$, the origin has become unstable and $x^* = r$ is now stable. Thus an exchange of stability conditions has taken place between the two fixed points.

The important difference between the saddle-node and transcritical bifurcations is that the two fixed points don't disappear after bifurcation; instead they just switch their stability.

Figure 16.1.6 shows the bifurcation diagram for the transcritical bifurcation.

Show that the first order system

$$\dot{x} = x(1 - x^2) - a(1 - \mathrm{e}^{-bx})$$

undergoes a transcritical bifurcation at $x = 0$ when the parameters a, b satisfy a certain equation to be determined.

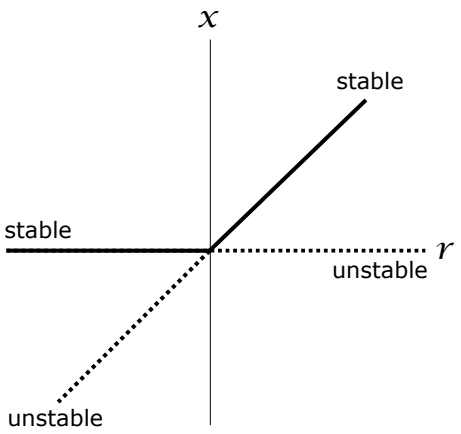


Figure 16.1.6: Bifurcation Diagram

Solution. Here, $x = 0$ is a fixed point for all $a,\ b$. For small x, we find,

$$\begin{aligned} 1 - \mathrm{e}^{-bx} &= 1 - \left[1 - bx + \frac{1}{2}b^2x^2 + O(x^3)\right] \\ &= bx - \frac{1}{2}b^2x^2 + O(x^3). \end{aligned}$$

Thus,

$$\begin{aligned} \dot{x} &= x - a\left(bx - \frac{1}{2}b^2x^2\right) + O(x^3) \\ &= (1 - ab)x + \frac{1}{2}b^2x^2 + O(x^3). \end{aligned}$$

Hence, the transcritical bifurcation occurs when $1 - ab = 0 \Rightarrow ab = 1$. This equation represents the equation of bifurcation curve. The non-zero critical point for small x is given by

$$(1 - ab) + \frac{1}{2}b^2x^* \simeq 0 \Rightarrow x^* \simeq \frac{2(ab - 1)}{ab^2}.$$

■

Show that the system

$$\dot{x} = r\ln(x) + x - 1$$

undergoes a transcritical bifurcation at a certain value of r.

Solution. Here $f(x) = r\ln(x) + x - 1$. Now, $f(1) = 0$. Hence, $x = 1$ is a critical point for all values of r. Since we are interested in the dynamics near the fixed point, we introduce a new variable $u = x - 1$, where u is very small. Then

$$\begin{aligned} \dot{u} = \dot{x} &= r\ln(u + 1) + u \\ &= r\left[u - \frac{1}{2}u^2 + O(u^3)\right] + u \\ &= (r + 1)u - \frac{1}{2}ru^2 + O(r^3). \end{aligned}$$

Hence the transcritical bifurcation occurs at $r + 1 = 0 \Rightarrow r = -1$. ■

16.1.3 Pitchfork Bifurcation

Here we discuss the third type of bifurcation, the so called pitchfork bifurcation. This bifurcation is common in physical problems that have a symmetry. There are two very different types of pitchfork bifurcation, namely supercritical bifurcation and subcritical bifurcation.

Supercritical Pitchfork Bifurcation

The normal form of the supercritical pitchfork bifurcation is

$$\dot{x} = rx - x^3. \tag{16.1.1}$$

This equation is invariant under the change of variable $x \to -x$. That is, if we replace x by $-x$ and then cancel the resulting minus sign on both sides of the equation, we get equation (16.1.1) again. This invariance is the mathematical expression of the left right symmetry.

The following figure shows the vector field for different values of r.

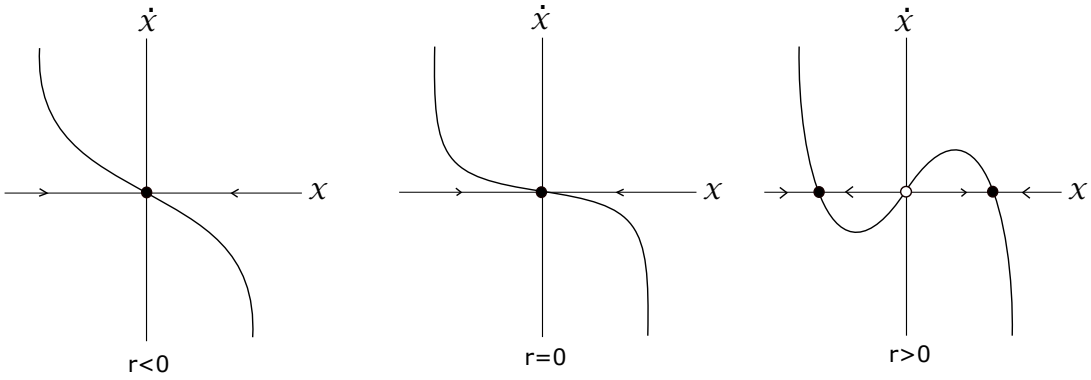


Figure 16.1.7: Supercritical Pitchfork Bifurcation

When $r < 0$, the origin is the only fixed point, and it is stable. When $r = 0$, the origin is still stable, but much weakly so, since linearization vanishes. Finally, when $r > 0$, the origin has become unstable. Two new stable fixed points appear on either side of the origin, symmetrically located at $x^* = \pm\sqrt{r}$.

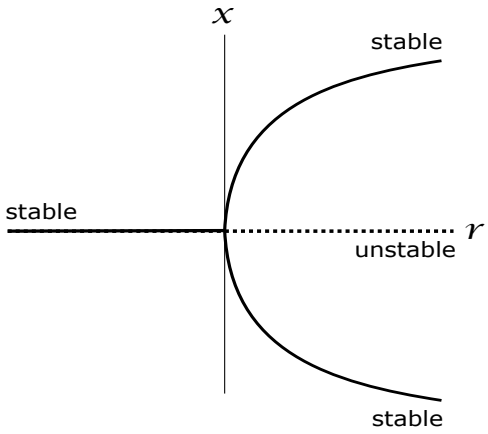


Figure 16.1.8: Bifurcation diagram for pitchfork bifurcation

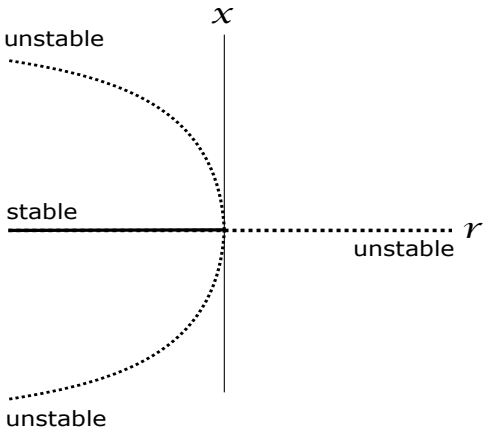


Figure 16.1.9: Bifurcation diagram for subcritical pitchfork bifurcation

Subcritical pitchfork bifurcation

The normal form of subcritical pitchfork bifurcation is given by

$$\dot{x} = rx + x^3.$$

16.2 Hopf Bifurcation

A Hopf Bifurcation occurs when a periodic solution or limit cycle, surrounding an equilibrium point, arises or goes away as a parameter varies. When a stable limit cycle surrounds an unstable equilibrium point, the bifurcation is called a *supercritical Hopf bifurcation*. If the limit cycle is unstable and surrounds a stable equilibrium point, then the bifurcation is called a *subcritical Hopf bifurcation.*

Theorem 16.3. Hopf Bifurcatoin Theorem: Consider the planar system

$$\begin{aligned} \dot{x} &= f_\mu(x, y), \\ \dot{y} &= g_\mu(x, y), \end{aligned} \tag{16.3.1}$$

where μ is a parameter. Suppose it has a fixed point, which without loss of generality we may assume to be located at $(x, y) = (0, 0)$. Let the eigenvalues of the linearized system about the fixed point be given by $\lambda(\mu), \overline{\lambda}(\mu) = \alpha(\mu) \pm i\beta(\mu)$. Suppose further that for a certain value of μ (which we may assumed to be 0) the following conditions are satisfied:

1. $\alpha(0) = 0,\ \beta(0) = \omega \neq 0$, where $\text{sgn}(\omega) = \text{sgn}\left[\left(\dfrac{\partial g_\mu}{\partial x}\right)\Big|_{\mu=0}(0,0)\right]$ (non-hyperbolicity condition: conjugate pair of imaginary eigenvalues)
2. $\dfrac{d\alpha(\mu)}{d\mu}\Big|_{\mu=0} = d \neq 0$ (transversality condition: the eigenvalues cross the imaginary axis with non-zero speed)
3. $a \neq 0$, where

$$\begin{aligned} a = \frac{1}{16}\left(f_{xxx} + f_{xyy} + g_{xxy} + g_{yyy}\right) + \frac{1}{16\omega}\left[f_{xy}(f_{xx} + f_{yy})\right. \\ \left. -g_{xy}(g_{xx} + g_{yy}) - f_{xx}g_{xx} + f_{yy}g_{yy}\right] \end{aligned}$$

with $\left[\left(\dfrac{\partial^2 f_\mu}{\partial_x \partial_y}\right)\Big|_{\mu=0}(0,0)\right]$, etc. (genericity condition)

Then a unique curve of periodic solutions bifurcates from the origin into the region $\mu > 0$ if $ad < 0$ or $\mu < 0$ if $ad > 0$. The origin is a stable fixed point for $\mu > 0$ (resp. $\mu < 0$) and an unstable fixed point for $\mu < 0$ (resp. $\mu > 0$) if $d < 0$ (resp. $d > 0$) whilst the periodic solutions are stable (resp. unstable) if the origin is unstable (resp. stable) on the side of $\mu = 0$ where the periodic solutions exist. The amplitude of the periodic orbits grows like $\sqrt{|\mu|}$ whilst their periods tend to $2\pi/|\omega|$ as $|\mu|$ tends to zero.

Illustration: Consider the two dimensional system

$$\begin{aligned} x_1' &= -x_2 + x_1(\mu - x_1^2 - x_2^2) \\ x_2' &= x_1 + x_2(\mu - x_1^2 - x_2^2). \end{aligned} \tag{16.3.2}$$

Using polar coordinates,

$$x_1 = r\cos\theta \quad \text{and} \quad x_2 = r\sin\theta.$$

We can rewrite the system (16.3.2) as

$$\begin{aligned} r' &= r(\mu - r^2) \\ \theta' &= 1. \end{aligned} \tag{16.3.3}$$

Note that the equation for r in (16.3.3) is the normal form for a pitchfork bifurcation. Thus as μ passes through the bifurcation value 0, the system (16.3.3) undergoes a pitchfork bifurcation.

The steady state $\overline{r} = 0$ corresponds to the steady state $(0, 0)$ while the other steady state $\overline{r} = \sqrt{\mu}$, corresponds to a periodic orbit

$$\sqrt{x_1^2 + x_2^2} = \sqrt{\mu}.$$

The corresponding bifurcation diagram is shown in the figure below.

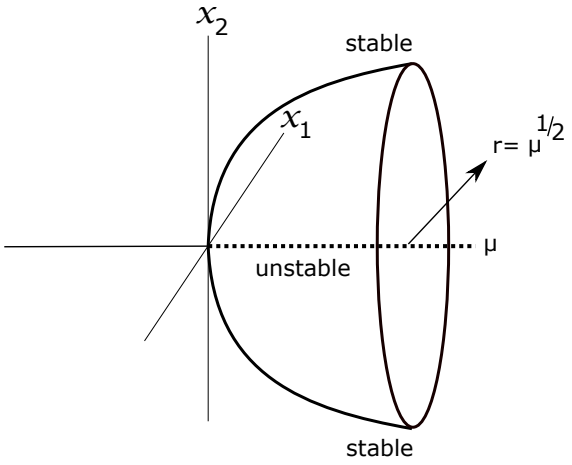


Figure 16.3.1: Hopf bifurcation diagram

Note that the Jacobian matrix $Df(0, 0)$ is given by

$$\begin{bmatrix} \mu & -1 \\ 1 & \mu \end{bmatrix}$$

which has a pair of complex eigen values, namely $\lambda = \mu \pm i$. At the bifurcation value $\overline{\mu} = 0$, the eigen values are purely imaginary. The occurance of purely imaginary eigen values for a set of parameter values is an important indicator of Hopf bifurcation.

Perform a bifurcation analysis for the Li$\acute{e}$nard equation

$$\ddot{x} - (\mu - x^2)\dot{x} + x = 0$$

If we let $u = x,\ v = \dot{x}$, we can rewrite the equation as a two-dimensional first order system

$$\begin{aligned} \dot{u} &= v \\ \dot{v} &= -u + (\mu - u^2)v \end{aligned}$$

The only equilibrium point is the origin. The Jacobian matrix for the linearized system about the origin is

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

The eigenvalues of the Jacobian matrix are

$$\alpha(\mu) + \beta(\mu) = \frac{\mu}{2} \pm i\sqrt{4 - \frac{\mu^2}{2}}.$$

Notice that

$$\alpha(0) = 0 \quad \text{and} \quad \omega = \beta(0) = -1.$$

Also,

$$d = \frac{d\alpha(\mu)}{d\mu}\Big|_{\mu=0} = \frac{1}{2} \neq 0.$$

Lastly, $a = -\frac{1}{8} \neq 0$. Hence, all the conditions of the Hopf Bifurcation Theorem are satisfied. Since $ad = -\frac{1}{16} < 0$, the origin is stable for $\mu < 0$ (see Fig. 16.3.2) and unstable for $\mu > 0$, where there is a stable periodic orbit (see Fig. 16.3.3). The system has a supercritical Hopf bifurcaton at $\mu = 0$.

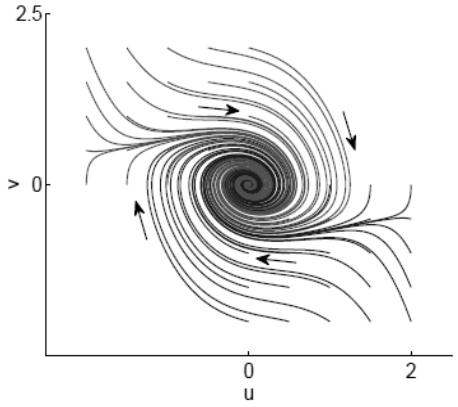


Figure 16.3.2: The origin is stable focus $\mu = -0.3$

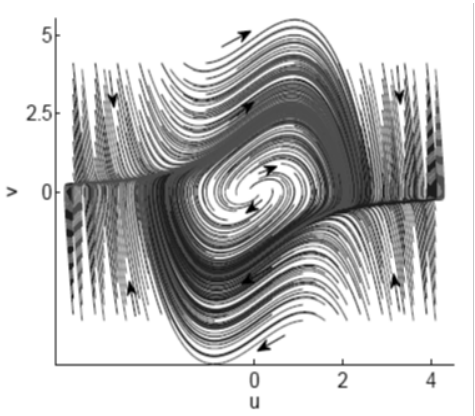


Figure 16.3.3: The origin is unstable focus $\mu = 1$

Perform a bifurcation analysis for the following logistic model.

$$\dot{x} = rx(1 - x) - h$$

where $r > 0$ is the rate of logistic growth and h is a harvesting component, say the amount of fishing allowed in a lake. If h is positive or the amount of stocked fish added to the lake per year if h is negative. Here, $f(x, r, h) = rx(1 - x) - h$. For critical point,

$$\begin{aligned} & f(x^*, r, h) = 0 \\ \Rightarrow \quad & rx^*(1 - x^*) - h = 0 \\ \Rightarrow \quad & r\left(x^*\right)^2 - rx^* + h = 0 \\ \Rightarrow \quad & x^* = \frac{r \pm \sqrt{r^2 - 4rh}}{2r} = \frac{1}{2}\left[1 \pm \sqrt{1 - \frac{4h}{r}}\right]. \end{aligned}$$

Letting $\mu = \frac{4h}{r}$, we have the critical points

$$x^* = \frac{1}{2}(1 \pm \sqrt{1 - \mu}).$$

When $\mu < 1$, we have two equilibrium points. When $\mu = 1$, we have only one equilibrium point. When $\mu > 1$, there is no equilibrium point.

In order to determine the stability, we need to look at the derivative of $f(x, r, h)$. Thus,

$$\frac{d}{dx}f(x, r, h) = -2rx + r = -r(2x - 1).$$

Now,

$$\frac{d}{dx}f(x^*, r, h) = -r[1 \pm \sqrt{1 - \mu} - 1] = \mp r\sqrt{1 - \mu}.$$

Since $r > 0$ when $\mu < 1$, $x^* = \frac{1}{2}\left[1 + \sqrt{1 - \frac{4h}{r}}\right]$ is stable while $x^* = \frac{1}{2}\left[1 - \sqrt{1 - \frac{4h}{r}}\right]$ is unstable. As μ increased towards 1, the two equilibria moves towards each other, eventually colliding each other. This point in the $(\mu; x^*)$ plane $\left(1, \frac{1}{2}\right)$ is called the bifurcation point. A qualitative bifurcation diagram is as follows.

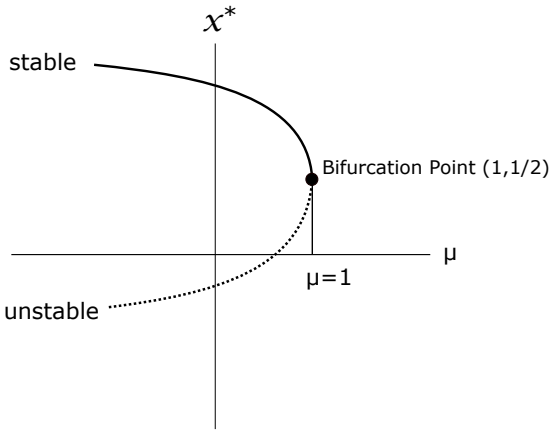


Figure 16.3.4: Bifurcation diagram

Show that the system

$$\frac{dx}{dt} = x(1 - x) - h\frac{x}{a + x}$$

can have one, two or three fixed points, depending on the values of a and h.

- Analyse the dynamics near $x = 0$ and show that a bifurcation occurs when $h = a$. What type of bifurcation is it?
- Show that another bifurcation occurs when $h = \frac{1}{4}(a + 1)^2$, for $a < a_c$, where a_c is to be determined. Classify this bifurcation.

References

1. D.M. Burton, *The History of Mathematics*, Allyn and Bacon, 5th edition.
2. Carl B. Boyer and Uta C. Merzbach, *A History of Mathematics*, 3rd Edition .
3. Florian Cajori , *A History of Mathematics*.
4. J.H. Eves, *An Introduction to the History of Mathematics*, Saunders, 1990.
5. H.A. Taha: *Operations Research*
6. J.G. Chakraborty and P.R. Ghosh: *Linear Programming and Game Theory*
7. P.K. Gupta and D.S. Hira: *Operations Research*
8. K. Swarup, P. K. Gupta and Man Mohan: *Operations Research.*
9. I. N. Herstein, *Topics in Algebra*.
10. K.Hoffman and R. Kunze, *Linear Algebra*.
11. B. C. Chatterjee, *Linear Algebra*.
12. L. Perko: *Differential Equations and Dynamical Systems, Springer Verlag.*
13. F. Verhulust, *Nonlinear Differential Equations and Dynamical Systems, Springer.*
14. S.H. Strogatz, *Nonlinear Dynamics and Chaos.*
15. M. Lakshmanan, S. Rajasekar, *Nonlinear Dynamics-Integrability, Chaos and Patterns.*